University of London
Imperial College of Science, Technology and Medicine
Department of Computing

Affect Recognition & Generation in-the-wild

Dimitrios Kollias

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of the University of London and
the Diploma of Imperial College, December, 2020

Abstract

Affect recognition based on a subject's facial expressions has been a topic of major research in the attempt to generate machines that can understand the way subjects feel, act and react. In the past, due to the unavailability of large amounts of data captured in real-life situations, research has mainly focused on controlled environments. However, recently, social media and platforms have been widely used. Moreover, deep learning has emerged as a means to solve visual analysis and recognition problems. This Ph.D. Thesis exploits these advances and makes significant contributions for affect analysis and recognition in-the-wild.

We tackle affect analysis and recognition as a dual knowledge generation problem: i) we create new, large and rich in-the-wild databases and ii) we design and train novel deep neural architectures that are able to analyse affect over these databases and to successfully generalise their performance on other datasets.

At first, we present the creation of Aff-Wild database annotated according to valence-arousal and an end-to-end CNN-RNN architecture, AffWildNet. Then we use AffWildNet as a robust prior for dimensional and categorical affect recognition and extend it by extracting low-/mid-/high-level latent information and analysing this via multiple RNNs. Additionally, we propose a novel loss function for DNN-based categorical affect recognition.

Next, we generate Aff-Wild2, the first database containing annotations for all main behavior tasks: estimate Valence-Arousal; classify into Basic Expressions; detect Action Units. We develop multi-task and multi-modal extensions of AffWildNet by fusing these tasks and propose a novel holistic approach that utilises all existing databases with non-overlapping annotations and couples them through co-annotation and distribution matching.

Finally, we present an approach for valence-arousal, or basic expressions' facial affect synthesis. We generate an image with a given affect, or a sequence of images with evolving affect, by annotating a 4-D database and utilising a 3-D morphable model.

Acknowledgements

I would like to thank my supervisor, Prof. Stefanos Zafeiriou for his invaluable help and guidance over these years and especially for always being there to support my enthusiasm in exploring new research avenues.

I would also like to thank Dr. Viktoriia Sharmanska for our fruitful collaboration and conversations, the guidance and all the help that she provided me with.

My warmest euharistie go to Dr. Krysia Broda for all her invaluable help and guidance as a Teaching Scholar Mentor.

I also feel extremely obliged to Dr. Amani El-Kholy for her valuable support (and timely responses to all my requests) in overcoming any obstacles that I met, including problems with this thesis submission.

I also feel extremely obliged to Ms. Teresa Ng for being always eager to take care of my administrative tasks.

I would also like to thank Konstantinos Gkoutzis for our collaboration in multiple courses taught at the Department.

I also need to thank all people in CSG and especially Loyd, Geoff and Duncan, for always being eager to help any hardware related problem I faced.

I thank Dr Grigorios Chrysos for the fruitful conversations and help in my first Ph.D. year, Evangelos Ververas for the assistance and help in a couple of papers, along with the help in the teaching load when needed (being the second/external marker) and Markos Georgopoulos for the many and of all kind conversations that we had over the years (being in near office tables).

I also thank Dr. Mihalis Nicolaou and Dr George Trigeorgis for their help in my first Ph.D. year.

I thank all my colleagues for any dialectic encounters, interactions and collaborations. I would like to express my gratitude to my close family and close friends, for supporting me and putting up with me.

Last but not at all least, I would like to especially thank my parents, Stefanos and Stavroula / Stavroula and Stefanos, as well as my sister, Ilianna, for constantly being an omnipotent core of love and support.

Dedication

This thesis is dedicated to pappous, who got too tired waiting for the completion of this thesis and decided to get a big rest instead.

‘Woodland, Chapter, C12A, C12F, B81E, Caledonian, Flat 1, Holloway, Football iBUG, trikala memories, chapteromania, exodoi trelo pareaki woodland, ellinofwnoi tou woodland, to klouvi me tis treles/trelous, basket caledonian, punch-lincoln, cut-finger, Greek OG, popular princess/devil/parthenopi, que serra serra, gizo, Palermo Experts, atlas, parties, deep render, brexit, covid19, 2 lockdowns’

London, M.Sc. & Ph.D. 2015-2020

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In this Thesis we deal with the problem of affect analysis and recognition, which constitutes a key issue in behavioural modelling, human machine interaction and affective computing. There are a number of related applications spread across a variety of fields, such as medicine, health, or driver fatigue, monitoring, e-learning, marketing, entertainment, lie detection and law. Some examples of these application fields are referenced below:

- determining patients' feeling and comfort level about treatment
- remote monitoring of elderly people's health
- detecting facial expressions and determining fatigue of a car driver; alerting drivers if they look sleepy or drowsy
- studying learners' emotions and adjusting learning techniques according to learners' reactions
- in call centers, determining anger and stress levels in the voice and prioritizing angry callers
- detecting positive or negative reactions of public in events
- detecting emotional state of candidates in interviews

- detecting facial expressions while playing a videogame can be a good metric to understand if the game is successful in making the experience enjoyable.

However, human facial expression and affect recognition constitutes a difficult problem, because emotion patterns are complex, time varying, user and context dependent. Due to this fact, affect analysis constitutes an open research problem, which has attracted great interest by researchers internationally.

Ekman [61] was the first to systematically study human facial expressions. His study categorizes the prototypical facial expressions, apart from neutral expression, into six classes representing anger, disgust, fear, happiness, sadness and surprise. This categorization is consistent across different ethnicities and cultures. Furthermore, facial expressions are related to specific movements of facial muscles, called Action Units (AUs). The Facial Action Coding System (FACS) was developed, in which facial changes are described in terms of AUs [41].

Apart from the above categorical definition of facial expressions and related emotions, in the last few years there has been great interest in dimensional emotion representations, which are of great interest in human computer interaction and human behaviour analysis. Dimensional emotion representations are used to tag emotional states in continuous mode, usually in terms of the arousal and valence dimensions, i.e. in terms of how active or passive, positive or negative is the human behaviour under analysis [67].

Moreover, differentiating between posed and spontaneous facial expressions is advantageous in many areas of human-computer interaction, or public security. Many annotated facial databases exist, which show human actors portraying facial expressions. A classification of these databases can be based on whether the basic annotation refers to categorical emotion categories, or to dimensional emotion representations, or to action units. However, existing databases have significant limitations: they contain data recorded in laboratory or controlled environments; their diversity is limited due to the small total number of subjects they contain, due to the limited variation in scene lighting, camera view, image resolution, background, subjects' head-pose and ethnicity, due to the lack of occlussions and due to the fact that the total duration of their included videos is rather short. Generating large databases showing spontaneous behaviours in real-life, uncontrolled, i.e., in-the-wild environments, overcoming these limitations has been a strong motivation for our work.

In addition, human emotion states in-the-wild do not have explicit temporal boundaries and their patterns often vary across individuals and contexts. Hence, existing affect recognition systems lack enough generality when used in uncontrolled human computer interaction. This makes necessary the development of new effective affect recognition systems which are able to operate across different real life environments. This has been another strong motivation for our work.

Major research has been given during the last few years to the development and use of deep learning techniques and deep neural networks [74, 125] in various applications, including affect recognition in-the-wild. The Emotion Recognition in the Wild Challenge (EmotiW), as well as the Challenges in Representation Learning: Facial Expression Recognition Challenge (FER2013) organized since 2013 [53] have focused on categorical emotion recognition. The Audio/Visual Emotion Challenge (AVEC), organized since 2011 [202] has focused on dimensional emotion representation, in two dimensions, i.e., arousal and valence. The winning methods in most of the above Challenges have been based on deep neural networks. However, no architecture has been shown able to analyse large amounts of audio-visual data and generalise well its performance on different data sources.

Moreover, apart from affect analysis and recognition, generation of facial affect is of great significance, in many real life applications, such as for synthesis of affect on avatars that interact with humans, in computer games, in augmented and virtual environments, in educational and learning contexts.

Finally, we need to mention that the problem of affect analysis and recognition is more easily tackled nowadays due to the advent of GPU technology. GPUs are optimized for training deep learning models as they can process multiple computations simultaneously. They have a large number of cores, which allows for better computation of multiple parallel processes. Computations in deep learning for affect analysis and recognition need to handle huge amounts of data - this makes GPUs' memory bandwidth most suitable.

1.2 Aim and Objectives

The above-described motivation led us to tackle affect analysis and recognition as a dual knowledge generation problem. Our aim is to: i) create new large rich databases in-the-wild and ii) design and train novel deep neural architectures that are able to analyse affect over these databases and to successfully generalise their performance on other datasets. We target to advance the state-of-the-art in affect analysis and recognition and provide the research and industrial communities with both data and trained deep neural architectures that they can effectively use in real life environments. This duality is an underlying basis for our specific objectives that are presented below:

A first objective is the generation of a new, large scale, captured in-the-wild, video database that includes annotations in terms of continuous, dimensional emotion representations. It will be able to represent both abrupt and subtle affect in human behaviour; consequently, it can be used to model a large variety of events in real life human machine interactions.

The dual objective is the design and training of a novel end-to-end neural architecture that is able to learn over this database, in-the-wild. It will learn to capture subtle spatial and time variations of affect on this video database and to generalise well in other relevant databases.

The second objective is to extend the generated, in-the-wild database, including annotations for all major emotion representations, i.e., dimensional, categorical and facial action units. It will provide the basis for developing unified affect analysis and recognition methodologies and frameworks.

The related dual objective is to design and train a new neural architecture that will be able to automatically extract cues for all representations and analyse affect in-the-wild through a unified dimensional, categorical and action unit recognition framework.

The third objective relates to generation of affect on faces, using the above-described emotion representations, the dimensional and categorical ones. A large face database is also needed, annotated in terms of facial affect, for the design of a system that learns to render affect on the faces. It should be added that large experimental and ablation studies are required for illustrating that the approaches to be developed meet the above objectives.

1.3 Contributions

A major contribution of our research is related to the first, as well as its dual, objectives described in the former subsection. It includes generation of Aff-Wild, a new in-the-wild database, annotated according to the dimensional emotion representation, using the arousal and valence dimensions and including a big variety of: emotional states; rapid emotional changes; ethnicities; head poses; illumination conditions; occlusions. Aff-Wild was generated, by capitalizing on the abundance of data available in video-sharing websites, such as YouTube and selecting videos that displayed the affective behavior of people, for example videos that displayed the behavior of people when watching a trailer, a movie, a disturbing clip, or reactions to pranks. It contains 298 videos displaying reactions of 200 subjects, with a total video duration of more than 30 hours. The database has been annotated by 8 lay experts. Aff-Wild was first introduced in the Aff-Wild Challenge [109, 226], in conjunction with International Conference on Computer Vision & Pattern Recognition (CVPR) 2017. Since then Aff-Wild has been used by many researchers all around the world.

A novel end-to-end Deep Neural Architecture, AffWildNet, was then designed and trained on Aff-Wild to provide on-line valence and arousal estimation, being able to capture the temporal dynamics and the in-the-wild nature of Aff-Wild. It is a Convolutional and Recurrent Network (CNN-RNN) that takes as input bpth subjects' faces and extracted facial landmarks and is trained end-to-end to minimize the Concordance Correlation Coefficient [124].

After achieving an excellent performance on Aff-Wild, transfer learning and domain adaptation principles have been adopted to use AffWildNet as a robust prior for dimensional affect recognition on other major databases, either controlled, or in-the-wild ones. State-of-the-art results have been achieved as illustrated by large experimental studies. Moreover, for the first time, AffWildNet, which was trained as a dimensional affect recognition model, was able to constitute a robust prior for categorical affect recognition, achieving state-of-the-art performance on the problem of seven basic facial expression recognition in-the-wild.

Our research then focused on advancing the state-of-the-art related to the design of AffWildNet. In more detail, we developed novel CNN plus multi-RNN architectures - multi-component extensions

of AffWildNet -, in which low-, mid- and high-level latent variables were extracted and appropriately fused. We participated in the OMG-Emotion Challenge [12] and ranked second in estimating valence from visual cues. Another research contribution was to adapt and use, for the first time, the ArcFace Loss [45] when developing novel DNNs for affect recognition. Excellent results were acquired, illustrating the effectiveness of additive angular margin in affect recognition.

Another major contribution of our research is related to the second objective, and its dual one, described in the previous subsection. Aff-Wild has been first extended with new videos, creating Aff-Wild2, containing 558 videos with a total duration of more than 43 hours, showing reactions of a large number of 458 subjects. Aff-Wild2 has been the first database that contains annotations for all three main behavior tasks: estimation of Valence and Arousal; classification in 7 Basic Expression Categories; detection of Facial Action Units; Aff-Wild2 is also the largest existing in-the-wild database in each of these tasks. This provides researchers with the unique ability to perform all three affect recognition tasks simultaneously, using the same audiovisual data. Aff-Wild2 is the main database used in the Affective Behavior Analysis in-the-wild (ABAW) Competition [106], and a subsequent workshop, that we organised in conjunction with IEEE International Conference on Face and Gesture Recognition (FG) 2020.

In its dual problem, we exploited the multiple annotations of Aff-Wild2 so as to develop a novel multi-task learning in-the-wild approach. We developed multi-task and multi-modal extensions of AffWildNet that provide affect recognition, by fusing the three behavior analysis tasks (MT-AffWildNet), or also including the audio component (A/V-MT-AffWildNet), when annotations for all three tasks are available – as in Aff-Wild2. We further proposed a novel holistic approach [107] for handling cases with incomplete and missing annotations in the multi-task problem. Such cases include all other existing databases, which include annotation for only one or two of the tasks. We developed the holistic (multi-task, multi-domain and multi-label) FaceBehaviorNet, in which the three studied tasks were coupled by the developed co-annotation and distribution matching losses. We showed that FaceBehaviorNet has learned features that encapsulate all aspects of facial behaviour and can successfully perform tasks, such as compound affect recognition, beyond the ones for which it has been trained, in a zero- and few-shot learning setting.

A third major contribution of our research refers to the final objective of synthesising facial affect, either in terms of valence and arousal (which is performed for the first time), or in terms of the six basic expressions. We generate either an image with a given affect, or a sequence of images with evolving affect, in controlled, or in-the-wild environments. A novel approach is proposed, based on annotation of a 4-D face database and utilisation of a 3-D morphable face model. We further used the high-quality synthesized facial images for data augmentation in training of deep neural architectures, over eight databases, annotated with either dimensional or categorical affect labels, achieving very high performance and advancing the respective state-of-the-art.

Large experimental studies have been performed in the above developments, including comparison with all respective state-of-the-art facial affect recognition and synthesis methods, over all existing major relevant databases and illustrating the improvement obtained through the use of the approaches developed in the Thesis.

As far as the ethical issues of affect analysis are concerned: i) the collection and generation of the Aff-Wild and Aff-Wild2 databases has been conducted under the scrutiny and approval of the Imperial College Ethical Committee (ICREC); additionally we have contacted the person who created the videos -that we used- and asked for their approval to be used in this research; ii) while constructing the two afore-mentioned databases in order to eliminate the bias in terms of ages, sexes and ethnicities, we collected videos of subjects coming from all these categories, therefore the diversity of the inputs and data points, makes the developed architectures and systems more fair and unbiased.

Chapter 2 presents the background and literature related to affect analysis and recognition. Chapter 3 presents the generation of Aff-Wild and Aff-Wild2, lists their attributes and compares them with those of all existing major databases showing dimensional, or categorical affect, or facial action units. Chapter 4 presents and analyses AffWildNet, its usage as a robust prior for dimensional and categorical affect recognition in-the-wild, as well as the multi-component AffWildNet extensions and the use of ArcFace Loss for training novel expression recognition networks. Chapter 5 focuses on multi-task learning of DNNs for affect recognition in-the-wild, i.e., valence and arousal estimation, seven basic expression classification and facial action unit detection. It presents multi-task extensions of AffWild-Net, as well as development of the holistic FaceBehaviorNet that is able to successfully perform affect

recognition on data with missing (task-specific) annotations. Chapter 6 presents a new approach for synthesis of facial affect and for generation of faces that can be effectively used for affect analysis by deep neural architectures. A summary of thesis achievements, conclusions and suggestions for further work are provided in Chapter 7 of the Thesis.

1.4 Statement of Originality & Copyright Declaration

This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

1.5 Publications

The work presented in this thesis has resulted in the following list of journal and conference publications:

- **Dimitrios Kollias**, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep Neural Network Augmentation: Generating Faces for Affect Analysis. *International Journal of Computer Vision (IJCV)*, 2020.

- **Dimitrios Kollias** and Stefanos Zafeiriou. Exploiting multi-CNN features in CNN-RNN based Dimensional Emotion Recognition on the OMG in-the-wild Dataset. *IEEE Transactions on Affective Computing (IEEE TAC)*, 2020.

- **Dimitrios Kollias**, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing Affective Behavior in the First ABAW 2020 Competition. *IEEE International Conference on Automatic Face and Gesture Recognition (IEEE FG)*, 2020.

- **Dimitrios Kollias** and Stefanos Zafeiriou. VA-StarGAN: Continuous Affect Generation. *International Conference on Advanced Concepts for Intelligent Vision Systems*, 2020.

- **Dimitrios Kollias**, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face Behavior a' la carte: Expressions, Affect and Action Units in a Single Network. *filed as a patent*, 2019.

- **Dimitrios Kollias**, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild:Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision (IJCV)*, 2019.

- **Dimitrios Kollias** and Stefanos Zafeiriou. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and Arcface. *British Machine Vision Conference (BMVC)*, 2019.

- **Dimitrios Kollias**, Shiyang Cheng, Maja Pantic, and Stefanos Zafeiriou. Photorealistic Facial Synthesis in the Dimensional Affect Space. *European Conference on Computer Vision (ECCV)*, 2018.

- **Dimitrios Kollias** and Stefanos Zafeiriou. Training Deep Neural Networks with Different Datasets in-the-wild: The Emotion Recognition Paradigm. *International Joint Conference on-Neural Networks (IJCNN)*, 2018.

- **Dimitrios Kollias** and Stefanos Zafeiriou. A Multi-Component CNN-RNN Approach for Dimensional Emotion Recognition in-the-wild. *arXiv preprint*, 2018.

- **Dimitrios Kollias** and Stefanos Zafeiriou. A Multi-Task Learning & Generation Framework: Valence-Arousal, Action Units & Primary Expressions. *arXiv preprint*, 2018.

- **Dimitrios Kollias** and Stefanos Zafeiriou. Aff-wild2: Extending the Aff-Wild Database for Affect Recognition. *arXiv preprint*, 2018.

- **Dimitrios Kollias**, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of Affect in-the-wild using Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, 2017.

- Stefanos Zafeiriou*, **Dimitrios Kollias***, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and Arousal 'in-the-wild' Challenge. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W) (* the first two authors contributed equally)*, 2017.

Chapter 2

Background - Literature Review

2.1 Models of Affect

2.1.1 Categorical Affect

For the past twenty years research in automatic analysis of facial behaviour was mainly limited to the recognition of the so-called six universal expressions (i.e., Anger, Disgust, Fear, Happy, Sad, Surprise), plus the Neutral state, influenced by the seminal work of Ekman [61]. Ekman defined these six basic emotions, shown in Figure 2.1, based on a cross-culture study [61], which indicated that humans perceive certain basic emotions in the same way regardless of culture. However, recently, advanced research on neuroscience and psychology argued that the model of six basic emotions are culture-specific and not universal [92]. Additionally, the affect model based on basic emotions is limited in the ability to represent the complexity and subtlety of our daily affective displays [140,178]. However, the categorical model that describes emotions in terms of discrete basic emotions is still the most popular perspective for FER, due to its pioneering investigations along with the direct and intuitive definition of facial expressions.

Besides ordinary facial expressions met in every day social communications, emotions can also manifest themselves as micro-expressions (ME) under certain circumstances. A ME is defined as a very brief, involuntary facial expression occurring in accordance with an experienced emotional state. Es-

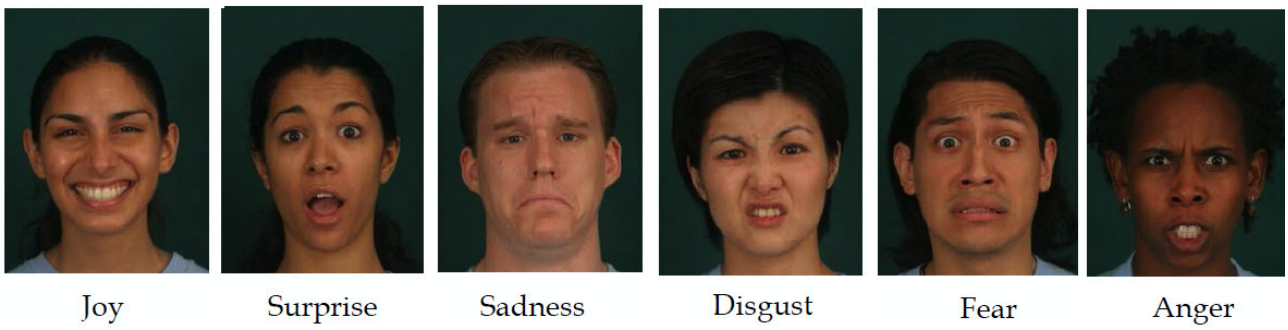


Figure 2.1: The six basic expressions

pecially in high-stake situations, humans are likely to display ME, despite their trying to conceal or mask their true feelings, e.g., so as to gain an advantage or avoid some loss. In comparison to ordinary facial expressions, a ME is very short (lasting $\frac{1}{25}$ to $\frac{1}{3}$ of a second, with the precise length varying in literature). Furthermore, the intensities of related muscle movements can be extremely subtle. The detection and interpretation of micro-expressions has been another area of current research.

Recently, research problems that focus on the recognition of spontaneous expressions including mental states, have also attracted attention, such as the recognition of pain intensity [7] and compound expressions [59].

2.1.2 Action Units

In this framework, detection of Facial Action Units has also attained much attention. The Facial Action Coding System (FACS) [61] provides a standardised taxonomy of facial muscles' movements and has been widely adopted as a common standard towards systematically categorising physical manifestation of complex facial expressions. Since any facial expression can be represented as a combination of action units, they constitute a natural physiological basis for face analysis. The existence of such a basis is a rare boon for a computer vision domain, as it allows focusing on the essential atoms of the problem and, by virtue of their exponentially large possible combinations, opens the door for studying a wide range of applications beyond prototypical emotion classification. Consequently, in the last years, there has been a shift of the scientific community towards the detection of action units. The expression of action units is typically brief and unconscious, and their detection requires analyzing subtle appearance changes in the human face. Furthermore, action units do not appear in isolation,

but as elemental units of facial expressions, and hence some AUs co-occur frequently while others are mutually exclusive. Figure 2.2 shows the most common action units and the corresponding facial action movement that defines them. A related problem that is recently gaining popularity relates to the estimation of the intensity of a particular activated action unit.

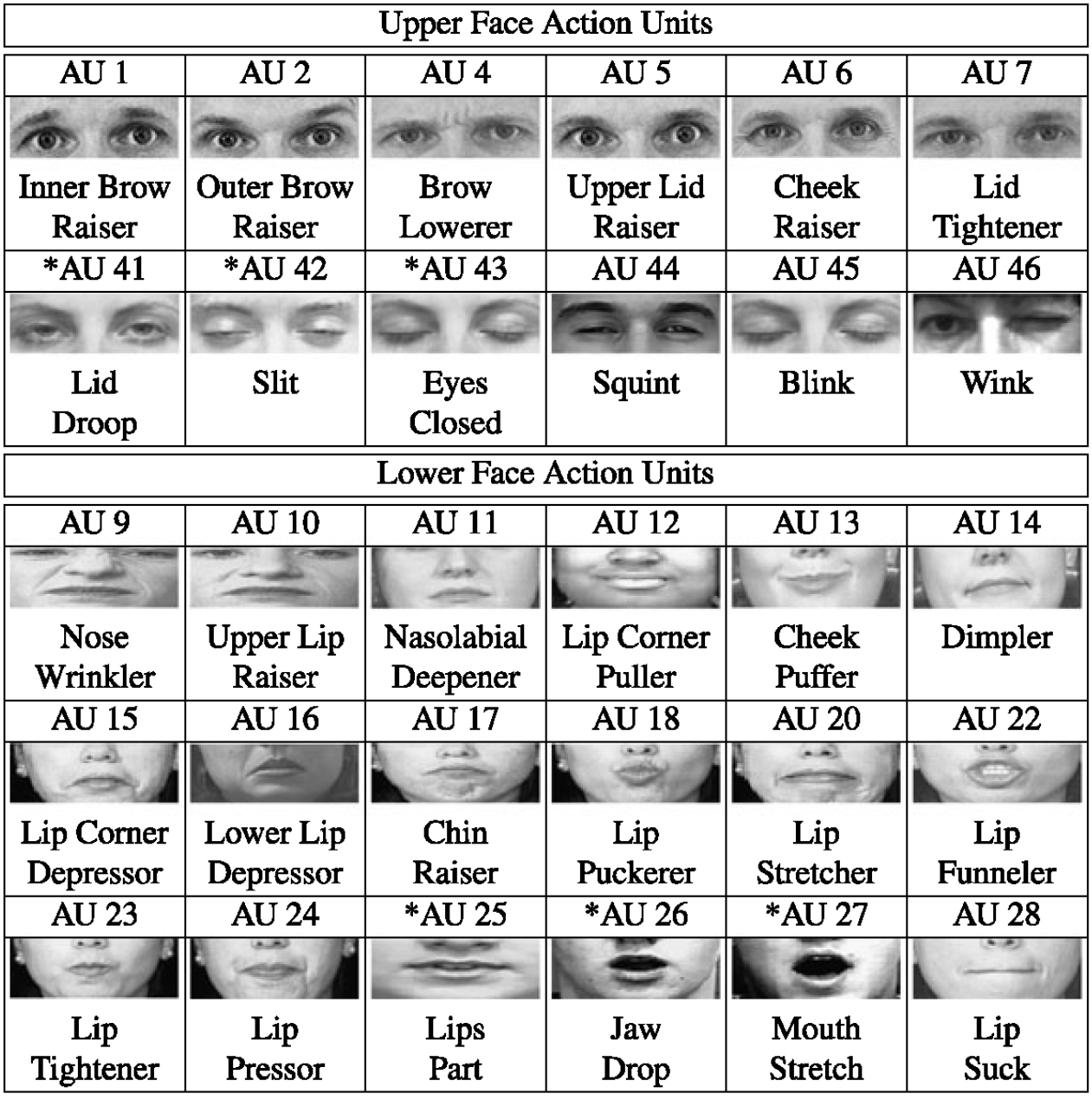


Figure 2.2: Some facial Action Units

2.1.3 Dimensional Affect

Finally, the dimensional model of affect, which is appropriate to represent not only extreme, but also subtle emotions appearing in everyday human-computer interactions, has also attracted significant attention over the last years. According to the dimensional approach [176] [214], affective behavior is described by a number of latent continuous dimensions. The most commonly used dimensions include valence (indicating how positive or negative an emotional state is) and arousal (measuring the power of emotion activation). Valence and arousal relate readily to specific functions of regions of the brain [90, 148, 198]; the parietal region of the right hemisphere appears to play a special role in the mediation of arousal, whereas the frontal regions appear to play a special role in emotional valence. A third dimension, tension, is also introduced but often excluded due to difficulties in consistently identifying what the dimension describes: tension, control, or potency (dominance). Figure 2.3 shows the 2D Valence-Arousal Space, introduced in [166], in which the horizontal axis is valence that ranges from very positive to very negative and the vertical one is arousal that ranges from very active to very passive.

2.2 Existing Datasets with Affect Annotation

Current research in automatic analysis of facial affect aims at developing systems, such as robots and virtual humans, that will interact with humans in a naturalistic way under real-world settings. To this end, such systems should automatically sense and interpret facial signals relevant to emotions, appraisals and intentions. Moreover, since real-world settings entail uncontrolled conditions, where subjects operate in a diversity of contexts and environments, systems that perform automatic analysis of human behavior should be robust to video recording conditions, the diversity of contexts and the timing of display. [1]

For the past twenty years research in automatic analysis of facial behavior was mainly limited to posed behavior which was captured in highly controlled recording conditions [135, 160, 197, 201].

[1] It is well known that the interpretation of a facial expression may depend on its dynamics, e.g. posed vs. spontaneous expressions [229].

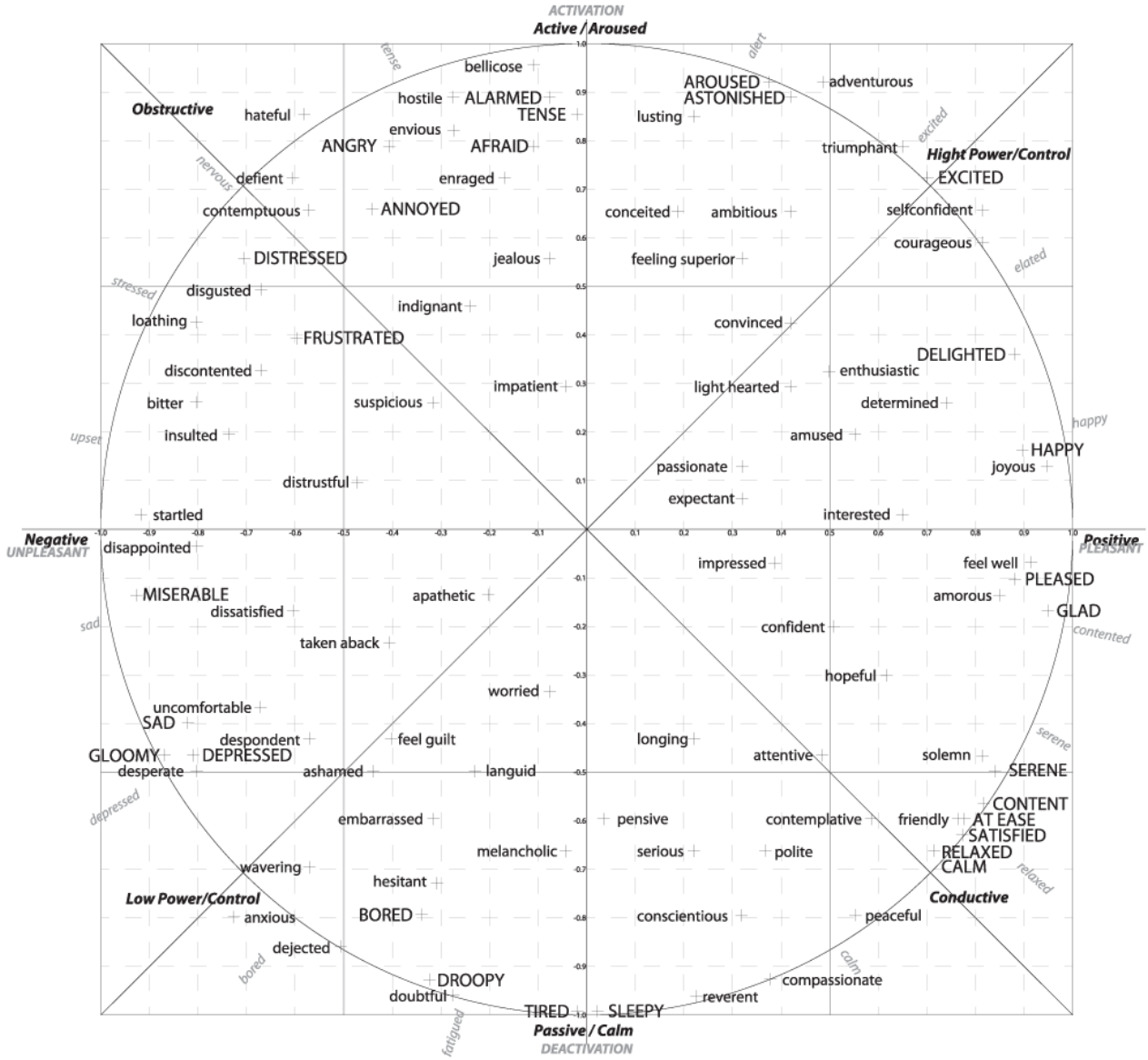


Figure 2.3: The 2D Valence-Arousal Space

Some representative datasets, which are still used in many recent works [94], are the Cohn-Kanade database [135, 197], MMI database [160, 201], Multi-PIE database [78] and the BU-3D and BU-4D databases [222, 223].

Nevertheless, it is now accepted by the community that facial expressions of naturalistic behaviors can be radically different from the posed ones [34, 178, 229]. Hence, efforts have been made in order to collect subjects displaying naturalistic behavior. Examples include the collected EmoPain [7] and UNBC-McMaster [136] databases for analysis of pain, the RU-FACS database of subjects participating in a false opinion scenario [15] and the SEMAINE corpus [143] which contains recordings of subjects interacting with a Sensitive Artificial Listener (SAL) in controlled conditions. All the above databases have been captured in well-controlled recording conditions and mainly under a strictly defined scenario eliciting pain.

However, with the development of large and diverse datasets in the field of computer vision (and the accompanying performance gains), it has become apparent that the diversity of human participants and spontaneous expressions have to become the prerogatives in deployment of the affective computing models in practice. Hence, it is now widely accepted, in both the computer vision and machine learning communities, that progress in a particular application domain is significantly catalysed when a large number of datasets are collected in unconstrained conditions (also referred as "in-the-wild" data). Therefore, facial analysis could not only focus on spontaneous behaviors, but also on behaviours captured in unconstrained conditions.

Some datasets with in-the-wild settings have been recently collected to study: i) facial expression analysis, such as the audiovisual AFEW [47], the static AffectNet [151] and the static RAF-DB [128]; ii) facial action units, such as the static EmotioNet [16]; and iii) continuous emotions of valence and arousal in-the-wild, such as the audiovisual OMG-Emotion Dataset [12], the audiovisual SEWA [172], the static AffectNet [151] and the static AFEW-VA [118]. Lets us note that the term 'static' means that the dataset contains only (static) images, neither video nor audio.

Next, we describe some, either controlled, or in-the-wild, databases that exist in literature (and are being utilised in our experiments in the next Sections) and are annotated in terms of either facial expressions, or action units, or valence-arousal. The controlled databases are captured in highly and

well controlled recording conditions, with good illumination, where subjects display posed affective states under a strictly defined scenario in cluttered backgrounds; there exist no occlusions on the face and not a big variety of head poses (mostly frontal ones exist). The in-the-wild databases are captured under different illumination conditions in uncluttered backgrounds and contexts, in which people have different head poses and there exist occlusions in the facial area.

The lack of in-the-wild databases and/or the fact that the existing ones are small in terms of size and subjects and/or have not been annotated by many experts and/or contain noisy annotations led to the creation of Aff-Wild and Aff-Wild2, as explained in the next Chapter.

2.2.1 Facial Expression Databases

1) BU-3DFE: The BU-3DFE database [223] is the first 3D facial expression database, which includes 2,500 expressive meshes from 100 subjects (56 females, 44 males) with age ranging from 18 to 70 years. The subjects are from various ethnic/racial ancestries. They recorded 6 articulated expressions (happiness, disgust, fear, angry, surprise and sadness) with 4 intensities; also, there is a neutral 3D scan per subject.

2) RaFD: The Radboud Faces Database [123] (RaFD) contains in total 5,880 portrait images of 49 models; 39 Caucasian Dutch adults and 10 Caucasian Dutch children. All models showed eight facial expressions with three gaze directions. Photos were taken against a uniform white background from five different camera angles simultaneously. Models wore black t-shirts, had no hair on the face and wore no glasses, makeup or jewellery.

3) Face place: This database [4] contains photographs of many different individuals in various types of disguises, such that, for each individual, there are multiple photographs in which hairstyle and/or eyeglasses have been changed/added. It consists of 1,284 images of Asian, 937 images of African-American, 3,362 images of Caucasian, 494 images of Hispanic and 497 images of multiracial people. All images show posed expressions.

[4] Stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/

Table 2.1: Existing Databases annotated in terms of facial expressions, along with their properties; 'static' means images, 'A/V' means audiovisual sequences, i.e., videos; '-' indicates no value is reported in the respective papers

DBs	DB Type	Model of Affect	Condition	DB Size	# of Subjects	Age Range
BU-3DFE [223]	static	6 Basic, Neutral + 4 levels of intensity	controlled	2,500	100 Male: 56 Female: 44	18-70
RaFD [123]	static	6 Basic, Neutral + Contempt	controlled	5,880	49 Male: 24 Female: 25	7-25
Face place[2]	static	6 Basic, Neutral + Confusion	controlled	6,574	235 Male: 143 Female: 92	-
2D Face Sets[3]: Pain	static	6 Basic, Neutral + 10 Pain Expr	controlled	599	23 Male: 13 Female: 10	-
2D Face Sets: Iranian	static	Neutral, Smile	controlled	369	34 Male: 0 Female: 34	-
2D Face Sets: Nottingham Scans	static	Neutral	controlled	100	100 Male: 50 Female: 50	-
KF-ITW [19]	static	Neutral, Happiness, Surprise	controlled	3,264	17	-
FEI [196]	static	Neutral, Smile	controlled	2,800	200 Male: 100 Female: 100	19-40
MULTI-PIE [78]	static	Neutral, Disgust, Smile Surprise + Squint, Scream	controlled	755,370	337 Male: 235 Female: 102	-
AFEW [51]	A/V	6 Basic, Neutral	in-the-wild	1809 videos 113,355 frames	-	-
FER2013 [76]	static	6 Basic, Neutral	in-the-wild	35,887	-	-
RAF-DB [128]	static	6 Basic, Neutral + 11 Compound	in-the-wild	15,339 + 3,954	-	0-70
AffectNet [151]	static	6 Basic, Neutral + Contempt	in-the-wild	291,651 manual, 400,000 automatic annotations	-	0 to >50

4) 2D Face Sets: The 2D Face Sets [5] database consists of the below three different sets of images.

Iranian women: It consists of 369 color images (1200 × 900) of 34 women. People display mostly smile and neutral expression in each of five poses under controlled conditions.

Nottingham scans: It has 100 monochrome images (50 men, 50 women) in neutral and frontal pose, under controlled settings. The image resolution varies from 358 × 463 to 468 × 536.

Pain expressions: It consists of 599 color images (720 × 576) of 13 women and 10 men, under laboratory recording conditions. They usually display two of the six basic emotions (anger, disgust, fear, sad, happy, surprise) plus 10 pain expressions. Profile neutral and 45 degrees images are also available.

5) Kinect Fusion ITW: The KF-ITW database [19] is the first Kinect 3D database captured under relatively unconstrained conditions. This database consists of 17 different subjects performing some expressions (neutral, happy, surprise) under various illumination conditions.

6) FEI: The FEI database [196] is a Brazilian face database that contains a set of face images taken in controlled conditions. 200 individuals were recorded, and each one has 14 images, resulting in 2,800 images of resolution 640 × 480. All images were color and taken against a white background in an upright frontal position with profile rotation of up to 180°. The subjects are mostly students and staff at FEI,and between 19 and 40 years old with distinct appearance, hairstyle and adorns. The number of male and female subjects are both 100.

7) Multi-PIE [78]: The CMU Multi-PIE face database contains 755,370 images (3072 × 2048) of 337 people recorded in up to four sessions under laboratory settings. Subjects were recorder under 15 view points and 19 illumination conditions while displaying a range of facial expressions. High resolution frontal images were acquired as well.

8) AFEW: This database is a dynamic facial expressions corpus used in the series of EmotiW Challenges [47–53] that focus on audiovisual classification of each video clip into the 7 basic emotion categories. It consists of 1,809 nearly real world scenes from movies and reality TV shows. There are over 330 subjects aging from 1 to 77. The database is split into three sets: training (773 videos),

[5] http://pics.stir.ac.uk

validation (383 videos) and test set (653 videos). It is a challenging database because both training and validation sets are mainly from the movies, while 114 out of 653 test videos are from TV. Semi-automatic annotations of neutral and 6 basic expressions are provided; the annotation is performed per video and not per frame.

9) FER2013: The in-the-wild FER2013 [76] dataset was utilized in the Facial Expression Recognition Challenge and contains 28,709 training images, 3,589 validation images and 3,589 test images. All images have resolution of 48×48 and are greyscale. The images are annotated in terms of the six basic expressions plus the neutral state. A curated version of FER2013, named FER+, was developed in which some of the original images were relabeled, while other images, e.g. not containing faces, were completely removed. The contempt class was added to the annotations of FER2013.

10) RAF-DB: The Real-world Affective Faces dataset [128] was prepared by collecting images from various search engines and was annotated manually by 40 independent labelers. The dataset contains 15,339 images labeled with seven basic emotion categories of which 3068 are to be used for testing and the rest 12,271 for training. It also contains 3,954 images annotated in terms of 11 coumpound expressions.

11) AffectNet: The AffectNet [151] database contains 287,651 training images and 4,000 validation images, which are manually annotated. The validation set contains 500 images for each of the 7 basic expressions and another 500 for the contempt one. The test set is not publicly available. The database also contains around 400,000 automatic annotations for the above expressions.

2.2.2 Action Unit Databases

1) CK+: The Extended Cohn-Kanade [135] (CK+) database is a laboratory-controlled one and contains 593 video sequences from 123 participants. The sequences vary in duration from 10 to 60 frames and show a shift from a neutral facial expression to the peak expression. Increases in AU intensity are monotonic. Pose is frontal with relatively little head motion.

2) MMI: The MMI database [201] is a laboratory-controlled one and includes 1,280 video sequences and over 250 images from 27 subjects. Many of the subjects wear accessories (e.g., glasses, mus-

Table 2.2: Existing Databases annotated in terms of action units, along with their properties; 'static' means images, 'dynamic' means image sequences (video without audio), 'A/V' means audiovisual sequences, i.e., videos; '-' indicates no value is reported in the respective papers

DBs	DB Type	Model of Affect	Condition	DB Size	# of Subjects	Age Range
CK+ [135]	static	30 action units	controlled	593 sequences	123 Male: 38 Female: 85	18-50
MMI [201]	static & A/V	12 action units	controlled	1,280 videos + over 250 images	25 Male: 12 Female: 13	20-32
DISFA [142]	dynamic	12 action units + intensities	controlled	54 videos 261,630 frames	27 Male: 15 Female: 12	18-50
BP4DS [231]	dynamic	27 action units + intensities	controlled	1,640 videos 222,573 frames	41 Male: 23 Female: 18	18-29
BP4D+ [239]	dynamic	34 action units + intensities	controlled	5,463 videos 967,570 frames	140 Male: 58 Female: 82	18-66
EmotioNet [16]	static	11 action units	in-the-wild	50,000 manual, 950,000 automatic annotations	-	0 to >40

tache). 205 video sequences are captured in frontal view. The sequences in MMI are onset-apex-offset labeled, i.e., the sequence begins with a neutral expression and reaches peak near the middle before returning to the neutral expression.

3) DISFA: The Denver Intensity of Spontaneous Facial Action [142] (DISFA) database is a lab controlled database with spontaneous emotion expressions. It has been annotated for the presence, absence and intensity of 12 AUs: 1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25, 26. It consists of 27 subjects, each recorded while watching a four minutes video clip by two cameras. It consists of 260K video frames (130K frames from each camera).

4) BP4DS: The BP4D Spontaneous database [231] (in the rest of the paper we refer to it as BP4DS) is a lab-controlled database with spontaneous expressionsa and is annotated for the occurence and intensity of 27 AUs appearing in a diverse group of young adults. There are 21 subjects with 75.6K images in the training, 20 subjects with 71.2K images in the development and 20 subjects with 75.7K images in the test partition. These sets have been part of the corresponding sets of the FERA 2015 Challenge [203], in which only AUs 1,2,4,6,7,10,12,14,15,17,23 were used. From the participants, 11 were Asian, 6 were African-American, 4 were Hispanic, and 20 were Euro-American.

5) BP4D+: The BP4D+ database [235] is an extension of BP4DS described above by incorporating

different modalities as well as more subjects (140). Ethnic/Racial Ancestries include Black, White, Asian (including East-Asian and Middle-East-Asian), Hispanic/Latino, and others (e.g., Native American). BP4D+ is annotated for occurrence of 34 AUs and intensity for 5 of them. It has been used as a part of the FERA 2017 Challenge [204]. Only AUs 1,4,6,7,10,12,14,15,17,23 have been used in the Challenge.

5) EmotioNet: The Emotionet database [65] is a large-scale database with around 1M facial expression images collected from the Internet. It was released for the EmotioNet Challenge in 2017 [17] [6]. A total of 950K images were annotated by the model of [65], and the remaining 50K images were manually annotated with 11 AUs, 1, 2, 4, 5, 6, 9, 12, 17, 20, 25, 26; around half of these constituted the validation and the other half the test set of the Challenge.

2.2.3 Valence-Arousal Databases

1) MAHNOB-HCI: The MAHNOB-HCI [185] database is a lab-controlled multimodal database. The peripheral physiological signals from 24 participants were recorded after eliciting their emotion by 20 affective movies.

2) DEAP: The Database for Emotion Analysis using Physiological signals [101] (DEAP) is a lab controlled database that contains the spontaneous bodily responses of 32 participants after inducing their emotional states by watching selected music videos clips.

3) SEMAINE: The Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression Dataset [143] (SEMAINE) presents richly annotated recordings of interactions in laboratory conditions between a human and a machine-like agent in three different Sensitive Active Listener (SAL) scenarios. It features 150 participants, most of which come from Caucasian background. The language of communication is predominantly English.

4) Belfast naturalistic: The Belfast Naturalistic Database contains 10 to 60 seconds–long audiovisual videos taken from English television chat shows, current affairs programmes and interviews. It features 125 subjects, of which 31 are male, and 94 are females.

[6] https://cbcsl.ece.ohio-state.edu/EmotionNetChallenge/index.html

Table 2.3: Existing Databases annotated in terms of valence and arousal, along with their properties; 'static' means images, 'dynamic' means image sequences (video without audio), 'A/V' means audiovisual sequences, i.e., videos; '-' indicates no value is reported in the respective papers

DBs	DB Type	Model of Affect	Condition	DB Size	# of Subjects
MAHNOB-HCI [185]	A/V	valence-arousal	controlled	20 videos	27 Male: 11 Female: 16
DEAP [101]	A/V	valence-arousal	controlled	40 videos	32
SAL [58]	A/V	valence-arousal	controlled	24 videos	4
SEMAINE [143]	A/V	valence-arousal	controlled	959 videos	150 Male: 57 Female: 93
Belfast naturalistic	A/V	valence-arousal	controlled	298 videos	125 Male: 31 Female: 94
Belfast induced [183]	A/V	valence-arousal	controlled	37 videos	37
RECOLA [173]	A/V	valence-arousal	controlled	46 videos 345,000 frames	46
AFEW-VA [118]	dynamic	valence-arousal	in-the-wild	600 videos 30,050 frames	240 Male: 120 Female: 120
AffectNet [151]	static	valence-arousal	in-the-wild	325,000 manual, 460,300 automatic annotations	-
SEWA [119]	A/V	valence-arousal	in-the-wild	538 videos	<398 Male: <201 Female: <197
OMG-Emotion [12]	A/V	valence-arousal	in-the-wild	495 videos 5,288 utterances	-

5) Belfast Induced: To collect Belfast Induced Natural Emotion Database [183], English speaking participants were asked to select set of tasks specifically designed to induce mild to moderately strong emotionally coloured responses. Mean age of subjects is 24 years with 6 years of deviation. Continuous values of valence and arousal were obtained for each clip by 6 to 258 raters using FeelTrace.

6) RECOLA: The REmote COLlaborative and Affective [173] (RECOLA) database contains natural and spontaneous emotions in the continuous domain (arousal and valence). The corpus consists of 46 French speaking subjects being recorded for 9.5 h recordings in total. The recordings were annotated for 5 minutes each by 6 French-speaking annotators (three male, three female). The dataset is divided into three parts, namely, training (16 subjects), validation (15 subjects) and test (15 subjects), in such a way that the gender, age and mother tongue are stratified (i.e., balanced).

7) AFEW-VA: A part of the AFEW database has been annotated -in September 2017- in terms of Valence and Arousal, thus creating the in-the-wild AFEW-VA [118] database. It includes 600 video clips selected from films with real-world conditions, i.e., occlusions, illumination and body movements. The length of each video ranges from around 10 frames to over 120 frames. This database consists of per-frame annotations of V-A. In total, more than 30,000 frames were annotated for dimensional affect prediction of V-A, using discrete values in the range of $[-10, +10]$.

7) AffectNet: The AffectNet [151] database contains 320,500 training images and 4,500 validation images, which are manually annotated. The test set is not publicly available. The database also contains around 460,300 automatic annotations. This in-the-wild database was created -in August 2017- by querying emotion related keywords from three search engines. The manual annotations were performed by 12 human labelers in total, but each image was annotated by one annotator.

7) SEWA: The SEWA [119] database consists of 538 short (10-30s) video-chat recording segments, annotated in a semi-automatic way by 3 annotators. It contains in-the-wild audio-visual data of 398 people coming from six cultures, spanning the age range of 18 to 65 years. Subjects were recorded in two different contexts: while watching adverts and while discussing adverts in a video chat.

7) OMG-Emotion: The One-Minute Gradual-Emotional Behavior [12] (OMG-Emotion) dataset contains in-the-wild videos from Youtube where emotion expressions emerge and develop over time

based on monologued scenarios. The dataset is split into training, validation and test sets. The training set consists of 231 videos composed of 2442 utterances, the validation set consists of 60 videos composed of 617 utterances and the test set consists of 204 videos composed of 2229 utterances. Each utterance has an average length of 8 seconds and each video has an average length of around 1 minute. Each utterance was given one specific valence and arousal value, based on the gold standard of the five annotations. Valence annotations range in $[-1,1]$, whereas arousal ones range in $[0,1]$

2.3 Existing Methodologies for Affect Recognition

In order to facilitate research on the three models of affect, many databases have been generated and annotated, as discussed before, most of which are in well-controlled conditions. It has not been until recently that the focus has turned into "in-the-wild" data. In the beginning, data and annotations were scarce, hence research relied on extracting highly engineered handcrafted features and designing ad-hoc learning strategies [73,146,147,157,169]. Naturally, as the amount of data and annotations grew, research has started to capitalise on data-intensive technologies, such as deep learning [26,28,98,130].

Regarding the pipeline of facial behavior analysis, the standard paradigm has been to: i) detect and/or track the face in an image sequence, ii) detect and/or track facial landmarks, iii) extract handcrafted features[7], either around the landmarks, or on the face region as a whole, and iv) use the features and the landmarks for classification/regression using affective labels. Recently this paradigm has shifted from utilizing handcrafted features to utilizing features learned by deep Convolutional Neural Networks (CNNs) and/or Recurrent Neural Networks (RNNs). This shift was motivated by the striking performance achieved when utilizing deep neural networks (DNNs) in a variety of emotion recognition tasks [56,79,154,238].

In Facial Expression Recognition, there are mainly two roads of research. The one is pre-training self-built networks from scratch by using auxiliary task-oriented data (so as to avoid overfitting when relatively small facial expression datasets are used), or using well-known pre-trained models (e.g., VGG/VGG-FACE [161, 182], ResNet [82], DenseNet [89], ResNext [218], SE-ResNet [87]) and

[7]Examples of handcrafted features include Histogram of Oriented Gradients (HoGs), Scale Invariant Feature Transform (SIFT), Local Binary Patterns (LBPs) and features from multiscale and multiorientation Gabor filterbanks

fine-tuning them on FER datasets. In [95] and [96], it is indicated that the use of auxiliary data helps to obtain models with high capacity without overfitting, consequently enhancing FER performance. In [100], it is shown that pre-training on large face recognition data positively affects emotion recognition accuracy and further fine-tuning with additional FER datasets boosts the performance even more.

The second road of research involved the addition of well-designed auxiliary layers or blocks to models (mainly CNN architectures) for enhancing the expression-related representation capability of learned features. One such example is HoloNet [221]. In HoloNet's middle layers, the authors used a modified Concatenated Rectified Linear Unit (CReLU) -instead of the typical ReLU- which they combined with residual block to maintain efficiency, increase network depth and obtain accuracy gain. In upper layers, the authors developed a variant of inception-residual block that learned multi-scale features so as to capture variations in expressions. Another example is the Supervised Scoring Ensemble (SSE) model [88] that was introduced to enhance the supervision degree for FER. In SSE, three types of supervised blocks were embedded in the early hidden layers of the mainstream CNN for shallow, intermediate and deep supervision, respectively. To filter irrelevant features and emphasize on correlated features according to learned feature maps of facial expression, a feature selection network (FSN) [237] was used by embedding a feature selection mechanism inside the AlexNet.

When referring to dimensional affect recognition, one of the first deep learning architectures for valence and arousal estimation was proposed in [98]. In this work, both frame-based CNN and CNN plus RNN architectures were proposed and compared. The CNN consisted of 3 convolutional layers; the first two layers were followed by max pooling layers and the third by a quadrant pooling layer. A fully connected layer was then used, followed by the output layer. The CNN plus RNN architectures consisted of the previously described CNN network (keeping its weights fixed) without the top regression layer, followed by a single RNN layer that gave the final estimates. This methodology achieved very high valence and arousal correlations in a part of the RECOLA database [173].

