ABSTRACT

Title of dissertation: FACIAL EXPRESSION RECOGNITION AND EDITING WITH LIMITED DATA

Hui Ding
Doctor of Philosophy, 2020

Dissertation directed by: Professor Rama Chellappa
Department of Electrical and Computer Engineering

Over the past five years, methods based on deep features have taken over the computer vision field. While dramatic performance improvements have been achieved for tasks such as face detection and verification, these methods usually need large amounts of annotated data. In practice, not all computer vision tasks have access to large amounts of annotated data. Facial expression analysis is such a task. In this dissertation, we focus on facial expression recognition and editing problems with small datasets. In addition, to cope with challenging conditions like pose and occlusion, we also study unaligned facial attribute detection and occluded expression recognition problems.

This dissertation has been divided into four parts. In the first part, we present FaceNet2ExpNet, a novel idea to train a light-weight and high accuracy classification model for expression recognition with small datasets. We first propose a new distribution function to model the high-level neurons of the expression network. Based on this, a two-stage training algorithm is carefully designed. In the pre-training

stage, we train the convolutional layers of the expression net, regularized by the face net; In the refining stage, we append fully-connected layers to the pre-trained convolutional layers and train the whole network jointly. Visualization shows that the model trained with our method captures improved high-level expression semantics. Evaluations on four public expression databases demonstrate that our method achieves better results than state-of-the-art.

In the second part, we focus on robust facial expression recognition under occlusion and propose a landmark-guided attention branch to find and discard corrupted feature elements from recognition. An attention map is first generated to indicate if a specific facial part is occluded and guide our model to attend to the non-occluded regions. To further increase robustness, we propose a facial region branch to partition the feature maps into non-overlapping facial blocks and enforce each block to predict the expression independently. Depending on the synergistic effect of the two branches, our occlusion adaptive deep network significantly outperforms state-of-the-art methods on two challenging in-the-wild benchmark datasets and three real-world occluded expression datasets.

In the third part, we propose a cascade network that simultaneously learns to localize face regions specific to attributes and performs attribute classification without alignment. First, a weakly-supervised face region localization network is designed to automatically detect regions (or parts) specific to attributes. Then multiple part-based networks and a whole-image-based network are separately constructed and combined together by the region switch layer and attribute relation layer for final attribute classification. A multi-net learning method and hint-based

model compression are further proposed to get an effective localization model and a compact classification model, respectively. Our approach achieves significantly better performance than state-of-the-art methods on unaligned CelebA dataset, reducing the classification error by 30.9%

In the final part of this dissertation, we propose an Expression Generative Adversarial Network (ExprGAN) for photo-realistic facial expression editing with controllable expression intensity. An expression controller module is specially designed to learn an expressive and compact expression code in addition to the encoder-decoder network. This novel architecture enables the expression intensity to be continuously adjusted from low to high. We further show that our ExprGAN can be applied for other tasks, such as expression transfer, image retrieval, and data augmentation for training improved face expression recognition models. To tackle the small size of the training database, an effective incremental learning scheme is proposed. Quantitative and qualitative evaluations on the widely used Oulu-CASIA dataset demonstrate the effectiveness of ExprGAN.

FACIAL EXPRESSION RECOGNITION AND EDITING WITH LIMITED DATA

by

Hui Ding

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Rama Chellappa, Chair/Advisor
Professor Gang Qu
Professor Ming Wu
Professor Behtash Babadi
Professor Ramani Duraiswami

Dedication

I dedicate this thesis to my family, my husband, Weisheng, and my lovely daughter, Eve for their constant support and unconditional love.

Acknowledgments

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

First and foremost I'd like to thank my advisor, Professor Rama Chellappa for his invaluable guidance through each stage of the process. He's the funniest advisor and one of the smartest people I know. I am immensely grateful to him for bringing me into the computer vision field. He has always been supportive and has given me the freedom to pursue various projects without objection. He has also provided insightful discussions about the research and constantly inspired me by his hardworking and passionate attitude. I have learned a great deal about the qualities needed to be a successful individual by observing him over the past six years. He was and remains my best role model for a great researcher, an excellent mentor, and a wonderful human being.

Next, I would like to thank Dr. Kevin Zhou who mentored me during my internships at Siemens. I appreciate all his contributions of time and ideas to make the internship experience productive and stimulating. I would also like to thank Dr. Kumar Sricharan for his research inputs during my internship at Palo Alto Research Center.

It is an honor to have Professor Gang Qu, Professor Min Wu, Professor Behtash Babadi and Professor Ramani Duraiswami in my dissertation committee. I am thankful to them for serving in my committee and providing invaluable advice to

improve this dissertation.

I would like to thank all my colleagues, roommates and friends for making my graduate life memorable. I am especially grateful to my labmate Dr. Mohammed Fathy for his invaluable advice and constant encouragement during the final stages of my doctoral studies. I would also like to thank the staff in UMIACS, ECE and Cfar for helping me in various different ways during my graduate life.

I owe my deepest gratitude to my parents and husband for always believing in me and encouraging me to follow my dreams. This dissertation would not have been possible without their warm love, continued patience, and endless support.

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA RD Contract No. 2019-022600002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Table of Contents

Dedication ii

Acknowledgements iii

Table of Contents v

List of Tables viii

List of Figures ix

1 Introduction 1
1.1 Motivation . . . 1
1.2 Proposed Algorithms and their Contributions . . . 3
1.3 Organization . . . 7

2 Transferring Knowledge from Face Recognition Network for Facial Expression Recognition 8
2.1 Introduction . . . 8
2.2 Related Works . . . 12
2.3 Approach . . . 14
2.3.1 Motivation . . . 14
2.3.2 Training Algorithm . . . 16
2.3.3 Which Layer to Transfer? . . . 18
2.4 Experiments . . . 20
2.4.1 Implementation . . . 22
2.4.2 Neuron Visualization . . . 23
2.4.3 CK+ . . . 23
2.4.4 Oulu-CAS VIS . . . 25
2.4.5 TFD . . . 27
2.4.6 SFEW . . . 28
2.4.7 RAF . . . 29
2.4.8 Ultrasound Abdomen Dataset . . . 32
2.5 Expression Feature Analysis . . . 35
2.6 Computational speed analysis . . . 36
2.7 Conclusions . . . 36

3 Occlusion Adaptive Deep Network for Robust Facial Expression Recognition 37
3.1 Introduction . . . 37
3.2 Related Work . . . 40
3.2.1 Deep Facial Expression Recognition . . . 40
3.2.2 Occlusive Facial Expression Recognition . . . 41
3.3 Occlusion Adaptive Deep Network . . . 42
3.3.1 Landmark-guided Attention Branch . . . 43
3.3.2 Facial Region Branch . . . 45
3.3.3 Relationship between the Two Branches . . . 46
3.4 Experiments . . . 47
3.4.1 Datasets . . . 47
3.4.2 Implementation Details . . . 48
3.4.3 Results Comparison . . . 49
3.4.4 Ablation Study . . . 55
3.4.5 Visualization . . . 57
3.5 Conclusions . . . 57

4 A Deep Cascade Network for Unaligned Face Attribute Classification 59
4.1 Introduction . . . 59
4.2 Related Works . . . 63
4.3 Proposed Method . . . 64
4.3.1 Face Region Localization (FRL) Network . . . 65
4.3.1.1 Multi-Net Learning . . . 66
4.3.2 Attribute Classification Network . . . 67
4.3.2.1 Parts and Whole (PaW) Classification Network . . . 67
4.3.2.2 Hint-based Model Compression . . . 68
4.3.3 Training Methodology . . . 70
4.4 Experiments . . . 71
4.4.1 Dataset . . . 71
4.4.2 Implementation details . . . 72
4.4.3 Ablative Analysis . . . 72
4.4.3.1 Face Region Localization . . . 72
4.4.3.2 Multi-Net Learning . . . 74
4.4.3.3 Hint-based Model Compression . . . 75
4.4.3.4 PaW Classification Network . . . 77
4.5 Conclusions . . . 80

5 Facial Expression Editing with Controllable Expression Intensity 81
5.1 Introduction . . . 81
5.2 Related Works . . . 84
5.2.1 Deep Generative Model . . . 84
5.2.2 Facial Expression Editing . . . 85
5.3 Proposed Method . . . 86
5.3.1 Conditional Generative Adversarial Network . . . 87
5.3.2 Adversarial Autoencoder . . . 87

5.3.3 Expression Generative Adversarial Network 88
5.3.3.1 Network Architecture 88
5.3.3.2 Expression Controller Networks $\boldsymbol{F_{ctrl}}$ and $\boldsymbol{Q}$ 89
5.3.3.3 Generator Network $\boldsymbol{G}$ 91
5.3.3.4 Discriminator on Identity Representation $\boldsymbol{D_z}$ 92
5.3.3.5 Discriminator on Image $\boldsymbol{D_{img}}$ 92
5.3.3.6 Overall Objective Function 93
5.3.3.7 Incremental Training 93
5.4 Experiments 94
5.4.1 Dataset 94
5.4.2 Implementation Details 95
5.4.3 Facial Expression Editing 97
5.4.4 Facial Expression Transfer 98
5.4.5 Face Image Generation for Data Augmentation 99
5.4.6 Feature Visualization 101
5.5 Conclusions 102

6 Conclusions and Directions for Future Work 104
6.1 Summary 104
6.2 Directions for Future Work 106

List of Tables

2.1 The number of low expressive score neurons for pre-trained network and fine-tuned network . . . 19
2.2 The number of images for different expression classes . . . 21
2.3 The number of images for different organ classes in the ultrasound abdomen dataset . . . 21
2.4 The Average Accuracy on CK+ dataset (Person-Independent) . . . 25
2.5 The Average Accuracy on Oulu-CAS dataset (Person-Independent) . . . 26
2.6 The Average Accuracy on TFD dataset (Person-Independent) . . . 27
2.7 The Average Accuracy on SFEW dataset (Person-Independent) . . . 29
2.8 The Average Accuracy on RAF dataset (Person-Independent) . . . 31
2.9 The Average Accuracy on Ultrasound Abdomen dataset . . . 35

3.1 Test Set Accuracy on RAF dataset . . . 49
3.2 Validation Set Accuracy on AffectNet dataset . . . 50
3.3 Validation Set Accuracy on Occlusion-AffectNet and Pose-AffectNet dataset . . . 51
3.4 Test Set Accuracy on FED-RO dataset . . . 53
3.5 Test Set Accuracy on Occlusion-FERPlus and Pose-FERPlus dataset 55

4.1 Average classification accuracy on uCelebA dataset. . . . 74
4.2 Fine-grained classification accuracy on CUB-200 dataset. . . . 74
4.3 Comparison of average accuracy and compactness between different compressed models on uCelebA dataset. . . . 76
4.4 Comparison of average accuracy and compactness on the aligned CelebA dataset. . . . 76
4.5 Performance comparison with state of the art methods on 40 binary facial attributes. The best results are shown in bold. . . . 78

5.1 Comparison of expression recognition accuracy with different numbers of synthesized images. . . . 101

List of Figures

2.1 The red-boxed images are generated by the model trained with our FaceNet2ExpNet method, while the black-boxed images are from the face network fine-tuned on the expression dataset. We can see the images produced by the face net are dominated with faces, while our model represents the facial expressions better. Models are visualized by DeepDraw [1]. 9

2.2 Two-stage Training Algorithm. In stage (a), the face net is frozen and provides supervision for the expression net. The regression loss is backpropped only to the expression net. The convolutional layers are trained in this stage. In stage (b), the randomly initialized fully-connected layers are attached to the trained convolutional blocks. The whole network is trained jointly with cross-entropy loss. The face net is normally much deeper than the expression net. 16

2.3 Histograms of neuron entropy scores from four different layers for pre-trained network (red) and fine-tuned network (blue). The X axis is the entropy value and the Y axis is the number of neurons. The first row plots are for the CK+ dataset, while the plots in the second row are for the Oulu-CASIA dataset. 20

2.4 Visualizes several neurons in the top hidden layer of our model on CK+ dataset. 21

2.5 Visualizes several neurons in the top hidden layer of our model on Oulu-CASIA dataset. 21

2.6 Confusion Matrix of CK+ for the Eight Classes problem. The darker the color, the higher the accuracy. 24

2.7 Confusion Matrix of Oulu-CASIA. The darker the color, the higher the accuracy. 26

2.8 Confusion Matrix of TFD. The darker the color, the higher the accuracy. 28

2.9 Confusion Matrix of SFEW. The darker the color, the higher the accuracy. 30

2.10 Confusion Matrix of RAF. The darker the color, the higher the accuracy. 31

2.11 Attention maps of three different methods. The images' expression labels are displayed on the leftmost. The left, middle and right columns show the predictions of networks train from scratch, fine-tune from FaceNet and FN2EN. A deep red denotes high attention. 33

2.12 Sample images from the Ultrasound Abdomen Dataset. 34
2.13 (**left**) The conv1 filters learned by training AlexNet from scratch on the abdomen dataset. (**right**) The conv1 filters learned with our method. Our model learns gabor-like conv1 filters. 34

3.1 Pipeline of the Occlusion Adaptive Deep Network. It consists of two branches: a Landmark-guided Attention Branch and a Facial Region Branch. The ResNet50 backbone is shared between the two branches to extract the global features. For the Landmark-guided Attention Branch, the facial landmarks are first detected. Then the interested points are computed to cover the most informative facial areas. The confidence scores of these points are further utilized to generate the attention maps, guiding the model to attend to the visible facial components. While for the Facial Region Branch, the feature maps are divided into non-overlapping facial blocks and each block is trained to be a discriminative expression classifier on its own. 42
3.2 We select 16 points from the original 68 landmarks (a) to cover the regions around eyes, eyebrows, nose and mouth. We further recompute 8 points to cover facial cheeks and the areas between eyes and eyebrows. 43
3.3 The interest points with confidence scores greater than the threshold T are shown in red points. We can see the occluded facial areas are removed. 47
3.4 Confusion Matrix of RAF-DB. The darker the color, the higher the accuracy. 51
3.5 Confusion Matrix of Affectnet. The darker the color, the higher the accuracy. 52
3.6 Confusion Matrix of FED-RO. The darker the color, the higher the accuracy. 54
3.7 The impacts of the confidence threshold T, number of regions K and the loss combination weight λ on the performance of OADN. 56
3.8 Comparison of the gACNN method and our OADN method on the FED-RO dataset. Red and green texts indicate the error and correct predictions. 57

4.1 Overview of our face attribute recognition framework. It consists of a facial region localization (FRL) network and a Parts and Whole (PaW) classification network. The localization network detects a discriminative part for each attribute. Then the detected face regions and the whole face image are fed into the PaW classification network. The region switch layer (RSL) selects the relevant subnet for predicting the attribute, while the attribute relation layer (ARL) models the attribute relationships. 62
4.2 face CNN resp 68
4.3 face CNN resp 71

4.4 Location heatmaps from the face region localization network. Face regions that correlate with facial attributes are discovered. 73
4.5 Visualization of the region switch layer weights. For each attribute, the blue and the red bar represent the weight values of RSL that corresponds to the part-based subnet and whole-image-based subnet respectively. It shows that the weights of the part-based subnets are higher for the local attributes. For global attributes, the whole-image-based subnet is assigned larger weight. 75
4.6 Attribute relation weights learned on uCelebA dataset. Red and yellow colors indicate high values while blue and green colors denote low values. 77

5.1 Comparison of previous GAN architectures and the proposed ExprGAN. 84
5.2 Visual comparison of facial expression editing results. For each input, we compare the ground truth images (top), the synthetic images of ExprGAN (middle) and CAAE (bottom). Zoom in for details. . . . 96
5.3 Face images are transformed to new expressions with different intensity levels. The top row contains the input faces with the original expressions, and the remaining rows show the synthesized results. Each column corresponds to a new expression with five intensity levels from weak to strong. The *Neutral* expression which is not in the training data is also generated. 98
5.4 Facial expression transfer. Expressions from the middle column are transferred to faces in the left column. The results are shown in the right column. 99
5.5 Random generated subjects displaying six categories of expressions. . 100
5.6 Identity feature space. Each color represents a different identity and the images for one identity are labeled. 102
5.7 Expression-based image retrieval. First column shows query images. Other columns show top one retrieval based on c, y and x. 103

Chapter 1: Introduction

1.1 Motivation

Facial expression plays an important role in social communication during our daily life. In recent years, automatically recognizing and editing expression have received increasing attention due to their numerous applications. Facial expression recognition is useful for driver safety, health care, video conferencing, virtual reality, cognitive science *etc.* Similarly, facial expression editing has applications in facial animation, human-computer interactions, entertainment, *etc.*

While Deep Convolutional Neural Networks (DCNN) have demonstrated impressive performance improvements for many problems in computer vision, one of the most important reasons behind its success is the availability of large-scale training databases. However, it is not uncommon to have small datasets in many application areas, facial expression recognition being one of them. With a relatively small set of training images, even when regularization techniques such as Dropout [2] and Batch Normalization [3] are used, the results are not satisfactory. Motivated by this, we propose FaceNet2ExpNet, a novel learning algorithm that incorporates face domain knowledge to regularize the training of an expression recognition network.

Although high accuracy classifiers have been obtained on datasets captured

in controlled environments, such as CK+ [4], MMI [5] and OULU-CASIA [6], they perform poorly when recognizing facial expressions under natural and uncontrollable variations like pose, illumination, and occlusion. Among all these factors, occlusion has been considered a highly challenging one. Previous works [7, 8] learn the importance weights for multiple facial regions. However, the self-attention based methods lack additional supervision information required to ensure the functionality. Thus, the network may not be able to locate these non-occluded facial regions accurately under large occlusions and poses. Motivated by this, we propose an Occlusion Adaptive Deep Network to overcome the occlusion problem for robust facial expression recognition in-the-wild.

Face attributes describe the characteristics observed from a face image. They include both identity-related attributes such as oval face and non-identity-related attributes like facial expression. Despite their wide applications, face attribute recognition is not an easy task. One reason is that recognizing different face attributes may require attentions to different regions of the face [9, 10]. For example, local attributes like *Mustache* could be recognized by just checking the region containing the mouth. Other parts of the face do not provide useful information and may even hamper this particular attribute recognition. However, recognizing global attributes like *Pale Skin* may require information from the whole face region. Motivate by this, we propose a learning-based method that dynamically selects different face regions for unaligned face attribute prediction.

Models based on generative adversarial networks (GAN) [11] have achieved great success for face synthesis over the past five years. Starting from DCGAN [12]

to StyleGAN [13], the generated images have higher resolution and quality. However, for facial expression generation, due to the small scale of expression datasets, GAN-based models are relatively unexplored. The synthesized images from existing methods have low resolution (48 x 48), lacking fine details and tend to be blurry. Moreover, these approaches can only transform the expression to different classes, like *Angry* or *Happy*. However, in reality, the intensity of facial expression is often displayed over a range. Motivated by this, we present a new expression editing model, Expression Generative Adversarial Network (ExprGAN) which has the unique property that multiple diverse styles of the target expression can be synthesized where the intensity of the generated expression can be continuously controlled from weak to strong, without the need for training data with intensity values.

1.2 Proposed Algorithms and their Contributions

In this section, we briefly describe the algorithms introduced in this dissertation and their key contributions.

1. **Transferring Knowledge from Face Recognition Network for Facial Expression Recognition:**

   In this part of the dissertation, we try to answer the following basic question: How to obtain a light-weight and high accuracy classification model for expression recognition with small datasets? Popular transfer learning methods utilize face recognition datasets to pre-train the network, which is then fine-tuned on the expression datasets. Although this strategy performs adequately, it has

two notable problems: (i) the fine-tuned face net may still contain information useful for subject identification. (ii) the network designed for the face recognition domain is often too big for the expression task, thus the overfitting issue is still severe. In this part of the dissertation, we address this issue by proposing a novel learning algorithm that incorporates face domain knowledge to regularize the training of an expression recognition network. Specifically, we propose a new distribution function to model the high-level neurons of the expression net using information derived from the fine-tuned face net. Such modeling naturally leads to a regression loss which serves as feature-level regularization that pushes the intermediate features of the expression net to be close to those of the fine-tuned face net. Next, to further improve the discriminativeness of the learned features, we refine the network with strong supervision from the label information. Experimental results show that the proposed method improves visual feature representation and outperforms various state-of-the-art methods on four public datasets.

2. **Occlusion Adaptive Deep Network for Robust Facial Expression Recognition:**

   In this part of the dissertation, we try to answer the following important question: How to achieve accurate facial expression recognition when faces are partially occluded? Previous expression recognition methods, either overlook this issue or resolve it based on extreme assumptions. In this part of the dissertation, we address this issue by proposing an Occlusion Adaptive Deep

Network to overcome the occlusion problem for robust facial expression recognition in-the-wild. It consists of two branches: a landmark-guided attention branch and a facial region branch. The landmark-guided attention branch discards feature elements that have been corrupted by occlusions and guides the model to focus on the non-occluded facial regions. To further enforce the robustness and learn complementary context information, we introduce a facial region branch to train multiple region-based expression classifiers. Experimental results on five challenging benchmark datasets show that our method obtains significantly better performance than existing methods.

3. **A Deep Cascade Network for Unaligned Face Attribute Classification:**

   In this part of the dissertation, we try to answer the following important question: How to classify facial attributes without face alignment? Inspired by the observation that humans focus attention on different face regions when recognizing face attributes, we propose a learning-based method that dynamically selects different face regions for unaligned face attribute prediction. It integrates two networks using a cascade: a face region localization network followed by an attribute classification network. The localization network detects face areas specific to attributes, especially those that have local spatial support. The classification network selectively leverages information from these face regions to make the final prediction. We show that with no use of alignment information, our method reduces the classification error by a significant

margin of compared with state-of-the-art. We also show the designed model could select the most relevant face region for predicting each face attribute.

4. **Facial Expression Editing with Controllable Expression Intensity:**

   In conventional methods, either paired training data is required or the synthetic face's resolution is low. Moreover, only the categories of facial expression can be changed. In this part of the dissertation, we address this issue by proposing a novel model called Expression Generative Adversarial Network (ExprGAN) that can change a face image to a target expression with multiple styles, where the expression intensity can also be controlled continuously. Our ExprGAN adopts the generator and discriminator framework in addition to the expression controller module and the regularizer network. To facilitate image editing, the generator is composed of an encoder and a decoder. The input of the encoder is a face image, the output of the decoder is a reconstructed one, and the learned identity and expression representations bridge the encoder and decoder. To preserve the most prominent facial structure, we adopt a multi-layer perceptual loss [14] in the feature space in addition to the pixel-wise L_1 loss. Moreover, to make the synthesized image look more photo-realistic, two adversarial networks are imposed on the encoder and decoder, respectively. Because it is difficult to directly train our model using the small training set, a three-stage incremental learning algorithm is also developed. We show that the synthesized face images have high perceptual quality, which can be used to improve the performance of an expression classifier. We also show that the

identity and expression representations are explicitly disentangled which can be exploited for tasks such as expression transfer, image retrieval, *etc.*

1.3 Organization

This dissertation is organized as follows. Chapter 2 presents a transfer learning algorithm for facial expression recognition with small datasets. Chapter 3 presents an occlusion adaptive deep network for in-the-wild facial expression recognition. Chapter 4 presents a facial region localization network and a Parts and Whole classification network for unaligned facial attribute classification. Chapter 5 presents a GAN-based model that can transform the face image to have a new expression where the expression intensity is allowed to be controlled continuously. Chapter 6 concludes the dissertation and discusses future research directions.

Chapter 2: Transferring Knowledge from Face Recognition Network for Facial Expression Recognition

2.1 Introduction

Deep Convolutional Neural Networks (DCNN) have demonstrated impressive performance improvements for many problems in computer vision. One of the most important reasons behind its success is the availability of large-scale training databases, for example, ImageNet [15] for image classification, Places [16] for scene recognition, CompCars [17] for fine-grained recognition and MegaFace [18] for face recognition.

However, it is not uncommon to have small datasets in many applications, like facial expression recognition and medical image classification. With a relatively small set of training images, even when regularization techniques such as Dropout [2] and Batch Normalization [3] are used, the results are not satisfactory. The popular approach is to fine-tune a network that has been pre-trained on a large dataset. Because of the generalizability of the pre-learned features, this approach has achieved great success [19].

Motivated by this observation, several previous works [20, 21] on expression

recognition utilize face recognition datasets to pre-train the network, which is then fine-tuned on the expression dataset. The large amount of labeled face data [18, 22], makes it possible to train a fairly complicated and deep network. Moreover, the close relationship between the two domains facilites the transfer learning of features.

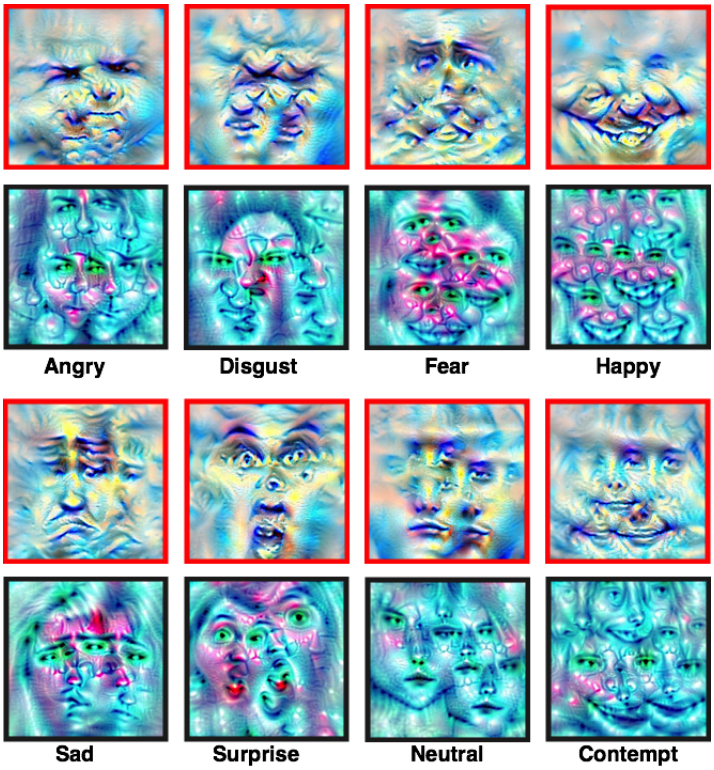


Figure 2.1: The red-boxed images are generated by the model trained with our FaceNet2ExpNet method, while the black-boxed images are from the face network fine-tuned on the expression dataset. We can see the images produced by the face net are dominated with faces, while our model represents the facial expressions better. Models are visualized by DeepDraw [1].

