Face analysis and deepfake detection

Han, J.

Publication date
2021
Document Version
Final published version

Link to publication

Citation for published version (APA):
Han, J. (2021). *Face analysis and deepfake detection*.

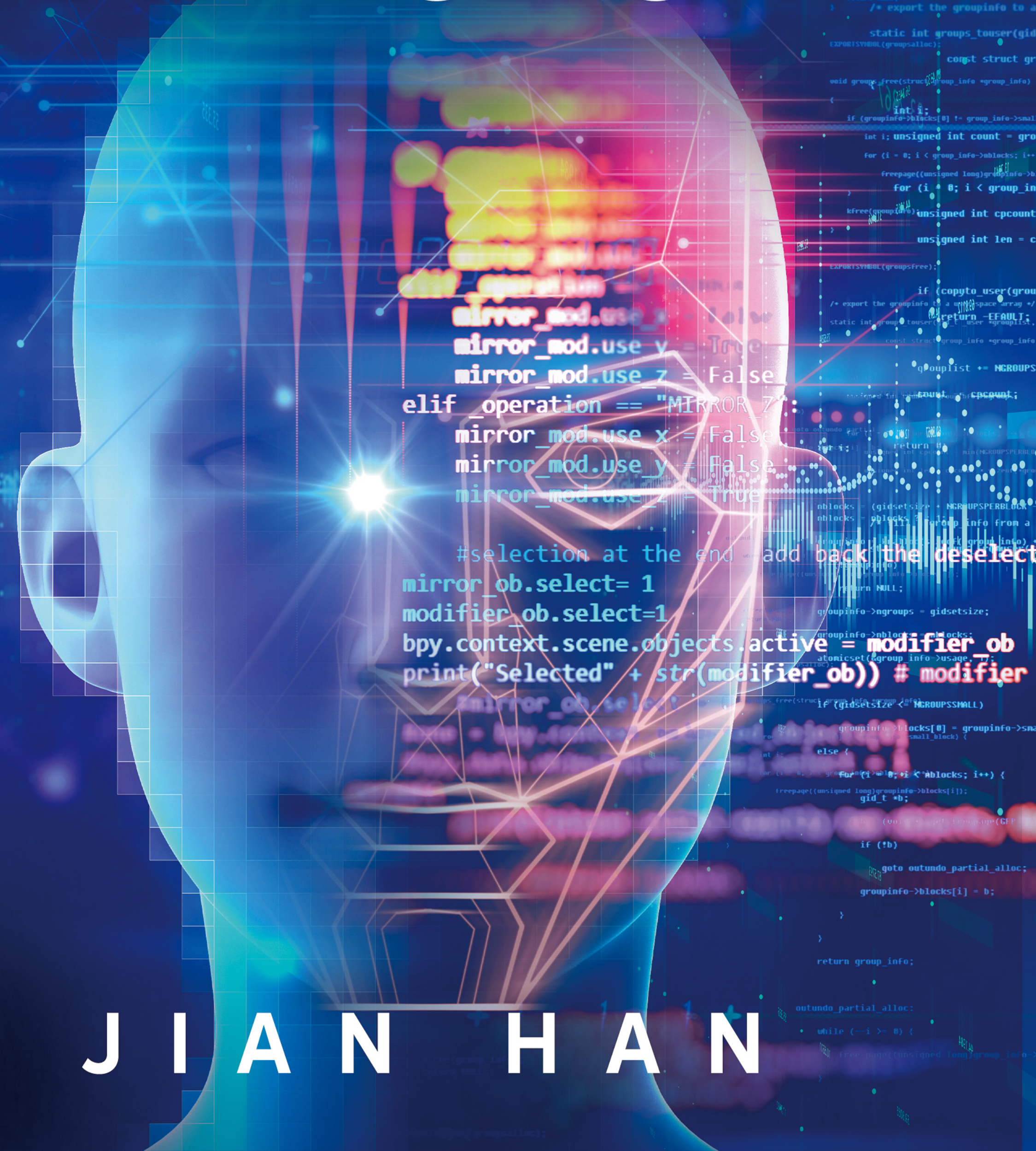
FACE ANALYSIS AND DEEPFAKE DETECTION

Face is one of the most popular biometrics used for recognition of body or behavioural characteristics. This dissertation is mainly about face detection, face forgery recognition, and age estimation.

FACE ANALYSIS AND DEEPFAKE DETECTION

JIAN HAN

FACE ANALYSIS AND DEEPFAKE DETECTION


mirror_mod.use_y = True
mirror_mod.use_z = False
elif _operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True
#selection at the end -add back the deselect
mirror_ob.select= 1
modifier_ob.select=1
bpy.context.scene.objects.active = modifier_ob
print("Selected" + str(modifier_ob)) # modifier

JIAN HAN

Face Analysis and Deepfake Detection

Jian Han

Face Analysis and Deepfake Detection

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 1 december 2021, te 14.00 uur

door

Jian Han

geboren te Shanxi

Promotiecommissie

Promotor:		
	prof. dr. T. Gevers	Universiteit van Amsterdam
Co-promotor:		
	dr. S. Karaoglu	Universiteit van Amsterdam
Overige leden:		
	prof. dr. D. Shen	Renmin University of China
	prof. dr. A. A. Salah	Universiteit Utrecht
	prof. dr. M. Worring	Universiteit van Amsterdam
	dr. S. You	Universiteit van Amsterdam
	dr. A. Visser	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Advanced School for Computing and Imaging

UNIVERSITEIT VAN AMSTERDAM

The research was supported by China Scholarship Council (CSC) under project number 201606880025.

The work described in this thesis has been carried out within the graduate school ASCI, dissertation number 430.

ISBN: 978-94-93197-94-7

Contents

1 Introduction

Face perception may be the most advanced part of our visual system. In our life, we see more faces than any other object [78]. We recognize and remember humans by their facial appearance and we sustain benign social interactions using our facial expressions. We are also able to identify in-group versus out-group members based on facial features [88]. The use and exchange of facial information is an indispensable part of our society enabling interpersonal interaction and group processes.

Information about faces may substantially influence people's social judgments and such influence may persist regardless of cognitive efforts. People can make dispositional inferences on how faces look like. These inferences can be irresistible even when people are explicitly urged against it [214]. Moreover, seemingly superficial judgments derived from faces can influence crucial decisions that are usually presumed as deliberate and calculative. Outsiders' facial perception of organizational leaders can predict the economic value and market performance of the leaders' affiliated organizations [154]. When voters have no prior knowledge about the electoral candidates, judgments merely based on faces can predict the outcomes of the United States Congress elections [183]. Such crucial decisions based on how faces look like can be made within 100 ms [198] and are the same regardless of culture [215].

In addition, the human face is one of the mostly used forms of biometrics and is applied to the recognition of body or behavioural characteristics. This is because a face, as a ubiquitous biometric, can be extracted in unconstrained (in-the-wild) environments while providing robust discriminative features [65, 124]. Automatic face analysis is of great importance because it can provide an automatic interpretation of interactions between humans and machines. With the development of computer hardware and new technology, face-based applications have been successfully applied in daily life such as access control, social networks, and video surveillance. In the last decade, major advances occurred in face related research, with many systems being capable of achieving high accuracy and speed in *constrained* environments. However, face applications in real-world scenarios are still very challenging, because the face acquisition process is dependent on a wide range of imaging conditions.

In summary, there are several key factors that may have significantly impeded performances on face analysis systems:

- Head pose variations may introduce projective deformations and self-occlusion. Extreme poses may lead to severe occlusion or skewed aspect ratios of face

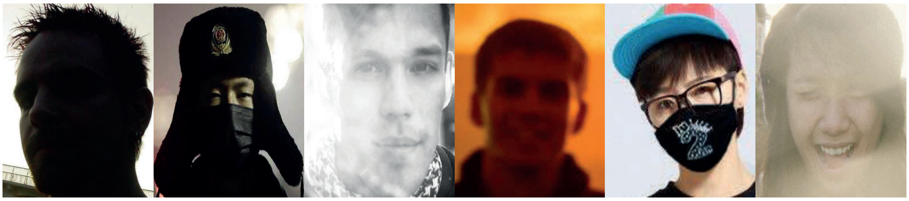


Figure 1.1: Examples of faces under extreme recording variations. Images are taken from UFDD [130] and MAFA [53].

bounding boxes [166, 209].

- Facial occlusion can negatively influence the performance of face-based applications. It may reduce the amount of information available for detection and may introduce noise [53, 190]. The challenging part is to obtain robustness for different occlusion types.

- Age variation is an important but easily ignored factor. Obviously, the appearance and geometrical variations of faces change over time. Although the aging process of human faces causes common features, like wrinkles and age spots, they may differ from person to person. This is because the facial aging process can be affected by personal (e.g., mental state, occupation) and social factors (e.g., living environment) [140].

- Imbalanced distribution of face datasets is an important factor for automatic face analysis. It originates from demographic characteristics of humans and usually manifests itself in a long-tailed distribution manner. Most samples concentrate around a small range of categories [84]. Learning systems which are trained on imbalanced datasets may be biased towards the majority part within the dataset. Hence, rare samples of face attributes may be considered as outliers or noise.

We now elaborate on the impact of each of the mentioned problems for different research tasks.

Face detection is the first step for all face applications. It it one of the first computer vision tasks with early works dating back decades ago [200]. The task can be generally defined as detecting and localizing an unknown number of faces in images. Most early studies are highly rigid because they are based on strong assumptions like no occlusion, plain backgrounds, and frontal facing. The increasing availability of data captured without constrained settings (‘in-the-wild’) has contributed to tremendous progress of face studies for varying imaging conditions (see Figure 1.1). The advancement of face detection methods also benefits from the rapid progress in deep learning [149].

In order to improve robustness against challenging variations in unconstrained datasets, data augmentation is a logical step [120, 125]. Data augmentation expands the dataset synthetically without changing the original labels [179, 189]. It can increase intra-class variations for imbalanced datasets [123]. Data augmentation for face-related datasets can be roughly classified into three categories: generic transformation, local, and global facial attribute manipulations. Local attribute manipulation mainly operates on local face areas. It includes modifications for hairstyle, makeup, and occlusion (e.g., facial hair, glasses). Global facial attributes manipulation includes pose, gender, expression, and age. Global attributes manipulations is more challenging because it affects all face regions [162].

Generic data augmentation may not always be effective. For example, a common way to generate facial occlusion is to randomly crop parts of the original face images. However, cropping can only reduces the information of faces but is unable to provide realistic occlusion samples. Simply over-sampling of original datasets may cause over-fitting in the training process.

For relatively complex data augmentation, parameterized model based methods have been proven to be very effective. Parameterized models are built based on previously collected face datasets. Early parametric 2D face models like Eigenfaces [170, 185], Fisherfaces [9], Active Shape Models [29], or Active Appearance Models [30] have facial shape and appearance which can be manipulated in linear spaces. Later, Blanz and Vetter [11] propose a 3D face morphable model (3DFMM), the first generative 3D face model that uses linear subspaces to model shape and appearance variations [12, 127]. A 3DFMM normally has a parameterized shape and appearance model. The model coefficients are moderated by probability densities [143]. A face can be defined by a set of coefficients. 3DFMM has been widely used in face reconstruction and face synthesis. From the data perspective, the creation of 3DFMM heavily relies on accurate 3D face scans which are collected under constrained settings. For example, the latest Basel Face Model [57] has only 300 individuals. Moreover, most of the subjects are young Caucasians. 3DFMM with unconstrained data is rarely considered. This leads to the main drawback of 3DFMM, that is, the lack of variations. From a methodology perspective, the generative power of 3DFMM is also limited by its own formulation. Most of 3DFMMs are still based on statistical PCA models. The facial variations are naturally nonlinear in the real world. For example, facial expressions or occlusion do not correspond to a linear assumption of PCA-based models. Thus, a PCA model does not have the capacity to simulate facial variations well [95]. The challenge of rendering photo-realistic face images lies in the difficulty of modeling hair, eyes, and the mouth cavity (e.g., teeth or tongue). Further challenges are the absence of facial details like wrinkles in the geometry. These factors degrade the quality of the final rendering results.

Another way to induce more variations is to use generative methods [45, 196, 203]. Generative adversarial network (GAN) has gained a lot of attention [61, 90]. The modification traits from GAN-based methods become more and more subtle. A number of GAN-based methods are used to manipulate face attributes [26, 39, 86]. Methods like StyleGAN [94] and its extensions [96] manage to control a number of complex facial attributes like identity, gender, and head pose during the training process. However, these facial characteristics are highly correlated (e.g. male and mustache). The manipulation of these models may not be sufficient to change attributes like facial appearance, shape (e.g., length, width, etc.) or expression (e.g., raise eyebrows, open mouth, etc.) independently. Although these methods have generated photo-realistic synthetic face images, the provided level of control of the features is unsatisfactory for real-world applications [58].

The aging effect on appearance and geometrical variations of human faces is irreversible [51]. Numerous factors have been shown to influence facial aging effects such as ethnicity, gender, dietary habits, occupation, and climate conditions [140]. Early work on automatic age estimation for still images mainly used biological-related features, while age estimation for videos often uses handcrafted features and temporal dynamics. The rapid development of deep neural networks have constantly improved the performance of age estimation. Despite the remarkable progress of deep learning based age estimation, their application in unconstrained scenarios is still imperfect. Challenging variations like extreme poses, occlusion, and blurring may limit the applicability of these age estimation methods. Moreover, most of the age estimation algorithms focus on still images. Only a limited number of methods are applied on face videos. One of the main reasons is the lack of large-scale face video datasets with age annotation. The underlying reason is that it is almost impossible to provide accurate annotation for image or videos collected from the Internet [72]. In contrast, the performance of age estimation does not benefit much from data augmentation. Compared to other relatively simple variations, generic data augmentation methods are not able to provide the necessary and complex features (like wrinkles or age spots) for age estimation systems. Recently, generative model based methods provide more realistic aging effects [45, 138, 196, 203]. However, the level of control from generative methods is not sufficient to systematically modify the aging effects in synthetic data.

With the rapid progress in 3D face reconstruction and face manipulation, it is exciting to see that more and more photo-realistic face variations can be synthesized. However, the question remains: can we distinguish which face is authentic? AI-synthesized face data is rising as a highly controversial topic [98, 102]. These AI-synthesized methods are often referred to as deep fakes. Using modest amounts of data and computing power, anyone is now able to automatically generate deep fake videos [3, 76, 146, 174]. For example, Channel 4 provided a controversial deep fake video of Queen Elizabeth giving a Christmas speech [10]. Most people in our society may not be aware of deepfakes. It can be easily used to spread disinformation [14, 35]. The misuse of deepfakes poses a substantial threat to social security and interpersonal trust [13, 93]. California state had already officially made it illegal to use deepfakes to impact political activities. It is crucial to distinguish manipulated faces from pristine face images [142].

Face manipulation methods can be generally classified into four categories:

- **Entire face synthesis**: This kind of manipulation methods focuses on generating entire non-existent face images, usually through powerful GAN based methods. Typical example of these methods is StyleGAN [95]. StyleGAN adds style vector to Progressive-GAN [94] to gain more control over the generation process. The application of StyleGAN to generate non-existent faces has drawn great public attention.

- **Facial reenactment**: This kind of methods transfers the expressions or motion from a source video to a target video while keeping the identity of the target person. One representative is Face2Face [180], which modifies the facial expressions of the target video by a source actor and generates the manipulated output video in a photo-realistic fashion in real time.

- **Attribute manipulation:** This type of methods, also known as face editing or face retouching, consists of modifying some facial attributes such as hair color [119], viewpoints [86], gender, age [6], glasses [163], etc. This manipulation process is usually carried out through generative model such as the StarGAN approach proposed by [26]. One typical example is the popular FaceApp application. Users could modify various types of makeup, glasses, or hairstyles in a virtual environment.

- **Identity manipulation**: Instead of changing facial local attributes or expressions, identity manipulation aims at replacing the facial identity information of a target subject with the face of source subject. This category is known as face swapping [122]. It became popular on social media with wide-spread consumer-level applications like Snapchat filter. There are also deep learning based methods which are known as DeepFakes [38], e.g., the recent viral mobile application ZAO.