The authors in [28] explored and fused different hand-crafted and deep learning features from all available modalities (acoustic, visual, textual). They also considered the interlocutor influence (a person's influence on the interacting partner's behaviors) for the acoustic features. Specifically, the authors

extracted: i) from the acoustic modality, hand-crafted features, such as MFCCs, loundness, F0, jitter, shimmer and features learned from the SoundNet [9], ii) from the visual modality, features learned from VGG-FACE [161] and DenseNet [89] that had been pre-trained on the FER+ [13] dataset, and iii) from the textual modality, word vectors that were used as features. All those features were fused and passed as input to a LSTM network that produced the estimates for valence, arousal and likability. This approach was the winning of AVEC 2017 Challenge that utilized the SEWA database.

In summary, Table 2.4 provides a summary of the performance of the afore-mentioned methods for valence & arousal estimation, in terms of the mean squared error (MSE), the Pearson correlation coefficient (PCC) and the Concordance Correlation Coefficient (CCC). A higher CCC or PCC and a lower MSE value indicate a better performance.

Table 2.4: State-of-the-art methods for valence-arousal estimation and their performance

Work	Databases Used	Methods	Results
[98]	part of RECOLA as used in the AVEC Challenge	CNN-RNN visual only: (conv + max-pool) x2 + conv + quadrant-pool + RNN	Valence: RMSE = 0.107 PCC = 0.554 CCC = 0.507
[28]	SEWA	(1) audio: handcrafted + SoundNet features (2) visual: VGG-FACE + DenseNet features (3) text: word vectors - features fusion of (1), (2), (3) + LSTM	Valence - Arousal: RMSE = 0.081 - 0.086 PCC = 0.758 - 0.702 CCC = 0.756 - 0.672

Chapter 3

Aff-Wild Databases

3.1 The Aff-Wild Database

Back in 2017, there existed some databases for dimensional emotion recognition. However, they were captured in laboratory settings and not in-the-wild (i.e., not in uncontrolled conditions). This urged us to create the benchmark Aff-Wild database and organize the Aff-Wild Challenge that utilised this database.

To tackle the aforementioned limitation, we collected the first, to the best of our knowledge, large scale captured in-the-wild database and annotated it in terms of valence and arousal. To do so, we capitalized on the abundance of data available in video-sharing websites, such as YouTube [224][1] and selected videos that display the affective behavior of people, for example videos that display the behaviour of people when watching a trailer, a movie, a disturbing clip, or reactions to pranks.

To this end we have collected 298 videos displaying reactions of 200 subjects, with a total video duration of more than 30 hours. This database has been annotated by 8 lay experts with regards to two continuous emotion dimensions, i.e. valence and arousal. We then organised the Aff-Wild Challenge based on the Aff-Wild database [226] [105], in conjunction with International Conference

[1]The collection has been conducted under the scrutiny and approval of the Imperial College Ethical Committee (ICREC). The majority of the chosen videos were under Creative Commons License (CCL). For those videos that were not under CCL, we have contacted the person who created them and asked for their approval to be used in this research.

on Computer Vision & Pattern Recognition (CVPR) 2017.

3.1.1 Limitations of Databases & Contributions of Aff-Wild

Currently, four databases exist, are widely used and are annotated in terms of valence and arousal: RECOLA, AFEW-VA, AffectNet and OMG-Emotion. Table 3.1 summarizes their limitations:

- the main limitations of the RECOLA dataset include the tightly controlled laboratory environment, as well as the small number of subjects (46). It should be noted that it contains a moderate total number of frames.
- the main limitations of AFEW-VA include its small size (in terms of total number of frames), the small number of annotators (only 2), the heavily imbalanced set in which mostly all annotations are in the second quadrant of the 2D VA Space (where valence is negative and arousal is positive), the fact that it contains only videoframes but no audio at all and finally the use of discrete values for valence and arousal. It should be noted that the 2D VA Space (Figure 2.3) is a continuous space. Therefore, using discrete only values for valence and arousal provides a rather coarse approximation of the behavior of persons in their everyday interactions. On the other hand, using continuous values can provide improved modelling of the expressiveness and richness of emotional states met in everyday human behaviors.
- the main limitations of AffectNet include its very imbalanced training set, the very small total number of images in the validation set (only 4,500 images), the small number of annotators (each image was annotated by one annotator) and finally the fact that it contains only static images (neither videoframes nor audio). The test set of this database is not released. It should be noted that it contains a moderate total number of frames.
- the main limitations of OMG-Emotion include the fact that the set consists of videos split into utterances, with their number being rather small (5,288) and that only one annotated valence-arousal value is given to each utterance and not to each frame (each utterance consists of a big number of frames; the annotation is per utterance and not per frame) and finally the annotated values of arousal are in the range [0,1].

Table 3.1 also provides the advantages of Aff-Wild over the previously described databases. When Aff-Wild was developed, it was the first time that a large in-the-wild database - with a big variety of: (1) emotional states, (2) rapid emotional changes, (3) ethnicities, (4) head poses, (5) illumination conditions and (6) occlusions - has been generated and used for affect recognition. Furthermore, Aff-Wild was the largest existing dimensionally annotated database consisting of 1,224,100 total number of frames. Aff-Wild is an audiovisual database (meaning that it contains both videoframes and audio). The annotation of Aff-Wild was performed by a large number of experts (8 lay experts), was done per-frame and the annotated values were continuous and ranged in [-1,1] for both valence and arousal. It should be mentioned that the total number of subjects in Aff-Wild is moderate (298).

Table 3.1: Current databases annotated in terms of valence and arousal, their disadvantages-limitations and comparison to Aff-Wild

Database	Year of Publication	Comments
RECOLA	2013	- laboratory environment - moderate total amount of frames (345,000); 120,000 frames in training set, 112,500 frames in validation & 112,500 frames in test set - small number of subjects (46)
AFEW-VA	09/2017	- very small total number of frames (30,050) - discrete valence and arousal values - heavily imbalanced set, with most annotations in the second quadrant of 2D VA Space - small number of annotators (2) - contains only video frames and no audio
AffectNet	08/2017	- moderate total number of manual annotations (325,000) - very imbalanced training set - contains only static images, neither video frames nor audio - test set is not released - very small total number of images in the validation set (4,500) - small number of annotators (each image was annotated by one annotator)
OMG-Emotion	2018	- one annotated value was given to each utterance and not to each frame - small number of utterances (5,288) - arousal ranges in [0,1] and thus no negative values exist
Aff-Wild	07/2017	+ the first in-the-wild database annotated for valence-arousal + the largest in-the-wild database (1,224,100 frames) + audio-visual database + annotation per frame + every frame was annotated by 8 experts + continuous valence-arousal values in [-1,1] - moderate number of subjects (298)

3.1.2 Collected Database and its Properties

We created a database consisting of 298 videos, with a total length of more than 30 hours. The aim was to collect spontaneous facial behaviors in arbitrary recording conditions. To this end, the videos were collected using the Youtube video sharing web-site. The main keyword that was used to retrieve the videos was "reaction". The database displays subjects reacting to a variety of stimuli, e.g. viewing an unexpected plot twist of a movie or series, a trailer of a highly anticipated movie, or tasting something hot or disgusting. The subjects display both positive or negative emotions (or combinations of them). In other cases, subjects display emotions while performing an activity (e.g., riding a rolling coaster). In some videos, subjects react on a practical joke, or on positive surprises (e.g., a gift). The videos contain subjects from different genders and ethnicities with high variations in head pose and lightning.

Most of the videos are in YUV 4:2:0 format, with some of them being in AVI format. Eight subjects have annotated the videos following a methodology similar to the one proposed in [37], in terms of valence and arousal. An online annotation procedure was used, according to which annotators were watching each video and provided their annotations through a joystick. Valence and arousal range continuously in $[-1, +1]$. All subjects present in each video have been annotated. The total number of subjects is 200, with 130 of them being male and 70 of them female. Table 3.2 shows the general attributes of the Aff-Wild database. Figure 3.1 shows some frames from the Aff-Wild database, with people from different ethnicities displaying various emotions, with different head poses and illumination conditions, as well as occlusions in the facial area.

Table 3.2: Attributes of the Aff-Wild Database

Attribute	Description
Length of videos	$0.10 - 14.47$ min
Video format	AVI , MP4
Average Image Resolution (AIR)	607×359
Standard deviation of AIR	85×11
Median Image Resolution	640×360

Figure 3.2 shows an example of annotated valence and arousal values over a part of a video in the Aff-Wild, together with corresponding frames. This illustrates the in-the-wild nature of our database,

Figure 3.1: Frames from the Aff-Wild database which show subjects in different emotional states, of different ethnicities, in a variety of head poses, illumination conditions and occlusions.

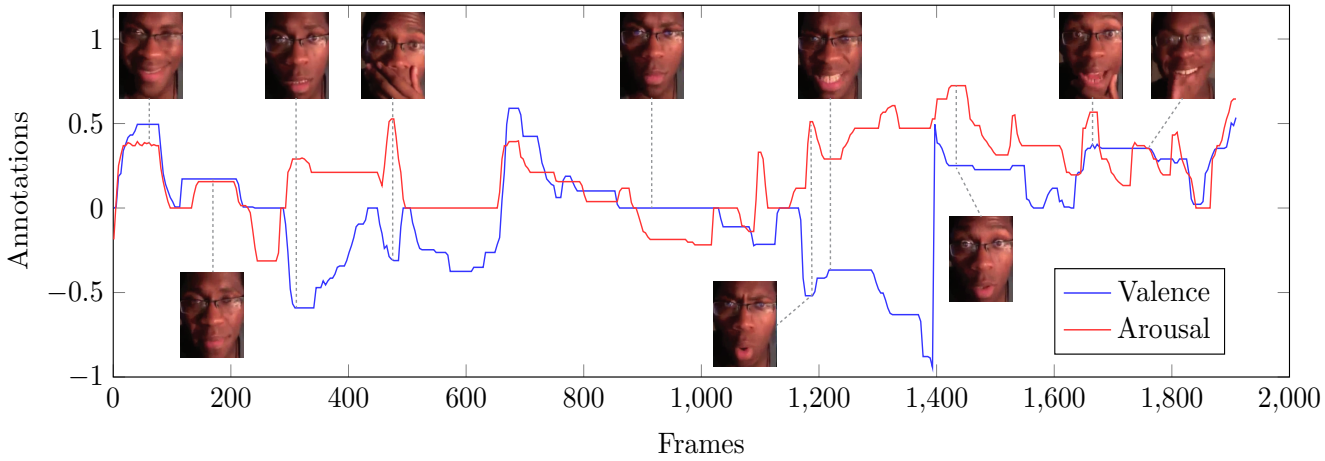


Figure 3.2: Valence and arousal annotations over a part of a video, along with corresponding frames; illustrating (i) the in-the-wild nature of Aff-Wild (different emotional states, rapid emotional changes, occlusions) and (ii) the use of continuous values for valence and arousal

namely, including many different emotional states, rapid emotional changes and occlusions in the facial areas. Figure 3.2 also shows the use of continuous values for valence and arousal annotation, which gives the ability to effectively model all these different phenomena. Figure 3.3 provides a histogram for the annotated values for valence and arousal in the generated database.

3.1.3 Partition Sets and Distributions

The Aff-Wild database was split into a training and a test set. The partitioning was performed in a subject independent manner, in the sense that a person/subject can appear either in the training set or in the testing set and not on both of them. As a consequence, the resulting training and test sets

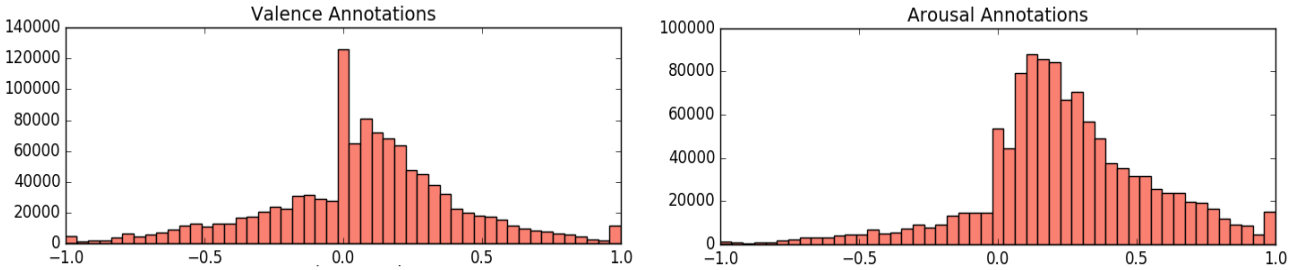


Figure 3.3: Histogram of valence and arousal annotations of the Aff-Wild database.

consist of *252* and *46* videos. Table 3.3 summarizes the specific attributes (numbers of males, females, videos, frames) of the training and test sets of Aff-Wild.

Table 3.3: Attributes of Training and Test sets of Aff-Wild.

Set	no of males	no of females	no of videos	total no of frames
Training	106	48	252	1,008,650
Test	24	22	46	215,450

3.1.4 Data Pre-processing and Annotation

The Aff-Wild database has been made public to the research community; Aff-Wild's videos and annotations along with face bounding boxes and landmarks are freely distributed. In this section we describe the pre-processing process of the Aff-Wild videos so as to perform face and facial landmark detection. Then we present the annotation procedure including:

(1) Creation of the annotation tool.

(2) Generation of guidelines for six experts to follow in order to perform the annotation.

(3) Post-processing annotation: the six annotators watched all videos again, checked their annotations and performed any corrections; two new annotators watched all videos and selected 2-4 annotations that best described each video; final annotations are the mean of the selected annotations by these two new annotators.

The detected faces and facial landmarks, as well as the generated annotations are publicly available with the Aff-Wild database.

Finally, we present a statistical analysis of the annotations created for each video, illustrating the consistency of annotations achieved by using the above procedure.

Aff-Wild video pre-processing

VirtualDub [126] was used first so as to trim the raw YouTube videos, mainly at their beginning and end-points, in order to remove useless content (e.g., advertisements). Then, we extracted a total of 1,224,100 video frames using the Menpo software [2]. In each frame, we detected the faces and generated corresponding bounding boxes, using the method described in [141]. Next, we extracted facial landmarks in all frames using the best performing method as indicated in [33].

During this process, we removed frames in which the bounding box or landmark detection failed. Failures occurred when either the bounding boxes, or landmarks, were wrongly detected, or were not detected at all. The former case was semi-automatically discovered by: (i) detecting significant shifts in the bounding box and landmark positions between consecutive frames and (ii) having the annotators verify the wrong detection in the frames.

Annotation tool

For data annotation, we developed our own application that builds on other existing ones, like Feeltrace [37] and Gtrace [38]. A time-continuous annotation is performed for each affective dimension, with the annotation process being as follows:

(a) the user logs in to the application using an identifier (e.g. his/her name) and selects an appropriate joystick;

(b) a scrolling list of all videos appears and the user selects a video to annotate;

(c) a screen appears that shows the selected video and a slider of valence or arousal values ranging in $[-1, 1]$;

(d) the user annotates the video by moving the joystick either up or down;

(e) finally, a file is created including the annotation values and the corresponding time instances that the annotations are generated.

It should be mentioned that the time instances generated in the above step (e), did not generally match the video frame rate. To tackle this problem, we modified/re-sampled the annotation time instances using nearest neighbor interpolation.

Figure 3.4 shows the graphical interface of our tool when annotating valence (the interface for arousal is similar); this corresponds to step (c) of the above described annotation process.

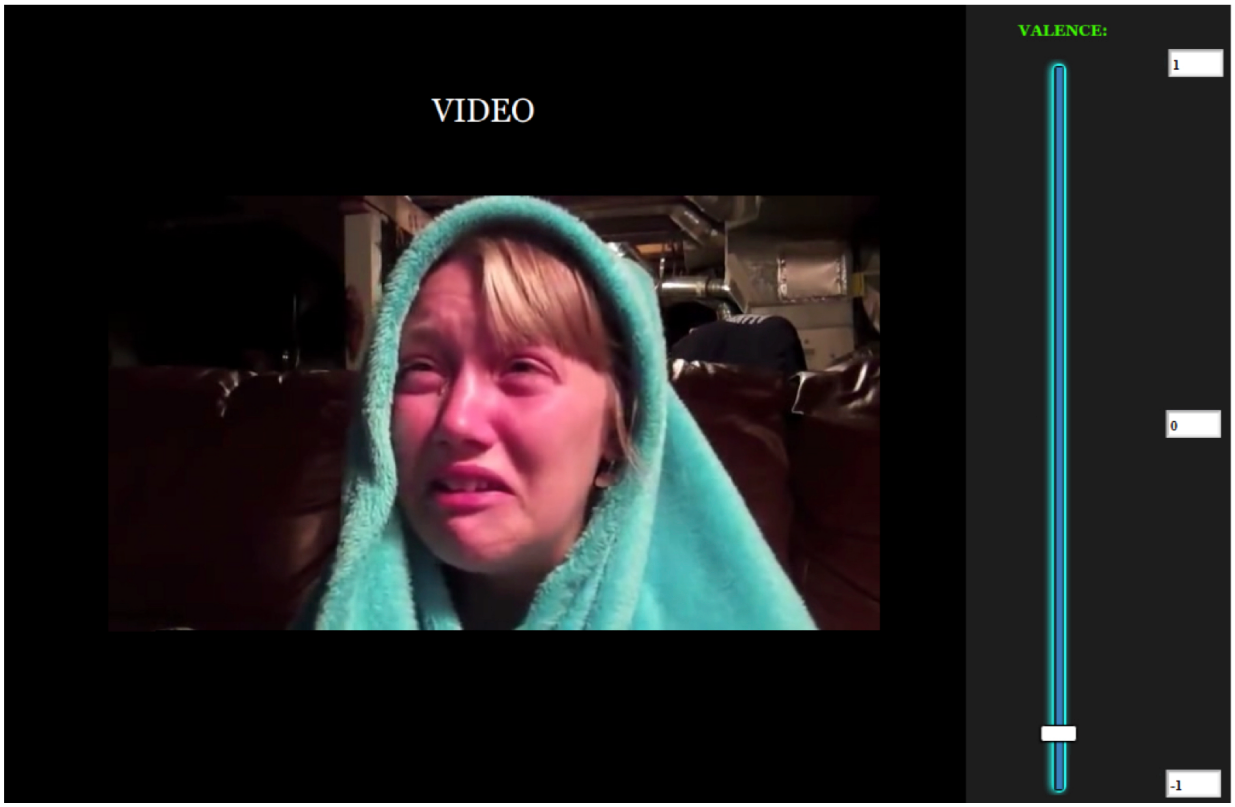


Figure 3.4: The GUI of the annotation tool when annotating valence (the GUI for arousal is exactly the same).

It should also be added that the annotation tool has also the ability to show the inserted valence and arousal annotation while displaying a respective video. This is used for annotation verification in a post-processing step.

Annotation guidelines

Six experts were chosen to perform the annotation task (including the author of this thesis). Each annotator was instructed orally and through a multi-page document on the procedure to follow for the task. This document included a list of some well identified emotional cues for both arousal and valence, providing a common basis for the annotation task. On top of that the experts used their own appraisal of the subject's emotional state for creating the annotations.[2] Before starting the annotation of each video, the experts watched the whole video so as to know what to expect regarding the emotions being displayed in the video.

Annotation Post-processing

A post-processing annotation verification step was also performed. Every expert-annotator watched all videos for a second time in order to verify that the recorded annotations were in accordance with the shown emotions in the videos or change the annotations accordingly. In this way, a further validation of annotations was achieved.

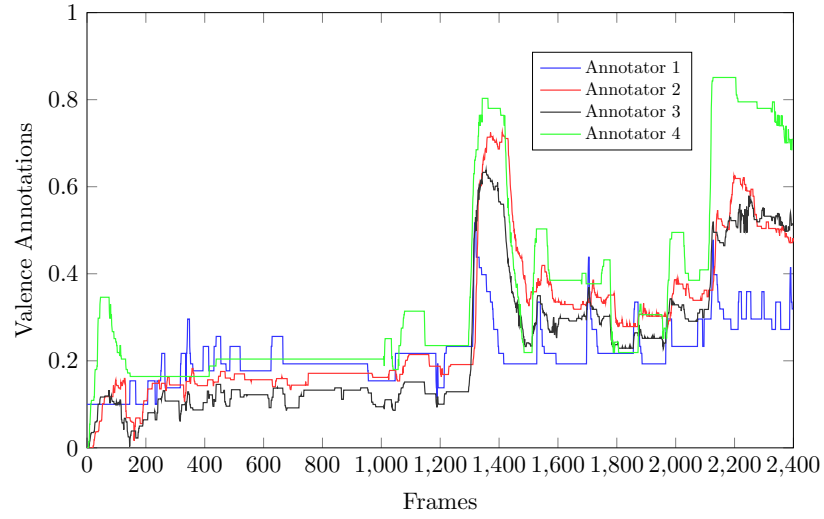
After the annotations have been validated by the annotators, a final annotation selection step followed. Two new experts watched all videos and, for every video, selected the annotations (between two and four) which best described the displayed emotions. The mean of these selected annotations constitute the final Aff-Wild labels.

This step is significant for obtaining highly correlated annotations, as shown by the statistical analysis presented next.

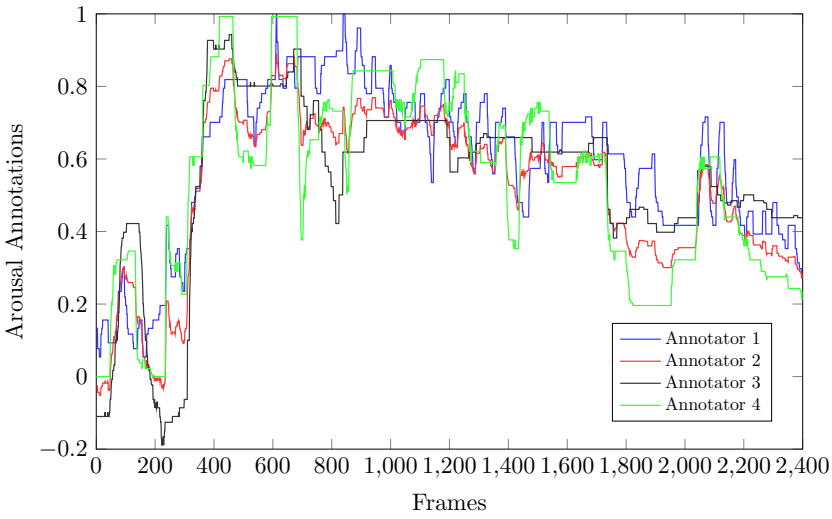
Statistical Analysis of Annotations

In the following we provide a quantitative and rich statistical analysis of the achieved Aff-Wild labeling. At first, for each video, and independently for valence and arousal, we computed:

[2] All annotators were computer scientists who were working on face analysis problems and all had a working understanding of facial expressions.



(a)



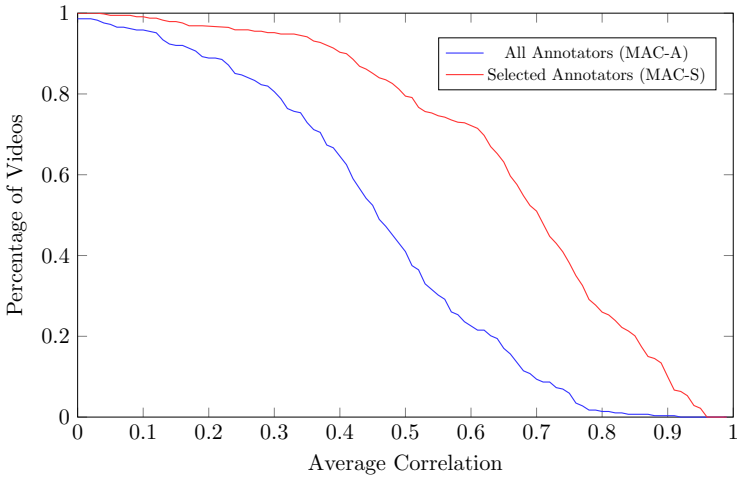
(b)

Figure 3.5: The four selected annotations in a video segment for (a) valence and (b) arousal. In both cases, the value of MAC-S (mean of average correlations between these four annotations) is 0.70. This value is similar to the mean MAC-S obtained over all Aff-Wild.

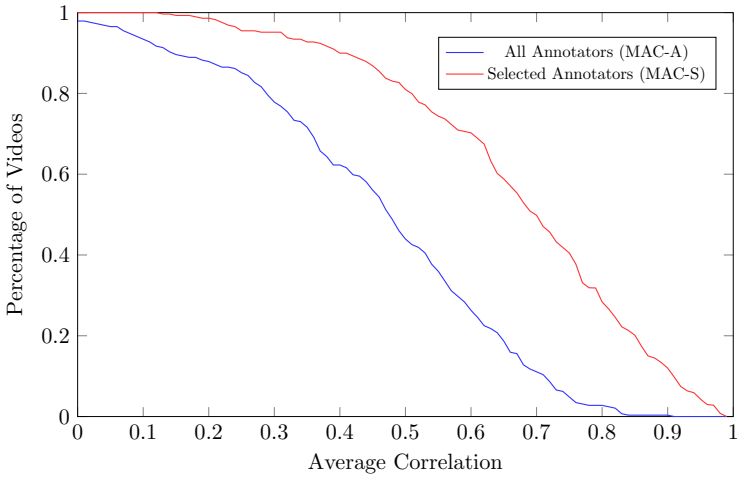
(i) the inter-annotator correlations, i.e., the correlations of each one of the six annotators with all other annotators, which resulted in five correlation values per annotator;

(ii) for each annotator, his/her average inter-annotator correlations, resulting in one value per annotator; the mean of those six average inter-annotator correlations value is denoted next as MAC-A;

(iii) the average inter-annotator correlations, across only the selected annotators, as described in the previous subsection, resulting in one value per selected annotator; the mean of those 2-4 average inter-selected-annotator correlations values is denoted next as MAC-S.



(a)

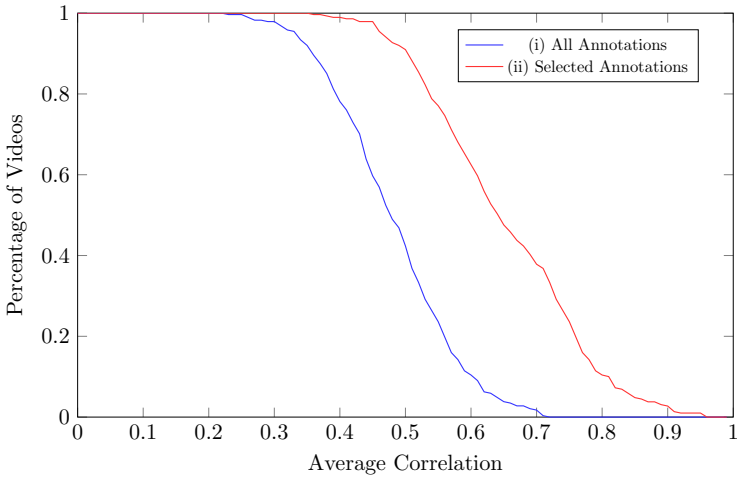


(b)

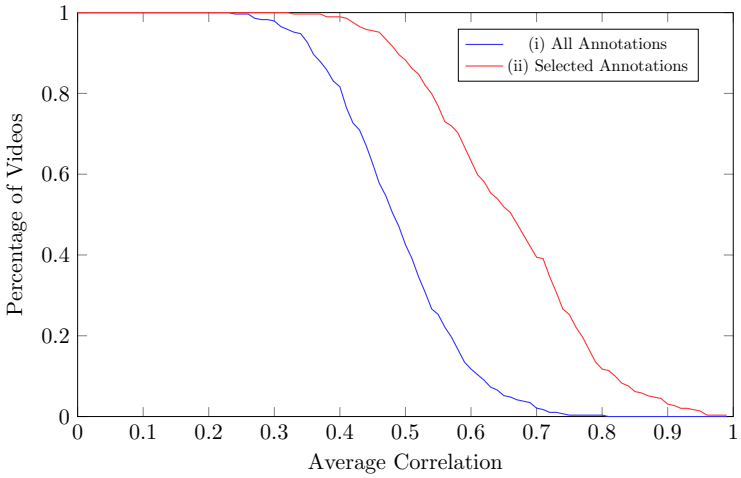
Figure 3.6: The cumulative distribution of MAC-S (mean of average inter-selected-annotator correlations) and MAC-A (mean of average inter-annotator correlations) values over all Aff-Wild videos for valence (Figure 3.6a) and arousal (Figure 3.6b). The Figure shows the percentage of videos with a MAC-S/MAC-A value greater or equal to the values shown in the horizontal axis. The mean MAC-S value, corresponding to a value of 0.5 in the vertical axis, is 0.71 for valence and 0.70 for arousal.

We then computed over all videos and independently for valence and arousal, the mean of MAC-A and the mean of MAC-S computed in (ii) and (iii) above. The mean MAC-A is 0.47 for valence and 0.46 for arousal, whilst the mean MAC-S for valence is 0.71 and for arousal 0.70. An example set of annotations is shown in Figure 3.5, in an effort to further clarify the obtained MAC-S values. It shows the four selected annotations in a video segment for valence and arousal, respectively, with MAC-S value of 0.70 (similar to the mean MAC-S value obtained over all Aff-Wild).

In addition, Figure 3.6 shows the cumulative distribution of MAC-S and MAC-A values over all Aff-Wild videos for valence (Figure 3.6a) and arousal (Figure 3.6b). In each case, two curves are shown.



(a)



(b)

Figure 3.7: The cumulative distribution of the correlation between landmarks and the average of (i) all or (ii) selected annotations over all Aff-Wild videos for valence (Figure 3.7a) and arousal (Figure 3.7b). The Figure shows the percentage of videos with a correlation value greater or equal to the values shown in the horizontal axis.

Every point (x, y) on these curves has a y value showing the percentage of videos with a (i) MAC-S (red curve) or (ii) MAC-A (blue curve) value greater or equal to x; the latter denotes an average correlation in $[0, 1]$. It can be observed that the mean MAC-S value, corresponding to a value of 0.5 in the vertical axis, is 0.71 for valence and 0.70 for arousal. These plots also illustrate that the MAC-S values are much higher than the corresponding MAC-A values in both valence and arousal annotation, verifying the effectiveness of the annotation post-processing procedure.

Next, we conducted similar experiments for the valence/ arousal average annotations and the facial landmarks in each video, in order to evaluate the correlation of annotations to landmarks. To this end,

we utilized Canonical Correlation Analysis (CCA) [81]. In particular, for each video and independently for valence and arousal, we computed the correlation between landmarks and the average of (i) all or (ii) selected annotations.

Figure 3.7 shows the cumulative distribution of these correlations over all Aff-Wild videos for valence (Figure 3.7a) and arousal (Figure 3.7b), similarly to Figure 3.6. Results of this analysis verify that the annotator-landmark correlation is much higher in the case of selected annotations than in the case of all annotations.

3.2 The Aff-Wild2 database

Up to the present, there was no database that contains annotations for all main behavior tasks (valence-arousal estimation, action unit detection, expression classification). Most of the existing databases contain annotations for only one task (AffectNet is the exception, that contains annotations for two tasks). Also the existing corpora have a number of other limitations; just to name a few: the not in-the-wild nature, the total number of annotations that is small (making it impossible to train deep neural networks and generalise to other databases), the automatic or semi-automatic annotation (which is error prone and makes the annotations noisy), or the small number of expert annotators (making the annotations biased).

These urged and led us to create the Aff-Wild2 database; the first and only database annotated in terms of valence and arousal (VA), action units (AUs) and expressions (Exprs). Aff-Wild2 is a significant extension of Aff-Wild, augmenting it with 260 more YouTube videos, which had a total duration of 13 hours and 5 minutes. In total, Aff-Wild2 consists of 558 videos of 58 subjects, with around 2,800,000 frames, showing both subtle and extreme human behaviours in real-world settings. Let us mention that we additionally organised the Affective Behavior Analysis in-the-wild (ABAW) Competition that utilised the Aff-Wild2 database, in conjunction with IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2020.

Aff-Wild2 [111, 113, 116] is described next. At first we present the existing databases' limitations and compare them with Aff-Wild2. Then, we present the new collected dataset and its properties, the

Table 3.4: Current databases used for affect recognition, their disadvantages-limitations and comparison to Aff-Wild2

Database	Comments
AFEW	- only seven basic expressions annotation - heavily imbalanced classes - small total number of frames (113,355) - semi-automatic annotation - small number of annotators (3) - one annotated value was given to the whole video and not to each frame
FER2013	- only seven basic expressions annotation - annotations are noisy/have mistakes - very small total number of images (35,887) - contains only static images - heavily imbalanced classes
RAF-DB	- only seven basic and 11 compound expressions annotation - very small total number of images (15,339) - validation set does not exist (only training and test set) - contains only static images - heavily imbalanced classes
AffectNet	- only 7 basic expressions plus contempt & valence-arousal annotations - moderate total number of manually annotated images - very imbalanced training set - contains only static images, neither video frames nor audio - test set is not released - very small total number of images in the validation set (4,500) - small number of annotators (each image was annotated by one annotator)
Aff-Wild	- only valence-arousal annotations - moderate number of subjects (298)
DISFA	- only action unit annotations - controlled conditions - small number of subjects (27) - dynamic video sequences, no audio exists - small total number of frames (130,815)
BP4DS	- only action unit annotations - controlled conditions - small number of subjects (41) - dynamic video sequences, no audio exists - moderate total number of frames (222,573)
BP4D+	- only action unit annotations - controlled conditions - small number of subjects (140) - dynamic video sequences, no audio exists
EmotioNet	- action unit annotations and 7 basic and 11 compound expressions - very small number of action unit manually annotated images (50,000) - very small number of expression annotated images (3,000) - very imbalanced expression categories (around 1,500 of the 3,000 images are happy) - no manually annotated training set (only validation and test sets) - contains only static images, neither dynamic video sequences nor audio
Aff-Wild2	+ first that contains annotations for: valence-arousal, expressions, action units + first action unit annotated with audio + first A/V in-the-wild for action units + first A/V in-the-wild for expressions with per frame annotation + largest in-the-wild database for valence-arousal + largest in-the-wild for expressions + largest in-the-wild for action units with manual annotations + contains only manual annotations

generated partition sets, their distributions and the annotation procedure.

3.2.1 Limitations of Databases & Contributions of Aff-Wild2

Table 3.4 shows the existing databases along with their limitations. The limitations of the AFEW dataset include its restriction to only seven expression categories, which are heavily imbalanced (the fear, disgust and surprise classes include a small number of samples), its small size (in terms of total number of frames which is 113,355), its annotations that are semi-automatic (with a small number of annotators, just 3) and also are per video and not per videoframe. The limitations of FER2013 include its restriction to only seven expression categories, which are heavily imbalanced, its small size (the database contains only 35,887 static images) and its annotations that are noisy (meaning that they contain mistakes). The limitations of RAF-DB include its restriction to only seven and eleven compound expression categories, all of which are heavily imbalanced, its very small size (it contains only 15,339 static images) and finally the fact that it does not contain a validation set (it contains only training and test sets).

The limitations of DISFA include its restriction to action unit annotations, its controlled environment, its small number of subjects (only 27) and its small total number of frames (130,815 videoframes; no audio exists). The limitations of BP4DS include its restriction to action unit annotations, its controlled environment, its small number of subjects (only 41) and its moderate total number of frames (222,573 videoframes; no audio exists). The limitations of BP4D+ include its restriction to action unit annotations, its controlled environment, its small number of subjects (140) and the fact that it contains dynamic video sequences and no audio. The limitations of EmotioNet include its restriction to action unit and basic and compound expression annotations, its very small number of manually annotated, with action units, static images (only 50,000), its very small number of basic and compound expression annotated images (only 3,000 static images; it doesn't contain video and/or audio), which are heavily imbalanced (around 1,500 image are annotated with the happy expression) and finally its not manually annotated training set (only the validation and test sets are).

All these urged and led us to create the Aff-Wild2 database; the first and only database annotated

Figure 3.8: Frames of Aff-Wild2, showing subjects of different ethnicities, age groups, emotional states, head poses, illumination conditions and occlusions

in terms of valence and arousal (VA), action units (AUs) and expressions (Exprs). Aff-Wild2 is a significant extension of Aff-Wild, augmenting it with more videos and annotations. Additionally Aff-Wild2 is: i) the first action unit annotated database with audio, ii) the first A/V in-the-wild database annotated with action units, iii) the first A/V in-the-wild database annotated with expressions with per frame annotation, iv) the largest in-the-wild database for valence-arousal, v) the largest in-the-wild database for expressions, vi) the largest in-the-wild database for action units with manual annotations, vii) contains only manual annotations. The above contributions of Aff-Wild2 (over the existing databases) are summarized in Table 3.4.

3.2.2 Collected Database and Properties

We extend the Aff-Wild database [109, 226], by collecting a new dataset consisting of *260* YouTube videos, with *1,413,000* frames and a total length of *13* hours and *5* minutes. The videos have been collected using the Youtube video sharing website. All of the collected videos are in MP4 format, with a frame rate of *30*, provided under the CC licence. Keywords for retrieving the videos were selected from the 2D VA Space, shown in Figure 2.3.

The new videos have wide range in subjects': age (from babies to elderly people); ethnicity (caucasian/hispanic/latino/asian/black/african american); profession (e.g. actors, athletes, politicians, journalists); head pose; illumination conditions; occlussions; emotions. Figure 3.8 shows frames from these new videos, verifying the above described ranges. These videos show subjects who: react on a surprise, on something that brings them happiness or fulfillment, on flirting or rejection, on important political issues, on funny or mean tweets; are stand-up comedians; give a really interesting

speech in ceremonies; are taking an oral exam; are giving lectures on depression, or other serious disorders; are performing passive, boring, apathetic, intense activities, etc.

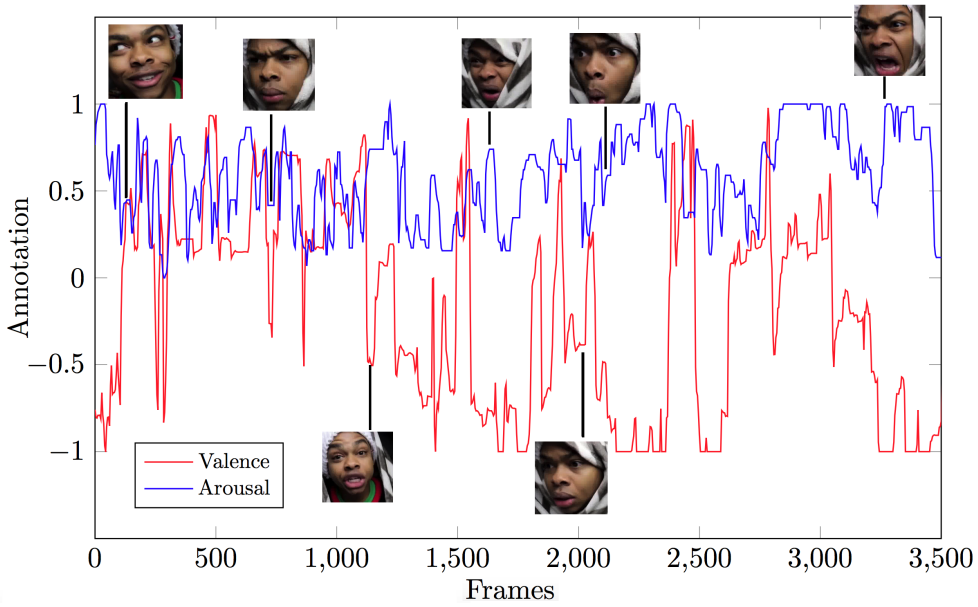


Figure 3.9: Valence and arousal annotations over a part of a video, along with corresponding frames, illustrating the in-the-wild nature of Aff-Wild2 (different emotional states, rapid emotional changes, occlusions)

Four experts annotated the new dataset in terms of valence and arousal, as in the case of Aff-Wild. Figure 3.9 shows an example of annotated valence and arousal values over a part of a video in the additional data, together with some corresponding frames. This illustrates the in-the-wild nature of the database, namely, including many different emotional states, rapid emotional changes and occlusions in the facial areas.

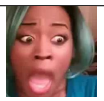
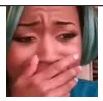
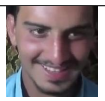
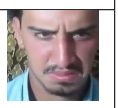
We then concatenated the Aff-Wild database with the new dataset, forming Aff-Wild2. In total, Aff-Wild2 consists of *558* videos with *2,786,201* frames, showing both subtle and extreme human behaviours in real-world settings. The total number of subjects is *458*; *279* of which are males and *179* females.

Two more tasks were implemented, in which we annotated parts of Aff-Wild2 with AUs and Exprs. In the first, three very experienced annotators annotated *63* videos, with *398,835* frames and a total length of *3* hours and *41* mins, in terms of AUs *1,2,4,6,12,15,20,25* - described in Figure 3.10. These videos contain *32* male and *31* female subjects. In the second, seven experts annotated *539* videos consisting of *2,595,572* frames, with a total length of *25* hours and *45* mins, in terms of the 7 basic

expressions. The videos show *431* subjects, *265* of which are male and *166* female.

Consequently, Aff-Wild2 contains 3 datasets (VA, AU, Expr); each contains annotations for a respective behavior task. Table 3.6 summarizes the attributes and properties of the three annotated sets of Aff-Wild2. Table 3.5 shows some images with their corresponding VA, AU and Expr annotations.

Table 3.5: Images with their corresponding VA, AU and Expr annotations

Annotation	Images			
				
Valence	-0.69	-0.54	0.38	-0.30
Arousal	0.92	0.52	0.35	0.51
AU 1	x			
AU 2				
AU 4		x		x
AU 6		x		
AU 12			x	
AU 15				x
AU 20				
AU 25	x		x	
Expression	Fear	Sadness	Happiness	-

AU#	Action
1	inner brow raiser
2	outer brow raiser
4	brow lowerer
6	cheek raiser
12	lip corner puller
15	Lip Corner Depressor
20	lip stretcher
25	lips part

Figure 3.10: The AUs annotated in Aff-Wild2, along with their corresponding facial actions

3.2.3 Partition Sets and Distributions

Each set (VA, AU, Expr) is split into three subsets: training, validation and test. Partitioning is done in a subject independent manner, in the sense that a person can appear only in one of those three

Table 3.6: General Attributes of Aff-Wild2; in the VA set, top row refers to the new dataset, while bottom row refers to Aff-Wild

Aff-Wild2	# frames	# videos	# annotators	Video Length	Mean Resolution
VA set	1,413,000 1,373,201	260 298	4 8	0.03 – 26.22 mins 0.10 – 14.47 mins	1450 × 900 607 × 359
AU set	398,835	63	3	0.03 – 26.22 mins	1500 × 900
Expr set	2,595,572	539	7	0.03 – 26.22 mins	1000 × 700

subsets. In the VA set, the resulting training, validation and test subsets consist of *349*, *68* and *131* videos respectively. In the AU set, the respective subsets consist of *42*, *7* and *14* videos respectively. In the Expr set, the corresponding subsets consist of *250*, *69* and *220* videos respectively.

Figure 3.11 shows the 2D VA histogram of Aff-Wild2. Figure 3.12 shows the distribution of the seven emotion categories in Aff-Wild2. Table 3.7 shows the distribution of the activated AUs.

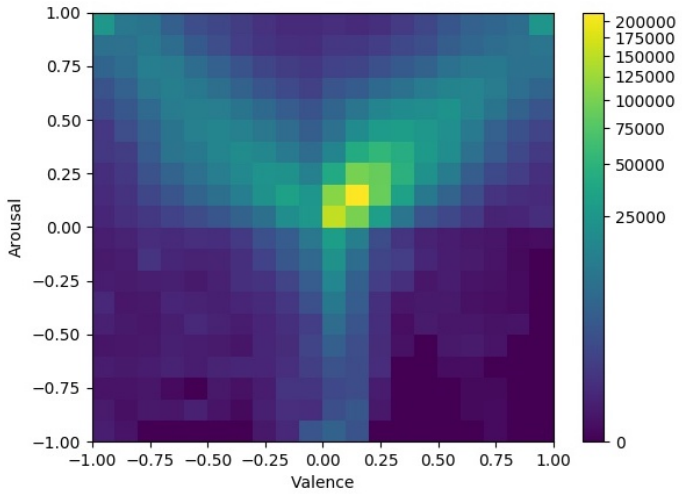


Figure 3.11: 2D Valence-Arousal Histogram of Aff-Wild2

Table 3.7: Distribution of AU annotations in Aff-Wild2

Action Unit #	Total Number of Activated AUs	Percentages of AUs
AU 1	86,677	43.9%
AU 2	4,166	2.1%
AU 4	56,327	28.5%
AU 6	25,226	12.8%
AU 12	35,675	18.1%
AU 15	3,340	1.7%
AU 20	5,695	2.9%
AU 25	9,048	4.6%

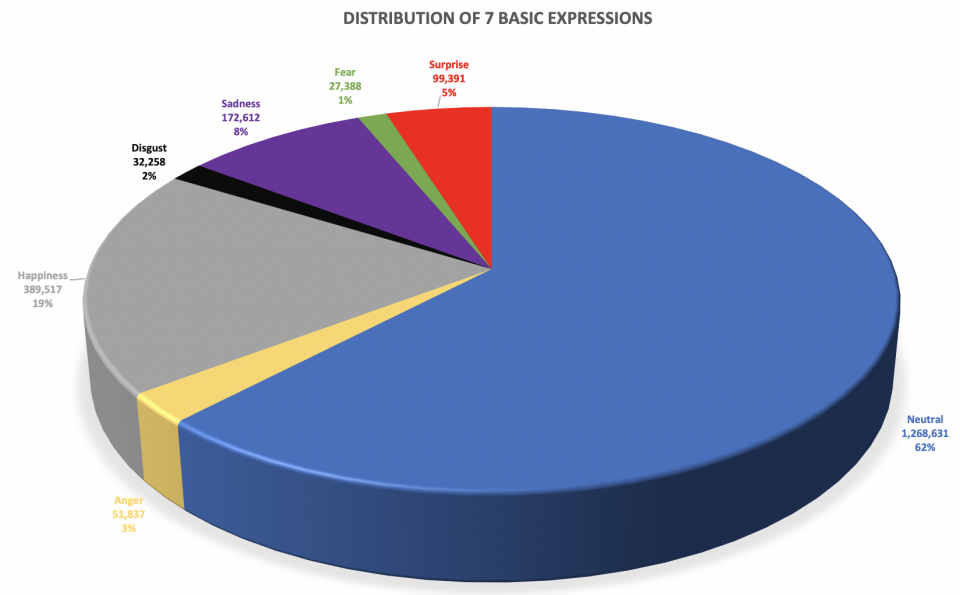


Figure 3.12: Histogram of the seven basic expressions in Aff-Wild2

3.2.4 Annotation

Four experts (including the author of this thesis) annotated Aff-Wild2 with respect to valence and arousal, using the method proposed in [37]. The annotators watched each video and provided their (frame-by-frame) annotations through a joystick. A time-continuous annotation was generated for each affective dimension. Valence and arousal values range continuously in [-1,1]. The final label values were the mean of those four annotations. The mean inter-annotation correlation is 0.63 for valence and 0.60 for arousal. An example set of annotations is shown in Figure 3.13. It shows the four annotations in a video segment for valence, with mean inter-annotation correlation of 0.64 (similar to the 0.63 mean inter-annotation correlation obtained over all Aff-Wild2).

Three experts performed the annotation of Aff-Wild2 for the occurrence of eight action units in a frame-by-frame basis; a platform-tool was developed in order to split each video into frames and let the experts annotate each videoframe. The annotation platform-tool is shown on Figure 3.14 and enabled the experts to annotate each Action Unit independently and frame-by-frame for each video. The agreement between the annotators has not always been 100%. Therefore, we decided to keep the annotations, on which all three experts agreed.

Seven experts performed the annotation of Aff-Wild2 for the seven basic expressions in a frame-

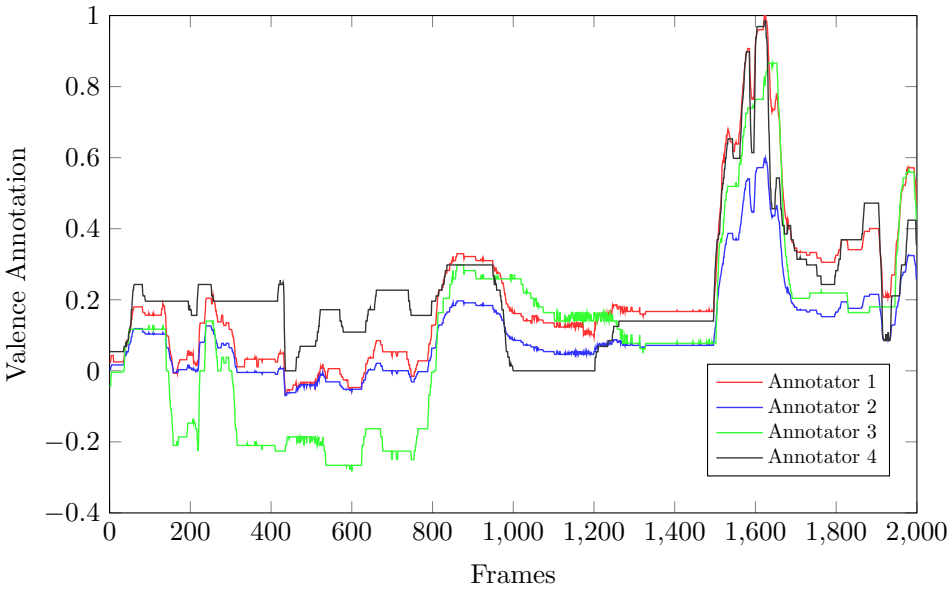


Figure 3.13: All four valence annotations in a video segment. The value of MAIC (mean of average inter annotation correlation) is 0.64 which is similar to the mean MAIC obtained over all additional data.

by-frame basis; a platform-tool (similar to the one used for annotating the eight action units) was developed in order to split each video into frames and let the experts annotate each videoframe. Let us mention that in this platform-tool, an expert could score a videoframe as having either one of the seven basic expressions or none (since there are affective states other than the seven basic expressions). Due to subjectivity of annotators and wide ranging levels of images' difficulty, there were some disagreements among annotators. We decided to keep only the annotations on which at least five (out of seven) experts agreed.

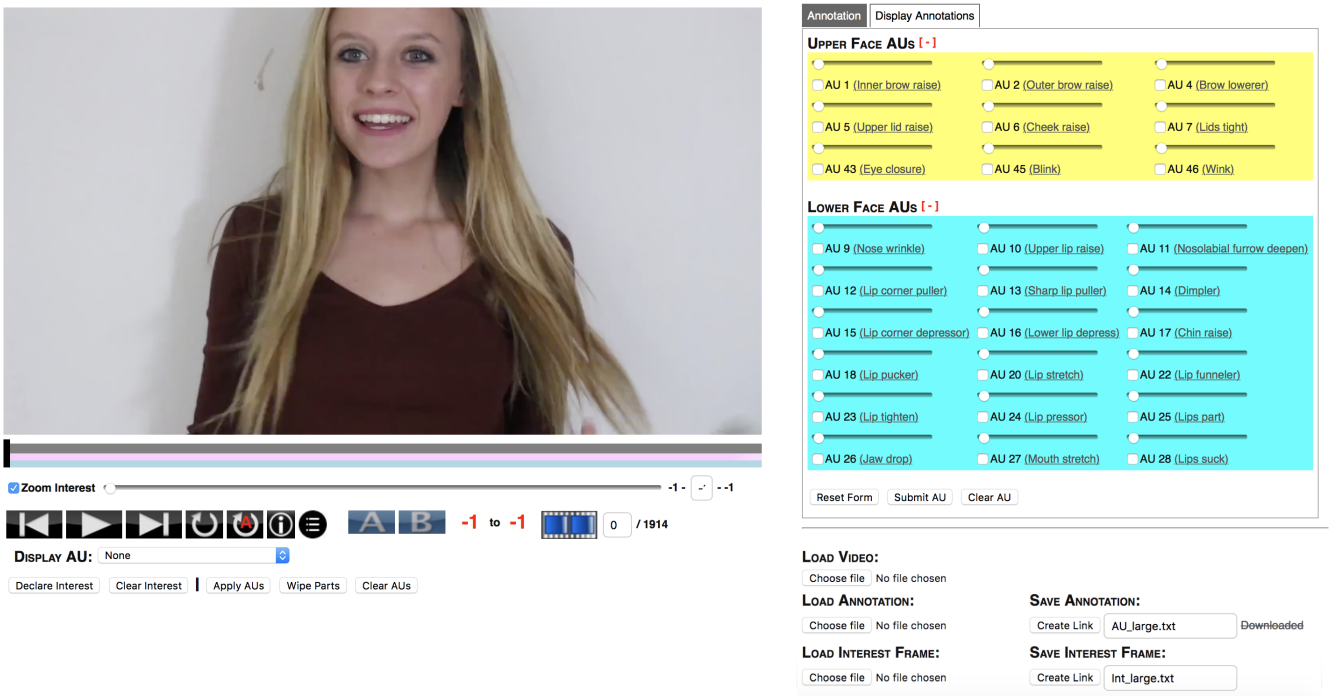


Figure 3.14: The GUI for the Action Unit annotation software. The GUI for the basic expression software was exactly the same; their difference being the titles in the annotation tabs

Chapter 4

Dimensional Affect Analysis in-the-wild

In the past, most of the traditional methods used handcrafted features or shallow learning (e.g., Local Binary Patterns (LBP) [180], LBP on Three Orthogonal Planes (LBP-TOP) [236], Histogram of Oriented Gradients (HoGs) [39], Scale Invariant Feature Transform (SIFT) [187], Multi-scale and Multi-orientation Gabor Filterbanks [132], Non-Negative Matrix Factorization (NMF) [240] and Sparse Learning [241]) for affect analysis and recognition.