Although this strategy performs well, it has two notable problems: (i) the features are 'sticky', that is, the fine-tuned face net may still contain information useful for subject identification. This is because of the large size gap (several orders of magnititudes) between face and expression datasets. As we see from Fig. 2.1, the images (black-boxed) generated by the face net are dominated by faces as they

should, which weakens the network's ability to represent the different expressions. (ii) the network designed for the face recognition domain is often too big for the expression task, thus the overfitting issue is still severe.

In this chapter, we present FaceNet2ExpNet, a novel transfer learning algorithm that incorporates face domain knowledge to regularize the training of an expression recognition network. Specifically, we first propose a new distribution function to model the high-level neurons of the expression net using information derived from the fine-tuned face net. This strategy naturally leads to a regression loss which serves as feature-level regularization that pushes the intermediate features of the expression net to be close to those of the fine-tuned face net. Next, to further improve the discriminativeness of the learned features, we refine the network with strong supervision from the label information. We adopt a conventional network architecture, consisting of convolutional blocks followed by fully-connected layers, to design our expression net. The training is carried out in two stages: in the first stage, only the convolutional layers are trained. We utilize the deep features from the face net as the supervision signal to make the learning easier. It also contains meaningful knowledge about human faces, which is important for expression recognition, too. After the first stage of learning is completed, we add randomly initialized fully-connected (FC) layers and jointly train the whole network using the label information in the second stage. As observed by previous works [23], FC layers generally capture domain-specific semantics. So we only utilize the face net to guide the learning of the convolutional layers and the FC layers are trained from scratch. Moreover, we empirically find that late middle layer (*e.g.* pool5 for VGG-16 [24])

is more suitable for training supervision due to the richness of low entropy neurons. In both training stages, only expression images are used.

From Fig. 2.1, we observe that the models trained with our method capture the key properties of different expressions. For example, the angry expression is displayed by frowned eye brows and a closed mouth; the surprise expression is represented by a large opened mouth and eyes. This method is different from knowledge distillation [25]. Here we do not have a large accurate network trained on the same domain to produce reliable outputs from softmax. It is also different from FitNets [26], which is mainly used to train a thinner and deeper network.

To validate the effectiveness of our method, we perform experiments on both constrained (CK+, Oulu-CASIA, TFD) and unconstrained expression datasets (SFEW, RAF). For all the five datasets, we achieve better results than the state-of-the-art. Moreover, we also conduct experiments on an Ultrasound Abdomen dataset for anatomical organ classification.

Contributions: We propose a two-stage training algorithm to develop a light-weight and high accuracy classification model for expression recognition with limited data. Our method performs better than all previously published works on four datasets. This method is very general, and can be applied to other domains which are short of training samples. Moreover, it can also be used as a model compression method.

Organization: Section 2.2 briefly introduces related works. The FaceNet2ExpNet algorithm is presented in Section 2.3. Experimental results and analysis are discussed in Section 2.4, Section 2.5 and Section 2.6, respectively. We conclude this

chapter in Section 2.7.

2.2 Related Works

In [27], Zhong *et al.* observed that only a few active facial patches are useful for expression recognition. These active patches include: common patches for the recognition of all expressions and specific patches that are only important for a single expression. To locate these patches, a two-stage multi-task sparse learning framework is proposed. In the first stage, multi-task learning with group sparsity is performed to search for the common patches. In the second stage, face recognition is utilized to find the specific patches. However, the sequential search process is likely to find overlapped patches. To solve this problem, Liu *et al.* [28] integrated the sparse vector machine and multi-task learning into a unified framework. Instead of performing the patch selection in two separate phrases, an expression specific feature selection vector and a common feature selection vector are employed together. To get more discriminative features instead of hand-crafted features, Liu *et al.* [29] used a patch-based learning method. Subsequently, a group feature selection scheme based on maximal mutual information and minimal redundancy criterion is presented. Lastly, three layers of restricted Boltzman machines (RBM) are stacked to learn hierarchical features. To further boost performance, a loopy boosted deep belief network (DBN) framework was explored in [30]. Feature learning, feature selection and classifier design are learned jointly. In the forward phase, several DBNs extract features from the overlapped facial patches. Then, AdaBoosting is adopted

to combine these patch-based DBNs. In the fine-tuning phase, the loss from both weak and strong classifiers are backproped. In [31], to utilize the temporal information for video-based expression recognition, a 3D CNN was applied to learn the low-level features. Then, a GMM model is trained on the features, and the covariance matrix for each component composes the expressionlet. Motivated by the domain knowledge that facial expression can be decomposed into a combination of facial action units (AU), a deformable facial part model was explored in [32]. Multiple part filters are learned to detect the locations of discriminative facial parts. To further cope with pose and identity variations, a quadratic deformation cost is used.

More recently, Jung *et al.* [33] trained a deep temporal geometry network and a deep temporal appearance network with facial landmarks and images. To effectively fuse these two networks, a joint fine-tuning method is proposed. Specifically, the weight values are frozen and only the top layers are trained. In [34], Mollahosseini *et al.* discovered that the inception network architecture works very well for expression recognition task. Multiple cross dataset experiments are performed to show the generality of the learned model. In [35, 36], a two-step training procedure is suggested, where in the first step, the network was trained using a relatively large expression dataset followed by training on the target dataset. Even though the image is of low resolution and the label of the relatively large dataset is noisy, this approach is effective. The work closely related to ours is [21], which employed a peak expression image (easy sample) to help the training of a network with input from a weak expression image (hard sample). Although both works propose to use feature maps as supervision signals, our work is different in the following aspects:

First, for our training, we do not need a pair of same identity and same expression face images. We adopt a FaceNet to guide the training of the ExpNet, using as inputs only the face images from the expression dataset. Second, to better learn expression specific features, we train our network from scratch on the expression dataset, instead of fine-tuning on a network pretrained for face recognition.

2.3 Approach

2.3.1 Motivation

We write our expression net as:

$$O = H_{\theta_2}(G_{\theta_1}(I))$$

where H represents the fully connected layers, and G corresponds to the convolutional layers. θ_2 and θ_1 are the parameters to be learned. I is the input image, and O is the output before softmax.

First, the parameters θ_1 of the convolutional layers are learned. In [37], Xie et al. observed that the high-level neurons decay exponentially. To be more specific, by denoting the outputs of the l_{th} layer as $x_{c,w,h}$, and the average response value over the spatial dimension as

$$x_c = \frac{1}{W \times H} \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} x_{c,w,h} \tag{2.1}$$

where C is the number of output channels in the l_{th} layer, and W, H are the width and height of the response maps, respectively. Then the distribution function can be formulated as follows:

$$f(X^l) = C_p \cdot e^{-||X^l||_p^p} \tag{2.2}$$

where $X^l = [x_1, ..., x_C] \in R^C$, and C_p is a normalization constant. $||\cdot||_p^p$ is the L_p norm.

To incorporate the knowledge of a face net, we propose to extend Equation (2.2) to have the following form, *i.e.*, :

$$f(X^l) = C_p \cdot e^{-||X^l - \mu||_p^p} \tag{2.3}$$

The mean is modeled by the face net, $\mu = F(I)$. And F represents the face net's convolutional layers. This is motivated by the observation that the fine-tuned face net already achieves competitive performance on the expression dataset, so it should provide a good initialization point for the expression net. Thus, we do not want the latter to deviate much from the former.

Using the maximum likelihood estimation (MLE) procedure, we can derive the loss function as:

$$\begin{aligned} \max_{\theta_1} L_1 &= \max_{\theta_1} \log f(X^l) \\ &= \max_{\theta_1} \log C_p \cdot e^{-||X^l - \mu||} \\ &= \min_{\theta_1} ||G_{\theta_1}(I) - F(I)||_p^p \end{aligned} \tag{2.4}$$

Note that if $p = 2$ and without G, this is the normal l_2 regularizer. Thus we can also view the face net as acting like a regularizer, which stabilizes the training step of the expression net.

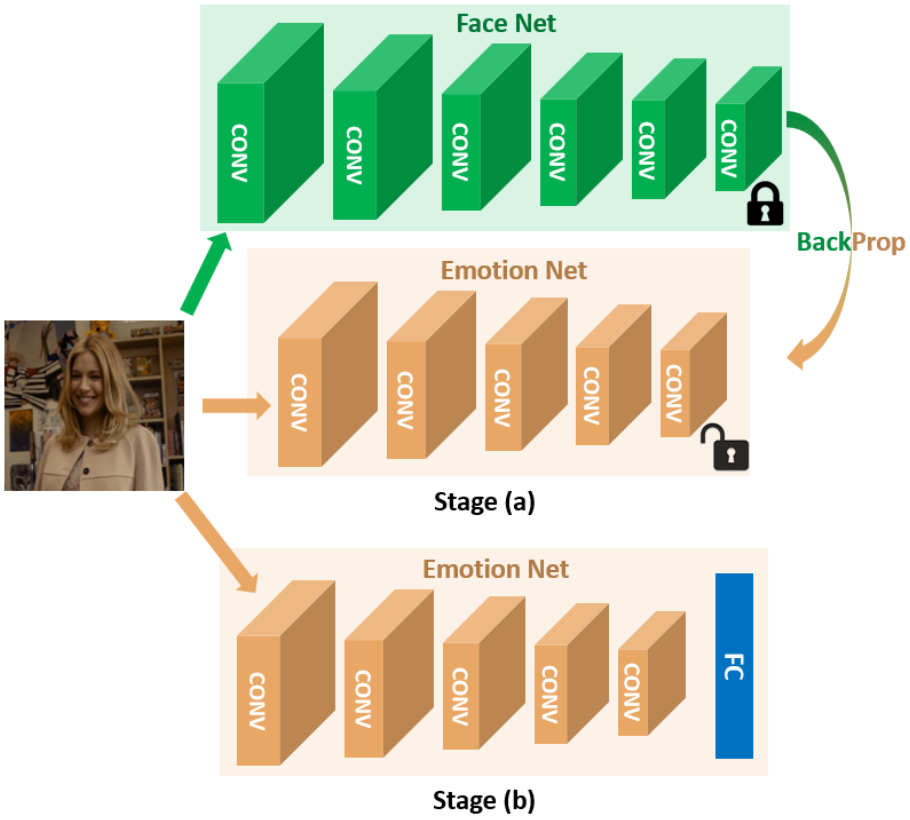


Figure 2.2: Two-stage Training Algorithm. In stage (a), the face net is frozen and provides supervision for the expression net. The regression loss is backpropped only to the expression net. The convolutional layers are trained in this stage. In stage (b), the randomly initialized fully-connected layers are attached to the trained convolutional blocks. The whole network is trained jointly with cross-entropy loss. The face net is normally much deeper than the expression net.

2.3.2 Training Algorithm

The training algorithm consists of the following two steps:

In the first stage, we train the convolutional layers using the loss function in Equation (2.4). The face net is frozen, and the outputs from the last pooling layer

are used to provide supervision for the expression net. We provide more explanations on this choice in the next section.

In the second stage, we append the fully connected layers to the trained convolutional layers. The whole network is jointly learned using the cross-entropy loss, defined as follows:

$$L_2 = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j} \log \hat{y}_{i,j}, \tag{2.5}$$

where $y_{i,j}$ is the ground truth for the image, and $\hat{y}_{i,j}$ is the predicated label. The complete training algorithm is illustrated in Fig. 2.2.

Our expression net consists of five convolutional layers, each followed by a non-linear activation function (ReLU) and a max-pooling layer. The kernel size of all the convolutional layers is a 3×3 window. For the pooling layer, it is 3×3 with stride 2. The numbers of the output channels are 64, 128, 256, 256, 512. After the last pooling layer, we add another 1×1 convolutional layer, which serves to bridge the gap between face and expression domains. Moreover, it also helps to adapt the dimension if the last pooling layer of the expression net does not match the face net. To reduce overfitting, we have only one fully-connected layer with dimension 256. Note, if the spatial size of the last pooling layer between the face net and expression net does not match exactly, then deconvolution (fractionally strided convolution) can be used for upsampling.

2.3.3 Which Layer to Transfer?

In this section, we explore the layer selection problem for the first stage supervision transfer. Since the fine-tuned face network outperforms the pre-trained network on expression recognition, we hypothesize that there may be interesting differences in the network before and after fine-tuning. These differences might help us understand better which layer is more suitable to guide the training of the expression network.

To this end, we first investigate the expression sensitivity of the neurons in the network, using VGG-16 as a working example. For each neuron, the images are ranked by the maximum response values. Then the top K ($K = 100$ in our experiments) images are binned according to the expression labels. We compute the entropy for the neuron x as $H(x) = -\sum_{i=1}^{n} p(i) \log p(i)$, where $p(i)$ denotes the histogram count for bin i and n denotes the number of quantized label bins (we normalize the histogram to a sum of 1). If the neuron has a low entropy, then it should be more expression-sensitive since its label distribution is peaky. To validate our assumption, we plot the histogram of the entropy for pool4, pool5, FC6 and FC7 layers. As shown in Fig. 2.3, the low-entropy neurons that are more expression-sensitive start to emerge in the pool5 layer *i.e.*, the blue and the red lines start to diverge. While for the pool4 and lower layers, there are few such high-level neurons, *i.e.*, the blue and the red lines almost overlap.

Since these low entropy neurons indicate layer discriminativeness, we next compute the number of low expressive score (LES) neurons for each layer (here

Table 2.1: The number of low expressive score neurons for pre-trained network and fine-tuned network

Model	Pool4	Pool5	FC6	FC7
Pre-trained (CK)	7763	2011	338	248
Fine-tuned (CK)	-57	+511	+658	+610
Pre-trained (Oulu-CASIA)	3009	605	48	33
Fine-tuned (Oulu-CASIA)	+194	+895	+952	+1086

low expressive score is the entropy lower than the minimum average entropy score among the four selected layers). In Table 2.1, we find that in comparison with the pre-trained network, the LES neurons increase dramatically in the fine-tuned network, especially starting from pool5 layer. For the CK+ dataset, the number of the low-entropy neurons in the pool4 layer is reduced, while for Oulu-CASIA, it increases only by 194. For the pool5 layer, it increases by 511 for CK+ and by 895 for Oulu-CASIA. Moreover, convolutional layers have a larger number of these neurons than FC layers. These results suggest that maybe late middle layer, such as pool5, is a good tradeoff between supervision richness and representation discriminativeness. For the first step of training, we would like the ExpNet to learn high-level expression semantics from the FaceNet, so we choose the pool5 layer. In the second step, in order to better adapt to the new domain task (facial expression recognition), we attach the fully-connected layer and train only with the expression labels.

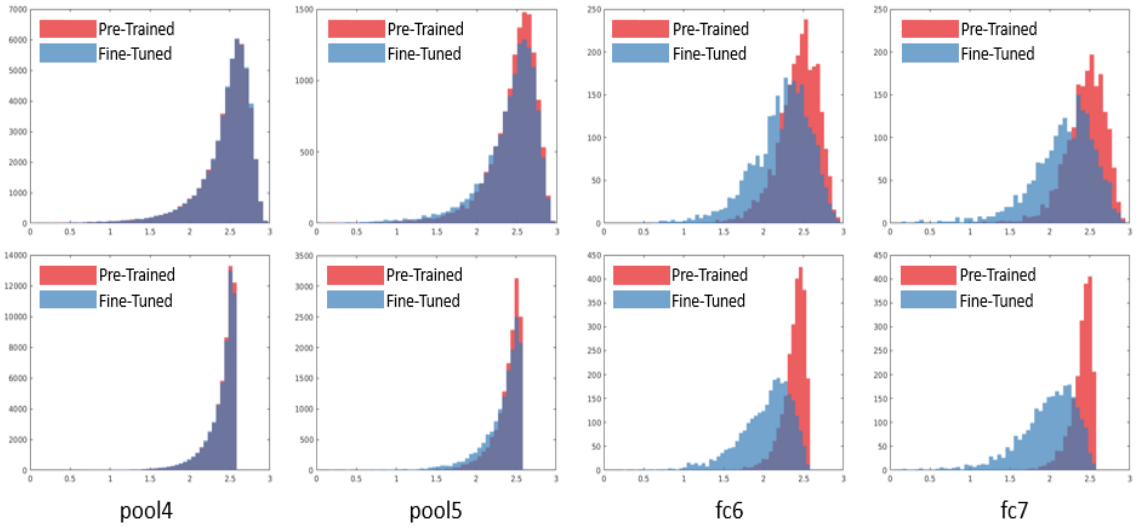


Figure 2.3: Histograms of neuron entropy scores from four different layers for pre-trained network (red) and fine-tuned network (blue). The X axis is the entropy value and the Y axis is the number of neurons. The first row plots are for the CK+ dataset, while the plots in the second row are for the Oulu-CASIA dataset.

2.4 Experiments

We validate the effectiveness of our method on five widely used expression databases: CK+ [4], Oulu-CASIA [6], Toronto Face Database (TFD) [38], Static Facial Expression in the Wild (SFEW) [39] and Real-world Affective Faces (RAF) [40]. The numbers of images for different expressions are shown in Table 2.2. To demonstrate the generality of our method, we also conduct one experiment on the Ultrasound (US) Abdomen dataset. The abdomen database contains US images for kidney, spleen and other anatomy organs. In total it has 131,000 frames from patients. The numbers of images for different organs are shown in Table 2.3. Some organs have four views, i.e., Left Transverse (LT), Left Longitudinal (LL), Right Transverse (RT) and Right Longitudinal (RL). In the following, we refer to our method FaceNet2ExpNet as FN2EN.

Table 2.2: The number of images for different expression classes

	An	Co	Di	Fe	Ha	Sa	Su	Ne	Total
CK+	135	54	177	75	147	84	249	327	1308
Oulu-CASIA	240		240	240	240	240	240		1444
TFD	437		457	424	758	441	459	1202	4178
SFEW	255		75	124	256	234	150	228	1322
RAF	705		717	281	4772	1982	1290	2524	12271

Table 2.3: The number of images for different organ classes in the ultrasound abdomen dataset

	Liver LT	Liver LL	Liver RT	Liver RL	Kidney LT	Kidney LL	Kidney RT	Kidney RL	Spleen Trans	Spleen Long	Aorta	Gallbladder	Iliac	IVC	Pancreas	Other
Train	9582	9015	9849	7671	10295	2827	9980	2840	3255	3115	20748	13747	10102	3228	8892	5787
Test	4354	4100	4645	3158	3555	615	4807	1285	1210	1065	8695	4754	3136	1172	4254	5411

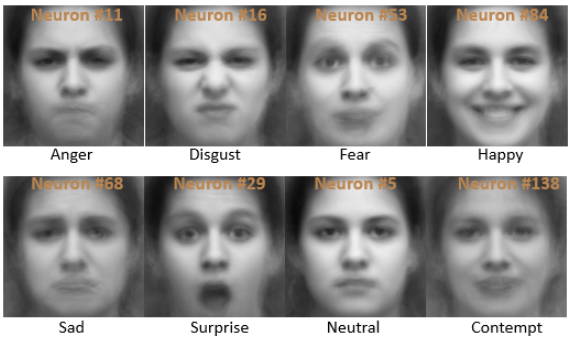


Figure 2.4: Visualizes several neurons in the top hidden layer of our model on CK+ dataset.

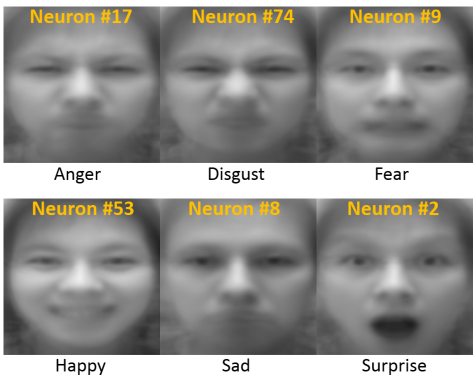


Figure 2.5: Visualizes several neurons in the top hidden layer of our model on Oulu-CASIA dataset.

2.4.1 Implementation

We apply Multi-task Cascade Convolutional Neural Networks (MTCNN) [41] for face detection and landmark detection. The faces are normalized, cropped, and resized to 256×256. We utilize conventional data augmentation in the form of random sampling and horizontal flipping. The min-batch size is 64, the momentum is fixed to be 0.9 and the dropout is set at 0.5.

For network training, in the first stage, the regression loss is very large. So we start with a very small learning rate 1e-7, and decrease it after 100 epochs. The total training epochs for this stage is 300. We also try gradient clipping, and find that though it enables us to use a bigger learning rate, the results are not better compared to when a small learning rate was used. In the second stage, the fully connected layer is randomly initialized from a Gaussian distribution, and the convolutional layers are initialized from the first stage. The learning rate is 1e-4 (bigger learning rate like 0.001 led to more sever overfitting because the training dataset size is small and the network is relatively deep), and decreased by 0.1 after 20 epochs. We train it for 50 epochs in total. The Stochastic Gradient Descent (SGD) is used as the optimization algorithm. For testing, a **single center crop** with size 224×224 is used. The settings are same for all the experiments. We use the face net (VGG-16) from [42], which is trained on 2.6M face images collected by the authors. For the VGG fine-tuning baseline, the fc8 layer is trained from scratch while the weights of the rest layers are initialized from the FaceNet. The learning rate of the fc8 layer is 1e-3 while the rest layers are 1e-4. All the experiments are performed using the deep

learning framework Caffe [43]. Upon publication, the trained expression models will be made publicly available.

2.4.2 Neuron Visualization

We first show that the model trained with our algorithm captures the semantic concepts related to facial expression very well. Given a hidden neuron, the face images that obtain high response are averaged. We visualize these mean images for several neurons in Fig. 2.4 and Fig. 2.5 on CK+ and Oulu-CASIA, respectively. Humans can easily assign each neuron with a semantic concept it measures (*i.e.* the text in black). For example, the neuron 11 in the first column in Fig. 4 corresponds to Anger, and the neuron 53 in Fig. 2.5 represents Happy. Interestingly, the high-level concepts learned by the neurons across the two datasets are very consistent.

2.4.3 CK+

CK+ consists of 529 videos from 123 subjects, 327 of them annotated with eight expression labels. Each video starts with a neutral expression, and reaches the peak in the last frame. As in other works [31], we extract the last three frames and the first frame of each video to compose the image-based CK+ database. The total number of images is 1308, which is split into 10 folds. The subjects are divided into ten groups by ID in ascending order.

In Table 2.4, we compare our approach with both traditional and deep learning-based methods in terms of average accuracy. We consider the fine-tuned VGG-16

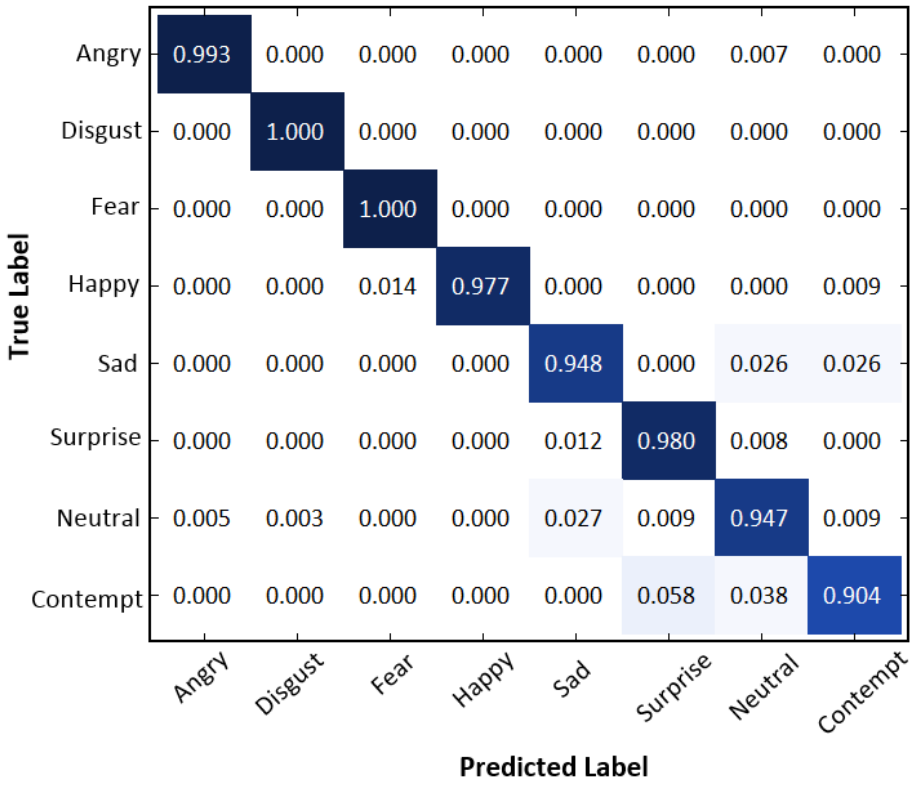


Figure 2.6: Confusion Matrix of CK+ for the Eight Classes problem. The darker the color, the higher the accuracy.

face net as our baseline. To further show the superiority of our method, we also include the results on training from scratch with batch normalization. The network architecture is same as FNEN. The first block shows the results for six classes, while the second block shows the results for eight classes, including both contempt and neutral expressions. Among them, 3DCNN-DAP [32], STM-ExpLet [31] and DTAGN [33] are image-sequence based methods, while others are image-based. For both cases, our method performs the best, achieving 98.6% vs the pervious best of 97.3% for six classes, and 96.8% vs 92.1% for eight classes.