Most of face manipulation methods normally need a pair of faces from source and target individuals. The final results are contingent on the source and target images. Even using the same pair of subjects, different modification methods may generate different outputs. This is because each method has different processing regions or architectures. On the other side, when the same modification method is applied on different pairs of data, the results can have various types of artifacts like face attribute mismatches. The reason is that each face has different variations in pose, lighting, or ethnicity. This process resembles the neural style transfer operation between content and reference images [52].

Recent research has shown that supervised deep learning approaches can achieve impressive face forgery detection performance. To detect manipulated faces, the common approach is data-driven. The generalization is normally restricted to a small range of manipulation artifacts. Another category of deep network based methods is to capture features from the generation process. The features include artifacts or cues introduced by the backbone network architecture [102]. These methods normally rely on a large amount of training data, and the performance decreases dramatically when new types of manipulations are presented, even though they are semantically close. The underlying neural networks quickly overfit to manipulation-specific artifacts. Thus, extracted features are highly discriminatory for a given task but lack generalization capabilities for unseen examples. This weakness can be alleviated by fine-tuning a pre-trained network with new task-specific data, but this also means that large amounts of new data are required [31]. Also, a significant performance gap can be observed when tested on compressed, low-resolution, or blurry data [151], which is crucial to detect fake media content on social media. All proposed learning-based methods need some form of fine-tuning on a dataset with manipulations aligning the samples from the training and test set. But the underlying datasets are limited in variations like face attribute, pose, or occlusion.

Based on existing research, it is feasible to accurately identify certain kinds of artifacts. However, artifacts such as imaging variations or face attributes do not persist across all generated results for a single generation method. Methods that simply try to detect certain artifacts are not able to handle unseen ones. The generalization of face forgery detection is the key to overcome the challenges posed by deep fakes.

Normally, face datasets have a relatively large volume. Just like the demographic information in the real world, the distribution of these datasets is always long-tailed and imbalanced. Therefore, there is always a large portion of subjects or facial attributes, of which the samples are insufficient and under-represented [210]. In conventionally trained deep networks, training with highly imbalanced data leads to biased classifiers. Although the minority categories constitute a small portion of the whole dataset, they can play a significant role in the prediction error of trained models. Such biased predictions can have severe consequences in applied settings. For instance, Gender Shades [16] found that gender classification systems from IBM, Microsoft, and Face++ had worse performance for darker-skinned females than lighter-skinned males. A similar study [171] also found that Amazon's face recognition system was more likely to misclassify colored than white Congress members. A government study further [63] demonstrates that most one-on-one matching systems had higher false positive rates for Asian and African faces than Caucasian faces. Concerning these consequential effects of face studies, the imbalanced issue of datasets should be better scrutinized.

Current methods for handling the imbalanced issue typically adopt class re-balancing strategies such as re-sampling and re-weighting the training scheme based on the information in the dataset [83, 97, 182, 224]. For re-sampling based methods, they can be classified into oversampling and down-sampling. Oversampling strategy focuses on the effect of minority samples. Intuitively, more data would lead to better performance of the model. However, it might cause over-fitting. Down-sampling normally reduces samples from majority categories, but this strategy has the risk to exclude useful feature variations in the sampling process. As for re-weighting based methods, they focus on the training process such as the design of loss functions. These methods normally rely on the frequency of each class from the dataset. The goal is to provide a smoother version of the training scheme towards the imbalanced dataset. However, the challenge is to determine the actual weights for different samples in various distributions. It is difficult to estimate the effective number of samples for each category [33].

1.1 Research Outline and Questions

In this thesis, we aim to address the following research questions:

Although face detection has already achieved impressive results for constrained datasets, face detection of images taken from faces with varying imaging conditions is still challenging. In existing datasets for face detection, the majority of data vary to a limited extent. Collecting and annotating real-world face datasets with various attributes is expensive and impractical. It is also difficult to fully control the imaging variations or to avoid errors during the annotation process. Current generative methods cannot provide the level of control to systematically manipulate different variations. We pose our first research question:

Question 1: Can we systematically manipulate variations in synthetic data to complement the real dataset and achieve better performance for the face detection task?

In Chapter 2, we provide an overview of how object features from images influence face detection performance, and how to choose synthetic data to address specific features based on 3D face models. First, we provide a 2D synthetic face data generator with fully controlled features. We systematically evaluate the influence of occlusion, scale, viewpoint, background, and noise by using this synthetic image generator. We consider three representative deep network face detectors for our analysis. Comparing different configurations of synthetic data on face detection systems, it shows that our synthetic dataset could complement face detectors to become more robust against specific features in the real world. Our analysis also reveals that a variety of data augmentation is necessary to address differences in performance.

Pose variation could considerably change the appearance of faces and may cause (self) occlusion. The lack of unconstrained face datasets with age labels is also a bottleneck for age estimation to improve robustness against challenging imaging conditions. One of the main difficulties is to accurately annotate a face image or video with age labels from the Internet. Therefore, we focus on our second question:

Question 2: How can we alleviate the negative influence of pose variations when predicting age?

In Chapter 3, we propose an age estimation method for handling large pose variations for unconstrained face images. To attenuate the effect of pose, our method is based on facial uv texture maps reconstructed from original video frames. A Wasserstein-based GAN model is used to complete the full uv texture presentation. Age is further predicted from the completed uv mappings such that the proposed AgeGAN method simultaneously learns to capture the facial uv texture map and age characteristics. In order to train our method, we created the UvAge dataset: the largest video dataset of face with age annotations (together with identity, gender, and ethnicity labels). The dataset contains in-the-wild videos from celebrities recorded in a variety of imaging settings. In total, we collected 6898 video segments from 516 celebrities in 57 events. This in-the-wild dataset contributes to future research in age estimation. Extensive experiments demonstrate that our proposed approach outperforms other advanced age estimation methods.

With the rapid development of generative methods, it becomes very challenging to distinguish real face images from modified images. Recently, supervised deep learning approaches show good performance. However, these methods normally train on large scale datasets. When facing new types of manipulations, the performance degrades dramatically. The backbone neural network architecture quickly overfits to modification-specific artifacts. The extracted features are highly discriminatory for the given task but lack transfer ability for unseen modification examples. In most cases, existing methods can handle certain types of methods or a limited range of artifacts. However, artifacts such as imaging variations or face attributes do not persist among all generated results for the same modification method. This problem leads to our third question:

Question 3: Can we find a robust method to distinguish deep fake data from multiple domains?

Our task is to distinguish manipulated from real face images from multiple domains. The main drawback of existing face forgery detection methods is their limited generalization ability due to differences in domains. Therefore, in Chapter 4, we propose a novel framework to address the domain gap induced by multiple deep fake datasets. Both neural style transfer and face manipulation need a pair of source and target for processing. Inspired by the application from style transfer task, we use maximum mean discrepancy (MMD) loss to align the different feature distributions. MMD loss is able to reduce the influence of manipulation specific artifacts. The center and triplet losses are also incorporated to enhance generalization of the network. This addition ensures that the learned features are shared by multiple domains and provides better generalization abilities to unseen deep fake samples. Evaluations on various deep fake benchmarks show that our method achieves the best overall performance.

Dataests for face related research often exhibit highly-skewed label distributions of face attributes. As for age estimation, the ages of the recorded persons in real-world videos usually have a long-tailed distribution. The majority of age groups are normally young people. The amount between the majority and minority normally have an imbalanced ratio. Therefore, the training process of age estimation systems becomes biased towards the majority age group. And the performance on the minority labels are largely ignored. Moreover, only a few methods consider the performance of age estimation on videos. Our fourth question is then postulated as follow:

Question 4: How can we mitigate the influence of imbalanced distribution and improve the performance of video based age predictions?

Most of the existing methods for age estimation largely ignore the negative influence of age imbalance. In Chapter 5, we address the problem of age imbalance in videos from a transfer learning perspective. We use a deep clustering module to both learn a proper data representation and transfer information from the majority groups. To provide a more balanced prediction, we use soft label assignment to represent the target age distribution. Each age is assigned with a specific degree of contribution. This avoids hard constraints on target age labels, leading to better solutions of the annealing clustering process. We also consider the influence of different variations (pose, expression etc.) for age estimation. Evaluations of our method on both constrained and unconstrained video datasets establish its effectiveness.

1.2 Origins

In this section, we list the publications each chapter is based on

Chapter 2 is based on the following paper:

- J. Han, S. Karaoglu, H.-A. Le, and T. Gevers. Object features and face detection performance: Analyses with 3d-rendered synthetic data. In *2020*

25th International Conference on Pattern Recognition (ICPR), pages 9959–9966. IEEE, 2021

Chapter 3 is based on the following paper:

- J. Han, W. Wang, S. Karaoglu, W. Zeng, and T. Gevers. Pose invariant age estimation of face images in the wild. *Computer Vision and Image Understanding*, 202:103123, .

Chapter 4 is based on the following paper:

- J. Han and T. Gevers. Mmd based discriminative learning for face forgery detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Chapter 5 is based on the following paper:

- J. Han, W. Wang, and T. Gevers. Deep imbalanced learning for age estimation from videos. *Under review at Computer Vision and Image Understanding*, .

1.3 Thesis Overview

In this thesis, we explore how to improve the performance of several face systems when confronting challenging variations. In Chapter 1, we provide an introduction and background information on the theme, and we list the main research questions of this dissertation.

In Chapter 2, we provide a synthetic face data generator with fully controlled variations and proposed an experimental comparison of main characteristics that influence face detection performance.

In Chapter 3, we introduce a new dataset UvAge for age estimation and provide a W-GAN based approach (AgeGAN) to simultaneously estimate the real age and complete the partial uv textures from original frames.

In Chapter 4, we propose a maximum mean discrepancy (MMD) based architecture for cross-domain face forgery detection.

In Chapter 5, we introduce an end-to-end framework to predict ages from face videos. Clustering based transfer learning is used to provide proper prediction for imbalanced datasets.

Chapter 6 concludes this thesis and proposes directions for future work.

2 Analysis for Object Features and Face Detection Performance

2.1 Introduction

Face detection is one of the most studied topics in the field of computer vision. It plays a fundamental role in basically all face related applications. Face detection requires first to determine whether there is a face in an image or a video and then to return the precise location of the face. A number of effective face detection systems have been rapidly emerged in recent years. It is impractical to evaluate on every recently proposed detection system, we are fortunate that most of the leading approaches shared a common methodology [85]. Deep network based object detection system can be roughly classified as two categories: 1) Two step face detector is the most representative detector based on deep network. Region proposal is the fundamental step for all these kind of methods. The first stage normally proposes candidate bounding boxes. In the second stage, features are extracted from each candidate box for the following classification and bounding-box regression tasks. 2) One step face detector doesn't need a region proposal unit before classification. Its light-weighted structure is more time efficient but less accurate.

Face detection still confronts challenges from features in scale, head pose, expression, facial occlusion and illumination. In existing datasets for face detection like FDDB [91], MAFA [53] and Wider Face [204], the majority of data normally belong to a limited range of variations. The faces did not sufficiently represent extreme poses, scale or heavy occlusion, to train a robust detector against all potential variations. Previous researchers designed different face detectors to address specific types of features in real-world situations. The rapid development of deep learning essentially relies on the availability of large-scale annotated datasets. Collecting and annotating real-world datasets with different attributes is unpractical. It is also difficult to fully control the imaging variations in such datasets, or to avoid errors during the annotation process. A bias from ground truth may lead to far-reaching impact in deep networks.

Data augmentation deals with aforementioned issue by artificially inflating the training set with label preserving transformations [179, 189]. A variety of data augmentation methods have shown effectiveness in face related tasks [120, 125]. In this paper, we aim to address the issue by using synthetic data, as complementary to real data, to

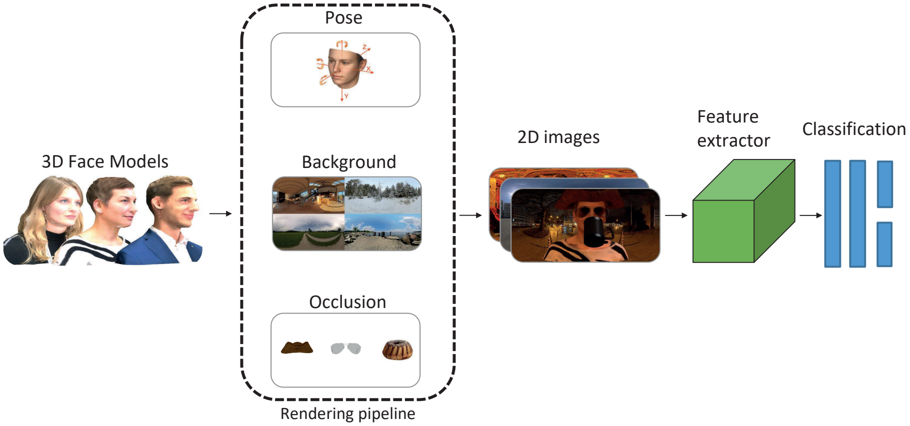


Figure 2.1: An overview of render pipeline. We manipulate pose, background, occlusion, and illumination on original 3D models, and then render 3D models into 2D images.

create fully controlled conditions with automatic and error-less annotation. We develop a synthetic data generator based on 3D face models. The 2D face synthesis process contains varied viewpoint, scale, illumination, occlusion and background. We manipulated all these features or attributes in 3D scenes, to make the rendered images more realistic than direct manipulation on 2D. With the help of synthetic data, we are able to systematically investigate the effects of different features. Then the face detectors are trained on the combination of real data and synthetic data to address the features. Based on our experiments, we also identified some potential deficiencies of the current face detection systems. Our work can also be an example to analyze other detectors.

Our contributions are:

- We provide a 2D face synthetic data generator with manipulated features (on pose, scale, background, illumination, and occlusion), which enables specified examinations of face detector performances.
- We conduct detailed analyses between feature and performance, which can be a guide to compare performances of other face detectors.
- Our analyses also reveal some weaknesses of the current face detectors and suggest using synthetic data for future improvement on robustness.

2.2 Related Work

2.2.1 Face detection

Face detection can be considered as a special case of object detection. Two thorough surveys related to object detection can be found in [80] and [85]. Most face detectors are designed to address specific characteristics in real-world scenarios, in terms of, for example, scale [75, 177, 208, 217, 218], occlusion [53, 190], pose [166] or lighting condition [221]. A certain face detector may be only suitable for datasets with corresponding characteristics and is almost impossible to be robust against features of all datasets.

We will briefly discuss several typical detectors and the features they mainly deal with. To handle large variations of scale, Hybrid Resolutions (HR) face detector is designed to detect faces with extreme scale by using contextual information and an image pyramid [81]; Single Stage Headless (SSH) detector [131] is a fast one-step detector based on scale-invariant design. To detect occluded faces, a local linear embedding method is used to reduce noise and recover the lost cues from occlusion in [53]; Face Attention Network [190] uses a special attention network with reduced background information and data augmentation to address face occlusion.

2.2.2 Face specific data augmentation

Basic geometric and photometric data augmentation methods, like flipping, rotation, resizing, cropping, color jittering, have been widely used in deep learning based face applications. Detailed surveys about face specific data augmentation can be found in [108, 193]. Previous research converges to support the effectiveness of synthetic data in improving the performance in face related applications [1, 100, 136]. Masi et al. [123] introduce face appearance variations with pose, shape and expression for effective face recognition. Lv et al. [120] propose multiple data augmentations for face recognition, including synthetic variations for hairstyle, glasses, poses and illumination.

Then, the problem shifts to the generation of synthetic face images. Face editing includes shape morphing [11, 158], relighting [167, 194], pose normalization [77, 209, 225], and expression modification [146, 180]. GAN-based methods can provide realistic results of facial attribute manipulation [26, 163] but is yet bounded with the limitation of its training images. The training images merely cover a narrow range of variations, and cause some artifacts in generation.

2.3 Problem Formalization

In order to investigate the influences of object features systematically, we generate synthetic face data targeting a specific feature for face detectors. In section 2.3.1, we introduce the influence of several major object features on face detectors. Then we provide basic information about face detectors in our experiments in section 2.3.2. In section 2.3.3, we explain how we synthesize face images based on 3D face models.

2.3.1 Challenging object features

We will briefly discuss several features which have major effect on face detection performance.

Pose could significantly change the appearance of faces. Extreme pose can lead to heavy occlusion or skewed aspect ratio of face bounding boxes [209].