However, since 2013, relatively sufficient datasets from real-world settings and scenarios have been collected and used in emotion recognition challenges such as Facial Expression Recognition (FER, 2013) [76] and Emotion Recognition in the Wild (EmotiW) [47–53]; as a consequence, FER was transitioned from lab-controlled to in-the-wild settings.

In the meanwhile, studies in various fields have shifted from utilizing handcrafted features to utilizing deep learning methods, such as features learned by Convolutional Neural Networks and/or Recurrent Neural Networks. This shift was motivated by the striking performance achieved when utilizing these in a variety of computer vision, speech and natural language processing tasks. Likewise, the availability of large amounts of data for facial expressions, increased the use of deep learning techniques for affect recognition in-the-wild.

4.1 Related Work

4.1.1 Baseline Model on Aff-Wild

In 2017, we organised the First Affect-in-the-wild Challenge, using the Aff-Wild database, for dimensional affect recognition. The baseline architecture for the challenge was based on the CNN-M [27] network, as a simple model that could be used to initiate the procedure. In particular, this architecture included the convolutional and pooling parts of CNN-M, having been trained on the FaceValue dataset [3]. On top of that we added a 4096-fully connected layer and a 2-fully connected layer that provided the valence and arousal predictions.

The exact structure of the network is shown in Table 4.1. In total, it consists of 5 convolutional, batch normalization and pooling layers and 2 fully connected (FC) ones. For each convolutional layer the parameters are the filter and the stride, in the form of (filter height, filter width, input channels , output channels/feature maps) and (1, stride height, stride width , 1), respectively, and for the max pooling layer the parameters are the ksize and stride, in the form of (pooling height, pooling width, input channels, output channels) and (1, stride height, stride width , 1), respectively. We follow the TensorFlow's platform notation for the values of all those parameters. Note that the activation function in the convolutional and batch normalization layers is the ReLU one; this is also the case in the first FC layer. The activation function of the second FC layer, which is the output layer, is a linear one.

Table 4.1: Baseline architecture based on CNN-M, showing the values of the parameters of the convolutional and pooling layers and the number of hidden units in the fully connected layers. We follow the TensorFlow's platform notation for the values of all those parameters.

Layer	filter	ksize	stride	padding	no of units
conv 1 batch norm max pooling	[7, 7, 3, 96]	 [1, 3, 3, 1]	[1, 2, 2, 1] [1, 2, 2, 1]	'VALID' 'VALID'	
conv 2 batch norm max pooling	[5, 5, 96, 256]	 [1, 3, 3, 1]	[1, 2, 2, 1] [1, 2, 2, 1]	'SAME' 'SAME'	
conv 3 batch norm	[3, 3, 256, 512]		[1, 1, 1, 1]	'SAME'	
conv 4 batch norm	[3, 3, 512, 512]		[1, 1, 1, 1]	'SAME'	
conv 5 batch norm max pooling	[3, 3, 512, 512]	 [1, 2, 2, 1]	[1, 1, 1, 1] [1, 2, 2, 1]	'SAME' 'SAME'	
fully connected 1					4096
fully connected 2					2

4.1.2 Dimensional Affect Recognition Algorithms on Aff-Wild

The three networks that were top-rated in the First Affect-in-the-wild challenge are briefly reported below, while Table 4.2 compares the acquired results (in terms of CCC and MSE) by these methods and by the baseline network. As one can see, FATAUVA-Net [26] provided the best results in terms of the mean CCC and mean MSE for valence and arousal.

Table 4.2: Concordance Correlation Coefficient (CCC) and Mean Squared Error (MSE) of valence & arousal predictions provided by the methods of the three participating teams and the baseline architecture. A higher CCC and a lower MSE value indicate a better performance.

Methods	CCC		
	Valence	Arousal	Mean Value
MM-Net	0.196	0.214	0.205
FATAUVA-Net	**0.396**	0.282	**0.339**
DRC-Net	0.042	**0.291**	0.167
Baseline	0.150	0.100	0.125

Methods	MSE		
	Valence	Arousal	Mean Value
MM-Net	0.134	**0.088**	0.111
FATAUVA-Net	**0.123**	0.095	**0.109**
DRC-Net	0.161	0.094	0.128
Baseline	0.130	0.140	0.135

In the MM-Net method [127], a variation of a deep convolutional residual neural network (ResNet) [82] was first used for affective level estimation of facial expressions. Then, multiple memory networks were used to model temporal relations between the video frames. Finally, ensemble models were used to combine the predictions of the multiple memory networks, showing that the latter steps improved the initially obtained performance, as far as MSE was concerned, by more than 10%.

In the FATAUVA-Net method [26], a deep learning framework was presented, in which a core layer, an attribute layer, an action unit (AU) layer and a valence-arousal layer were sequentially trained. The core layer was a series of convolutional layers, followed by the attribute layer which extracted facial features. These layers were applied for learning of AUs. Finally, AUs were employed as mid-level representations to estimate the intensity of valence and arousal.

In the DRC-Net method [138], three neural network-based methods which were based on Inception-ResNet [190] modules, redesigned specifically for the task of facial affect estimation, were pre-

sented and compared. These methods were: Shallow Inception-ResNet, Deep Inception-ResNet, and Inception-ResNet with Long Short Term Memory [84]. Facial features were extracted in different scales and both, the valence and arousal, were simultaneously estimated in each frame. Best results were obtained by the Deep Inception-ResNet method.

4.1.3 Transfer Learning & Domain Adaptation

Conventional machine learning methodologies assume that the training and test data are taken from the same domain, such that the input feature space and data distribution characteristics are the same. Under this assumption, transfer learning [192] has been the main approach to train Deep Neural Networks with small amounts of annotated data. Transfer learning uses networks previously trained with large datasets (even of generic patterns) and fine-tunes the whole, or parts of them, using the small training datasets. However, this is not the case in many real-world machine learning scenarios. Due to many factors (e.g., illumination, pose, and image quality), there is always a distribution change or domain shift between two domains that can degrade the performance of the methodologies. Additionally, collecting and annotating datasets for every new task and domain are extremely expensive and time-consuming processes, so that sufficient training data are not always available.

Fortunately, the big data era makes a large amount of data available in other domains and tasks. Mimicking the human vision system, domain adaptation [207] utilises labeled data in one or more relevant source domains to execute new tasks in a target domain. Deep domain adaptation has emerged as a new learning technique to address the lack of massive amounts of labeled data. Compared to traditional methods that learn shared feature subspaces, or reuse important source instances with shallow representations, deep domain adaptation methods leverage deep networks to learn more transferable representations, by embedding domain adaptation in the pipeline of deep learning.

We can draw from real-world non-technical experiences to understand why domain adaptation is possible. Consider an example of two people who want to learn to play the piano. One person has no previous experience playing music, and the other person has extensive music knowledge through playing the guitar. The person with an extensive music background will be able to learn the piano in

a more efficient manner by transferring previously learned music knowledge to the task of learning to play the piano. One person is able to take information from a previously learned task and use it in a beneficial way to learn a related task.

4.1.4 The OMG-Emotion Challenge

The OMG-Emotion Challenge was organised in conjunction with WCCI/IJCNN in 2018, based on the the One-Minute Gradual-Emotional Behavior dataset, targeting single, or multimodal dimensional affect (in terms of valence and arousal) recognition. Some relevant approaches are listed below.

The authors of [163] developed the VNet and ANet models. VNet is a SphereFace [133] network, followed by a BLSTM, followed by a temporal pooling and the output layer. ANet is a VGG16 network with average pooling and accepts as input STFT maps extracted from the audio. In their fusion, the features extracted from VNet's temporal pooling and ANet's average pooling layers, were concatenated and passed to the output layer.

The authors of [199] developed two models. In the first model, denoted as openSMILE + LSTMs, features extracted from audio using openSMILE [64] were passed through six 2-layer LSTMs, each predicting valence, arousal or both; the final prediction was their average. In the second model, denoted as VGG-FACE-BLSTM, the visual modality was used; frames from the utterances were passed through a fixed and pre-trained VGG-FACE followed by a 2-layer BLSTM that gave the final valence prediction.

The authors of [239] developed both single and ensemble networks, consisting of three models. In the first model, denoted as Single Multi-Modal, acoustic features were extracted using openSmile; visual features were extracted from a fixed and pre-trained VGG16 followed by 1-layer LSTM with attention mechanism; visual and acoustic features were passed into an SVM that performed the final predictions. The second model was similar to the first and extracted similar visual and acoustic features, but it also extracted acoustic features from SoundNet. All these features were passed to an SVM that performed the predictions. The late fusion of the two afore-mentioned models, is denoted as Ensemble I; the final predictions were a weighted sum of the models' predictions. The third model

was an end-to-end trained VGG16 followed by 1-layer LSTM with attention mechanism that takes as input only visual data. The late fusion of the three developed models, is denoted as Ensemble II; again the final predictions were a weighted sum of the models' predictions.

4.2 The AffWildNet for Dimensional Affect Recognition

By utilising the Aff-Wild database, we built a novel network that could capture the dynamics and the in-the-wild nature of the database. In the following, we present the developed AffWildNet and evaluate its performance in a large variety of contexts.

At first, AffWildNet is a CNN-RNN network. The CNN part is based on the VGG-Face or ResNet-50 network's convolutional and pooling layers. Low- and middle-level facial features are common and important in both face recognition and facial affect recognition. Therefore we used the VGG-Face network that has been pre-trained with a large dataset for face recognition and therefore many human faces have been used in its construction.

This CNN part is followed by a single FC layer. The inputs of this layer are: a) the outputs of the last pooling layer of the CNN part; b) the facial landmarks, which are directly passed as inputs to this FC layer. As a consequence, this layer has the role to map its two types of inputs to the same feature space, before forwarding them to the RNN part. The facial landmarks, which are provided as additional input to the network, in this way, contribute to boosting the performance of our model. Feeding the landmarks to the network in this way has been a new development for dimensional affect analysis. The output of the fully connected layer is then passed to the RNN part.

The RNN is used in order to model the contextual information in the data, taking into account temporal variations. The RNN is composed of 2-layers, with GRU units in each layer; the first layer processes the FC layer outputs; the second layer is followed by the output layer that gives the final estimates for valence and arousal. GRU units have been chosen instead of LSTM ones as they are less complex, more efficient and as shown in the experimental evaluation provided the best results.

Table 4.3 shows the configuration of the AffWildNet, including the respective number of units for the

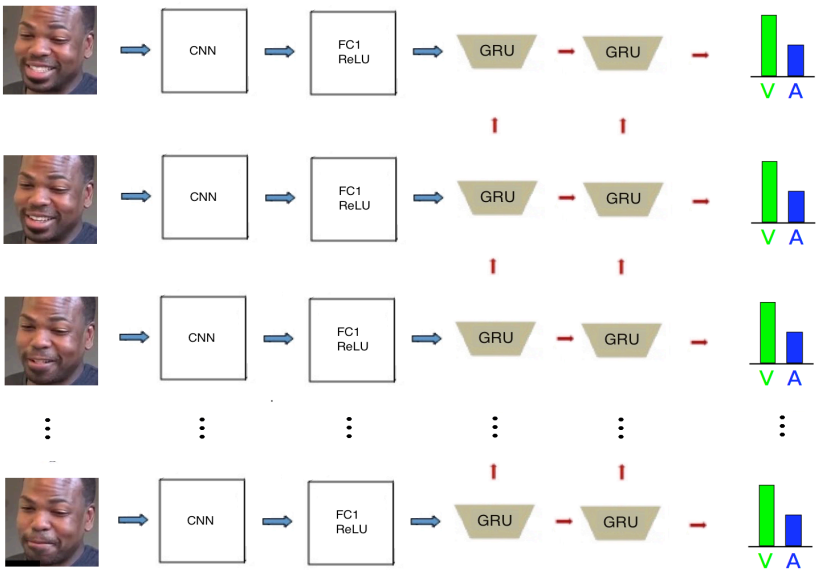


Figure 4.1: The AffWildNet: it consists of convolutional and pooling layers of either VGG-Face or ResNet-50 structures (denoted as CNN), followed by a fully connected layer (denoted as FC1) and two RNN layers with GRU units (V and A stand for valence and arousal respectively).

GRU and the fully connected layers. Additionally, Figure 4.1 illustrates AffWildNet.

Table 4.3: The AffWildNet architecture: the fully connected 1 layer has 4096, or 1500 hidden units, depending on whether VGG-Face or ResNet-50 is used.

block 1	VGG-Face or ResNet-50 conv & pooling parts	
block 2	fully connected 1 dropout	4096 or 1500
block 3	GRU layer 1 dropout	128
block 4	GRU layer 2	128
block 5	fully connected 2	2

Furthermore, defining the loss function used for network training was of great significance as well. The selected loss function was based on the Concordance Correlation Coefficient (CCC) [124], as this was the main evaluation criterion of the Aff-Wild database and Challenge; state-of-the-art research works used to utilize the Mean Squared Error (MSE) as the loss function. CCC is widely used in measuring the performance of dimensional emotion recognition methods, e.g., in the series of AVEC challenges. CCC evaluates the agreement between two time series (e.g., all video annotations and predictions) by scaling their correlation coefficient with their mean square difference. In this way, predictions that are well correlated with the annotations but shifted in value are penalised in proportion

to the deviation. CCC takes values in the range $[-1, 1]$, where $+1$ indicates perfect concordance and -1 denotes perfect discordance. The highest the value of the CCC the better the fit between annotations and predictions, and therefore high values are desired. CCC is defined as follows:

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} = \frac{2s_x s_y \rho_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}, \tag{4.1}$$

where ρ_{xy} is the Pearson Correlation Coefficient (Pearson CC), s_x and s_y are the variances of all video valence/arousal annotations and predicted values, respectively and s_{xy} is the corresponding covariance value.

Consequently, the defined loss function was:

$$\mathcal{L}_{total} = 1 - \frac{\rho_a + \rho_v}{2}, \tag{4.2}$$

where ρ_a and ρ_v is the CCC for arousal and valence, respectively.

Finally, as far as network training is concerned, AffWildNet has been trained as an end-to-end architecture, by jointly training its CNN and RNN parts; previous works used to either train separately the CNN and the RNN parts, or use fixed pre-trained weights for the CNN and train only the RNN part.

4.2.1 Pre-Processing and Network Training Details

Data pre-processing consists of all processing steps that are required for starting the extraction of meaningful features from the data. The usual steps are face detection, face alignment, image resizing and image normalization. At first, we extracted a total of 1,224,100 video frames using the Menpo software [2]. Then we generated face bounding boxes in all video frames. In order to do so, we used the Deformable Part Model (DPM) detector ffld2 [141] that has proven to be highly efficient and accurate for face detection in-the-wild.

For face alignment, we extracted facial landmarks and implemented the Generalized Procrustes Analysis [77]. Facial landmarks are defined as distinctive face locations, such as the corners of the eyes,

centre of the bottom lip, or the tip of the nose. If they are aggregated in sufficient numbers, they can effectively describe the face shape. In our implementations, we used the facial landmark detector inside the dlib library [97] to locate 68 facial landmarks in all frames.

We used as reference and rigid points, 5 anchor points that corresponded to the location of the left eye, right eye, nose and mouth in a prototypical frontal face. For every frame, we used its 5 facial landmarks corresponding to the location of the same facial components; we performed Procrustes transformation, which eliminates in-plane rotation, isotropic scaling and translation, on the coordinates of these 5 landmarks and the coordinates of the 5 landmarks of the frontal face; we imposed this transformation to the whole new frame to perform the alignment.

All cropped and aligned images were then resized to $96 \times 96 \times 3$ pixel resolution and their intensity values were normalized to the range $[-1, 1]$. The input to AffWildNet were the facial images resized to resolution of $96 \times 96 \times 3$ and the facial landmarks. We normalized the facial images' pixel intensities to the range $[-1, 1]$. In order to train the network, we utilized the Adam optimizer algorithm; the batch size was set to 4 and the sequence length to 80; the initial learning rate was set to 0.0001 and was decaying exponentially after 10 epochs; the dropout probability value has been set to 0.5. Training was performed on a single GeForce GTX TITAN X GPU and the training time was about 2-3 days. The platform used for this implementation was Tensorflow.

4.2.2 AffWildNet Performance Evaluation and Ablation Study

Next, we evaluated the performance of AffWildNet, also comparing it to the performance of other CNN and standard CNN-RNN architectures.

For the CNN architectures, we considered the ResNet-50 and VGG-16 networks, pre-trained on the ImageNet [44] dataset that has been broadly used for state-of-the-art object detection. We also considered the VGG-Face network, pre-trained for face recognition on the VGG-Face dataset [161].

The first architecture we utilized was the deep residual network (ResNet) of 50 layers [82], on top of which we stacked a 2-layer fully connected (FC) network. For the first FC layer, best results have

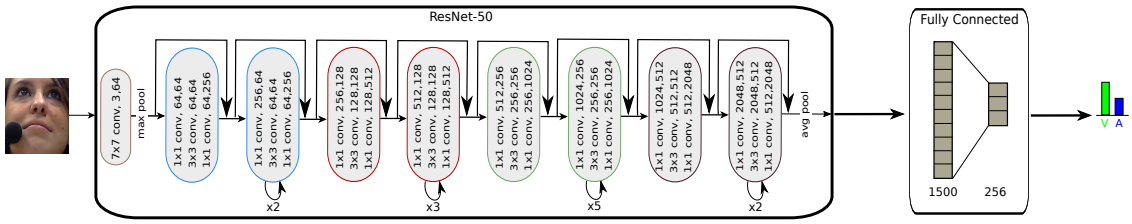


Figure 4.2: The CNN-only architecture for valence and arousal estimation, based on ResNet-50 structure and including two fully connected layers (V and A stand for valence and arousal respectively). Each convolutional layer is in the format: filter height $\times$ filter width, number of input feature maps, number of output feature maps.

been obtained when using 1500 units. For the second FC layer, 256 units provided the best results. An output layer with two linear units followed providing the valence and arousal predictions.

Residual learning was adopted in these models by stacking multiple blocks of the form:

$$\mathbf{o}_k = \mathcal{B}(\mathbf{x}_k, \{\mathbf{W}_k\}) + h(\mathbf{x}_k), \tag{4.3}$$

where $\mathbf{x}_k$, $\mathbf{W}_k$ and $\mathbf{o}_k$ indicate the input, the weights, and the output of layer k, respectively, $\mathcal{B}$ indicates the residual function that is learnt and h is the identity mapping between the residual function and the input. The h identity mapping is a projection of $\mathbf{x}_k$ to match the dimensions of $\mathcal{B}(\mathbf{x}_k, \{\mathbf{W}_k\})$ (done by 1×1 convolutions), as in [82].

The first layer of the ResNet-50 model is comprised of a 7×7 convolutional layer with 64 feature maps, followed by a max pooling layer of size 3×3. Next, there are 4-bottleneck blocks, where a shortcut connection is added after each block. Each of these blocks is comprised of 3 convolutional layers of sizes 1×1, 3×3, and 1×1 with different number of feature maps.

The architecture of the network is depicted in Figure 4.2. Each convolutional layer is in the format: filter height $\times$ filter width, number of input feature maps, number of output feature maps.

The other architecture that we used was based on the convolutional and pooling layers of VGG-Face or VGG-16 networks, on top of which we stacked a 2-layer FC network. For the first and second FC layers, best results have been obtained when using 4096 units. An output layer followed, including two linear units, providing the valence and arousal predictions.

Table 4.4 shows the configuration of the CNN architecture based on VGG-Face or VGG-16. In total, it is composed of thirteen convolutional and pooling layers and three fully connected ones. For all those layers the form of the parameters is the same, as described above in the baseline architecture. We follow the TensorFlow's platform notation for the values of all those parameters. The output number of units is also shown in the Table.

A linear activation function was used in the last FC layer, providing the final estimates. All units in the remaining FC layers used the ReLU activation function. Dropout has been added after the first FC layer, in order to avoid over-fitting. The architecture of the network is depicted in Figure 4.3.

Table 4.4: CNN architecture based on VGG-Face/VGG-16, showing the values of the parameters of the convolutional and pooling layers and the number of hidden units in the fully connected layers. We follow the TensorFlow's platform notation for the values of all those parameters.

Layer	filter	ksize	stride	padding	no of units
conv 1	[3, 3, 3, 64]		[1, 1, 1, 1]	'SAME'	
conv 2	[3, 3, 64, 64]		[1, 1, 1, 1]	'SAME'	
max pooling		[1, 2, 2, 1]	[1, 2, 2, 1]	'SAME'	
conv 3	[3, 3, 64, 128]		[1, 1, 1, 1]	'SAME'	
conv 4	[3, 3, 128, 128]		[1, 1, 1, 1]	'SAME'	
max pooling		[1, 2, 2, 1]	[1, 2, 2, 1]	'SAME'	
conv 5	[3, 3, 128, 256]		[1, 1, 1, 1]	'SAME'	
conv 6	[3, 3, 256, 256]		[1, 1, 1, 1]	'SAME'	
conv 7	[3, 3, 256, 256]		[1, 1, 1, 1]	'SAME'	
max pooling		[1, 2, 2, 1]	[1, 2, 2, 1]	'SAME'	
conv 8	[3, 3, 256, 512]		[1, 1, 1, 1]	'SAME'	
conv 9	[3, 3, 512, 512]		[1, 1, 1, 1]	'SAME'	
conv 10	[3, 3, 512, 512]		[1, 1, 1, 1]	'SAME'	
max pooling		[1, 2, 2, 1]	[1, 2, 2, 1]	'SAME'	
conv 11	[3, 3, 512, 512]		[1, 1, 1, 1]	'SAME'	
conv 12	[3, 3, 512, 512]		[1, 1, 1, 1]	'SAME'	
conv 13	[3, 3, 512, 512]		[1, 1, 1, 1]	'SAME'	
max pooling		[1, 2, 2, 1]	[1, 2, 2, 1]	'SAME'	
fully connected 1					4096
dropout					
fully connected 2					4096
fully connected 3					2

In order to consider the contextual information in the data, we developed a CNN-RNN architecture, in which the RNN part was fed with the outputs of either the first, or the second fully connected layer of the respective CNN networks.

The structure of the RNN, which we examined, consisted of one or two hidden layers, with $100 - 150$ units, following the LSTM neuron model with peephole connections. Using one fully connected layer

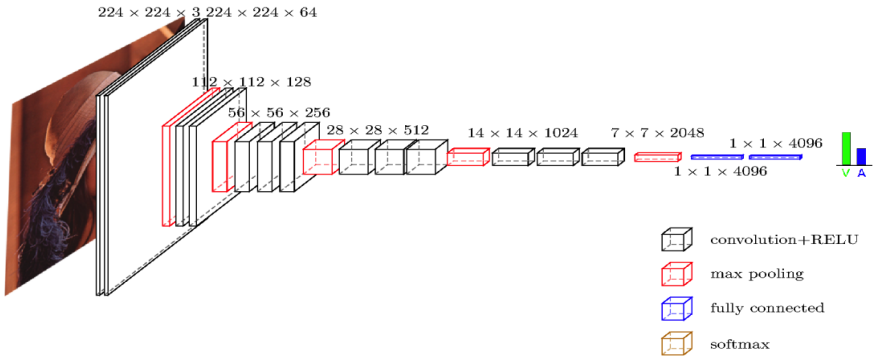


Figure 4.3: The CNN-only architecture for valence and arousal estimation, based on VGG-Face structure (V and A stand for valence and arousal respectively).

in the CNN part and two hidden layers in the RNN part has been found to provide the best results. An output layer followed, including two linear units, providing the valence and arousal predictions.

Table 4.5 summarizes the CCC and MSE values obtained when applying all the above architectures, to the Aff-Wild test set (Table 3.3 shows the total number of frames in the training and testing sets of Aff-Wild). It shows the improvement in the CCC and MSE values obtained when using the AffWildNet compared to all other developed architectures. This improvement clearly indicates the ability of the AffWildNet to better capture the dynamics in Aff-Wild. Table 4.5 also compares the performance of AffWildNet to that of the FATAUVA-NET, which was the winner of the Aff-Wild Challenge. AffWildNet outperformed this network, as well.

Table 4.5: CCC and MSE based evaluation of valence & arousal predictions provided by: 1) the CNN architecture when using three different pre-trained networks for initialization (VGG-16, ResNet-50, VGG-Face), 2) the winner of Aff-Wild Challenge, FATAUVA-NET and 3) the VGG-Face-LSTM and AffWildNet architectures (2 RNN layers with 128 units each). A higher CCC and a lower MSE value indicate a better performance.

	CCC		
	Valence	Arousal	Mean Value
FATAUVA-Net	0.40	0.28	0.34
VGG-16	0.40	0.30	0.35
ResNet-50	0.43	0.30	0.37
VGG-Face	**0.51**	**0.33**	**0.42**
VGG-Face-LSTM	0.52	0.38	0.45
AffWildNet	**0.57**	**0.43**	**0.50**

Next, we performed an ablation study on the use of various numbers of hidden layers and hidden units per layer when training and testing the AffWildNet. Some characteristic selections and their

	MSE		
	Valence	Arousal	Mean Value
FATAUVA-Net	0.12	0.10	0.11
VGG-16	0.13	0.11	0.12
ResNet-50	0.11	0.11	0.11
VGG-Face	**0.10**	**0.08**	**0.09**
VGG-Face-LSTM	0.10	0.09	0.10
AffWildNet	**0.08**	**0.06**	**0.07**

corresponding performances are shown in Table 4.6. It can be seen that the best results have been obtained when the RNN part of the network consisted of 2 layers, each of 128 hidden units.

Table 4.6: Obtained CCC values for valence & arousal estimation, when changing the number of hidden units & hidden layers in the VGG-Face-GRU architecture. A higher CCC value indicates a better performance.

CCC	1 Hidden Layer		2 Hidden Layers	
Hidden Units	Valence	Arousal	Valence	Arousal
100	0.44	0.36	0.50	0.41
128	0.53	0.40	**0.57**	**0.43**
150	0.46	0.39	0.51	0.41

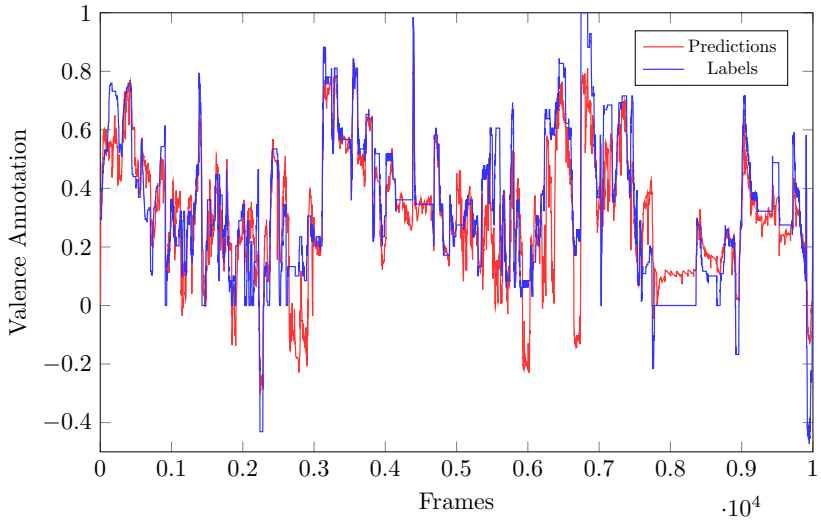
Next, we performed an ablation study on the use of the 68 2-D landmarks as additional input to our developed network. The results are summarized in Table 4.7, which shows that there is a notable improvement in the performance, when we also used the 68 2-D landmark positions as input data.

Table 4.7: CCC and MSE based evaluation of valence & arousal predictions provided by the AffWildNet when landmarks were or were not given as input to the network. A higher CCC and a lower MSE value indicate a better performance.

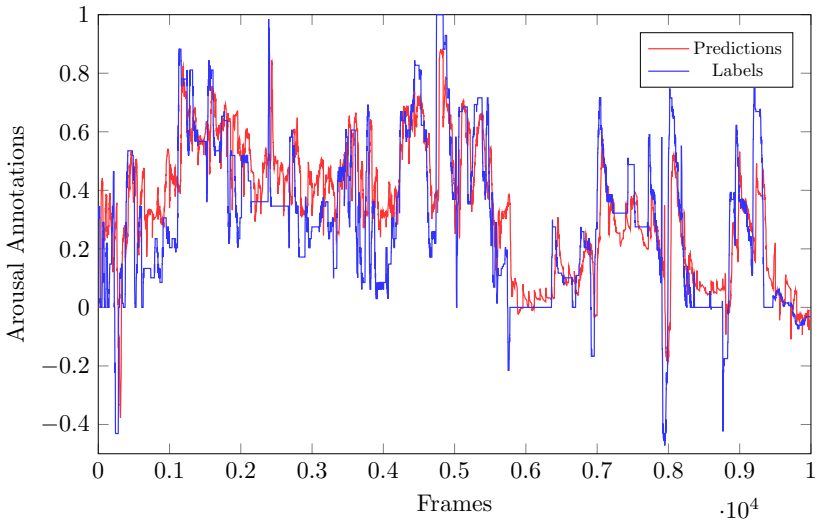
AffWilldNet	With Landmarks		Without Landmarks	
	Valence	Arousal	Valence	Arousal
CCC	**0.57**	**0.43**	0.50	0.41
MSE	**0.08**	**0.06**	0.10	0.09

In Figures 4.4(a) and 4.4(b), we qualitatively illustrate some of the obtained results by comparing a segment of the obtained valence/arousal predictions to the ground truth values, in 10000 consecutive frames of test data.

Moreover, in Figures 4.5(a) and 4.5(b), we illustrate, in the 2-D valence & arousal space, the histograms of the ground truth labels of the test set and the corresponding predictions of AffWildNet.



(a) valence



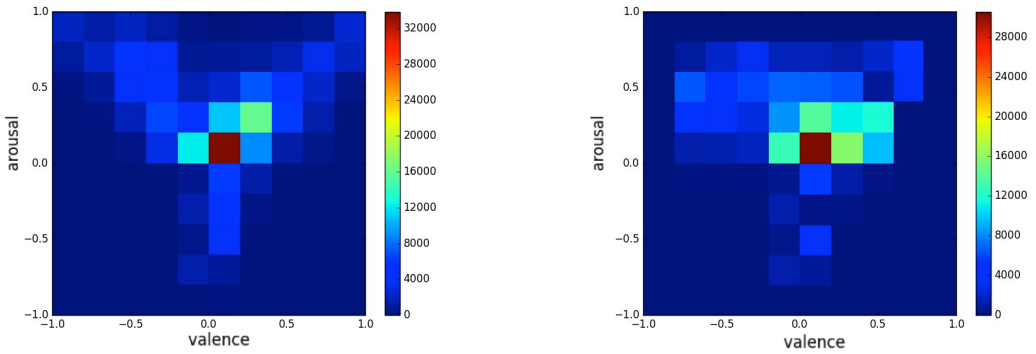
(b) arousal

Figure 4.4: Predictions vs Labels for (a) valence and (b) arousal over a video segment of the Aff-Wild.

The results shown in Table 4.5 and in the above Figures verify the excellent performance of AffWild-Net. They also show that it greatly outperformed all methods submitted to the Aff-Wild Challenge.

4.3 Robust Prior for Dimensional & Categorical Affect Analysis

In this Section, we focus on the use of domain adaptation in the context of both dimensional affect recognition and FER. In the past, there did not exist a lot of FER databases and from the existing ones there was no in-the-wild. Therefore, FER approaches that were using DNNs were pre-training



(a) annotations (b) predictions

Figure 4.5: Histogram in the 2-D valence & arousal space of: (a) annotations and (b) predictions of AffWildNet, on the test set of the Aff-Wild Challenge.

the networks on large but diverse and generic datasets, such as ImageNet [44]. In 2013, the first in-the-wild FER database was developed (Facial Expression Recognition 2013, FER-2013) and FER approaches started using this database. However FER2013 proved to contain noisy labels [13]. What is more, FER2013 is a very small database, containing around $38,000$ samples and thus training of deep networks on this database is prone to overfitting. To mitigate this problem, works exploited the development of large-scale in-the-wild face recognition datasets, e.g., Celebrity Face in the Wild (CFW) [232], FaceScrub dataset [155] and VGG-FACE [161]. It was shown in [100] that pre-training on larger face recognition data positively affects the emotion recognition accuracy, and further fine-tuning with additional FER datasets can help improve the performance of DNNs.

With the development of the first and largest in-the-wild database with dimensional emotion annotation, i.e., Aff-Wild, and the corresponding development of AffWildNet, we were the first to perform domain adaptation experiments from the in-the-wild dimensional affect recognition domain (using AffWildNet trained on the dimensionally annotated Aff-Wild) to other, either in-the-wild or lab-controlled, dimensional affect recognition domains (utilizing the RECOLA and AFEW-VA datasets). Additionally, to the best of our knowledge, we were the first to perform domain adaptation experiments from in-the-wild dimensional affect recognition domain (again using AffWildNet) to the categorical emotion recognition domain (utilizing the AFEW database as used in the EmotiW Challenge).

It is shown that the AffWildNet has been capable of generalising its knowledge to other emotion recognition datasets and contexts. By learning complex and emotionally rich features of the AffWild,

the AffWildNet constitutes a robust prior for both dimensional and categorical emotion recognition. It is the first time that state-of-the-art performance has been achieved in this way.

4.3.1 AffWildNet as Prior for Valence and Arousal Prediction

Experimental Results on the RECOLA database

In this subsection, we fine-tuned AffWildNet on RECOLA and for comparison purposes, we also trained on RECOLA an architecture comprising a ResNet-50 and a 2-layer GRU stacked on top (let us call it ResNet-GRU network). Table 4.8 shows the results, according to CCC score, as our minimization loss was depending on this metric. It is clear that the performance on both arousal and valence of the fine-tuned AffWildNet model was much higher than the performance of the ResNet-GRU model.

Table 4.8: CCC based evaluation of valence & arousal predictions provided by the fine-tuned AffWildNet and the ResNet-GRU on the RECOLA test set. A higher CCC value indicates a better performance.

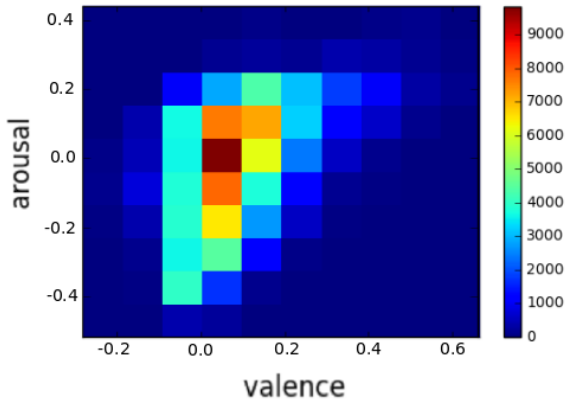
	CCC	
	Valence	Arousal
Fine-tuned AffWildNet	**0.526**	**0.273**
ResNet-GRU	0.462	0.209

To further demonstrate the benefits of using AffWildNet to predict valence and arousal, we demonstrate a histogram in the 2-D valence & arousal space of the annotations (Fig. 4.6(a)) and predictions of the fine-tuned AffWildNet (Fig. 4.6(b)) for the whole test set of RECOLA.

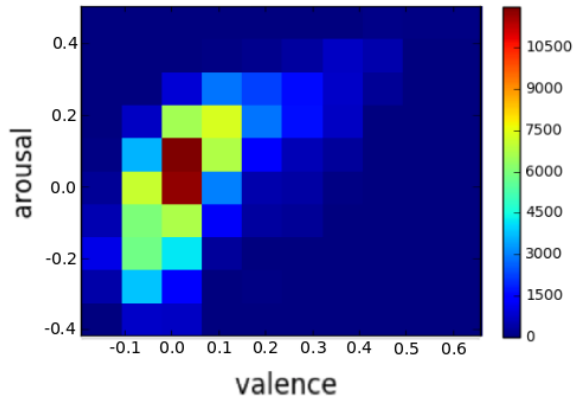
Finally, we also illustrate in Figs. 4.7(a) and 4.7(b) the network prediction and ground truth for one test video of RECOLA, for the valence and arousal dimensions, respectively.

Experimental Results on the AFEW-VA database

In this subsection, we focus on recognition of affect in the AFEW-VA database; these annotation is somewhat different from the annotation of the Aff-Wild database. In particular, the labels of the



(a)

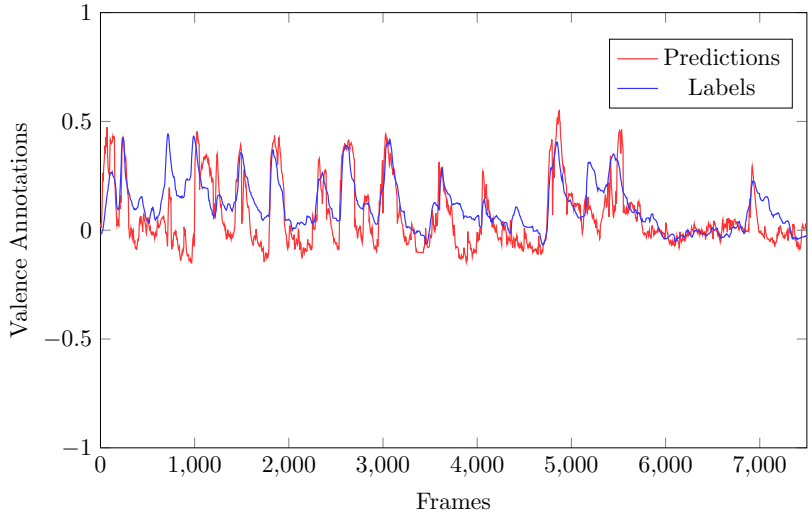


(b)

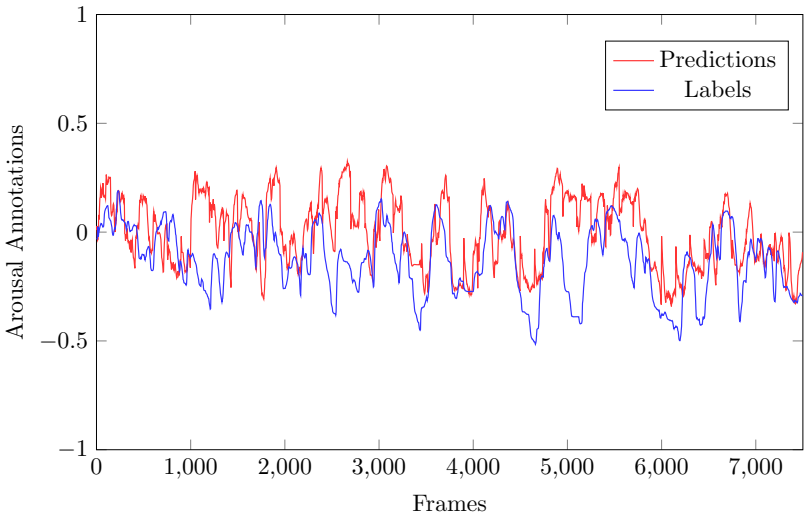
Figure 4.6: Histogram in the 2-D valence & arousal space of (a) annotations and (b) predictions for the test set of the RECOLA database.

AFEW-VA database are in the range $[-10, +10]$, while the labels of the Aff-Wild database are in the range $[-1, +1]$. To tackle this problem, we scaled the range of the AFEW-VA labels to $[-1, +1]$. Moreover, differences were observed, due to the fact that the labels of the AFEW-VA are discrete, while the labels of the Aff-Wild are continuous. Figure 4.8(a) shows the discrete valence and arousal values of the annotations in AFEW-VA database, whereas Figure 4.8(b) shows the corresponding histogram in the 2-D valence & arousal space.

We then performed fine-tuning of the AffWildNet to the AFEW-VA database and tested the performance of the generated network. Similarly to [118], we used a 5-fold person-independent cross-validation strategy. Table 4.9 shows a comparison of the performance of the fine-tuned AffWildNet with the best results reported in [118]. Those results are in terms of the Pearson CC. It can be easily seen that the fine-tuned AffWildNet greatly outperformed the best method reported in [118].



(a)

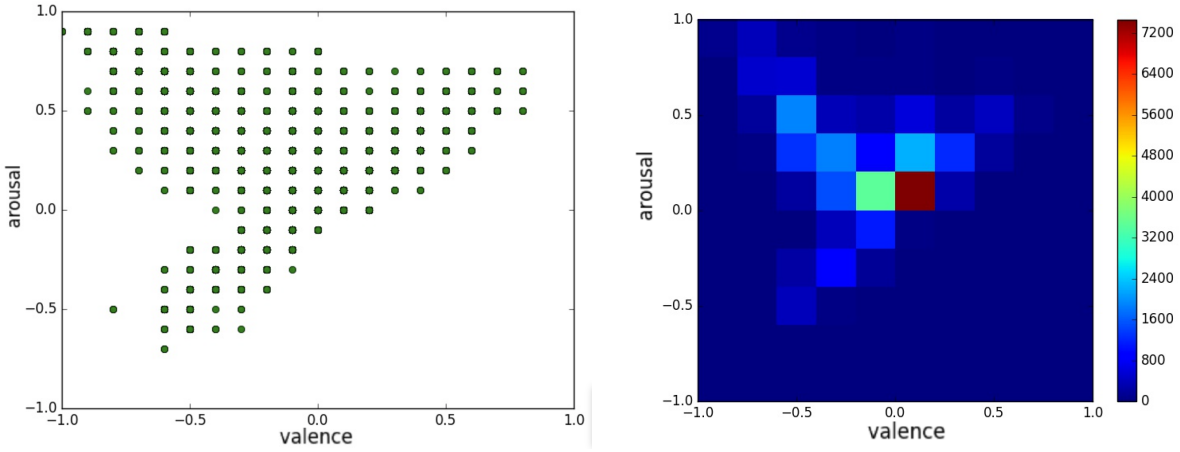


(b)

Figure 4.7: Fine-tuned AffWildNet's Predictions vs Labels for (a) valence and (b) arousal for a single test video of the RECOLA database.

For comparison purposes, we also trained a CNN network on the AFEW-VA database. This network's architecture was based on the convolution and pooling layers of VGG-Face followed by 2 fully connected layers with 4096 and 2048 hidden units, respectively. As shown in Table 4.10, the performance of the fine-tuned AffWildNet, in terms of CCC, greatly outperformed this network as well.

All these verify that AffWildNet can be used as a pre-trained network to yield excellent results across different dimensional databases.



(a) (b)

Figure 4.8: AFEW-VA database's: (a) discrete values of annotations and (b) histogram of annotations in the 2-D valence & arousal space.

Table 4.9: Pearson Correlation Coefficient (Pearson CC) based evaluation of valence & arousal predictions provided by the best architecture in [118] vs our AffWildNet fine-tuned on the AFEW-VA. A higher Pearson CC value indicates a better performance.

Group	Pearson CC	
	Valence	Arousal
best of [118]	0.407	0.45
Fine-tuned AffWildNet	**0.514**	**0.575**

4.3.2 AffWildNet as Prior for Facial Expression Recognition

To further show the strength of AffWildNet, we used AffWildNet - which is trained for dimensional emotion recognition task - in a very different problem, i.e., categorical in-the-wild emotion recognition, focusing on the EmotiW 2017 Grand Challenge. To tackle categorical emotion recognition, we modified the AffWildNet's output layer to include 7 neurons (one for each basic emotion category) and performed fine-tuning on the AFEW 5.0 dataset.

In this study, we compared the fine-tuned AffWildNet's performance with that of other state-of-the-

Table 4.10: CCC based evaluation of valence & arousal predictions provided by the CNN architecture based on VGG-Face and the fine-tuned AffWildNet on the AFEW-VA training set. A higher CCC value indicate a better performance.

	CCC AFEW-VA	
	Valence	Arousal
only CNN	0.44	0.474
Fine-tuned AffWildNet	**0.515**	**0.556**

art CNN and CNN-RNN networks; the CNN part of which was based on the ResNet 50, VGG-16 and VGG-Face architectures, trained on the same AFEW 5.0 dataset. The accuracy achieved by all networks on the validation set of the EmotiW 2017 Grand Challenge are shown in Table 4.11. A higher accuracy value indicates better performance for the model. We can easily see that the AffWildNet outperformed all those other networks in terms of total accuracy.

Table 4.11: Accuracies on the EmotiW validation set obtained by different CNN and CNN-RNN architectures vs the fine-tuned AffWildNet. A higher accuracy value indicates better performance.

Architectures	Accuracy							
	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	Total
VGG-16	0.327	0.424	**0.102**	**0.093**	0.476	0.138	**0.133**	0.263
VGG-16 + RNN	0.431	0.559	0.026	0.07	0.444	0.259	0.044	0.293
ResNet	0.31	0.153	0.077	0.023	0.534	0.207	0.067	0.211
ResNet + RNN	0.431	0.237	0.077	0.07	0.587	0.155	0.089	0.261
VGG-Face + RNN	0.552	0.593	0.026	0.047	**0.794**	0.259	0.111	0.384
fine-tuned AffWildNet	**0.569**	**0.627**	0.051	0.023	0.746	**0.709**	0.111	**0.454**

We should note that:

(i) the AffWildNet was trained to classify only video frames (and not audio) and then video classification based on frame aggregation was performed;

(ii) the cropped faces provided by the challenge were only used (and not our own detection and/or normalization procedure);

(iii) no data-augmentation, post-processing of the results or ensemble methodology have been conducted.

It should also be mentioned that the fine-tuned AffWildNet performance, in terms of total accuracy, is:

(i) much higher than the baseline total accuracy of 0.3881 reported in [47]

(ii) better than all vanilla architectures' performance that were reported by the three winning methods in the audio-video emotion recognition EmotiW 2017 Grand Challenge [88] [100] [205]

Table 4.12: Overall accuracy of the best architectures of the three winning methods in the EmotiW 2017 Grand Challenge, reported on the validation set, vs that of fine-tuned AffWildNet. A higher accuracy value indicates better performance.

Group	Architecture	Total Accuracy		
		Original	After Fine-Tuning on FER2013	Data augmentation
[88]	DenseNet-121	0.414	-	-
	HoloNet	0.41		
	ResNet-50	0.418		
[100]	VGG-Face	0.379	0.483	-
	FR-Net-A	0.337	0.446	-
	FR-Net-B	0.334	**0.488**	-
	FR-Net-C	0.376	0.452	-
	LSTM + FR-NET-B	-	0.465	**0.504**
[205]	Weighted C3D (no overlap)	-	-	0.421
	LSTM C3D (no overlap)			0.432
	VGG-Face			0.414
	VGG-LSTM 1 layer			0.486
Our	Fine-tuned AffWildNet	**0.454**	-	-

(iii) comparable and better in some cases than the rest of the results obtained by the three winning methods [88] [100] [205]

The above are shown in Table 4.12. Those results verify that the AffWildNet was appropriately fine-tuned and successfully used for dimensional, as well as for categorical emotion recognition.

4.4 Multi-Component Extensions of AffWildNet

Next we addressed the issue of estimating valence and arousal utilising the One-Minute-Gradual Emotion Dataset (OMG-Emotion Dataset), focusing only on visual information. We present novel extensions of the AffWildNet architecture that provided best performance in valence and arousal estimation. It should be mentioned that the submissions we made to the OMG-Emotion Challenge were ranked at second position for valence estimation [112, 115].

The first extension was to extract latent information from the trained AffWildNet and use them for improving its performance in affect recognition. In general, features extracted from low CNN layers

contain rich, complete and time varying information, whilst high-level features are highly specific and characteristic of the specific problem studied. Taking this into account, we have developed and used CNN plus Multi-RNN networks; these networks extract low-, mid- and high- level features from different layers of the CNN and pass them as inputs to RNNs. The best performing networks were split into two different types based on the adopted methodology: the first, referred as CNN-1RNN, concatenates the extracted features from 3 CNN layers and passes them to a single RNN, whereas the other, referred as CNN-3RNN, processes them independently through 3 RNN subnets.

Our work deviates from others, such as [28, 43, 130], that either: i) use standard CNN-RNN networks in which the output of the CNN is passed to the RNN, or ii) apply ensemble methodologies, using features extracted from many CNN networks (but not using features from multiple layers of the same network) and fusing them.

Both facial images and landmarks (after applying a Procrustes Analysis) were provided as inputs to these architectures. Additionally, ensemble formulations were also developed, using different levels of fusion (Model- or Decision-level) in the proposed architectures; these formulations are shown to further boost the obtained performance. In model-level fusion, our approach was to perform fusion through an RNN instead of a typical fully connected layer.

Another contribution was the adaptation of AffWildNet to the OMG-Emotion dataset characteristics and in particular to the dataset's annotation at utterance level. To deal with this, we split each utterance into sequences, which were individually processed by the above architectures. The mean or median of the predicted valence-arousal values were computed per sequence. Then, the means/medians were averaged at utterance level to provide the final valence and arousal estimates. This procedure deviates from related works that uniformly (or randomly) sample a constant number of frames from each utterance, assign to each of them the annotation value of the utterance and compute the prediction per frame [43].

An additional contribution was the pre-training of the proposed architectures on the large-scale emotionally rich Aff-Wild database and on its larger extension, the Aff-Wild2. Other works [163, 199, 239] used networks that were not pre-trained on same task (i.e., valence-arousal estimation), but on other tasks (face recognition, object detection). The pre-training on these specific databases provided our

developed architectures with the ability to effectively capture the dynamics of the OMG-Emotion in-the-wild dataset and thus provided a better performance.

Training of the CNN networks was performed as shown in Fig.4.12. In more detail, each CNN was provided with an input sequence and was trained to predict, for each frame in the sequence, the respective valence-arousal pair of values. The 68 facial landmarks (per each frame of the input sequence) were also provided as additional inputs to the CNN networks. The final valence (arousal) prediction was computed as the mean, or median (both approaches were considered) of the per-frame valence (arousal) values in that sequence.

4.4.1 CNN-3RNN networks

The CNN-3RNN networks include the convolutional and pooling layers of VGG-FACE, followed by a fully connected layer of 4096 units. The 68 facial landmarks were concatenated with the features extracted from the last pooling layer of VGG-FACE and were passed to this fully connected layer. Then, low-, mid- and high-level features were extracted and each one was processed by a 2-layer GRU network that predicted the valence and arousal values. Each GRU layer comprised 128 units. The CNN-3RNN networks were provided with an input sequence of frames (and the corresponding landmarks of each frame), predicting, for each frame, the valence-arousal values; their mean, or median constitute the final estimates.

Fig.4.9, presents an example of CNN-3RNN networks, named CNN-3RNN-2nd-pool_last-pool_fc. In this network: i) the features extracted from the fully connected layer are passed as input to a RNN network, denoted RNN_1 in Fig.4.9; ii) the features extracted from the last pooling layer (before being concatenated with the landmarks) are passed as input to a second RNN network, denoted RNN_2 in Fig.4.9; iii) the features extracted from the second pooling layer (following the fourth convolutional layer) are passed as input to another RNN network, denoted RNN_3 in Fig.4.9. Fig.4.10 depicts the exact structure of the afore-mentioned RNN_i, $i \in \{1,2,3\}$, networks. All networks have the same structure; a 2-layer GRU network, with each layer having 128 units. Next, the outputs of the 3 RNNs are concatenated and passed to the output layer that performs the valence-arousal prediction.

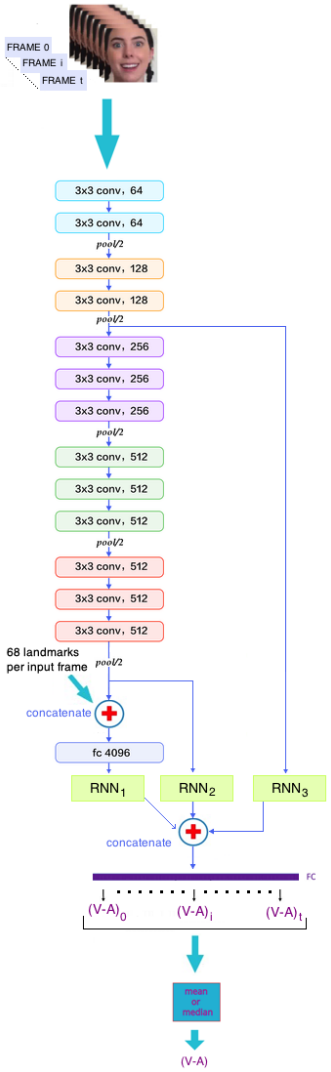


Figure 4.9: The CNN-3RNN-2nd-pool_last-pool_fc. It provides a valence-arousal (V-A) estimate per input sequence of consecutive frames. The '68 landmarks' are concatenated with the features of the last 'pool' layer and passed as input to the 'fc' layer. This architecture provided the best results.