Because of the high accuracy on the six class problem, here we only show the confusion matrix for eight class problem. From Fig. 2.6 we can see that both disgust

Table 2.4: The Average Accuracy on CK+ dataset (Person-Independent)

Method	Average Accuracy	#Exp. Classes
CSPL [27]	89.9%	Six Classes
AdaGabor [44]	93.3%	
LBPSVM [45]	95.1%	
3DCNN-DAP [32]	92.4%	
BDBN [30]	96.7%	
STM-ExpLet [31]	94.2%	
DTAGN [33]	97.3%	
Inception [34]	93.2%	
LOMo [46]	95.1%	
PPDN [21]	97.3%	
FN2EN	**98.6%**	
AUDN [29]	92.1%	Eight Classes
Train From Scratch (BN)	88.7%	
VGG Fine-Tune (baseline)	89.9%	
FN2EN	**96.8%**	

and fear expressions are perfectly classified, while contempt is the most difficult to classify. It is because this expression has the least number of training images, and the way people show it is very subtle. Surprisingly, from the visualization in Fig. 2.1, the network is still able to capture the speciality of contempt: the conner of the mouth is pulled up. This demonstrates the effectiveness of our training method.

2.4.4 Oulu-CAS VIS

Oulu-CASIA data set has 480 image sequences taken under Dark, Strong, Weak illumination conditions. In this experiment, only videos with strong conditions captured by a VIS camera are used. There are 80 subjects and six expressions in total. Similar to CK+, the first frame is always neutral while the last frame has the peak expression. Only the last three frames are used, and the total number of

Table 2.5: The Average Accuracy on Oulu-CAS dataset (Person-Independent)

Method	Average Accuracy
HOG 3D [47]	70.63%
AdaLBP [6]	73.54%
Atlases [48]	75.52%
STM-ExpLet [31]	74.59%
DTAGN [33]	81.46%
LOMo [46]	82.10%
PPDN [21]	84.59%
Train From Scratch (BN)	76.87%
VGG Fine-Tune (baseline)	83.26%
FN2EN	**87.71%**

images is 1440. A ten-fold cross validation is performed, and the split is subject independent.

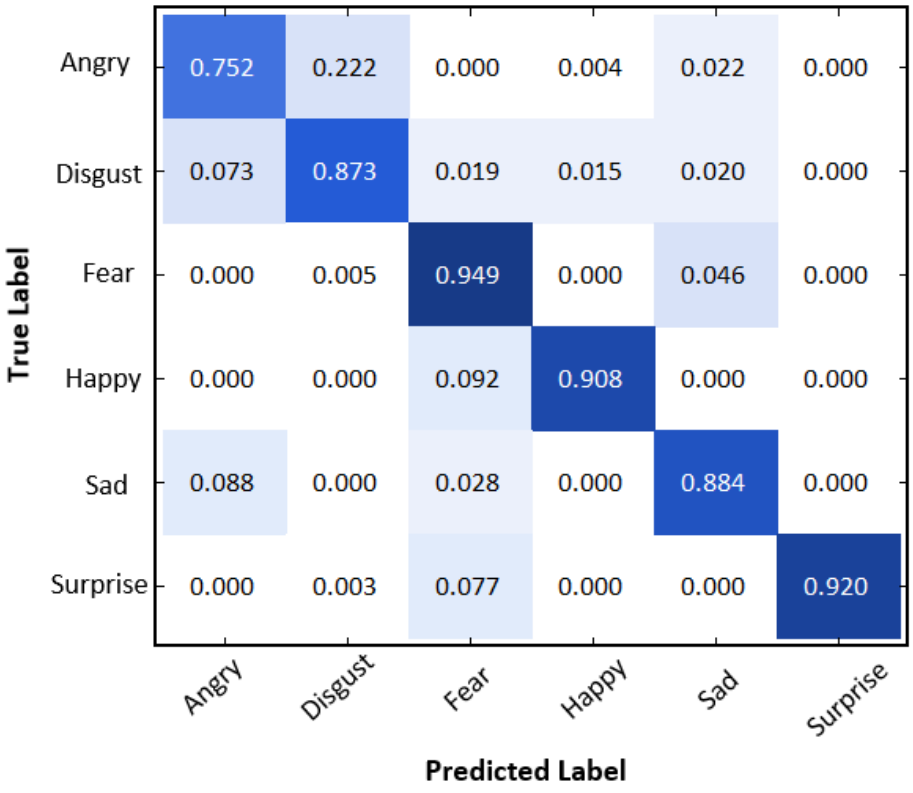


Figure 2.7: Confusion Matrix of Oulu-CASIA. The darker the color, the higher the accuracy.

Table 2.6: The Average Accuracy on TFD dataset (Person-Independent)

Method	Average Accuracy
Gabor + PCA [49]	80.2%
Deep mPoT [50]	82.4%
CDA+CCA [51]	85.0%
disRBM [52]	85.4%
bootstrap-recon [53]	86.8%
Train From Scratch (BN)	82.5%
VGG Fine-Tune (baseline)	86.7%
FN2EN	**88.9%**

Table 2.5 reports the results of average accuracy for the different approaches. As can be seen, our method achieves substantial improvements over the previous best performance achieved by PPDN [21], with a gain of **3.1%**. The confusion matrix is shown in Fig. 2.7. The proposed method performs well in recognizing fear and happy, while angry is the hardest expression, which is mostly confused with disgust.

2.4.5 TFD

The TFD is the largest expression dataset so far, which is comprised of images from many different sources. It contains *4178* images, each of which is assigned one of seven expression labels. The images are divided into 5 separate folds, each containing train, valid and test partitions. We train our networks using the training set and report the average results over five folds on the test sets.

Table 2.6 summarizes our TFD results. As we can see, the fine-tuned VGG face is a fairly strong baseline, which is almost on par with the current state-of-the-art, 86.7% vs 86.8%. Our method performs the best, significantly outperforming

bootstrap-recon [53] by 2%. From the confusion matrix, we find that fear has the lowest recognition rate and is easy to be confused with surprise. When inspecting the dataset, we find the images from the two expressions indeed have very similar facial appearances: mouth and eyes are wide open.

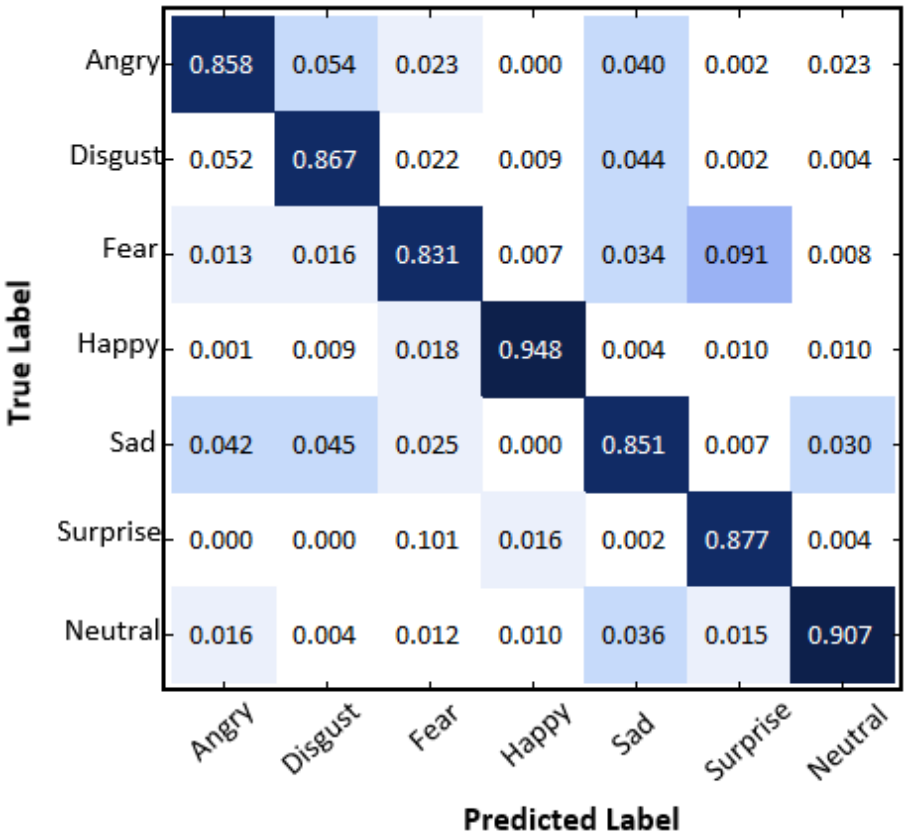


Figure 2.8: Confusion Matrix of TFD. The darker the color, the higher the accuracy.

2.4.6 SFEW

Different from the previous three datasets, SFEW is targeted for unconstrained expression recognition. So the images are all extracted from films clips, and labeled with seven expressions. The poses are large, and the expression is much more difficult to recognize. Furthermore, it has only 891 training images. Because we do

Table 2.7: The Average Accuracy on SFEW dataset (Person-Independent)

Method	Average Accuracy	Extra Train
AUDN [29]	26.14%	None
STM-ExpLet [31]	31.73%	
Inception [34]	47.70%	
Mapped LBP [20]	41.92%	
Train From Scratch (BN)	39.55%	
VGG Fine-Tune (baseline)	41.23%	
FN2EN	**48.19%**	
Transfer Learning [36]	48.50%	FER2013
Multiple Deep Network [35]	52.29%	
FN2EN	**55.15%**	

not have access to the test data, here we report the results on the validation data.

In Table 2.7, we divide the methods into two blocks, where the first block only uses the training images from SFEW, while the second block utilizes FER2013 [54] as additional training data. For both settings, our method achieves best recognition rates. Especially with more training data, we surpass Multiple Deep Network Learning [35] by almost **3%**, which is the runner-up in EmotiW 2015. We do not compare the result with the winner [55] since they use 216 deep CNNs to get 56.40%, while we only use a single CNN (1.25% higher than our method). From the confusion matrix Fig. 2.9, we can see the accuracy for fear is much lower than other expressions. This is also observed in other works [36].

2.4.7 RAF

To further explore our method on large-scale facial expression dataset, we conduct experiments on the recently proposed RAF dataset. RAF contains 30,000 in-the-wild facial expression images, annotated with basic or compound expressions

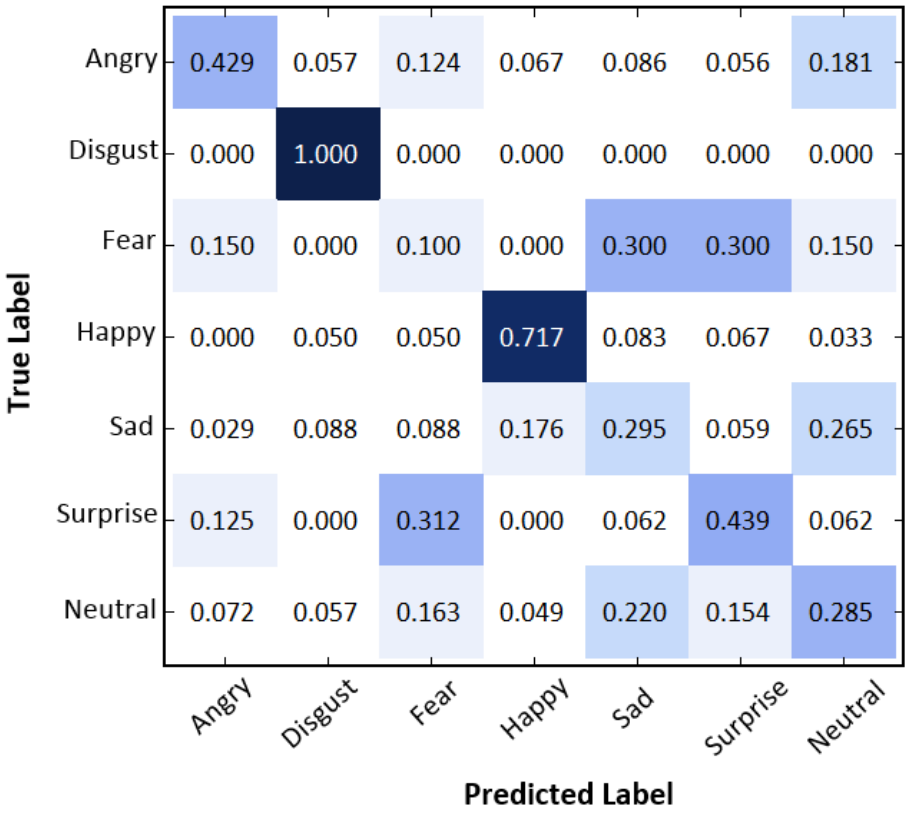


Figure 2.9: Confusion Matrix of SFEW. The darker the color, the higher the accuracy.

by 40 independent human labelers. In this experiment, only images with basic expressions are used, including 12,271 for training and 3,068 for testing.

As we can see from Table 2.8, our method achieves comparable performance in terms of total accuracy, which is 86.18% vs. 87.00%. However, Covariance Pooling [56] requires more computational resource since it needs to compute the covariance matrix, while our model is more compact and light-weight. Notably, LTNet [57] obtains accuracy of 86.77% by pretraining the model on a much larger dataset Affectnet [58]. This validates that our method is also effective on relative large dataset. From the confusion matrix in Fig. 3.4 we observe that the expression Disgust is the most difficult category due to the subtleness, which is easily confused

Table 2.8: The Average Accuracy on RAF dataset (Person-Independent)

Method	Average Accuracy	Extra Train
DLP-CNN [40]	84.70%	None
Covariance Pooling [56]	87.00%	
PG-CNN [59]	83.27%	
Train From Scratch (BN)	81.75%	
VGG Fine-Tune (baseline)	85.56%	
FN2EN	**86.18%**	
LTNet [57]	86.77%	AffectNet

with the expression Sad or Neutral.

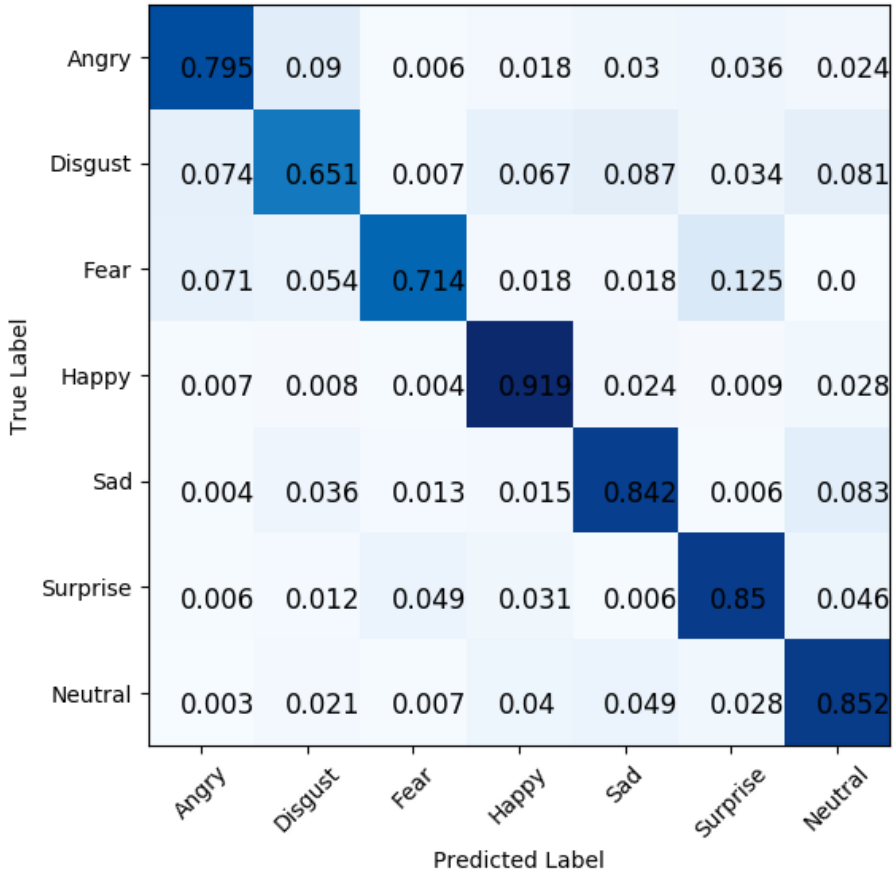


Figure 2.10: Confusion Matrix of RAF. The darker the color, the higher the accuracy.

To investigate how the regularization of FaceNet influence the learning of ExpNet, we visualize the attention maps of Train From Scratch (BN), VGG Fine-

Tune and FN2EN using the gradient weighed class activation mapping (Grad-CAM) method in [60]. The results are shown in Fig. 2.11. We observe that without the guidance of FaceNet, the focused facial region by train from scratch is not very meaningful. For example, in the first row of the Angry expression, the high attention region includes the eyes with sunglasses. While FN2EN learns to focus on the cheek and inner brow since these are the most discriminative facial regions. We also find that FN2EN can avoid the mistake of FaceNet by looking at the correct facial areas. For example, in the fifth row of the Surprise expression, both train from scratch and fine-tune FaceNet have a strong attention on the eyes region, and make wrong predictions (Sad and Happy). While FN2EN is able to extract the expression features from the mouth region and make the correct prediction.

2.4.8 Ultrasound Abdomen Dataset

The abdomen database contains US images for 16 classes of different anatomy organs. It includes Liver, Kidney, Spleen, Aorta, Gallbladder, Iliac, IVC, Pancreas and others. Some organs have four different views: left transverse, left longitudinal, right transverse and right longitudinal. In total it has 131,003 and 56216 frames from patients for training and testing. Ultrasound (US) images are especially difficult to analyze because of low contrast and large intra-variance. Some images from the dataset are shown in Fig. 2.12.

To show the effectiveness of our method, we adopted two different network architectures, *i.e.*, AlexNet and VGG16, as our teacher nets. For comparison, three

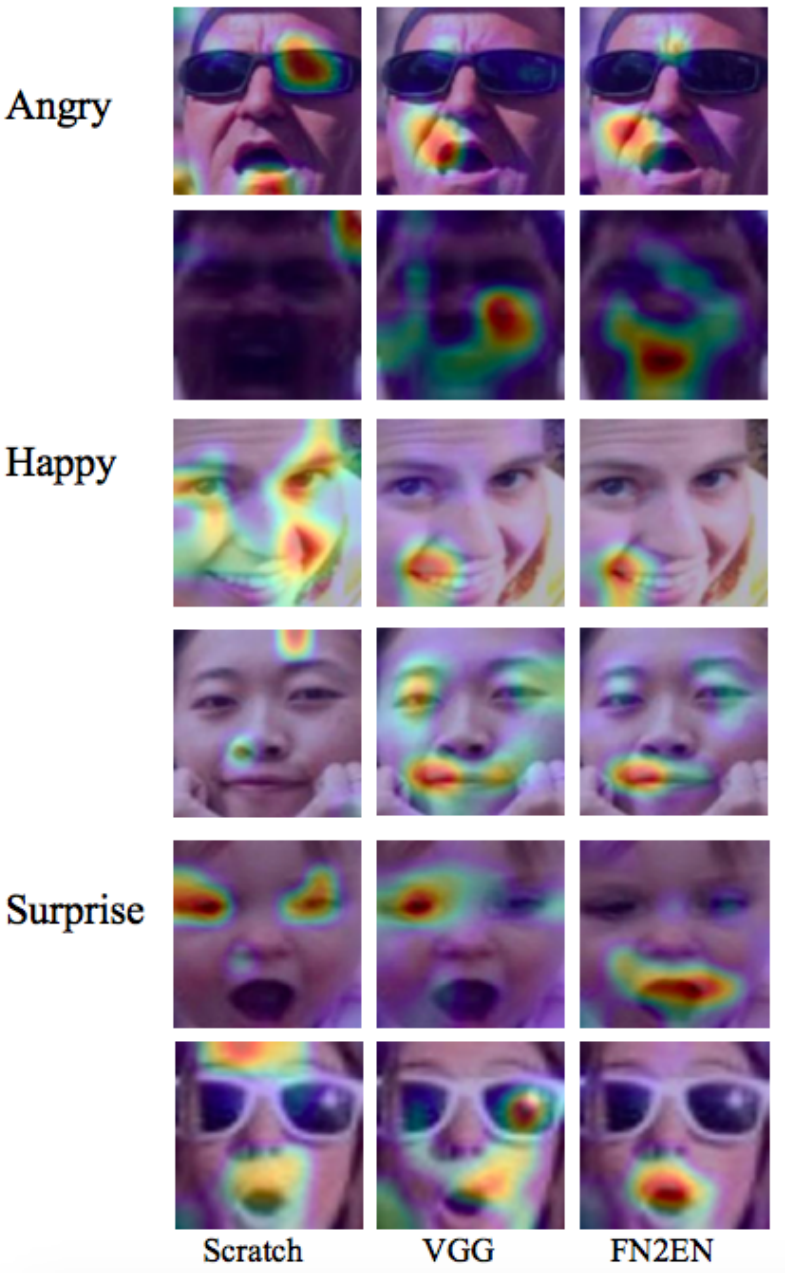


Figure 2.11: Attention maps of three different methods. The images' expression labels are displayed on the leftmost. The left, middle and right columns show the predictions of networks train from scratch, fine-tune from FaceNet and FN2EN. A deep red denotes high attention.

baselines are adopted: training teacher net from scratch, fine-tuning teacher net and training student net from scratch. The recognition results are shown in Table 2.9. We observed that our method achieves significant performance boost, 6%

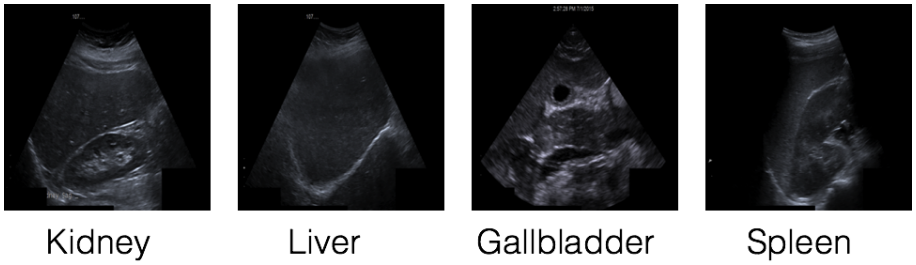


Figure 2.12: Sample images from the Ultrasound Abdomen Dataset.

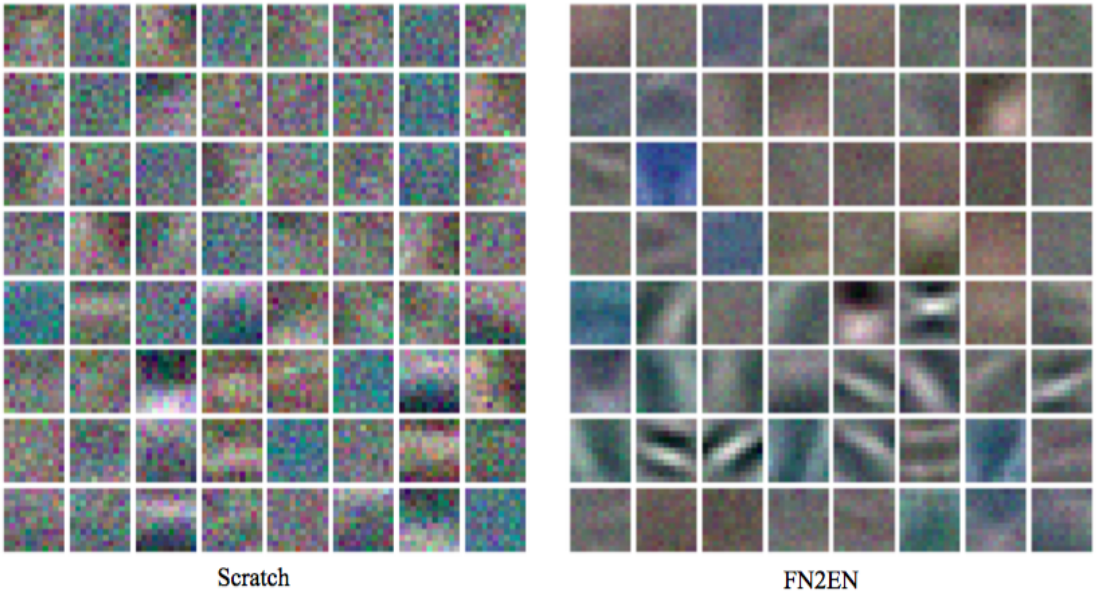


Figure 2.13: (**left**) The conv1 filters learned by training AlexNet from scratch on the abdomen dataset. (**right**) The conv1 filters learned with our method. Our model learns gabor-like conv1 filters.

improvement when AlexNet used as teacher net and 4% when VGG16 as the teacher net.

To gain further insight into what our model learns, we visualize the first convolution layer of AlexNet training from scratch and the network learned by FN2EN, respectively. We concatenate the ultrasound image along the channel axis to make it a color input to the network. From Fig. 2.13, we observed that our model learns gabor-like fiters. Since ultrasound images are gray scale, our network does not learn

Table 2.9: The Average Accuracy on Ultrasound Abdomen dataset

Method	Average Accuracy
AlexNet	74.92%
AlexNet Fine-Tune	79.97%
Train From Scratch (BN)	75.70%
FN2EN (AlexNet)	**81.01%**
VGG16	80.70%
VGG16 Fine-Tune	82.98%
Train From Scratch (BN)	81.75%
FN2EN (VGG16)	**85.02%**

any color blob filters.

2.5 Expression Feature Analysis

We analyze how well different facial attributes are being captured in the expression representation by computing the mutual information (MI). For this, we adopt the Mutual Information Neural Estimator (MINE) [61] which provides unbiased estimation of mutual information on n i.i.d samples by leveraging a neural network. We conduct our experiments on RAF dataset since it also contains labels for gender, race and age. The neural network has two fully-connected layers, each has 50 hidden unites. We find that the MI between the expression feature and the expression label is the highest, which is 1.43. While the MI between the expression feature and race is higher than gender and age, which is 1.05.