Scale is very challenging to deep network based modern object detectors. For example, the features between a 10px tall face and a 1000px tall face are essentially different [81]. Pyramid architecture and multi-scale inference are currently the common approaches to detect faces of extreme scales [169].

Context information plays a fundamental role in providing the precise location of faces. Normally, surrounding regions of faces provide complementary information on object appearance and high-level features [25]. However, faces in unconstrained settings may be surrounded or occluded by different distracting objects.

Facial occlusion decreases information available for detection and introduces additional noise. Facial occlusion can be divided into two different categories: landmark occlusion and heavy occlusion. Landmark occlusion means that only a few landmarks like eyes or mouths are occluded, while most parts of the faces are still visible. In contrast, heavy occlusion represents situations where more than half of the face is missing due to occlusion, image border or extreme pose. It is most challenging when the occlusion comes from other faces. A detector may identify several partially occluded faces as one face [53].

Blur and low resolution usually impede face detectors from retrieving available information. In some practical applications, images may be distorted in collection, storage, or transmission, leading to degraded quality of images [223]. In some extreme cases, mere outline of faces can be identified.

2.3.2 Face detectors

We provide an overview about the face detectors in our experiment.

Faster RCNN is the most representative object detector based on a deep network. However, its initial design does not have additional settings targeted at challenging features [149] in the face detection task.

Single Stage Headless (SSH) [131] detector is an extremely fast one-step face detector, designed to be scale invariant. To accelerate the inference process, it has a light-weighted structure. This strategy jeopardizes the detector's performance when confronted with other potential variations.

Hybrid Resolutions (HR) face detector [81] has good performance on tiny face detection by using wide-range contextual information and testing on multiple resolutions. Its architecture resembles RPN [149] and uses both feature pyramid and image pyramid. However, HR face detector is extremely sensitive to tiny distracting objects from the background. HR also heavily relies on contextual information to locate faces. For faces with limited information (e.g., heavily occluded, extremely small or blurry), complex background could hinder precise detection. Even though HR can sometimes perform well when dealing with blurry or extreme pose, it is insufficiently robust in the detection of occluded faces, especially when occlusion stems from other faces.

2.3.3 Face synthesis

Here we give a brief introduction on how we rendered images from 3D face models. Our synthetic data generation is based on a new 3D face dataset called 3DU Face Dataset. It has 700 3D face mesh models with high-resolution texture of 435 different individuals.

Feature	Faster RCNN	SSH	HR
landmark occlusion	✓	✓	✓
complex background	✓	✓	✓
extreme pose		✓	✓
extreme scale		✓	✓
heavy occlusion			✓
blur			✓
extreme illumination			✓
misleading objects			

Table 2.1: Three advanced face detectors and challenging characteristics they can handle.

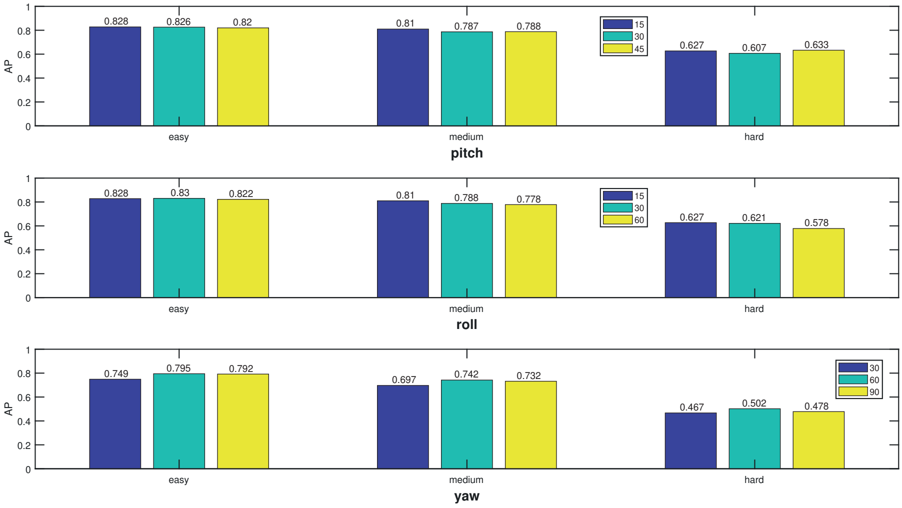


Figure 2.2: Performance comparison on pose variation. Across pitch, yaw, and roll, different colors represent the rotated degree, as labeled at the top right (e.g., “15” means “-15” to “15”).

Some people have multiple records taken at different times. Most models of this dataset are taken in varying conditions. For future research and applications, each model is annotated by humans with 50 landmarks.

An overview of our render pipeline can be found in Figure 2.1. The rendering pipeline is built on Blender. To change the viewpoints, we rotated the model with different Euler angles with the camera staying in the same position. The parameters of pitch, yaw and roll are selected randomly within different ranges. For face scale variation, the distance between camera and face models is selected randomly from a uniform distribution. The ground truth for face detection is generated from 3D landmarks. The annotation policy is the same as in Wider Face. Our rendering pipeline can also be applied on other 3D face models.

In the current research, we consider face occlusion as a crucial factor in face detection. The common way to add occlusion is to crop face images. However, cropping reduces the information of faces and cannot provide reasonable noise as in real occlusion samples. We randomly add different 3D objects like sunglasses, hats, and helmets in the 3D scenes before we start rendering. With the anchoring of landmarks, all the objects can be placed in a reasonable location to simulate the landmark occlusion. Face region has been divided into three different parts of head, eye and mouth, to simulate occlusion. We can generate more than 1000 different combinations of occlusion for each model.

2.4 Experiments

2.4.1 Experimental setup

We conduct experiments to systematically investigate how configuration of data augmentation influences performance of face detection. We test on advanced face detection benchmarks MAFA, UFDD and Wider Face. The face detection methods include Faster RCNN, SSH and HR. Despite the common practices of training and testing on the same datasets, here we first train the face detectors on synthetic data and then test on real data. We validate on a subset of the real data training part. After comparing the performance on real data with different rendering parameters, we attain a suitable configuration for one specific dataset. We use these augmented synthetic data to improve performance on real data. The metric for all the experiments is AP (average precision). We keep the original parameters and settings including data augmentation for each detector.

2.4.2 Datasets

We briefly introduce three face detection datasets used in our experiments and their features. In Table 2.2, we listed and compared features of these face detection benchmarks, which are derived from their official introductions. For all three datasets, we follow official settings for splitting train and test set.

MAFA is a representative dataset of facial occlusion, which is mainly composed of various level of occlusions [53]. Most informative features from facial attributes are missing in faces with heavy occlusion. The highly diversified masks which generate the occlusion can bring in diversified types of noises. To exclude the interference of pose,

Feature	MAFA	UFDD	Wider
landmark occlusion	✓	✓	✓
complex background			✓
extreme pose			✓
extreme scale		✓	✓
heavy occlusion	✓	✓	✓
blur	✓	✓	✓
extreme illumination		✓	✓
misleading objects		✓	✓

Table 2.2: Three face detection benchmarks and their challenging characteristics.

MAFA only includes a narrow range of head poses. Faces are labeled as “Ignore” if they are very difficult to be detected. In total, MAFA has 30811 images which include at least one masked face. In order to evaluate the occlusion degree, four major regions (eyes, nose, mouth and chin) are considered. Depends on the number of facial regions have been occluded, three occlusion levels are defined, including weak occlusion (1 or 2 regions), medium occlusion (3 regions), and heavy occlusion (4 regions).
Unconstrained Face Detection Dataset (UFDD) contains faces in different weather conditions (Rain, Snow, and Haze) and other challenging features concerning lens impediments, motion blur and defocus blur [130]. Additionally, it has a collection of distracting images to enhance difficulty. In total, it has 6425 images with 10897 face bounding boxes annotated. For most of previous datasets, each image normally has at least one annotation for face. These confounding images from UFDD might contain objects which look similar to human faces such as animal faces or include no faces. Therefore, UFDD is very good target to fully analyze the robustness of face detectors. In UFDD, the most challenging part is extreme lighting condition and blur.
Wider Face has been the most demanding benchmark for face detection till now [204]. It includes diverse events with a variety of backgrounds. The massive number of faces included has extreme poses, exaggerated expressions, heavy occlusion and extreme lighting conditions. All these features, especially scale, are difficult to handle for most face detectors. Table 2.3 shows the basic characteristics of faces, irrespective of invalid faces, in the Wider Face validation partition. Successively, the easy partition only has large faces; the medium partition additionally contains medium faces; and the hard partition includes the whole dataset.

Partition	Large	Medium	Tiny
Height	50-400 (96.6%)	30-50 (99%)	10-30 (99%)
Width	20-300 (96.3%)	10-70 (99.7%)	8-20 (95%)
Number	7211	6108	18636

Table 2.3: Face scale information of the validation set in Wider Face. We distinguish three face categories based on height and width. Proportion information represents the percentage of faces that fits within the scale interval.

Feature	Faster RCNN	SSH	HR
landmark occlusion	✓	✓	✓
complex background	✓	✓	✓
extreme pose		✓	✓
extreme scale		✓	✓
heavy occlusion			✓
blur			✓
extreme illumination			✓
misleading objects			

Table 2.4: Three advanced face detectors and challenging characteristics they can handle.

2.5 Experiment Setup

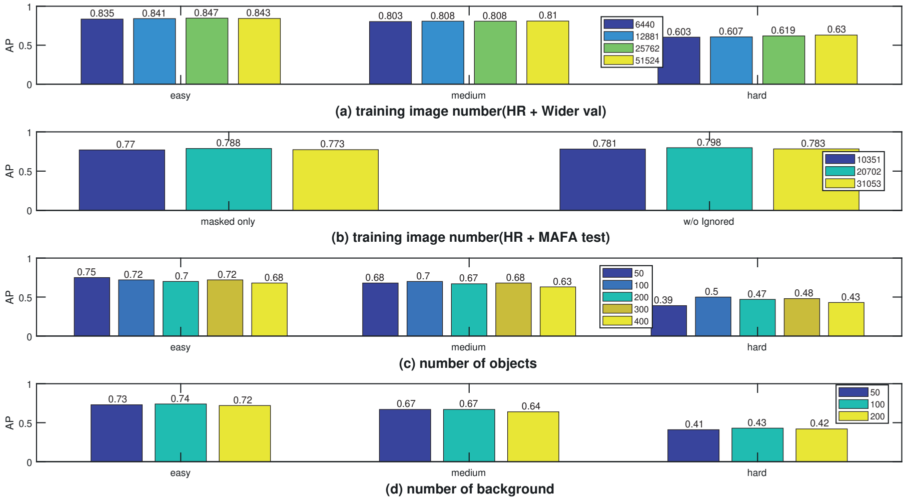
2.5.1 Settings of rendering process

This section is to demonstrate the basic settings of our rendering process. The 3D models are not changed in terms of shape or texture in experiments. We choose 100 fixed 3D models as experimental subjects and set the default for pitch, roll and yaw randomly in ranges of (-15, 15), (-15, 15), (-60, 60), respectively. For each face model, the rotation origin is the center of its landmarks, irrespective of invalid landmarks. All the face models are aligned by using landmarks with an anchor model. The anchor model is aligned to the global axis in Blender. Within a fixed range (from 1 to 48 faces), we randomize the number of faces in each image. The distance between each model and camera is randomly selected within the range of 0 to 20 meters. The size of the faces is larger than 10×8 pixels. 50 HDR images (no humans) are taken from Shape Net [19] and used as backgrounds for rendering. These images provide environment lighting and background variations for the synthetic images. The background dataset includes both indoor and outdoor scenes. Back face culling is applied to avoid artifacts in the rendered results.

2.5.2 Settings of face detectors

We provide a summary how face detectors differ in their performances in tackling different variations in Table 2.4. This information is based on experiments of their methods. We presume a good detection performance if we manage to fill the gap between real and synthetic data. There are massive parameters in our rendering pipeline, it is therefore highly unlikely to find the optimal setting to simulate one specific real dataset. However, after comparing the performance on real data with different rendering parameters, we attain a suitable and effective configuration for testing on one specific dataset. At last, we use these augmented synthetic data to improve the performance on real data.

We do not employ Faster RCNN or SSH. Faster RCNN is a generic detector without multi-scale testing, of which the performance can not reflect all the changes in variations.



(a) training image number(HR + Wider val)

(b) training image number(HR + MAFA test)

(c) number of objects

(d) number of background

Figure 2.3: Performance comparison on the basic settings of our render pipeline. It includes the effect of training data size on both (a) Wider Face validation set and (b) MAFA test set. (c) and (d) respectively show the effects of objects and background on Wider Face validation with HR.

SSH is designed to be a scale-invariant one-step detector, which can handle scale well but not other variations. For all the following comparison experiments, when we focus on one variation in the dataset, the rest variations are kept the same.

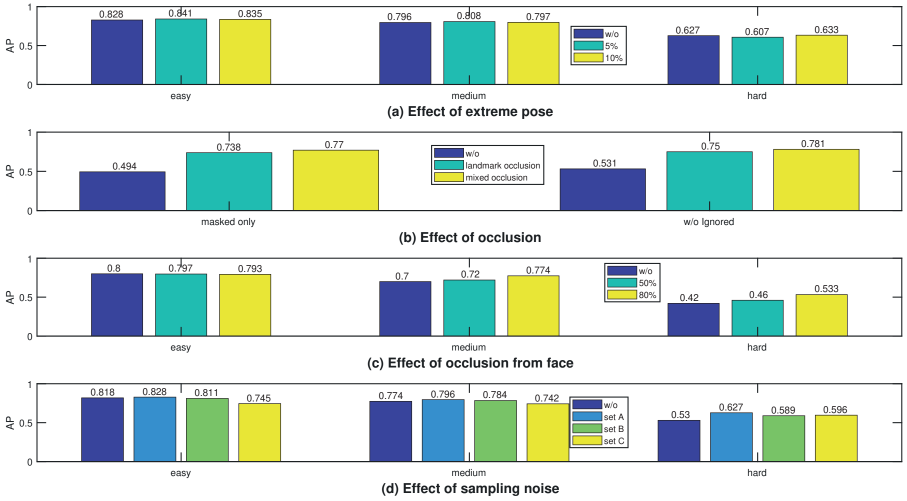
Different face detectors have their own approaches to augment data, like flipping, cropping, or transforming. We keep their original operations or hyper-parameters as much as possible. For all the following experiments with SSH and HR, we deploy on one single GPU [7]. For Faster RCNN, we use the implementation from [34].

2.6 Evaluation on Object Features

Based on our own rendering pipeline, we are able to generate all kinds of synthetic datasets with fully controlled configuration. We can investigate in-depth the effect of various data augmentations. Compared with other detectors, the architecture of HR is specially designed for face detection. HR would be a suitable option to reveal all the effects of different data augmentations. The following experiments in this section are all based on HR. For all experiments, we study one feature at the time. The other features are kept the same.

2.6.1 Effect of render settings

First, we test some basic settings of our rendering pipeline. We diversify image numbers to test the influence of training dataset size on performance. As shown in Figure 2.3.(a), the quantitative increase of synthetic images cannot continuously improve performance.



(a) Effect of extreme pose

(b) Effect of occlusion

(c) Effect of occlusion from face

(d) Effect of sampling noise

Figure 2.4: Performance comparison on other features. Only (b) tests on MAFA test set, while the remaining are on Wider Face validation set. (a) shows the results after adding small-portion extreme pose into training dataset; (b) shows the results of adding different types of occlusion. “w/o Ignored” means face with label “Ignored” are not included; (c) shows the results of adding occlusion from other faces. (d) shows the results of adding different noise from down-sampling or up-sampling;

Since the variation in synthetic data are not as much as in real data, simply increasing synthetic images may lead to over-fit. In the training process, large faces generate more positive samples than tiny faces. A larger training sample is more helpful for tiny faces. Due to the fact that most of faces from Wider Face have extreme scale, we also test on MAFA with different training dataset sizes in Figure 2.3.(b). It shows that the performances on both occluded faces and entire dataset (including occluded and unoccluded faces) saturate after adding more training images. Moreover, background is crucial for object detection tasks. HDR images provide higher resolution and less sharp results than real images. HDR images also have their own bias in respect to other images; the quantitative increase of HDR images does not improve performance constantly (see Figure 2.3.(d)). In Figure 2.3.(c), we further investigate the influence of the number of 3D objects. It shows that the number of 3D models does not strongly influence the performance, probably because of potential bias of our 3D models.