4.4.2 CNN-1RNN Networks

The CNN-1RNN types of networks consisted of the convolutional and pooling layers of VGG-FACE, followed by a fully connected layer of 4096 units. The 68 facial landmarks were concatenated with the features extracted from the last pooling layer of VGG-FACE and were passed to this fully connected layer. Then, low-, mid- and high-level features were extracted, concatenated and passed to a 2-layer GRU network that predicted the valence and arousal values. Each GRU layer comprised 128 units. Similarly to the other architectures described above, the CNN-1RNN networks were provided with an input sequence of frames (and the corresponding landmarks of each frame), predicting, for each frame, the valence-arousal values; their mean, or median, were the final estimates.

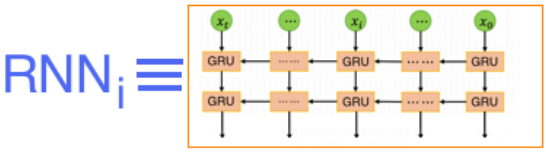


Figure 4.10: Structure of each RNN network in the CNN-3RNN architecture displayed in Fig. 4.9.

Fig.4.11 presents one example of CNN-1RNN networks, which we call CNN-1RNN-2nd-pool_last-pool_fc. In this network, the features extracted from: i) the second pooling layer (following the 4th convolutional), ii) the last pooling layer (following the 13th convolutional and before being concatenated with the landmarks) and iii) the fully connected layer, are concatenated and passed to the RNN.

4.4.3 Ensemble Methodology

In this Subsection we describe an ensemble approach which fuses the developed networks at: i) Model-level and ii) Decision-level. Model-level fusion is based on concatenating the high level features extracted by different networks, whilst Decision-level fusion is based on weighted averaging the predictions provided by different networks. On the one hand side, Model-level fusion takes advantage of the mutual information in the data. On the other hand side, the averaging procedure in Decision-level fusion reduces variance in the ensemble regressor (thus achieving higher robustness), while preserving the relative importance of each individual model.

Model-level Fusion

Let us consider the CNN-1RNNs and CNN-3RNNs described in the previous Subsection. We concatenate the outputs of all the RNNs in the above networks and provide them, as input, either: i) to another single RNN layer with 128 GRU units, or ii) to a fully connected layer with 128 units; the output layer follows. We denote the resulting networks as Model-level Fusion + RNN and Model-level Fusion + FC, respectively. Similarly to the previous Subsections, for each frame in the input sequence of frames, this model-level fusion network predicts the valence-arousal values and then computes their mean, or median, as final estimates.

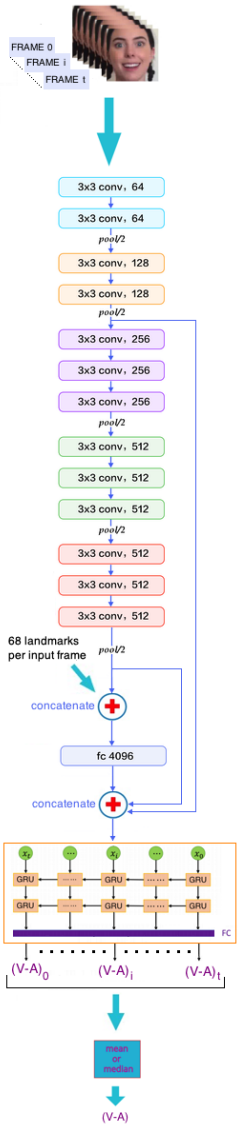


Figure 4.11: The CNN-1RNN-2nd-pool_last-pool_fc architecture. It provides a valence-arousal (V-A) estimate per input sequence of consecutive frames. The '68 landmarks' are concatenated with the features of the last 'pool' layer and passed as input to the 'fc' layer.

Decision-level Fusion

Let us consider again the CNN-1RNNs and CNN-3RNNs described above. The final valence (arousal) estimate $O_v^{dec.-level}$ ($O_a^{dec.-level}$), is computed as a weighted average of the final valence (arousal) estimates, $o_v^n(o_a^n)$, of these networks; each weight is proportional to the corresponding network performance on the validation set:

$$O_i^{dec.-level} = \frac{1}{\sum_n t_i^n} \sum_n t_i^n \cdot o_i^n, \tag{4.4}$$

where $i \in \{v, a\}$ (v stands for valence, a stands for arousal), t_i^n is equal to the Concordance Correlation Coefficient (CCC), ρ_i, for valence or arousal, computed on the validation set, with n denoting the CNN-1RNNs or CNN-3RNNs; the CCC has been the evaluation criterion of the OMG-Emotion Challenge, taking values in $[-1, 1]$ and is defined as follows :

$$\rho_i = \frac{2s_{i,xy}}{s_{i,x}^2 + s_{i,y}^2 + (\bar{x}_i - \bar{y}_i)^2}, \tag{4.5}$$

where $i \in \{v, a\}$, $s_{i,x}$ and $s_{i,y}$ are the variances of the valence/arousal labels and predicted values respectively, $\bar{x}_i$ and $\bar{y}_i$ are the corresponding mean values and $s_{i,xy}$ is the covariance value.

4.4.4 Experimental Study on the OMG-Challenge

In all developed CNN plus Multi-RNN architectures, dropout with 0.5 probability value was applied to the fully connected layers that were on top of the convolutional and pooling layers of the CNN networks (VGG-FACE, ResNet-50). Additionally, dropout with 0.8 probability value was applied after the first GRU layer of the RNNs.

These networks were first pre-trained either on the Aff-Wild or the Aff-Wild2 database and then trained on the OMG-Emotion training set.

For training these networks, we used a batch size of 4 and sequence length of 80 consecutive frames. The training was end-to-end, with a learning rate of either 10^{-4} or 10^{-5}. All networks were trained using Tensorflow on a Quadro GV100 Volta GPU and the training time was about a day.

Since the evaluation criterion of the OMG-Emotion Challenge was the CCC, our loss function was based on that criterion and was defined as:

$$\mathcal{L}_{total} = 1 - \frac{\rho_a + \rho_v}{2}, \tag{4.6}$$

where ρ_a and ρ_v are the CCC for the arousal and valence.

Finally, for all developed architectures, a chain of post-processing steps was applied. These steps included: i) median filtering of the - per frame - predictions within a sequence and ii) smoothing of the - per utterance - predictions (especially to those that consisted of too few frames). Any of these post-processing steps was kept when an improvement was observed on the CCC over the validation set, and applied then, with the same configuration to the test partition.

In all conducted experiments, best results were obtained when the final estimates were the median of the, per frame, valence and arousal estimates within a sequence. It should also be stated, that all reported results refer to the test set. In our developments, we trained the DNNs with the training set, evaluated them on the respective validation set and selected the best networks according to the validation performance. There were no significant differences between training the DNN multiple times and then averaging the predictions, or using a 10-fold cross validation (training on 8 folds, testing on the remaining 2 and in the end averaging the predictions of the networks).

We examined to include a level of encoding for matching the size of landmarks with the size of the CNN features before fusing them. We first passed the 68 landmarks to a fully connected layer of 512, 1024, or 2048 units and then fused this output with the features extracted from the CNN. However, we did not notice any significant difference in performance, although the developed architectures were more complex and bigger in terms of learnable parameters.

Evaluation of CNN plus Multi-RNN Extensions of AffWildNet

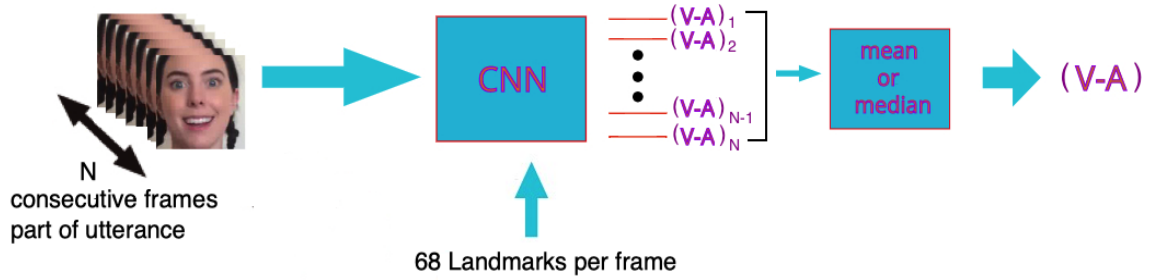


Figure 4.12: Standard CNN structures providing a single valence-arousal (V-A) estimate per input sequence of consecutive frames. They can be any of the VGG-FACE, ResNet-50 and DenseNet-121 networks. The 68 landmarks are concatenated with the extracted features from the last pooling layer of the CNN component and are passed to the fully connected layer that precedes the output layer.

Comparison with standard Deep Neural Architectures First we compared the CNN plus Multi-RNN architectures to three state-of-the-art CNN networks: VGG-Face, ResNet-50 and DenseNet-121. These networks were pre-trained either on the Aff-Wild or the Aff-Wild2 database and then trained on the OMG-Emotion training set. To design the structure of these networks, we took into account the procedure used to annotate the OMG-Emotion dataset. According to this, each utterance was labeled with a single pair of valence and arousal values. We split each utterance into smaller parts-sequences, each consisting of the same number of consecutive frames. Then, we assigned to each of those parts-sequences of frames, the label of the corresponding utterance.

Training of the CNN networks was performed as shown in Fig.4.12. In more detail, each CNN was provided with an input sequence and was trained to predict, for each frame in the sequence, the respective valence-arousal pair of values. The 68 facial landmarks (per each frame of the input sequence) were also provided as additional inputs to the CNN networks. The final valence (arousal) prediction was computed as the mean, or median (both approaches were considered) of the per-frame valence (arousal) values in that sequence.

In Fig.4.12, the CNN structure can be any of the VGG-FACE, ResNet-50 and DenseNet-121 ones. In the VGG-FACE CNN case, the landmarks were concatenated with the outputs of the last pooling layer of the network and were given as input to the first fully connected layer, that consisted of 4096 units. In this way, both outputs and landmarks were mapped to the same feature space, before performing the prediction. In the ResNet-50 (and DenseNet-121) case, the landmarks were concatenated with the averaged pooled features of the ResNet-50 (DenseNet-121) network and were given as input to a fully connected layer consisting of 1500 units. This layer was followed by the output layer which provided the final estimates for valence-arousal pair.

Then we compared the CNN plus Multi-RNN architectures to standard CNN plus RNN architectures, so as to take into account the contextual information in the data and more specifically the temporal dependencies of facial expressions in each utterance. In these architectures, the output of the CNN's last pooling layer is being fed to a fully connected layer, whose output constitutes the input of the RNN layers. These architectures were pre-trained on either the Aff-Wild, or the Aff-Wild2 databases. We then used two different strategies for training these architectures: i) keeping the CNN weights fixed

and training the remaining architecture (i.e., the fully connected layers and the RNNs), or ii) training the whole architecture in an end-to-end manner (by jointly training the CNN and RNN parts). The latter approach provided the best results.

We also used a DenseNet-RNN structure that is quite similar to that of the AffWildNet described before. The only difference was that it uses the DenseNet-121 network's convolutional and pooling layers.

Table 4.13 shows the performance of the developed CNN, standard CNN plus RNN, CNN plus Multi-RNN and ensemble architectures, pre-trained on the Aff-Wild2 database, with and without the post-processing steps (for all networks: p-value $\leqslant 10^{-20} \ll 0.05$). VGG-FACE achieved the best performance compared to the ResNet-50 and DesNet-121 networks. This was expected as the VGG-FACE network has been pre-trained with a large dataset for face recognition (many human faces have been, therefore, used in its construction), thus better filters are already established in comparison to the ResNet-50 and DesNet-121 that have been pre-trained on objects. Additionally, after further pre-training on Aff-Wild2, a better tuning of these filters was attained in the VGG-FACE case.

Additionally, AffWildNet and DenseNet-RNN networks achieved a better performance than all CNN networks. The former networks were standard CNN plus RNNs in which the RNN is used in order to model the contextual information in the data, taking into account temporal variations and thus a better performance was expected.

One can also note that both CNN-1RNN-2nd-pool_last-pool_fc and CNN-3RNN-2nd-pool_last-pool_fc exhibit a much improved performance (between 6% and 10% on average) when compared to CNN plus RNN architectures. This validates our essence that low-level CNN features together with high-level ones provide useful information for our task. Additionally, CNN-3RNN-2nd-pool_last-pool_fc outperformed CNN-1RNN-2nd-pool_last-pool_fc showing that it is better to exploit the low- and high-level features' time variations via RNNs, independently, and then concatenate them, rather than concatenate them first and process them through the use of a single RNN.

Table 4.13 validates that using the ensemble methodology is better than using a single network. This is because different networks produce quite different features; fusing them exploits all these repre-

Table 4.13: CCC based evaluation, on the OMG test set, of valence & arousal predictions provided by our developed CNN, CNN plus RNN, CNN plus Multi-RNN and ensemble architectures. All networks are pre-trained on Aff-Wild2 with (without) post-processing. A higher CCC value indicates a better performance.

CCC	With (Without) Post-Processing		Mean
	Valence	Arousal	
VGG-Face	*0.378 (0.361)*	*0.203 (0.193)*	*0.291 (0.277)*
DenseNet-121	0.365 (0.350)	0.191 (0.184)	0.278 (0.267)
ResNet-50	0.359 (0.344)	0.195 (0.189)	0.277 (0.267)
AffWildNet	*0.409 (0.390)*	*0.224 (0.219)*	*0.317 (0.305)*
DenseNet-RNN	0.394 (0.378)	0.211 (0.209)	0.303 (0.294)
CNN-1RNN-2nd-pool_last-pool_fc	*0.449 (0.441)*	*0.303 (0.297)*	*0.376 (0.369)*
CNN-3RNN-2nd-pool_last-pool_fc	**0.472 (0.463)**	**0.329 (0.322)**	**0.401 (0.393)**
Decision-Level Fusion	0.501 (0.482)	0.332 (0.321)	0.417 (0.402)
Model-Level Fusion + FC	*0.518 (0.500)*	*0.348 (0.328)*	*0.433 (0.414)*
Model-Level Fusion + RNN	***0.535 (0.512)***	***0.365 (0.340)***	***0.450 (0.426)***

sentations that include rich information. It can also be observed that Model-level fusion method has a superior performance compared to that of the Decision-level one, since the features from different networks that are concatenated, contain richer information about the raw data than the final decision. In particular, in Model-level fusion, we concatenate these features and pass them through an RNN and the whole ensemble is trained end-to-end and optimised so that the concatenation of features can provide the best overall result. Moreover, in Model-level fusion, a better performance is achieved when a RNN, instead of a fully connected layer, is used for the fusion.

One can also notice that the post-processing steps helped to achieve a better performance, mainly in valence estimation. The median filter size that we used was 81 for valence (similar to the sequence length), whereas only 3 for the arousal. The arousal window size was small, but, when it was increased, the performance decreased. This in essence means that for the frames within an utterance, the emotional state itself did not change, but the intensity did change. Our final observation is that the performance of the networks in arousal estimation was worse than their performance in valence estimation. This was expected because we only used the visual modality for training our networks; for arousal the audio cues appear to include more discriminating capabilities than facial features in terms of correlation coefficient; this conclusion confirms previous findings [156].

Comparison with the State-of-the-Art Here we compare the performance of our best CNN plus Multi-RNN networks to the performances of state-of-the-art methods submitted to the OMG-Emotion Challenge, utilizing the OMG Emotion database.

Table 4.14 shows that our Model-level Fusion + RNN method outperforms all other methods -even those that have been trained using the audio modality as well- on both the valence and arousal estimation. Table 4.14 also shows that the CNN-3RNN-2nd-pool_last-pool_fc outperformed all state-of-the-art networks, regardless whether they additionally used the audio modality, except for: i) the Single Multi-Modal method that outperformed it on average by 0.015 (however this network used the audio modality as well; since the audio and speech contribute more to arousal estimation, this small difference is justified) and ii) Ensembles I and II, which are a fusion of many different networks that used the visual and audio modalities and thus again the difference in performance was expected.

Table 4.14: CCC based evaluation, on the OMG test set, of VA predictions provided by our best performing networks vs the state-of-the-art. V,A stand for valence and arousal. A higher CCC value indicates a better performance. The results are taken from https://www2.informatik.uni-hamburg.de/wtm/omgchallenges/omg_emotion2018_results2018.html

Methods	Modality	CCC	
		Valence	Arousal
VNet [163]	V,A: visual	0.438	0.244
ANet + VNet [163]	V,A: audio + visual	0.442	0.236
openSMILE + LSTMs, VGG-FACE-BLSTM [199]	A: audio, V: visual	0.258	0.277
openSMILE + LSTMs, VGG-FACE-BLSTM + openSMILE + LSTMs [199]	A: audio, V: audio + visual	0.369	0.286
openSMILE + LSTMs [199]	V,A: audio	0.361	0.293
Single Multi-Modal [239]	V,A: audio + visual	0.484	0.345
Ensemble I [239]	V,A: audio + visual	0.496	0.356
Ensemble II [239]	V,A: audio + visual	0.499	0.361
CNN-3RNN-2nd-pool_last-pool_fc	V,A: visual	0.472	0.329
Model-level Fusion + RNN	V,A: visual	**0.535**	**0.365**

Ablation Study

Pre-Training In the following, we compare the performance of the best performing networks of Table 4.13 with post-processing to that of networks trained from scratch, or being pre-trained with the Aff-Wild or the Aff-Wild2 database. Table 4.15 presents the results of this comparison. The Aff-

Wild2 database, due to its big size and emotion diversity, boosted the performance of all networks pre-trained with it, in comparison to the performance of the networks trained directly with the OMG-Emotion set. This was also the case when we pre-trained the networks with the Aff-Wild database. Overall, networks pre-trained with the Aff-Wild2 achieved a better performance in comparison to networks pre-trained with the Aff-Wild database. Between CNN-1RNN and CNN-3RNN types of architectures, a better performance was acquired when using the latter one.

Table 4.15: CCC based evaluation, on the OMG test set, of valence & arousal predictions provided by various networks when: they are trained from scratch or are pre-trained with the Aff-Wild and Aff-Wild2 databases. A higher CCC value indicates a better performance.

Methods	Trained from Scratch		Pre-trained on Aff-Wild		Pre-trained on Aff-Wild2	
	Valence	Arousal	Valence	Arousal	Valence	Arousal
CNN-1RNN-2nd-pool_last-pool_fc	0.371	0.210	0.419	0.278	**0.449**	**0.303**
CNN-3RNN-2nd-pool_last-pool_fc	0.385	0.192	0.448	0.302	**0.472**	**0.329**
Model-level Fusion + RNN	0.431	0.265	0.511	0.342	**0.535**	**0.365**

Extracted Components Next, we present an ablation study on extracting different CNN low-, mid- and high-level features in CNN-3RNN networks. Table 4.16 compares their performance (in all cases: p-value $\leqslant 10^{-25} \ll 0.05$). The first four rows of Table 4.16 show the performance of networks where a combination of low-, mid- and high-level features are extracted, whereas the next rows show the performance of networks where only low-, or only mid-, or only high-level features are extracted. Let us note that worst performances among all these types of networks were obtained when features were extracted from mid- CNN levels (convolutional layers 6-9). Generally, best performances were obtained when features were extracted from high- and from low-levels. The optimal combination (that provided the best performance) was through the use of CNN-3RNN-2nd-pool_last-pool_fc. One more observation is that low-level features (convolutional layers 3-5), especially when combined with high-level, significantly affected the performance in predicting both valence and arousal.

Utilisation of Landmarks Next, we present an ablation study on the use of landmarks as additional input to various networks. Table 4.17 compares the performance of the CNN-1RNN-2nd-pool_last-pool_fc, CNN-3RNN-2nd-pool_last-pool_fc and Model-level Fusion + RNN networks when land-

Table 4.16: Effect on CCC (on the OMG test set) of using features from different layers in the CNN-3RNN case. All networks are post-processed & pre-trained on Aff-Wild2. A higher CCC value indicates a better performance.

CNN-3RNN	CCC		Mean
	Valence	Arousal	
8th conv + last pool + fc	0.416	0.261	0.339
5th conv + last pool + fc	0.455	0.322	0.389
2nd pool + last pool + fc	**0.472**	**0.329**	**0.401**
3rd conv + 7th conv + fc	0.402	0.267	0.335
last conv + last pool + fc	0.440	0.248	0.344
6th conv + 7th conv + 8th conv	0.328	0.162	0.245
7th conv + 8th conv + 9th conv	0.334	0.172	0.253
3rd conv + 4th conv + 5th conv	0.345	0.185	0.265

marks were/were not used as additional input. In all cases, using landmarks increased their performance by 1.2% - 1.9%.

Table 4.17: Effect on CCC (on the OMG test set) of (not) using landmarks as additional input to various networks. All networks are post-processed & pre-trained on Aff-Wild2. A higher CCC value indicates a better performance. V,A stand for Valence and Arousal

CCC	Without Landmarks			With Landmarks		
	V	A	Mean	V	A	Mean
CNN-1RNN-2nd-pool_last-pool_fc	0.429	0.291	0.360	**0.449**	**0.303**	**0.376**
CNN-3RNN-2nd-pool_last-pool_fc	0.454	0.310	0.382	**0.472**	**0.329**	**0.401**
Model-level Fusion + RNN	0.524	0.352	0.438	**0.535**	**0.365**	**0.450**

Performance Analysis on 2D VA-Space

Finally, to give more insight on the performance of the best CNN-3RNN (CNN-3RNN-2nd-pool_last-pool_fc) network, we analysed its performance at different parts of the 2D Valence-Arousal Space. Table 4.18 presents the obtained valence and arousal performance in terms of Mean Squared Error (MSE) across 4 different regions of this Space. It can be seen that better results have been obtained in the region with high arousal and positive valence; however the obtained MSE is not far away from the MSE across the whole 2D Valence-Arousal Space.

In summary, the main findings of the proposed approach have been: i) low-level features when combined with high-level ones in the CNN plus multi-RNN architectures, helped in boosting the networks' performance in arousal estimation; ii) CNN plus multi-RNN architectures outperformed stan-

Table 4.18: Valence and Arousal MSE in areas of the 2D VA Space for the best CNN-3RNN. A lower MSE indicates a better performance. V,A stand for Valence and Arousal

2D VA-Space	V ∈ [0,1] A ∈ [0,0.5)	V ∈ [0,1] A ∈ [0.5,1]	V ∈ [-1,0) A ∈ [0,0.5)	V ∈ [-1,0) A ∈ [0.5,1]	V ∈ [-1,1] A ∈ [0,1]
CNN-3RNN- 2nd-pool_ last-pool_fc	MSE-V = 0.101 MSE-A = 0.031	MSE-V = 0.055 MSE-A = 0.021	MSE-V = 0.154 MSE-A = 0.061	MSE-V = 0.110 MSE-A = 0.040	MSE-V = 0.110 MSE-A = 0.041

dard CNN plus RNN ones showing that features extracted from previous layers contain useful and rich information for valence-arousal prediction; iii) better results were obtained when the features extracted from previous layers were processed by independent RNNs instead of being concatenated and fed to a single RNN; iv) better results were obtained when using a RNN instead of a fully connected layer for model-level fusion; v) when using the visual modality, network performance for valence estimation was much higher than the corresponding for arousal estimation.

4.5 Expression Recognition Variants with ArcFace Loss

The traditional and most widely used loss function in CNN training for FER is the softmax one that minimizes the cross entropy between the estimated (by the model) class probabilities and the ground truth distribution. This loss simply forces features of different classes to remain apart, but FER in real-world scenarios suffers from not only high inter-class similarity but also high intra-class variation.

In the related face recognition field, it has been shown [133, 212] that categorical cross entropy loss (aka softmax loss) is insufficient to acquire discriminating power for face classification. Several loss functions have been proposed for maximising inter-class and minimising intra-class variance. [32, 85] propose multi-loss learning to increase feature discriminating power. These, require thorough mining of pair/triplet samples, which is a time-consuming procedure. [133] projects the original Euclidean space of features to an angular space, introducing an angular margin for larger inter-class variance. [206] directly adds a cosine margin penalty to the target logit, showing better performance than [133]. [45] further improved the discriminative power of face recognition models, stabilising the training process.

Therefore, for FER, some works have proposed novel losses. Inspired by the center loss [212], which

penalises the distance between deep features and their corresponding class centers, two variations were proposed to assist the supervision of the softmax loss for more discriminative features for FER: (i) island loss [22] was formalized to further increase the pairwise distances between different class centers, and (ii) deep locality-preserving loss [128] was formalised to pull the locally neighboring features of the same class together so that the intra-class local clusters of each class are compact.

Given the success of the ArcFace loss [45] and the boost it brought to face recognition models' performance, we adopted the ArcFace loss and adapted it for emotion recognition. To the best of our knowledge, this is the first time that this loss designed for face recognition, has been used in the context of affect recognition. Another contribution we made wass the design of two networks, the one based on residual units and the other based on VGG-FACE layers and their training with the ArcFace loss. At first, we pre-trained them on Aff-Wild2 and then re-trained them, one at a time, on a plethora of other databases. Our results outperformed all state-of-the-art networks, illustrating: i) the richness of Aff-Wild2 (providing it with the ability to be used as robust prior for network pre-training) and ii) that the ArcFace loss can be used in the affect recognition field, yielding state-of-the-art results. In fact, this was the very first proof of the effectiveness of additive angular margin in affect recognition.

4.5.1 The ArcFace Loss Function

The softmax cross-entropy loss can be modified as follows:

$$\mathcal{L} = \frac{-1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{7} e^{W_j^T x_i}} = \frac{-1}{N}\sum_{i=1}^{N}\log\frac{e^{\|W_{y_i}\|\cdot\|x_i\|\cdot\cos\theta_{y_i}}}{\sum_{j=1}^{7} e^{\|W_j\|\cdot\|x_i\|\cdot\cos\theta_j}} \overset{\substack{\|x_i\|=s,\\ \|W_j\|=1}}{=\!=\!=} \frac{-1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cdot\cos\theta_{y_i}}}{\sum_{j=1}^{7} e^{s\cdot\cos\theta_j}} \tag{4.7}$$

where the embedding feature $x_i \in \mathcal{R}^d$ denotes the deep feature of the i-th sample belonging to the y_i-th class, $W_j \in \mathcal{R}^d$ denotes the j-th column of the weight $W \in \mathcal{R}^{d\times 7}$, N is the batch size, θ_j is the angle between weight W_j and feature x_i, $\|W_j\|$ is fixed to 1 by l_2 normalization, $\|x_i\|$ is fixed by l_2 normalisation and re-scaled to s.

From eq.4.7, it can be seen that the embedding features are distributed around each feature centre on the hypersphere. In our case, we adopt the ArcFace loss, where an angular margin penalty m between x_i and W_{y_i} is added to simultaneously enhance the intra-class compactness and inter-class discrepancy (eq.4.7: $\theta_{y_i} \rightarrow \theta_{y_i} + m$). m is equal to the geodesic distance margin penalty in the normalised hypersphere. We refer the interested reader to [45] for more details and explanation of this loss.

4.5.2 The ArcRes and ArcVGG Deep Neural Architectures

Next, we develop two networks that will be trained with this loss. The first CNN architecture, called ArcFace-Residual (ArcRes) uses residual units and is depicted in Fig.4.13; 'bn' stands for batch normalization, the convolution layer is in the format: filter height $\times$ filter width 'conv.', number of output feature maps; the stride is equal to 2, everywhere; the fc layer is the embedding layer; the output layer provides the seven expresion class logits ($W_j^T x_i$, $j = 1..7$). Table 4.19 shows the exact network configuration of ArcRes.

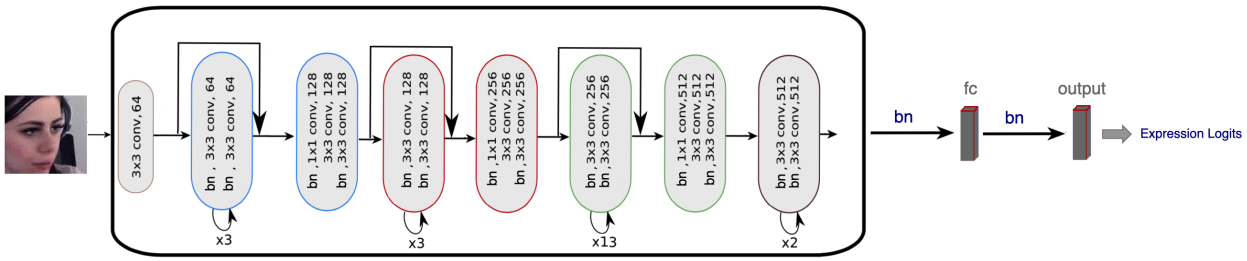


Figure 4.13: The ArcRes network that has been trained with the ArcFace loss

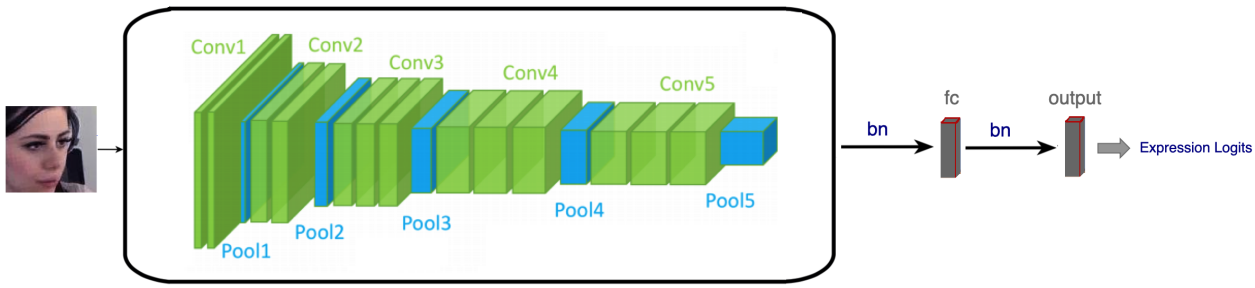


Figure 4.14: The ArcVGG network that has been trained with the ArcFace loss

The second network is called Arcface-VGG (ArcVGG) and is depicted in Fig.4.14; the difference

Table 4.19: ArcRes: the developed network with residual units for seven basic expression classification

Block	Layer	filter , # feature maps	stride	# units
Normal 1	conv	3×3 , 64	2×2	-
Residual 1	batch norm, conv batch norm, conv	$\begin{bmatrix} 3 \times 3 , 64 \\ 3 \times 3 , 64 \end{bmatrix} \times 3$	2×2	-
Normal 2	batch norm, conv conv batch norm, conv	1×1 , 128 3×3 , 128 3×3 , 128	2×2	-
Residual 2	batch norm, conv batch norm, conv	$\begin{bmatrix} 3 \times 3 , 128 \\ 3 \times 3 , 128 \end{bmatrix} \times 3$	2×2	-
Normal 3	batch norm, conv conv batch norm, conv	1×1 , 256 3×3 , 256 3×3 , 256	2×2	-
Residual 3	batch norm, conv batch norm, conv	$\begin{bmatrix} 3 \times 3 , 256 \\ 3 \times 3 , 256 \end{bmatrix} \times 13$	2×2	-
Normal 4	batch norm, conv conv batch norm, conv	1×1 , 512 3×3 , 512 3×3 , 512	2×2	-
Residual 4	batch norm, conv batch norm, conv	$\begin{bmatrix} 3 \times 3 , 512 \\ 3 \times 3 , 512 \end{bmatrix} \times 2$	2×2	-
Normal 5	batch norm, fully connected	-	-	32/512
Normal 6	batch norm, fully connected	-	-	7

with ArcRes is that the rectangular area in the Figure 4.13 contains VGGFace's convolutional and pooling layers. In more detail, Table 4.20 shows the exact network configuration of ArcVGG.

In both ArcRes and ArcVGG, during testing we keep the feature embedding layer denoted as 'fc' in Figures 4.13 and 4.14, discarding the output layer. For all training images, we extract features from the embedding layer and split them in seven clusters. Then, for each test image, we compute its distance (based on cosine similarity) from all cluster centers and assign it to the center for which this distance is minimum.

4.5.3 Pre-Processing & Network Training Details

The SSH detector [152] based on the ResNet and trained on the WiderFace dataset [220] was used to extract face bounding boxes from all images. Also, 5 facial landmarks (two eyes, nose and two

Table 4.20: ArcVGG: the developed network with residual units for seven basic expression classification

Layer	filter	# feature maps	stride	# units
conv 1	3×3	64	1×1	-
conv 2	3×3	64	1×1	-
max pooling	2×2	-	2×2	-
conv 3	3×3	128	1×1	-
conv 4	3×3	128	1×1	-
max pooling	2×2	-	2×2	-
conv 5	3×3	256	1×1	-
conv 6	3×3	256	1×1	-
conv 7	3×3	256	1×1	-
max pooling	2×2	-	2×2	-
conv 8	3×3	512	1×1	-
conv 9	3×3	512	1×1	-
conv 10	3×3	512	1×1	-
max pooling	2×2	-	2×2	-
conv 11	3×3	512	1×1	-
conv 12	3×3	512	1×1	-
conv 13	3×3	512	1×1	-
max pooling	2×2	-	2×2	-
batch normalisation	-	-	-	-
fully connected	-	-	-	32/512
batch normalisation	-	-	-	-
fully connected	-	-	-	7

mouth corners) were extracted and used to perform similarity transformation (for face alignment). After that we obtained the cropped faces which were then resized to dimension $96 \times 96 \times 3$. The pixel intensities were normalized to take values in $[-1, 1]$.

As far as specific details about hyperparameters of the developed architectures are concerned, they can be found in Table 4.21. For both networks, the optimal learning rate was 10^{-4}, the batch size was 300 and dropout with 0.4 has been used. The best angular margin penalty m was either 0.1 or 1, s was 32 or 64 and d was either 32 or 512 meaning that the embedding layer had either 32 or 512 features. All experiments were implemented in TensorFlow, on a Tesla V100 32GB GPU, using SGD with momentum (0.9).

Let us also note that ArcRes and ArcVGG networks were first trained on Aff-Wild2 using the ArcFace loss as we defined before and then, they were re-trained end-to-end on each of the examined databases, which were AffectNet, RAF-DB, IMFDB and FER2013 (again using the ArcFace loss).

Table 4.21: Network Configurations

	ArcRes / ArcVGG
learning rate	$[10^{-4}, 10^{-5}], best: 10^{-4}$
batch size	300
parameters	dropout=0.4, $d \in \{32, 512\}, s \in \{32, 64\}$, $m \in \{0.1, 0.5, 1, 1.5, 2, 2.5, 3\}, best: 0.1/1$

4.5.4 Performance Evaluation

The best performing network in AffectNet database is an AlexNet [120] [151]. AlexNet consists of five convolution layers, followed by max-pooling and normalization layers, and three fully-connected one. The authors in [151] took into account the data imbalance existing in the training set and thus weighted the loss for each of the classes by their relative proportion in the training dataset. This loss heavily penalised AlexNet for misclassifying examples from under-represented classes, while penalised AlexNet less for misclassifying examples from well-represented classes.

The best performing network in RAF-DB database is the DLP-CNN [128] network consisting of six convolutional layers followed by max-pooling ones and then by a fully connected layer before the output that performs the classification into the seven basic expressions. The authors of [128] developed a loss function that preserved the locality of each sample and made the local neighborhoods within each class as compact as possible. The DLP-CNN was trained for discriminative feature learning using the joint supervision of the softmax loss that characterises the global scatter and the locality preserving loss that characterises the local scatters within class.

The best performing method in FER2013 was developed by [70], where automatic features learned by VGGFACE, VGG-f [27] and VGG-13 [14] are combined with handcrafted features computed by the bag-of-visual-words model. For training the CNN models, the authors used the Dense-Sparse-Dense [80] training procedure. After fusing the two types of features, a local learning framework was employed that included k-nearest neighbors and an one-versus-all Support Vector Machines (SVM) classifier, which provided the final prediction.

Table 4.22 presents a performance comparison between the ArcRes and ArcVGG networks trained with the ArcFace loss on Aff-Wild2 and re-trained on each of the examined databases and the state-of-

the-art in these databases (whose results are taken from the respective papers). At first, it can be seen that both networks outperformed the state-of-the-art by a large margin, on all examined databases. In AffectNet and FER2013 ArcRes outperformed ArcVGG by 1%, whereas in RAF-DB and IMFDB ArcVGG outperformed ArcRes by 1%.

Table 4.22: Performance evaluation of ArcRes and ArcVGG

Databases	ArcRes	ArcVGG	AlexNet [151]	DLP-CNN [128]	VGG [71]
AffectNet	**0.63**	0.62	0.58	-	-
RAF-DB	0.75	**0.76**	-	0.74	-
IMFDB	0.55	**0.56**	-	-	-
FER2013	**0.8**	0.79	-	-	0.75

Chapter 5

Multi-Task Learning for Affect Analysis

5.1 A Multi-Task Approach to Affect Recognition

In the former Chapters of this Thesis, we have described the development of Aff-Wild and Aff-Wild2 databases, which have extended early efforts towards collecting large scale datasets of naturalistic behaviour captured in uncontrolled conditions, *in-the-wild* [16, 151, 226]. We have described the three main affect recognition tasks: recognition of the seven basic expressions, i.e., anger, disgust, fear, happiness, sadness, surprise and neutral [62], in-the-wild [36, 40]; estimation of continuous affect dimensions, i.e. valence and arousal; detection of facial action units (AU) [63], which code facial motion with respect to activation of facial muscles, using automatic AU annotation toolboxes [11, 16].

Up to the present, these three tasks have been generally tackled individually from each other, despite the fact that they are interconnected. In [63], the facial action coding system (FACS) has been built, indicating, for each of the basic expressions, the respective prototypical action units. In [59], a dedicated user study has been conducted to study the relationship between AU activations and emotion expressions beyond basic types, dealing with compound emotions (e.g., happily surprised). In [99], the authors showed that neural networks trained for expression recognition implicitly learn to detect facial action units as well. Moreover, in [145] the authors have discovered that valence and arousal dimensions could be interpreted through AUs; for example, AU12 (lip corner puller) is related to positive valence.

Multi-task learning (MTL) is an approach that can be used to jointly learn all three behaviour analysis tasks. MTL was first studied in [25], where the authors proposed to jointly learn parallel tasks sharing a common representation and transferring part of the knowledge - learned to solve one task - to improve learning of the other related task. Since then, several approaches have adopted MTL for solving different problems in computer vision and machine learning. In the face analysis domain, the use of MTL is somewhat limited. In [211], MTL was tackled through a neural network that jointly handled face recognition and facial attribute prediction tasks. MTL helped to capture global feature and local attribute information simultaneously. In the following we develop a new MTL approach to affect recognition in-the-wild, by using Aff-Wild2 which includes annotations for all three tasks and appropriately extending AffWildNet that was described in the previous Chapter.

5.1.1 Related Work

In 2020 we organised the Affective Behavior Analysis in-the-wild (ABAW) Competition [106], in conjunction with the IEEE Conference on Face and Gesture Recognition. The ABAW 2020 Competition was the first Competition aiming at automatic analysis of the three main behaviour tasks of valence-arousal estimation, basic expression recognition and action unit detection. It was split into three Challenges, based on the Aff-Wild2 database, with each one addressing a respective behaviour analysis task. In the following we make reference to five approaches that displayed the best performance in each Challenge and ranked in the top-3 positions; we also present each Challenge's baseline system.

The NISL2020 team [42] participated in all three Challenges, ranking first, third and first in Valence-Arousal Estimation, Seven Basic Expression Classification and Eight Action Unit Detection Challenges, respectively. They used multi-task learning of the three tasks, through an algorithm that learnt from partial labels. At first, they trained a teacher model to perform all three tasks, where each instance was trained by the ground truth label of its corresponding task. Then, they used the outputs of the teacher model and the ground truth to train the student model, so that the latter outperformed the teacher model. They also used an ensemble methodology so as to further boost the performance of the model.

The TNT team [121] participated in all three Challenges and ranked second, first and second in Valence-Arousal Estimation, Seven Basic Expression Classification and Eight Action Unit Detection Challenges, respectively. They also used multi-task learning of the three tasks using both provided video and audio inputs. They developed a two-stream aural-visual analysis model in which audio and image streams were first processed separately and fed into a convolutional neural network. They did not use recurrent architectures for temporal analysis, but instead used temporal convolutions. Furthermore, the model was given access to additional features extracted during face-alignment in the pre-processing stage. At training time, correlations between different emotion representations were exploited so as to improve the model's performance.

The ICT-VIPL-VA team [234] participated in the Valence-Arousal Estimation Challenge and ranked third. Their methodology fused both visual features extracted from videos and acoustic features extracted from audio tracks. To extract visual features, they followed a CNN-RNN paradigm, in which spatio-temporal visual features were extracted by a 3D convolutional network and / or a pretrained 2D convolutional network, and were fused through a bidirectional recurrent neural network. The audio features were extracted by a GRU-MLP network.

The ICT-VIPL-Expression team [131] participated in the Seven Basic Expression Classification Challenge and ranked second. Their methodology combined a Deep Residual Network with convolutional block attention module and Bidirectional Long-Short-Term Memory Units. They provided visualisation of the learned attention maps and analysed the importance of different regions in facial expression recognition.

The SALT team [158] participated in the Eight Action Unit Detection Challenge and ranked third. Their methodology included a multi-label class balancing algorithm as a pre-processing step for overcoming the imbalanced occurrences of Action Units in the training dataset. Then a ResNet was trained using the augmented training dataset.

The architecture of the baseline system that we generated for estimating valence and arousal was based on that of PatchGAN [31, 91, 243]. PatchGAN is a deep convolutional neural network initially designed to classify patches of an input image, rather than the entire image, as real or fake. The PatchGAN was the discriminator of the pix2pix architecture [91]. The output of the network is a

single feature map of real/fake predictions that is averaged to give a single score. In StarGAN [31], PatchGAN was additionally used as a classifier. Here, we adopted PatchGAN for valence-arousal regression. The exact architecture used, can be seen in Table 5.1.

Table 5.1: PatchGAN adopted for valence-arousal estimation. Leaky Rely follows each convolutional layer.

Name	Type	Filter	# Feature Maps
conv 1	weights	4×4	64
conv 2	weights	4×4	128
conv 3	weights	4×4	256
conv 4	weights	4×4	512
conv 5	weights	4×4	1024
conv 6	weights	4×4	1024
conv 7	weights	4×4	2048
conv 8	weights/D-label	1×1	2

The baseline systems for the tasks of classification into the seven basic expressions and detection of eight action units, were based on the architecture of MobileNetV2 [177]. MobileNetV2 belongs to the class of efficient models called MobileNets [86] that are light-weight deep neural networks. They are based on a streamlined architecture that uses depth-wise separable convolutions which dramatically reduce the complexity, cost and model size of the network. For more detail regarding this class of architectures and the MobileNetV2 network, we refer the interested reader to [177]. Table 5.2 shows the basic structure of MobileNetV2.

Let us note that: i) batch normalization was applied after each convolutional or expanded convolutional layer, ii) the non-linear activation used was the Relu6 and iii) no average pooling was conducted in the end. After the final convolutional layer (shown in Table 5.2), a fully connected layer followed (with 7 units if the task was to predict the 7 basic expressions, or 8 units if the task was to detect the 8 action units) and on top of that was a softmax or sigmoid layer, respectively.

5.1.2 MT Extensions of AffWildNet

Here, we present multi-task CNN, CNN-RNN and audiovisual CNN-RNN networks, including extension of the AffWildNet architecture, taking into account the fact that the three tasks of facial behaviour

Table 5.2: The MobileNetV2 network

Name	Type	Filter
conv	weights	(3, 3, 3, 32)
expanded conv	depthwise	(3, 3, 32, 1)
expanded conv	project	(1, 1, 32, 16)
expanded conv 1	expand	(1, 1, 16, 96)
expanded conv 1	depthwise	(3, 3, 96, 1)
expanded conv 1	project	(1, 1, 96, 24)
expanded conv 2	expand	(1, 1, 24, 144)
expanded conv 2	depthwise	(3, 3, 144, 1)
expanded conv 2	project	(1, 1, 144, 24)
expanded conv 3	expand	(1, 1, 24, 144)
expanded conv 3	depthwise	(3, 3, 144, 1)
expanded conv 3	project	(1, 1, 144, 32)
expanded conv 4	expand	(1, 1, 32, 192)
expanded conv 4	depthwise	(3, 3, 192, 1)
expanded conv 4	project	(1, 1, 192, 32)
expanded conv 5	expand	(1, 1, 32, 192)
expanded conv 5	depthwise	(3, 3, 192, 1)
expanded conv 5	project	(1, 1, 192, 32)
expanded conv 6	expand	(1, 1, 32, 192)
expanded conv 6	depthwise	(3, 3, 192, 1)
expanded conv 6	project	(1, 1, 192, 64)
expanded conv 7	expand	(1, 1, 64, 384)
expanded conv 7	depthwise	(3, 3, 384, 1)
expanded conv 7	project	(1, 1, 384, 64)
expanded conv 8	expand	(1, 1, 64, 384)
expanded conv 8	depthwise	(3, 3, 384, 1)
expanded conv 8	project	(1, 1, 384, 64)
expanded conv 9	expand	(1, 1, 64, 384)
expanded conv 9	depthwise	(3, 3, 384, 1)
expanded conv 9	project	(1, 1, 384, 64)
expanded conv 10	expand	(1, 1, 64, 384)
expanded conv 10	depthwise	(3, 3, 384, 1)
expanded conv 10	project	(1, 1, 384, 96)
expanded conv 11	expand	(1, 1, 96, 576)
expanded conv 11	depthwise	(3, 3, 576, 1)
expanded conv 11	project	(1, 1, 576, 96)
expanded conv 12	expand	(1, 1, 96, 576)
expanded conv 12	depthwise	(3, 3, 576, 1)
expanded conv 12	project	(1, 1, 576, 96)
expanded conv 13	expand	(1, 1, 96, 576)
expanded conv 13	depthwise	(3, 3, 576, 1)
expanded conv 13	project	(1, 1, 576, 160)
expanded conv 14	expand	(1, 1, 160, 960)
expanded conv 14	depthwise	(3, 3, 960, 1)
expanded conv 14	project	(1, 1, 960, 160)
expanded conv 15	expand	(1, 1, 160, 960)
expanded conv 15	depthwise	(3, 3, 960, 1)
expanded conv 15	project	(1, 1, 960, 160)
expanded conv 16	expand	(1, 1, 160, 960)
expanded conv 16	depthwise	(3, 3, 960, 1)
expanded conv 16	project	(1, 1, 960, 320)
conv 1	weights	(1, 1, 320, 1280)

analysis (valence-arousal estimation, action unit detection and basic expression classification) are interconnected.

MT-VGGFACE

At first, we developed a multi-task CNN network based on the VGGFACE structure (MT-VGGFACE). We kept the convolutional and pooling layers of VGGFACE, discarded all its fully connected layers and added on top of it 2 fully connected layers with 4096 units. A (linear) output layer followed that provided final estimates for valence and arousal; it also produced 7 basic expression logits that were passed through a softmax function to get the final 7 basic expression predictions; lastly, it produced 8 AU logits that were passed through a sigmoid function to get the final 8 AU predictions. Table 5.3 shows the structure of MT-VGGFACE (except of the output layer).

Table 5.3: MT-VGGFACE: the multi-task developed CNN model

Layer	Filter	# Feature Maps	stride	no of units
conv 1	3×3	64	1×1	-
conv 2	3×3	64	1×1	-
max pooling	2×2	-	2×2	-
conv 3	3×3	128	1×1	-
conv 4	3×3	128	1×1	-
max pooling	2×2	-	2×2	-
conv 5	3×3	256	1×1	-
conv 6	3×3	256	1×1	-
conv 7	3×3	256	1×1	-
max pooling	2×2	-	2×2	-
conv 8	3×3	512	1×1	-
conv 9	3×3	512	1×1	-
conv 10	3×3	512	1×1	-
max pooling	2×2	-	2×2	-
conv 11	3×3	512	1×1	-
conv 12	3×3	512	1×1	-
conv 13	3×3	512	1×1	-
max pooling	2×2	-	2×2	-
fully connected 1 dropout	-	-	-	4096
fully connected 2	-	-	-	4096

MT-AffWildNet

Next we exploited the fact that the developed AffWildNet has shown best performance in capturing the dynamics and the in-the-wild nature of the Aff-Wild database. Moreover, since it has a CNN-RNN structure, it effectively models contextual information in the data, taking into account temporal affect variations. As a consequence, we have developed a network, extending AffWildNet - initially developed for valence-arousal estimation - to additionally account for action unit detection and for basic expression classification. The developed MT-AffWildNet, a multi-task CNN-RNN network in which the CNN-RNN part was the AffWildNet (the CNN part included 13 convolutional and pooling layers of VGG-FACE, followed by a fully connected layer of 4096 hidden units; the RNN was a 2-layer GRU with 128 cells each and was stacked on top of the CNN); the output layer followed on top of it, which was exactly the same as in MT-VGGFACE. It is therefore evident that the predictions for all tasks are pooled from the same feature space, since there exist correlations between the three different tasks.

A/V-MT-AffWildNet

Since Aff-Wild2 is an audiovisual (A/V) database, we additionally developed a network for handling both the video and audio modalities. We based this developed network on MT-AffWildNet, which was above-described, generating A/V-MT-AffWildNet. A/V-MT-AffWildNet took as input frames extracted from the video and spectrograms extracted from the audio sequences. We used a feature level fusion strategy, illustrated in Figure 5.1. A/V-MT-AffWildNet consisted of two identical streams that extracted features directly from raw input images and spectrograms, respectively. Each stream consisted of a MT-AffWildNet, without an output layer. The features from the two streams were concatenated, forming a 256-dimensional feature vector that was passed through a 2-layer GRU layer with 128 units in each layer, so as to fuse the information of the audio and visual streams. The output layer followed on top of it, being exactly the same as in MT-AffWildNet. No feature normalisation was performed on the two concatenated streams as they were from the same scale/numerical range and the training was end-to-end. A/V-MT-AffWildNet is a multi-modal and multi-task network. Let us note here that it was the first time - in our Challenge - that audio was taken into account for action

unit detection.

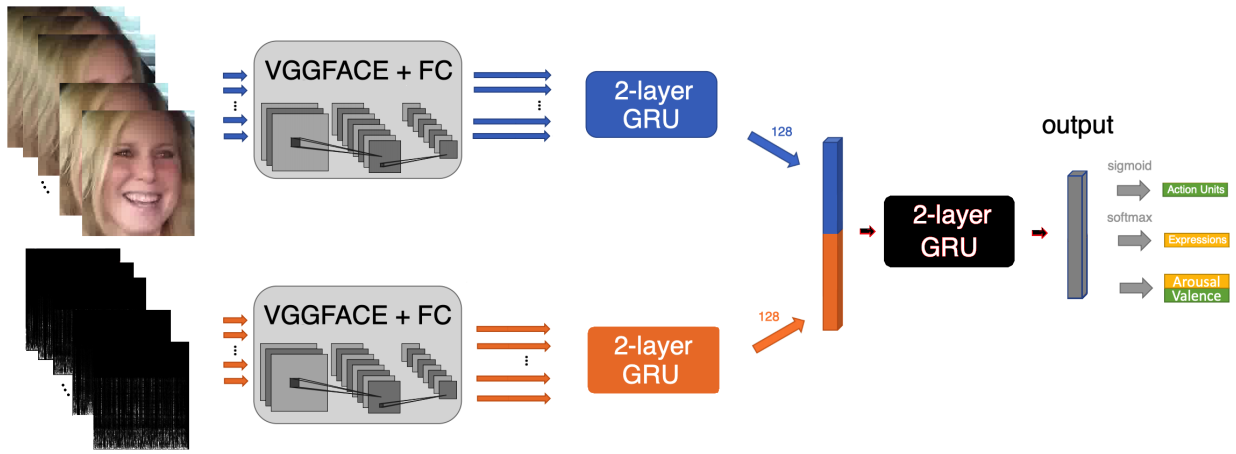


Figure 5.1: A/V-MT-AffWildNet: the Multi-Modal and Multi-Task developed model

5.1.3 Pre-Processing, Performance Measures & Network Training Details

At first, we describe the two pre-processing steps, applied to the visual and audio modalities respectively, that have been used to generate the input data for affect analysis. Next, we present the loss function used for training the multi-task networks, as well as the evaluation metrics used in each affect recognition task. These metrics have been used across 11 databases, including Aff-Wild2, Aff-Wild, AFEW-VA, AffectNet, RAF-DB, FER2013, EmotioNet, DISFA, BP4DS and BP4D+. Finally, the network implementation details are described.

Pre-Processing

Visual Modality The SSH detector [152], based on ResNet and trained on the WiderFace dataset [220] was used to extract face bounding boxes from all images. Also, 5 facial landmarks (two eyes, nose and two mouth corners) were extracted and used to perform similarity transformation (for face alignment). As a result, we obtained cropped faces which were then resized to dimension $96 \times 96 \times 3$. The pixel intensities were normalized to take values in $[-1, 1]$.

Audio Modality The audio signal (mono) was sampled at $44,100$Hz. Then spectrograms were extracted; spectrogram frames were computed over a 33ms window with 11ms overlap. The resulting

intensity values were normalized in $[-1, 1]$ to be consistent with the visual modality.