2.6 Computational speed analysis

Compared with networks adopted in previous works [21, 33, 34], AlexNet [62] or VGG-M [63], the size of our network is fairly small. The number of parameters is 11M vs. VGG-16 baseline 138M. The learned expression representation is also very compact with only 256 dimensions. This is 20 times less compared with VGG-16. For testing, our approach takes only 3ms per image using a single Titan X GPU.

2.7 Conclusions

In this chapter, we present FaceNet2ExpNet, a novel two-stage training algorithm for expression recognition. In the first stage, we propose a probabilistic distribution function to model the high level neuron response based on already fine-tuned face net, thereby leading to feature level regularization that exploits the rich face information in the face net. In the second stage, we perform label supervision to boost the final discriminative capability. As a result, FaceNet2ExpNet improves visual feature representation and outperforms various state-of-the-art methods on five public expression datasets and one medical dataset.

Chapter 3: Occlusion Adaptive Deep Network for Robust Facial Expression Recognition

3.1 Introduction

Facial expression plays an important role in social communication during our daily life. In recent years, automatically recognizing expression has received increasing attention due to its wide applications, including driver safety, health care, video conferencing, virtual reality, cognitive science, *etc.*

Existing methods that address expression recognition can be divided into two categories. One category utilizes synthesis techniques to facilitate the discriminative feature learning [64, 65, 66, 67]; while the other tries to boost the performance by designing new loss functions or network architectures [40, 56, 57, 68]. In the first category, de-expression residue learning [64] leverages the neutral face images to distill the expression information from the corresponding expressive images. Zhang *et al.* [65] explore an adversarial autoencoder to generate facial images with different expressions under arbitrary poses to enlarge the training set. However, those works mainly focus on datasets captured in controlled environments, such as CK+ [4], MMI [5] and OULU-CASIA [6]. Although high accuracy classifier has been obtained

on these datasets, it performs poorly when recognizing facial expressions in-the-wild. In the second category, Li *et al.* [40] propose a locality preserving loss to enhance deep features by preserving the locality closeness while maximizing the inter-class scatters. To address the annotation inconsistency among different facial expression datasets, Zeng *et al.* [57] introduce a probability transition layer to recover the latent truths from the noisy labels. Although expression datasets under nature and uncontrollable variations are explored, facial expression recognition under partial occlusions is still a challenging problem that has been relatively unexplored. In real-life images or videos, facial occlusions can often be observed, *e.g.* facial accessories including sunglasses, scarves, and masks or other random objects like hands, hairs and cups.

Recently, some related works have been proposed to solve this challenge. Patch-gated Convolutional Neural Network [59] decomposes a face into different patches and explicitly predicts the occlusion likelihood of the corresponding patch using a patch-gated unit. Wang *et al.* [8] propose a self-attention scheme to learn the importance weights for multiple facial regions. However, the unobstructed scores are learned without any ground truth of the occlusion information and may be biased. In this work, we present an Occlusion Adaptive Deep Network (OADN) to overcome the occlusion problem for robust facial expression recognition in-the-wild. It consists of two branches: a landmark-guided attention branch and a facial region branch.

In order to pay attention to the non-occluded facial areas and ignore the occluded areas, we propose a landmark-guided attention branch to discard feature

elements that have been corrupted by occlusions. The interest points covering the most distinctive facial areas for facial expression recognition are computed based on the domain knowledge. Then the meta information of these points is utilized to generate the attention maps. The global features are modulated by the attention maps to guide the model to focus on the non-occluded facial regions and filter out the information of occluded regions.

To further enhance the robustness and learn complementary context information, we introduce a facial region branch to train multiple region-based expression classifiers. This is achieved by first partitioning the global feature maps into non-overlapping facial blocks. Then each block is trained by backpropgating the recognition loss independently. Thus even the face is partially occluded, the classifiers from other non-occluded regions are still able to function properly. Furthermore, since the expression datasets are usually small, having multiple region-based classifiers adds more supervision and acts as a regularizer to alleviate the overfitting issue.

Contributions: We propose OADN, an effective method to deal with the occlusion problem for facial expression recognition in-the-wild. We introduce a landmark-guided attention branch to guide the network to attend to the non-occluded regions for representation learning. We design a facial region branch to learn region-based classifiers for complementary context features and further increasing the robustness. Experimental results on five challenging benchmark datasets show that our proposed OADN obtains significantly better performance than existing methods.

Organization: Section 3.2 provides an overview of existing works on deep learning-based facial expression recognition and facial expression recognition under occlu-

sions. Section 3.3 presents the proposed occlusion adaptive deep network. Experiment results and conclusions are presented in Section 3.4 and Section 3.5, respectively.

3.2 Related Work

3.2.1 Deep Facial Expression Recognition

Deep learning methods [27, 28, 29, 30, 31, 40, 56, 57, 64, 65, 67, 68, 69] for facial expression recognition have achieved great success in the past few years. Based on the assumptions that a facial expression is the combination of a neutral face image and the expressive component, Yang *et al.* [64] proposed a de-expression residue learning to learn the residual expressive component in a generative model. To reduce the inter-subject variations, Cai *et al.* [67] introduced an identity-free generative adversarial network [11] to generate an average identity face image while keep the expression unchanged. Considering the pose variation, Zhang *et al.* [65] leveraged an adversarial autoencoder to augment the training set with face images under different expression and poses. However, these methods mainly focus on datasets captured in controlled environments. The facial images are near frontal without any occlusion. Thus the models generalize poorly when recognizing human expressions under nature and uncontrollable variations.

Another line of works focus on designing advanced network architectures [56] or loss functions [40, 57, 68, 69]. Li *et al.* [40] proposed a deep locality-preserving Convolutional Neural Network, which preserved the locality proximity by minimiz-

ing the distance to the K-nearest neighbors within the same class. Building on this, Cai *et al.* [68] further introduced an island loss to simultaneously reduce intra-class variations and augment inter-class differences. Zeng *et al.* [57] studied the annotation error and bias problem among different facial expression datasets. Each image is predicted with multiple pseudo labels and a model is learned to fit the latent truth from these inconsistent labels. Acharya *et al.* [56] explored a covariance pooling layer to better capture the distortions in regional facial features and temporal evolution of per-frame features. Although the aforementioned approaches achieve good performance on the data from the wild, facial expression recognition is still challenging due to the existence of partially occluded faces. As a result, only few methods are proposed to address this challenging issue.

3.2.2 Occlusive Facial Expression Recognition

Recently, there are some works starting to investigate the occlusions issue. Li *et al.* [7] proposed a gate unit to enable the model to shift attention from the occluded patches to other visible facial regions. The gate unit estimates how informative a face patch is through an attention net, then the features are modulated by the learned weights. Similarly, region attention network [8] cropped multiple face regions and utilized a self-attention based model to learn an important weight for each region. However, the self-attention based methods lack additional supervision information to ensure the functionality. Thus, the network may not be able to locate these non-occluded facial regions accurately under large occlusions and poses.

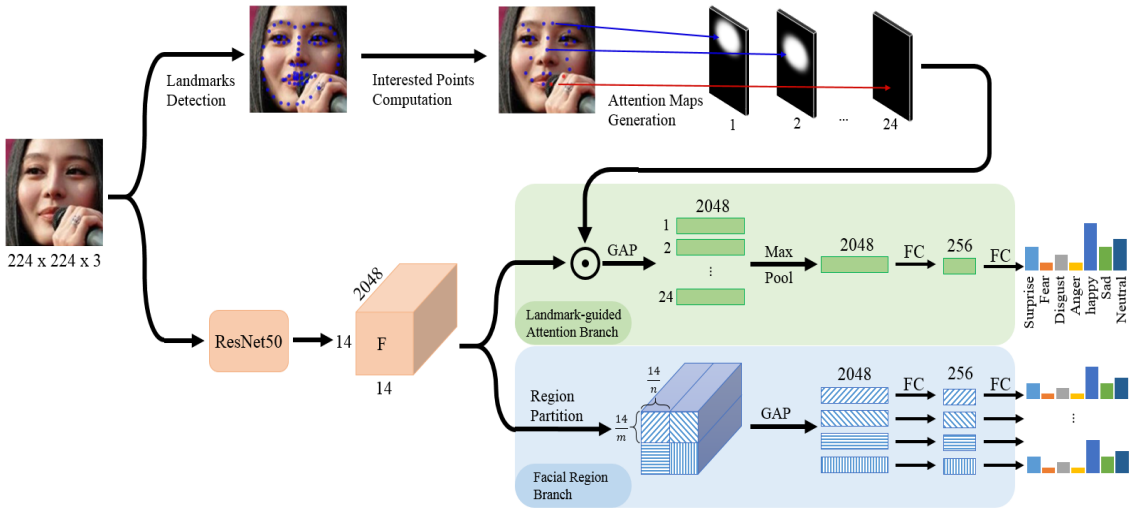


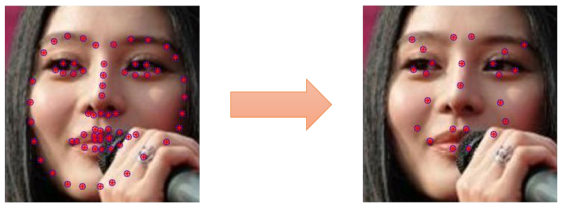
Figure 3.1: Pipeline of the Occlusion Adaptive Deep Network. It consists of two branches: a Landmark-guided Attention Branch and a Facial Region Branch. The ResNet50 backbone is shared between the two branches to extract the global features. For the Landmark-guided Attention Branch, the facial landmarks are first detected. Then the interested points are computed to cover the most informative facial areas. The confidence scores of these points are further utilized to generate the attention maps, guiding the model to attend to the visible facial components. While for the Facial Region Branch, the feature maps are divided into non-overlapping facial blocks and each block is trained to be a discriminative expression classifier on its own.

3.3 Occlusion Adaptive Deep Network

In this chapter, we propose OADN for robust facial expression recognition in-the-wild. To be specific, we use ResNet50 [70] without the average pooling layer and fully connected layer as the backbone to extract global feature maps F from given images. The feature map is denoted as $F \in h \times w \times c$, where h, w, c are the height, width and channel dimensions. We set the stride of conv4_1 to be 1, so a larger feature map is obtained. For an input image with height H and width W, the resolution of the output feature F will be $H/16 \times W/16$ instead of $H/32 \times W/32$. This is beneficial to identify the occlusion information and focus on the visible facial

regions.

As illustrated in Fig. 3.1, OADN mainly consists of two branches: one is the landmark-guided attention branch, which utilizes a landmark detector to estimate landmarks and to guide the network to attend to the non-occluded facial areas. The other one is the facial region branch to divide the global feature maps into blocks and train region-based classifiers to increase robustness. We describe each branch and the structural relationship among the two branches in details below.



(a) Original 68 detected landmarks (b) Recomputed 24 points

Figure 3.2: We select 16 points from the original 68 landmarks (a) to cover the regions around eyes, eyebrows, nose and mouth. We further recompute 8 points to cover facial cheeks and the areas between eyes and eyebrows.

3.3.1 Landmark-guided Attention Branch

OADN employs a facial landmark detector [71] to obtain landmarks from face images. The landmark detector is pre-trained on the 300W dataset [72]. Given an input image, OADN utilizes the detector to extract $N = 68$ landmarks. For each landmark, the detector predicts its coordinates and confidence score. Then based on the detected 68 points, we select or recompute $M = 24$ interested points that cover the distinctive regions of face, including the eyes, nose, mouth and cheeks. Fig. 3.2 illustrates the computation results. For those recomputed points (mainly around eyes and cheeks), we set their confidence scores to be the minimum confidence score

of landmark points that used to compute them. To remove the occluded facial regions, we set a threshold T to filter out the landmarks that have confidence scores smaller than T. Specifically, the interested points are obtained by:

$$p_i = \begin{cases} (x_i, y_i) & if \quad s_i^{conf} \geq T \\ 0 & else \end{cases} \tag{3.1}$$

where p_i denotes the ith interested point, and x_i, y_i denote the coordinates of the ith point. s_i is the confidence score ranged from 0 to 1 and T is the threshold.

We then generate the attention heatmaps consisting of a 2D Gaussian distribution, where the centers are the ground truth locations of the visible landmarks. For those occluded landmarks, the corresponding attention maps are set to be all zeros. We further downsample the attention maps by linear interpolation to match the size of the output feature maps. As shown in Fig. 3.1, the attention map A_i modulates the global feature maps F to obtain the re-weighted features F_i^A. To achieve this, the feature map F from the backbone is multiplied by each attention map A_i, $i = 1, ..., M$ element-wisely, resulting M landmark-guided feature maps F_i^A:

$$F_i^A = F \odot A_i, i = 1, ..., M \tag{3.2}$$

where A_i is the ith heatmap, and $\odot$ is element-wise product. Since the attention map indicates the visibility of each facial component, the landmark-guided feature map F_i^A can attend to the non-occluded facial parts and remove the information from the occluded regions. Thus, the feature from the visible region is signified and

occluded part is canceled.

Then global average pooling is applied to each landmark-guided feature map F_i^A to obtain a 2048-D feature $f_i^A, i = 1, ..., M$, corresponding to the facial component containing the specific interested point. Finally, the component-wise feature f_i^A is max-pooled to fuse the features from the non-occluded facial areas and reduce the redundant partial information. A fully-connected layer is further used to reduce the dimension from 2048 to 256, and the output is fed into a softmax layer to predict the expression category of each input face image. We utilize cross-entropy loss to train the landmark-guided attention branch, which is expressed as follows:

$$L_{LAB} = -\sum_{i=1}^{C} y_i \log \hat{y}_i \tag{3.3}$$

where $\hat{y}_i$ is the prediction, y_i is the ground truth and C is the number of expression classes.

3.3.2 Facial Region Branch

When the face is seriously occluded, the landmark detection results may not be accurate. Thus relying on the landmark-guided attention branch solely is not enough. OADN utilizes a Facial Region Branch (FRB) to learn useful context information and further increase the robustness.

Given the global feature maps F, we first divide them into small $m \times n$ non-overlapping blocks. Each facial region feature $F_i^R \in m \times n \times c, i = 1, ..., K$, with $K = \lceil \frac{h}{m} \rceil \cdot \lceil \frac{w}{n} \rceil$ is then fed into a global average pooling layer to obtain a region-level

feature f_i^R. Afterwards, a fully-connected layer is employed to reduce the dimension of f_i^R from 2048 to 256. Finally, a softmax layer is applied to each region to get a set of predictions y_i^R, where $i = 1, ..., K$.

To train the facial region branch, we minimize the cross-entropy loss over the K regions independently. Formally, the loss is expressed as:

$$L_{FRB} = -\sum_{i=1}^{C}\sum_{j=1}^{K} y_i \log \hat{y}_{i,j}^R \tag{3.4}$$

where K is the number of facial regions, $\hat{y}_{i,j}^R$ is the probability of the jth region prediction, and y_i is the ground truth expression category.

To be able to make an accurate prediction based on facial region only, OADN learns more discriminative and diverse features at a finer-level. Thus the partial occlusion will have a less effect on the network compared with standard model. Moreover, the size of the expression recognition dataset is usually not very large. Training multiple region-based classifiers adds more supervision and reduces the overfitting.

3.3.3 Relationship between the Two Branches

OADN is specifically designed to handle the occlusion problem for in-the-wild facial expression recognition. The landmark-guided attention branch explicitly guides the model to focus on the non-occluded facial areas, learning a clean global feature. While the facial region branch promotes part-level features and enables the model to work robustly when the face is largely occluded. Combining the benefits

from each branch, we train OADN by the following loss:

$$L = \lambda L_{LAB} + (1 - \lambda) L_{FRB} \tag{3.5}$$

where λ is the loss combination weight. L_{LAB} and L_{FRB} are defined in Equation (3.3) and (3.4).

3.4 Experiments

3.4.1 Datasets

We validate the effectiveness of our method on two largest in-the-wild expression datasets: RAF-DB [40] and AffectNet [58]. The in-the-wild datasets contain facial expression in real world with various poses, illuminations, intensities, and other uncontrolled conditions. We also evaluate our method on three recently proposed real-world occlusion datasets: Occlusion-AffectNet[8], Occlusion-FERPlus [8] and FED-RO [7]. The occlusions are diverse in color, shape, position and occlusion ratio.

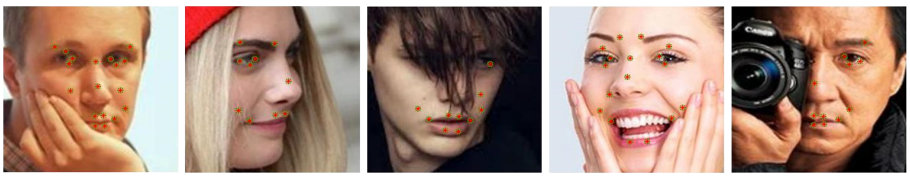


Figure 3.3: The interest points with confidence scores greater than the threshold T are shown in red points. We can see the occluded facial areas are removed.

3.4.2 Implementation Details

Preprocessing. The standard MTCNN [41] is used to detect five face landmarks for all the images. After performing similarity transformation accordingly, we obtain the aligned face images and resize them to be 224 x 224 pixels. To detect landmarks from occluded images, we use SAN [71] pre-trained on the 300W dataset [72] to get 68 face landmarks. We also try another landmark detector [73] and similar results are obtained. Then we select 18 points covering eyebrows, eyes, nose and mouth, and recompute eight points related with facial cheeks. The confidence scores of these recomputed points are the minimum score of the points that used to compute them. In all experiments, we set the threshold T of the confidence score to be 0.6, thus the landmarks with confidence scores smaller than it are removed. Fig. 3.3 shows the computed interested points after thresholding. From it we can see the occluded facial regions are discarded. Finally, we generate attention maps consisting of a Gaussian with the centers to be the coordinates of the visible points. For those occluded points, the attention maps are all zeros. We resize the attention maps to be 14×14 to match the size of the global feature maps F.

Training and Testing. We employ the ResNet50 as our backbone, removing the average pooling layer and the fully connected layer. We modify the stride of conv4_1 from 2 to 1, so a larger feature map with size 14×14 is obtained. We initialize the model with the weights pre-trained on ImageNet [74]. The mini-batch size is set to be 128, the momentum is 0.9, and the weight decay is 0.0005. The learning rate starts at 0.1, and decreased by 10 after 20 epochs. We train the

Table 3.1: Test Set Accuracy on RAF dataset

Method	Average Accuracy
RAN [8]	86.90%
OADN(ours)	**89.83%**
ResiDen [76]	76.54%
ResNet-PL [77]	81.97%
PG-CNN [59]	83.27%
Center Loss [78]	83.68%
DLP-CNN [79]	84.13%
ALT [80]	84.50%
gACNN [7]	85.07%
OADN(ours)	**87.16%**

model for a total of 60 epochs. Stochastic Gradient Descent (SGD) is adopted as the optimization algorithm. During training, only random flipping is used as data augmentation. For testing, a single image is used and the predication scores from the landmark-guided attention branch and the facial region branch are averaged to get the final prediction score. The settings are same for all the experiments. For evaluation, the total accuracy metric is adopted. Considering the imbalance of the expression classes, confusion matrix is also employed to show the average class accuracy. The deep learning framework Pytorch [75] is used to conduct the experiments. Upon publication, the codes and trained expression models will be made publicly available.

3.4.3 Results Comparison

RAF [40] contains 30,000 in-the-wild facial expression images, annotated with basic or compound expressions by 40 independent human labelers. In this experi-

Table 3.2: Validation Set Accuracy on AffectNet dataset

Method	Average Accuracy
RAN [8]	59.50%
OADN(ours)	**64.06%**
VGG16 [82]	51.11%
GAN-Inpainting [83]	52.97%
DLP-CNN [40]	54.47%
PG-CNN [59]	55.33%
ResNet-PL [77]	56.42%
gACNN [7]	58.78%
OADN(ours)	**61.89%**

ment, only images with seven basic expressions are used, including 12,271 for training and 3,068 for testing.

Table. 3.1 shows the results of our method and previous works. Our OADN achieves 87.16% in terms of total accuracy on the test set, outperforming all the previous methods. Compared with the strongest competing method in the same setting gACNN [7], OADN surpasses it by 2.1%. This is because OADN explicitly utilizes the meta information of landmarks to depress the noisy information from the occluded regions and enhances the robustness with multiple region-based classifiers. To have a fair comparison with [8], we also pre-trained our model on a large-scale face recognition dataset VGGFace2 [81]. OADN achieves a new state-of-the-art result with an accuracy of 89.83% to the best of our knowledge, outperforming RAN by 2.93%. This validates the superiority of the proposed method.

We show the confusion matrix in Fig. 3.4. The average class accuracy is computed using the mean diagonal values of the confusion matrix. From the figure, we can see OADN achieves an average class accuracy of 83.21%, surpassing DLP-

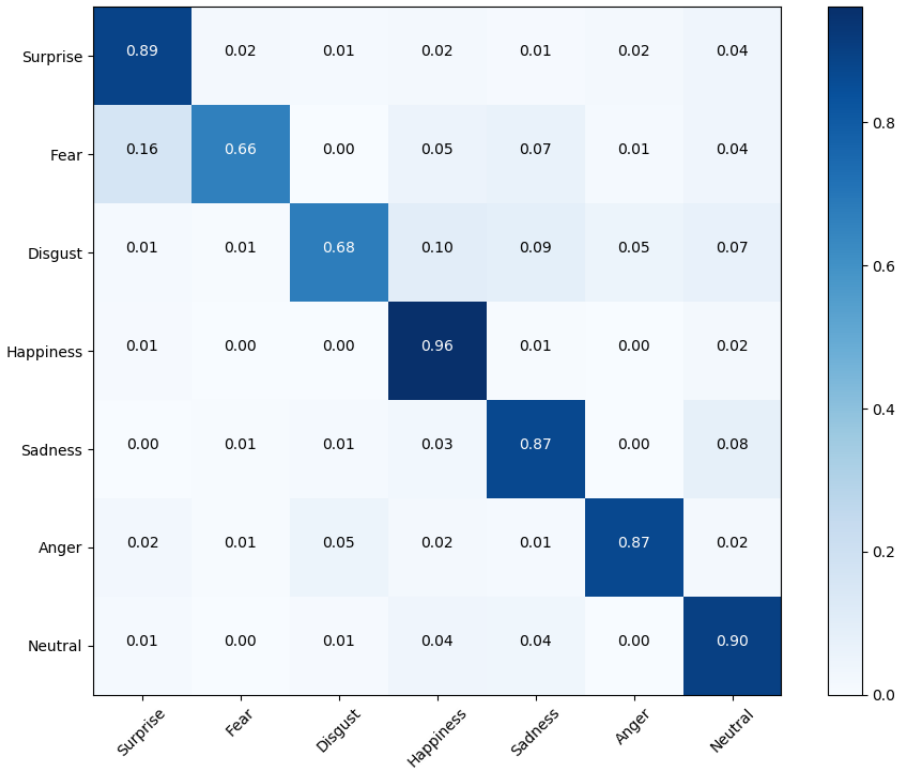


Figure 3.4: Confusion Matrix of RAF-DB. The darker the color, the higher the accuracy.

CNN [40] by 9%, which is 74.20%. In addition, it is observed that *Fear* and *Disgust* are the two most confusing expression, where *Fear* is easily confused with *Surprise* because of the similar facial appearance While *Disgust* is mainly confused by *Neutral* due to the subtleness of the expression.

AffectNet [58] is currently the largest expression dataset. There are about 400,000 images manually annotated with seven discrete facial expressions and the in-

Table 3.3: Validation Set Accuracy on Occlusion-AffectNet and Pose-AffectNet dataset

Method	Occ. Acc.	Pose>30 Acc.	Pose>45 Acc.
RAN [8]	58.50%	53.90%	53.19%
OADN(ours)	**64.02%**	**61.12%**	**61.08%**

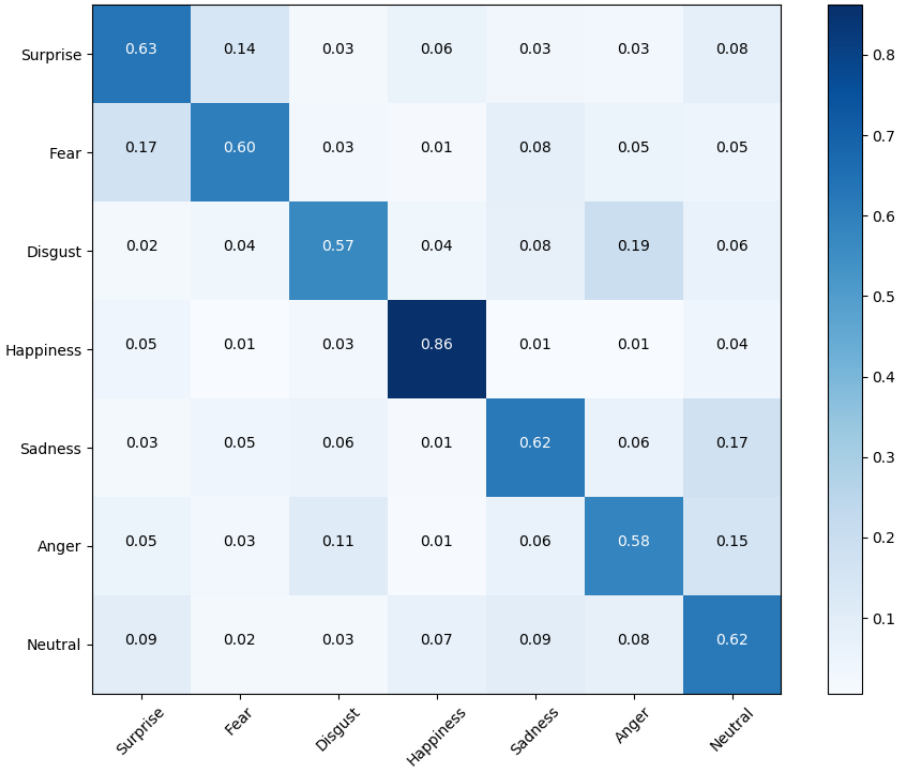


Figure 3.5: Confusion Matrix of Affectnet. The darker the color, the higher the accuracy.

tensity of valence and arousal. Following the experiment setting in [7], we only used the images with neutral and six basic emotions, containing 280,000 images for training and 3,500 images from the validation set for testing since the the test set is not publicly available. Very recently, Wang *et al.* [8] released the **Occlusion-AffectNet** and **Pose-AffectNet** datasets where only images with challenging conditions are selected as the test sets. For the Occlusion-Affectnet, each image is occluded with at least one type of occlusion: wearing mask, wearing glasses, *etc.* There are a total of 682 images. For the Pose-AffectNet, images with pose degrees larger than 30 and 45 are collected. The number of images are 1,949 and 985, respectively.