2.6.2 Effect of pose

We investigate the effect of head pose on Wider Face because it has a wide range of pose. To this end, we set different ranges for pitch, roll and yaw. As shown in Figure 2.2, a narrow range of head pose provides better performance. Despite some extremes, most faces only have a small range of pose. When head pose becomes too extreme, performance starts degrading. Then we add different portions of face images with extreme orientation to the training dataset in Figure 2.4 (a). A lower ratio of extreme data boosts the performance on “easy” and “medium” faces.

2.6.3 Effect of occlusion

We test two different kinds of face occlusion respectively. The first kind is from other objects except faces. MAFA concentrates on occluded face images, so we test on MAFA test set in three different settings: baseline condition with no occlusion, landmark occlusion setting, and mixed occlusion setting (including landmark and heavy occlusion). As shown in Figure 2.4 (b), the performance on MAFA test set improves drastically after adding occlusion in the synthetic training dataset. HR becomes more robust after training on synthetic faces with landmark and heavy occlusions. Some occlusion examples can be found in Figure 2.1.

The second kind of occlusion is from other faces or human body parts. We choose Wider Face validation set to test the effect. This is because many images in Wider Face are group pictures, and some image may have hundreds of overlapped faces. We only set the threshold for the overlap between synthetic faces, to avoid other large faces to cover the tiny faces. As shown in Figure 2.4 (c), after adding occlusion from other faces, the performance of HR for hard level in Wider validation set improves substantially.

2.6.4 Effect of noise

Every benchmark has its own configurations when established. Wider Face dataset has a bias during its collection process: the original images downloaded from search engine are resized to one predetermined width of 1024 pixels, which causes every image to have noise from down-sampling or up-sampling. Therefore, we first render images with multiple resolutions (as Set A, B and C as below) , and then re-size them to one fixed resolution (1024×768). Set A includes various high-resolution images (4096, 3072, 2048). Set B has various high- and low-resolution images (4096, 3072, 2048, 512, 256, 128). And Set C has various low-resolution images (512, 256, 128). We demonstrate the influence of noise at all the difficulty levels of Wider Face in Figure 2.4 (d). The performance at all the difficulty levels of Wider Face has been improved, especially for tiny faces. In general, set A achieved the best performance because its pattern resembles tiny faces in Wider Face.

2.7 Improve Performance through Synthetic Data

In this part, we show how to use our synthetic data to improve performance on real datasets. We train on a combination of Wider Face and synthetic data and then test on another dataset. Visualization of our detection results can be found in Figure 2.6.

2.7.1 Performance comparison on synthetic data

Before we start using synthetic data to complement real data, we first investigate the performance of different sets of synthetic data. The configurations for each set are as follows: s_1 is our basic settings for rendering with light occlusion. s_2 combines s_1 with

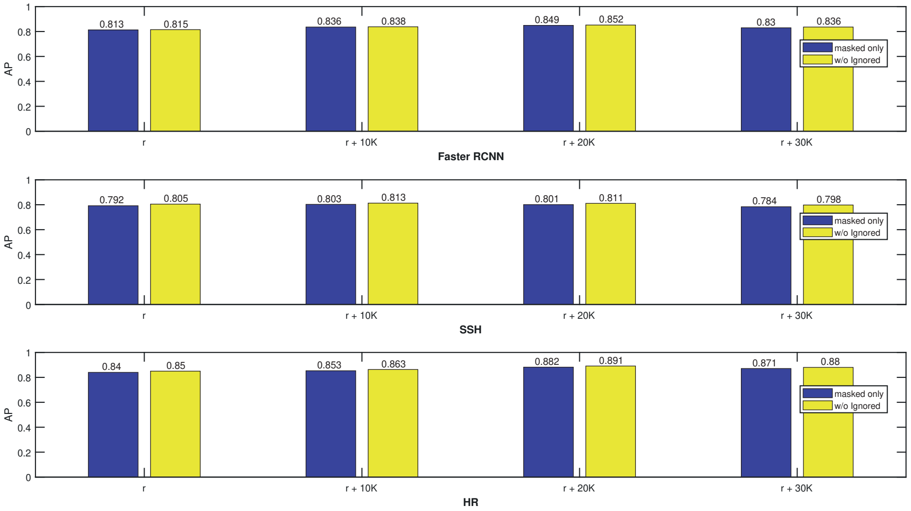


Figure 2.5: Performance comparison on different sizes of synthetic data in training on MAFA test set with different detectors. “w/o Ignored” means face with label “Ignored” are not included;

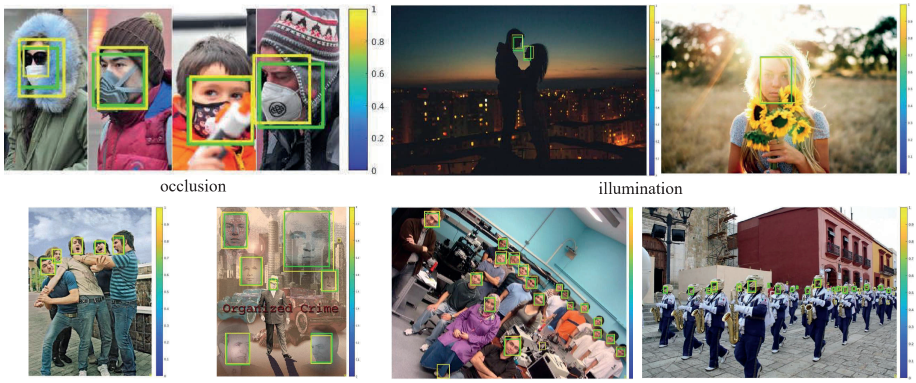
extra occlusion from other faces in the rendering process. s_3 adds additional blurry results from down-sampled high resolution images into s_2. In Table 2.5, we compare the performance of three synthetic data sets on Wider Face validation set. These three synthetic datasets s_1, s_2, s_3 are combined with real data to improve detection performance for UFDD and Wider Face respectively in Section 2.7.3 and 2.7.4.

Set	Easy	Medium	Hard
s_1	0.795	0.742	0.502
s_2	0.818	0.774	0.53
s_3	0.828	0.796	0.627

Table 2.5: Average precision from HR trained on different sets of synthetic data. These three sets are combined with real data to improve detectors’ performance.

2.7.2 Evaluation on MAFA

We only use mixed face occlusion (that is landmark and heavy occlusion as in Section 2.6). The synthetic images for data augmentation follow the setting of MAFA training set as closely as possible. We presume the training data size would have an effect on the performance. As shown in Figure 2.5, the performances of different detectors improve to some extent with the increase of synthetic data. We do note that when we add more synthetic data, the performance saturates and drops, however. Our tentative explanation is the inherent bias in our synthetic data.



occlusion

illumination

expression

blur

pose

scale

Figure 2.6: Qualitative results on different features of a real dataset. We visualize examples of each feature. The green bounding boxes are ground truth. The bounding boxes with other colors are predictions with different confidence intervals.

2.7.3 Evaluation on UFDD

We show the influence of our data augmentation in Table 2.6. Three different synthetic sets are combined with real data to improve detectors' performance. "r" denotes the detector is trained on real data. "$r + s_1$" denotes the detector is trained on a combination of real data and synthetic data set s_1. s_1 is our basic settings for rendering with light occlusion. s_2 combines s_1 with extra occlusion from other faces in the render process. s_3 adds additional blurry results from down-sampled high resolution images into s_2. After merging synthetic data and real data together, the performance of Faster RCNN, which was trained on real data, has been improved significantly on $r + s_1$ and $r + s_2$. Given that Faster RCNN was not trained on different scales, the noise of s_3 impedes it performance. As for SSH, its architecture and parameters heavily relies on Wider Face. The features in UFDD is too difficult for its light-weight structure. Unsurprisingly, its performance becomes saturated after being trained on real data. After adding synthetic data, its performance is even worse than Faster RCNN. Faces in UFDD are not very challenging to HR; the performance therefore only changes slightly after using our data augmentation.

Detectors	r	$r + s_1$	$r + s_2$	$r + s_3$
Faster RCNN	0.64	0.745	0.742	0.64
SSH	0.681	0.682	0.672	0.674
HR	0.721	0.733	0.721	0.736

Table 2.6: Performance comparison on different data augmentations on UFDD test set with different detectors.

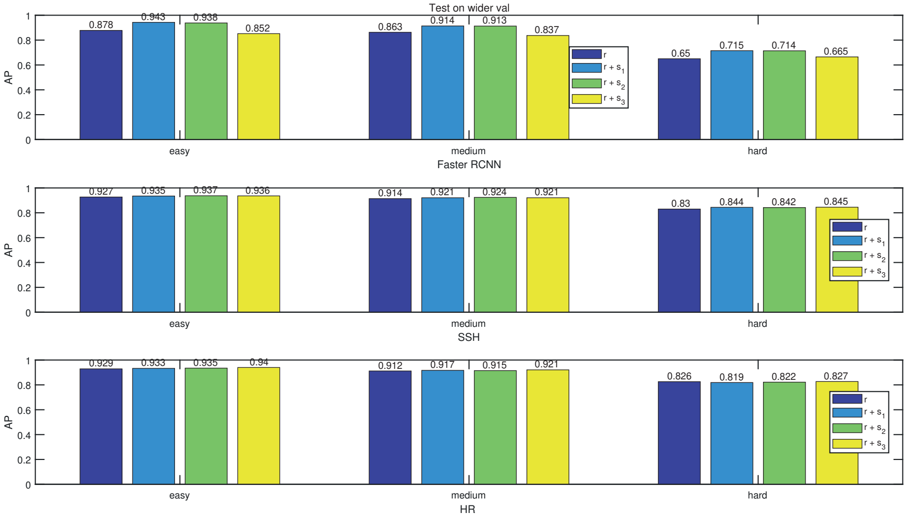


Figure 2.7: Performance comparison on different data augmentations on Wider Face validation set with different detectors.

2.7.4 Evaluation on Wider Face

We still use the same setting as in Section 2.6 to perform data augmentation for Wider Face. The performance comparisons are showed in Figure 2.7. Faster RCNN is a generic object detector without multi-scale inference, so it is supposed to generate fewer predictions than HR and SSH. After we add synthetic data, the performance substantially improves on all difficult levels. The performances of HR and SSH nearly saturate after being trained on real data. Although their architectures aimed at Wider Face dataset, our synthetic data can still improve performance to a certain extent.

2.8 Analysis

2.8.1 Analysis of object features

In general, proper amount of well-structured synthetic data can be a good complement to real data in training face detectors. The settings of synthetic data need to be similar with configurations of real data. If targeting at a single feature, the increase of data quantity cannot yield a consistent improvement on performances. For a more complex dataset like Wider Face, synthetic face images are generated with a comprehensive combination of several features.

The advantage of synthetic data is that the variations in dataset can be fully controlled in different practical situations. The dataset could be adjusted depending on specific requirements. Although, admittedly, there is always a domain gap between synthetic data and real data, synthetic data can provide a large-scale dataset with annotation conveniently and precisely. Our results on multiple challenging benchmarks with different advanced detectors highlight the applicability of synthetic data as complement to real data, to equip face detectors against various features.

2.8.2 Analysis of face detectors

Based on our detection results, we analyze the performance of three detectors respectively. Faster RCNN is an object, instead of face, detector. It does not adjust settings of anchors for any face detection benchmarks, and has much fewer predictions without multi-scale inference. Despite that, our synthetic data augmentation substantially improves its performance on multiple challenging datasets.

SSH is the representative of one-step face detector. Even though SSH is a face-targeted detector, adding our synthetic data augmentation cannot help it outperform Faster RCNN in most detection tasks except in hard level of Wider Face. Specialized in scale, SSH sometimes has imperfect performance when encountering other features. Detectors have a trade-off between speed and performance. SSH pursues fast speed in inference process so that its light weight architecture cannot handle other features.

Of crucial importance, although HR face detector already has excellent performance in terms of all kinds of features, our synthetic data still boosts its performance. However, HR has a drawback that is extremely sensitive to tiny distracting objects given its tiny-face-targeted architecture. In general, HR generates more false positives than other detectors. The design of HR restricts the generalization on normal faces.

2.8.3 Analysis of false positives

False positive is a primary factor that jeopardizes performance. In Figure 2.8, we plot some false positive examples from our detection results. There are two major sources of false positive in our detection results.

The first source is annotation. Different datasets have their own annotation policy. This deviation may turn our reasonable predictions into false positives. For example, annotation of occluded faces in Wider Face estimated the region of entire faces (Figure 2.8 (c)). In comparison, UFDD often annotates the visible part of the occluded faces (Figure 2.8 (b)), while MAFA uses square annotation, which may contain background information surrounding faces (Figure 2.8 (a)). Moreover, a human annotator sometimes cannot annotate all the tiny and blurry faces in the background. In Figure 2.8 (f), for example, there are only three annotations for nearly a hundred faces. We detected more blurry and tiny faces than those annotated.

The second source of false positives is misleading objects, such as round-shaped objects and human body parts. Because real data has much more diverse and complex background than synthetic data. All of our synthetic images are rendered from 3D face models. Most 3D face models only depict the upper part of human bodies. Therefore, our rendering results are inherently unrepresentative of other human body parts or clothes. As a result, other human body parts (see Figure 2.8 (e)) and some accessories from clothes would become a major source of false positives.

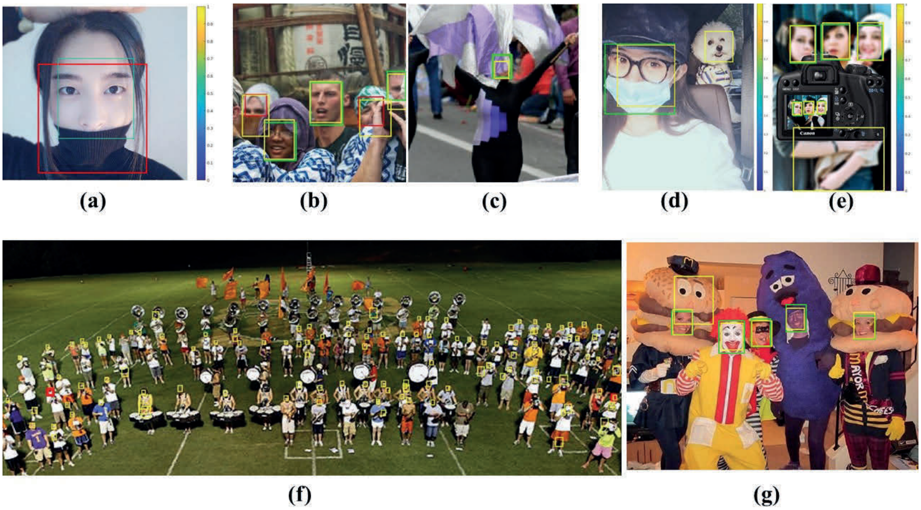


Figure 2.8: False positive examples of our detection results. The green bounding boxes are ground truth. The bounding boxes with other colors are predictions with different confidence intervals, and the red bounding boxes are false negative examples. (a) is a square annotation example from MAFA test set. (b) and (c) are occlusion annotations respectively in UFDD and Wider Face. (f) only has three annotations in ground truth but actually has way more unlabeled faces. Please zoom in to see some small detections. (d), (e) and (g) are some examples from real dataset.

2.9 Conclusion

In this chapter, we proposed an experimental comparison of main characteristics that influence face detection performance. We customized synthetic dataset to address specific types of features (scale, pose, occlusion, blur, etc.), and systematically investigated the influence of different features on face detection performance. Through our analyses, we also identified some potential deficiencies of the current face detection architectures. To conclude, there are often challenging features in real-world face detection. By providing an overview of the relationship between object features and face detection performances, we hope to assist researchers to choose more appropriate synthetic data when addressing challenging real-life variations.

3

Pose Invariant Age Estimation of Face Images in-the-wild

3.1 Introduction

Age estimation is an important topic in computer vision with various applications in areas of, for example, human-computer interaction [68], surveillance and social networking sites. Aging is an irreversible process that causes both appearance and geometrical variations on human faces [5]. Facial aging process can be both affected by personal (e.g., mental/physical states) and situational factors (e.g., living environment).