Loss Function

The objective function minimized during training of the multi-task networks was the sum of the individual task losses:

$$\mathcal{L}_{CCE} = \mathbb{E}[-\log \frac{e^{p_p}}{\sum_{i=1}^{7} e^{p_i}}] \tag{5.1}$$

$$\mathcal{L}_{BCE} = \mathbb{E}[-\sum_{i=1}^{8} (t_i \cdot \log p_i + (1 - t_i) \cdot \log (1 - p_i))] \tag{5.2}$$

$$\mathcal{L}_{CCC} = 1 - \frac{(\rho_a + \rho_v)}{2}, \text{ with } \rho_{a,v} = \frac{2 s_{xy}}{[s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2]} \tag{5.3}$$

where $\mathcal{L}_{CCE}$ is the categorical cross entropy loss, $\mathcal{L}_{BCE}$ is the binary cross entropy loss, p_p is the prediction of positive class, p_i is the prediction of AU_i or $Expr_i$, $t_i \in \{0, 1\}$ is the label of AU_i, $\rho_{a,v}$ is the Concordance Correlation Coefficient (CCC) of arousal/valence, s_x and s_y are the variances of arousal/valence labels and predicted values respectively and s_{xy} is the corresponding covariance value.

Evaluation Metrics

Valence-Arousal Estimation The mean value of CCC for valence and arousal estimation is adopted as the main evaluation criterion of the performance of systems providing valence and arousal estimation.

$$\mathcal{E}_{total} = \frac{\rho_a + \rho_v}{2}, \tag{5.4}$$

Basic Expression Classification The F_1 score is a weighted average of the recall (i.e., the ability of the classifier to find all positive samples) and precision (i.e., the ability of the classifier not to label as

positive a sample that is negative). The F_1 score reaches its best value at 1 and its worst score at 0. The F_1 score is defined as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{5.5}$$

The F_1 score for affect recognition is computed based on a per-frame prediction (an emotion category is specified in each frame).

Total accuracy (denoted as $\mathcal{T}Acc$) is defined on all test samples and is the fraction of predictions that the model got right. Total accuracy reaches its best value at 1 and its worst score at 0. It is defined as:

$$\mathcal{T}Acc = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{5.6}$$

When comparing our developed architectures' performance to the ABAW Challenge participating teams' systems performance, the weighted average between the F_1 score and the total accuracy, $\mathcal{T}Acc$, is the main evaluation criterion:

$$\mathcal{E}_{total} = 0.67 \times F_1 + 0.33 * \mathcal{T}Acc, \tag{5.7}$$

When comparing our developed architectures' performance, to the state-of-the-art methods' performance in different databases, the evaluation criterion is the usual F1 score, because of its robustness to the imbalance in positive and negative samples, which is very common in the case of AUs. Exceptions are the RAF-DB and FER2013 databases, in which the mean diagonal value of the confusion matrix and the accuracy metric, respectively, are the default performance measures.

Action Unit Detection We first obtain the F_1 score for each AU independently, and then compute the (unweighted) average over all 8 AUs (denoted as $\mathcal{A}F_1$) :

$$\mathcal{A}F_1 = \frac{1}{8}\sum_{i=1}^{8} F_1^i \tag{5.8}$$

The F_1 score for AUs is computed based on a per-frame detection (whether each AU is present or absent).

When comparing our developed architectures' performance to the ABAW Challenge participating teams' systems performance, the average between the $\mathcal{A}F_1$ score and the total accuracy, $\mathcal{T}Acc$, is the main evaluation criterion:

$$\mathcal{E}_{total} = 0.5 \times \mathcal{A}F_1 + 0.5 * \mathcal{T}Acc \tag{5.9}$$

When comparing our developed architectures' performance, to the state-of-the-art methods' performance in different databases, the evaluation criterion is the usual F1 score.

Network Training Details

Specific details about hyperparameters of the developed architectures can be found in Table 5.4. All experiments were implemented in TensorFlow, on a Tesla V100 32GB GPU, using Adam optimizer (with default values). Details follow:

I) *MT-VGGFACE*: The network has first been pre-trained for VA estimation on the Aff-Wild database, then the output layer was discarded and substituted by a multi-task one. Then it was trained end-to-end on Aff-Wild2. Best results have been provided with a learning rate of 10^{-4}, with batch size 256. Dropout with value 0.4 has been used.

II) *MT-AffWildNet*: The CNN-RNN part was initialized with the weights of the AffWildNet. Then the whole architecture was trained end-to-end on Aff-Wild2. Learning rate used was in the range $[10^{-4}, 10^{-6}]$. Best results have been provided for 10^{-5}, with batch size 10 and sequence length 90. Dropout with value 0.4 has been used after all fully connected layers.

III) *A/V-MT-AffWildNet*: Training was divided in two phases: first the audio/visual streams were trained independently and then the audiovisual network was trained end-to-end. To train each stream individually, we followed the same procedure as in the MT-AffWildNet case: for each stream, the CNN-RNN part was initialized with the weights of the AffWildNet, then we appended on top an

output layer; then the whole stream was trained end-to-end on Aff-Wild2. Once each single stream has been trained, we discarded their output layers and they were used for initializing the corresponding streams in the multi-stream architecture. Finally, the entire audiovisual network was trained end-to-end. The learning rate was in the range $[10^{-3}, 10^{-6}]$. Best results have been provided for 10^{-5}, with batch size 5 and sequence length 90 (for each modality). Dropout with value 0.4 has been used after all fully connected layers.

Table 5.4: Network Configurations: MT = Multi-Task, A/V = audiovisual

	MT-VGGFACE	MT-AffWildNet	A/V-MT-AffWildNet
learning rate	$[10^{-4}, 10^{-5}], best: 10^{-4}$	$[10^{-4}, 10^{-6}], best: 10^{-5}$	$[10^{-3}, 10^{-6}], best: 10^{-5}$
batch size	256	10	5
sequence length	-	90	90
parameters	dropout=0.4	dropout=0.4	dropout=0.4

Finally, let us mention that for the Basic Expression Classification task, we did not tackle the data imbalanced problem (the majority of the frames are labelled as neutral) since the evaluation metric for that task contained the F_1 score which in general is known to take into account the data imbalance and since the results -as shown in the Experimental Study that follows- have been great outperforming all state of the art.

5.1.4 Experimental Study

The experimental study consists of three parts. In the first, we compare our developed networks' performance to each single- and multi-task CNN, showing that our MT networks outperform: their single-task counterparts; other single-task networks; other multi-task networks. In the second, we compare our developed networks' performance to the state-of-the-art methods developed on Aff-Wild2, showing that our networks outperform the methods of the teams that participated in the ABAW Competition on Aff-Wild2. In the final part, we perform a cross database experimental study, by testing our developed networks on 10 different databases and comparing the achieved performance to that of state-of-the-art methods on these databases, showing that our networks provide the best pre-trained framework for a large variety of affect recognition settings.

Results: Developments vs Single- & Multi-Task CNNs

At first, we employed 3 state-of-the-art networks, SphereFace-20 [133], VGGFACE [161], and Inception ResNet [190]. We train these networks to perform a single behavior task (VA estimation, or AU detection, or Expr classification), or jointly perform all 3 tasks. The predictions for all tasks were pooled from the same feature space.

Table 5.5 compares the performance of the developed MT-VGGFACE and MT-AffWildNet (in two settings: i) when trained only with video frames and thus with the visual modality; ii) when trained only with spectrograms and thus with the audio modality); it also compares A/V-MT-AffWildNet to single- and multi-task SphereFace-20, VGGFace and Inception ResNet, denoted as ST-SphereFace/MT-SphereFace, ST-VGGFACE and ST-InceptionResNet/MT-InceptionResNet. The utilised database was Aff-Wild2. Table 5.5 shows the corresponding performance for each task: valence-arousal estimation (denoted as VA), seven basic expression classification (denoted as Expr) and eight action unit detection (denoted as AU). We note that the corresponding performance metrics are the CCC for valence and arousal and the F1 score for the action units and the expression categories.

It is evident from Table 5.5 that MT-VGGFACE outperformed in all tasks: i) its single-task counterpart network, ST-VGGFACE and ii) all other developed either single- or multi-task CNN networks. It is also evident that the MT-AffWildNet either when trained with the audio modality or the visual one, outperformed in all tasks the MT-VGGFACE network. When trained with the visual modality, since it is a CNN-RNN network and the RNN is used to model temporal variations, a better performance over MT-VGGFACE was expected in all three tasks.

When trained with the audio modality, in the VA estimation task, one can observe a big performance difference (9/% increase to the MT-VGGFACE and 6% increase to the MT-AffWildNet that used the visual modality) when estimating arousal; this is expected because for arousal the audio cues appear to include more discriminating capabilities than facial features in terms of correlation coefficient [156]. Additionally, for valence, facial features appear to include more discriminating capabilities than audio cues; as expected, in valence estimation, the MT-AffWildNet, when trained with the visual modality, outperformed the MT-AffWildNet, when trained with the audio modality.

It is interesting to note that, in expression recognition, the performance was the same regardless of which of the two modalities was used. In action unit detection, the visual modality provided MT-AffWildNet with better performance than the audio one; a result which is intuitive. Finally, the A/V-MT-AffWildNet displayed the best performance on all tasks.

These results illustrate the ability of the developed networks to capture the underline affect state (either valence-arousal, or basic expression or action units) when using only the audio modality. Let us mention that it is the first time that it is shown and proved that action units detection can benefit by the audio modality.

Table 5.5: Evaluation on Aff-Wild2 for the developed networks and other single- and multi-task CNNs. 'ST' means single-task, 'MT' means multi-task; VA evaluation is shown as CCC_V and CCC_A; AU and Expr evaluation corresponds to F1 score

Networks	Aff-Wild2			
	CCC-V	CCC-A	F1 Expr	F1 AU
ST-SphereFace	0.33	0.33	0.37	0.28
MT-SphereFace	0.35	0.34	0.39	0.29
ST-InceptionResNet	0.35	0.35	0.36	0.29
MT-InceptionResNet	0.37	0.36	0.38	0.30
ST-VGGFACE	0.40	0.39	0.38	0.30
MT-VGGFACE	0.43	0.42	0.41	0.32
MT-AffWildNet (audio modality)	0.44	0.51	0.44	0.34
MT-AffWildNet (visual modality)	0.46	0.45	0.44	0.36
A/V-MT-AffWildNet	**0.47**	**0.52**	**0.46**	**0.37**

Results: Developments vs State-of-the-art ABAW Competition Teams

Table 5.6 compares the performance, in all tasks, of the developed MT-VGGFACE and MT-AffWildNet (in two settings: i) when trained only with video frames and thus with the visual modality; ii) when trained only with spectrograms and thus with the audio modality); it also compares A/V-MT-AffWildNet to state-of-the-art methods in Aff-Wild2, developed by the top-3 performing teams of the ABAW Competition. We note that the corresponding performance metrics are the CCC for valence and arousal, a weighted average between the F1 score and the total accuracy, as discussed in Section 5.1.3 for expression recognition and the average between the F1 score and the total accuracy, as discussed in Section 5.1.3 for action unit detection.

It can be easily observed that the MT-AffWildNet, either when trained with the audio, or visual modality, outperformed, in all tasks, all other methods; the same happened with the best performing A/V-MT-AffWildNet. It can also be seen that the MT-VGGFACE displayed a slightly worse performance in all tasks than the best performing other methods, which was expected given that the other methods used either an ensemble methodology of CNN-RNN networks, or fused the visual and audio modalities.

Table 5.6: Performance Comparison on Aff-Wild2 for the VA, Expr, AU tasks between our developments and the state-of-the-art developed by the top-3 performing teams of ABAW Competition (and the baseline method); '-' means that no result is reported in the corresponding paper; $\mathcal{E}_{total}^{Expr} = 0.67 \times F_1 + 0.33 * \mathcal{T}Acc$; $\mathcal{E}_{total}^{AU} = 0.5 \times \mathcal{A}F_1 + 0.5 * \mathcal{T}Acc$

Network	Team Name	Aff-Wild2			
		CCC-V	CCC-A	$\mathcal{E}_{total}^{Expr}$	$\mathcal{E}_{total}^{AU}$
Baseline	-	0.11	0.27	0.30	0.26
CNN-RNN Ensemble	NISL2020	0.440	0.454	0.405	0.607
Two-Stream Aural-Visual Model	TNT	0.448	0.417	0.509	0.601
Multi-Modal MultiFeature Network	ICT-VIPL	0.361	0.408	0.408	-
MT-VGGFACE	-	0.430	0.420	0.495	0.601
MT-AffWildNet (audio modality)	-	0.440	0.510	0.513	0.615
MT-AffWildNet (visual modality)	-	0.460	0.450	0.521	0.619
A/V-MT-AffWildNet	-	**0.470**	**0.520**	**0.532**	**0.624**

Results: Cross Database Experiments & Comparison with State-of-the-Art

Table 5.7 presents a cross-database comparison for the three tasks on 10 databases, between the state-of-the-art in these databases and our developed networks. Cross-database means that the models were trained on Aff-Wild2 and were then evaluated on the other databases. Let us note that the mean diagonal value of the confusion matrix was the evaluation criterion for RAF-DB, the accuracy metric was the one for FER2013, CCC was the criterion in all dimensionally annotated databases and the F1 score was the criterion for all other databases. Let us also note that AffectNet, RAF-DB, FER2013 and EmotioNet are static databases, meaning that they contain only images and not videos and thus we could only test the MT-VGGFACE on them. AFEW-VA database does not contain audio and thus we could not test the A/V-MT-AffWildNet on it. DISFA, BP4DS and BP4D+ do not contain audio as well.

Table 5.7: Cross-database evaluation (models trained on Aff-Wild2 and tested on other databases) for the three tasks on 10 databases, between the state-of-the-art and our developed networks; VA evaluation is shown as CCCV-CCCA; the mean diagonal value of the confusion matrix (denoted as 'Diag.') was the evaluation criterion for RAF-DB; 'Acc' stands for Accuracy; '-' means that either the database did not contain audio or the database is a static one consisting of only images or the network was not trained on this database or the network was not trained for this task

Network	Aff-Wild	AFEW-VA	AffectNet		RAF-DB	FER2013	EmotioNet	DISFA	BP4DS	BP4D+
	CCC	CCC	CCC	F1	Diag.	Acc.	F1	F1	F1	F1
best CNN [105, 109]	0.51-0.33	0.49-0.52	0.51-0.36	-	-	-	-	-	-	-
AffWildNet [105, 109]	0.57-0.43	0.52-0.56	-	-	-	-	-	-	-	-
AlexNet [151]	-	-	0.6-0.34	**0.58**	-	-	-	-	-	-
VGGFACE [128]	-	-	-	-	0.58	-	-	-	-	-
VGG [70]	-	-	-	-	-	0.75	-	-	-	-
ResNet-34 [57]	-	-	-	-	-	-	0.51	-	-	-
FVGG [129]	-	-	-	-	-	-	-	0.52	-	-
R-T1 [129]	-	-	-	-	-	-	-	0.6	-	-
DLE extension [225]	-	-	-	-	-	-	-	-	0.54	-
LGBP+Geometric [203]	-	-	-	-	-	-	-	-	0.53	-
VGG+SVM [193]	-	-	-	-	-	-	-	-	-	0.51
Geometric+CRF [204]	-	-	-	-	-	-	-	-	-	0.34
MT-VGGFACE	0.56-0.35	0.58-0.53	**0.61-0.46**	*0.54*	**0.61**	**0.76**	**0.52**	0.61	0.66	0.49
MT-AffWildNet (audio modality)	0.54-0.47	-	-	-	-	-	-	-	-	-
MT-AffWildNet (visual modality)	0.6-0.45	**0.6-0.6**	-	-	-	-	-	**0.63**	**0.67**	**0.52**
A/V-MT-AffWildNet	**0.62-0.49**	-	-	-	-	-	-	-	-	-

The A/V-MT-AffWildNet achieved the best performance in Aff-Wild for both valence and arousal estimation, outperforming the existing state-of-the-art AffWildNet. Moreover, it can be seen that MT-AffWildNet performed best for valence estimation when trained, on Aff-Wild, with the visual modality, whilst performing best for arousal when trained with the audio modality. This is because audio tends to have thematic constancy. Consider, for example, two fight sequences in a movie, one being a flashy fight scene and the other a one-sided fight with a person being injured. In both cases, arousal can be high due to loud and pronounced music, but valence will be positive in the former and negative in the latter sequence.

Table 5.7 shows that the MT-AffWildNet, trained with the visual modality, outperformed the fine-tuned AffWildNet in AFEW-VA database. It can also be observed that the MT-VGGFACE outperformed: i) the state-of-the-art AlexNet [151] on AffectNet both in valence and arousal estimation, ii) the state-of-the-art VGGFACE [128] in RAF-DB, iii) the state-of-the-art VGG [70] in FER2013 and iv) the winner [57] of Emotionet 2017 Challenge, ResNet-34. Only, in expression recognition in AffectNet, the obtained performance of the MT-VGGFACE is lower to the state-of-the-art. Finally, the MT-AffWildNet trained with the visual modality outperformed: i) the fine-tuned VGG (FVGG)

and R-TI methods [129] in DISFA, ii) the baseline LGBP+Geometric [203] and the winner DLE extension [225] of FERA 2015 Challenge and iii) the baseline Geometric+CRF [204] and the winner VGG+SVM [193] of FERA 2017 Challenge.

5.2 A Holistic Approach to Affect Recognition in-the-wild

In the previous subsection we described and tested the proposed multi-task networks, trained on Aff-Wild2. There, we exploited the fact that Aff-Wild2 contains annotations for all three affect recognition tasks. It should be, however, mentioned that the other existing databases contain annotations for only one or two of the tasks and not for all three of them; it is not, therefore, straightforward how to take advantage of the knowledge developed by the multi-task networks in these other cases.

In this Section, we present the first and largest study of all facial behaviour tasks learned jointly in a single holistic framework, deriving the novel FaceBehaviorNet architecture. To achieve this, we utilise all publicly available datasets (including over 5 million images) that study facial behaviour tasks in-the-wild. At first, we demonstrate that by training an end-to-end network jointly on all tasks, we consistently achieve a better performance than by training each of the single-task networks. Then, we propose a new approach for coupling all three tasks during training, through co-annotation and distribution matching. We show that this approach performs well under partial, or non-overlapping, annotation of the datasets. Finally we show that FaceBehaviorNet learns features that encapsulate all aspects of facial behaviour; being capable of successfully performing tasks (such as compound emotion recognition) beyond the ones that it has been trained, in an efficient zero- or few-shot learning setting.

5.2.1 Related Work

Holistic frameworks, where several parts, e.g. learning tasks, are interconnected and explicable by the reference to the whole, are common in computer vision. The diverse examples range from the scene understanding framework that reasons about 3D object detection, pose estimation, semantic

segmentation and depth reconstruction [208], the face analysis framework that addresses face detection, landmark localization, gender recognition, age estimation [170], to the universal networks for low-, mid-, high-level vision [102] and for various visual tasks [228]. Most if not all of these prior works rely on building a multi-task framework where learning is done based on the ground truth annotations with full or partial overlap across tasks. During training, all tasks are optimised simultaneously aiming at representation learning that supports the holistic view.

In this work we propose the first holistic framework for emotional behaviour analysis in-the-wild, where different emotional states such as binary action units activations, basic categorical emotions and continuous dimensions of valence and arousal constitute the interconnected tasks that are explicable by the human's affective state. What makes it different from the aforementioned holistic approaches is exploring the idea of task-relatedness, given explicitly either as external expert knowledge or from empirical evidence. In this form, it is similarly motivated to the classical multi-task literature exploring feature sharing [6] and task relatedness [93] during training; more examples can be found in the surveys [159, 233]. However in the multi-task setting, one typically assumes homogeneity of the tasks, i.e. tasks of the same type such as object classifiers or attribute detectors. The main difference and novelty of our work is that the proposed holistic framework (i) explores the relatedness of non-homogeneous tasks, e.g. tasks for classification, detection, regression; (ii) operates over datasets with partial or non-overlapping annotations of the tasks; (iii) encodes explicit relationship between tasks to improve transparency and to enable expert input.

Works exist in literature that use emotion labels to complement missing AU annotations or increase generalization of AU classifiers [175, 209, 219]. Our work deviates from such methods, as we target a joint learning of three facial behaviour tasks via a single holistic framework, whilst these works perform only AU detection and not emotion recognition (nor valence-arousal estimation).

One of the closest goals to ours is [26], where an integrated deep learning framework (FATAUVA-Net) for sequential facial attribute recognition, AU detection, and valence-arousal estimation was proposed. This framework employed face attributes as low-level (first component) and AUs as mid-level (second component) representations for predicting quantized valence-arousal values (third component). However training of this model is made of transfer learning and fine-tuning steps, is hierarchical and

not end-to-end. In a similar work of [210], a two-level attention with two stage multi-task learning framework was constructed for emotion recognition and valence-arousal estimation; this work was based on a database (AffectNet [151]) annotated for both tasks. In the first attention level, a CNN extracted position-level features and then in the second an RNN with self-attention was proposed to model the relationship between layer-level features.

5.2.2 The Proposed Approach

In the following:

- We propose a flexible holistic framework that can accommodate non-homogeneous tasks with encoding prior knowledge of tasks relatedness. In our experiments we evaluate two effective strategies of task relatedness: a) obtained from a cognitive and psychological study, e.g. how action units are related to basic emotion categories [59], and b) inferred empirically from external dataset annotations.
- We propose an effective algorithmic approach of coupling the tasks via co-annotation and distribution matching and show its effectiveness for facial behaviour analysis.
- We present the first, to the best of our knowledge, holistic network for facial behaviour analysis (FaceBehaviorNet) and train it, in end-to-end manner, for simultaneously predicting 7 basic expressions, 17 action units and continuous valence-arousal, in-the-wild. For network training we utilise all publicly available in-the-wild databases that, in total, consist of over 5M images with partial and/or non-overlapping annotations for different tasks.
- We show that FaceBehaviorNet greatly outperforms each of the single-task networks, validating that the network's affect recognition capabilities are enhanced when it is jointly trained for all related tasks. We further explore the feature representation learned during joint training and show its good generalisation on the task of compound expression recognition, when no, or little, training data is available (corresponding to zero-shot and few-shot learning).

As in the former subsection, we start with the multi-task formulation of the facial behaviour model. In this model we have three objectives: (1) learning seven basic emotions, (2) detecting activations

of 17 binary facial action units, (3) learning the intensity of the valence and arousal continuous affect dimensions. Our target is to train a multi-task neural network model to jointly achieve objectives (1)-(3). However, now we assume that for a given image $x \in \mathcal{X}$, we can have a single type of label annotations; i.e., in terms of either the seven basic emotions $y_{emo} \in \{1,2,\ldots,7\}$, or the 17[1] binary action units activations $y_{au} \in \{0,1\}^{17}$, or the two continuous affect dimensions, valence and arousal, $y_{va} \in [-1,1]^2$. For simplicity of presentation, we use the same notation x for all images leaving the context to be explained by the label notations. We train the multi-task model by minimizing the following total objective, which is similar to (4.1)-(4.3), with a slight change in the symbols and notations used, so as to fit the following developments:

$$\begin{aligned} \mathcal{L}_{MT} &= \mathcal{L}_{Emo} + \lambda_1 \mathcal{L}_{AU} + \lambda_2 \mathcal{L}_{VA} \\ \mathcal{L}_{Emo} &= \mathbb{E}_{x,y_{emo}}[-\log\, p(y_{emo}|x)] \\ \mathcal{L}_{AU} &= \mathbb{E}_{x,y_{au}}[-\log\, p(y_{au}|x)] \\ \mathcal{L}_{VA} &= 1 - CCC(y_{va}, \bar{y}_{va}), \end{aligned} \tag{5.10}$$

where: the first term is the cross entropy loss computed over images with a basic emotion label; the second term is the binary cross entropy loss computed over images with 17 AU activations, $\log\, p(y_{au}|x) := [\sum_{k=1}^{17} \delta_k]^{-1} \cdot \sum_{i=1}^{17} \delta_i \cdot [y^i_{au} \log\, p(y^i_{au}|x) + (1 - y^i_{au}) \log\, (1 - p(y^i_{au}|x))]$, with $\delta_i \in \{0,1\}$ indicating whether the image contains annotation for AU_i; the third term measures the concordance correlation coefficient between the ground truth valence and arousal y_{va} and the predicted $\bar{y}_{va}$, $CCC(y_{va}, \bar{y}_{va}) = \frac{\rho_a + \rho_v}{2}$, in which for $i \in \{v,a\}$, y_i is the ground truth, $\bar{y}_i$ is the predicted value and

$$\rho_i = \frac{2 \cdot \mathbb{E}[(y_i - \mathbb{E}_{y_i}) \cdot (\bar{y}_i - \mathbb{E}_{\bar{y}_i})]}{\mathbb{E}^2[(y_i - \mathbb{E}_{y_i})^2] + \mathbb{E}^2[(\bar{y}_i - \mathbb{E}_{\bar{y}_i})^2] + (\mathbb{E}_{y_i} - \mathbb{E}_{\bar{y}_i})^2} \tag{5.11}$$

Coupling of basic emotions and AUs via co-annotation In the seminal work of [59], the authors conduct a study on the relationship between emotions (basic and compound) and facial action unit

[1] 17 is an aggregate of action units in all datasets; typically each dataset has from 10 to 12 AUs labelled by purposely trained annotators.

Table 5.8: Basic emotions and their prototypical and observational AUs from [59]. The weights w in brackets correspond to the fraction of annotators that observed the AU activation.

Emotion	Prototypical AUs	Observational AUs (with weights w)
happiness	12, 25	6 (0.51)
sadness	4, 15	1 (0.6), 6 (0.5), 11 (0.26), 17 (0.67)
fear	1, 4, 20, 25	2 (0.57), 5 (0.63), 26 (0.33)
anger	4, 7, 24	10 (0.26), 17 (0.52), 23 (0.29)
surprise	1, 2, 25, 26	5 (0.66)
disgust	9, 10, 17	4 (0.31), 24 (0.26)

activations. The summary of the study is a Table of the emotions and their prototypical and observational actions units (Table 1 in [59]), which we include in Table 5.8 for completeness. Prototypical ones are action units that are labelled as activated across all annotators' responses, observational are action units that are labelled as activated by a fraction of annotators. For example, in emotion *happiness* the prototypical are AU12 and AU25, the observational is AU6 with weight 0.51 (observed by 51% of the annotators).

Here let us mention that Table 5.8 constitutes the relatedness between the emotion categories and action units obtained from a cognitive study. In our experiments, in Section 5.2.4, we also show that such relatedness can be inferred empirically from external dataset annotations.

We propose a *co-annotation* strategy to couple the training of emotions and action unit predictions. Given an image x with the ground truth basic emotion y_{emo}, we enforce the prototypical and observational AUs of this emotion to be activated. We co-annotate the image (x, y_{emo}) with y_{au}; this image contributes to both $\mathcal{L}_{Emo}$ and $\mathcal{L}_{AU}$[2] in eq. 5.10. We re-weight the contributions of the observational AUs with the annotators' agreement score (from Table 5.8).

Similarly, for an image x with ground truth action units y_{au}, we check whether we can co-annotate it with an emotion label. For an emotion to be present, all its prototypical and observational AUs have to be present. In cases when more than one emotion is possible, we assign the label y_{emo} of the emotion with the largest requirement of prototypical and observational AUs. The image (x, y_{au}) that is co-annotated with the emotion label y_{emo} contributes to both $\mathcal{L}_{AU}$ and $\mathcal{L}_{Emo}$ in eq. 5.10. We use this approach to develop FaceBehaviorNet, with co-annotation.

[2]Here we overload slightly our notations; for co-annotated images, y_{au} has variable length and only contains prototypical and observational AUs.

Coupling of basic emotions and AUs via distribution matching The aim here is to align the *predictions* of emotions and action units tasks during training. For each sample x we have the predictions of emotions $p(y_{emo}|x)$ as the softmax scores over seven basic emotions and we have the prediction of AU activations $p(y^i_{au}|x)$, $i = 1,\ldots,17$ as the sigmoid scores over 17 AUs.

The distribution matching idea is the following: we match the distribution over AU predictions $p(y^i_{au}|x)$ with the distribution $q(y^i_{au}|x)$, where the AUs are modeled as a mixture over the basic emotion categories:

$$q(y^i_{au}|x) = \sum_{y_{emo}\in\{1,\ldots,7\}} p(y_{emo}|x)\; p(y^i_{au}|y_{emo}), \tag{5.12}$$

where $p(y^i_{au}|y_{emo})$ is defined in a deterministic way from Table 5.8 and is equal to 1 for prototypical/observational action units, or to 0 otherwise. For example, AU2 is prototypical for emotion *surprise* and observational for emotion *fear* and thus $q(y_{\text{AU2}}|x) = p(y_{\text{surprise}}|x) + p(y_{\text{fear}}|x)$[3].

This matching aims to make the network's predicted AUs consistent with the prototypical and observational AUs of the network's predicted emotions. So if, e.g., the network predicts the emotion *happiness* with probability 1, i.e., $p(y_{\text{happiness}}|x) = 1$, then the prototypical and observational AUs of *happiness*, i.e., AUs 12, 25 and 6- need to be activated in the distribution q: $q(y_{\text{AU12}}|x) = 1$; $q(y_{\text{AU25}}|x) = 1$; $q(y_{\text{AU6}}|x) = 1$; $q(y^i_{au}|x) = 0, i \in \{1,..,14\}$.

In spirit of the distillation approach [83], we match the distributions $p(y^i_{au}|x)$ and $q(y^i_{au}|x)$ by minimizing the cross entropy with the soft targets loss term[4]:

$$\mathcal{L}_{DM} = \mathbb{E}_x \sum_{i=1}^{17} [-p(y^i_{au}|x) \log q(y^i_{au}|x)], \tag{5.13}$$

where all available training samples are used to match the predictions. We use this approach to develop FaceBehaviorNet, with distr-matching.

A mix of the two strategies, co-annotation and distribution matching, is also possible. Given an image x with the ground truth annotation of the action units y_{au}, we can first co-annotate it with a *soft label* in form of the distribution over emotions and then match it with the predictions of emotions

[3] We also tried a variant with reweighting for observational AUs, i.e. $p(y^i_{au}|y_{emo}) = w$

[4] This can be seen as minimizing the KL-divergence $KL(p||q)$ across the 17 action units.

$p(y_{emo}|x)$. More specifically, for each basic emotion, we compute the score of its prototypical and observational AUs being present. For example, for emotion *happiness*, we compute $(y_{\text{AU12}} + y_{\text{AU25}} + 0.51 \cdot y_{\text{AU6}})/(1+1+0.51)$, or se all weights to be equal to 1, when no reweighting is used. We take a softmax over the scores to produce the probabilities over emotion categories. In this variant, every single image that has ground truth annotation of AUs will have a *soft* emotion label assigned to it. Finally we match the predictions $p(y_{emo}|x)$ and the soft label by minimizing the cross entropy with soft targets similarly to eq. 5.13. We use this approach to develop FaceBehaviorNet, with soft co-annotation.

Coupling of categorical emotions and AUs with continuous affect In our work, continuous affect (valence and arousal) is implicitly coupled with the basic expressions and action units via a joint training procedure. Also one of the datasets we used has annotations for categorical and continuous emotions (AffectNet [151]). Studying an explicit relationship between them is a novel research direction beyond the scope of this work.

FaceBehaviorNet structure Fig.5.2 shows the structure of the holistic (multi-task, multi-domain and multi-label) FaceBehaviorNet, based on the 13 convolutional and pooling layers of VGG-FACE [161] (its fully connected layers are discarded), followed by 2 fully connected layers, each with 4096 hidden units. The structure of FaceBehaviorNet is based on the VGG-Face as it has been pre-trained with a large dataset for face recognition and therefore many human faces have been used in its construction. A (linear) output layer follows that gives final estimates for valence and arousal; it also gives 7 basic expression logits that are passed through a softmax function to get the final 7 basic expression predictions; lastly, it gives 17 AU logits that are passed through a sigmoid function to get the final 17 AU predictions. One can see that the predictions for all tasks are pooled from the same feature space.

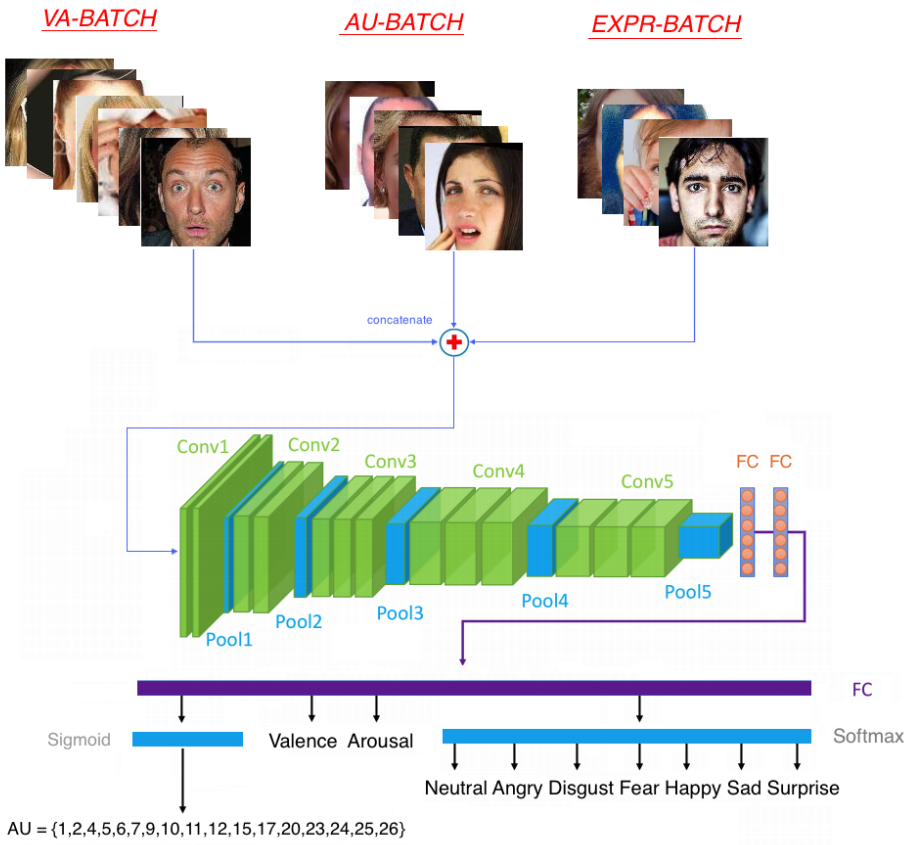


Figure 5.2: The holistic (multi-task, multi-domain, multi-label) FaceBehaviorNet; 'VA/AU/EXPR-BATCH' refers to batches annotated in terms of VA/AU/7 basic expressions

5.2.3 Pre-Processing, Performance Measures & Network Training Details

At first, we describe the two pre-processing steps that has been used to generate the input data for affect analysis. Next, we present the databases that we utilised, as well as the evaluation metrics used across these databases. Finally, we present the network implementation details.

Pre-Processing

We used the SSH detector [152] based on ResNet and trained on the WiderFace dataset [220] to extract, from all images, face bounding boxes and 5 facial landmarks; the latter were used for face alignment. All cropped and aligned images were resized to $96 \times 96 \times 3$ pixel resolution and their intensity values were normalized to $[-1, 1]$.

Databases

We utilised the following databases in our experiments: Aff-Wild, AffectNet, AFEW, RAF-DB, BP4D-Spontaneous (denoted as BP4DS), BP4D+, DISFA, EmotioNet. We selected to work with these databases because they provide a large number of samples with accurate annotations of valence-arousal, basic expressions and AUs. Through training with these datasets, the networks learn to recognize affective states under a large number of image conditions, e.g., different resolutions, poses, orientations and lighting conditions. These datasets also include a variety of samples in gender, ethnicity and race.

Performance measures

We use:

- the CCC for Aff-Wild (as CCC was the evaluation criterion of Aff-Wild Challenge) and AffectNet (for valence and arousal estimation, we use CCC to be consistent)
- the total accuracy for AFEW (as this metric was the evaluation criterion of the EmotiW Challenges), the mean diagonal value of the confusion matrix for RAF-DB (as this criterion was selected for evaluating the performance in this database by [128]), the F1 score for AffectNet (for evaluating the 7 basic expressions, as this metric is widely used in classification task)
- the F1 score for BP4D-Spontaneous, BP4D+ (as this metric was the evaluation criterion of the respective FERA 2015 and 2017 Challenges) and DISFA (for consistency purposes) For AU detection in Emotionet, the Challenge's metric was the average between: a) the mean F1 score, across all AUs, b) the mean accuracy, across all AUs; regarding the emotion classification it was the average between: a) the mean F1 score, b) the unweighted average recall (UAR), over all emotion categories.

Training Implementation Details

At this point let us describe the strategy that was used for feeding images from different databases to FaceBehaviorNet. At first, the training set was split into three different sets, each of which contained

images that were annotated in terms of either valence-arousal, or action units, or seven basic expressions; let us denote these sets as VA-Set, AU-Set and EXPR-Set, respectively. During training, at each iteration, three batches, one from each set (as can be seen in Fig.5.2), were concatenated and fed to FaceBehaviorNet. This step was important for network training, because: i) the network minimizes the objective function of eq. 5.10; at each iteration, the network has seen images from all categories and thus all loss terms contribute to the objective function, ii) since the network sees an adequate number of images from all categories, the weight updates (during gradient descent) are not based on noisy gradients; this in turn prevents poor convergence behaviors; otherwise, we would need to tackle these problems, e.g. do asynchronous SGD as proposed in [102] to make the task parameter updates decoupled, iii) the CCC cost function needs an adequate sequence of predictions.

Since VA-Set, AU-Set and EXPR-Set had different sizes, they needed to be 'aligned'. To do so, we selected the batches of these sets in such a manner, so that after one epoch we have sampled all images in the sets. In particular, we chose batches of size 401, 247 and 103 for the VA-Set, AU-Set and EXPR-Set, respectively. The training of FaceBehaviorNet was performed in an end-to-end manner, with a learning rate of 10^{-4}. A 0.5 Dropout value was used in the fully connected layers. Training was performed on a Tesla V100 32GB GPU and training time was about 2 days. The Tensorflow platform has been used.

5.2.4 Task-Relatedness from Empirical Evidences

Table 5.8 was created using a cognitive and psychological study with human participants. Next, we created another Table inferred empirically from external dataset annotations. In particular, we used the recently proposed Aff-Wild2 database, which is the first in-the-wild database that contains annotations for all three behavior tasks that we are dealing with.

At first, we trained a network for AU detection on the union of Aff-Wild2 and GFT databases [72]. Next, this network was used for automatically annotating all Aff-Wild2 videos with AUs. Table 5.9 shows the distribution of AUs for each basic expression. In parenthesis next to each AU (e.g. AU12) is the percentage of images (0.82) annotated with the specific expression (happiness) in which this

AU (AU12) was activated.

Table 5.9: Relatedness between basic emotions and AUs, inferred from Aff-Wild2

Emotion	AUs (with weights w)
happy	12 (0.82), 25 (0.7), 6 (0.57), 7 (0.83), 10 (0.63)
sad	4 (0.53), 15 (0.42), 1 (0.31), 7 (0.13), 17 (0.1)
fearful	1 (0.52), 4 (0.4), 25 (0.85), 5 (0.38), 7 (0.57), 10 (0.57)
angry	4 (0.65), 7 (0.45), 25 (0.4), 10 (0.33), 9 (0.15)
surprised	1 (0.38), 2 (0.37), 25 (0.85), 26 (0.3), 5 (0.5), 7 (0.2)
disgusted	9 (0.21), 10 (0.85), 17 (0.23), 4 (0.6), 7 (0.75), 25 (0.8)

5.2.5 Experimental Results

Ablation Study on Loss Functions

At first, we compare the performance of FaceBehaviorNet when trained: i) with only the loss functions of eq. 5.10 and without using the coupling losses described in Section 5.2.2, ii) with co-annotation coupling loss, iii) with soft co-annotation coupling loss, iv) with distr-matching coupling loss, v) with soft co-annotation and distr-matching coupling losses. Table 5.10 shows the results obtained using all these approaches, whilst Tables 5.8 and 5.9 are used for the task relatedness.

Table 5.10: Performance evaluation of valence-arousal, seven basic expression and action units predictions on all used databases provided by the FaceBehaviorNet when trained with/without the coupled losses, under the two task relatedness scenarios.

Databases	Relatedness	Aff-Wild		AffectNet			AFEW	RAF-DB	EmotioNet		DISFA	BP4DS	BP4D+
FaceBehaviorNet		CCC-V	CCC-A	CCC-V	CCC-A	F1 Score	Total Accuracy	Mean diag. of conf. matrix	F1 Score	Accuracy	F1 Score	F1 Score	F1 Score
no coupling loss	-	0.55	0.36	0.56	0.46	0.54	0.38	0.67	0.49	0.94	0.52	0.61	0.57
co-annotation	[59]	0.56	0.38	0.56	0.46	0.55	0.40	0.67	0.49	0.94	0.54	0.64	0.58
soft co-annotation	[59]	0.56	0.39	0.57	0.47	0.57	0.41	0.67	0.50	0.94	0.54	0.64	0.60
distr-matching	[59]	0.56	0.37	0.57	0.49	0.57	0.42	0.68	0.50	0.94	0.56	0.66	0.58
soft co-annotation and distr-matching	[59]	0.59	**0.41**	0.59	0.50	**0.60**	**0.43**	0.70	0.51	**0.95**	0.57	**0.67**	**0.60**
co-annotation	Aff-Wild2	0.55	0.37	0.56	0.47	0.54	0.40	0.67	0.50	0.93	0.54	0.61	0.57
soft co-annotation	Aff-Wild2	0.56	0.37	0.57	0.47	0.55	0.42	0.68	0.52	0.94	0.58	0.63	0.59
distr-matching	Aff-Wild2	0.57	0.39	0.60	**0.51**	0.57	0.42	0.69	0.50	0.94	0.57	0.62	0.58
soft co-annotation and distr-matching	Aff-Wild2	**0.60**	0.40	**0.61**	**0.51**	**0.60**	0.42	**0.71**	**0.54**	0.94	**0.60**	0.66	**0.60**

Many deductions can be made. *Firstly*, when FaceBehaviorNet is trained with any coupling loss, or any combination of these, it displays a better (or in the worst case equal) performance on all databases, in both task relatedness scenarios. This validates the fact that the proposed losses help to couple the three studied tasks regardless of which relatedness scenario was followed; this shows the generality

of the proposed losses that boosted the performance of the network. *Secondly*, the performance in estimation of valence and arousal improved, although we did not explicitly designed a coupling loss for this; we only coupled emotion categories and action units. We conjecture that when action unit detection and expression classification accuracy is improving (due to coupling), valence and arousal performance also improves, because valence and arousal are implicitly coupled with emotions via joint dataset annotations for both emotion types.

Thirdly, in all scenarios, the co-annotation loss results in FaceBehaviorNet having the worst performance when compared to all other coupling losses. *Furthermore*, in both settings, when the network was trained with the soft co-annotation loss, the performance increase in AUs was bigger than the corresponding increase in expressions; whereas, when the network was trained with the distr-matching loss the performance increase in expressions was bigger than the corresponding increase in AUs. *Finally*, overall best results have been achieved, in both scenarios, when FaceBehaviorNet was trained with both soft co-annotation and distr-matching losses. In particular, in both settings, an average performance increase of more than 2% has been observed when using both coupling losses, compared to the (two) cases when only one of them was used.

Comparison with State-of-the-Art and Single-Task Methods

Table 5.11: Performance evaluation of valence-arousal, seven basic expression and action units predictions on all utilised databases provided by the FaceBehaviorNet and state-of-the-art methods.

Databases	Aff-Wild		AffectNet			AFEW	RAF-DB	EmotioNet		DISFA	BP4DS	BP4D+
	CCC-V	CCC-A	CCC-V	CCC-A	F1 Score	Total Accuracy	Mean diagonal of conf. matrix	F1 Score	Mean Accuracy	F1 Score	F1 Score	F1 Score
best performing CNN [105, 109]	0.51	0.33	-	-	-	-	-	-	-	-	-	-
FATAUVA-Net [26]	0.40	0.28	-	-	-	-	-	-	-	-	-	-
(2 ×) AlexNet [151]	-	-	0.60	0.34	0.58	-	-	-	-	-	-	-
non-linear SVM [47]	-	-	-	-	-	0.38	-	-	-	-	-	-
VGG-FACE-mSVM [128]	-	-	-	-	-	-	0.58	-	-	-	-	-
AlexNet [17]	-	-	-	-	-	-	-	0.39	0.83	-	-	-
ResNet-34 [57]	-	-	-	-	-	-	-	**0.64**	0.82	-	-	-
DLE extension [225]	-	-	-	-	-	-	-	-	-	-	0.59	-
[193]	-	-	-	-	-	-	-	-	-	-	-	0.58
(3 ×) VGG-FACE single-task	0.52	0.31	0.53	0.43	0.51	0.37	0.59	0.41	0.92	0.47	0.56	0.54
FaceBehaviorNet, no coupling loss	*0.55*	*0.36*	0.56	*0.46*	0.54	*0.38*	*0.67*	0.49	*0.94*	*0.52*	*0.61*	0.57
FaceBehaviorNet, soft co-annotation and distr-matching, [59]	0.59	**0.41**	0.59	0.50	**0.60**	**0.43**	0.70	0.51	**0.95**	0.57	**0.67**	**0.60**
FaceBehaviorNet, soft co-annotation and distr-matching, Aff-Wild2	**0.60**	0.40	**0.61**	**0.51**	**0.60**	0.42	**0.71**	0.54	0.94	**0.60**	0.66	**0.60**

Next, we trained a VGG-FACE network on all the dimensionally annotated databases to predict valence and arousal; we also trained another VGG-FACE network on all categorically annotated

databases, to perform seven basic expression classification; finally we trained a third VGG-FACE network on all databases annotated with action units, so as to perform AU detection. For brevity these three single-task networks are denoted as '(3 $\times$) VGG-FACE single-task' in one row of Table 5.11.

We compared these networks' performance with the performance of FaceBehaviorNet when trained with and without the coupling losses. We also compare them with the performances of the state-of-the-art methodologies of each utilised database: i) FATAUVA-Net [26], which was the winner of Aff-Wild Challenge; ii) the best performing CNN (VGG-FACE) on Aff-Wild [105, 109]; iii) the best performing networks (AlexNet) on AffectNet [151] (in Table 5.11 they are denoted as '(2 $\times$) AlexNet' as they are two different networks: one for VA estimation and another for expression classification); iv) the baseline network (non-linear Chi-square kernel based SVM) [51] on EmotiW Challenges; v) VGG-FACE-mSVM [128] on RAF-DB; vi) the baseline network (AlexNet) on EmotioNet [17]; vii) ResNet-34, which was the winner of the EmotioNet Challenge [57]; viii) Discriminant Laplacian Embedding extension (DLE extension) [225], which was the winner of FERA 2015 on BP4DS; ix) [193], which was the winner of FERA 2017 on BP4D+. Table 5.11 displays the performance of all these networks.

Here, let us mention that on RAF-DB the best performing network is the Deep Locality-preserving CNN (DLP-CNN) of [128] with a performance metric value of 0.74; this network was trained using a joint classical softmax loss - which forces different classes to stay apart - and a newly created loss - that pulls the locally neighboring faces of the same class together. For the task of expression recognition, our approach used the standard cross entropy loss; therefore a fair comparison cannot be made with our model because DLP-CNN uses a different cost function that we do not use and thus DLP-CNN is not listed in Table 5.11.

It might be argued that the more data used for network training (even if they contain partial or non-overlapping annotations), the better network performance will be in all tasks. However this may not be true, as the three studied tasks are non-homogeneous and each one of them contains ambiguous cases: i) there is generally discrepancy in the perception of the disgust, fear, sadness and (negative) surprise emotions across different people and across databases; ii) the exact valence and arousal value for a particular affect is also not consistent among databases; iii) the AU annotation process is a hard to do

and error prone one. Nevertheless, from Table 5.11, it can be verified that FaceBehaviorNet achieved a better performance on all databases than the independently trained VGG-FACE single-task models. This shows that, all described facial behavior understanding tasks are coherently correlated to each other; training an end-to-end architecture with heterogeneous databases simultaneously, therefore, leads to improved performance.

In Table 5.11, it can be observed that FaceBehaviorNet trained with no coupling loss: i) ouperforms the state-of-the-art by 3.5% (average CCC) on Aff-Wild, 4% (average CCC) on AffectNet, 9% on RAF-DB and 2% on BP4DS; ii) has the same performance on AFEW; iii) shows inferior performance by 4% on AffectNet and 1.5% (on average) on EmotioNet, 1% on BP4D+. However, when FaceBehaviorNet is trained with soft co-annotation and distr-matching losses (either when task relatedness is inferred from Aff-Wild2 or from [59]), it shows superior performance to all state-of-the-art methods. The fact that it outperforms these methods and the single-task networks, in both task relatedness settings, verifies the generality of the proposed losses; network performance is boosted independently of the Table of task relatedness which was used.

Zero-Shot and Few-Shot Learning

In order to further prove and validate that FaceBehaviorNet learned good features encapsulating all aspects of facial behaviour, we conducted zero-shot learning experiments for classifying compound expressions. Given that there exist only 2 datasets (EmotioNet and RAF-DB) annotated with compound expressions and that they do not contain a lot of samples (less than 3,000 each), at first, we used the predictions of FaceBehaviorNet together with the rules from [59] to generate compound emotion predictions. Additionally, to demonstrate the superiority of FaceBehaviorNet, we used it as a pre-trained network in a few-shot learning experiment. We took advantage of the fact that our network has learned good features and used them as priors for fine-tuning the network to perform compound emotion classification.

RAF-DB database At first, we performed zero-shot experiments on the 11 compound categories of RAF-DB. We computed a candidate score, $C_s(y_{emo})$, for each class y_{emo}:

$$\mathcal{C}_s(y_{emo}) = [\sum_{k=1}^{17} p(y_{au}^k|y_{emo})]^{-1} \cdot \sum_{k=1}^{17} p(y_{au}^k|x)\, p(y_{au}^k|y_{emo}) + p(y_{emo1}) + p(y_{emo2}) + 0.5 \cdot (\frac{p(y_v|x)}{|p(y_v|x)|} + 1), p(y_v|x) \neq 0, \quad (5.14)$$

where: i) the first term of the sum is FaceBehaviorNet's predictions of only the prototypical (and observational) AUs that are associated with this compound class according to [59]; in this manner, every AU acts as an indicator for this particular emotion class; this terms describes the confidence (probability) of AUs that this compound emotion is present; ii) $p(y_{emo1})$ and $p(y_{emo2})$ are FaceBehaviorNet's predictions of only the basic expression classes $emo1$ and $emo2$ that are mixed and form the compound class (e.g., if the compound class is happily surprised then $emo1$ is happy and $emo2$ is surprised); iii) the last term of the sum is added only to the happily surprised and happily disgusted classes and is either 0 or 1 depending on whether FaceBehaviorNet's valence prediction is negative or positive, respectively; the rationale is that only happily surprised and (maybe) happily disgusted classes have positive valence; all other classes are expected to have negative valence as they correspond to negative emotions. Our final prediction was the class that had the maximum candidate score.

Table 5.12 shows the results of this approach when we used the predictions of FaceBehaviorNet trained with and without the soft co-annotation and distr-matching losses. Best results have been obtained when the network was trained with the coupling losses. One can observe, that this approach outperformed by 4.8% the VGG-FACE-mSVM [128] which has the same architecture as our network and it has been trained for compound emotion classification.

Next, we target few-shot learning. In particular, we fine-tune the FaceBehaviorNet (trained with and without the soft co-annotation and distr-matching losses) on the small training set of RAF-DB. In Table 5.12 we compare its performance to a state-of-the-art network. It can be seen that our fine-tuned FaceBehaviorNet, trained with and without the coupling losses, outperformed by 1.2% and 3.7%, respectively, the best performing network, DLP-CNN, that was trained with a loss designed for

this specific task.

EmotioNet database Next, we performed zero-shot experiments on the EmotioNet basic and compound set that was released for the related Challenge. This set includes 6 basic plus 10 compound categories, as described at the beginning of this Section. Our zero-shot methodology was similar to the one described above for the RAF-DB database.

The results of this experiment can be found in Table 5.12. Best results have also been obtained when the network was trained with the two coupling losses. It can be observed that this approach outperformed by 5.7% and 8.6% in F1 score and Unweighted Average Recall (UAR), respectively, the state-of-the-art NTechLab's [17] approach, which used the Emotionet's images with compound annotation.

Table 5.12: Performance evaluation of generated compound emotion predictions on EmotioNet and RAF-DB databases.

Databases	EmotioNet		RAF-DB
Methods	F1 Score	Unweighted Average Recall	Mean diagonal of conf. matrix
zero-shot, FaceBehaviorNet, no coupling loss	0.243	0.260	0.342
zero-shot, FaceBehaviorNet, both coupling losses	**0.312**	**0.329**	**0.364**
NTechLab [17]	0.255	0.243	-
VGG-FACE-mSVM [128]	-	-	0.316
DLP-CNN [128]	-	-	0.446
fine-tuned FaceBehaviorNet, no coupling loss	-	-	0.458
fine-tuned FaceBehaviorNet, both coupling losses	-	-	**0.483**

Chapter 6

Affect Synthesis

Rendering photorealistic facial expressions from single static faces while preserving the identity information is an open research topic which has significant impact on the area of affective computing. Generating faces of a specific person with different facial expressions can be used in various applications, including face recognition [24, 161], face verification [188, 191], emotion prediction [104, 108, 110], expression database generation, facial expression augmentation and entertainment.