As shown in Table. 3.3, OADN achieves the best performance with an accuracy

Table 3.4: Test Set Accuracy on FED-RO dataset

Method	Average Accuracy
RAN [8]	67.98%
OADN(ours)	**71.17%**
VGG16 [82]	51.11%
ResNet18 [70]	64.25%
GAN-Inpainting [83]	58.33%
DLP-CNN [40]	60.31%
PG-CNN [59]	64.25%
gACNN [7]	66.50%
OADN(ours)	**68.11%**

of 61.89% on the validation set. Compared to the strongest competing method in the same setting gACNN [7], OADN surpasses it by 3.1%, which is a large margin. OADN also significantly outperforms RAN [8] by 4.56%, when both pre-trained on a large-scale face recognition dataset. On the Occlusion-AffectNet and Pose-AffectNet datasets, the performance gap between OADN and RAN is further increased. As a comparison, OADN exceeds RAN by 5.52%, 7.22% and 7.89% on the test sets with occlusion, pose degree greater than 30 and 45, respectively. This validates the effectiveness of the proposed method on the occluded facial expression recognition problem. The confusion matrix is shown in Fig. 3.5. From it we can find both *Disgust* and *Anger* are the most difficult expressions to classify.

FED-RO [7] is a recently released facial expression dataset with real world occlusions. Each image has natural occlusions including sunglasses, medical mask, hands and hair. It contains 400 images labeled with seven expressions for testing. We train our model on the joint training data of RAF and AffectNet, following the method [7].

As shown in Table. 3.4, OADN achieves the best performance with an accuracy of 68.11%, improving gACNN by 1.61%. OADN also significantly outperforms RAN by 3.19%. This validates the superiority of the proposed approach. From the confusion matrix shown in Fig. 3.6, we can see both *Surprise* and *Happy* have high accuracy, while *Fear* and *Disgust* are easily confused with *Surprise* and *Sad*.

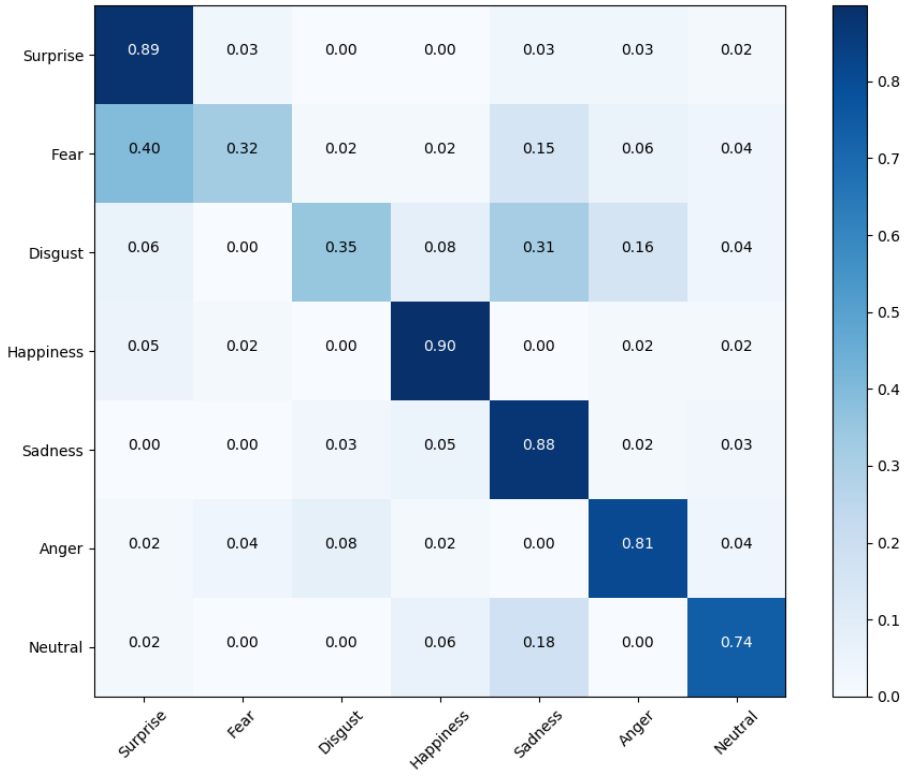


Figure 3.6: Confusion Matrix of FED-RO. The darker the color, the higher the accuracy.

FERPlus [84] is a real-world facial expression dataset initially introduced during ICML 2013 Challenge [54]. It consists of 28,709 training images, 3,589 validation images and 3,589 test images. Each image is labeled with one of the eight expressions by 10 independent taggers. Recently, Wang *et al.* [8] released the **Occlusion-FERPlus** and **Pose-FERPlus** datasets, where images under occlusion and large

Table 3.5: Test Set Accuracy on Occlusion-FERPlus and Pose-FERPlus dataset

Method	Occ. Acc.	Pose>30 Acc.	Pose>45 Acc.
RAN [8]	83.63%	82.23%	80.40%
OADN(ours)	**84.57%**	**88.52%**	**87.50%**

pose (>30 and >45) are collected from the FERPlus test sets. Following [8], we trained our model on the training data of FERPlus and test on these challenging datasets.

Table 3.5 reports the test accuracy. Our OADN significantly surpasses RAN by a large margin with 6.29% and 7.10% improvements on the Pose-FERPlus datasets. OADN also achieves better performance on the Occlusion-FERPlus dataset. This validates the effectiveness of our method on recognizing facial expressions under challenging conditions.

3.4.4 Ablation Study

In this section, we conduct extensive ablation studies on RAF dataset to analyze each component of OADN.

The impact of the landmark confidence threshold T. The confidence scores of the interested points are utilized to select the points from the non-occluded facial areas. From Equation (3.1), the points with confidence scores higher than T are kept. We can see from Fig. 3.7 (a) that with $T = 0.6$, OADN achieves the best performance. When T is further increased, the performance drops quickly since some important facial areas which may not be occluded are also thrown away. On

the other hand, when T becomes less than 0.6, OADN starts to perform worse. This is because noisy information from the occluded areas are also included, which deteriorates the clean features.

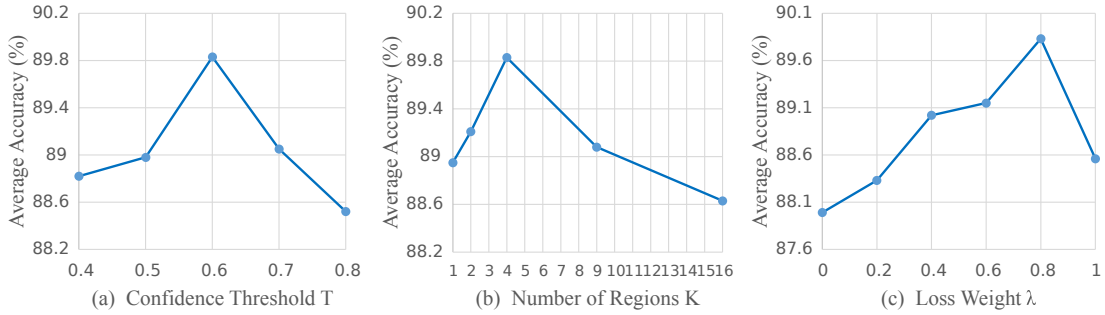


Figure 3.7: The impacts of the confidence threshold T, number of regions K and the loss combination weight λ on the performance of OADN.

The impact of the number of regions K. In the facial region branch, we partition the global feature maps into K blocks and train an expression classifier from each block independently. So K decides the granularity of the part-level features. From Fig. 3.7 (b), it is observed that the best accuracy is achieved at $K = 4$. When $K = 1$, the facial region branch equals to the standard ResNet50 classifier. The worse performance indicates the necessity to learn features at part-level. However, increasing K to be a large number like 16 does not bring further increasement. This is because when the facial region is too small, it lacks enough information to make the prediction due to the occlusion. Thus the classifiers are confused and the training is stagnated.

The impact of the loss combination weight λ. To train OADN, we jointly optimize the loss from the landmark-guided attention branch (LAB) and the facial region branch (FRB) as defined in Equation (3.5). The loss weight λ controls the

relative importance of each loss. When λ equals 1, only LAB is utilized. While $\lambda = 0$ means only FRB is used. From Fig. 3.7 (c), we can find that LAB obtains better performance since the network is guided to attend to the most discriminative facial areas. While combining the two branches achieves better performance than using either one branch alone. This validates the effectiveness of the complementary features learned by the two branches.

3.4.5 Visualization

Fig. 3.8 shows some expression recognition examples of the gACNN [7] and our OADN method on the FED-RO dataset. The classification results show that gACNN is vulnerable to large head poses and heavy facial occlusions. On the contrary, our OADN can work successfully in the same situation.

Figure 3.8: Comparison of the gACNN method and our OADN method on the FED-RO dataset. Red and green texts indicate the error and correct predictions.

3.5 Conclusions

In this chapter, we present an occlusion adaptive deep network to tackle the occluded facial expression recognition problem. The network is composed of two branches: the landmark-guided attention branch guides the network to learn clean

features from the non-occluded facial areas. While the facial region branch increases the robustness by dividing the last convolutional layer into several part classifiers. We conduct extensive experiments on both challenging in-the-wild expression datasets and real-world occluded expression datasets. The superior results show that our method outperforms existing methods and achieves robustness against occlusion and various poses.

Chapter 4: A Deep Cascade Network for Unaligned Face Attribute Classification

4.1 Introduction

Face attributes describe the characteristics observed from a face image. They were first introduced by Kumar et al. [85] as mid-level features for face verification [86] and since then have attracted much attention. The last few years have witnessed their successful applications in hashing [87], face retrieval [88], and one-shot face recognition [89]. Recently, researchers have begun to investigate the possibility of synthesizing face images based on face attributes [12, 90].

Despite their wide applications, face attribute recognition is not an easy task. One reason is that recognizing different face attributes may require attentions to different regions of the face [9, 10]. For example, local attributes like *Mustache* could be recognized by just checking the region containing the mouth. Other areas of the face do not provide useful information and may even hamper this particular attribute recognition. However, recognizing global attributes like *Pale Skin* may require information from the whole face region. Most current studies do not pay special attention to this problem. They either detect facial landmarks and extract

hand-crafted features from patches around them [85, 91] or train a deep network to classify the attributes by taking a whole face as input [92, 93, 94, 95].

In this chapter, we propose a learning-based method that dynamically selects different face regions for unaligned face attribute prediction. It integrates two networks using a cascade: a face region localization (FRL) network followed by an attribute classification network. The localization network detects face areas specific to attributes, especially those that have local spatial support. The classification network selectively leverages information from these face regions to make the final prediction.

For accurate face region detection, our localization network is constructed under a multi-task learning framework. The lower layers which are used to extract low level features are shared by all the tasks while the high-level semantics are learned separately. Moreover, a global average pooling step is applied to force the network to learn location-sensitive information [96]. Although the network is trained in a weakly-supervised manner with attribute labels only, the detected face regions are consistent with what one may expect. As a result, face alignment algorithms which are usually sensitive to occlusion, variations of pose and illumination are not needed.

For each face region (also called a part) detected by our localization network, we train a separate attribute classification network, called a part-based subnet. The localized face parts may not contain enough contextual information for predicting global attributes. Thus, a whole-image-based subnet is also trained. To combine the information from the part-based and whole-image-based subnets, a two-layer

fully-connected classifier is built on top of the output attribute scores. The first layer is used to select the relevant subnet for predicting each attribute, while the second layer is designed to model the rich attribute relations. The integrated system is called the parts and whole (PaW) network.

Since the face region localization network is supervised by attribute labels, it is appealing to adapt its weights to initialize the subnets in PaW. However, features from the localization network, which are mainly designed for localization purpose, are generally not very discriminative for attribute classification. To this end, a multi-net learning method is proposed. It utilizes a network with enhanced attribute classification capability to train the localization network to find a more discriminative solution.

A naive implementation of the PaW network is problematic since the number of total parameters increases linearly with the number of attributes, and the subnet adapted from the FRL network is not very compact. To jointly train the PaW network end-to-end, a hint-based model compression technique is further proposed. This not only leads to a compact model with only $11M$ parameters, but also reduces the training time significantly.

We applied the proposed method to CelebA dataset [92]. With no use of alignment information, our method achieves an accuracy of **91.23**%, reducing the classification error by a significant margin of **30.9**% compared with state-of-the-art [92]. Moreover, our model could select the most relevant face region for predicting each face attribute.

Contributions: We design a weakly-supervised localization network to accurately

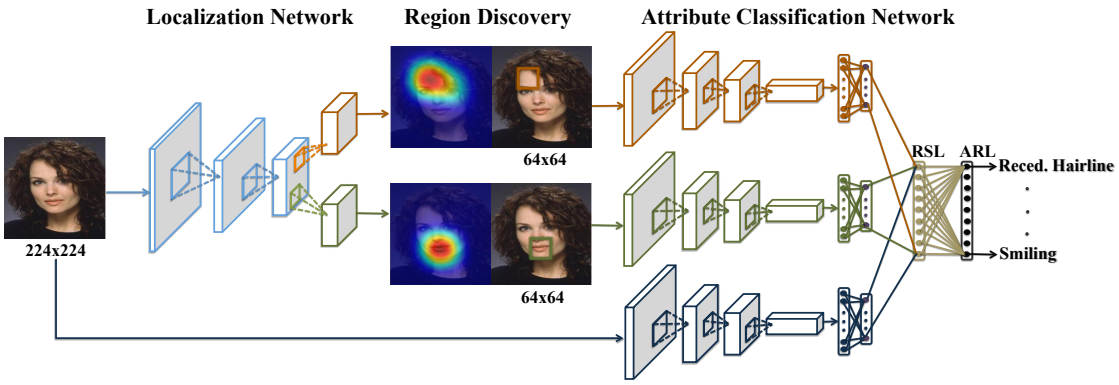


Figure 4.1: Overview of our face attribute recognition framework. It consists of a facial region localization (FRL) network and a Parts and Whole (PaW) classification network. The localization network detects a discriminative part for each attribute. Then the detected face regions and the whole face image are fed into the PaW classification network. The region switch layer (RSL) selects the relevant subnet for predicting the attribute, while the attribute relation layer (ARL) models the attribute relationships.

locate attribute regions. We also propose a hybrid classification network to dynamically choose the pertinent face regions for predicting different attributes. A hint-based model compression technique is explored to obtain a compact model. We show that the performance of unaligned face attribute classification is significantly improved by the proposed method.

Organization: Section 4.2 provides an overview of existing works on face attribute recognition, weakly supervised object localization and model compression. Section 4.3 presents the proposed face region localization network and attribute classification network. Experiment results and conclusions are presented in Section 4.4 and Section 4.5, respectively.

4.2 Related Works

Face Attribute Recognition Early works [85, 91] on face attribute recognition used manually defined face parts to extract features and then train a linear SVM classifier. This strategy though is well suited for near-frontal faces, is heavily dependent on the accuracy of landmark detection. Recently, with the emergence of large-scale data and deep neural networks, holistic methods [92, 93, 97] have produced better performance than the part-based method. Liu et al. [92] noticed that a deep model pre-trained for face recognition implicitly learns attributes. Huang et al. [97] employed a quintuplet loss to combat the imbalanced data distribution problem. These methods typically use the whole face image to train a deep network, ignoring the fact that different facial attributes have different attentional facial regions. This problem has been recently noticed in [98, 99]. Murrugarra-Llerena and Ko-vashka [99] created human gaze maps for each attribute such that only features within the saliency maps are used for attribute recognition. Our method differs from the aforementioned approaches in the sense that *the face parts are localized automatically without relying on detected landmarks or human gaze data.* Moreover, our classification network can dynamically select the relevant face regions for predicting different attributes.

Weakly Supervised Object Localization Despite training with only image-level labels, recent works [100, 101, 102] showed that deep Convolutional Neural Networks (CNN) have remarkable object localization ability. Zhou et al. [101] proposed a class activation mapping method to localize the objects with class labels only. The design

of our face region localization network is motivated by this work. However, to fully utilize the correlations among different face attributes, the localization network is designed in a multi-task learning framework.

Model Compression To obtain a compact model, several methods including network distillation [103], parameter pruning [104] have been proposed. Recently, knowledge distillation [25] has been shown to be very effective to teach a small student model. However, it can not be directly applied to our problem: the teacher net uses soft labels which contain rich ambiguous information to supervise the student net, while for attribute classification, the output has only one logit for each attribute. Thus, a new loss function based on hints is proposed to replace soft label supervision.

4.3 Proposed Method

The proposed method contains two networks: a localization network and an attribute classification network. An overview of the framework is shown in Fig. 4.1. First, we adopt the multi-net learning method to train a face region localization (FRL) network. Then one attentional region is detected for each attribute by the FRL network, which is fed into the PaW network for attribute prediction. To train the PaW end-to-end, a hint-based method is further applied to compress the model. The details of the proposed approach are discussed below.

4.3.1 Face Region Localization (FRL) Network

One challenge in designing a face region localization algorithm is that we do not have the labeled regions available. Murrugarra-Llerena and Kovashka [99] used human gaze to label the related region for each attribute, however, this is both time consuming and expensive. Inspired by the success in weakly supervised object localization [101], we apply a global average pooling (GAP) network for the localization task, and train it in a weakly-supervised way where only face attribute labels are needed. In this network structure, a GAP layer is used to pool features from the last convolutional layer, and a fully-connected layer is followed to predict the attribute score. A localization heatmap, H_j, for the j-th attribute, is obtained by applying the class activation mapping method. $H_j = \sum_{i=1}^{N} w_{j,i} F_i, i = 1, ..., N$, where F_i is the output feature maps from the last convolutional layer and $w_{j,i}$ is the i-th weight of the fully connected layer for predicting the j-th attribute. N is chosen to be 32 in our experiments.

We design the FRL network using multi-task learning [105] strategy, where each attribute can be seen as one separate task. It has five VGGNet [82] convolutional modules shared by all the attributes, and a domain adapted convolutional layer which has M different branches for each attribute, where $M = 40$ is the number of face attributes. The weights of the network are initialized from the VGG-Face CNN [42] which is trained on a large-scale face recognition dataset.

4.3.1.1 Multi-Net Learning

Since the supervision of the FRL network comes from the attribute tags, it is appealing to transfer its weights to the subnets in PaW for faster convergence and better performance. However, training the FRL net in a plain way leads to less discriminative features due to GAP regularization [101]. This is also verified in our experiments. To this end, a multi-net learning (MNL) method is proposed to boost the classification performance of the GAP feature, which yield improved final attribute classification.

The network architecture for MNL is shown in Fig. 4.2. Except for the FRL network (blue and red boxes), another two fully-connected layers (gray box) are also attached to the output of the fifth convolutional module. We call it a classification branch because of its improved performance on the classification task compared with the localization branch. The idea is to simultaneously train the two different types of networks with the same attributes loss. Meanwhile the first several convolutional layers are constructed to be shared between them. The gradients from both classification and localization branches are backpropagated to the shared layers. This extra supervision from the classification branch regularizes the training process to search for a more discriminative solution. Interestingly, we find this simple learning strategy is beneficial for both branches in terms of classification performance. *After the multi-net training is completed, the classification branch is removed, and only the localization branch is kept for extracting attribute-specific heatmaps.*

To localize the face region, we upsample the location heatmap to the original

image size 224×224, and find the position that corresponds to the maximum value. Then, a 64×64 patch centered around this position is cropped from the original image as the detected face region. We empirically found this patch size to be sufficient for most face parts. This process is repeated for each attribute and M face regions are obtained.

4.3.2 Attribute Classification Network

As shown in Fig. 4.1, the proposed attribute classification network PaW contains M part-based subnets and one whole-image-based subnet. After getting the predicted attributes scores from each subnet, a two-layer fully-connected classifier is adopted to combine them.

4.3.2.1 Parts and Whole (PaW) Classification Network

Suppose x_0 represents the whole face image, $x_1, ..., x_M$ represent face region related to each face attribute. $g_i, i \in 0, .., M$ represent the $(M+1)$ subnets. Each x_i is first fed into its corresponding subnet g_i to predict the M attribute scores $\{s_{i,j}\}$, where $s_{i,j}$ represents the predicted score of the j-th attribute by the i-th subnet. The reason why we train each part-based subnet to predict M attributes instead of the one related to the input region is based on the observation that some attributes can usually be predicted by other ones [106]. The predicted scores $s_{i,j}$ will be fed into a region switch layer (RSL) which is designed as $r_j = \sum_{i=0}^{M} W_{ij} s_{ij}, j = 1, ..., M, W \in R^{(M+1)\times M}$ whose element in the i-th row and j-th column is W_{ij}. RSL adopts a

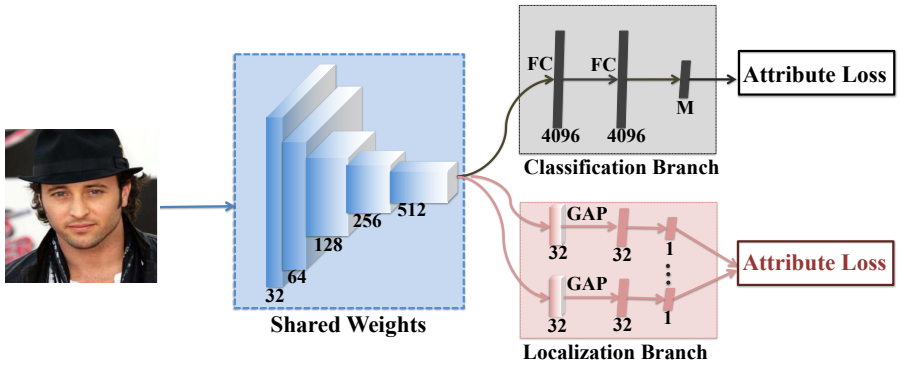


Figure 4.2: Multi-Net Learning.

group fully-connected structure, where the j-th output is only connected with the j-th attribute scores predicted by all subnets. Especially, it could balance the scores from the part-based and whole-image-based subnets by putting more weight to the one that is more important. An attribute relation layer (ARL), which is a fully-connected layer, then takes these $r_j, j \in 1, ..., M$ as input to predict the final score for each face attribute. ARL here is used to further model the high correlations among the face attributes. The PaW network is trained end-to-end with the sigmoid cross entropy loss: $L_{attr} = \sum_{j=1}^{M} y_j \log o_j + (1-y_j) \log(1-o_j)$, where y_j's are the attributes labels, and o_j's are the outputs from the ARL layer.

4.3.2.2 Hint-based Model Compression

Training the PaW network in a naive way is both memory demanding and time consuming, since the total number of network parameters increases substantially as the number of attributes becomes large, and the subnet architecture adapted from the FRL network is not very compact. To obtain a compact subnet model, we

further propose a model compression technique. Motivated by [107, 108], we design a hint loss to make the student net (SNet) reconstruct the feature maps from the teacher net (TNet). It can be expressed as:

$$L_{hint}(w) = ||T_k(I) - S_l(I, w)||_2, \tag{4.1}$$

where k (l) is the chosen layer of the teacher (student) net to transfer (add) supervision, w are the weights of the student net to be learned, and I is the whole face image. The network architecture is shown in Fig. 4.3. Besides the hint loss, the student network is also supervised by the attributes loss. Thus, the total loss function can be written as $L_S = \lambda_1 L_{hint} + \lambda_2 L_{attr}$. The FRL network trained by MNL is adopted as the teacher network to teach the whole-image-based subnet (or the student net). Since it is fully-convolutional and deeper layer generally captures high-level semantics [109, 110], we set the supervision layer k to be the teacher network's last convolutional layer. During training, the weights of the teacher network are frozen, and only the student network is learned. The training algorithm is carried out in two stages: first setting $\lambda_1 = 1, \lambda_2 = 0$, and training S with only the hint loss. In this way, the knowledge of the teacher network could help the student network find a good initialization. Then we set $\lambda_1 = 0, \lambda_2 = 1$ and train S with attribute loss only. After the whole-image-based subnet is learned, its weights are used to initialize all the part-based subnets in PaW.

4.3.3 Training Methodology

The training process is carried out as follows:

1. First, MNL is adopted to train the FRL network with superior classification performance;
2. Then hint-based compression method is applied to train a compact whole-image-based subnet g_0 using the learned FRL network as the teacher net.
3. Initialize each part-based subnet $\{g_i\}_{i=1}^{M}$ using the weights from g_0 and then train each subnet g_i independently using the corresponding attentional face region;
4. By fixing all the part-based subnets and the whole-image-based subnet, the RSL and ARL are learned;
5. Finally, the PaW network is fine-tuned by back-propagating errors from ARL to all the lower layers of the part-based subnets and the whole-image-based subnet.

All the subnets and the two layer fully-connected model are trained under the supervision of attribute labels. The third and forth steps initialize the classification model to be close to a good local minimum, which is important for the successful training of PaW.