Age estimation tasks can be classified into two categories: biological and apparent age estimation. Traditional age estimation for still images widely employs biologically inspired features [66, 103], while estimation using videos often utilizes handcrafted features and temporal dynamics [40, 42].

In recent years, deep neural network based methods have achieved remarkable performances in multiple face-related tasks. Despite the advancement of deep learning based age estimation, the application of these methods in unconstrained scenarios is still far from ideal. Image variations like extreme occlusion, illumination and blur may negatively influence the performance of these age estimation methods. Moreover, most of the age estimation algorithms focus on still images. Only a limited number of methods are applied to face videos. One of the main reasons is the lack of large-scale face video datasets with age annotation.

In this paper, we introduce a new video dataset named UvAge, which contains face videos in-the-wild for the purpose of age estimation. This dataset comprises videos of celebrities from the Internet. Information extracted from Wikipedia provides us with the identity, age, gender and ethnicity labels for each video. Compared with previous single-setting video datasets (e.g., UvA-Nemo dataset from [41], the newly created dataset is a one-of-a-kind dataset containing recordings in-the-wild for a substantial variety of scenarios. In addition to this new dataset, we also propose a new method for age estimation of face videos.

Our method is using a pose invariant representation. To obtain such a representation, a face uv texture representation is computed from the video frames [48], which results in a pose invariant texture image containing the estimated frontal view of the face. As head pose changes may cause self-occlusion, the uv map may be negatively affected by

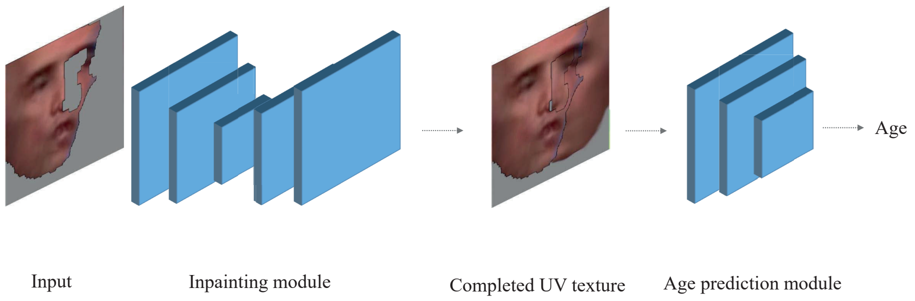


Figure 3.1: Overview of our proposed method. The original frames are obtained from the video segments. 3D face uv texture is retrieved from reconstruction pipeline. Input of the network is the self-occluded uv texture. Age is predicted based on the completed uv texture from inpainting module

missing face parts. To address this problem, we use a Wasserstein GAN based network to complete the missing parts of the uv texture map. Compared to standard inpainting tasks, the occlusion-included missing face regions are highly irregular. Therefore, the GAN based approach (AgeGAN) learns to complete missing regions and estimates the age at the same time.

Our contributions are:

- The creation of the UvAge Dataset, the largest video dataset containing face videos in-the-wild with identity, age, gender, ethnicity labels for each video.
- A new GAN-based approach (AgeGAN) to complete the facial uv texture and to estimate the age.
- Our proposed method outperforms state-of-the-art age estimation methods on the UvAge dataset.

3.2 Related Work

3.2.1 Age datasets

Deep learning for age estimation heavily depends on the availability of large-scale age datasets with annotation. A number of early datasets like the Productive Aging Laboratory (PAL) database [129], Adience [46], VADANA [172], Gallagher group photos [51] and PubFig [101] categorize images into different age groups. In the FG-NET dataset [140], each subject has more than 10 photos taken at different ages. MORPH-II [150] is one of the largest longitudinal face image dataset. Cross-Age Celebrity Dataset (CACD) [23] is a cross-age dataset of celebrities. IMDB-WIKI dataset [153] is composed of celebrities' images from IMDB and Wikipedia. CLAP2016 [47] is collected via crowd sourcing, which provides the mean age and variance (considered as ground truth) for each image. The MegaAge [219] dataset used the FG-NET data as reference to label images from other large scale face datasets. AFAD [135] dataset focused on Asian people age to mitigate ethnicity homogeneity in existing datasets. FIA [59] database was created with subjects' age and gender while only a limited number of participants and a narrow range of ages are avaliable. AGFW-v2 [45] is an extension of AGFW [134], which contains still images and 100 videos of celebrities. The UvA-Nemo dataset is developed to analyze the temporal dynamics of spontaneous/pose smile videos for different ages [41], collected under constrained conditions (regarding camera setting, face pose and illumination). When face video datasets are recorded under constrained (vs. unconstrained) settings, it is more feasible for methods to concentrate on age related features, which, however, fails to capture the influence of other variations (e.g., occlusion, pose, blur) in real-world scenarios. Moreover, these previous video datasets are largely confined to the collection policies with constraints on ethnicity, occupation and living environment in subjects' aging process.

3.2.2 Age estimation

Systematical reviews on age estimation can be found in [51, 140, 148]. Most age estimation methods are based on still images. Regression based methods are the most straightforward approach to estimate ages [116], while they usually do not adopt the ordinal property embedded in the age information. Ranking based approaches instead tend to make use of informative ordinal relationship between age labels [20, 21, 24]. In order to adapt to the recent progress of deep neural network, some methods deem age estimation as classification task by considering age or age groups as independent classes [105]. Label distribution learning (LDL) is an alternative to the classification-based age estimation problem [55, 56, 206]. Compared to the standard classification problem, age labels are ordinal and comparable. LDL based methods are capable of predicting a range with different confidence levels instead of a single value. [139] introduced the mean-variance loss to provide concentrated estimation. [110] add bridge connections into a random forest model to enforce the continuity among neighbor nodes of trees. [216] proposes a light-weighted and effective age estimation method based on cascade network.

Nonetheless, only a few methods uses videos for age estimation till now. [67] used local binary features to extract spatio-temporal information of faces in videos. [144] combined spatial and temporal features in age estimation of videos. [40, 42] investigated aging characteristics in facial dynamics.

3.2.3 Face attributes completion and manipulation

Most completion approaches rely on extracted features to search for patches from context and then synthesize the contents to the matched patches. This strategy performs particularly well for background inpainting. Performance of these approaches are bounded to existing features of the known regions. However, similar patterns may not always appear in the original partial images when completing missing face images. Inpainting for faces is extremely challenging when the missing attribute is unique and does not replicate any part of the background. [113] directly generates contents to complete missing regions for faces instead of searching relevant patches. [39] proposes a GAN based approach to complete missing parts in facial uv texture.

Face attributes manipulation is a promising direction for generative methods. Effective manipulation on facial attributes can be considered as data augmentation to improve the performance of face-related tasks. [196] proposes a conditional GAN-based method with an identity preserved module to simulate the aging process. [203] used a pyramid architecture of GAN to generate face age progression. [138] considered face aging problem as an image style transfer task. [45] proposed a deep reinforce learning based approach to generate facial age progressing effects in videos.

3.3 The UvAge dataset

3.3.1 Collection

To acquire biological ages of subjects, our approach focuses on videos of celebrities. It is generally difficult to obtain accurate age information of people from videos because it is almost impossible to obtain the precise recording time of a specific video or people's birth information. To solve this problem, we collected videos of celebrities which were recorded in identifiable events. First, from Wikipedia pages, we selected a vast number of traceable events of different topics (e.g World Cup, Academy awards, G20 summit). The Wikipedia pages of selected events provide name lists of the related celebrities. From the celebrities' own Wikipedia pages subsequently, we obtained their respective birth information. Accordingly, their age information is inferred from the time of the event and their year of birth. Finally, videos are collected from the Internet using key words like "2018 World Cup Luka Modrić interview", "2016 G20 Hangzhou summit Barack Obama speech". After the raw videos are collected, irrelevant videos are manually removed. Each video is segmented into separate video shots, while each segment captures a single person. The final dataset contains persons with different professions (e.g., politicians, entrepreneurs, sportsman, actors, and the like).

To better understand the influence of extreme cases on the performance of age estimation, hard videos are identified with extreme imaging conditions like image blur, extreme scale and strong illumination. The number of videos for each celebrity is not constrained. For each video segment, the following labels are provided: identity, biological age, gender, ethnicity and occupation.

3.3.2 Statistics

The UvAge dataset consists of 6898 video segments from 516 subjects. There are 2176 videos of 212 female subjects. We split the dataset into 396, 65 and 55 subjects for training, validation and testing respectively. We used the pre-trained weights from IMDB-WIKI to guide our splitting process. Specific information about different partitions are presented in Table 3.2. The video distribution for each individual was not strictly balanced due to the fact that available videos of each celebrity on the Internet were considerably different. We try to balance all the partitions in subjects, videos and performance domains. The splitting process followed the subject-exclusive protocol.

Dataset	MMI [187]	CMU FIA [59]	UvA-Nemo [41]	UvAge
subject number	25	180	400	516
video number	2005	6470	1240	6898
age range	20-32	18-57	8-76	16-83
illumination	constrained	constrained	constrained	unconstrained
pose	constrained	unconstrained	mostly frontal	unconstrained
ethnicity	-	-	mostly Caucasian	all
occupation	no	no	no	yes

Table 3.1: Comparison among different face video datasets with age annotation. Here we compare the proposed UvAge dataset with previous datasets on crucial indicators.

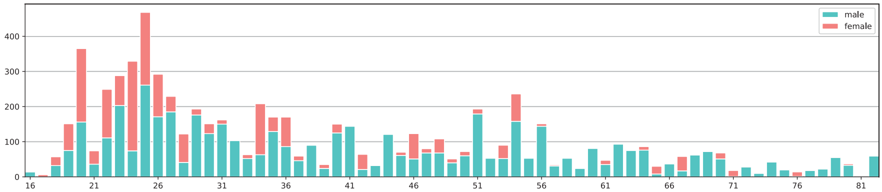


Figure 3.2: Age distribution of UvAge dataset. The horizontal axis is age. The vertical axis is the number of subjects. Blue bars represent males and red bars represent females.

3.3.3 Comparison with existing datasets

Most face video datasets with age annotation are not designed for age estimation; Therefore, they often fail to consider all the aspects that could affect the age estimation performance. We compare UvAge dataset with previous datasets in Table 3.1. In comparison with other face age video datasets like MMI Facial Expression Database [187] and CMU Face In Action (FIA) [59] database, our dataset includes a larger diversity of subjects and a wider age range. The video source and age labeling process of AGFW-v2 [45] video subset is quite similar to our collection process. However, their data (with 100 videos) is restricted in investigating the influence of different variations. Another dataset UvA-Nemo is primarily targeted at collecting facial expression videos. Therefore, the videos of UvA-Nemo dataset are recorded under constrained lighting conditions with high resolution cameras. The poses of subjects are mostly frontal and do not change much across frames. Due to its recording policy, participants comprising UvA-Nemo are largely homogeneous with relatively low inter-subject variation.

3.4 Our Method

Our age estimation method aims to provide a pose-invariant age estimation pipeline. Pose variation could significantly change the appearance of faces and add extra occlusion. Compared with 2D images, 3D face uv textures are insusceptible to face orientation, which can considerably attenuate the interference of pose.

PRNet [48] is used to reconstruct the face uv texture to represent the 2D faces from the original frames. uv texture assigns 3D texture into 2D space with universal per-pixel alignment for all textures. A common approach to create uv texture is to unwrap 3D texture in a cylindrical manner. Each vertex in a 3D shape has a corresponding 2D texture coordinate. The reconstructed uv texture is partially invisible due to the effect of self-occlusion from the face itself. We assume that a completed uv texture map provides better visual features for age estimation. Using a uv texture map as input of a network facilitates concentration on the feature associated with ages and reduces the influence of pose variation. We presume that the central region of the face uv texture map is the most influential feature for ages. Thus, we use the cropped face uv in our experiments.

3.4.1 Inpainting module

The application of inpainting in real-world scenarios can be extremely challenging. While most previous inpainting networks train on large-scale datasets with synthesized missing regions, of which the size and shape of missing regions are controllable, the shape of missing regions in our task is highly irregular, more so when some uv textures have more than 50% parts missing. Moreover, the ground truth of synthesized missing regions is available so that the network can be forced to learn how to extract the background to fill in the missing parts. In contrast, the uv textures of faces in original frames are not available in our task. Therefore, we need a stable inpainting network as basic architecture.

We use a modified Wasserstein GAN-based inpainting network from [211] to complete the missing part in face uv texture. It is a coarse-to-fine network achieving good results on different inpainting tasks with regular shape masks. A binary mask image M is used to represent the missing region. The value is 0 for the missing pixels and 1 for the rest. The inpainting network is pre-trained on a subset of uvdb dataset [39] with an irregular shape mask to contract the convergence progress. The inpainting output is

$$I_{gen} = I_{ori} * M + I_{pred} * (1 - M). \tag{3.1}$$

where I_{ori} is the original input uv image, I_{pred} is the output from the inpainting network, and I_{gen} is the final uv texture for predicting ages. N denotes the batch size.

In order to complete large missing regions, we also introduce symmetry loss on completed images. Human faces share horizontal features. Therefore, we enforce the generated part to resemble the symmetrical part horizontally. As shown in equation 3.2, I_{flip} is the horizontally flipped version of I_{ori} and I_{gen} is the output image from the inpainting network. Total variational loss [92] is also added into the final loss to improve the quality of generated images.

$$L_{sym} = \frac{1}{N} \| M * (I_{flip} - I_{gen}) \|_2. \tag{3.2}$$

3.4.2 Age prediction module

The age prediction module of our architecture is adapted from VGG16 [168]. Similar to [139], the age loss function has three parts: softmax loss, mean loss and variance loss.

Normal classification task would penalize all the prediction values equivalently except for the ground truth. We additionally penalize the mean and variance of predicted age distribution, which can lead the network to output a compact age prediction distribution towards the ground truth. Age loss is added to the generator part of the total loss function. Age is predicted based on the completed uv texture map.

The mean and variance of the predicted distribution from network is calculated by equations 3.3. $y_i \in \{1, 2, ...K\}$ denotes the ground truth age for each input image, $j \in \{1, 2, ...K\}$ is the age label and $p_{i,j}$ is the probability of input image i belonging to class j.

$$\bar{y}_i = \sum_{j=1}^{K} j * p_{i,j}, \quad var_i = \sum_{j=1}^{K} p_{i,j} * (\bar{y}_i - y_i)^2 . \tag{3.3}$$

Mean loss and variance loss are further obtained by equations 3.4 and 3.5, respectively.

$$L_{mean} = \frac{1}{N} \sum_{i=1}^{N} (\bar{y}_i - y_i)^2 , \tag{3.4}$$

$$L_{var} = \frac{1}{N} \sum_{i=1}^{N} var_i . \tag{3.5}$$

Our loss function for the age prediction task is in equation 3.6.

$$L_{age} = \lambda_1 L_{cls} + \lambda_2 L_{mean} + \lambda_3 L_{var} . \tag{3.6}$$

In equation 3.7, L_G is the final loss function of the generator. L_{W-G} is the WGAN-GP loss [64].

$$L_G = L_{W-G} + L_{age} + \lambda_4 L_{sym} + \lambda_5 L_{tv} . \tag{3.7}$$

3.5 Experiments

3.5.1 Implementation details

We use the Dlib frontal face detector to obtain a face bounding box for each frame. The side length of the bounding box is extended by 20% before inputting it to the network. For our AgeGAN, the input size of the cropped uv is 192 * 192. The mask image has the same shape as the cropped uv. The subset of uvdb for pre-training is 2000 images. The parameters for our loss function are given by $\lambda_1 = 10$, $\lambda_2 = 2$, $\lambda_3 = 0.05$, $\lambda_4 = 1e^{-5}$, $\lambda_5 = 1e^{-4}$. The learning rate is 0.0001. The batch size is 16. The age prediction module uses pretrained models on Imagenet.

	Train	Validation	Test
Subjects number	396	65	55
Video number	5469	712	717
MAE	9.23	9.01	9.99

Table 3.2: Here we compare some basic information between different partitions of UvAge dataset. The MAE (mean average error) is from pretrained weights on IMDB-WIKI.