In this Chapter, we present a novel approach that uses an arbitrary face image with a neutral expression and synthesizes a new face image of the same person, but with a different expression, generated according to a categorical or dimensional emotion representation model. This problem cannot be tackled using small databases with labeled facial expressions, as it would be really difficult to disentangle facial expressions and identity information through them. Our approach is based on analysis of a large 4D facial database, the 4DFAB [29]. A dimensional emotion model, in terms of valence and arousal [215] [176], has been used to annotate a large amount of 600,000 facial images in 4DFAB. I should be mentioned that it is the first time that the dimensional model of affect is used when synthesizing face images. A categorical emotion model, in terms of the six basic facial expressions (Anger, Disgust, Fear, Happiness, Sadness, Surprise), has also been used, according to which 12,000 expressions were generated from the 4DFAB, including 2,000 cases for each of the six basic expressions.

The proposed approach for facial affect synthesis can accept, either a pair of valence-arousal values and synthesise the respective facial affect, or a path of affect in the 2D VA Space and synthesise the

respective temporal facial affect sequence, or a value indicating the desired basic facial expression and synthesise it. A given neutral 2D image of a person is used in all cases to appropriately transfer the synthesised affect. The affect synthesis is implemented by fitting a 3D Morphable Model on the neutral image, then deforming the reconstructed face, adding the inputted affect and blending the new face with given affect into the original image.

Qualitative experiments illustrate the synthesis of realistic images, when the neutral image is sampled from 15 well known lab-controlled and in-the-wild databases; also showing the achieved higher quality when compared to Generative Adversarial Network (GAN) generated facial affect. Then, we use the synthesized facial images for data augmentation and for training Deep Neural Networks over eight databases, annotated with either dimensional or categorical affect labels. We show that improved affect recognition, when compared to state-of-the-art methods, as well as to GAN-based data augmentation is achieved, over all databases.

6.1 Related Work

Facial expression transfer is a research field for mapping and generating desired images of specified subject and facial expression. Many methods achieved significant results for high-resolution images and are applied to a wide range of applications, such as facial animation, facial editing, and facial expression recognition.

There are mainly two categories of methods for facial expression transfer from a single image: traditional graphic-based methods and emerging generative methods. In the first case, some methods directly warp the input face to create the targeted expression, by either 2D warps [66, 68], or 3D warps [18, 23, 134]. Other methods construct parametric global models. In [150], a probabilistic model is learned, in which existing and generated images obey structural constraints. [8] added fine-scale dynamic details, such as wrinkles and inner mouth, that are associated with facial expressions. Although these methods have achieved some positive results in high-resolution and one-to-many image synthesis, they are still limited due to their sophisticated design and expensive computation.

In [194], the authors developed a real-time face-to-face expression transfer system, with an extra

blending step for mouth. This 2D-to-3D approach shows promising results, but due to the nature of its formulation, it is unable to retrieve fine-details, and its applicability is limited to expressions lying in a linear shape subspace with known rank. The authors extended this system to human portrait video transfer [195]. They captured facial expressions, eye gaze, rigid head pose, and motions of the upper body of a source actor and transferred them to a target actor in real time.

The second category of methods is based on data-driven generative models. At the beginning, some generative models, such as deep belief nets (DBN) [189] and higher-order Boltzmann machines [171], had been applied to facial expression synthesis. However, these models faced problems such as blurry generated images, incapability of fine control of facial expression and low-resolution outputs.

With the recent development of Generative Adversarial Networks (GANs) [75], these networks have been applied to facial expression transfer; due to the fact that the generated images are of high-quality, these provided positive results. A generative model is trained according to a dataset, including all information about identity, expression, viewing angle, etc, while performing facial expression transfer. Generative modeling methods reduce the complicated design of the connection between facial textures and emotional states and encode intuitionistic facial features into parameters of data distribution. However, the main drawback of GANs is the training instability and the trade-off between visual quality and image diversity.

Since the original GAN could not generate facial images with a specific facial expression referring to a specific person, some methods conditioned on expression categories have been proposed. Conditional GANs (cGANs) [149] (and conditional variational autoencoders (cVAEs) [184]) can generate samples conditioned on attribute information, when this is available. Those networks require large training databases so that identity information can be properly disambiguated. Otherwise, when presented with an unseen face, the networks tend to generate faces which look like the “closest” subject in the training datasets. During training, those networks require the knowledge of the attribute labels; it is not clear how to adapt them to new attributes without retraining from scratch. Finally, these networks suffer from mode-collapse (e.g., the generator only outputs samples from a single mode, or with extremely low variety) and blurriness.

The conditional difference adversarial autoencoder (CDAAE) [242] aims at synthesizing specific ex-

pressions for unseen persons with a targeted emotion or facial action unit label. However, such GAN-based methods are still limited to discrete facial expression synthesis, i.e., they cannot generate a face sequence showing a smooth transition from an emotion to another. [55] proposed an Expression Generative Adversarial Network (ExprGAN) in which the expression intensity could be controlled in a continuous manner from weak to strong. The identity and expression representation learning were disentangled and there was no rigid requirement of paired samples for training. The authors developed a three-stage incremental learning algorithm to train the model on small datasets.

In [165], the authors proposed a weakly supervised adversarial learning framework for automatic facial expression synthesis based on continuous action unit coefficients. In [167], the GANimation was proposed that additionally controlled the generated expression by AU labels, and allowed a continuous expression transformation. The authors introduced an attention-based generator to promote the robustness of their model for distracting backgrounds and illuminations.

There are some differences between continuous expression synthesis based on AUs and VA. Firstly, AUs are related to some facial muscles, with only a small number of them being mapped to facial expression modelling. On the contrary, the VA model covers the whole spectrum of emotions. Moreover, mapping AUs to emotions is not straightforward (different psychological studies provide different results). GANimation is solely based on automatic annotation of AUs, whilst the proposed methodology is based on manual, i.e., more robust and trusted, VA annotation of the 4DFAB database. Finally, it can be mentioned that annotation of AUs needs experienced FACS coders; especially in in-the-wild datasets. That is why, there exists only one in-the-wild database annotated for AUs (existence and not intensity information), the EmotioNet, which only contains 50,000 annotations, in terms of 12 AUs.

Recently, [186] utilized landmarks and proposed the geometry-guided GAN (G2GAN) to generate smooth image sequences of facial expressions. G2GAN uses geometry information based on dual adversarial networks to express face changes and synthesizes facial images. Through manipulating landmarks, smoothly changed images can also be generated. However, this method demands a neutral face of the targeted person as the intermediate of facial expression transfer. Although the expression removal network could generate a neutral expression of a specific person, this procedure brings additional artifacts and degrades the performance of expression transition.

[168] used geometry (facial landmarks) to control the expression synthesis with a facial geometry embedding network and proposed a Geometry-Contrastive Generative Adversarial Network (GC-GAN) to transfer continuous emotions across different subjects, even if there existed big difference in shapes. [217] proposed a boundary latent space and boundary transformer. They mapped the source face into the boundary latent space, and transformed the source face's boundary to the target's boundary, which was the medium to capture facial geometric variances during expression transfer.

In [137], an unpaired learning framework was developed to learn the mapping between any two facial expressions in the facial blendshape space. This framework automatically transforms the source expression in an input video clip to a specified target expression. This work lacks the capability to generate personalized expressions; individual-specific expression characteristics, such as wrinkles and creases, are ignored. Also, the transitions between different expressions are not taken into consideration. Finally, this work is limited in the sense that it cannot produce highly exaggerated expressions.

Both the graphic-based methods and the genererative methods of facial expression transfer have been used to create synthetic data that are used as auxiliary data in network training, augmenting the training dataset. A synthetic data generation system with a 3D convolutional neural network (CNN) was created in [1] to confidentially create faces with different levels of saturation in expression. [5] proposed the Data Augmentation Generative Adversarial Network (DAGAN) which is based on cGAN and tested its effectiveness on vanilla classifiers and one shot learning. DAGAN is a basic framework for data augmentation based on cGAN.

[244] presented another basic framework for face data augmentation based on CycleGAN [243]. Similar to cGAN, CycleGAN is also an general-purpose solution for image-to-image translation, but it learns a dual mapping between two domains simultaneously with no need for paired training examples, because it combines a cycle consistency loss with adversarial loss. The authors used this framework to generate auxiliary data for imbalanced datasets, where the data class with fewer samples was selected as transfer target and the data class with more samples was the reference.

6.2 Materials & Methods

In the following, we first describe the 4DFAB database, its annotation in terms of valence-arousal and the selection of expressive categorical sequences from it. The annotated 4DFAB database has been used for constructing the 3D facial expression gallery that is the basis of our affect synthesis pipeline described in the next Section. Then we describe the methods we have used: a) for registering and correlating all components of the 3D gallery into a universal coordinate frame; b) for constructing the 3D Morphable Model used in this work.

The 4DFAB Database

The 4DFAB database [29] is the first large scale 4D face database designed for biometric applications and facial expression analysis. It consists of 180 subjects (60 females, 120 males) aging from 5 to 75 years. 4DFAB was collected over a period of 5 years under four different sessions, with over 1,800,000 3D faces. The database was designed to capture articulated facial actions and spontaneous facial behaviors, the latter being elicited by watching emotional video clips. In each of the four sessions, different video clips were shown that stimulated different spontaneous behaviors. We used all 1,580 spontaneous expression sequences (video clips) for dimensional emotion analysis and synthesis. The frame rate of 4DFAB database is 60 FPS and the average clip length for spontaneous expression sequences is 380 frames. Consequently the 1,580 expression sequences correspond to 600,000 frames, which we annotated in terms of valence and arousal (details follow in the next subsection). These sequences cover a wide range of expressions as shown in Figs. 6.2 and 6.3.

Moreover, to be able to develop the categorical emotion synthesis model, we used the 2,000 expressive 3D meshes per basic expression (12,000 meshes in total) that were provided along with 4DFAB. Those 3D meshes corresponded to (annotated) apex frames of posed expression sequences in 4DFAB. Such examples are shown in Fig.6.1.

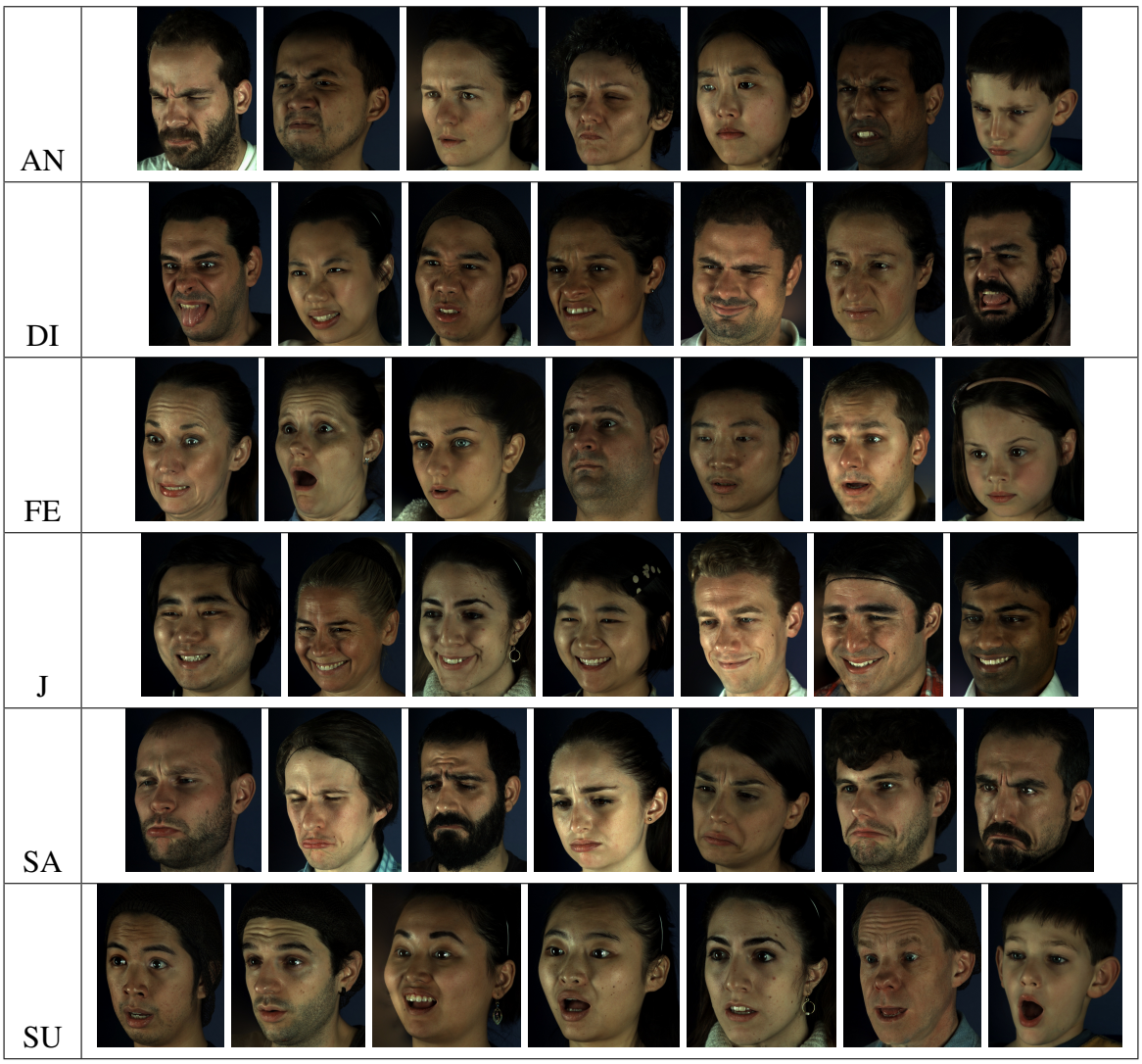


Figure 6.1: Examples from the 4DFAB of apex frames with posed expressions for the six basic expressions: Anger (AN), Disgust (DI), Fear (FE), Joy (J), Sadness (SA), Surprise (SU)

4DFAB Dimensional Annotation

Targeting to develop the novel dimensional expression synthesis method, all 1,580 dynamic 3D sequences (i.e., over 600,000 frames) of 4DFAB have been annotated in terms of valence and arousal emotion dimensions. In total, three experts were chosen to perform the annotation task. Each expert performed a time-continuous annotation for both affective dimensions. The application-tool described in [226] was used in the annotation process.

Each expert logged into the application-annotation tool using an identifier (e.g. his/her name) and selected an appropriate joystick; then the application showed a scrolling list of all videos and the expert selected a video to annotate; then a screen appeared that showed the selected video and a

slider of valence or arousal values ranging in $[-1, 1]$; the expert annotated the video by moving the joystick either up or down; finally, a file was created with the annotations. The mean inter-annotation correlation per annotator was 0.66, 0.70, 0.68 for valence and 0.59, 0.62, 0.59 for arousal. The average of those mean inter-annotation correlations was 0.68 for valence and 0.60 for arousal. Those values are high, indicating a very good agreement between annotators. As a consequence, the final label values were chosen to be the mean of those three annotations.

Examples of frames from the 4DFAB along with their annotations, are shown in Fig. 6.2. Fig. 6.3 shows the 2D histogram of annotations of 4DFAB. In the following, we refer to the 4DFAB database either as: i) the 600,000 frames with their corresponding 3D meshes, which have been annotated with 2D valence and arousal (VA) emotion values, or ii) the 12,000 apex frames of posed expressions with their corresponding 3D meshes, which have categorical annotation.

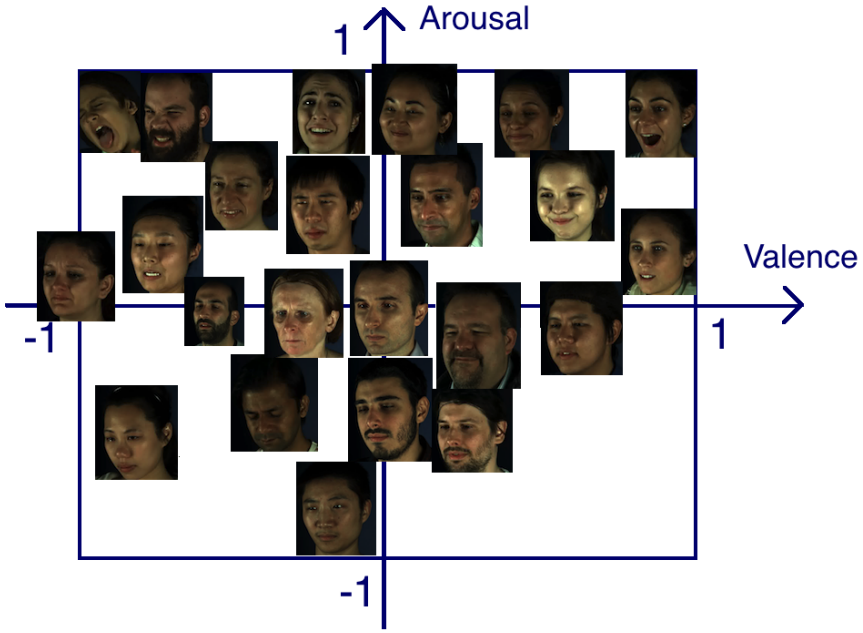


Figure 6.2: The 2D Valence-Arousal Space and some representative frames of 4DFAB

Mesh Pre-Processing: Establishing Dense Correspondence

Each 3D mesh is first re-parameterised into a consistent form where the number of vertices, the triangulation and the anatomical meaning of each vertex are made consistent across all meshes. For example, if the vertex with index i in one mesh corresponds to the nose tip, it is required that the

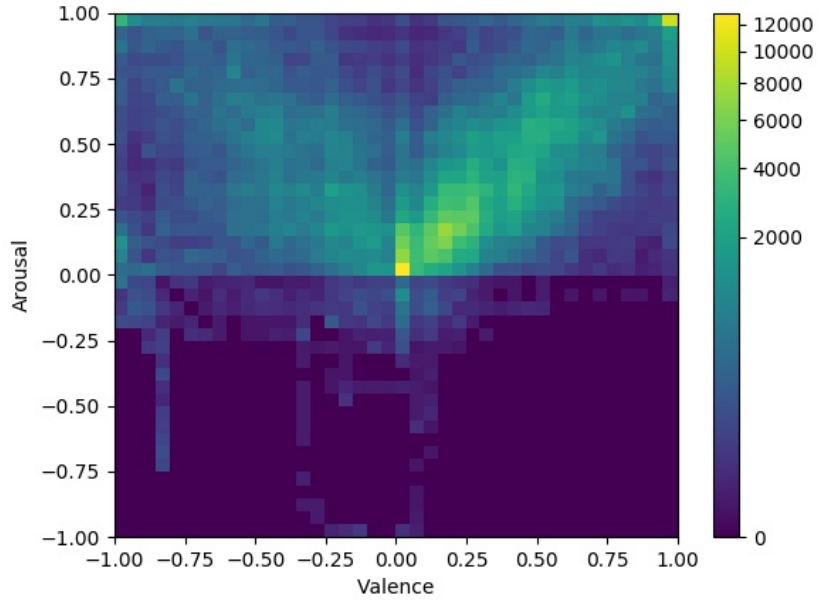


Figure 6.3: The 2D histogram of annotations of 4DFAB

vertex with the same index in every mesh corresponds to the nose tip too. Meshes satisfying the above properties are said to be in dense correspondence with one another. So, correlating all these meshes with a universal coordinate frame (viz. a 3D face template) is a step to follow so as to establish dense correspondence.

In order to do so, we need to define a 2D *UV* space for each mesh, which in fact is a contiguous flattened atlas that embeds the 3D facial surface. Such a UV space is associated with its corresponding 3D surface through a bijective mapping; thus, establishing dense correspondence between two UV images implicitly establishes a 3D-to-3D correspondence for the mapped mesh. UV mapping is the 3D modelling process of projecting a 2D image to a 3D model's surface for texture mapping. The letters U and V denote the axes of the 2D texture, since X, Y and Z are already taken to denote the axes of the 3D object in model space.

We employ an optimal cylindrical projection method [21] to synthetically create a UV space for each mesh. A UV map (which is an image I), with each pixel encoding both spatial information (X, Y, Z) and texture information (R, G, B), is produced, on which we perform non-rigid alignment. Non-rigid alignment is performed through the UV-TPS method that utilises key landmarks fitting and Thin Plate Spline (TPS) warping [35]. Following [29], we perform several modifications to [35], to suit

our data. Firstly, we build session-and-person-specific Active Appearance Models (AAMs) [144] to automatically track feature points in the UV sequences. This means that 4 different AAMs are built and used separately for one subject. Main reasons behind this are: i) textures of different sessions differ due to several facts (i.e. aging, beards, make-ups, experiment lighting condition), ii) person-specific model is proven more accurate and robust in specific domains [30].

In total, 435 neutral meshes and 1047 expression meshes (1 neutral and 2-3 expressive meshes per person and session) in 4DFAB were selected; these contained annotations with 79 3D landmarks. They were unwrapped and rasterised to UV space, then grouped for building the corresponding AAMs. Each UV map was flipped to increase fitting robustness. Once all the UV sequences were tracked with 79 landmarks, they were then warped to the corresponding reference frame using TPS, thus achieving the 3D dense correspondence. For each subject and session, one specific reference coordinate frame from his/her neutral UV map was built. From each warped frame, we could uniformly sample the texture and 3D coordinates. Eventually, a set of non-rigidly corresponded 3D meshes under the same topology and density were obtained.

Given that meshes have been aligned to their designated reference frame, the last step was to establish dense 3D-to-3D correspondences between those reference frames and a 3D template face. This is a 3D mesh registration problem, solved by Non-rigid ICP [4]. We employed it to register the neutral reference meshes to a common template, the Large Scale Facial Model (LSFM) [20]. We brought all 600,000 3D meshes into full correspondence with the mean face of LSFM. As a result, we created a new set of 600,000 3D faces that share identical mesh topology, while maintaining their original facial expressions. In the following, this set constitutes the 3D facial expression gallery which we use for facial affect synthesis.

Constructing a 3D Morphable Model

General Pipeline A common 3DMM consists of three parametric models: the shape, the camera and the texture models.

To build the shape model, the training 3D meshes should be put in dense correspondence (similarly to

the previous Mesh Pre-Processing subsection). Next, Generalized Procrustes Analysis is performed to remove any similarity effects, leaving only shape information. Finally, Principal Component Analysis (PCA) is applied to these meshes, which generates a 3D deformable model as a linear basis of shapes. This model allows for the generation of novel shape instances. The model can be expressed as:

$$\mathcal{S}(\mathbf{p}) = \bar{\mathbf{s}} + \mathbf{U}_s\mathbf{p} \tag{6.1}$$

where $\bar{\mathbf{s}} \in \mathbb{R}^{3N}$ is the mean component of 3D shape (in our case it is the mean of shape models from the LSFM model described in the next subsection) with N denoting the number of vertices in the shape model; $\mathbf{U}_s \in \mathbb{R}^{3N\times n_s}$ is the shape eigenbase (in our case it is the identity subspace of LSFM) with $n_s << 3N$ being the number of principal components (n_s is chosen to explain a percentage of the training set variance; generally, this percentage is 99.5%); and $\mathbf{p} \in \mathbb{R}^{n_s}$ is a vector of parameters which allows for the generation of novel shape instances.

The purpose of camera model is to project the object-centered Cartesian coordinates of a 3D mesh instance into 2D Cartesian coordinates in an image plane. At first, given that the camera is static, the 3D mesh is rotated and translated using a linear view transformation, which results in 3D rotation and translation components. Then, a nonlinear perspective transformation is applied. Note that quaternions [122, 213] are used to parametrise the 3D rotation, which ensures computational efficiency, robustness and simpler differentiation. In this manner we construct the camera parameters (i.e., 3D translation components, quaternions and parameter of linear perspective transformation). The camera model of the 3DMM applies the above transformations on the 3D shape instances generated by the shape model. Finally, the camera model can be written as:

$$\mathcal{W}(\mathbf{p},\mathbf{c}) = \mathcal{P}(\mathcal{S}(\mathbf{p}),\mathbf{c}), \tag{6.2}$$

where $\mathcal{S}(\mathbf{p})$ is a 3D face instance; $\mathbf{c} \in \mathbb{R}^{n_c}$ are the camera parameters (for rotation, translation and focal length; n_c is 7); and $\mathcal{P} : \mathbb{R}^{3N} \to \mathbb{R}^{2N}$ is the perspective camera projection.

For the texture model, large facial "in-the-wild" data-bases annotated for sparse landmarks are needed. Let us assume that the meshes have corresponding camera and shape parameters. These images are

passed through a dense feature extraction function that returns feature-based representations for each image. These are then sampled from the camera model at each vertex location so as to build a texture sample, which will be nonsensical for some regions mainly due to self occlusions present in the mesh projected in the image space. To complete the missing information of the texture samples, Robust PCA (RPCA) with missing values [181] is applied. This produces complete feature-based textures that can be processed with PCA to create the statistical model of texture, which can be written as:

$$\mathcal{T}(\boldsymbol{\lambda}) = \bar{\mathbf{t}} + \mathbf{U_t}\boldsymbol{\lambda}, \tag{6.3}$$

where $\bar{\mathbf{t}} \in \mathbb{R}^{3N}$ is the mean texture component (in our case it is the mean of texture model from LSFM); $\mathbf{U}_t \in \mathbb{R}^{3N \times n_t}$ and $\boldsymbol{\lambda} \in \mathbb{R}^{n_t}$ are the texture subspace (eigenbase) and texture parameters, respectively, with $n_t << 3N$ being the number of principal components. This model can be used to generate novel 3D feature-based texture instances.

The Large Scale Facial Model (LSFM) We have adopted the LSFM model constructed using the MeIn3D dataset [20]. The construction pipeline of LSFM starts with a robust approach to 3D landmark localization resulting in generating 3D landmarks for the meshes. The 3D landmarks are then employed as soft constraints in Non-rigid ICP to place all meshes in correspondence with a template facial surface; the mean face of the Basel Face Model [162] has been chosen. However, the large cohort of data could result in convergence failures. These are an unavoidable byproduct of the fact that both landmark localization and NICP are non-convex optimization problems sensitive to initialization.

A refinement post-processing step weeds out problematic subjects automatically, guaranteeing that the LSFM models are only constructed from training data for which there exist a high confidence of successful processing. Finally, the LSFM models are derived by applying PCA on the corresponding training sets, after excluding the shape vectors that have been classified as outliers. In total, 9,663 subjects are used to build LSFM, which covers a wide variety of age (from 5 to over 80s), gender (48% male, 52% female), and ethnicity (82% White, 9% Asian, 5% Mixed Heritage, 3% Black and 1% other).

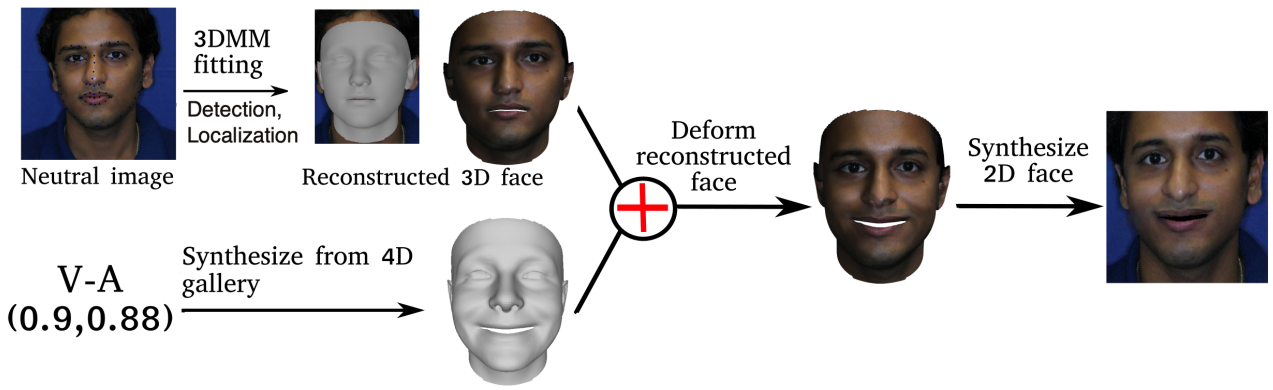


Figure 6.4: The facial affect synthesis framework: the user inputs an arbitrary 2D neutral face and the affect to be synthesized (a pair of valence-arousal values in this case)

6.3 The Proposed Approach

In this Section, we present the fully automatic facial affect synthesis framework. The user needs to provide a neutral image and an affect, which can be a VA pair of values, a path in the 2D VA space, or one of the basic expression categories. Our approach: 1) performs face detection and landmark localization on the input neutral image, 2) fits a 3D Morphable Model (3DMM) on the resulting image [19], 3) deforms the reconstructed face and adds the input affect, and 4) blends the new face with the given affect into the original image. Here let us note that the total time needed for the first two steps is about 400ms; this has to be performed only once, if generating multiple images from the same input image. Specific details regarding the described steps of our approach follow. This procedure is shown in Fig. 6.4.

Face Detection & Landmark Localization

The first step to edit an image is to locate landmark points that will be used for fitting the 3DMM. We first perform face detection with the face detection model from [230] and then utilize [46] to localize 68 2D facial landmark points which are aware of the 3D structure of the face, in the sense that points on occluded parts of the face (most commonly part of the jawline) are correctly localized.

3DMM-Fitting: Cost Function & Optimization

The goal of this step is to retrieve a reconstructed 3D face with the texture sampled from the original image. In order to do so, we first need a 3DMM; we select the LSFM.

Fitting a 3DMM on face images is an inverse graphics approach to 3D reconstruction and consists of optimizing three parametric models of the 3DMM, the *shape*, *texture* and *camera* models. The optimization aims at rendering a 2D image which is as close as as possible to the input one. In our pipeline we follow the 3DMM fitting approach of [19]. As is already noted, we employ the LSFM [20] $\mathcal{S}(\mathbf{p})$ to model the identity deformation of faces. Moreover, we adopt the robust, feature-based texture model $\mathcal{T}(\boldsymbol{\lambda})$ of [19], built from in-the-wild images. The employed camera model is a perspective transformation $\mathcal{W}(\mathbf{p},\mathbf{c})$, which projects shape $\mathcal{S}(\mathbf{p})$ on the image plane.

Consequently, the objective function that we optimize can be formulated as:

$$\underset{\mathbf{p},\boldsymbol{\lambda},\mathbf{c}}{\operatorname{argmin}} \|\mathbf{F}(\mathcal{W}(\mathbf{p},\mathbf{c})) - \mathcal{T}(\boldsymbol{\lambda})\|^2 + c_l\|\mathcal{W}_l(\mathbf{p},\mathbf{c}) - \mathbf{s}_l\|^2 + \; c_s\|\mathbf{p}\|^2_{\Sigma_s^{-1}} + c_t\|\boldsymbol{\lambda}\|^2_{\Sigma_t^{-1}}, \tag{6.4}$$

where the first term denotes the pixel loss between the feature based image $\mathbf{F}$ sampled at the projected shape's locations and the model generated texture; the second term denotes a sparse landmark loss between the image 2D landmarks and the corresponding 2D projected 3D points, where the 2D shape, $\mathbf{s}_l$, is provided by [46]; the rest two terms are regularization terms which serve as counter over-fitting mechanism, where $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Sigma}_t$ are diagonal matrices with the main diagonal being eigenvalues of the shape and texture models respectively; c_l, c_s and c_t are weights used to regularize the importance of each term during optimization and were empirically set to 10^5, 3×10^6 and 1, respectively, following [19]. Note also, that the 2D landmarks term is useful as it drives the optimization to converge faster. Eq. 6.4 is solved by the Project-Out variation of Gauss-Newton optimization as formulated in [19].

From the optimized models, the optimal shape instance constitutes the neutral 3D representation of the input face. Moreover, by utilizing the optimal shape and camera models, we are able to sample the input image at the projected locations of the recovered mesh and extract a UV texture, that we

later use for rendering.

Deforming Face & Adding Affect

Given an affect and an arbitrary 2D image $\mathbf{I}$, we first fit the LSFM to this image using the aforementioned 3DMM fitting method. After that, we can retrieve a reconstructed 3D face $\mathbf{s}_{orig}$ with the texture sampled from the original image (texture sampling is simply extracting image pixel value for each projected 3D vertex in image plane). Let us assume that we have created an affect synthesis model $\mathbf{M}_{Aff}$ that takes the affect as input and generates a new expressive face (denoted as $\mathbf{s}_{gen}$), i.e., $\mathbf{s} = \mathbf{M}_{Aff}(affect)$ (specific details regarding the generation of the expressive face, can be found in subsection 6.3). Next, we calculate the facial deformation $\Delta\mathbf{s}$ by subtracting the synthesized face $\mathbf{s}_{gen}$ from the LSFM template $\bar{\mathbf{s}}$, i.e., $\Delta\mathbf{s} = \mathbf{s}_{gen} - \bar{\mathbf{s}}$, and impose this deformation on the reconstructed mesh, i.e., $\mathbf{s}_{new} = \mathbf{s}_{orig} + \Delta\mathbf{s}$. Therefore, we obtain a 3D face (dubbed $\mathbf{s}_{new}$) with facial affect.

Synthesizing 2D Face

The final step in our pipeline is to render the new 3D face $\mathbf{s}_{new}$ back to the original 2D image. To do that we employ the mesh that we have deformed according to the given affect, the extracted UV texture and the optimal camera transformation of the 3DMM. For rendering, we pass the three model instances to a renderer and we use as background the background of the input image. Lastly, the rendered image is fused with the original image via poisson blending [164] to smooth the boundary between foreground face and image background so as to produce a natural and realistic result. In our experiments, we used both a CPU-based renderer [2] and a GPU-based renderer [69]. The GPU-based renderer greatly decreases the rendering time, as it needs 20ms to render a single image, while the CPU-based renderer needs 400ms.

Synthesising Expressive Faces with Given Affect

VA & Basic Expression cases: Building Blendshape Models & Computing Mean Faces Let us first describe the VA case. We have 600,000 3D meshes (established into dense correspondence) and

their VA annotations. We want to appropriately discretize the VA Space into classes, so that each class contains a sufficient number of data. This is due to the fact that if classes contain only few examples, it is more likely to include identity information. However, the synthesized facial affect should only describe the expression associated with the VA pair of values, rather than information for the person's identity, gender, or age. We have chosen to perform agglomerative clustering [139] on the VA values, using the euclidean distance as metric and the ward as linkage criterion (keeping the correspondence between VA values and 3D meshes). In this manner, we created 550 clusters, i.e., classes. Then we built blendshape models and computed the mean face per class. Fig. 6.5 illustrates the mean faces of various classes. It should be mentioned that the majority of classes correspond to the first two quadrants of the VA Space, namely the regions of positive valence (as can be seen in the 2D histogram of Fig. 6.3).

Figure 6.5: Some mean faces of the 550 classes in the VA Space

As far as the basic expression case is concerned, based on the derived 12,000 3D meshes, 2,000 for each of the six basic expressions, we built six blendshape models and six corresponding mean faces.

User Selection: VA/Basic Expr & Static/Temporal Synthesis The user first chooses the type of affect that our approach will generate. The affect could be either a point, or a path in the VA space, or one the six basic expression categories. If the user chooses the latter, then we retrieve the mean face of this category and add it on the 3D face reconstructed from the user's input neutral image. In this case, the only difference in Fig. 6.4 would be for the user to input a basic expression, the happy one, instead of a VA pair of values. If the user chooses the former, then (s)he needs to additionally clarify if our approach should generate an image ('static synthesis') or a sequence of images ('temporal synthesis') with this affect.

Static synthesis If the user selects 'static synthesis', then the user should input a specific VA pair of values. Then, we retrieve the mean face of the class to which this VA value belongs. We use this mean face as the affect to be added on the 3D face reconstructed from the provided neutral image. Fig. 6.4 shows the proposed approach for this specific case. Fig. 6.6 illustrates the procedure described in 6.3 given that the 550 VA classes are already created.

Temporal synthesis If the user selects 'temporal synthesis', then, (s)he should provide a path in the VA space (for instance by drawing) that the synthesized sequence should follow. Then, we retrieve the mean faces of the classes to which the VA values of the path belong. We use each of these mean faces as the affect to be added on the 3D faces reconstructed from the provided neutral image. As a consequence, an expressive sequence is generated that shows the evolution of affect on the VA path specified by the user.

Here let us mention that the fact that the 4DFAB used in our approach is a temporal database, ensures that successive video frames' annotations are adjacent in the VA Space, since they generally show the same or slightly different states of affect. Thus, the 3D meshes of successive video frames will lie in the same and in adjacent classes in the 2-D VA space. Thus mean faces from adjacent classes can be used to show temporal evolution of affect as was above described.

Expression Blendshape Models Expression blendshape models provide an effective way to parameterize facial behaviors. The localized blendshape model [153] has been used to describe the selected VA samples. To build this model, we first bring all meshes into full correspondence following the

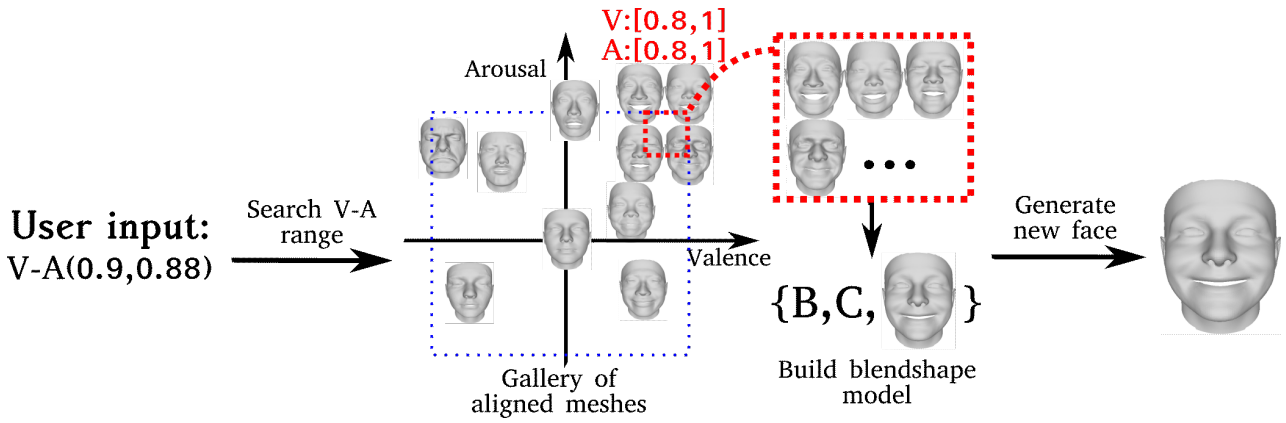


Figure 6.6: Generation of new facial affect from the 4D face gallery; the user provides a target VA pair

dense registration approach described in Section 6.2. As a result, we have a set of training meshes with the same number of vertices and identical topology. Note that we have also selected one neutral mesh for each subject, which should have full correspondence with the rest data. Next, we subtract each 3D mesh from the respective neutral mesh, and create a set of m difference vectors $\mathbf{d}_i \in \mathbb{R}^{3N}$. We then stack them into a matrix $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_m] \in \mathbb{R}^{3N \times m}$, where N is number of vertices in the mesh. Finally, a variant of sparse Principal Component Analysis (PCA) is applied to the data matrix $\mathbf{D}$, so as to identify sparse deformation components $\mathbf{C} \in \mathbb{R}^{h \times 1}$:

$$\arg\min \|\mathbf{D} - \mathbf{B}\mathbf{C}\|_F^2 + \Omega(\mathbf{C}) \quad \text{s.t. } \mathcal{V}(\mathbf{B}), \tag{6.5}$$

where, the constraint $\mathcal{V}$ can be either $\max(|\mathbf{B}_k|) = 1, \ \forall k$ or $\max(\mathbf{B}_k) = 1, \ \mathbf{B} \geq 1, \ \forall k$, with $\mathbf{B}_k \in \mathbb{R}^{3N \times 1}$ denoting the k^{th} components of sparse weight matrix $\mathbf{B} = [\mathbf{B}_1, \cdots, \mathbf{B}_h]$. Selection of these two constraints depends on the actual usage; the major difference is that the latter one allows for negative weights and therefore enables deformation towards both directions, which is useful for describing shapes like muscle bulges. In this paper, we have selected the latter constraint, as we wish to enable bidirectional muscle movement and synthesise a rich variety of expressions. The regularization of sparse components $\mathbf{C}$ was performed with $\ell 1/\ell 2$ norm [10, 216]. To permit more local deformations, additional regularization parameters were added into $\Omega(\mathbf{C})$. To compute optimal $\mathbf{C}$ and $\mathbf{B}$, an iterative alternating optimization was employed (please refer to [153] for more details).

6.4 Databases

To evaluate our facial affect synthesis method in different scenarios (e.g. controlled laboratory environment, uncontrolled in-the-wild setting), we utilized neutral facial images from as many as 15 databases (both small and large in terms of size). Table 6.1 briefly presents the Multi-PIE [78], Aff-Wild [109, 226], AFEW 5.0 [47], AFEW-VA [118], BU-3DFE [223], RECOLA [173], AffectNet [151], RAF-DB [128], KF-ITW [19], Face place, FEI [196], 2D Face Sets and Bosphorus [179] databases that we used in our experimental study. Let us note that for AffectNet no test set is released and thus we use the released validation set to test on and randomly divide the training set into a training and a validation subset (with a 85/15 split).

Table 6.1 presents these databases by showing: i) the model of affect they use, their condition, their type (static images or audiovisual image sequences), the total number of frames and (male/female) subjects that they contain and the range of ages of the subjects, and ii) the total number of images that we synthesized using our approach (both in the valence-arousal and the six basic expressions cases).

6.5 Qualitative evaluation of achieved facial affect synthesis

This section provides a qualitative evaluation of the proposed approach by showing many synthesized images or image sequences from all fifteen databases described in the previous Section; as well as by comparing images generated by state-of-the-art GANs (StarGAN, GANimation) and the proposed approach [54, 103].

We used all databases mentioned in Section 6.4 to supply the proposed approach with 'input' neutral faces. We then synthesized the emotional state corresponding to specific affects (both in VA case and in the six basic expressions one) for these images. At first we show many generated images (static synthesis) according to different VA values, then we illustrate examples of generated image sequences (temporal synthesis) and next we present some synthesized (static) images according to the six basic expressions. Finally, we visually compare images generated by our approach with synthesized images by StarGAN and GANimation.

Table 6.1: Databases used in our approach, along with their properties and the number of synthesized images in the valence-arousal case and the six basic expressions one; 'static' means images, 'A/V' means audiovisual sequences, i.e., videos

Databases (DBs)	DB Type	Model of Affect	Condition	DB Size	# of Subjects	Age Range	Total # of Synthesized Images	
							VA	Basic Expr
MULTI-PIE [78]	static	Neutral, Surprise, Disgust, Smile + Squint, Scream	controlled	755,370	337 Male: 235 Female: 102	-	52,254	5,520
Kinect Fusion ITW [19]	static	Neutral, Happiness , Surprise	in-the-wild	3,264	17	-	116,235	12,236
FEI [196]	static	Neutral, Smile	controlled	2,800	200 Male: 100 Female: 100	19-40	11,400	1,200
Face place[1]	static	6 Basic Expr, Neutral, Confusion	controlled	6,574	235 Male: 143 Female: 92	-	59,736	6,288
AFEW 5.0 [47]	A/V	6 Basic Expr, Neutral	in-the-wild	41,406	>330	1-77	705,649	56,514
RECOLA [173]	A/V	VA	controlled	345,000	46 Male: 19 Female: 27	-	46,455	4,890
BU-3DFE [223]	static	6 Basic Expr, Neutral	controlled	2,500	100 Male: 56 Female: 44	18-70	5,700	600
Bosphorus [179]	static	6 Basic Expr	controlled	4,666	105 Male: 60 Female: 45	25-35	17,018	1,792
AffectNet [151]	static	VA + 6 Basic Expr, Neutral + Contempt	in-the-wild	450,000 manually annotated	-	0 to >50	2,476,235	176,425
Aff-Wild [109] [226]	A/V	VA	in-the-wild	1,224,094	200 Male: 130 Female: 70	-	60,135	6,330
AFEW-VA [118]	A/V	VA	in-the-wild	30,050	<600	-	108,864	11,460
RAF-DB [128]	static	6 Basic, Neutral + 11 Compound Expr	in-the-wild	15,339 + 3,954	-	0-70	121,866	12,828
2D Face Sets[2]: Pain	static	6 Basic, Neutral + 10 Pain Expr	controlled	599	23 Male: 13 Female: 10	-	2,736	288
2D Face Sets: Iranian	static	Neutral, Smile	controlled	369	34 Male: 0 Female: 34	-	2,679	282
2D Face Sets: Nottingham Scans	static	Neutral	controlled	100	100 Male: 50 Female: 50	-	5,700	600

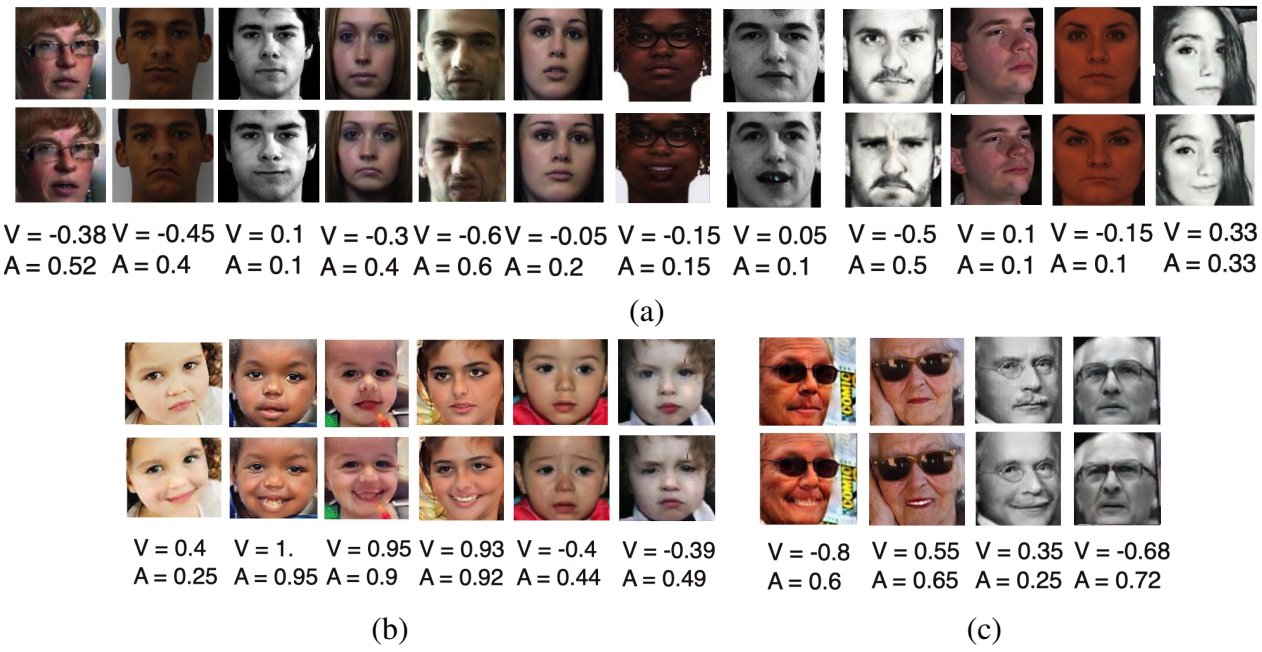


Figure 6.7: **(a)-(c).** VA Case of static (facial) synthesis across all databases; first rows show the neutral, second ones show the corresponding synthesized images and third rows show the corresponding VA values. Images of: (b) kids, (c) elderly people and (a) in-between ages, are shown.

6.5.1 Results on Static & Temporal Affect Synthesis

Fig. 6.7 shows representative results of facial affect synthesis, when user inputs a VA pair and selects to generate a static image. These results are organized in three age groups: Fig. 6.7(b) kids, Fig. 6.7(c) elderly people and Fig. 6.7(a) in-between ages. In each part, the first row illustrates neutral images sampled from each of the aforementioned databases, the second one shows the respective synthesized images and the third shows the respective VA values that were synthesized. Moreover, Fig. 6.8 shows neutral images on the left hand side (first column) and synthesized images, with various valence and arousal values, on the right hand side (following columns). It can be observed that the synthesized images are identity preserving, realistic and vivid. Fig. 6.9 refers to the basic expression case; it shows neutral images on the left hand side of (a) and (b) and synthesized images with basic expressions on the right hand side. Fig. 6.10 illustrates the VA case for temporal synthesis, as was described in Section 4.5.2. Neutral images are shown on the left hand side, while synthesized face sequences with time-varying levels of affect are shown on the right hand side.

All these Figs. show that the proposed framework works well, when using images from either in-the-wild, or controlled databases. This indicates that we can effectively synthesize facial affect irregard-

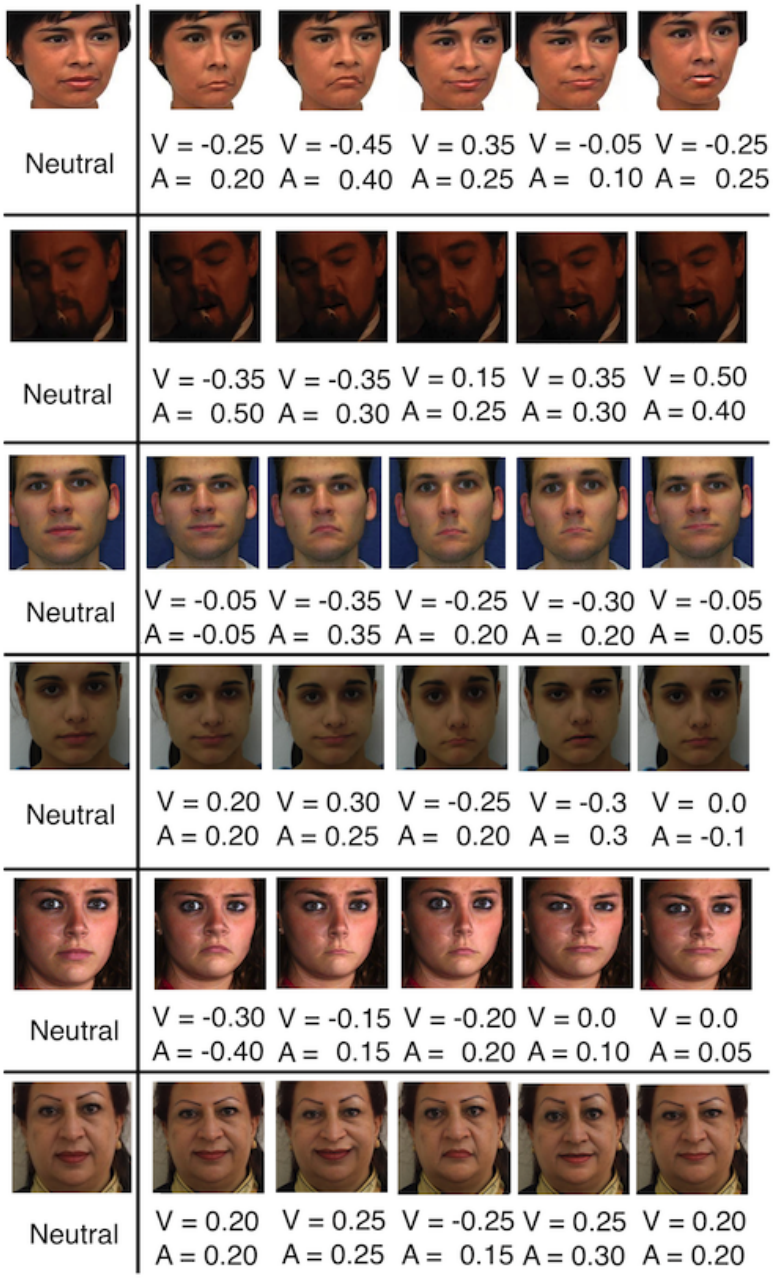
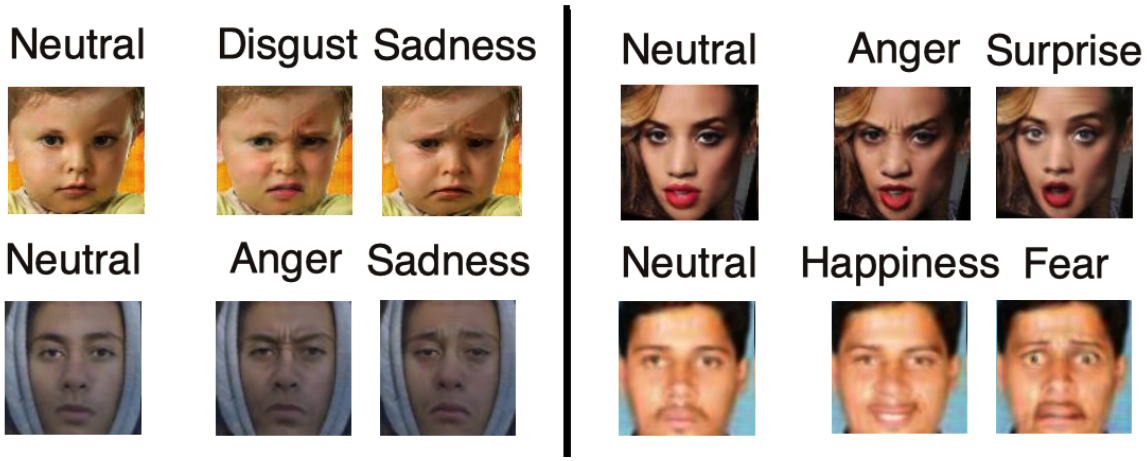


Figure 6.8: VA case of facial synthesis: on the left hand side are the neutral 2D images and on the right the synthesized images with different levels of affect



(a) (b)

Figure 6.9: Basic Expression Case of facial synthesis: on the left hand side of (a) and (b) are the neutral 2D images and on the right the synthesized images with some basic expressions

less of image conditions (e.g., occlusions, illumination and head poses).