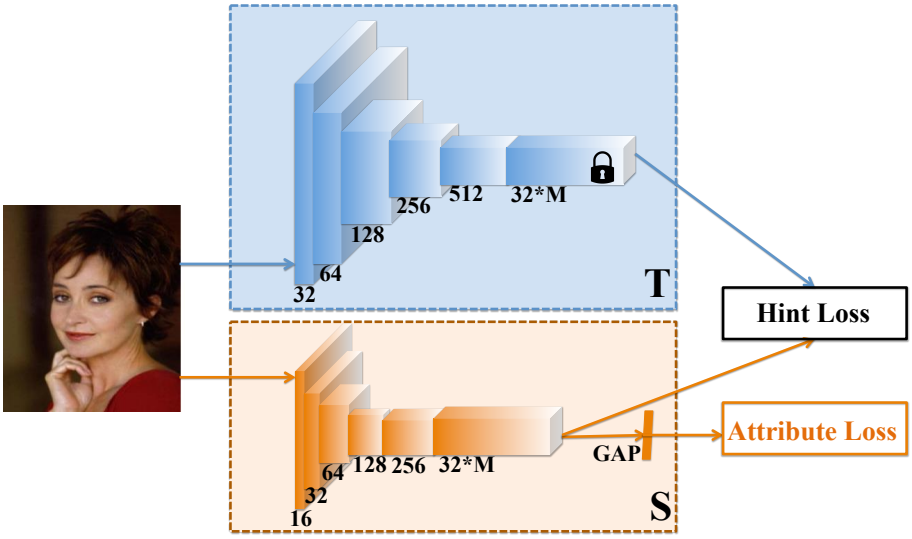


Figure 4.3: Hint-based Model Compression.

4.4 Experiments

4.4.1 Dataset

We use the CelebA dataset [92] in our experiments, since it has been widely used for face attributes classification. It consists of 202,599 face images collected from the Internet and annotated with 40 binary attributes. As suggested in [92], 162,770 of these images are used for training, 19,867 and 19,962 are reserved for validation and testing respectively. Both unaligned and aligned sets are provided and we applied our method on the unaligned one (**uCelebA**). To conduct experiments on uCelebA, we use the publicly available face detector [41] to detect faces. For 560 images which have no face detected, we use the provided landmarks to get the groundtruth bounding box (we empirically expand the minimum bounding box containing all landmarks twice to cover the neck and hair region). For 15,181 images

with multiple faces detected, we select the bounding box that has maximum overlap with the groundtruth bounding box. This is the only preprocessing step applied to the unaligned images.

4.4.2 Implementation details

We applied MNL to train the FRL network. The learning rate is fixed to be 0.0001, and the network is trained for 10 epochs with batch size of 128. The FRL network is then compressed with a learning rate of $1e^{-7}$ for the hint loss training and 0.0001 for the attribute loss training. The part-based subnets are trained for 15 epochs with the weights initialized from the whole-image-based subnet. After that, the RSL and ARL are trained with a learning rate of 0.1 with all subnets fixed. Finally, a learning rate of 0.001 is applied to train the PaW network in an end-to-end manner. Stochastic gradient descent (SGD) is used to train all the networks. The momentum and weight decay are set at 0.9 and 0.0005 for all the experiments respectively. Horizontal flipping is applied for data augmentation. We use Caffe [43] to implement our networks.

4.4.3 Ablative Analysis

4.4.3.1 Face Region Localization

In this section, we evaluate the FRL network qualitatively. Fig. 4.4 shows the location heatmaps corresponding to several attributes. We observe that the localized parts are quite semantically meaningful, even though some face images

have large pose variations or under occlusion. For example, the eye area produces the highest response for the *Arched Eyebrow* attribute even though the woman wears sunglasses. While for the attribute of *Wavy Hair*, the network localizes the head region although the man wears a hat. We also examine it quantitatively in the **Classification Results** section to show that accurate region localization is essential for good classification results.

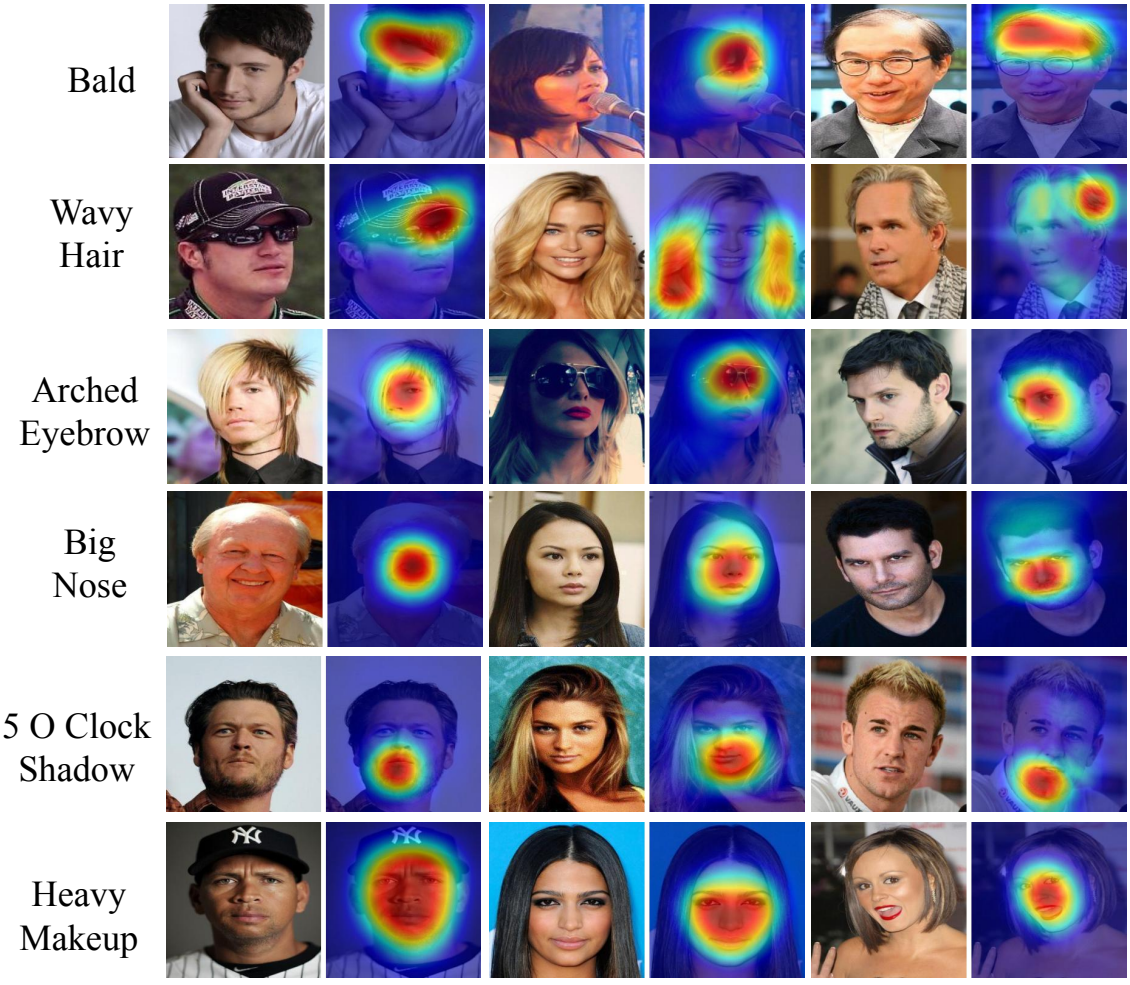


Figure 4.4: Location heatmaps from the face region localization network. Face regions that correlate with facial attributes are discovered.

Table 4.1: Average classification accuracy on uCelebA dataset.

Methods	Classif. Branch	Loc. Branch
Without MNL	-	91.01
MNL	91.05	**91.07**

Table 4.2: Fine-grained classification accuracy on CUB-200 dataset.

Methods	Classif. Branch	Loc. Branch
Without MNL on full image	-	67.40
MNL on full image	72.10	**71.66**
Without MNL on crop	-	71.90
MNL on crop	75.76	**76.03**

4.4.3.2 Multi-Net Learning

In this section, we study the ability of MNL for obtaining a localizable and discriminative deep representation. Table 5.1 summarizes the attribute classification results from classification and localization branches. We find that MNL consistently improves the classification performance of the localization branch, achieving an accuracy of 91.07% vs. 91.01% with/without MNL.

To further test the proposed MNL, we applied it on the popular CUB-200-2011 dataset [111] for fine-grained object recognition. The dataset contains 11,788 images, with 5,994 images for training and 5,794 for testing. The network architecture is same as the one used in uCelebA, except that the last layer is replaced with 200 output nodes (the number of classes). The weights are initialized from VGGNet [82]. Table 4.2 summarizes the results. We find that the localization branch performs worse than the classification branch, with almost 4% performance gap. After applying MNL, the accuracy of the localization branch is improved from 67.40% to 71.66% when using the full image. We also adopt the same localization

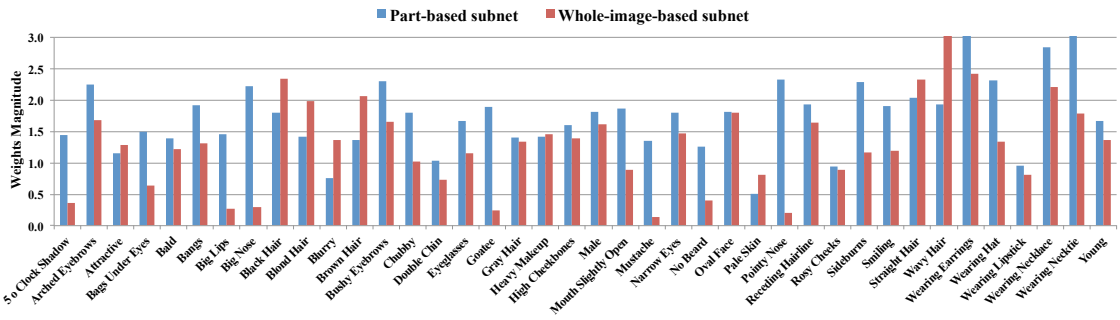


Figure 4.5: Visualization of the region switch layer weights. For each attribute, the blue and the red bar represent the weight values of RSL that corresponds to the part-based subnet and whole-image-based subnet respectively. It shows that the weights of the part-based subnets are higher for the local attributes. For global attributes, the whole-image-based subnet is assigned larger weight.

technique as [101] to identify the bounding box of the birds in both the training and testing sets. With the cropped bird images as training data, the performance of the localization branch is further improved from 71.90% to 76.03%. This further demonstrates that MNL is able to improve the discriminativeness of the GAP-based localization network.

4.4.3.3 Hint-based Model Compression

In this section, we analyze the effectiveness of our model compression technique. To show the flexibility and robustness of our method, we experiment with three student nets (SNet1, SNet2 and SNet3) with different sizes. Table 4.3 shows the network architectures and their classification results. We use $s \times s \times n(t)$ to denote kernel size $s \times s$ with n output feature maps, where t is the number of repeated convolution modules. We observe that the proposed method is able to compress a deep network to a relatively shallow network, with little performance drop. For SNet3, which achieves an accuracy of 90.60%, the depth is shortened from 14 to 5,

Table 4.3: Comparison of average accuracy and compactness between different compressed models on uCelebA dataset.

Layer	TNet	SNet1	SNet2	SNet3
Conv1	3x3x32(2)	3x3x32	3x3x32	3x3x16
Pool1	2x2x32	2x2x32	2x2x32	2x2x16
Conv2	3x3x64(2)	3x3x64	3x3x64	3x3x32
Pool2	2x2x64	2x2x64	2x2x64	2x2x32
Conv3	3x3x128(3)	3x3x128	3x3x128	3x3x64
Pool3	2x2x128	2x2x128	2x2x128	2x2x64
Conv4	3x3x256(3)	3x3x256	3x3x256	3x3x128
Pool4	2x2x256	2x2x256	2x2x256	2x2x128
Conv5	3x3x512(3)	3x3x512	3x3x512	1x1x1280
Conv6	3x3x1280	3x3x1280	1x1x1280	n/a
Classifier	GAP	GAP	GAP	GAP
	FC40	FC40	FC40	FC40
Accuracy	91.07	91.02	90.89	90.60
Param.	19M	6M	2M	0.27M

Table 4.4: Comparison of average accuracy and compactness on the aligned CelebA dataset.

Method	Accuracy	Param.
SOMP [112]-thin-32	89.96	0.22M
SOMP [112]-branch-32	90.74	1.49M
Low Rank [113]	90.88	4.52M
SNet3	**90.89**	**0.27M**

and the number of parameters is reduced from 19M to 0.27M.

To further compare our approach with existing methods, we also train our models on the *aligned* CelebA dataset. The results are summarized in Table 4.4. We find that our SNet3 model achieves similar or better accuracy compared to these state-of-the-art methods, while being much more compact and thus faster.

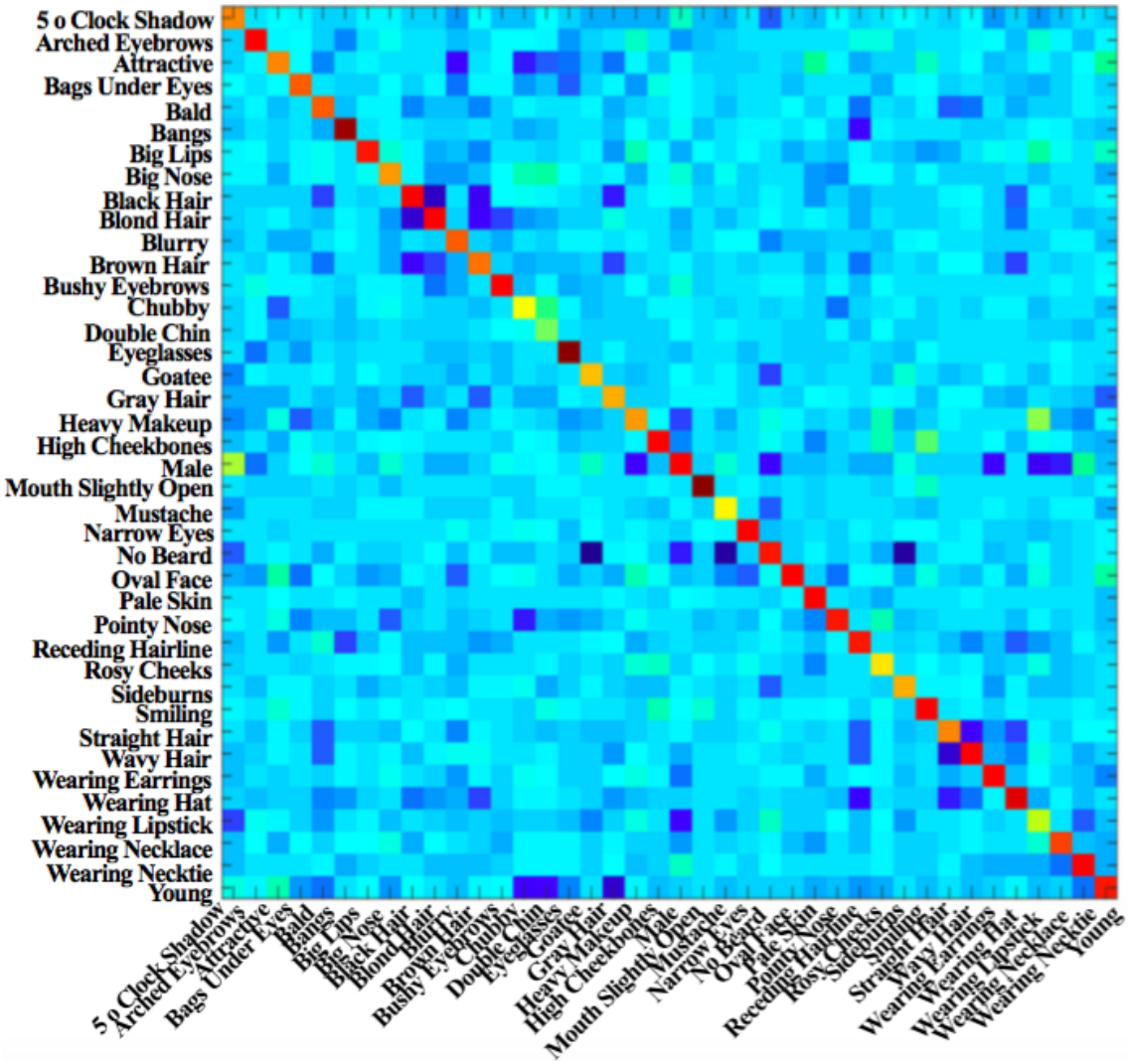


Figure 4.6: Attribute relation weights learned on uCelebA dataset. Red and yellow colors indicate high values while blue and green colors denote low values.

4.4.3.4 PaW Classification Network

In this section, we evaluate the classification performance of the proposed PaW network. Before showing the results, we first explore whether RSL assigns appropriate weights to different subnets for attribute prediction and whether ARL learns meaningful attributes correlations.

Face Region Selection We visualize the weights of RSL in Fig. 4.5. Although

Table 4.5: Performance comparison with state of the art methods on 40 binary facial attributes. The best results are shown in bold.

		5 o Clock Shadow	Arched Eyebrows	Attractive	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	High Cheekbones	Male
uCelebA	LNets+ANet [92]	91.00	79.00	81.00	79.00	98.00	95.00	68.00	78.00	88.00	95.00	84.00	80.00	90.00	91.00	92.00	99.00	95.00	97.00	90.00	87.00	98.00
	Part-only	93.90	81.86	81.88	84.07	98.72	95.71	70.63	83.48	87.97	95.16	95.83	87.53	91.73	95.05	95.92	99.46	97.19	97.93	90.26	86.20	96.65
	Whole-only	93.95	81.43	82.06	84.11	98.57	95.45	70.66	82.91	89.08	95.52	96.01	88.63	92.32	95.12	95.98	99.40	96.90	98.07	90.67	86.57	97.10
	PaW	**94.64**	**83.01**	**82.86**	**84.58**	**98.93**	**95.93**	**71.46**	**83.63**	**89.84**	**95.85**	**96.11**	**88.50**	**92.62**	**95.46**	**96.26**	**99.59**	**97.38**	**98.21**	**91.53**	**87.44**	**98.39**

		Mouth Slightly Open	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Receding Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wearing Earrings	Wearing Hat	Wearing Lipstick	Wearing Necklace	Wearing Necktie	Young		**Average**
uCelebA	LNets+ANet [92]	92.00	95.00	81.00	95.00	66.00	91.00	72.00	89.00	90.00	96.00	92.00	73.00	80.00	82.00	99.00	93.00	71.00	93.00	87.00		87.30
	Part-only	93.55	96.63	86.96	95.71	73.03	96.86	76.40	92.87	94.77	97.63	91.98	82.53	81.29	89.07	98.75	92.96	87.13	96.69	86.51		90.46
	Whole-only	93.24	96.59	87.19	95.40	74.48	96.85	76.06	92.95	94.83	97.50	91.61	82.18	82.63	89.13	98.50	93.58	87.14	96.77	87.14		90.60
	PaW	**94.05**	**96.90**	**87.56**	**96.22**	**75.03**	**97.08**	**77.35**	**93.44**	**95.07**	**97.64**	**92.73**	**83.52**	**84.07**	**89.93**	**99.02**	**94.24**	**87.70**	**96.85**	**88.59**		**91.23**

each subnet predicts M attribute scores simultaneously, only the weights of the corresponding part-based subnet against the whole-image-based subnet are shown here. The weight magnitude indicates the importance of the subnet for predicting the attribute. Interestingly, we find that the part-based subnet related to the local attribute, *e.g. 5 o Clock Shadow* and *Bushy Eyebrows*, is always assigned the largest weight among the $M + 1$ subnets. We also observe that for global attributes, *e.g. Attractive*, *Blurry*, *Heavy Makeup*, and *Pale Skin*, the whole-image-based subnet achieves the highest weight. Intuitively those global attributes should obtain more information from the whole-image-based subnet. This validates the region selection ability of the RSL.

Face Attribute Correlation The learned ARL weights are visualized in Fig. 4.6. We find that attribute pairs that are mutually exclusive such as (*Attractive, Blurry*), (*Black Hair, Blond Hair*) and (*No Beard, Goatee*) are assigned lowest weights. Rarely co-occurring attribute pairs like (*Male, Heavy Makeup*) are also assigned low weights. Pairs of attributes such as (*Chubby, Double Chin*), (*Heavy Makeup, Wearing Lipstick*) and (*Smiling, High Cheekbones*) that commonly co-occur are

given relatively higher weights. Moreover, the weights are asymmetric, for example, a person who wears lipstick is very unlikely to have a beard, but not the other way round. This is also reflected in the learned weights. This shows that ARL captures the attribute relationships.

Classification Results We show that our model achieves state-of-the-art results on uCelebA dataset. In the following experiments, each subnet adopts the architecture of SNet3 in Table 4.3.

We compare PaW with two baselines:

1. Part-only: each part net is trained on the detected face region to predict all face attributes. Then the attribute score from the most related part-based subnet is adopted for testing.

2. Whole-only: this method does not have part nets. It is trained with the whole face image only and is used to directly predict all attributes.

Table 4.5 summarizes the classification performances. We observe that the PaW net performs consistently better than either the Part-only or Whole-only method alone, achieving an accuracy of 91.23% vs. 90.60% for Part-only and 90.46% for Whole-only on uCelebA. This shows that RSL learns to selectively combine information from part-based and whole-image-based subnets. For unaligned face attribute classification on uCelebA dataset, we achieve the highest recognition rates across the board on all attributes and decrease the average recognition error from 12.70% to 8.77%, a reduction of 30.9%. Our method on the aligned CelebA also achieves an accuracy of 91.33% vs. 90.94% compared with the state-of-the-art [94]. This validates the effectiveness of the proposed attribute classification

network. Also, the small performance gap on uCelebA and the aligned CelebA means that we practically eliminate the alignment step, and hence no special annotations are needed. Although the PaW network contains multiple part-based and whole-image-based subnets, the total number of parameters is only 11 M.

To test the importance of the FRL network, we further employ a baseline that divides each image into 4×4 non-overlapping blocks to simulate crude part detectors. Then part-based subnets and whole-image-based subnet are trained the same way as before. It achieves an average accuracy of 90.95% on uCelebA. However, we found that the weights corresponding to the whole-image-based net in the RSL are always higher than those of the part-based subnets for predicting *all* the attributes. This is because coarse region localization makes the part-based subnets unreliable, thus all the predictions are essentially made by the whole-image-based subnet only. This validates the effectiveness of the proposed FRL network.

4.5 Conclusions

In this chapter, we propose to learn attentional face regions to improve attribute classification under unaligned condition. To this end, a weakly-supervised face region localization network is first designed. Then the information from those detected regions are selectively combined by the hybrid classification network. Visualization shows our method not only discovers semantic meaningful attributes regions, but also captures rich correlations among attributes. Moreover, our results outperform previous methods on the unaligned CelebA dataset by a large margin.

Chapter 5: Facial Expression Editing with Controllable Expression Intensity

5.1 Introduction

Facial expression editing is the task that transforms the expression of a given face image to a target one without affecting the identity properties. It has applications in facial animation, human-computer interactions, entertainment, etc. The area has been attracting considerable attention both from academic and industrial research communities.

Existing methods that address expression editing can be divided into two categories. One category tries to manipulate images by reusing parts of existing ones [114, 115, 116] while the other resorts to synthesis techniques to generate a face image with the target expression [52, 117, 118]. In the first category, traditional methods [114] often make use of the expression flow map to transfer an expression by image warping. Recently, Yeh et al. [116] applied the idea to a variational autoencoder to learn the flow field. Although the generated face image has high resolution, paired data where one subject has different expressions are needed to train the model. In the second category, deep learning-based methods are mainly

used. The early work by Susskind et al. [117] used a deep belief network to generate emotional faces, which can be controlled by the Facial Action Coding System (FACS) labels. In [52], a three-way gated Boltzmann machine was employed to model the relationships between the expression and identity. However, the synthesized images of these methods have low resolution (48 x 48), lacking fine details and tend to be blurry.

Moreover, existing works can only transform the expression to different classes, like *Angry* or *Happy*. However, in reality, the intensity of facial expression is often displayed over a range. For example, humans can express the *Happy* expression either with a huge grin or by a gentle smile. Thus it is appealing if both the type of the expression and its intensity can be controlled simultaneously. Motivated by this, in this chapter, we present a new expression editing model, Expression Generative Adversarial Network (ExprGAN) which has the unique property that multiple diverse styles of the target expression can be synthesized where the intensity of the generated expression can be continuously controlled from weak to strong, without the need for training data with intensity values.

To achieve this goal, we specially design an expression controller module. Instead of feeding in a deterministic one-hot vector label like previous works, the expression code generated by the expression controller module is used. It is a real-valued vector conditioned on the label, thus more complex information such as expression intensity can be described. Moreover, to force each dimension of the expression code to capture a different factor of the intensity variations, the conditional mutual information between the generated image and the expression code is

maximized by a regularizer network.

Our work is inspired by the recent success of the image generative model, where a generative adversarial network [11] learns to produce samples similar to a given data distribution through a two-player game between a generator and a discriminator. Our ExprGAN also adopts the generator and discriminator framework in addition to the expression controller module and the regularizer network. However, to facilitate image editing, the generator is composed of an encoder and a decoder. The input of the encoder is a face image, the output of the decoder is a reconstructed one, and the learned identity and expression representations bridge the encoder and decoder. To preserve the most prominent facial structure, we adopt a multi-layer perceptual loss [14] in the feature space in addition to the pixel-wise L_1 loss. Moreover, to make the synthesized image look more photo-realistic, two adversarial networks are imposed on the encoder and decoder, respectively. Because it is difficult to directly train our model using the small training set, a three-stage incremental learning algorithm is also developed.