Methods	Valid	Test
DEX [153]	10.49	8.39
SSR-Net [205]	9.18	7.71
MV [139]	8.83	6.65
DRF [164]	7.57	6.15
Ours	7.45	5.82

Table 3.3: Comparison between the different methods. Here we compare our method with other age estimation architectures.

3.5.2 Comparision with other methods

For all the following experiments, we use the Mean Absolute Error (MAE; unit: year) to measure the performance of the age estimation methods. This is the average of the absolute deviation between the prediction and the ground truth age.

We train and evaluate our method on UvAge dataset. To focus on the most representative feature of age estimation, we crop the central region from the original uv texture. Random noise is added to replace the missing part. Some irregular masks are then selected from other real partial uv's to simulate the missing effect. The missing region in each mask image or input uv image is not allowed to be larger than 50%.

The performance comparison of the different age estimation methods is shown in Table 3.5.2. We report performance on both validation and test set of UvAge dataset. All experiments are based on the same split of the UvAge dataset. We compare our method with DEX [153], SSR-Net [205], DRF [164, 165] and MV [139]. All the methods are using pretrained weights from ImageNet. The performance of SSR-Net is slightly worse than DRF and MV. The architecture of SSR-Net avoids large amount of neurons and leads to a more compact model. However, the light-weight structure may not be robust enough to handle the challenging variations in UvAge dataset. DRF uses a VGG based random forest architecture for age estimation. The convergence of both the backbone architecture and the random forest method is crucial to the final performance.

Compared to other methods, our method facilitates improved accuracy in age estimation. A completed and frontal uv texture can help the network to retrieve the most influential feature for age estimation. Our method is robust to the influence of pose changes.

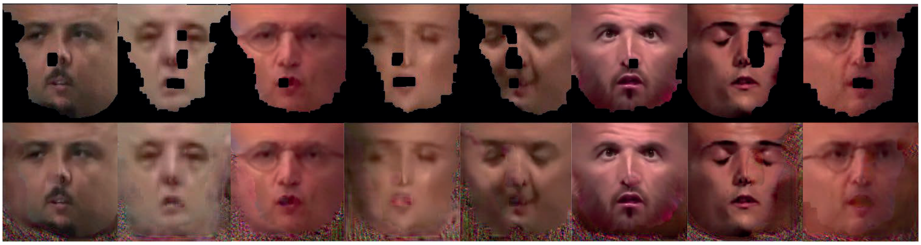


Figure 3.3: Inpainting results from our method. The first row contains the self-occluded uv textures and the second row shows the inpainted results.

Methods	DEX	DRF	Ours
MAE	4.785	4.610	4.53

Table 3.4: We compare our approach to other age estimation methods on CACD test subset.

3.5.3 Ablation study

To further analyze the improvement of the proposed method, we conduct more experiments as listed in Table 3.5. We remove the inpainting module from our method to examine the effect of simply training on cropped uv images. It shows that training on completed uv images is better than training on self-occluded uv images. The inpainting module generates more useful information for age estimation.

3.5.4 Evaluation on CACD

We run additional experiments on Cross-Age Celebrity Dataset (CACD) [23] dataset. The purpose of this dataset to collect face age images with large gaps. Compared to other existing dataset, age gaps from the same subject are larger. Images from CACD are collected based on images from Internet, it contains some noisy images. We train our age estimation pipeline on reconstructed uv textures from CACD train subset. In Table 3.4, we show the performance comparison on CACD test subeset with other methods.

Train set	Valid	Test
Crop uvs	9.90	7.59
Completed uvs	7.45	5.82

Table 3.5: Comparison of different training settings.

3.5.5 Quantitative analysis

To investigate the performance on extreme challenging cases, we select all the images with a MAE larger than 20 years from our experiment results. Most of these poor predictions are images of senior people. The network tends to give younger predictions for these individuals. The UvAge dataset follows a subject-exclusive protocol. For some large age label (larger than 70), only a few individuals are included. Due to personal and situational factors, each individual's aging progress could be significantly different from others.

3.5.6 Qualitative analysis

Face completion

In Figure 3.3, we plot a number of results for our inpainting process. The first row contains the self-occluded uv textures and the second row is the result after inpainting. The missing regions for face textures are irregular shapes. Generally, the missing part in the central region of faces is better recovered by the inpainting network. Local face attributes like nose or mouth can be inpainted in a reasonable way. Some artifacts are found around the boundary of missing regions. There are also some failing cases when missing regions on faces are too extreme. The network cannot extract enough information from the background and may add some background noise to the missing part.

Age estimation

In Figure 3.4, we plot a number of original 2D frames with their predictions and ground truth. Generally, the hard cases are from two categories: imaging variation and personal factor. Imaging variation like blur or extreme illumination can impact every step in our method.

As for personal factors, human faces share some common features for different ages. However, the aging process for each person can be different. Even people of the same age could have quite different appearance. These effects are caused by personal or situational factors. Additionally, makeup is negatively influencing the age estimation pipeline.

3.6 Conclusion

In this chapter, we introduced the largest in-the-wild video dataset for age estimation. It contains unconstrained videos from celebrities in different events. To make age prediction more robust against pose variation, face uv textures are reconstructed from the 2D frames of videos. We provide a W-GAN based approach (AgeGAN) to simultaneously estimate the real age and complete the partial uv textures. Serial experiments demonstrate the effectiveness of our proposed method.

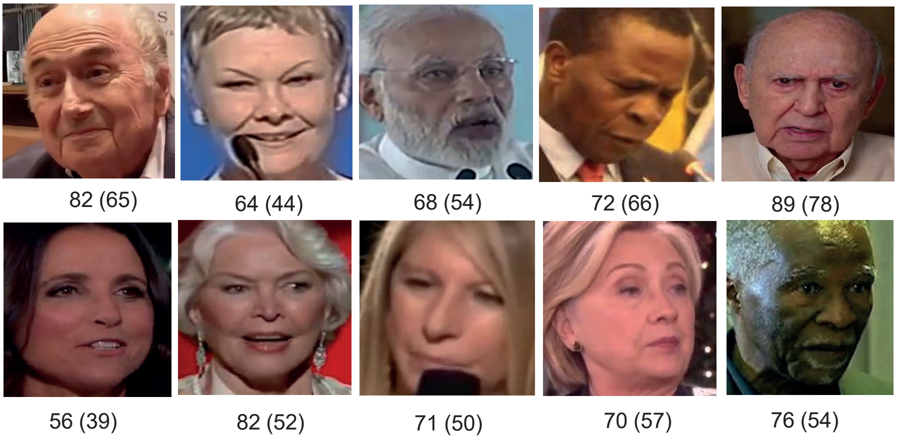


Figure 3.4: Hard video examples from UvAge Dataset. We plot the face region which is cropped from original frames. The numbers below each image show the ground-truth and prediction, i.e., ground-truth age (estimated age).

4

MMD based Discriminative Learning for Face Forgery Detection

4.1 Introduction

With the rapid development of face manipulation and generation, more and more photo-realistic applications have emerged. These modified images or videos are commonly known as deep fakes [152]. Even human experts find it difficult to make a distinction between pristine and manipulated facial images. Different generative methods exist nowadays to produce manipulated images and videos. In fact, it's easy to generate new types of synthetic face data by simply changing the architectural design or hyper parameters. Attackers don't need to have profound knowledge about the generation process of deep fake (face) attacks [3]. Therefore, it is of crucial importance to develop robust and accurate methods to detect manipulated face images.

Face forensic detection is to distinguish between manipulated and pristine face images. Using the same pair of subjects, different manipulation methods may generate significantly different outcomes (see Figure 4.1). If the same modification method is applied on different pairs of data, the results can have quite diverse artifacts due to the variations in pose, lighting, or ethnicity. Because these artifacts do not exist in all samples, simple artifacts-based detection systems are not sufficiently robust to unseen artifacts in the test set. Other methods choose to exploit cues which are specific for the generative network at hand [191, 212]. When the dataset is a combination of multiple domains of deep fake data like FaceForensics++, each category of manipulated face images can be a different domain compared to the rest of the data. Performance may be negatively affected by this cross-domain mismatch. In summary, the major challenges of face forensics detection are: 1) The difference among positive and negative samples is much smaller than the difference among positive examples. 2) The artifacts including imaging variations and face attributes do not persist across all generated results for a single generation method.

Our paper focuses on detecting manipulated face images which are produced by generative methods based on neural networks. To distinguish real and fake face images is equivalent to performance evaluations of different generative methods. To this end, the maximum mean discrepancy (MMD) is used to measure different properties and to analyze the performance of different generative adversarial networks [202]. The face

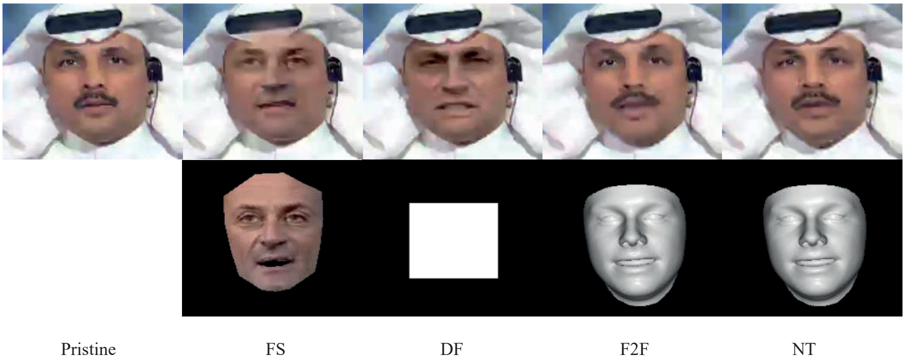


Figure 4.1: Visualization of a number of samples from FaceForensics++. The first row shows pristine and generated face images and the second row contains face masks used to add the modifications. “DF”: “DeepFakes”; “NT”: “Neural Textures”; “FS”: “Face Swap”; “F2F”: “Face2Face”. Although NT and F2F share the same face mask, NT only modifies the region around the mouth.

manipulation process requires a pair of faces from source and target subjects. This process resembles the neural style transfer operation between content and reference images [52]. The final results are contingent on the source and target images. Inspired by [114], we use a MMD loss to align the extracted features from different distributions. A triplet loss is added to maximize the distance between real and fake samples and to minimize the discrepancies among positive samples. Center loss is further integrated to enhance the generalization ability. In order to fully investigate the performance of the proposed method, we evaluate our method on several deep fake benchmarks: DF-TIMIT [99], UADFV [111], Celeb-DF [115], and FaceForensics++ [152].

Our main contributions are:

- We propose a deep network based on a joint supervision framework to detect manipulated face images.
- We systematically examine the effect of style transfer loss on the performance of face forgery detection.
- The proposed method achieves the overall best performance on different deep fake benchmarks.

4.2 Related Work

4.2.1 Face manipulation methods

In general, face manipulation methods can be classified into two categories: facial reenactment and identity modification [152]. Deep fake has become the name for all face modification methods. However, it is originally a specific approach based on an auto-encoder architecture. Face swap represents methods that use information of 3D face models to assist the reconstructing process. Face2Face [180] is a facial reenactment framework that transfers the expressions of a source video to a target video while maintaining the identity of the target person. NeuralTextures [181] is a GAN-based rendering approach which is applied to the transformation of facial expressions between faces.

4.2.2 Face forgery detection

A survey of face forensics detection can be found in [184]. Several methods are proposed to detect manipulated faces [28, 36, 147, 192]. While previous literature often relies on hand-crafted features, more and more ConvNet-based methods are proposed. There are two main directions to detect face forgery.

The most straightforward approach is data-driven. Forensic transfer [31] uses the features learned from face forensics to adapt to new domains. [132, 173] combine forgery detection and location simultaneously. [87] uses a modified semantic segmentation architecture to detect manipulated regions. Peng et al. [222] provide a two stream network to detect tampered faces. Shruti et al. [3] concentrate on detecting fake videos of celebrities. Ghazal et al. [128] proposes a deep network based image forgery detection framework using full-resolution information. Ekraam et al. [155] propose a recurrent model to include temporal information. Irene et al. [4] propose an optical flow based CNN for deep fake video detection.

Another category of deep network-based methods is to capture features from the generation process. The features include artifacts or cues introduced by the network [102]. The Face Warping Artifacts (FWA) exploit post processing artifacts in generated videos [111]. Falko et al. [126] use visual artifacts around the face region to detect manipulated videos. [133] proposes a capsule network based method to detect a wide range of forged images and videos. Xin et al. [207] propose a fake video detection method based on inconsistencies between head poses. [18, 37] exploit the effect of illumination. [191] monitors neuron behavior to detect synthetic face images. [212] uses fingerprints from generative adversarial networks to achieve face forensic detection. And [109] proposes a framework based on detecting noise from blending methods.

4.2.3 Domain adaptation

Domain adaptation has been widely used in face-related applications. It aims to transfer features from a source to a target domain. The problem is how to measure and minimize the difference between source and target distributions. Several deep domain generalization methods are proposed [106, 186] to improve the generalization ability. Rui et al. [160] propose a multi-adversarial based deep domain generalization with a triplet constraint and depth estimation to handle face anti-spoofing. Maximum mean discrepancy (MMD) [62] is a discrepancy metric to measure the difference in a Reproducing Kernel Hilbert Space. [107] uses MMD-based adversarial learning to align multiple source domains with a prior distribution. [114] considers neural style transfer as a domain adaptation task and theoretically analyzes the effect of the MMD loss.

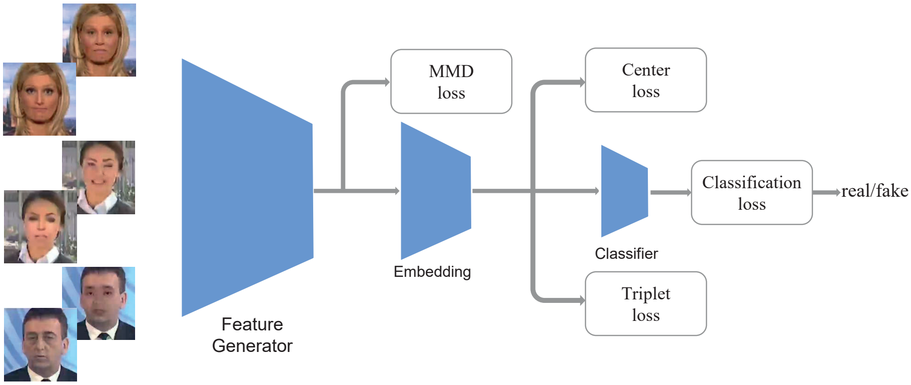


Figure 4.2: Overview of the proposed method. Inputs of the network are frames of manipulated face videos. A deep network is used to extract features. Here we use the cross-entropy loss for binary classification. A MMD loss is added to learn a generalized feature space for different domains. Moreover, the triplet and center losses are integrated to provide a discriminative embedding.

4.3 Method

4.3.1 Overview

Most current forensic detection approaches fail to address unique issues in face manipulation results. The learned features may not generalize well to unseen deep fake samples. Some approaches choose to extract features from the modification process (e.g., detecting artifacts in manipulated results). Nevertheless, artifacts are dependent on the discrepancy between source and target face images. The discrepancy may originate from differences in head pose, occlusion, illumination, or ethnicity. Therefore, artifacts may differ depending on the discrepancies between source and target face images. Other methods choose to exploit characteristic cues induced by different generative models. However, any minor changes in the architecture or hyper-parameter setting may negatively influence the forgery detection performance. In contrast, our aim is a generic approach to forgery detection.

To this end, we propose a ConvNet-based discriminative learning model to detect forgery faces. A maximum mean discrepancy (MMD) loss is used to penalize the difference between pristine and fake samples. As a result, extracted features are not biased to the characteristics of a single manipulation method or subject. A center loss is introduced to guide the network to focus on more influential regions of manipulated faces. Furthermore, a triplet loss is incorporated to minimize the intra-distances. We consider the task as a binary classification problem for each frame from real or manipulated videos. In Figure 4.2, we provide an overview of the proposed framework. Input images are pristine and fake face samples from a deep fake dataset.