6.5.2 Comparison with GANs

In order to characterize the value that the proposed approach imparts, we provide qualitative comparisons with two state-of-the-art GANs that have been widely used for affect generation, namely StarGAN [31] and GANimation. Like CycleGAN, Star-GAN performs image-to-image translation, but adopts a unified approach such that a single generator is trained to map an input image to one of multiple target domains, selected by the user. By sharing the generator weights among different domains, a dramatic reduction of the number of parameters is achieved.

At first, it should be mentioned that, the original StarGAN synthesized images according to the basic expressions (apart from facial attributes) and the GANimation synthesized images according to AUs. However, in psychology, there does not exist any mapping between AUs - VA and no consistent mapping (across studies) between AUs - expressions, or VA - expressions. In order to achieve a fair comparison of our method with these networks, we applied them - for the first time - to the VA Space [117]; we trained them with the same 600,000 frames of 4DFAB that we used in our approach. In both networks, pre-processing was conducted, which included face detection and alignment. For a fair comparison, in all presented results (both qualitative and quantitative), the GANs were provided with the same neutral images and the same VA values.

Figure 6.10: VA Case of temporal (facial) synthesis: on the left hand side are the neutral 2D images and on the right the synthesized image sequences

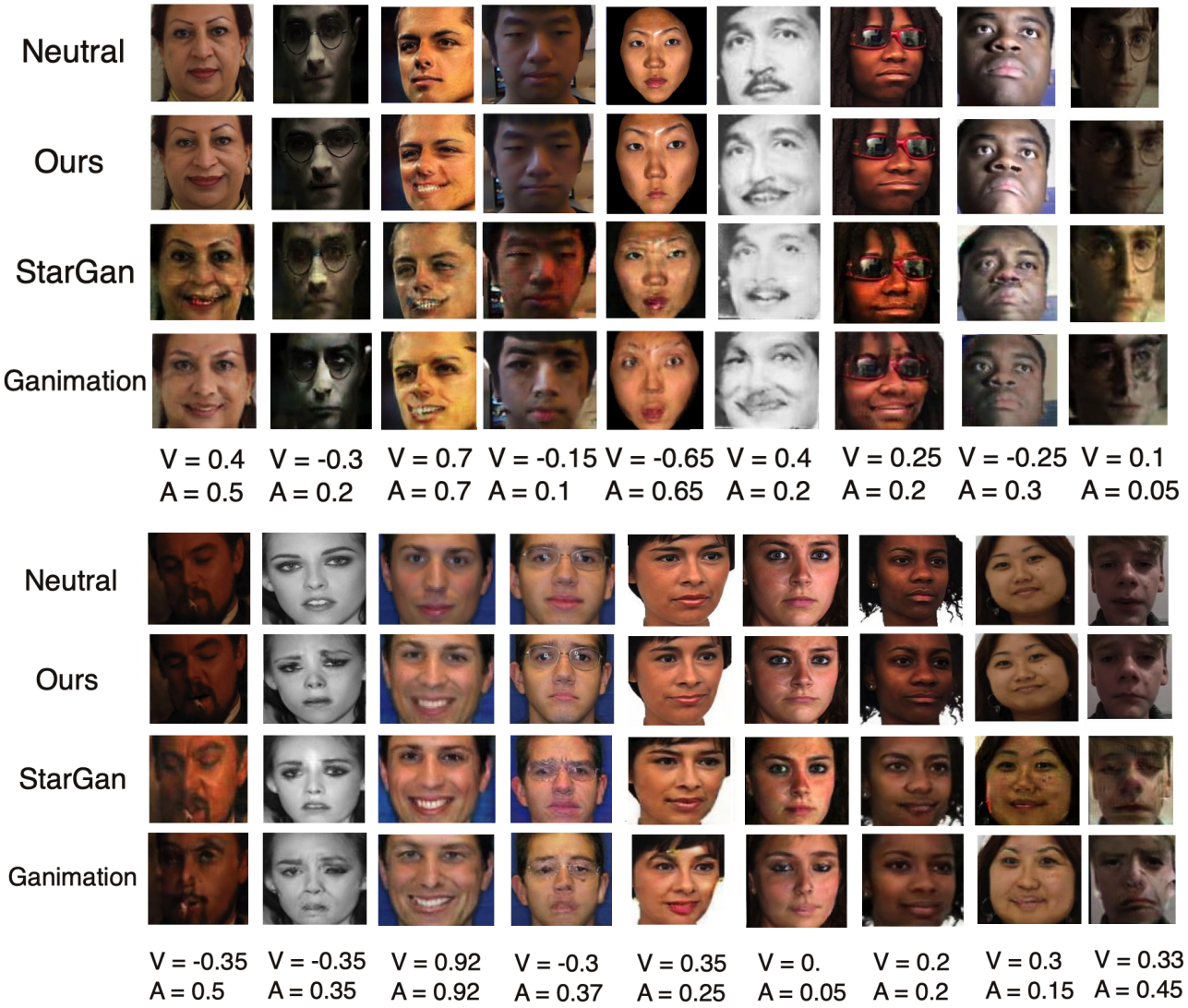


Figure 6.11: Generated results by our approach, StarGAN and GANimation

Fig. 6.11 presents a visual comparison between images generated by our approach, StarGAN and GANimation. It shows the neutral images, the synthesized VA values and the resulting images. It is evident that our approach synthesizes samples that: i) look much more natural and realistic, ii) maintain the degree of sharpness of the original neutral image, and iii) combine visual accuracy with spatial resolution.

Some further deductions can be made from Fig. 6.11. StarGAN does not perform well when tested on different in-the-wild and controlled databases that include variations in illumination conditions and head poses. StarGAN is unable to reflect detailed illumination; unnatural lighting changes were observed on the results. These can be explained because in the original StartGAN paper [31], its capability to generate affect has not been tested on in-the-wild facial analysis (we refer only to the case of emotion recognition). In general, StarGAN yields more realistic results when it is trained simultaneously with multiple datasets annotated for different tasks.

Additionally, in [31], when referring to emotion recognition, StarGAN was trained and evaluated on Radboud Faces Database (RaFD) [123] which: i) is very small in terms of size (around 4,800 images) and ii) is a lab-controlled and posed expression database. Last but not least, StarGAN has been tested to change only a particular aspect of a face among a discrete number of attributes/emotions defined by the annotation granularity of the dataset. As can be seen in Fig. 6.11, StarGAN cannot accurately provide realistic results when tested in the much broader and more difficult task of valence and arousal generation (and estimation).

As far as GANimation is concerned, its results are also worse than the results of our approach. In most cases, it shows artifacts and in some cases certain levels of blurriness. When compared to StarGAN, GANimation seems more robust to changing backgrounds and lighting conditions; this is due to the attention and color masks that it contains. Nevertheless, in general, errors in the attention mechanism occur when the input contains extreme expressions. The attention mechanism does not seem to sufficiently weight the color transformation, causing transparencies. It is interesting to note that on the Leonardo DiCaprio image, the synthesized image by GANimation shows open eyes, whereas on the neutral image (and the one synthesized by our approach) eyes are closed; this illustrates errors of the mask. For example, in Fig. 6.11, images produced by GANimation in columns 1, 3, 4, 5, 6, 9 show

the discussed problems.

6.6 Quantitative evaluation of the facial affect synthesis

Next, in order to assess the quality of the synthesized images, we perform a quantitative evaluation by using them as additional data to train Deep Neural Networks (DNNs). If the synthesized images are of good quality, using them as additional data will lead to better performance of the DNNs (compared to the case when the DNNs are trained without these additional data). Such data augmentation methodologies have been widely used in DNN training so as to increase the effective size of the training dataset. We, therefore, present a data augmentation strategy which uses the synthesized data produced by our approach, as additional data to train DNNs, for both valence-arousal prediction, as well as classification into the basic expression categories.

In particular, we describe experiments performed on eight databases, presenting the adopted evaluation criteria, the networks we used and the obtained results. We also report the performances of the networks trained -in a data augmentation manner- with synthesized images from StarGAN and GANimation. It is shown that the DNNs trained with the proposed data augmentation methodology outperform both the state-of-the-art techniques and the DNNs trained with StarGAN and GANimation, in all experiments, validating the effectiveness of the proposed facial synthesis approach. Let us first explain some notations. In the following, by reporting 'network_name trained using StarGAN', 'network_name trained using GANimation' and 'network_name trained using the proposed approach', we refer to networks trained with the specific database's training set augmented with data synthesized by StarGAN, GANimation and the proposed approach, respectively.

6.6.1 Leveraging synthesised data for training DNNs: Valence-Arousal case

In this set of experiments we consider four facial affect databases annotated in terms of valence and arousal, the Aff-Wild, RECOLA, AffectNet and AFEW-VA data-bases. At first, we selected neutral frames from these databases, i.e., frames with zero valence and arousal values (human inspection was

also conducted to make sure that they represented neutral faces). For every frame, we synthesized facial affect according to the methodology described in Section 6.3. We start by first describing the evaluation criteria used in our experiments.

The adopted evaluation criteria

The main evaluation criterion that we use is the Concordance Correlation Coefficient (CCC) [124], which has been widely used in related Challenges (e.g., [200]); we also report the Mean Squared Error (MSE), since this has been also frequently used in related research.

As already mentioned previously, CCC evaluates the agreement between two time series by scaling their correlation coefficient with their mean square difference. CCC takes values in the range $[-1, 1]$, where $+1$ indicates perfect concordance and -1 denotes perfect discordance. Therefore high values are desired. CCC is defined as follows:

$$\rho_c = \frac{2 s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}, \tag{6.6}$$

where s_x and s_y are the variances of the ground truth and predicted values respectively, $\bar{x}$ and $\bar{y}$ are the corresponding mean values and s_{xy} is the respective covariance value.

The Mean Squared Error (MSE) provides a simple comparative metric, with a small value being desirable. MSE is defined as follows:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2, \tag{6.7}$$

where x and y are the ground truth and predicted values respectively and N is the total number of samples.

In some cases we also report the Pearson-CC (P-CC) and the Sign Agreement Metric (SAGR), since they have been reported by respective state-of-the-art methods.

The P-CC takes values in the range [-1,1] and high values are desired. It is defined as follows:

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}, \tag{6.8}$$

where s_x and s_y are the variances of the ground truth and predicted values respectively and s_{xy} is the respective covariance value.

The SAGR takes values in the range [0,1], with high values being desirable. It is defined as follows:

$$SAGR = \frac{1}{N}\sum_{n=1}^{N}\delta(sign(x_i), sign(y_i)), \tag{6.9}$$

where N is the total number of samples, x and y are the ground truth and predicted values respectively, δ is the Kronecker delta function and $\delta(sign(x), sign(y))$ is defined as:

$$\delta(sign(x), sign(y)) = \begin{cases} 1, & x \geqslant 0 \text{ and } y \geqslant 0 \\ 1, & x \leqslant 0 \text{ and } y \leqslant 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.10}$$

1) Experiments on Aff-Wild We synthesized 60,135 images from the Aff-Wild database and added those images to the training set of the first Affect-in-the-wild Challenge. The employed network architecture was the AffWildNet (VGG-FACE-GRU) [105, 109] described in Chapter 3 of the Thesis.

Table 6.2 shows a comparison of the performance of: the VGG-FACE-GRU trained using: i) our approach, ii) StarGAN, iii) GANimation; AffWildNet; the winner of the Aff-Wild Challenge [26] (FATAUVA-Net).

From Table 6.2, it can be verified that the network trained on the augmented dataset, with synthesised by our approach images, outperformed all other networks. It should be noted that the number of synthesised images (around $60K$) was small compared to the size of Aff-Wild's training set (around $1M$), the latter being already sufficient for training the best performing DNN; consequently, the improvement was not large, about 2%. An interesting observation is that the network trained using StarGAN displayed worse performance than AffWildNet. This means that the 68 landmark points that were

Table 6.2: Aff-Wild: CCC and MSE evaluation of valence & arousal predictions provided by the VGG-FACE-GRU trained using our approach vs state-of-the-art networks and methods. Valence and arousal values are in $[-1, 1]$.

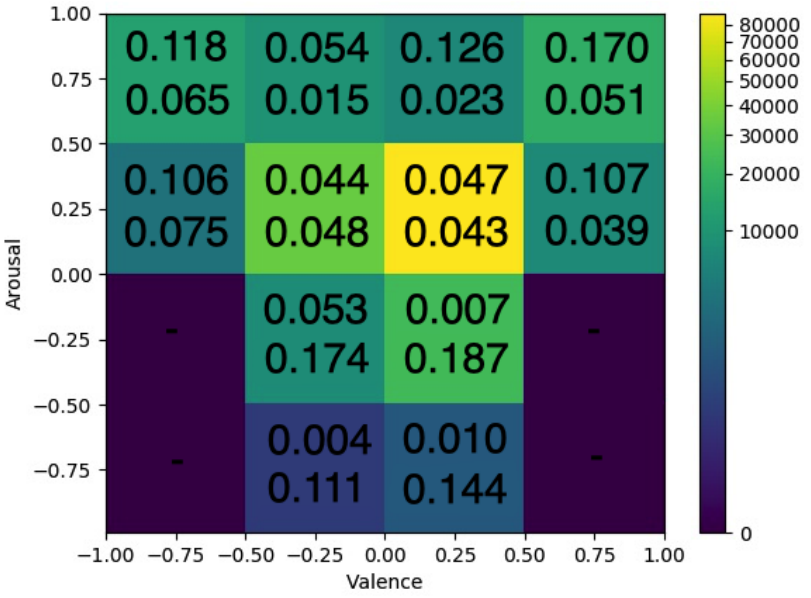
Networks	CCC		MSE	
	Valence	Arousal	Valence	Arousal
FATAUVA-Net [26]	0.396	0.282	0.123	0.095
VGG-FACE-GRU trained using StarGAN	0.556	0.424	0.085	0.060
VGG-FACE-GRU trained using GANimation	0.576	0.433	0.077	0.057
AffWildNet [105, 109]	0.570	0.430	0.080	0.060
VGG-FACE-GRU trained using the proposed approach	**0.595**	**0.445**	**0.074**	**0.051**

passed as additional input to the AffWildNet helped the network in reaching a better performance than just adding a small amount (compared to the training set size) of auxiliary synthesized data. The MSE error improvement on Valence and Arousal estimation provided by the augmented training vs the AffWildNet one, over the different areas of the VA space, is shown through the 2D histograms presented in Fig. 6.12. It can be seen that the improvement on MSE was better in areas in which a larger number of new samples was generated, i.e., in the positive valence regions.

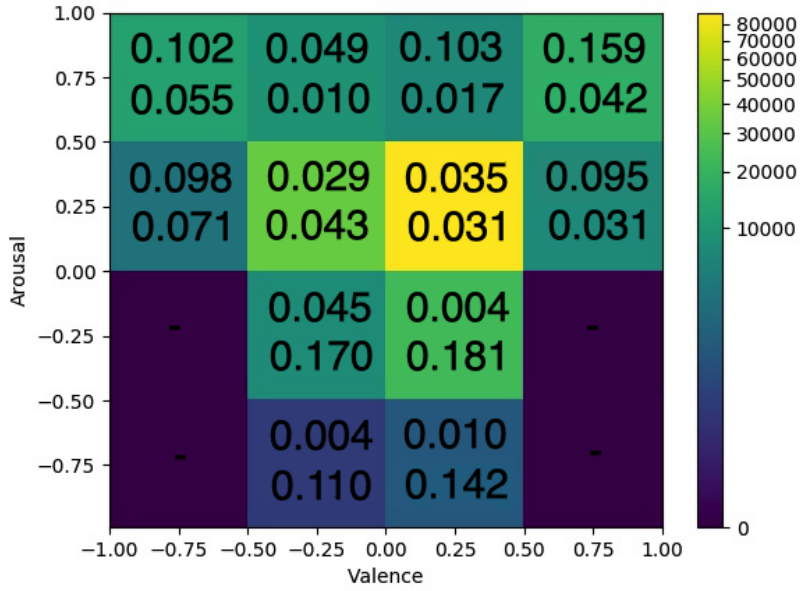
2) Experiments on RECOLA We generated 46,455 images from RECOLA; this number corresponds to around 40% of its training data set size. The employed network architecture was the ResNet-GRU described in [109].

Table 6.3 shows a comparison of the performance of: the ResNet-GRU network trained using: i) our approach, ii) StarGAN, and iii) GANimation; AffWildNet fine-tuned on RECOLA, as reported in [109]; a ResNet-GRU directly trained on RECOLA, as reported in [109].

From Table 6.3, it can be verified that the network trained using the proposed approach outperformed all other networks. The gain in performance can be justified by the fact that the number of synthesised images (around 46,500) was significant compared to the size of RECOLA's training set (around 120,000) and that the original training set size was not very sufficient to train the DNNs. It is worth mentioning that the GAN based methods have not managed to provide a sufficiently enriched dataset so that a similar boost in the achieved performances could be obtained. The MSE error improvement on Valence and Arousal estimation provided by the augmented training vs the original one (which



(a)



(b)

Figure 6.12: The 2D histogram of valence and arousal Aff-Wild's test set annotations, along with the MSE per grid area, in the case of (a) AffWildNet and (b) VGG-FACE-GRU trained using the proposed approach

was 0.045-0.100 vs 0.055-0.160), over the different areas of the VA space, is shown through the 2D histograms presented in Fig. 6.13. Big reduction of MSE value was achieved in all covered VA areas.

3) Experiments on AffectNet The AffectNet database contains around 450,000 manually annotated images and around 550,000 automatically annotated images for valence-arousal. We only used the manually annotated images so as to be consistent with the state-of-the-art networks that were also

Table 6.3: RECOLA: CCC evaluation of valence & arousal predictions provided by the ResNet-GRU trained using the proposed approach vs other state-of-the-art networks and methods.

Networks	CCC	
	Valence	Arousal
ResNet-GRU [109]	0.462	0.209
ResNet-GRU trained using StarGAN	0.503	0.245
ResNet-GRU trained using GANimation	0.486	0.222
fine-tuned AffWildNet [109]	0.526	0.273
ResNet-GRU trained using the proposed approach	**0.554**	**0.312**

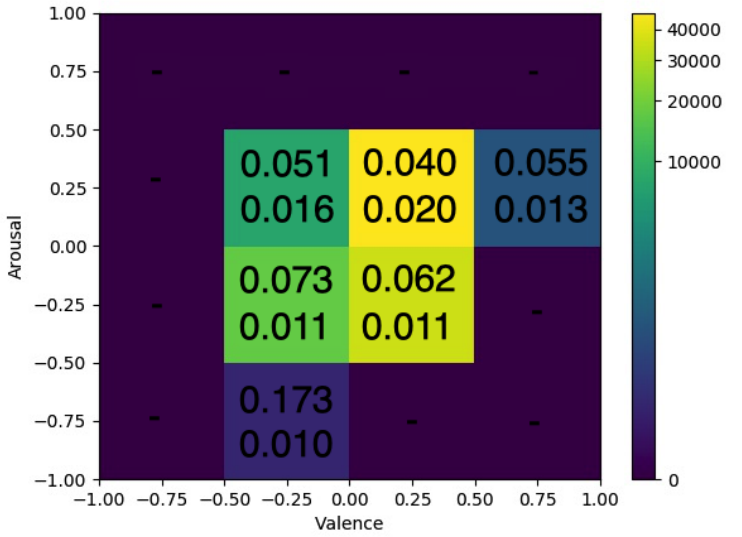
trained using this set. Additionally, the manually annotated set ensures that the images used by our approach to synthesise new, are indeed neutral. We created 2,476,235 synthesised images from the AffectNet database, a number that is more than 5 times bigger than the training data size. The employed network architecture was VGG-FACE. For comparison purposes, we trained the network using the original training data set (let us call this network 'the VGG-FACE baseline').

Table 6.4: AffectNet: CCC, P-CC, SAGR and MSE evaluation of valence & arousal predictions provided by the VGG-FACE trained using the proposed approach vs state-of-the-art networks and methods. Valence and arousal values are in $[-1, 1]$.

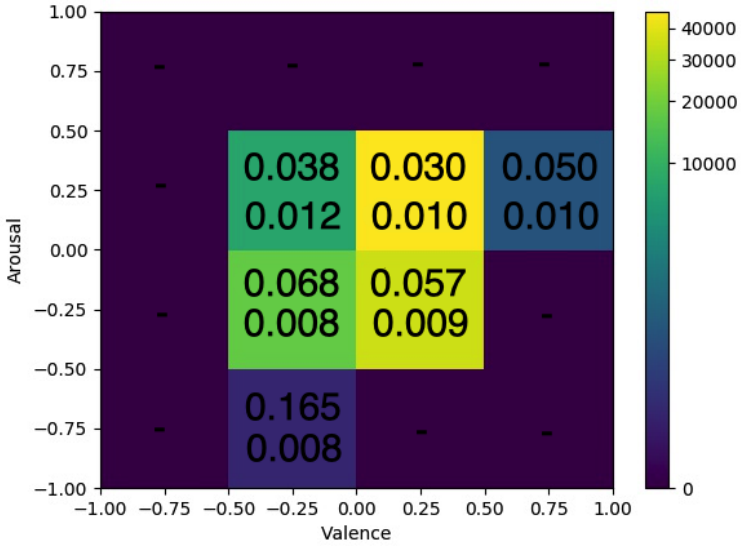
Networks	CCC		P-CC		SAGR		MSE	
	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal
AlexNet [151]	0.60	0.34	**0.66**	0.54	0.74	0.65	**0.14**	0.17
the VGG-FACE baseline	0.50	0.37	0.54	0.48	0.65	0.60	0.19	0.18
VGG-FACE trained using StarGAN	0.55	0.42	0.58	0.49	0.74	0.73	0.17	0.16
VGG-FACE trained using GANimation	0.56	0.45	0.59	0.51	0.74	0.74	0.15	0.16
VGG-FACE trained using the proposed approach	**0.62**	**0.54**	**0.66**	**0.55**	**0.78**	**0.75**	**0.14**	**0.15**

Table 6.4 shows a comparison of the performance of: the VGG-FACE baseline; the VGG-FACE trained using: i) our approach, ii) StarGAN, and iii) GANimation; AlexNet, which is the state-of-the-art network of the AffectNet database [151].

From Table 6.4, it can be verified that the network trained by the proposed methodology outperformed all other networks. This boost in performance has been large, in all evaluation criteria, compared to the VGG-FACE baseline network, with spread of this improvement over the VA space shown in Fig. 6.14. The explanation arises from the large number of synthesized images that helped the network



(a)

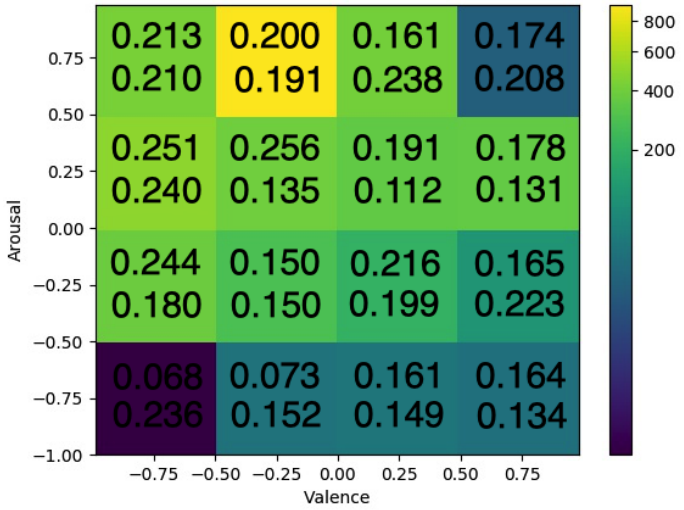


(b)

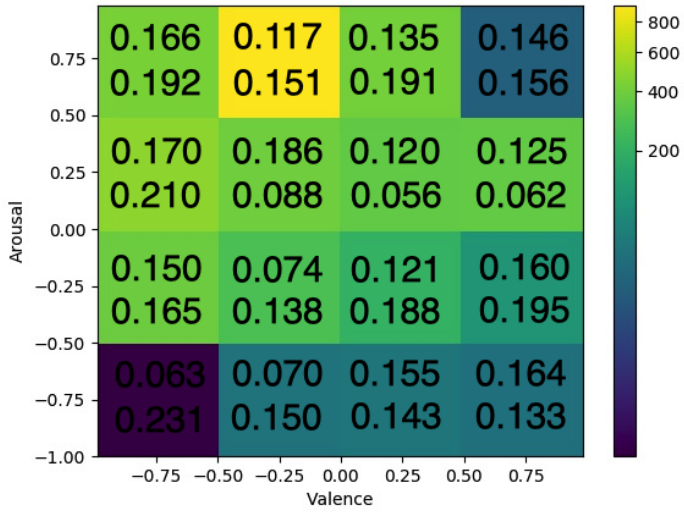
Figure 6.13: The 2D histogram of valence and arousal RECOLA's test set annotations, along with the MSE per grid area, in the case of (a) ResNet-GRU and (b) ResNet-GRU trained using the proposed approach

train and generalize better, since in the training set there existed a lot of ranges that were poorly represented. This is shown in the histogram of the -manually annotated- training set, for valence and arousal, in Fig. 6.15. Our network also outperformed the AffectNet's database baseline. For the arousal estimation, the performance gain was remarkable, mainly in CCC and SAGR evaluation criteria, whereas for the valence estimation the performance gain was also significant.

4) Experiments on AFEW-VA. We synthesised 108,864 images from the AFEW-VA database, a



(a)



(b)

Figure 6.14: The 2D histogram of valence and arousal AffectNet's test set annotations, along with the MSE per grid area, in the case of (a) the VGG-FACE baseline, (b) the VGG-FACE trained using the proposed approach

number that is more than 3.5 times bigger than its original size. For training, we used the VGG-FACE-GRU architecture described in [109]. Similarly to [118], we used a 5-fold person-independent cross-validation strategy and at each fold we augmented the training set with the synthesised images of people appearing only in that set (preserving person independence).

Table 6.5 shows a comparison of the performance of: the VGG-FACE-GRU network trained using: i) our approach, ii) StarGAN, and iii) GANimation; the best performing network as reported in [118].

From Table 6.5, it can be verified that the network trained using the proposed approach outperformed

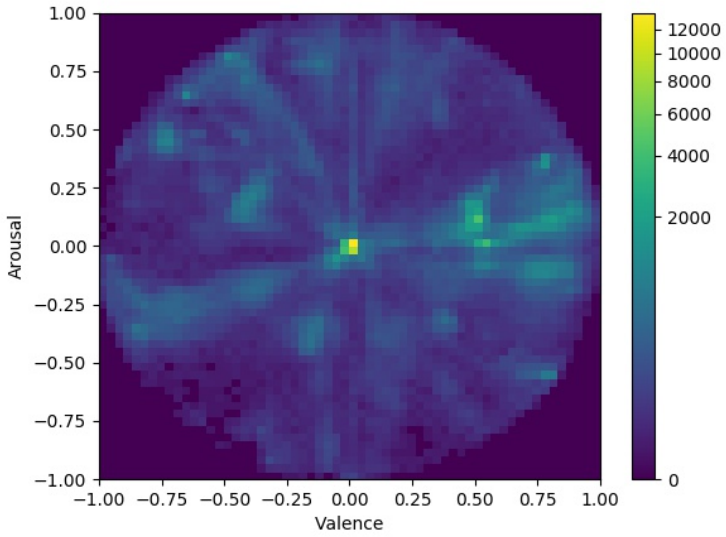


Figure 6.15: The 2D histogram of valence and arousal AffectNet's annotations for the manually annotated training set

Table 6.5: AFEW-VA: P-CC and MSE evaluation of valence & arousal predictions provided by the VGG-FACE trained using the proposed approach vs state-of-the-art network and methods. Valence and arousal values are in $[-1, 1]$.

Networks	Pearson CC		MSE	
	Valence	Arousal	Valence	Arousal
best of [118]	0.407	0.450	0.484	0.247
VGG-FACE trained using StarGAN	0.512	0.489	0.262	0.097
VGG-FACE trained using GANimation	0.491	0.453	0.308	0.151
VGG-FACE-GRU trained using the proposed approach	**0.562**	**0.614**	**0.226**	**0.075**

all other networks. Great boost in performance was achieved. The general gain in performance can be justified by the fact that the number of synthesised images (around 109,000) is much greater than the number of images in the dataset (around 30,000), with the latter being rather small for effectively training the DNNs. The 2D histogram in Fig. 6.16 shows the achieved MSE when using the proposed approach over the different areas of the VA space.

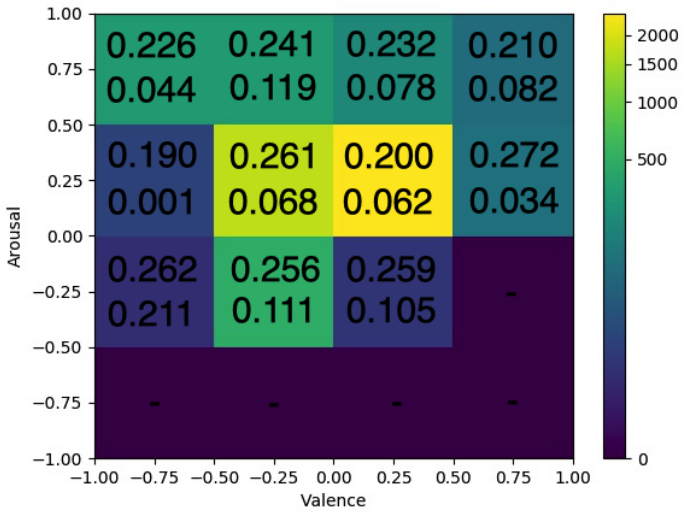


Figure 6.16: The 2D histogram of valence and arousal AFEW-VA's test set annotations, along with the MSE per grid area, in the case of the VGG-FACE trained using the proposed approach

6.6.2 Leveraging synthesised data for training DNNs: Basic Expressions case

In the following experiments we used the synthesized faces to train DNNs, for classification into the six basic expressions, over four facial affect databases, RAF-DB, AffectNet, AFEW and BU-3DFE. Our first step has been to select neutral frames from these four databases. Then, for each frame, we synthesised facial affect according to the methodology described in Section 6.3. We start by first describing the evaluation criteria used in our experiments.

The adopted evaluation criteria

One evaluation criterion used in the experiments is total accuracy, defined as the total number of correct predictions divided by the total number of samples. Another criterion is the F_1 score, which is a weighted average of the recall (= the ability of the classifier to find all the positive samples) and precision (= the ability of the classifier not to label as positive a sample that is negative). The F_1 score reaches its best value at 1 and its worst score at 0. In our multi-class problem, F_1 score is the unweighted mean of the F_1 scores of the expression classes. F_1 score of each class is defined as:

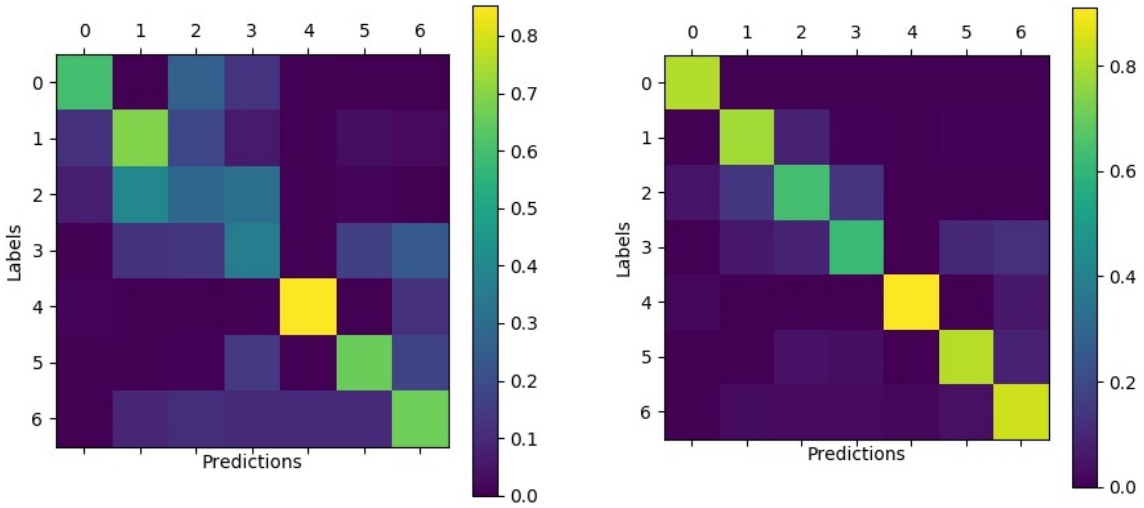
$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{6.11}$$

Another criterion that is used is the average of the diagonal values of the confusion matrix for the seven basic expressions.

One, or more of the above criteria are used in our experiments, so as to illustrate the comparison with other state-of-the-art methods.

Table 6.6: RAF-DB: The diagonal values of the confusion matrix for the seven basic expressions and their average, using the VGG-FACE trained using the proposed approach, as well as using other state-of-the-art networks.

Networks	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral	Average
LDA-VGG-FACE [128]	0.661	0.250	0.378	0.731	0.515	0.535	0.472	0.506
mSVM-VGG-FACE [128]	0.685	0.275	0.351	0.853	0.649	0.663	0.599	0.582
the VGG-FACE baseline	0.691	0.287	0.363	0.853	0.661	0.666	0.600	0.589
mSVM-DLP-CNN [128]	0.716	0.522	**0.622**	**0.928**	0.801	0.812	0.803	0.742
VGG-FACE trained using the proposed approach	**0.784**	**0.644**	**0.622**	0.911	**0.812**	**0.845**	**0.806**	**0.775**



(a) (b)

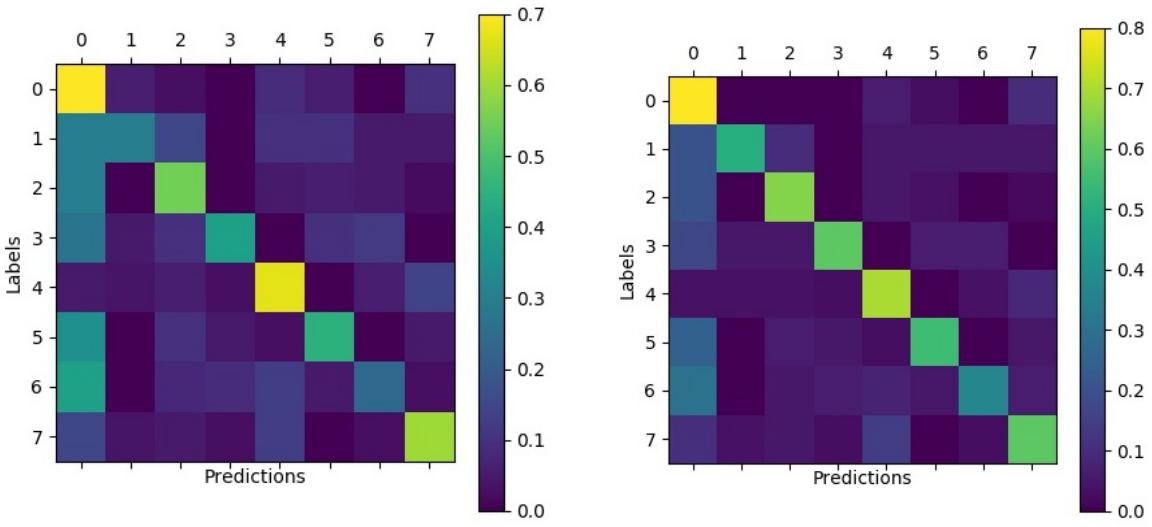
Figure 6.17: The confusion matrix of (a) the VGG-FACE baseline and (b) the VGG-FACE trained using the proposed approach for the RAF-DB database; 0: Neutral, 1: Anger, 2: Disgust, 3: Fear, 4: Joy, 5: Sadness, 6: Surprise

1) Experiments on RAF-DB. In this database we only considered the six basic expression categories, since our approach synthesizes images based on these categories; we ignored compound expressions that were included in the original dataset. We created 12,828 synthesized images, which are slightly more than the training images (12,271). We employed the VGG-FACE network. For comparison purposes, we trained the network using the original training dataset (let us call this network 'the VGG-FACE baseline').

For further comparison purposes, we used the networks defined in [128]: i) mSVM-VGG-FACE: first the VGG-FACE was trained on the RAF-DB database and then features from the penultimate fully connected layer were extracted and fed into a Support Vector Machine (SVM) that performed the classification, ii) LDA-VGG-FACE: same as before: LDA was applied on the features which were extracted from the penultimate fully connected layer and performed the final classification and iii) mSVM-DLP-CNN: the designed Deep Locality Preserving CNN network (we refer the interested reader for more details to [128]) was first trained on the RAF-DB database and then a SVM performed the classification using the features extracted from the penultimate fully connected layer of this architecture.

Table 6.6 shows a comparison of the performance of the above described networks. From Table 6.6, it can be verified that the network trained using the proposed approach outperformed all state-of-the-art nets. When compared to the mSVM-VGG-FACE and LDA-VGG-FACE networks, the boost in performance has been significant. This can be explained by the fact that the disgust and fear classes, originally, did not contain a lot of training images, but after adding the synthesised data, they did. This resulted in obtaining a better performance in the other classes, as well. Interestingly, there was also a considerable performance gain in the neutral class, that did not contain any synthesised images. This can be explained by considering the fact that the network trained with the augmented data could distinguish better the classes, since it had more samples in the two above described categories. Fig. 6.17 illustrates the whole confusion matrix of the VGG-FACE baseline and the VGG-FACE trained using the proposed approach, giving a better insight on the improved performance and verifying the above explanations.

2) Experiments on AffectNet. We synthesised 176,425 images from the AffectNet database, a number that is almost 40% of its size. It should be mentioned that the AffectNet database contained the six basic expressions and another one, contempt. Our approach synthesized images only for the basic expressions, so for the contempt class we only kept the original training data. The network architecture that we employed here was VGG-FACE. For comparison purproses, we trained a VGG-FACE network using the training set of the AffectNet database (let us call this network 'the VGG-FACE baseline').



(a) (b)

Figure 6.18: The confusion matrix of (a) the VGG-FACE baseline and (b) the VGG-FACE trained using the proposed approach for the AffectNet database; 0: Neutral, 1: Anger, 2: Disgust, 3: Fear, 4: Joy, 5: Sadness, 6: Surprise, 7: Contempt

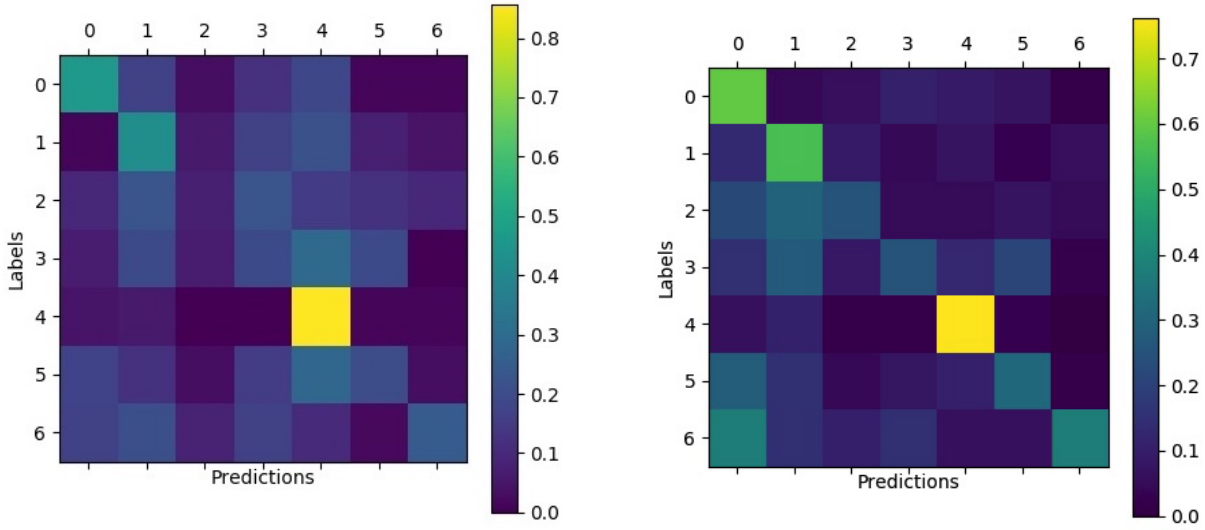
Table 6.7 shows a comparison of the performance of: i) the VGG-FACE baseline, ii) the VGG-FACE network trained using the proposed approach and iii) AlexNet, the baseline network of the AffectNet database [151].

Table 6.7: AffectNet: Total accuracy and F_1 score of the VGG-FACE trained using the proposed approach vs state-of-the-art networks

Networks	Total Accuracy	F_1 score
AlexNet [151]	0.58	0.58
the VGG-FACE baseline	0.52	0.51
VGG-FACE trained using the proposed approach	**0.60**	**0.59**

From Table 6.7, it can be verified that the network trained using the proposed approach outperformed all the other networks. In more detail, when compared to the VGG-FACE baseline network, the boost in performance was significant, as also shown in Fig. 6.18 in terms of the confusion matrices obtained by the two networks. This can be explained by the big size of the added synthesized images. When compared to the AffectNet's baseline, a slightly improved performance was also obtained; this could be higher, if we had synthesized images for the contempt category as well.

3) Experiments on AFEW. We synthesized 56,514 images from the AFEW database; this number was almost 1.4 times bigger than its training set size (41,406). The employed network architecture



(a) (b)

Figure 6.19: The confusion matrix of (a) the VGG-FACE baseline and (b) the VGG-FACE trained using the proposed approach for the AFEW database; 0: Neutral, 1: Anger, 2: Disgust, 3: Fear, 4: Joy, 5: Sadness, 6: Surprise

was VGG-FACE. For comparison purposes, we first trained a baseline network on AFEW's training set, which we call the VGG-FACE baseline. For further comparisons, we used the following networks developed by the three winning methods of the EmotiW 2017 Grand Challenge: i) VGG-FACE-FER: the VGG-FACE was first fine-tuned on the FER2013 database [76] and then trained on the AFEW as described in [100], ii) VGG-FACE-external: the VGG-FACE was trained on the union of the AFEW database and some external data as described in [205] and iii) VGG-FACE-LSTM-external-augmentation: the VGG-FACE-LSTM was trained on the union of the AFEW database and some external data; then data augmentation was performed, as described in [205].

Table 6.8: AFEW: Total accuracy of the VGG-FACE trained using the proposed approach vs state-of-the-art networks

Networks	Total Accuracy
the VGG-FACE baseline	0.379
VGG-FACE-external [205]	0.414
VGG-FACE-FER [100]	0.483
VGG-FACE-LSTM-external-augmentation [205]	**0.486**
VGG-FACE trained using the proposed approach	0.484

Table 6.8 shows a comparison of the performance of the above described networks. From Table 6.8,

one can see that the VGG-FACE trained using the proposed approach performed much better than the same network trained on, either only the AFEW database, or the union of the AFEW database with some external data whose size in terms of videos was the same as that of AFEW. The boost in performance can be explained taking into account the fact that the fear, disgust and surprise classes contained few data in AFEW and that our approach augmented the data size of those classes; in total the large number of synthesised images assisted to improve the performance of the network. This is evident when comparing the confusion matrix of the VGG-FACE baseline to the one of VGG-FACE trained using the proposed approach, as can be seen in Fig.6.19. The diagonal of the two confusion matrices indicates that there is an increase in the performance in almost all basic categories.

Additionally, performance of our network is slightly better than the performance of the same VGG-FACE network first fine-tuned on the FER2013 database and then trained on the AFEW. FER2013 is a database of around 35,000 still images and different identities, annotated with the six basic expressions. In this case, the network that was first fine-tuned on the FER2013 database has seen more faces, since the tasks were similar. However, still our network provided a slightly better performance. On the other hand, our network had a slightly worse performance than a VGG-FACE-LSTM network that was trained with the same external data mentioned before and was also trained with data augmentation. Here, it was the LSTM network, which due to the time recurrent nature could better exploit the fact that AFEW consists of video sequences.

4) Experiments on BU-3DFE. We synthesized 600 images from the BU-3DFE database. This number was almost one fourth of its size (2,500). BU-3DFE is a small database and is not really suited for training DNNs. The network architecture that we employed here was VGG-FACE, with a modification in the number of hidden units in the two first fully connected layers. Since we did not have a lot of data for training the network, we i) used 256 and 128 units in the two fully connected layers and ii) kept the convolutional weights fixed, training only the fully connected ones. For training the network on this database, we used a 10-fold person-independent cross-validation strategy; in each fold, we augmented the training set with the synthesised images of people appearing only in that set (preserving person independence). The reported total accuracy of the model has been the average of the total accuracies over the 10-folds.

At first, we trained the above described VGG-FACE network (let us call this network 'the VGG-FACE baseline'). Next, we trained the above described VGG-FACE network, but also applied on-the-fly data augmentation techniques, such as: small rotations, left and right flipping, first resize and then random crop to original dimensions, random brightness and saturation (let us call this network 'VGG-FACE-augmentation'). Finally, we trained the above described VGG-FACE network using the proposed approach.

Table 6.9: BU-3DFE: Total accuracy of the VGG-FACE trained using the proposed approach vs the VGG-FACE baseline and the VGG-FACE trained with on-the-fly data augmentation.

Networks	Total Accuracy
the VGG-FACE baseline	0.528
VGG-FACE-augmentation	0.588
VGG-FACE trained using the proposed approach	**0.768**

Table 6.9 shows a comparison of the performance of those networks. From Table 6.9, it can be verified that the network trained using the proposed approach greatly outperformed the networks trained without it. This indicates that the proposed approach for synthesising images can be used for data augmentation in cases of small amount of DNN training data, being able to significantly improve the obtained performance.

6.7 Ablation Studies

6.7.1 Quantitative evaluation of facial affect synthesis in testing or training

Results in the previous section show that the data generated using our approach provide improvements in network performance in both valence-arousal and basic expressions settings, when used for data augmentation. In the following, we perform further analysis (two different settings) to assess the quality of our generated data, compared to the data synthesised by StarGAN and GANimation, focusing only on the synthesised data.

In the first setting, the synthesised data are evaluated as a test set, for each database, against models trained on real data/images. The AffWildNet that has been trained solely on Aff-Wild's training

set, the ResNet-GRU trained on the RECOLA's training set and the VGG-FACE baseline trained on AffectNet's training set, have been used as emotion regressors and were evaluated on each of the three afore-mentioned synthesised datasets. From Table 6.10 it is evident that the networks trained on the afore mentioned databases displayed a much better performance (in all databases) when tested on data produced by the proposed approach in comparison to data produced by StarGAN, or GANimation.

We further conducted a second setting, using the synthesised data to train respective DNN models. These models were then evaluated on the real test set of Aff-Wild, RECOLA and AffectNet. Table 6.11 shows the results of this setting. The performance in terms of both CCC and MSE was much higher in all databases when the networks were trained with data synthesised by the proposed approach. This difference in achieved performance, along with the former results, reflect the direct value of our generated data in enhancing regression accuracy.

Table 6.10: CCC and MSE evaluation of valence & arousal predictions provided by the: i) AffWildNet (trained on Aff-Wild), ii) ResNet-GRU (trained on RECOLA) and iii) the VGG-FACE baseline (trained on AffectNet); these networks are tested on images produced by StarGAN, GANimation and our approach. Each score is shown in the format: Valence value-Arousal value

Databases	Methods	Evaluation Metrics	Networks		
			AffWildNet [109]	ResNet-GRU [109]	the VGG-FACE baseline
Aff-Wild	StarGAN	CCC MSE	0.33-0.26 0.21-0.19	-	-
	GANimation	CCC MSE	0.35-0.28 0.19-0.16	-	-
	Ours	CCC MSE	**0.43-0.33** **0.15-0.13**	-	-
RECOLA	StarGAN	CCC	-	0.29-0.23	-
	GANimation	CCC	-	0.28-0.22	-
	Ours	CCC	-	**0.34-0.33**	-
AffectNet	StarGAN	CCC MSE	-	-	0.23-0.23 0.34-0.37
	GANimation	CCC MSE	-	-	0.26-0.21 0.31-0.38
	Ours	CCC MSE	-	-	**0.39-0.31** **0.27-0.28**

6.7.2 Effect of synthesised data granularity on performance improvement

In this subsection we performed experiments using a subset of our synthesised data for augmenting the data-bases. Our aim was to see if all synthesised data were needed for augmenting network

Table 6.11: CCC and MSE evaluation of valence & arousal predictions provided by the: i) AffWild-Net, ii) ResNet-GRU and iii) the VGG-FACE baseline; these networks are trained on the synthesized images by StarGAN, GANimation and our approach; these networks are evaluated on the Aff-Wild, RECOLA and AffectNet test sets. Each score is shown in the format: Valence value-Arousal value

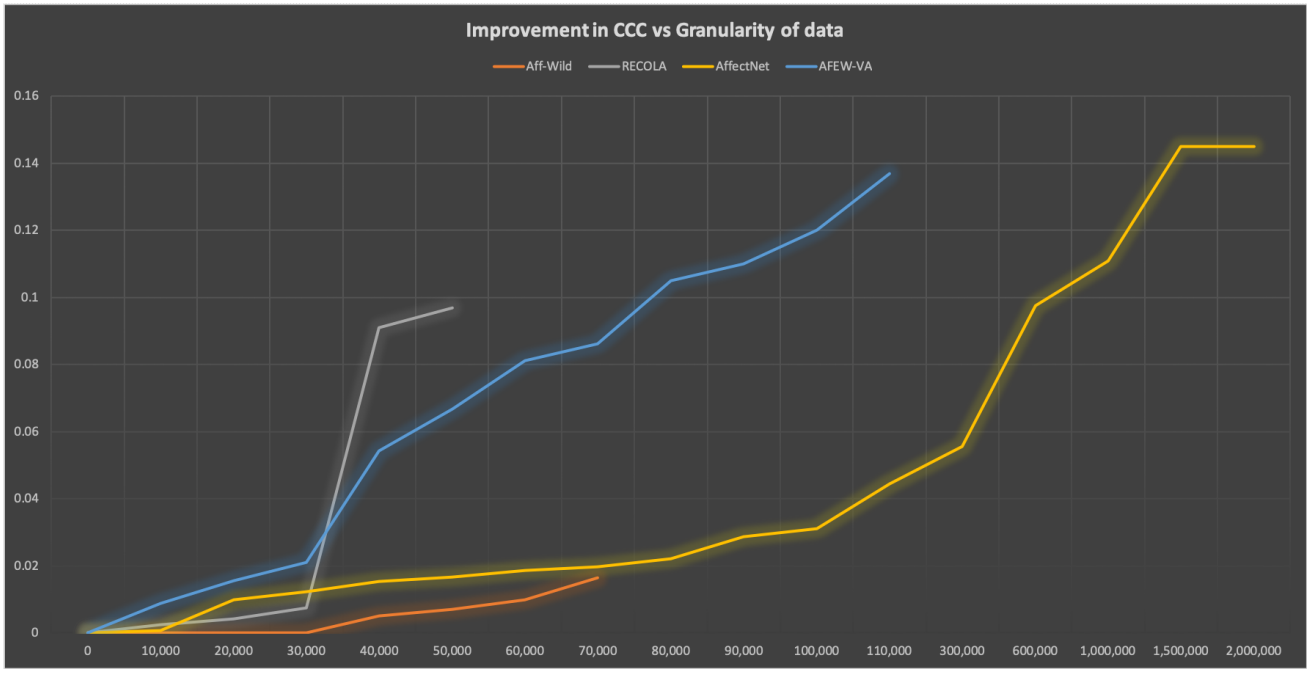
Databases	Methods	Evaluation Metrics	Networks		
			AffWildNet	ResNet-GRU	VGG-FACE baseline
Aff-Wild	StarGAN	CCC MSE	0.16-0.13 0.18-0.17	-	-
	GANimation	CCC MSE	0.17-0.14 0.17-0.15	-	-
	Ours	CCC MSE	**0.21-0.20** **0.15-0.12**	-	-
RECOLA	StarGAN	CCC	-	0.19-0.10	-
	GANimation	CCC	-	0.17-0.10	-
	Ours	CCC	-	**0.23-0.14**	-
AffectNet	StarGAN	CCC MSE	-	-	0.37-0.29 0.23-0.21
	GANimation	CCC MSE	-	-	0.40-0.31 0.20-0.19
	Ours	CCC MSE	-	-	**0.45-0.35** **0.18-0.17**

training and more generally to see how the improvement in classification and regression scaled with the granularity of synthesised data. In more detail, for each database used in our experiments, we used a subset of N synthesised data from this database to augment its training set. Table 6.12 shows the databases and the corresponding N values.