Contributions: We propose a novel model called ExprGAN that can change a face image to a target expression with multiple styles, where the expression intensity can also be controlled continuously. We show that the synthesized face images have high perceptual quality, which can be used to improve the performance of an expression classifier. Our identity and expression representations are explicitly disentangled which can be exploited for tasks such as expression transfer, image retrieval, etc. We develop an incremental training strategy to train the model on a relative small dataset without the rigid requirement of paired samples.

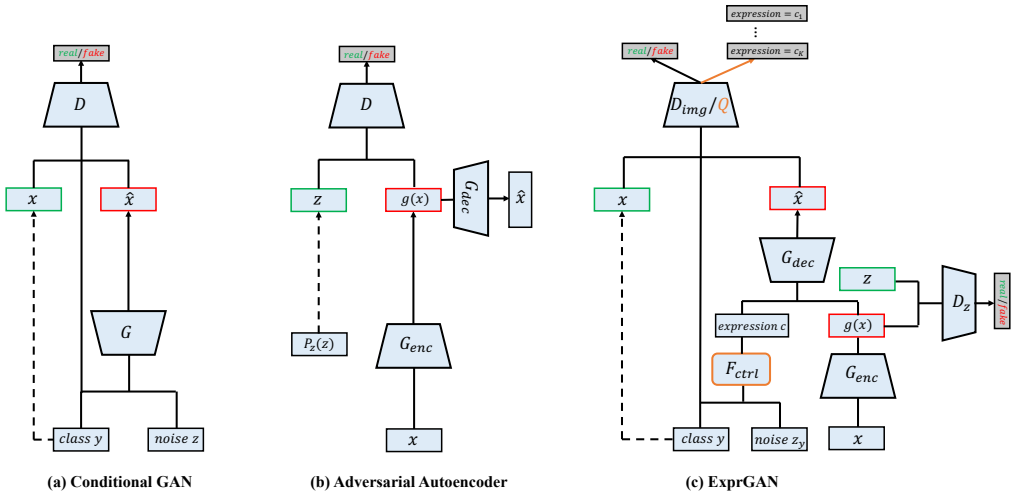


Figure 5.1: Comparison of previous GAN architectures and the proposed ExprGAN.

Organizations: Section 5.2 provides an overview of existing works on deep generative model and facial expression editing. Section 5.3 presents the proposed facial expression generative adversarial network. Experiment results and conclusions are presented in Section 5.4 and Section 5.5, respectively.

5.2 Related Works

5.2.1 Deep Generative Model

Deep generative models have achieved impressive success in recent years. There are two major approaches: generative adversarial network (GAN) [11] and variational autoencoder (VAE) [119]. GAN is composed of a generator and a discriminator, where the training is carried out with a minimax two-player game. GAN has been used for image synthesis [12], image superresolution [120], etc. One interesting extension of GAN is Conditional GAN (CGAN) [121] where the generated image can be controlled by the condition variable. On the other hand, VAE is a probabilistic

model with an encoder to map an image to a latent representation and a decoder to reconstruct the image. A reparametrization trick is proposed which enables the model to be trained by backpropogation [122]. One variant of VAE is Adversarial Autoencoder [123], where an adversarial network is adopted to regularize the latent representation to conform to a prior distribution. ExprGAN also adopts an autoencoder structure, but there are two main differences: First, an expression controller module is specially designed, so a face with different types of expressions across a wide range of intensities can be synthesized. Second, to improve the generated image quality, a face identity preserving loss and two adversarial losses are incorporated.

5.2.2 Facial Expression Editing

Facial expression editing has been actively investigated in computer graphics. Traditional approaches include 3D model-based [124], 2D expression mapping-based [125] and flow-based [114]. Recently, deep learning-based methods have been proposed. Susskind et al. [117] studied a deep belief network to generate facial expression given high-level identity and facial action unit (AU) labels. In [52], a higher-order Boltzman machine with multiplicative interactions was proposed to model the distinct factors of variation. Cheung et al. [118] proposed a decorrelating regularizer to disentangle the variations between identity and expression in an unsupervised manner. However, the generated image is low resolution with size of 48 x 48, which is not visually satisfying. Recently, Yeh et al. [116] proposed to edit the facial expression by image warping with appearance flow. Although the model can

generate high-resolution images, paired samples as well as the labeled query image are required.

The most similar work to ours is CFGAN [126], which uses a filter module to control the generated face attributes. However, there are two main differences: First, CFGAN adopts the CGAN architecture where an encoder needs to be trained separately for image editing, while for the proposed ExprGAN, the encoder and the decoder are constructed in a unified framework. Second, the attribute filter of CFGAN is mainly designed for a single class, while our expression controller module works for multiple categories. Most recently, Zhang et al. [127] proposed a conditional AAE (CAAE) for face aging, which can also be applied for expression editing. Compared with these studies, ExprGAN has two main differences: First, in addition to transforming a given face image to a new facial expression, our model can also control the expression intensity continuously without the intensity training labels; Second, photo-realistic face images with new identities can be generated for data augmentation, which is found to be useful to train an improved expression classifier.

5.3 Proposed Method

In this section, we describe the architecture of ExprGAN. We first describe the Conditional Generative Adversarial Network (CGAN) [121] and the Adversarial Autoencoder (AAE) [123], which form the basis of ExprGAN. Then the design of ExprGAN is detailed. The architectures of the three models are shown in Fig. 5.1.

5.3.1 Conditional Generative Adversarial Network

CGAN is an extension of a GAN [11] for conditional image generation. It is composed of two networks: a generator network G and a discriminator network D that compete in a two-player minimax game. Network G is trained to produce a synthetic image $\hat{x} = G(z, y)$ to fool D to believe it is an actual photograph, where z and y are the random noise and condition variable, respectively. D tries to distinguish the real image x and the generated one $\hat{x}$. Mathematically, the objective function for G and D can be written as follows:

$$\begin{aligned} \min_G \max_D \; & \mathbb{E}_{x,y\sim P_{data}(x,y)}[\log D(x,y)] \\ & + \mathbb{E}_{z\sim P_z(z),y\sim P_y(y)}[\log(1 - D(G(z,y),y))] \end{aligned} \tag{5.1}$$

5.3.2 Adversarial Autoencoder

AAE [123] is a probabilistic autoencoder which consists of an encoder G_{enc}, a decoder G_{dec} and a discriminator D. Apart from the reconstruction loss, the hidden code vector $g(x) = G_{enc}(x)$ is also regularized by an adversarial network to impose a prior distribution $P_z(z)$. Network D aims to discriminate $g(x)$ from $z \sim P_z(z)$, while G_{enc} is trained to generate $g(x)$ that could fool D. Thus, the AAE objective function becomes:

$$\begin{aligned} & \min_{G_{enc},G_{dec}} \max_D L_p(G_{dec}(G_{enc}(x)),x) \\ & + \mathbb{E}_{z\sim P_z(z)}[\log D(z)] + \mathbb{E}_{x\sim P_{data}(x)}[\log(1 - D(G_{enc}(x))] \end{aligned} \tag{5.2}$$

where $L_p(,)$ is the p_{th} norm: $L_p(x', x) = ||x' - x||_p^p$

5.3.3 Expression Generative Adversarial Network

Given a face image x with expression label y, the objective of our learning problem is to edit the face to display a new type of expression at different intensities. Our approach is to train a ExprGAN conditional on the original image x and the expression label y with its architecture illustrated in Fig. 5.1 (c).

5.3.3.1 Network Architecture

ExprGAN first applies an encoder G_{enc} to map the image x to a latent representation $g(x)$ that preserves identity. Then, an expression controller module F_{ctrl} is adopted to convert the one-hot expression label y to a more expressive expression code c. To further constrain the elements of c to capture the various aspects of the represented expression, a regularizer Q is exploited to maximize the conditional mutual information between c and the generated image. Finally, the decoder G_{dec} generates a reconstructed image $\hat{x}$ combining the information from $g(x)$ and c. To further improve the generated image quality, a discriminator D_{img} on the decoder G_{dec} is used to refine the synthesized image $\hat{x}$ to have photo-realistic textures. Moreover, to better capture the face manifold, a discriminator D_z on the encoder G_{enc} is applied to ensure the learned identity representation is filled and exhibits no "holes" [123].

5.3.3.2 Expression Controller Networks $\boldsymbol{F_{ctrl}}$ and $\boldsymbol{Q}$

In previous conditional image generation methods [127, 128], a binary one-hot vector is usually adopted as the condition variable. This is enough for generating images corresponding to different categories. However, for our problem, a stronger control over the synthesized facial expression is needed: we want to change the expression intensity in addition to generating different types of expressions. To achieve this goal, an expression controller module F_{ctrl} is designed to ensure the expression code c can describe the property of the expression intensity except the category information. Furthermore, a regularizer network Q is proposed to enforce the elements of c to capture the multiple levels of expression intensity comprehensively.

Expression Controller Module $\boldsymbol{F_{ctrl}}$ To enhance the description capability, F_{ctrl} transforms the binary input y to a continuous representation c by the following operation:

$$c_i = F_{ctrl}(y_i, z_y) = |z_y| \cdot (2y_i - 1) \qquad i = 1, 2, \ldots, K \tag{5.3}$$

where the inputs are the expression label $y \in \{0, 1\}^K$ and uniformly distributed $z_y \sim U(-1, 1)^d$, while the output is the expression code $c = [c_1^T, \ldots, c_K^T]^T \in R^{Kd}$, K is the number of classes. If the i_{th} class expression is present, *i.e.*, $y_i = 1$, $c_i \in R^d$ is set to be a positive vector within 0 and 1, while $c_j, j \neq i$ has negative values from -1 to 0. Thus, in testing, we can manipulate the elements of c to generate the desired expression type. This flexibility greatly increases the controllability of c over synthesizing diverse styles and intensities of facial expressions.

Regularizer on Expression Code Q It is desirable if each dimension of c could learn a different factor of the expression intensity variations. Then faces with a specific intensity level can be generated by manipulating the corresponding expression code. To enforce this constraint, we impose a regularization on c by maximizing the conditional mutual information $I(c; \hat{x}|y)$ between the generated image $\hat{x}$ and the expression code c. This ensures that the expression type and intensity encoded in c is reflected in the image generated by the decoder. The direct computation of I is hard since it requires the posterior $P(c|\hat{x}, y)$, which is generally intractable. Thus, a lower bound is derived with variational inference which extends [129] to the conditional setting:

$$
\begin{aligned}
& I(c; \hat{x}|y) \\
& = H(c|y) - H(c|\hat{x}, y) \\
& = \mathbb{E}_{\hat{x}\sim G_{dec}(g(x),c)}[\mathbb{E}_{c'\sim P(c'|\hat{x},y)}[\log P(c'|\hat{x}, y)]] + H(c|y) \\
& = \mathbb{E}_{\hat{x}\sim G_{dec}(g(x),c)}[D_{KL}(P(\cdot|\hat{x}, y)||Q(\cdot|\hat{x}, y))+ \\
& \quad \mathbb{E}_{c'\sim P(c'|\hat{x},y)}[\log Q(c'|\hat{x}, y)]] + H(c|y) \\
& \geq \mathbb{E}_{\hat{x}\sim G_{dec}(g(x),c)}[\mathbb{E}_{c'\sim P(c'|\hat{x},y)}[\log Q(c'|\hat{x}, y)]] + H(c|y) \\
& = \mathbb{E}_{c\sim P(c|y),\hat{x}\sim G_{dec}(g(x),c)}[\log Q(c|\hat{x}, y)] + H(c|y)
\end{aligned}
\tag{5.4}
$$

For simplicity, the distribution of c is fixed, thus $H(c|y)$ is treated as a constant. Here the auxiliary distribution Q is parameterized as a neural network, thus the

final loss function is defined as follows:

$$\min_{Q} L_Q = -\mathbb{E}_{c\sim P(c|y),\hat{x}\sim G_{dec}(g(x),c)}[\log Q(c|\hat{x},y)] \tag{5.5}$$

5.3.3.3 Generator Network $\boldsymbol{G}$

The generator network $G = (G_{enc}, G_{dec})$ adopts the autoencoder structure where the encoder G_{enc} first transforms the input image x to a latent representation that preserves as much identity information as possible. After obtaining the identity code $g(x)$ and the expression code c, the decoder G_{dec} then generates a synthetic image $\hat{x} = G_{dec}(G_{enc}(x), c)$ which should be identical as x. For this purpose, a pixel-wise image reconstruction loss is used:

$$\min_{G_{enc},G_{dec}} L_{pixel} = L_1(G_{dec}(G_{enc}(x), c), x) \tag{5.6}$$

To further preserve the face identity between x and $\hat{x}$, a pre-trained discriminative deep face model is leveraged to enforce the similarity in the feature space:

$$\min_{G_{enc},G_{dec}} L_{id} = \sum_{l} \beta_l L_1(\phi_l(G_{dec}(G_{enc}(x), c)), \phi_l(x)) \tag{5.7}$$

where ϕ_l are the l_{th} layer feature maps of a face recognition network, and β_l is the corresponding weight. We use the activations at the $conv1\_2$, $conv2\_2$, $conv3\_2$, $conv4\_2$ and $conv5\_2$ layer of the VGG face model [42].

5.3.3.4 Discriminator on Identity Representation $\boldsymbol{D_z}$

It is a well known fact that face images lie on a manifold [130, 131]. To ensure that face images generated by interpolating between arbitrary identity representations do not deviate from the face manifold [127], we impose a uniform distribution on $g(x)$, forcing it to populate the latent space evenly without "holes". This is achieved through an adversarial training process where the training objective is:

$$\begin{aligned}\min_{G_{enc}} \max_{D_z} L^z_{adv} =& \mathbb{E}_{z\sim P_z(z)}[\log D_z(z)] \\ &+ \mathbb{E}_{x\sim P_{data}(x)}[\log(1 - D_z(G_{enc}(x))]\end{aligned} \tag{5.8}$$

5.3.3.5 Discriminator on Image $\boldsymbol{D_{img}}$

Similar to existing methods [128, 132], an adversarial loss between the generated image $\hat{x}$ and the real image x is further adopted to improve the photorealism:

$$\begin{aligned}\min_{G_{enc},G_{dec}} \max_{D_{img}} L^{img}_{adv} = \mathbb{E}_{x,y\sim P_{data}(x,y)}[\log D_{img}(x, y)]+ \\ \mathbb{E}_{x,y\sim P_{data}(x,y),z_y\sim P_{z_y}(z_y)} \\ [\log(1 - D_{img}(G_{dec}(G_{enc}(x), F_{ctrl}(z_y, y)), y))]\end{aligned} \tag{5.9}$$

5.3.3.6 Overall Objective Function

The final training loss function is a weighted sum of all the losses defined above:

$$\begin{aligned}\min_{G_{enc},G_{dec},Q}\max_{D_{img},D_z} L_{ExprGAN} = L_{pixel} + \lambda_1 L_{id} + \lambda_2 L_Q \\ + \lambda_3 L_{adv}^{img} + \lambda_4 L_{adv}^{z} + \lambda_5 L_{tv}\end{aligned} \tag{5.10}$$

We also impose a total variation regularization L_{tv} [133] on the reconstructed image to reduce spike artifacts.

5.3.3.7 Incremental Training

Empirically we find that jointly training all the subnetworks yields poor results as we have multiple loss functions. It is difficult for the model to learn all the functions at one time considering the small size of the dataset. Therefore, we propose an incremental training algorithm to train the proposed ExprGAN. Overall our incremental training strategy can be seen as a form of curriculum learning, and includes three stages: controller learning stage, image reconstruction stage and image refining stage. First, we teach the network to generate the image conditionally by training G_{dec}, Q and D_{img} where the loss function only includes L_Q and L_{adv}^{img}. $g(x)$ is set to be random noise in this stage. After the training finishes, we then teach the network to learn the disentangled representations by reconstructing the input image with G_{enc} and G_{dec}. To ensure that the network does not forget what is

already learned, Q is also trained but with a decreased weight. So the loss function has three parts: L_{pixel}, L_{id} and L_Q. Finally, we train the whole network to refine the image to be more photo-realistic by adding D_{img} and D_z with the loss function defined in Equation (5.10). We find in our experiments that stage-wise training is crucial to learn the desired model on the small dataset.

5.4 Experiments

We first describe the experimental setup and then the three main applications: expression editing with continuous control over intensity, facial expression transfer and conditional face image generation for data augmentation .

5.4.1 Dataset

We evaluated the proposed ExprGAN on the widely used Oulu-CASIA [6] dataset. Oulu-CASIA has 480 image sequences taken under Dark, Strong, Weak illumination conditions. In this experiment, only videos with Strong condition captured by a VIS camera are used. There are 80 subjects and six expressions, *i.e.*, *Angry*, *Disgust*, *Fear*, *Happy*, *Sad* and *Surprise*. The first frame is always neutral while the last frame has the peak expression. Only the last three frames are used, and the total number of images is 1440. Training and testing sets are divided based on identity, with 1296 for training and 144 for testing. We aligned the faces using the landmarks detected from [41], then cropped and resized the images to dimension of 128 x 128. Lastly, we normalized the pixel values into range of [-1, 1]. To alleviate

overfitting, we augmented the training data with random flipping.

5.4.2 Implementation Details

ExprGAN mainly builds on multiple upsampling and downsampling blocks. The upsampling block consists of the nearest-neighbor upsampling followed by a 3 x 3 stride 1 convolution. The downsampling block consists of a 5 x 5 stride 2 convolution. Specifically, G_{enc} has 5 downsampling blocks where the numbers of channels are 64, 128, 256, 512, 1024 and one FC layer to get the identity representation $g(x)$. For G_{dec}, it has 7 upsampling blocks with 512, 256, 128, 64, 32, 16, 3 channels. D_z consists of 4 FC layers with 64, 32, 16, 1 channels. We model $Q(c|\hat{x}, y)$ as a factored Gaussian, and share many parts of Q with D_{img} to reduce computation cost. The shared parts have 4 downsampling blocks with 16, 32, 64, 128 channels and one FC layer to output a 1024-dim representation. Then it is branched into two heads, one for D_{img} and one for Q. Q has K branches $\{Q_i\}_{i=1}^{K}$ where each Q_i has two individual FC layers with 64, d channels to predict the expression code c_i. Leaky ReLU [134] and batch normalization [3] are applied to D_{img} and D_z, while ReLU [62] activation is used in G_{enc} and G_{dec}. The random noise z is uniformly distributed from -1 to 1. We fixed the dimensions of $g(x)$ and c to be 50 and 30, and found this configuration sufficient for representing the identity and expression variations.

We train the networks using the Adam optimizer [135], with learning rate of 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$ and mini-batch size of 48. In the image refining stage, we empirically set $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.01$, $\lambda_4 = 0.01$, $\lambda_5 = 0.001$. The model is

implemented using Tensorflow [136].

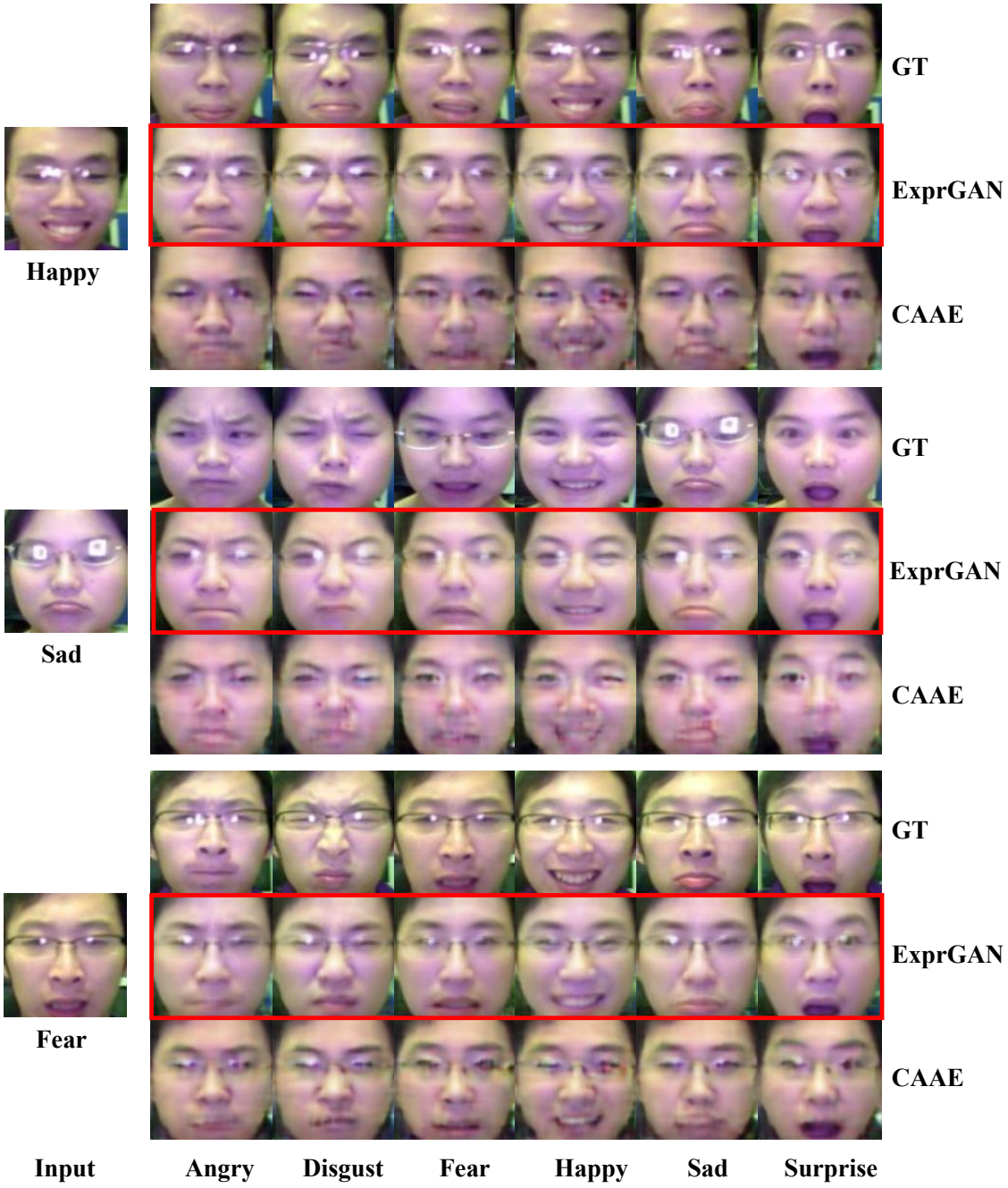


Figure 5.2: Visual comparison of facial expression editing results. For each input, we compare the ground truth images (top), the synthetic images of ExprGAN (middle) and CAAE (bottom). Zoom in for details.

5.4.3 Facial Expression Editing

In this part, we demonstrate our model's ability to edit the expression of a given face image. To do this, we first input the image to G_{enc} to obtain an identity representation $g(x)$. Then with the decoder G_{dec}, a face image of the desired expression i can be generated by setting c_i to be positive and $c_j, j \neq i$ to be negative. A positive (negative) value indicates the represented expression is present (absent). Here 1 and -1 are used. Some example results are shown in Fig. 5.2. The left column contains the original input images, while the middle row in the right column contains the synthesized faces corresponding to six different expressions. For comparison, the ground truth images and the results from the recent proposed CAAE [127] are also shown in the first and third row, respectively. We see that faces generated by ExprGAN preserve the identities well. Even some subtle details like the transparent eyeglasses are also kept. Moreover, the synthesized expressions look natural. In comparison, CAAE failed to transform the input faces to new expressions with fine details, and the generated faces are blurry.

We now demonstrate that our model can transform a face image to new types of expressions with continuous intensity. This is achieved by exploiting the fact that each dimension of the expression code captures a specific level of expression intensity. In particular, to vary the intensity of the desired class i, we set the individual element of the expression code c_i to be 1, while the other dimensions of c_i and all other $c_j, j \neq i$ to be -1. The generated results are shown in Fig. 5.3. Take the *Happy* expression in the forth column as an example. The face in the first row

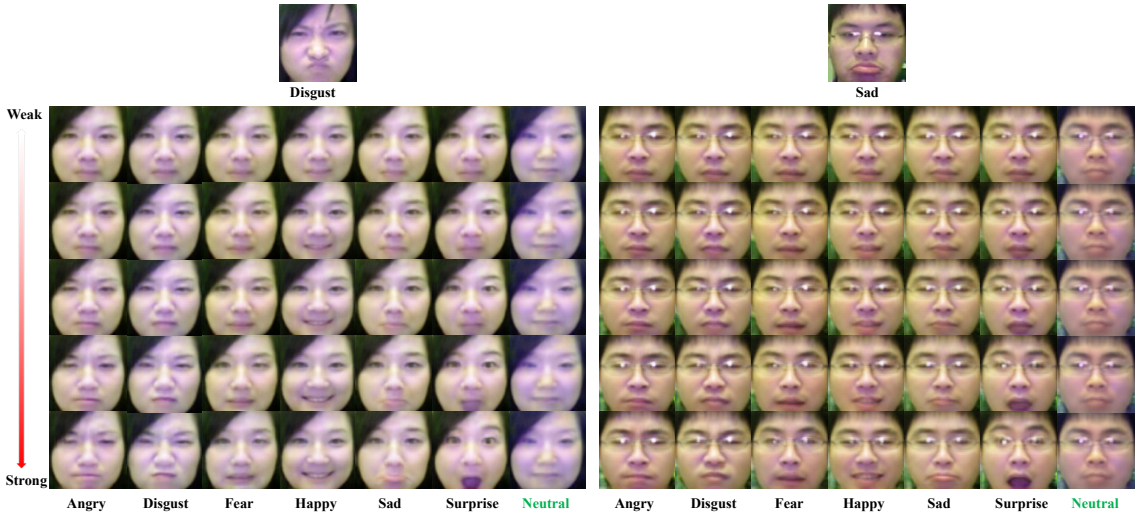


Figure 5.3: Face images are transformed to new expressions with different intensity levels. The top row contains the input faces with the original expressions, and the remaining rows show the synthesized results. Each column corresponds to a new expression with five intensity levels from weak to strong. The *Neutral* expression which is not in the training data is also generated.

which corresponds to the first element of c_i being 1 displays a gentle smile with mouth closed, while a big smile with white teeth is synthesized in the last row that corresponds to the fifth element of c_i being 1. Moreover, when we set all c_i to be -1, a *Neutral* expression is able to be generated even though this expression class is not present in the training data. This validates that the expression code discovers the diverse spectrum of expression intensity in an unsupervised way, *i.e.*, without the training data containing explicit labels for intensity levels.