4.3.2 MMD based domain generalization

Maximum mean discrepancy (MMD) measures the difference between two distributions. MMD provides many desirable properties to analyze the performance of GAN [202]. In this paper, we use MMD to measure the performance of forgery detection as follows. Suppose that there are two sets of sample distributions P_s and P_t for a single face manipulation method. The MMD between the two distributions is measured with a finite sample approximation of the expectation. It represents the difference between distribution P_s and P_t based on the fixed kernel function k. A lower MMD means that P_s is closer to P_t. MMD is expressed by

$$MMD^2(P_s, P_t) = \mathbb{E}_{x_s, x'_s \sim P_s, x_t, x'_t \sim P_t}\left[k(x_s, x_s^{'}) - 2k(x_s, x_t) + k(x_t, x_t^{'})\right]. \quad (4.1)$$

where $k\left(a,b\right) = \left\langle \phi\left(a\right), \phi\left(b\right)\right\rangle$ denotes the kernel function defining a mapping. ϕ is an explicit function. s, t denote the source and target domains respectively. x_s and x_t are data samples from the source and target distributions.

MMD is used to measure the discrepancies among feature embeddings. The MMD loss has been used in neural style transfer tasks [114]. Different kernel functions (Gaussian, linear, or polynomial) can be used for MMD. The MMD loss is defined by:

$$L_{mmd} = \frac{1}{W_k^l}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(k(f_i^l, f_i^l) + k(r_j^l, r_j^l) - 2k(f_i^l, r_j^l)\right). \quad (4.2)$$

where W_k^l denotes the normalization term based on the kernel function and the feature map l. M and N are numbers of fake and real examples in one batch, respectively. f and r are the features for fake and real examples. The MMD loss can supervise the network to extract more generalized features to distinguish real from fake samples. It can be considered as an alignment mechanism for different distributions.

4.3.3 Triplet constraint

For several manipulated videos which are generated from the same video, the background of most samples is the same. Face reenactment methods can manage to keep the identity of the original faces constant. Meanwhile, modifications from generative methods become more and more subtle. This makes the negative examples look more similar than the positives ones for the same subject. As shown in Figure 4.1, images which are generated from F2F and NT are nearly the same as the original image. Therefore, intra-distances between positive samples are larger than their inter-distances.

To learn a more generalized feature embedding, a triplet loss is added to architecture. Illustration of the triplet process can be found in Figure 4.3. It is introduced in [141, 159] and used in various face related applications. We aim to improve the generalization ability by penalizing the triplet relationships among batches. The triplet loss is defined by

$$L_{triplet} = \|g(x_i^a) - g(x_i^p)\|_2^2 - \|g(x_i^a) - g(x_i^n)\|_2^2 + \alpha. \quad (4.3)$$

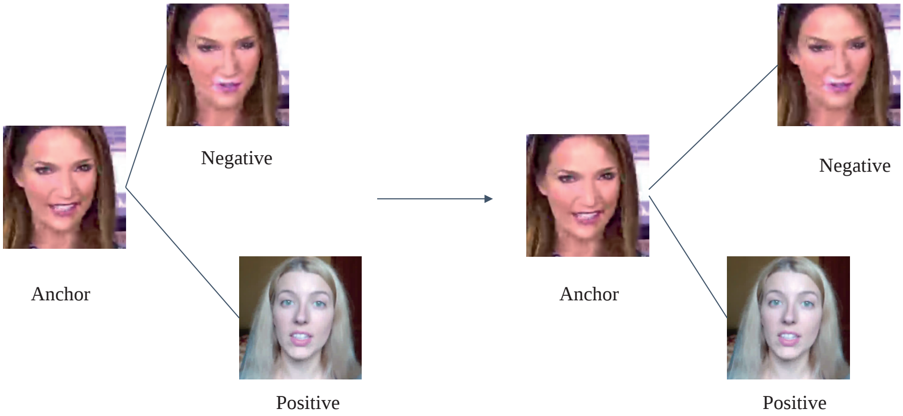


Figure 4.3: Visualization of the triplet loss. Images are from FaceForensics++. Normally, the pristine and manipulated images from the same subject look similar. Through the triplet loss, we attempt to minimize the distance between positive examples while maximizing the distance between positive and negative examples.

where α denotes the margin and i represents the batch index. x_i^a, x_i^p and x_i^n are anchor, positive samples, and negative samples respectively. They are selected online in each batch. g is the embedding learned from the network. The triplet loss can force the network to minimize the distance between an anchor and a positive sample and maximize the distance between the anchor and a negative sample. It can also contribute to higher robustness when the input is an unseen facial attribute or identity.

4.3.4 Center loss

Different modification methods may select different face regions for manipulation (see Figure 4.1). When this region is very small compared to the entire image, the majority of the features may exclude information about the manipulation. Our aim is that the network focuses on influential regions around faces instead of the background. To extract more discriminative embeddings, the center loss is used. It has been applied to face recognition [197], and proven effective in measuring intra-class variations. The center loss is defined by:

$$L_{center} = \sum_{k=1}^{M} ||\theta_k - c_k||^2, \tag{4.4}$$

where θ is extracted from feature maps by global average pooling. c denotes the center of feature. Theoretically, the feature center needs to be calculated based on the entire dataset. From [82], a more practical way is used to iteratively update the feature center:

$$c_{k+1} = c_k + \gamma(\theta_k - c_k). \tag{4.5}$$

where γ defines the learning rate of the feature center $c_k \in R^{N \times S}$. k denotes the iteration index, N is the batch size, and S is the dimension of the embedding. This iterative learning process provides a more smooth prediction for the feature center.

Our final loss function is given by:

$$L = L_{cls} + \lambda_1 L_{mmd} + \lambda_2 L_{triplet} + \lambda_3 L_{center}. \tag{4.6}$$

where λ_1, λ_2 and λ_3 are balancing factors. L_{cls} is a cross-entropy loss for binary classification.

4.4 Experiments

4.4.1 Implementation details

Our implementation is based on TensorFlow. Adam is used for optimization. We use Dlib to detect the face bounding boxes for each frame of the videos. The cropped face image is 300 x 300. All images contain a single face. For all the experiments, we follow subject-exclusive protocol meaning that each subject exists in one split of the dataset. Inception v4 [175] is used as the backbone architecture. The pre-trained model on Imagenet is used. The batch size is 16. Learning rate is 10^{-5}. The training set has a balanced distribution of real and fake data. λ_1 , λ_2, and λ_3 are set to 0.1, 0.05, 10. For the MMD loss, we use a Gaussian kernel function. The kernel bandwidths σ are 1 and 10. Feature maps $Mixed3a$, $Mixed4a$, $Mixed5a$ from the Inception net are used to calculate the MMD loss. For the triplet loss, we use the implementation of [159]. The margin is 2. Triplets are generated online. For every batch, we select hard positive/negative examples. For center loss, γ is 0.095. The dimension for embedding is 1024. The center is randomly initialized. For the experiments on FaceForensics++, our settings are aligned with [152]. As for experiments on Celeb-DF, our settings follow [115].

4.4.2 Evaluation results

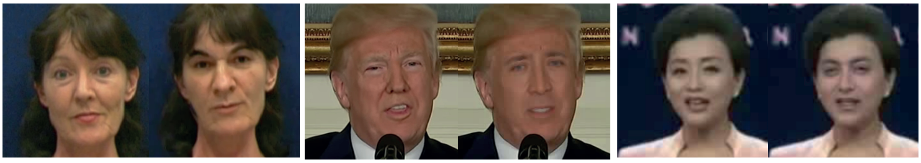
We evaluate our approach on several deep fake datasets. A summary of the datasets is shown in Table 4.1. Visualization of the data samples for each dataset are given in Figure 4.1 and 4.4.

Results on UADFV and DF-TIMIT

Both UADFV [111] and DF-TIMIT [99] are generated by identity swap methods. UADFV has 49 manipulated videos of celebrities obtained from the Internet. DF-TIMIT is created based on VidTIMIT [157] under constrained settings. DF-TIMIT has two different settings: low- and high-quality. We choose to evaluate our method on the high-quality subset because it is more challenging. Some pristine videos of the same subjects are selected from VidTIMIT to compose a balanced training dataset. We compare our method with Mesonet [2], XceptionNet [27], Capsule [133], and DSP-FWA [115]. We provide a brief introduction of each method below. Mesonet is based

Dataset	UADFV [111]	DF-TIMIT [99]	FaceForensics++	Celeb-DF [115]
Number of videos	98	300	5000	6000
Number of frames	34k	68k	2500k	2341k
Method	DF	FS	FS, DF, F2F, NT	DF

Table 4.1: Main contrasts of several deep fake datasets. “DF”: “DeepFakes”; “NT”: “Neural Textures”; “FS”: “Face Swap”; “F2F”: “Face2Face”. The “deep fakes” is an overarching name representing a collection of these methods. For each dataset, the manipulation algorithm and process can be different.



DFTIMIT UADFV Celeb-DF-v2

Figure 4.4: Visualization of data samples from DF-TIMIT (high quality), UADFV and Celeb-DF. For each pair of images, the left one is the real image and the right one is the modified image.

on Inception Net, which is also used in our architecture. XceptionNet is a deep network with separable convolutions and skip connections. Capsule uses a VGG network with capsule structures. DSP-FWA combines a spatial pyramid pooling with FWA [111]. In Table 4.2, we report all the performances following the same setting as in [115]. Two datasets are relatively small and not very challenging for forensic detection.

Results on Celeb-DF

Celeb-DF [115] is one of the largest deep fake video datasets. It is composed of more than 5,000 manipulated videos taken from celebrities. Data is collected from publicly available YouTube videos. The videos include a large range of variations such as face sizes, head poses, backgrounds and illuminants. In addition, subjects show large variations in gender, age, and ethnicity. The generation process of fake faces focuses on reducing the visual artifacts and providing a high-quality synthetic dataset. Table 4.2 shows that the area under curve (AUC) score of the proposed method on Celeb-DF outperforms all other approaches. The experimental settings remain the same as [115]. Compared to other deep fake datasets, Celeb-DF has fewer artifacts and better quality. The majority of failure cases are false positives. Typical false positives are shown in Figure 4.5. Most cases have relatively large poses.

AUC	UADFV	DF-TIMIT	FF-DF	Celeb-DF
MesoNet [2]	82.1	62.7	83.1	54.8
Xception [27]	83.6	70.5	93.7	65.5
Capsule [133]	61.3	74.4	96.6	57.5
DSP-FWA[115]	97.7	99.7	93.0	64.6
Ours	**98.1**	**99.8**	**97.2**	**88.3**

Table 4.2: Evaluation on UADFV [111], DF-TIMIT [99], FF-DF, Celeb-DF [115]. Each datset is evaluated separately. The metric is the Area Under Curve (AUC) score. "FF-DF" is the deep fake subest from FaceForensics++. We follow the same setting of [115].

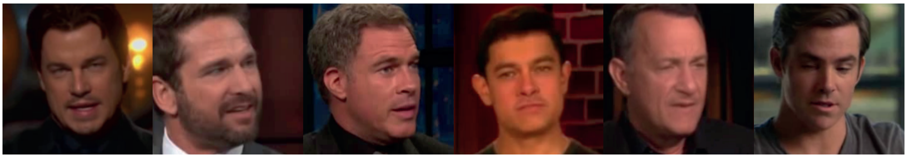


Figure 4.5: Visualization of false negative predictions of Celeb-DF from our method. These cropped images are from frames of deep fake videos.

Results on FaceForensics++

FaceForensics++ [151, 152] is one of the largest face forgery dataset. It includes pristine and synthetic videos manipulated by Face2Face [180], NeuralTextures [181], Deepfakes and Faceswap. The modified videos are generated from a pair of pristine videos. In Figure 4.1, we plot all pristine and manipulated examples from one subject within the same frame. Even though the pair of source and target are the same, different methods lead to different results. The performance on raw and high-quality images from FaceForensics++ are already good (accuracy exceeding 95%); we therefore focus on the performance on low quality images. For all experiments, we follow the same protocol as in [152] to split the dataset into a fixed training, validation, and test set, consisting of 720, 140, and 140 videos respectively. All the evaluations are based on the test set.

We compare our method with Mesonet [2], XceptionNet [27], and other methods [8, 147]. In Table 4.3, we report the performance while training all the categories together with our pipeline. The total f1 score is 0.89. In Table 4.4, we show the performance while training each category separately. In general, a more balanced prediction is obtained among the pristine and generated examples. The overall performance is better than the other methods. As expected, training FaceForensics++ separately (Table 4.4) results in a better performance than combined training (Table 4.3). This is because each generation method is seen as a different domain to the rest. The modified face images contain different types of artifacts and features. When training entirely, manipulated faces from facial reenactment method is extremely similar to real faces. The forgery detector tends to confuse real faces with deep fake data. Our method successfully

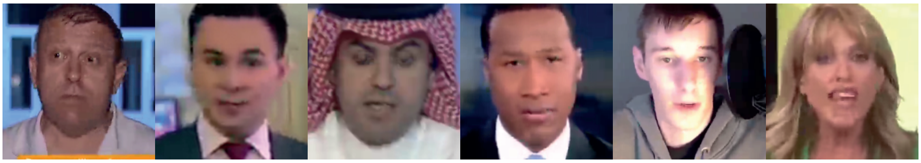


Figure 4.6: Visualization of false negative predictions of FaceForensics++ for the proposed method. These cropped images are frames taken from the deep fake videos.

Accuracy	DF	F2F	FS	NT	Real	Total
Rahmouni et al [147]	80.4	62.0	60.0	60.0	56.8	61.2
Bayar and Stamm [8]	86.9	83.7	74.3	74.4	53.9	66.8
MesoNet [2]	80.4	69.1	59.2	44.8	77.6	70.5
XceptionNet [27]	93.4	**88.1**	**87.4**	78.1	75.3	81.0
Ours	**98.8**	78.6	80.8	**97.4**	**89.5**	**89.7**

Table 4.3: Evaluation on the test set of FaceForensics++. The training and test set includes all the categories of manipulated dataset. “DF”: “DeepFakes”, “NT”: “Neural Textures”, “FS”: “Face Swap”, “F2F”: “Face2Face”.

improves the performance on pristine faces without impairing the performance on each deep fake category. When training each category separately, the main challenge becomes the image variations like blur. A number of false negative predictions are shown in Figure 4.6. In general, the performance degrades significantly when the face is blurry or the modification region is relatively tiny.

4.4.3 Analysis

Performance on a single type of deep fake dataset is better than on a dataset containing multiple domains. This is because the extracted features for different manipulated results are diverse. In general, face reenactment may have fewer artifacts than identity

Accuracy	DF	F2F	FS	NT
Bayar and Stamm [8]	81.0	77.3	76.8	72.4
Rahmouni et al [147]	73.3	62.3	67.1	62.6
MesoNet [2]	89.5	84.4	83.6	75.8
XceptionNet [27]	94.3	**91.6**	93.7	82.1
Ours	**99.2**	89.8	**94.5**	**97.3**

Table 4.4: Evaluation on each category of the FaceForensics++ test set. Each category has a balanced distribution between pristine data and fake data.

Method	Data augmentation	MMD	Center	Triplet	F1
Basic					0.826
Ours	✓				0.846
Ours	✓	✓			0.881
Ours	✓	✓	✓		0.889
Ours	✓	✓		✓	0.887
Ours	✓	✓	✓	✓	0.897

Table 4.5: Performance comparison with different components of our method. We evaluate our method on test set of FaceForensics++. Metric is F1 score.

Style Loss	GRAM	BN	MMD
F1	0.862	0.887	0.897

Table 4.6: Performance comparison with different style transfer losses. We evaluate our method on test set of FaceForensics++.

modification methods because the transfer of the expressions may require less facial alternations. It results in better performance on detecting identity modification results. We further calculate the prediction accuracy based on each video in the test set of FaceForensics++. On average, the prediction for pristine and fake videos is higher than 80%. Although most of the datasets have many frames, the number of videos is relatively small. In most videos, faces have a limited range of variations like pose, illumination, or occlusion. This can also cause the network to predict negatives when pristine face images are relatively blurry or partially occluded. Also, the number of different subjects for the deep fake dataset is relatively small compared to other face-related datasets. This leads to biased results when testing an unseen identity with unique facial attributes.

In our framework, we combine several losses to jointly supervise the learning process of the network. MMD loss can be considered as aligning the distributions of different domains. The style of each image can be expressed by feature distributions in different layers of deep network. Network is constrained to learn a more discriminative feature embedding through different domains. Center loss forces network to concentrate on more influential features rather than background noise. Triplet loss can be considered as an additional constraint to reduce the intra-distance effectively among positive examples.