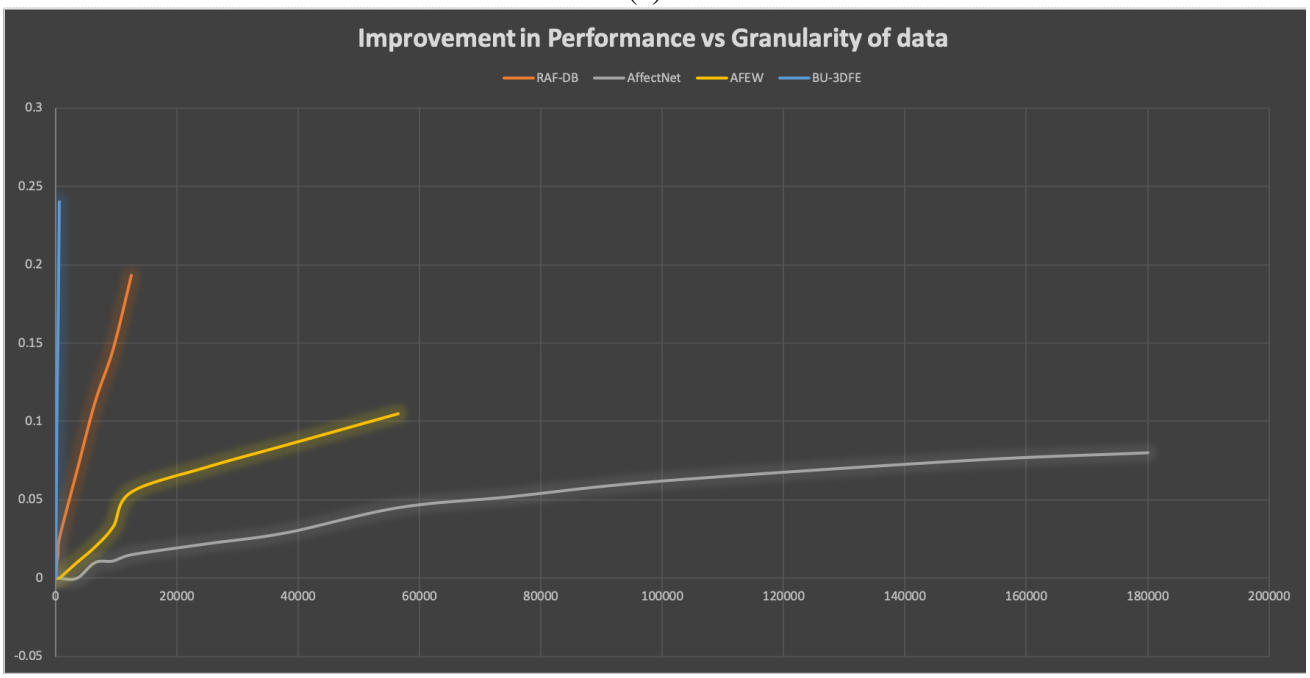
Fig. 6.20 shows the improvement in network performance when training using additionally auxiliary data; the improvement shown per database is the difference in performance when training networks with only the database's training set and when training them with the union of the training and auxil-

Table 6.12: Databases used in our approach and the different values of N for each one; N denotes a subset of the synthesised data (per database) by the proposed approach

Databases	N synthesized data
Aff-Wild	$N \in \{10K, 20K, 30K, 40K, 50K, 60K\}$
RECOLA	$N \in \{10K, 20K, 30K, 40K, 50K\}$
AffectNet (VA)	$N \in \{10K, 20K, 30K, 40K, 50K, 60K, 70K, 80K, 90K, 100K, 110K, 300K, 600K, 1M, 1.5M, 2M, 2.5M\}$
AFEW-VA	$N \in \{10K, 20K, 30K, 40K, 50K, 60K, 70K, 80K, 90K, 100K, 110K\}$
RAF-DB	$N \in \{200, 400, 600, 3.5K, 6.5K, 9.5K, 12.5K\}$
AffectNet (Expressions)	$N \in \{6.5K, 12.5K, 25K, 38K, 56.5K, 75K, 100K, 150K, 180K\}$
AFEW	$N \in \{3.5K, 6.5K, 12.5K, 25K, 38K, 56.5K\}$
BU-3DFE	$N \in \{200, 400, 600\}$



(a)



(b)

Figure 6.20: Improvement in network performance vs amount of synthesized data; criteria: (a) mean/average CCC of VA in Aff-Wild, RECOLA, AffectNet, AFEW-VA and (b) mean diagonal value of the confusion matrix for RAF-DB, F1 score for AffectNet, Total Accuracy for AFEW and BU-3DFE.

iary datasets. Fig. 6.20 illustrates for each database the difference in network performance, when N data generated by our approach (N defined in Table 6.12) were used as auxiliary data.

The performance measure for Aff-Wild, RECOLA, AffectNet and AFEW-VA is the average of va-

lence CCC and arousal CCC. The performance measure for the rest databases depends on the database, as reported next.

Dimensional affect generation

For the Aff-Wild database, we used the VGG-FACE-GRU network. When augmenting the dataset with $30K$ or less synthesised images, no performance improvement was observed, whereas when augmenting it with more than $30K$, the performance increased, following the increase in the granularity of synthesised data. Adding synthesised data to the training set seemed to be beneficial for improving performance and thus improvement would be much greater if we added more than $60K$ (if we had more neutral expressions), although probably at a given point, a plateau would be reached (considering the large training set that consisted of around $1M$ images).

For the RECOLA database, we used the ResNet-GRU network. When augmenting the dataset with up to $30K$ synthesized images, there appeared small performance improvement, whereas when augmenting it with more than $30K$, the performance was continuously increasing following the increase in the granularity of synthesised data; this increase is large. This is expected, since $120K$ frames are not sufficient for training a network for regression and additionally, $170K$ frames are not either.

For the AffectNet database, we used the VGG-FACE network. After adding $10K$ synthesised images, performance started to increase. This increase continued as more data were added until the training set had been augmented with $1.5M$ data. If more data were added, the performance did not change, implying that a plateau had been reached. The final performance improvement was large.

For the AFEW-VA database, we used the VGG-FACE-GRU network. The improvement was systematically very significant. When adding more than $30K$ data, the increase in performance was more rapid. The performance is expected to continue increasing while more data are added, as both the initial training set of around $23K$ frames and the augmented set of around $135K$ frames are not large enough to train a DNN for regression.

Categorical affect generation

For the RAF-DB database, we used the VGG-FACE network and the performance was measured in terms of the mean diagonal value of the confusion matrix. The increase in performance was almost linear as more data were used. The final gain in performance was large. RAF-DB is a very small database (of size about $12K$ images) and therefore if we had more data to add, the performance would further improve.

In the AffectNet database, we used the VGG-FACE network and performance was measured in terms of the F1 score. Increasing the amount of added data provided a respective increase in performance. After adding $60K$ images the performance was increasing at a lower rate. It should be mentioned that the results included erroneous classification of the contempt class. If we synthesised samples of the contempt class as well, the network would provide a higher performance; but this was beyond the scope of our work.

In the AFEW database, we used the VGG-FACE network; the performance measure was total accuracy. The performance increased with the addition of more data. This increase was significant. The AFEW database is a small database (of size about $40K$ images) and therefore adding data is expected to improve performance.

In the BU-3DFE database, we used the VGG-FACE network; the performance measure was total accuracy. There was a huge and rapid increase in network performance with the addition of data. This is explained by the very small size of BU-3DFE (around $2K$) which makes it impossible to train a neural network on it.

General deductions that can be made from Fig. 6.20:

- the smaller the size of the database, the bigger and faster the increase in performance would be, when augmenting it with data synthesised by our approach
- the improvement in performance is small if we augment the training set with few data in proportion to its size
- in dimensionally annotated databases, a plateau is reached and no further improvement is seen when a lot of data (about $\geq 1.5M$ in our case) are added

Table 6.13: Age Analysis in terms of CCC and MSE for the dimensionally annotated databases

Databases	Ages	# Test Samples	# Synthesized Samples	Network-Augmented		Network	
				CCC	MSE	CCC	MSE
Aff-Wild	20-29	29,013	5,301	**0.61-0.38**	**0.101-0.063**	0.59-0.37	0.102-0.066
	30-39	99,962	23,427	**0.66-0.47**	**0.077-0.054**	0.61-0.44	0.088-0.066
	40-49	44,727	21,831	**0.50-0.48**	**0.048-0.033**	0.46-0.44	0.054-0.044
	50-59	41,748	9,120	**0.58-0.40**	**0.074-0.054**	0.57-0.38	0.075-0.057
	total	215,450	59,679	**0.60-0.45**	**0.074-0.051**	0.57-0.43	0.080-0.060
RECOLA	30-39	90,000	11,001	**0.61-0.38**	-	0.60-0.34	-
	40-49	15,000	16,188	**0.43-0.24**	-	0.36-0.19	-
	50-59	7,500	11,742	**0.49-0.20**	-	0.44-0.10	-
	total	112,500	38,931	**0.55-0.31**	-	0.53-0.27	-
AffectNet	0-19	172	118,902	**0.67-0.55**	**0.105-0.156**	0.61-0.41	0.127-0.181
	20-29	1,179	714,232	**0.60-0.53**	**0.128-0.159**	0.51-0.36	0.170-0.193
	30-39	1,218	814,588	**0.64-0.54**	**0.139-0.145**	0.50-0.39	0.193-0.169
	40-49	762	452,504	**0.64-0.61**	**0.149-0.134**	0.49-0.44	0.202-0.166
	50-59	569	229,938	**0.58-0.53**	**0.161-0.149**	0.47-0.34	0.216-0.181
	60-89	600	146,091	**0.62-0.44**	**0.145-0.167**	0.51-0.29	0.200-0.195
	total	4,500	2,476,235	**0.62-0.54**	**0.141-0.150**	0.50-0.37	0.190-0.180
AFEW-VA	20-29	766	17,466	0.46-0.60	0.192-0.084	-	-
	30-39	1,990	36,388	0.51-0.62	0.254-0.080	-	-
	40-49	1,558	34,906	0.59-0.47	0.211-0.076	-	-
	50-59	946	15,102	0.74-0.85	0.215-0.045	-	-
	60-79	396	4,102	0.63-0.45	0.236-0.100	-	-
	total	5,646	108,864	0.57-0.59	0.226-0.075	-	-

- the performance due to data augmentation does not increase commensurately; in the AffectNet database (mainly in the valence-arousal case) the gain yielded by data augmentation saturates as N increases
- generally, the performance increase is larger in categorically annotated databases in comparison to dimensionally annotated ones. This is an interesting result, since it indicates that synthesising more data is needed in the latter case, to make the data distribution more dense.

Effect of subjects' age in classification & regression results

It is interesting to quantitatively assess the effect of age on the performance of the proposed approach. However, not all databases contain age information about their subjects. To achieve this, we trained an age estimator on them. In more detail, we trained a Wide Residual Network (WideResNet) [227] on the union of IMDB [174] and Adience datasets [60] (so that the training dataset contained an adequate number of images of people under the age of 25) and tested it on WIKI [174]. Then we applied this estimator on the test sets of the examined databases.

Table 6.14: Age Analysis for the categorically annotated databases; criterion for RAF-DB & AffectNet is F1 score, for AFEW & BU-3DFE is total accuracy; AFEW test samples refer to number of videos (frames)

Databases	Ages	# Test Samples	# Synthesized Samples	VGG-FACE-Augmented	VGG-FACE
				Performance Metric	Performance Metric
RAF-DB	10-19	168	210	**0.631**	0.446
	20-29	911	2,250	**0.813**	0.556
	30-39	998	4,320	**0.739**	0.498
	40-49	516	3,606	**0.744**	0.511
	50-59	258	1,776	**0.709**	0.440
	60-69	149	552	**0.657**	0.550
	70-79	68	128	**0.904**	0.635
	total	3,068	12,828	**0.738**	0.505
AffectNet	0-19	152	12,516	**0.593**	0.453
	20-29	882	45,182	**0.584**	0.477
	30-39	962	55,513	**0.593**	0.518
	40-49	594	27,632	**0.586**	0.532
	50-59	431	20,204	**0.648**	0.606
	60-69	289	11,178	**0.564**	0.498
	70-79	161	3,582	**0.466**	0.398
	80-89	29	618	**0.448**	0.410
	total	3,500	176,425	**0.590**	0.510
AFEW	20-29	29 (1,536)	6,474	**0.379**	0.241
	30-39	156 (8,568)	22,518	**0.455**	0.333
	40-49	132 (7,803)	17,934	**0.553**	0.439
	50-59	57 (3,202)	7,482	**0.474**	0.456
	60-79	16 (764)	2,106	**0.438**	0.313
	total	390 (21,873)	56,514	**0.484**	0.379
BU-3DFE	20-29	115	192	**0.800**	0.600
	30-39	100	240	**0.820**	0.570
	40-49	100	120	**0.800**	0.550
	50-59	100	30	**0.790**	0.490
	60-70	85	18	**0.600**	0.400
	total	500	600	**0.768**	0.528

Table 6.13 shows, for each dimensionally annotated database (Aff-Wild, RECOLA, AffectNet and AFEW-VA), the estimated age groups (we split the age values into appropriate groups so that each group contained a significant amount of samples), the number of test samples that are within the age groups, the number of samples synthesised by our approach for each age group, different evaluation metrics (CCC and MSE) for each age group in two cases: when a network trained only with the training set of each database was used (denoted as 'Network' in Table 6.13) and when the same network was trained with the training set augmented with data synthesised by our approach (denoted as 'Network-Augmented' in Table 6.13). For Aff-Wild and AFEW-VA, the VGG-FACE-GRU network was used, for RECOLA the ResNet-GRU and for AffectNet the VGG-FACE.

Table 6.14 is similar to Table 6.13 with the difference being that it refers to categorically annotated databases (RAF-DB, AffectNet, AFEW and BU-3DFE). In this case, the evaluation metrics are the F1 score for RAF-DB and AffectNet, and the total accuracy for AFEW and BU-3DFE. The 'VGG-FACE-Augmented' refers to the case in which the VGG-FACE network is trained on the union of training set of each database and data synthesised by our approach.

By observing the two Tables (6.13 and 6.14), it is seen that augmenting the training dataset with the images generated by our approach is beneficial in all age groups, both for regression and classification. It would be interesting to focus on specific groups, such as very young ($<$20 years old) in RAF-DB and AffectNet, each containing more than 150 subjects, or elderly (e.g., 70-79 years old) in AffectNet, also containing more than 150 subjects. In the former case, the F1 value improved from about 0.45 to 0.6; the F1 values over all categories improved from about 0.51 to 0.66. Although the F1 values in the very young category were lower than the mean F1 values over all ages, the improvement in both cases was similar. A similar observation can be made in the latter case, of elderly persons, with the F1 value in the category being improved from about 0.4 to 0.47. Although these values were lower than the total F1 values over all ages, which were 0.51 and 0.59 respectively, the improvement in these cases was similar as well. This verifies the observation that the proposed approach for data augmentation is also beneficial in cases where the number of available samples is rather small.

Chapter 7

Conclusions & Future Work

7.1 Summary of Thesis Achievements

The current thesis has managed to create new knowledge on affect analysis, recognition and synthesis. This knowledge contains new large databases in-the-wild, annotated in terms of: dimensional variables, i.e., valence and arousal; seven basic expression categories; facial action units. It also contains novel deep neural architectures that are trained with these databases, providing state-of-the-art performance on them and constitute robust priors for both dimensional and categorical recognition over all other main datasets in-the-wild. Moreover, it includes a new approach for facial affect synthesis that can be used for data augmentation and improvement of the performance of deep neural networks in dimensional and categorical affect recognition.

In particular, we first presented the generation of Aff-Wild, a new, very large in-the-wild database. Aff-Wild has been introduced in a respective Workshop and Challenge in CVPR 2017 and used, thereafter, by many researchers in the field. We also presented the design of the AffWildNet, which produced the best performance for valence and arousal estimation on Aff-Wild, both in terms of the Concordance Correlation Coefficient and the Mean Squared Error criteria, when compared to other deep learning networks trained on the same database.

We then showed that the AffWildNet, which has been trained on the Aff-Wild database, constitutes

a robust prior for affect recognition on other datasets and other environments. Using appropriate retraining methodologies, AffWildNet was able to produce the best performance when retrained and applied to other dimensional databases, when compared to other state-of-the-art pre-trained and fine-tuned networks. Furthermore, we have been able to show that AffWildNet can be a robust prior, not only for dimensional, but also for categorical affect recognition. This was the first time that the same deep neural architecture has been successfully trained for valence-arousal estimation (which is a regression analysis problem) and then used for categorical affect analysis (which is a classification, with seven basic expression categories, problem).

We extended AffWildNet by extracting low-, mid- and high-level latent information from it and analysing this by multiple RNN subnets. Moreover, we used an ensemble approach so as to perform model-level fusion, which produced excellent results for visual affect recognition on the OMG-Emotion Challenge.

In the following, we presented the extension of Aff-Wild and the generation of Aff-Wild2, which is the largest in-the-wild, audiovisual, database, being annotated in terms of valence-arousal dimensions, seven basic expressions and facial action units. We also presented the design of multi-task and multi-modal deep neural architectures that extend AffWildNet, being trained on Aff-Wild2. We tested their performance -in a cross-database setting- on ten other databases, illustrating that they beat all state-of-the-art methods for affect recognition. We further trained new deep neural architectures on Aff-Wild2, by adapting the ArcFace Loss Function. By using these as priors for expression recognition on all existing databases, we improved the existing state-of-the art.

Moreover, we presented the development of a new deep neural architecture, named FaceBehaviorNet, which is the first holistic framework for behaviour analysis in-the-wild. FaceBehaviorNet is an end-to-end network trained for joint: basic expression recognition, action unit detection and valence-arousal estimation, over all publicly available databases, containing over 5M images. Additionally we developed a novel strategy for coupling all the tasks during training, based on co-annotation and on distribution matching, consistently outperforming all existing methodologies. By exploring the feature representations learned through the joint training, we illustrated the good generalisation abilities for recognition of compound expressions, under zero-shot or few-shot learning settings.

A novel approach was then proposed so as to generate facial affect on faces. It leverages a dimensional emotion model in terms of valence and arousal or the six basic expressions, and a large scale 4D face database, the 4DFAB. We performed dimensional annotation of the 4DFAB and used the facial images with their respective annotations to generate mean faces on a discretised 2-D affect space. A methodology has been proposed using these mean faces to synthesise faces with affect, both categorical or dimensional, both static or dynamic. Using a given neutral image and the desired affect, which can be a Valence Arousal pair of values, a path in the 2D VA space, or one of the basic expression categories, the proposed approach performs face detection and landmark localization on the input neutral image, fits a 3D Morphable Model on the resulting image, deforms the reconstructed face, adds the input affect and blends the new face with the given affect into the original image.

An extensive experimental study has been conducted, providing both qualitative and quantitative evaluation of the proposed approach. The qualitative results showed the achieved higher quality of the synthesised data compared to GAN-generated facial affect. The quantitative results were based on using the synthesised facial images for data augmentation and training of Deep Neural Networks over eight databases, annotated with either dimensional or categorical affect labels. It has been shown that, over all databases, the achieved performance was much higher than i) the performance of the respective state-of-the-art methods, ii) the performance of the same DNNs with data augmentation provided by the StarGAN and GANimation networks.

7.2 Applications

There are numerous applications of the technologies and data developed in this Thesis, in the areas of human computer interaction, computer vision, robotics, security and biomedical, as well as consumer applications. In all these fields, it is of great significance, if agents can, on the one hand, detect and analyse the affect of their user, patient, or customer, and on the other hand, adapt their behaviour, e.g., their 'facial expression' to appropriately react to this affect.

Behaviour analysis is a main component in these applications. We have already developed systems implementing our developments that are able to capture the facial affect of users and represent it in

the 2-D affect space, analyse the main expression involved in it, as well as extract the active facial action units. They are also possible to detect faces and extract this affect information from videos in real time.

Using this technology for customer behaviour analysis is a feasible application that can be used, for example in big supermarkets, or in banks, where the security system can be used for analysing the affect in customers' faces. Similarly, for surveillance, for example, in train stations, or in airports, detection and analysis of travellers' faces can be performed in real time. This can provide alerts when some unusual behaviour, e.g., negative affect, or expression of anger or fear, are detected. We have tested the technologies with images and image sequences from CCTV cameras and the results have been very good. An issue that needs to be mentioned is that affect recognition can be performed without the need for face recognition. It is well known that face recognition in surveillance applications does not match persons' privacy right. However, the anonymous affect recognition operation can be set so as not to have conflict with person's privacy requirement.

Through specific adaptations, the developed technology can be used by robots, so that they can successfully operate in human robot interactions. It is essential that a 'companion' robot can understand an elderly person's behavioural state, to be able to assist them in a friendly and effective way.

Similarly, a robot, taking care of a child, should be able to understand its behaviour and treat it accordingly. Other applications include detection of a driver's behaviour for providing alert in case of loss of attention, or tiredness. Similarly lie detection can take advantage of the rich facial affect analysis capabilities produced by the developed technologies.

7.3 Future Work

Our future plans include further improvement of the generated state-of-the-art in the field, both related to data generation, as well as to development of deep neural architectures that are able to learn over the data and generalise well in other datasets or environments.

Our data generation plans include extending Aff-Wild2 with large numbers of high-quality 4K videos

in-the-wild. The target is to produce a large testbed for developing new scalable architectures that can learn to analyse affect, by extracting coarse-to-fine information from visual inputs in-the-wild. Moreover, focusing on specific types of affect, for example, related to negative or reluctant behaviours, or on compound emotions, will require extending the databases, and/or performing domain adaptation of the developed architectures in these frameworks.

Most of the approaches developed in this Thesis, and in the affect recognition field in general, are based on supervised learning, by ensuring that experts in the field provide the required annotation of the aggregated, or generated data. However, labeling large number of datasets can be infeasible due to lack of experts in a continuous base. Unsupervised learning will be further investigated in our future work for handling non-annotated data cases, while focusing on related problems, such as uncertainty of the estimation procedure. Extending domain adaptation approaches by introducing and using modified Loss Functions in training over non-annotated data constitutes our first step in this direction [114].

Moreover, although it has been shown that deep neural architectures are capable of analysing large datasets for affect recognition, they lack transparency in their decision making, in the sense that it is not straightforward to justify their prediction. In this Thesis we have investigated the extraction of latent variables, containing low-, medium- and high-level semantic information, from deep neural architectures during training, and further exploring them through multiple networks, for improving the performance in affect recognition.

In our future work we will further analyse this latent information, through unsupervised learning, so as to develop respective representations, clusters, graphs. We will use the latter to provide transparency, visualisation and explainability of the decision making procedure adopted by the deep neural architectures. Moreover, we will be able to extend the zero-, or one-shot learning approaches we introduced in the Thesis to provide effective and efficient training over the extracted representations.

Another future direction will be the exploitation of the analysis by synthesis approach that was presented in Chapter 5 in broader contexts and environments. The ability to generate dimensional, or specific types of affect on faces and use them for training deep neural architectures can be applied in a variety of contexts. Blending generation of dimensional affect with generation of facial action units

will be of major interest, since it can provide a local-to-global facial synthesis of affect.

Finally, our future work will include adaptation of the developed deep neural architectures and use of them in other applications, including the ones described in the previous subsection. Of major interest is healthcare prediction, through the analysis of medical images and image sequences. The presented architectures can be adapted and used in analysis of both time varying and volumetric medical information, such as Magnetic Resonance Images, or CT scan series, for early prediction of diseases.

Bibliography

[1] Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., Lucey, S.: Using synthetic data to improve facial expression analysis with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1609–1618 (2017)

[2] Alabort-i-Medina, J., Antonakos, E., Booth, J., Snape, P., Zafeiriou, S.: Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In: Proceedings of the ACM International Conference on Multimedia, MM '14, pp. 679–682. ACM, New York, NY, USA (2014). DOI 10.1145/2647868.2654890. URL http://doi.acm.org/10.1145/2647868.2654890

[3] Albanie, S., Vedaldi, A.: Learning grimaces by watching tv. In: Proceedings of the British Machine Vision Conference (BMVC) (2016)

[4] Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid icp algorithms for surface registration. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)

[5] Antoniou, A., Storkey, A., Edwards, H.: Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340 (2017)

[6] Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: Advances in neural information processing systems, pp. 41–48 (2007)

[7] Aung, M.S., Kaltwang, S., Romera-paredes, B., Martinez, B., Singh, A., Cella, M., Valstar, M.F., Meng, H., Kemp, A., Elkins, A.C., Tyler, N., Watson, P.J., Williams, A.C., Pantic, M.,

Berthouze, N.: The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset. IEEE Transactions on Affective Computing (2016)

[8] Averbuch-Elor, H., Cohen-Or, D., Kopf, J., Cohen, M.F.: Bringing portraits to life. ACM Transactions on Graphics (TOG) **36**(6), 196 (2017)

[9] Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems, pp. 892–900 (2016)

[10] Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning **4**(1), 1–106 (2012). DOI 10.1561/2200000015. URL http://dx.doi.org/10.1561/2200000015

[11] Baltrušaitis, T., Mahmoud, M., Robinson, P.: Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, pp. 1–6 (2015)

[12] Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., Wermter, S.: The omg-emotion behavior dataset. arXiv preprint arXiv:1803.05434 (2018)

[13] Barsoum, E., Zhang, C., Canton Ferrer, C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ACM International Conference on Multimodal Interaction (ICMI) (2016)

[14] Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 279–283 (2016)

[15] Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Fully automatic facial action recognition in spontaneous behavior. In: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, pp. 223–230. IEEE (2006)

[16] Benitez-Quiroz, C., Srinivasan, R., Martinez, A.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of

IEEE International Conference on Computer Vision & Pattern Recognition (CVPR'16). Las Vegas, NV, USA (2016)

[17] Benitez-Quiroz, C.F., Srinivasan, R., Feng, Q., Wang, Y., Martinez, A.M.: Emotionet challenge: Recognition of facial expressions of emotion in the wild. arXiv preprint arXiv:1703.01210 (2017)

[18] Blanz, V., Basso, C., Poggio, T., Vetter, T.: Reanimating faces in images and video. In: Computer graphics forum, vol. 22, pp. 641–650. Wiley Online Library (2003)

[19] Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., Zafeiriou, S.: 3d face morphable models "in-the-wild". In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017). URL https://arxiv.org/abs/1701.05360

[20] Booth, J., Roussos, A., Ponniah, A., Dunaway, D., Zafeiriou, S.: Large scale 3d morphable models. International Journal of Computer Vision **126**(2-4), 233–254 (2018)

[21] Booth, J., Zafeiriou, S.: Optimal uv spaces for facial morphable model construction. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 4672–4676. IEEE (2014)

[22] Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y.: Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 302–309. IEEE (2018)

[23] Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Transactions on graphics (TOG) **33**(4), 43 (2014)

[24] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on, pp. 67–74. IEEE (2018)

[25] Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)

[26] Chang, W.Y., Hsu, S.H., Chien, J.H.: Fatauva-net : An integrated deep learning framework for facial attribute recognition, action unit (au) detection, and valence-arousal estimation. In:

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (2017)

[27] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)

[28] Chen, S., Jin, Q., Zhao, J., Wang, S.: Multimodal multi-task learning for dimensional and continuous emotion recognition. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 19–26. ACM (2017)

[29] Cheng, S., Kotsia, I., Pantic, M., Zafeiriou, S.: 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

[30] Chew, S.W., Lucey, P., Lucey, S., Saragih, J., Cohn, J.F., Matthews, I., Sridharan, S.: In the pursuit of effective affective computing: The relationship between features and registration. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **42**(4), 1006–1016 (2012)

[31] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)

[32] Chopra, S., Hadsell, R., LeCun, Y., et al.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR (1), pp. 539–546 (2005)

[33] Chrysos, G.G., Antonakos, E., Snape, P., Asthana, A., Zafeiriou, S.: A comprehensive performance evaluation of deformable face tracking "in-the-wild". International Journal of Computer Vision **126**(2-4), 198–232 (2018)

[34] Corneanu, C., Oliu, M., Cohn, J., Escalera, S.: Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. IEEE transactions on pattern analysis and machine intelligence (2016)

[35] Cosker, D., Krumhuber, E., Hilton, A.: A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In: 2011 International Conference on Computer Vision, pp. 2296–2303. IEEE (2011)

[36] Cowie, R., Cornelius, R.R.: Describing the emotional states that are expressed in speech. Speech communication **40**(1), 5–32 (2003)

[37] Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., Schröder, M.: 'feeltrace': An instrument for recording perceived emotion in real time. In: ISCA tutorial and research workshop (ITRW) on speech and emotion (2000)

[38] Cowie, R., McKeown, G., Douglas-Cowie, E.: Tracing emotion: an overview. International Journal of Synthetic Emotions (IJSE) **3**(1), 1–17 (2012)

[39] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, pp. 886–893. IEEE (2005)

[40] Dalgleish, T., Power, M.: Handbook of cognition and emotion. John Wiley & Sons (2000)

[41] Darwin, C., Prodger, P.: The expression of the emotions in man and animals. Oxford University Press, USA (1998)

[42] Deng, D., Chen, Z., Shi, B.E.: Fau, facial expressions, valence and arousal: A multi-task solution. arXiv preprint arXiv:2002.03557 (2020)

[43] Deng, D., Zhou, Y., Pi, J., Shi, B.E.: Multimodal utterance-level affect analysis using visual, audio and text features. arXiv preprint arXiv:1805.00625 (2018)

[44] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 248–255. IEEE (2009)

[45] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. arXiv preprint arXiv:1801.07698 (2018)

[46] Deng, J., Zhou, Y., Cheng, S., Zaferiou, S.: Cascade multi-view hourglass model for robust 3d face alignment. pp. 399–403 (2018). DOI 10.1109/FG.2018.00064

[47] Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T.: From individual to group-level emotion recognition: Emotiw 5.0. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 524–528. ACM (2017)

[48] Dhall, A., Goecke, R., Joshi, J., Hoey, J., Gedeon, T.: Emotiw 2016: Video and group-level emotion recognition challenges. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 427–432. ACM (2016)

[49] Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T.: Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 461–466. ACM (2014)

[50] Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T.: Emotion recognition in the wild challenge 2013. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp. 509–516. ACM (2013)

[51] Dhall, A., Kaur, A., Goecke, R., Gedeon, T.: Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 2018 on International Conference on Multimodal Interaction, pp. 653–656. ACM (2018)

[52] Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 423–426. ACM (2015)

[53] Dhall, A., et al.: Collecting large, richly annotated facial-expression databases from movies

[54] Dimitrios, K., Shiyang, C., Evangelos, V., Irene, K., Stefanos, Z.: Deep neural network augmentation: Generating faces for affect analysis. International Journal of Computer Vision **128**(5), 1455–1484 (2020)

[55] Ding, H., Sricharan, K., Chellappa, R.: Exprgan: Facial expression editing with controllable expression intensity. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)

[56] Ding, H., Zhou, S.K., Chellappa, R.: Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 118–126. IEEE (2017)

[57] Ding, W., Huang, D.Y., Chen, Z., Yu, X., Lin, W.: Facial action recognition using very deep networks for highly imbalanced class distribution. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1368–1372. IEEE (2017)

[58] Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., Heylen, D.K.: The sensitive artificial listner: an induction technique for generating emotionally coloured conversation. In: LREC Workshop on Corpora for Research on Emotion and Affect. ELRA (2008)

[59] Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proceedings of the National Academy of Sciences **111**(15), E1454–E1462 (2014)

[60] Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security **9**(12), 2170–2179 (2014)

[61] Ekman, P.: Facial action coding system (facs). A human face (2002)

[62] Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of personality and social psychology **17**(2), 124 (1971)

[63] Ekman, R.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA (1997)

[64] Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia, pp. 1459–1462. ACM (2010)

[65] Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5562–5570 (2016)

[66] Fried, O., Shechtman, E., Goldman, D.B., Finkelstein, A.: Perspective-aware manipulation of portrait photos. ACM Transactions on Graphics (TOG) **35**(4), 128 (2016)

[67] Frijda, N.H., et al.: The emotions. Cambridge University Press (1986)

[68] Garrido, P., Valgaerts, L., Rehmsen, O., Thormahlen, T., Perez, P., Theobalt, C.: Automatic face reenactment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4217–4224 (2014)

[69] Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

[70] Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. IEEE Access **7**, 64,827–64,836 (2019)

[71] Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. IEEE Access **7**, 64,827–64,836 (2019)

[72] Girard, J.M., Chu, W.S., Jeni, L.A., Cohn, J.F.: Sayette group formation task (gft) spontaneous facial expression database. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 581–588. IEEE (2017)

[73] Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., et al.: Multiple classifier systems for the classification of audio-visual emotional states. In: International Conference on Affective Computing and Intelligent Interaction, pp. 359–368. Springer (2011)

[74] Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)

[75] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680 (2014)

[76] Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three

machine learning contests. In: International Conference on Neural Information Processing, pp. 117–124. Springer (2013)

[77] Gower, J.C.: Generalized procrustes analysis. Psychometrika **40**(1), 33–51 (1975)

[78] Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28**(5), 807–813 (2010)

[79] Han, S., Meng, Z., Khan, A.S., Tong, Y.: Incremental boosting convolutional neural network for facial action unit recognition. In: Advances in neural information processing systems, pp. 109–117 (2016)

[80] Han, S., Pool, J., Narang, S., Mao, H., Gong, E., Tang, S., Elsen, E., Vajda, P., Paluri, M., Tran, J., et al.: Dsd: Dense-sparse-dense training for deep neural networks. arXiv preprint arXiv:1607.04381 (2016)

[81] Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis; an overview with application to learning methods. Technical report, Royal Holloway, University of London (2003)

[82] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[83] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)

[84] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[85] Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition, pp. 84–92. Springer (2015)

[86] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[87] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141 (2018)

[88] Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y.: Learning supervised scoring ensemble for emotion recognition in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 553–560. ACM (2017)

[89] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708 (2017)

[90] Iordan, A., Dolcos, F.: Brain activity and network interactions linked to valence-related differences in the impact of emotional distraction. Cerebral cortex **27**(1), 731–749 (2017)

[91] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134 (2017)

[92] Jack, R.E., Garrod, O.G., Yu, H., Caldara, R., Schyns, P.G.: Facial expressions of emotion are not culturally universal. Proceedings of the National Academy of Sciences **109**(19), 7241–7244 (2012)

[93] Jayaraman, D., Sha, F., Grauman, K.: Decorrelating semantic visual attributes by resisting the urge to share. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1629–1636 (2014)

[94] Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2983–2991 (2015)

[95] Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., et al.: Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp. 543–550. ACM (2013)

[96] Kaneko, T., Hiramatsu, K., Kashino, K.: Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks. In: Proceedings of the 24th ACM international conference on Multimedia, pp. 327–331 (2016)

[97] Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)

[98] Khorrami, P., Le Paine, T., Brady, K., Dagli, C., Huang, T.S.: How deep neural networks can improve emotion recognition on video data. In: Image Processing (ICIP), 2016 IEEE International Conference on, pp. 619–623. IEEE (2016)

[99] Khorrami, P., Paine, T., Huang, T.: Do deep neural networks learn facial action units when doing expression recognition? In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 19–27 (2015)

[100] Knyazev, B., Shvetsov, R., Efremova, N., Kuharenko, A.: Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. arXiv preprint arXiv:1711.04598 (2017)

[101] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing **3**(1), 18–31 (2012)

[102] Kokkinos, I.: Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6129–6138 (2017)

[103] Kollias, D., Cheng, S., Pantic, M., Zafeiriou, S.: Photorealistic facial synthesis in the dimensional affect space. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 0–0 (2018)

[104] Kollias, D., Marandianos, G., Raouzaiou, A., Stafylopatis, A.G.: Interweaving deep learning and semantic techniques for emotion analysis in human-machine interaction. In: 2015 10th In-

ternational Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 1–6. IEEE (2015)

[105] Kollias, D., Nicolaou, M.A., Kotsia, I., Zhao, G., Zafeiriou, S.: Recognition of affect in the wild using deep neural networks. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pp. 1972–1979. IEEE (2017)

[106] Kollias, D., Schulc, A., Hajiyev, E., Zafeiriou, S.: Analysing affective behavior in the first abaw 2020 competition. arXiv preprint arXiv:2001.11409 (2020)

[107] Kollias, D., Sharmanska, V., Zafeiriou, S.: Face behavior\a la carte: Expressions, affect and action units in a single network. arXiv preprint arXiv:1910.11111 (2019)

[108] Kollias, D., Tagaris, A., Stafylopatis, A.: On line emotion detection using retrainable deep neural networks. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–8. IEEE (2016)

[109] Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. International Journal of Computer Vision **127**(6-7), 907–929 (2019)

[110] Kollias, D., Yu, M., Tagaris, A., Leontidis, G., Stafylopatis, A., Kollias, S.: Adaptation and contextualization of deep neural network models. In: 2017 IEEE symposium series on computational intelligence (SSCI), pp. 1–8. IEEE

[111] Kollias, D., Zafeiriou, S.: Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770 (2018)

[112] Kollias, D., Zafeiriou, S.: A multi-component cnn-rnn approach for dimensional emotion recognition in-the-wild. arXiv preprint arXiv:1805.01452 (2018)

[113] Kollias, D., Zafeiriou, S.: A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. arXiv preprint arXiv:1811.07771 (2018)

[114] Kollias, D., Zafeiriou, S.: Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2018)

[115] Kollias, D., Zafeiriou, S.: Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. arXiv preprint arXiv:1910.01417 (2019)

[116] Kollias, D., Zafeiriou, S.: Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. arXiv preprint arXiv:1910.04855 (2019)

[117] Kollias, D., Zafeiriou, S.: Va-stargan: Continuous affect generation. In: International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 227–238. Springer (2020)

[118] Kossaifi, J., Tzimiropoulos, G., Todorovic, S., Pantic, M.: Afew-va database for valence and arousal estimation in-the-wild. Image and Vision Computing (2017)

[119] Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B.W., et al.: Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)

[120] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)

[121] Kuhnke, F., Rumberg, L., Ostermann, J.: Two-stream aural-visual affect analysis in the wild. arXiv preprint arXiv:2002.03399 (2020)

[122] Kuipers, J.B., et al.: Quaternions and rotation sequences, vol. 66. Princeton university press Princeton (1999)

[123] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition and emotion **24**(8), 1377–1388 (2010)

[124] Lawrence, I., Lin, K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics pp. 255–268 (1989)

[125] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)

[126] Lee, A.: Welcome to virtualdub. org!-virtualdub. org (2002)

[127] Li, J., Chen, Y., Xiao, S., Zhao, J., Roy, S., Feng, J., Yan, S., Sim, T.: Estimation of affective level in the wild with multiple memory networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (2017)

[128] Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2852–2861 (2017)

[129] Li, W., Abtahi, F., Zhu, Z.: Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1841–1850 (2017)

[130] Liu, C., Tang, T., Lv, K., Wang, M.: Multi-feature based emotion recognition for video clips. In: Proceedings of the 2018 on International Conference on Multimodal Interaction, pp. 630–634. ACM (2018)

[131] Liu, H., Zeng, J., Shan, S., Chen, X.: Emotion recognition for in-the-wild videos. arXiv preprint arXiv:2002.05447 (2020)

[132] Liu, W., Wang, Z.: Facial expression recognition based on fusion of multiple gabor features. In: 18th International Conference on Pattern Recognition (ICPR'06), vol. 3, pp. 536–539. IEEE (2006)

[133] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 212–220 (2017)

[134] Liu, X., Mao, T., Xia, S., Yu, Y., Wang, Z.: Facial animation by optimized blendshapes from motion capture data. Computer Animation and Virtual Worlds **19**(3-4), 235–245 (2008)

[135] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pp. 94–101. IEEE (2010)

[136] Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Matthews, I.: Painful data: The unbc-mcmaster shoulder pain expression archive database. In: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp. 57–64. IEEE (2011)

[137] Ma, L., Deng, Z.: Real-time facial expression transformation for monocular rgb video. In: Computer Graphics Forum, vol. 38, pp. 470–481. Wiley Online Library (2019)

[138] Mahoor, M., Hasani, B.: Facial affect estimation in the wild using deep residual and convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (2017)

[139] Maimon, O., Rokach, L.: Data mining and knowledge discovery handbook (2005)

[140] Martinez, B., Valstar, M.F.: Advances, challenges, and opportunities in automatic facial expression recognition. In: Advances in face detection and facial image analysis, pp. 63–100. Springer (2016)

[141] Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. In: European Conference on Computer Vision, pp. 720–735. Springer (2014)

[142] Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: A spontaneous facial action intensity database. Affective Computing, IEEE Transactions on **4**(2), 151–160 (2013)

[143] McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schröder, M.: The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. Affective Computing, IEEE Transactions on **3**(1), 5–17 (2012)

[144] Alabort-i Medina, J., Zafeiriou, S.: A unified framework for compositional fitting of active appearance models. International Journal of Computer Vision **121**(1), 26–64 (2017)

[145] Mehu, M., Scherer, K.R.: Emotion categories and dimensions in the facial communication of affect: An integrated approach. Emotion **15**(6), 798 (2015)

[146] Meng, H., Bianchi-Berthouze, N.: Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In: International Conference on Affective Computing and Intelligent Interaction, pp. 378–387. Springer (2011)

[147] Meng, H., Huang, D., Wang, H., Yang, H., Ai-Shuraifi, M., Wang, Y.: Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 21–30. ACM (2013)

[148] Mickley Steinmetz, K.R., Kensinger, E.A.: The effects of valence and arousal on the neural activity leading to subsequent memory. Psychophysiology **46**(6), 1190–1199 (2009)

[149] Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

[150] Mohammed, U., Prince, S.J., Kautz, J.: Visio-lization: generating novel facial images. ACM Transactions on Graphics (TOG) **28**(3), 57 (2009)

[151] Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. arXiv preprint arXiv:1708.03985 (2017)

[152] Najibi, M., Samangouei, P., Chellappa, R., Davis, L.: SSH: Single stage headless face detector. In: The IEEE International Conference on Computer Vision (ICCV) (2017)

[153] Neumann, T., Varanasi, K., Wenger, S., Wacker, M., Magnor, M., Theobalt, C.: Sparse localized deformation components. ACM Transactions on Graphics (TOG) **32**(6), 179 (2013)

[154] Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 443–449. ACM (2015)

[155] Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: 2014 IEEE international conference on image processing (ICIP), pp. 343–347. IEEE (2014)

[156] Nicolaou, M.A., Gunes, H., Pantic, M.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. Affective Computing, IEEE Transactions on **2**(2), 92–105 (2011)

[157] Nicolle, J., Rapp, V., Bailly, K., Prevost, L., Chetouani, M.: Robust continuous prediction of human emotions using multiscale dynamic cues. In: Proceedings of the 14th ACM international conference on Multimodal interaction, pp. 501–508. ACM (2012)

[158] Pahl, J., Rieger, I., Seuss, D.: Multi-label class balancing algorithm for action unit detection. arXiv preprint arXiv:2002.03238 (2020)

[159] Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010)

[160] Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pp. 5–pp. IEEE (2005)

[161] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC, vol. 1, p. 6 (2015)

[162] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301. Ieee (2009)

[163] Peng, S., Zhang, L., Ban, Y., Fang, M., Winkler, S.: A deep network for arousal-valence emotion prediction with acoustic-visual cues. arXiv preprint arXiv:1805.00638 (2018)

[164] Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003 Papers, SIGGRAPH '03, pp. 313–318. ACM, New York, NY, USA (2003). DOI 10.1145/1201775.882269. URL http://doi.acm.org/10.1145/1201775.882269

[165] Pham, H.X., Wang, Y., Pavlovic, V.: Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network. arXiv preprint arXiv:1803.07716 (2018)

[166] Plutchik, R.: Emotion: A psychoevolutionary synthesis. Harpercollins College Division (1980)

[167] Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 818–833 (2018)

[168] Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., Wang, H.: Geometry-contrastive gan for facial expression transfer. arXiv preprint arXiv:1802.01822 (2018)

[169] Ramirez, G.A., Baltrušaitis, T., Morency, L.P.: Modeling latent discriminative dynamic of multi-dimensional affective signals. In: International Conference on Affective Computing and Intelligent Interaction, pp. 396–406. Springer (2011)

[170] Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 17–24. IEEE (2017)

[171] Reed, S., Sohn, K., Zhang, Y., Lee, H.: Learning to disentangle factors of variation with manifold interaction. In: International Conference on Machine Learning, pp. 1431–1439 (2014)

[172] Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmi, M., Pantic, M.: Avec 2017–real-life depression, and affect recognition workshop and challenge (2017)

[173] Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pp. 1–8. IEEE (2013)

[174] Rothe, R., Timofte, R., Van Gool, L.: Dex: Deep expectation of apparent age from a single image. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 10–15 (2015)

[175] Ruiz, A., Van de Weijer, J., Binefa, X.: From emotions to action units with hidden and semi-hidden-task learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3703–3711 (2015)

[176] Russell, J.A.: Evidence of convergent validity on the dimensions of affect. Journal of personality and social psychology **36**(10), 1152 (1978)

[177] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520 (2018)

[178] Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **37**(6), 1113–1133 (2015)

[179] Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: European Workshop on Biometrics and Identity Management, pp. 47–56. Springer (2008)

[180] Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and vision Computing **27**(6), 803–816 (2009)

[181] Shang, F., Liu, Y., Cheng, J., Cheng, H.: Robust principal component analysis with missing data. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14, pp. 1149–1158. ACM, New York, NY, USA (2014). DOI 10.1145/2661829.2662083. URL http://doi.acm.org/10.1145/2661829.2662083

[182] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[183] Sneddon, I., McRorie, M., McKeown, G., Hanratty, J.: The belfast induced natural emotion database. IEEE Transactions on Affective Computing **3**(1), 32–41 (2012)

[184] Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, pp. 3483–3491 (2015)

[185] Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing **3**(1), 42–55 (2012)

[186] Song, L., Lu, Z., He, R., Sun, Z., Tan, T.: Geometry guided adversarial facial expression synthesis. In: 2018 ACM Multimedia Conference on Multimedia Conference, pp. 627–635. ACM (2018)

[187] Soyel, H., Demirel, H.: Facial expression recognition based on discriminative scale invariant feature transform. Electronics letters **46**(5), 343–345 (2010)

[188] Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems, pp. 1988–1996 (2014)

[189] Susskind, J.M., Hinton, G.E., Movellan, J.R., Anderson, A.K.: Generating facial expressions with deep belief nets. In: Affective Computing. InTech (2008)

[190] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017)

[191] Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)

[192] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: ICANN (2018)

[193] Tang, C., Zheng, W., Yan, J., Li, Q., Li, Y., Zhang, T., Cui, Z.: View-independent facial action unit detection. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 878–882. IEEE (2017)

[194] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)

[195] Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., Nießner, M.: Headon: real-time reenactment of human portrait videos. ACM Transactions on Graphics (TOG) **37**(4), 164 (2018)

[196] Thomaz, C.E., Giraldi, G.A.: A new ranking method for principal components analysis and its application to face image analysis. Image and Vision Computing **28**(6), 902–913 (2010)

[197] Tian, Y.l., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23**(2), 97–115 (2001)

[198] Tom, N.L.S.B.L., et al.: Psychological and biological approaches to emotion. Psychology Press (1990)

[199] Triantafyllopoulos, A., Sagha, H., Eyben, F., Schuller, B.: audeering's approach to the one-minute-gradual emotion challenge. arXiv preprint arXiv:1805.01222 (2018)

[200] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M.: Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 3–10. ACM (2016)

[201] Valstar, M., Pantic, M.: Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, p. 65 (2010)

[202] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 3–10. ACM (2013)

[203] Valstar, M.F., Almaev, T., Girard, J.M., McKeown, G., Mehu, M., Yin, L., Pantic, M., Cohn, J.F.: Fera 2015-second facial expression recognition and analysis challenge. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, vol. 6, pp. 1–8. IEEE (2015)

[204] Valstar, M.F., Sánchez-Lozano, E., Cohn, J.F., Jeni, L.A., Girard, J.M., Zhang, Z., Yin, L., Pantic, M.: Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 839–847. IEEE (2017)

[205] Vielzeuf, V., Pateux, S., Jurie, F.: Temporal multimodal fusion for video emotion classification in the wild. arXiv preprint arXiv:1709.07200 (2017)

[206] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018)

[207] Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312**, 135–153 (2018)

[208] Wang, S., Fidler, S., Urtasun, R.: Holistic 3d scene understanding from a single geo-tagged image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3964–3972 (2015)

[209] Wang, S., Gan, Q., Ji, Q.: Expression-assisted facial action unit recognition under incomplete au annotation. Pattern Recognition **61**, 78–91 (2017)

[210] Wang, X., Peng, M., Pan, L., Hu, M., Jin, C., Ren, F.: Two-level attention with two-stage multi-task learning for facial emotion recognition. arXiv preprint arXiv:1811.12139 (2018)

[211] Wang, Z., He, K., Fu, Y., Feng, R., Jiang, Y.G., Xue, X.: Multi-task deep neural network for joint face recognition and facial attribute prediction. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 365–374. ACM (2017)

[212] Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision, pp. 499–515. Springer (2016)

[213] Wheeler, M.D., Ikeuchi, K.: Iterative estimation of rotation and translation using the quaternion. Carnegie-Mellon University. Department of Computer Science (1995)

[214] Whissel, C.: The dictionary of affect in language, emotion: Theory, research and experience: vol. 4, the measurement of emotions, r. Plutchik and H. Kellerman, Eds., New York: Academic (1989)

[215] Whissell, C.M.: The dictionary of affect in language. In: The measurement of emotions, pp. 113–131. Elsevier (1989)

[216] Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Transactions on Signal Processing **57**(7), 2479–2493 (2009). DOI 10.1109/TSP.2009.2016892

[217] Wu, W., Zhang, Y., Li, C., Qian, C., Change Loy, C.: Reenactgan: Learning to reenact faces via boundary transfer. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 603–619 (2018)

[218] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500 (2017)

[219] Yang, J., Wu, S., Wang, S., Ji, Q.: Multiple facial action unit recognition enhanced by facial expressions. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 4089–4094. IEEE (2016)

[220] Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[221] Yao, A., Cai, D., Hu, P., Wang, S., Sha, L., Chen, Y.: Holonet: towards robust emotion recognition in the wild. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 472–478 (2016)

[222] Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3d dynamic facial expression database. In: Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On, pp. 1–6. IEEE (2008)

[223] Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on, pp. 211–216. IEEE (2006)

[224] YouTube, L.: Youtube. Retrieved **27**, 2011 (2011)

[225] Yüce, A., Gao, H., Thiran, J.P.: Discriminant multi-label manifold embedding for facial action unit detection. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, pp. 1–6. IEEE (2015)

[226] Zafeiriou, S., Kollias, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-wild: Valence and arousal 'in-the-wild'challenge. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pp. 1980–1987. IEEE (2017)

[227] Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)

[228] Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3712–3722 (2018)

[229] Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**(1), 39–58 (2009)

[230] Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016). DOI 10.1109/LSP.2016.2603342

[231] Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32**(10), 692–706 (2014)

[232] Zhang, X., Zhang, L., Wang, X.J., Shum, H.Y.: Finding celebrities in billions of web images. IEEE Transactions on Multimedia **14**(4), 995–1007 (2012)

[233] Zhang, Y., Yang, Q.: A survey on multi-task learning. arXiv preprint arXiv:1707.08114 (2017)

[234] Zhang, Y.H., Huang, R., Zeng, J., Shan, S., Chen, X.: m^3 t: Multi-modal continuous valence-arousal estimation in the wild. arXiv preprint arXiv:2002.02957 (2020)

[235] Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3438–3446 (2016)

[236] Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence **29**(6), 915–928 (2007)

[237] Zhao, S., Cai, H., Liu, H., Zhang, J., Chen, S.: Feature selection mechanism in cnns for facial expression recognition. In: BMVC, p. 317 (2018)

[238] Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., Yan, S.: Peak-piloted deep network for facial expression recognition. In: European conference on computer vision, pp. 425–442. Springer (2016)

[239] Zheng, Z., Cao, C., Chen, X., Xu, G.: Multimodal emotion recognition for one-minute-gradual emotion challenge. arXiv preprint arXiv:1805.01060 (2018)

[240] Zhi, R., Flierl, M., Ruan, Q., Kleijn, W.B.: Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **41**(1), 38–52 (2010)

[241] Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.N.: Learning active facial patches for expression analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2562–2569. IEEE (2012)

[242] Zhou, Y., Shi, B.E.: Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 370–376. IEEE (2017)

[243] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232 (2017)

[244] Zhu, X., Liu, Y., Li, J., Wan, T., Qin, Z.: Emotion classification with data augmentation using generative adversarial networks. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 349–360. Springer (2018)