5.4.4 Facial Expression Transfer

We now demonstrate our model's ability to transfer the expression of another face image x_B to a given face image x_A. To do this, we first input x_A to G_{enc} to get the identity representation $g(x_A)$. Then we train an expression classifier to

predict the expression label y_B of x_B. With y_B and x_B, the expression code c_B can be obtained from Q. Finally, we can get an image with identity A and expression B from $G_{dec}(g(x_A), c_B)$. The generated images are shown in Fig. 5.4. We observe that faces having the source identities and expressions similar to the targets can be synthesized even for some very challenging cases. For example, when the expression *Happy* is transferred to an *Angry* face, the teeth region which does not exist in the source image is also able to be generated.

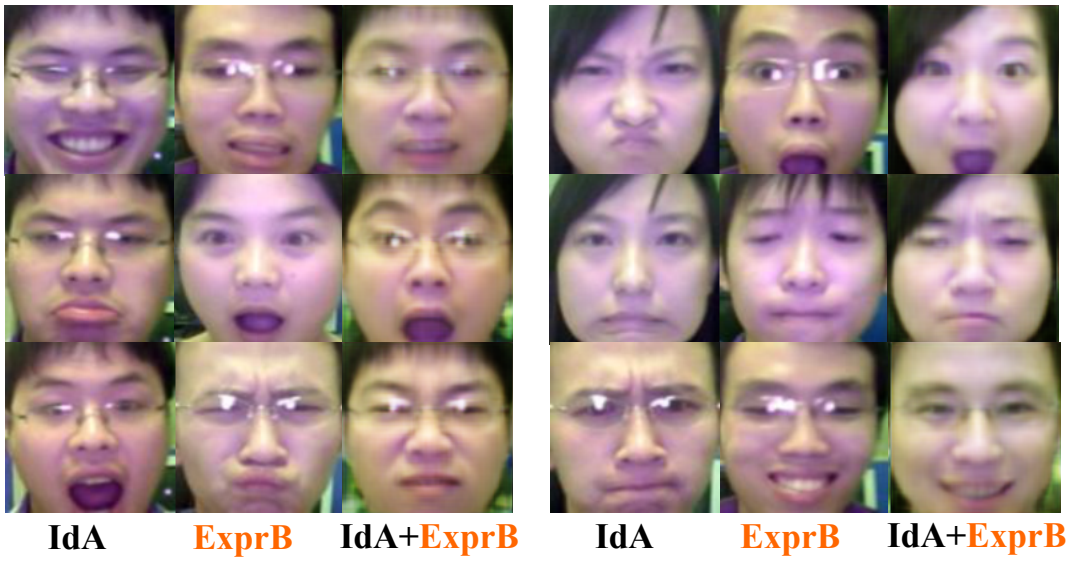


Figure 5.4: Facial expression transfer. Expressions from the middle column are transferred to faces in the left column. The results are shown in the right column.

5.4.5 Face Image Generation for Data Augmentation

In this part, we first show our model's ability to generate high-quality face images controlled by the expression label, then quantitatively demonstrate the usefulness of the synthesized images. To generate faces with new identities, we feed in random noise and expression code to G_{dec}. The results are shown in Fig. 5.5. Each

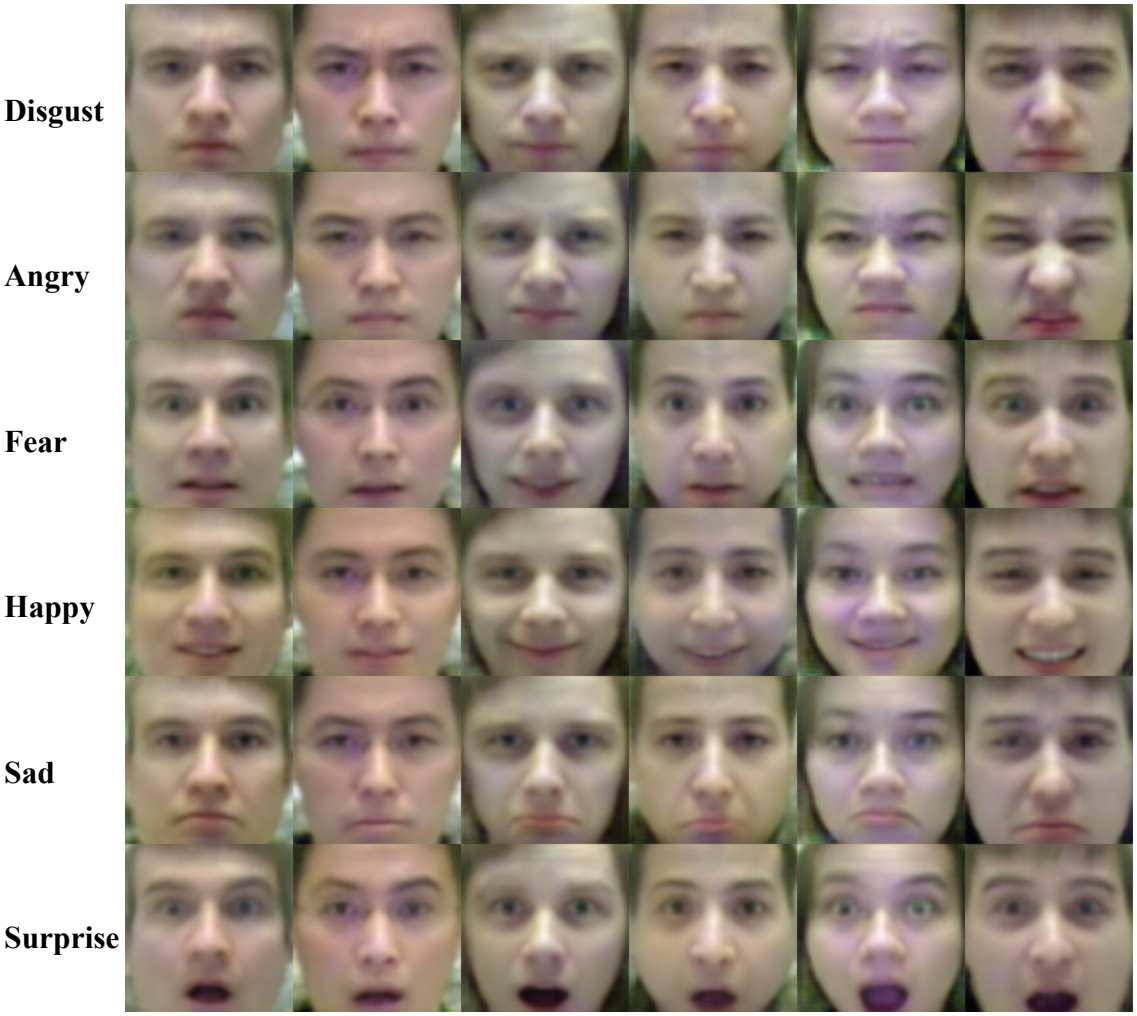


Figure 5.5: Random generated subjects displaying six categories of expressions.

column shows the same subject displaying different expressions. We can see that the synthesized face images look realistic. Moreover, because of the design of the expression controller module, the generated expressions for the same class are also diverse. For example, for the class *Happy*, there are big smile showing teeth and a gentle smile with mouth closed.

We further demonstrate that images synthesized by our model can be used for data augmentation to train a robust expression classifier. Specifically, for each expression category, we generate $0.5K$, $1K$, $5K$, and $10K$ images, respectively. The

Table 5.1: Comparison of expression recognition accuracy with different numbers of synthesized images.

# Syn. Images	0	$3K$	$6K$	$30K$	$60K$
Accuracy (%)	77.78	78.47	81.94	**84.72**	84.72

classifier has the same network architecture as G_{enc} except one additional FC layer with six neurons is added. The results are shown in Table 5.1. We can see by only adding $3K$ synthetic images, the improvement is marginal, with an accuracy of 78.47% vs. 77.78%. However, when the number is increased to $30K$, the recognition accuracy is significantly improved, reaching **84.72%** with a relative error reduction by **31.23%**. The performance starts to saturate when more images ($60K$) are utilized. This validates the high perceptual quality of the synthetic face images.

5.4.6 Feature Visualization

In this part, we demonstrate that the identity $g(x)$ and expression c representations learned by our model are disentangled. To show this, we first use t-SNE [137] to visualize the 50-dim identity feature $g(x)$ on a two dimensional space. The results are shown in Fig. 5.6. We can see that most of the subjects are well separated, which confirms that the latent identity features $g(x)$ learn to preserve the identity information.

To demonstrate that the expression code c captures the high-level expression semantics, we perform image retrieval experiment based on c in terms of Euclidean distance. For comparison, the results with expression label y and image pixel space x are also provided in Fig. 5.7. As expected, the pixel space x sometimes fails to

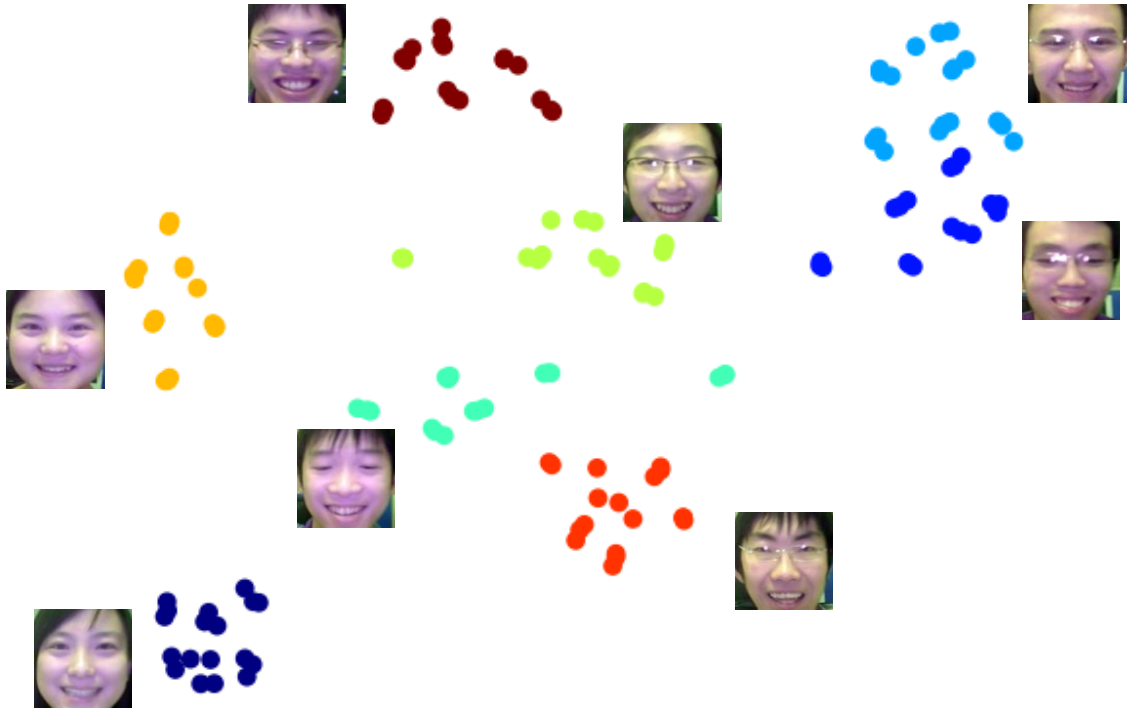


Figure 5.6: Identity feature space. Each color represents a different identity and the images for one identity are labeled.

retrieve images from the same expression. Similarly, the images retrieved by y do not always have the same *style* of expressions as the queries. For example, the query face in the second row shows a big smile with teeth, but the retrieved image by y only has a mild smile with mouth closed. However, with the expression code c, we observe that face images with similar expressions are always retrieved. This validates that the expression code learns a rich and diverse feature representation.

5.5 Conclusions

In this chapter, we present ExprGAN for facial expression editing. To the best of our knowledge, it is the first GAN-based model that can transform the face image to have a new expression where the expression intensity is allowed to

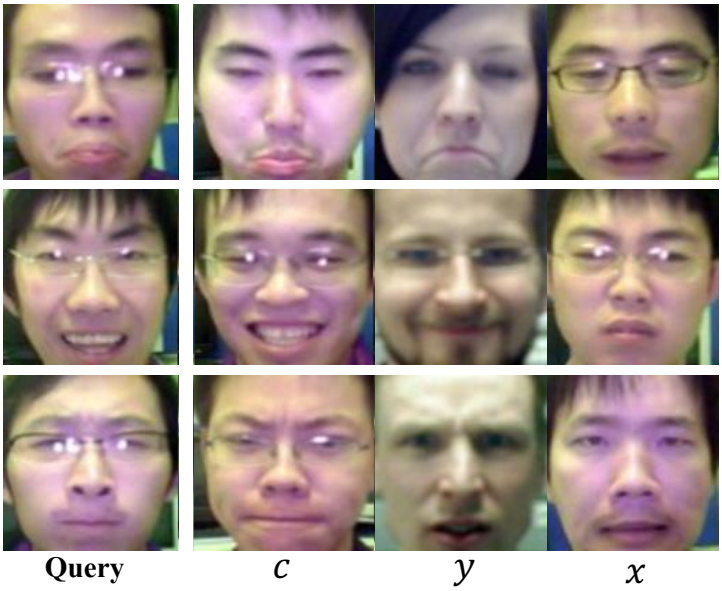


Figure 5.7: Expression-based image retrieval. First column shows query images. Other columns show top one retrieval based on c, y and x.

be controlled continuously. The proposed model learns the disentangled identity and expression representations explicitly, allowing for a wide variety of applications, including expression editing, expression transfer, and data augmentation for training improved face expression recognition models. We further develop an incremental learning scheme to train the model on small datasets. Our future work will explore how to apply ExprGAN to a larger and more unconstrained facial expression dataset.

Chapter 6: Conclusions and Directions for Future Work

6.1 Summary

In this dissertation, we focused on facial expression recognition and editing with limited data, and made novel contributions by proposing a two-stage training algorithm for expression recognition, and introducing the first GAN-based model that can transform the face image to have a new expression where the expression intensity is allowed to be controlled continuously. In addition, to tackle the challenges due to occlusion and poses, we also proposed an occlusion adaptive deep network to recognize expressions when faces are partially occluded and proposed a method that learns attentional face regions to improve attribute classification performance under unaligned condition.

In the first part of this dissertation, we presented FaceNet2ExpNet to train a light-weight and high accuracy expression classifier on small datasets. In the first stage, we proposed a probabilistic distribution function to model the high level neuron response based on already fine-tuned face net, leading to feature level regularization that exploits the rich face information in the face net. In the second stage, we performed label supervision to boost the final discriminative capability. As a result, FaceNet2ExpNet improves visual feature representation and outper-

forms various state-of-the-art methods on five public expression datasets and one medical dataset.

In the second part of this dissertation, we introduced an occlusion adaptive deep network to tackle the occluded facial expression recognition problem, which is composed of two branches. The landmark-guided attention branch guides the network to learn clean features from the non-occluded facial areas. While the facial region branch increases the robustness by dividing the last convolutional layer into several part classifiers. We conducted extensive experiments on both challenging in-the-wild expression datasets and real-world occluded expression datasets. The superior results show that our method outperforms existing methods and achieves robustness against occlusion and various poses.

In the third part of this dissertation, we proposed a parts and whole framework for unaligned facial attributes classification. A weakly-supervised face region localization network is first designed. Then the information from those detected regions are selectively combined by the hybrid classification network. Visualization shows that our method not only discovers semantically meaningful attributes regions, but also captures rich correlations among attributes. Moreover, our results outperform state-of-the-art by a significant margin on the unaligned CelebA dataset.

In the last part of this dissertation, we presented ExprGAN for facial expression editing. The proposed model learns the disentangled identity and expression representations explicitly, allowing for a wide variety of applications, including expression editing, expression transfer, and data augmentation for training improved face expression recognition models. We further developed an incremental learning

scheme to train the model on small datasets. Our future work will explore how to apply ExprGAN to a larger and more unconstrained facial expression dataset.

6.2 Directions for Future Work

In Chapter 2, we proposed a transfer learning algorithm to utilize a face recognition network for expression recognition. We only used feature maps of the last convolution layer to provide supervision in the first stage. Adding other layers may further improve the performance. Another possible direction of future work is to explore activation-based spatial attention maps [138] instead of the simple feature maps used in this dissertation.

In Chapter 3, we presented an occlusion-adaptive deep network for occluded facial expression recognition. We used the meta information of facial landmarks to guide the model to learn representations from the non-occluded facial regions. One possible direction of future work is to train the network to predict a face segmentation mask [139] directly instead of manually setting the threshold of the confidence score as done in this dissertation.

In Chapter 4, we introduced a cascade network for unaligned facial attributes classification. Though we focus on facial attributes in this dissertation, the proposed framework is very general and can be applied to other tasks like facial action unit detection. In the parts and whole classification network, we proposed an attribute relation layer to model the relationship between different attributes. Another possible direction of future work is to explore a more complex relation model like the

Graph Convolutional Neural Network [140].

In Chapter 5, we proposed a generative model for continuous expression editing. In this dissertation, we mainly focused on frontal faces. The model can be extended to generate faces with different poses by incorporating a pose variable. Moreover, another interesting future direction is to enable the model to generate faces with a relative control on the expression intensities [141].

Bibliography

[1] Deepdraw. Deepdraw on github.com/auduno/ deepdraw.

[2] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 2014.

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015.

[4] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010.

[5] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010.

[6] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9), 2011.

[7] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, 28(5), 2018.

[8] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 2020.

[9] Jeffrey Moran and Robert Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715), 1985.

[10] Michael I Posner and Steven E Petersen. The attention system of the human brain. *Annual Review of Neuroscience*, 1990.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[12] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2015.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using Places database. In *Advances in Neural Information Processing Systems*, 2017.

[17] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[18] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[20] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *ACM on International Conference on Multimodal Interaction*, 2015.

[21] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European Conference on Computer Vision*, 2016.

[22] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[23] Sirion Vittayakorn, Takayuki Umeda, Kazuhiko Murasaki, Kyoko Sudo, Takayuki Okatani, and Kota Yamaguchi. Automatic attribute discovery with neural activations. In *European Conference on Computer Vision*, 2016.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2014.

[25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[26] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[27] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. Learning active facial patches for expression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[28] Ping Liu, Joey Tianyi Zhou, Ivor Wai-Hung Tsang, Zibo Meng, Shizhong Han, and Yan Tong. Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In *European Conference on Computer Vision*, 2014.

[29] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. AU-aware deep networks for facial expression recognition. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013.

[30] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[31] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[32] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian Conference on Computer Vision*, 2014.

[33] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE International Conference on Computer Vision*, 2015.

[34] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[35] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *ACM on International Conference on Multimodal Interaction*, 2015.

[36] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *ACM on International Conference on Multimodal Interaction*, 2015.

[37] Lingxi Xie, Liang Zheng, Jingdong Wang, Alan L Yuille, and Qi Tian. Interactive: Inter-layer activeness propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[38] Josh M Susskind, Adam K Anderson, and Geoffrey E Hinton. The toronto face database. *Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep*, 3, 2010.

[39] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *ACM on International Conference on Multimodal Interaction*, 2015.

[40] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10), 2016.

[42] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Association*, 2015.

[43] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM international conference on Multimedia*, 2014.

[44] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[45] Xiaoyi Feng, Matti Pietikäinen, and Abdenour Hadid. Facial expression recognition based on local binary patterns. *Pattern Recognition and Image Analysis*, 17(4), 2007.

[46] Karan Sikka, Gaurav Sharma, and Marian Bartlett. Lomo: Latent ordinal model for facial analysis in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[47] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3D-gradients. In *British Machine Vision Conference*, 2008.

[48] Yimo Guo, Guoying Zhao, and Matti Pietikäinen. Dynamic facial expression recognition using longitudinal facial expression atlases. In *European Conference on Computer Vision*, 2012.

[49] Matthew N Dailey, Garrison W Cottrell, Curtis Padgett, and Ralph Adolphs. Empath: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, 14(8), 2002.

[50] Marc'Aurelio Ranzato, Joshua Susskind, Volodymyr Mnih, and Geoffrey Hinton. On deep generative models with applications to recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[51] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, 2012.

[52] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, 2014.

[53] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[54] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, 2013.

[55] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 2016.

[56] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Covariance pooling for facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[57] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *European Conference on Computer Vision*, 2018.

[58] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017.

[59] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Patch-gated CNN for occlusion-aware facial expression recognition. In *International Conference on Pattern Recognition*, 2018.

[60] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 2017.

[61] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. In *International Conference on Machine Learning*, 2018.

[62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[63] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[64] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[65] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[66] Behzad Bozorgtabar, Mohammad Saeed Rad, Hazım Kemal Ekenel, and Jean-Philippe Thiran. Using photorealistic face synthesis and domain adaptation to improve facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2019.

[67] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Identity-free facial expression recognition using conditional generative adversarial network. *arXiv preprint arXiv:1903.08051*, 2019.

[68] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[69] Zimeng Luo, Jiani Hu, and Weihong Deng. Local subclass constraint for facial expression recognition in the wild. In *International Conference on Pattern Recognition*, 2018.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[71] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[72] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshops*, 2013.

[73] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem?(and a dataset of 230,000 3D facial landmarks). In *IEEE International Conference on Computer Vision*, 2017.

[74] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[75] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[76] Shreyank Jyoti, Garima Sharma, and Abhinav Dhall. Expression empowered residen network for facial action unit detection. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2019.

[77] Bowen Pan, Shangfei Wang, and Bin Xia. Occluded facial expression recognition enhanced through privileged information. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[78] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 2016.

[79] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1), 2018.

[80] Corneliu Florea, Laura Florea, Mihai Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In *British Machine Vision Association*, 2019.

[81] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[82] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[83] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[84] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016.

[85] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, 2009.

[86] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10), 2011.

[87] Yan Li, Ruiping Wang, Haomiao Liu, Huajie Jiang, Shiguang Shan, and Xilin Chen. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *IEEE International Conference on Computer Vision*, 2015.

[88] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE International Conference on Computer Vision*, 2011.

[89] Aishwarya Jadhav, Vinay P Namboodiri, and KS Venkatesh. Deep attributes for one-shot face recognition. In *European Conference on Computer Vision*, 2016.

[90] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, 2016.

[91] Thomas Berg and Peter Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *IEEE International Conference on Computer Vision*, 2013.

[92] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015.

[93] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[94] Ethan M Rudd, Manuel Günther, and Terrance E Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision*, 2016.

[95] Emily M Hand and Rama Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *AAAI Conference on Artificial Intelligence*, 2017.

[96] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *International Conference on Learning Representations*, 2014.

[97] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[98] Max Ehrlich, Timothy J Shields, Timur Almaev, and Mohamed R Amer. Facial attributes classification using multi-task representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[99] Nils Murrugarra-Llerena and Adriana Kovashka. Learning attributes from human gaze. In *IEEE Winter conference on Applications of Computer Vision*, 2017.

[100] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[101] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[102] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1), 2016.

[103] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *ACM International Conference on Knowledge Discovery and Data Mining*, 2006.

[104] Yann LeCun, John S Denker, Sara A Solla, Richard E Howard, and Lawrence D Jackel. Optimal brain damage. In *Advances in Neural Information Processing Systems*, 1989.

[105] Rich Caruana. Multitask learning. *Machine learning*, 28(1), 1997.

[106] Robert Torfason, Eirikur Agustsson, Rasmus Rothe, and Radu Timofte. From face images and attributes to attributes. In *Asian Conference on Computer Vision*, 2016.

[107] Yaser S Abu-Mostafa. A method for learning from hints. In *Advances in Neural Information Processing Systems*, 1992.

[108] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017.

[109] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2014.

[110] Victor Escorcia, Juan Carlos Niebles, and Bernard Ghanem. On the relationship between visual attributes and convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[111] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

[112] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[113] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*, 2014.

[114] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. Expression flow for 3D-aware face component transfer. In *ACM Special Interest Group on GRAPHics and Interactive Techniques*. 2011.

[115] Umar Mohammed, Simon JD Prince, and Jan Kautz. Visio-lization: generating novel facial images. *ACM Transactions on Graphics*, 28(3), 2009.

[116] Raymond Yeh, Ziwei Liu, Dan B Goldman, and Aseem Agarwala. Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv:1611.09961*, 2016.

[117] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. Generating facial expressions with deep belief nets. *Affective Computing, Emotion Modelling, Synthesis and Recognition*, 2008.

[118] Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

[119] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Advances in Neural Information Processing Systems*, 2014.

[120] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[121] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[122] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088), 1986.

[123] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. In *International Conference on Learning Representations*, 2016.

[124] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Computer graphics forum*, volume 22, 2003.

[125] Zicheng Liu, Ying Shan, and Zhengyou Zhang. Expressive expression mapping with ratio images. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001.

[126] Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *IEEE conference on computer vision and pattern recognition*, 2017.

[127] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE conference on computer vision and pattern recognition*, 2017.

[128] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *IEEE conference on computer vision and pattern recognition*, 2017.

[129] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2016.

[130] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3), 2005.

[131] Kuang-Chih Lee, Jeffrey Ho, Ming-Hsuan Yang, and David Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[132] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In *IEEE International Conference on Computer Vision*, 2017.

[133] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[134] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, 2013.

[135] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014.

[136] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[137] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), 2008.

[138] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

[139] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[140] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, 2018.

[141] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. RELGAN: Multi-domain image-to-image translation via relative attributes. In *IEEE International Conference on Computer Vision*, 2019.