4.4.4 Ablation study

To investigate the effect of each component of our method, we evaluate the performance of the proposed method with different components, see Table 4.5. We start with the baseline architecture and add different components separately.

Kernel Function	Polynomial	Linear	Gaussian
F1	0.841	0.876	0.897

Table 4.7: Performance comparison with different kernel functions in the MMD loss. We evaluate our method on the test set of FaceForensics++.

Comparison with other style transfer losses

First, we test our method with other neural style transfer losses for distribution alignment. Here, we choose the GRAM matrix-based style loss and batch normalization (BN) statistics matching [112]. The GRAM-based loss is defined by

$$L_{GRAM} = \frac{1}{W_l} \sum \left(G_R^l - G_F^l \right)^2 , \tag{4.7}$$

where the Gram matrix G^l is the inner product between the vectorized feature maps in layer l. G_R and G_F are GRAM matrix for real and fake samples respectively. W_l is the normalization term.

The BN style loss is described by

$$L_{BN} = \frac{1}{W_l} \sum \left[\left(\mu_{F_l} - \mu_{R_l} \right)^2 + \left(\sigma_{F_l} - \sigma_{R_l} \right)^2 \right] , \tag{4.8}$$

where μ and σ is the mean and standard deviation of the vectorized feature maps. μ_{R_l} and σ_{R_l} are corresponding to real face samples. From Table 4.6, performance of the network with the MMD loss outperforms other types of losses.

Comparison with different kernel functions

A different kernel function k can provide different mapping spaces for the MMD loss. In Table 4.7, we investigate the effect of different kernel functions. Linear and polynomial kernel functions are defined as $k(a, b) = a^T b + c$, $k(a, b) = (a^T b + c)^d$, respectively. We choose $d = 2$ for polynomial kernel function. The Gaussian kernel outperforms other kernels for the MMD loss.

Comparison with different feature maps

Different levels of feature maps capture different type of style information. We further examine how different combinations of feature maps influence the face forensic detection performance. In Table 4.8, we illustrate the performances of using multiple sets of feature maps. Feature maps $Mixed\,3a$, $Mixed\,4a$, $Mixed\,5a$ are slightly better than other options.

4.5 Conclusions

This chapter focused on face forgery detection and proposed a deep network based architecture. Maximum mean discrepancy (MMD) loss has been used to learn a more

Feature map	F1
$Mixed\,3a$	0.884
$Mixed\,4a$	0.881
$Mixed\,5a$	0.883
$Mixed\,3a, Mixed\,4a$	0.890
$Mixed\,4a, Mixed\,5a$	0.891
$Mixed\,3a, Mixed\,4a, Mixed\,5a$	0.897

Table 4.8: Performance comparison with different combinations of feature maps from our method. We evaluate our method on the test set of FaceForensics++.

generalized feature space for multiple domains of manipulation results. Furthermore, triplet constraint and center loss have been integrated to reduce the intra-distance and to provide a discriminative embedding for forensics detection.

Our proposed method achieved the best overall performance on UADFV, DF-TIMIT, Celeb-DF and FaceForensics++. Moreover, we provided a detailed analysis of each component in our framework and exploited other distribution alignment methods. Extensive experiments showed that our algorithm has high capacity and accuracy in detecting face forensics.

5

Deep Imbalanced Learning for Age Estimation from Videos

5.1 Introduction

Datasets for face related research often exhibit highly-skewed class distributions. Most samples belong to only a few majority classes, while the minority classes include fewer cases [84]. In Figure 5.1, age distributions are shown for the three largest face video datasets for age estimation. Hence, age estimation datasets are often highly imbalanced and with long-tailed distributions. When a model is trained on a dataset containing a strongly imbalanced distribution, the model may extract features which are biased towards the majority classes. For example, age related features of children may be different than those computed from elderly people. Thus, a model trained mainly on teenagers may consider senior people as outliers. Although the minority classes constitute a small portion of the whole dataset, they play a predominant role in the prediction error. In Table 5.3, most methods show high accuracy for the majority groups. The performance for senior minority groups is nearly random. Moreover, subtle discrepancies between classes are ignored by the deep learning networks. The set of features extracted from the minority classes is insufficiently represented. Hence, the generalization capabilities are restricted when facing a large span of age distributions.

Aging is a natural and irreversible process that causes both appearance and geometrical variations on human faces [5]. Both personal (e.g., physical/mental states) and situational factors (e.g., occupation, living environment) can influence the facial aging process. Age estimation from human faces has various real-world applications in, for example, human-computer interaction, social network application, and surveillance monitoring. Age estimation tasks can be classified into two categories: biological and apparent age estimation. Traditional age estimation methods for still images widely employ biologically inspired features [66, 103]. Recently, more and more deep network based methods have been applied in age estimation [24, 89]. The methods have achieved notable progress in image-based age estimation. However, their performances on video-based age estimation are still far from ideal. Face age estimation using videos has the possibility to incorporate both the facial structure and its dynamics. Dynamic information from video sequences is crucial to handle challenging imaging variations like facial occlusion and extreme face poses. Age estimation from unconstrained video sequences have largely been ignored so far.

Therefore, in this paper, we addresses the problem of age imbalance in videos from a transfer learning perspective. A deep clustering module is proposed to both learn a proper data representation and transfer information from the majority groups. In fact, the deep clustering module is used to leverage knowledge of majority age groups to improve the performance on minority ones. The scarcity of data leads to insufficient variations for the system to learn. Given the relevance between age labels, the majority age groups is the proper complementary source for minority classes. To mitigate the impact of imbalanced data, the target distribution is constructed by a linear combination of age predictions. Compared to previous label distribution learning methods for age estimation [139], the proposed method doesn't incorporate hard constraints on the margin of the age distribution. In addition, temporal ensemble and consistency constraints are added to further improve the smoothness of the annealing process.

Our main contributions are:

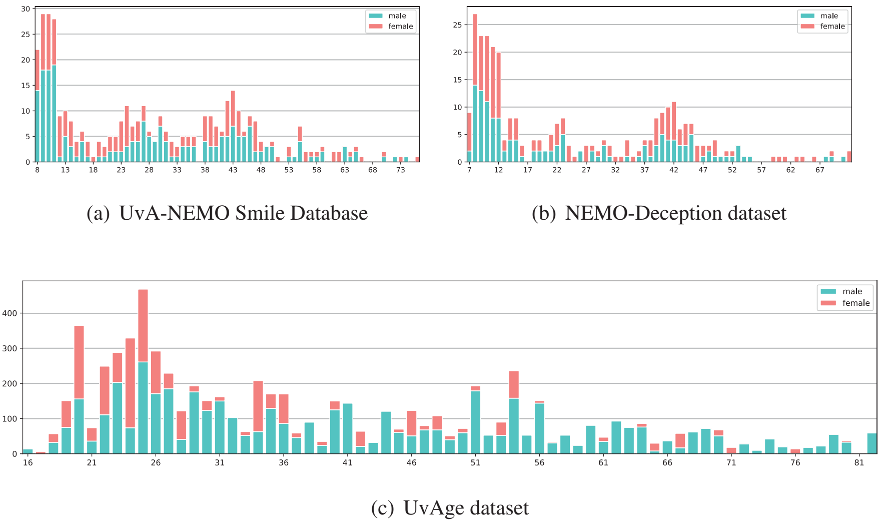
- We propose an end-to-end framework to predict ages from face videos. Clustering based transfer learning is used to provide proper prediction for imbalanced datasets.
- We investigate the influence of different variations on age estimation. We propose a new video dataset NEMO-Deception.
- Detailed experiments on age estimation and imbalanced learning are provided to show the effectiveness of our method. The proposed method achieves the best performance on three face age estimation datasets (UvA-NEMO Smile Database, NEMO-Deception, and UvAge dataset).

5.2 Related Work

5.2.1 Age estimation

Detailed reviews about age estimation can be found in [51, 140, 148]. Most existing age estimation methods focus on still images. They can be classified into two categories: classification and regression. Regression based methods ar the most straightforward approaches to predict age [116]. Classification based approaches consider age or age groups as separate classes [105]. Compared to the standard classification task, human age information is ordinal. Label distribution learning (LDL) based approaches are capable of predicting a range with different confidence levels [55, 56, 79, 206]. [139] introduces the mean-variance loss to provide concentrated distributions for age estimation.

Most of the work focuses on image-based age estimation. Only a few methods use videos for age estimation. Traditional video-based methods usually pre-design hand-crafted features. Hadid and Abdenour [67] extract local binary features for age estimation from videos. Dibeklioglu et al. [40, 42] use face dynamic information to improve the performance of age prediction. Pei et al. [145] propose an attention based network to predict age and disgust expression in videos. Previous approaches are



(a) UvA-NEMO Smile Database

(b) NEMO-Deception dataset

(c) UvAge dataset

Figure 5.1: Age distributions of UvA-NEMO Smile Database, NEMO-Deception dataset, and UvAge dataset. The horizontal axis denotes the age. The vertical axis is the number of subjects. Green bars represent males and red bars represent females.

established based on restrictions of, for example, faces with a moderately pose changes throughout a video. They are not considered in the context of unconstrained data.

5.2.2 Deep imbalanced learning

Previous papers on imbalanced data distribution can be divided into two categories: re-sampling [15, 54, 161, 227] (including over- and under-sampling) and cost-sensitive learning [83, 97, 182, 224]. Over-sampling adds samples repeatedly from minor classes. Novel data samples are generated by interpolation or synthetic data [22]. Based on the effective number of samples for each class, Cui et al. [33] propose a re-weighting mechanism to re-balance the loss function. Class rectification loss [44] searches hard minority classes among every batch and adds regularization in feature space to rectify the learning bias. Huang et al. [84] propose a cluster-based local embedding to the improve performance of face recognition and attribute prediction. Wang et al. [195] employ a meta network to regress network weights between different classes. [220] proposes a Bilateral-Branch Network (BBN) using a novel cumulative learning scheme to jointly compute both representation and classifier learning. [176] uses causal intervention and counterfactual reasoning to select the proper momentum causal effect. [17] uses label distributions based on the margin loss (LDAM), and a deferred re-weighting (DRW) schedule for training imbalanced datasets. [199] re-balances the weights to alleviate the influence of label co-occurrence and uses a negative regularization to reduce the over-suppression of negative labels. These class-balanced approaches enable the minority classes to play a re-weighted role in determining the decision boundaries of the models. Oversampling may cause over-fitting by emphasising on minority samples. Down-sampling discards many majority samples but may fail to exploit useful feature variations. Some methods use a cost sensitive learning scheme which is applied to image based age estimation [213, 226]. However, the challenge is to determine the actual cost for different samples in various distributions. In contrast to previous approaches, our method relies on clustering based transfer learning to fully exploit the shared features between majority and minority samples.

Methods are proposed to approach the data imbalance including transferring the information learned from major to minor classes [32, 137]. Yin et al. [210] transfer intra-class variances from the head to tail for face recognition tasks. Liu et al. [118] add a memory module to the neural networks to transfer semantic features. Compared to these methods, our method doesn't rely on large scale dataset pre-training or a complex training scheme.

5.2.3 Joint clustering and representation learning

Clustering-based representation learning [201] shows great potential in unsupervised and semi-supervised learning (SSL) tasks. It can simultaneously cluster the data samples and learn a proper data representation. π-Model [104] incorporates the consistency regularization between the prediction for an unlabeled instance and its stochastic perturbation sample. Mean teacher [178] models further improves the target distribution for unlabeled instances by an exponential moving average (EMA) of parameters from previous training information. [121] applies deep temporal clustering on unsupervised time-domain tasks. [74] transfers features from the known classes to improve the quality of newly found categories in the unlabelled dataset. Although deep clustering methods are able to extract proper representations from large-scale unlabeled datasets, the alternating update scheme of feature clustering may lead to instabilities in the training process.

5.3 Method

We propose a novel end-to-end deep framework for video based age estimation. We aim to address two main limitations of previous work: 1) methods are negatively influenced by imbalanced age distributions. To provide a balanced prediction, soft label assignment is used to construct the target distribution. Each age is assigned a different degree of contribution. This avoids hard constraints on target labels, leading to better solutions of the annealing process of clustering. 2) different impacts of variations (pose, expression etc.) are considered for age estimation, which are typically neglected in traditional methods. An overview of our pipeline is shown in Figure. 5.2.

5.3.1 Clustering module

Our transfer learning module is based on a deep embedded clustering (DEC) algorithm that clusters the data while jointly learning a proper data representation. Data clustering in our task is that for a video sequence $I = \{x_i\}, i = 1, ..., M$, the goal is to produce

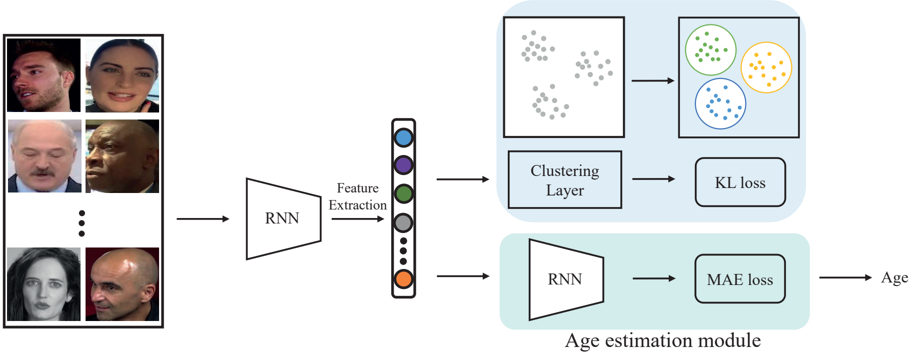


Figure 5.2: Overview of our pipeline. Input is a video sequence from a face dataset. The RNN for extracting features is LRCN. It allows for recognizing temporal dynamics in sequential inputs. “MAE loss” represents mean the absolute error. “KL loss” means the Kullback–Leibler divergence loss.

output class assignments $\{y_i\}, i \in 1, ..., K$. Effective latent representation is a crucial step in the clustering process. We achieve this by using a Long-term Recurrent Convolutional Network (LRCN) [43] architecture. It facilitates the learning of the temporal information of video sequences resulting in compressing videos into a compact latent space. The latent representations of videos are assigned to clusters of the clustering layer.

Clusters are composed of a set of vectors or prototypes, where $C = \{c_k\}, k = 1, ..., K$ represent the cluster “centers”. The aim is to determine the clusters and learn the data representations simultaneously. p_{ij} is the probability of assigning data sample $i \in 1, ..., N$ to cluster $j \in 1, ..., K$. Integrating clustering into representation learning is a challenging task. Following [188], the Student’s t-distribution is used as the kernel to measure the similarity between feature embedding z_i and center c_j. The probability assignment of the latent feature belonging to k th cluster is as follows:

$$q_{ij} = \frac{\left(1 + \frac{sim(z_i, c_j)}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\sum_{j=1}^{k}\left(1 + \frac{sim(z_i, c_j)}{\alpha}\right)^{-\frac{\alpha+1}{2}}}, \tag{5.1}$$

sim is the similarity function used to compute the distance between feature z_i and each centroid c_j. The Euclidean distance is used as the similarity metric. q_{ij} denotes the probability of input i belonging to cluster j, z_i corresponds to the input in the feature space Z, obtained from the recurrent module after encoding the input signal $x_i \in X$. Instead of maximizing the likelihood of model q directly, the model is matched to distribution p by minimizing the Kullback–Leibler (KL) divergence between the joint distribution of q_{ij} and p_{ij}.

$$L_{cluster} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij} log \frac{p_{ij}}{q_{ij}}. \tag{5.2}$$

Soft label assignment

In previous papers [188, 201], the target distribution p_{ij} is often constructed by first raising q_{ij} to the second power, and then normalizing it by the frequency. In this way, the initial classifier's high confidence predictions are normally corrected. However, for age estimation, a weighted average of the predictions is more suitable than a single value. Given that age labels are ordinal and related to each other, one can be represented by a linear combination of the other labels. The target distribution is constructed by:

$$P = F * Q + B. \tag{5.3}$$

P and Q are matrices for p_{ij} and q_{ij}, respectively. F denotes the global structure of labels. Each part in the correlation vector represents the impact of each cluster on this instance. B is used for regularization between target and ground truth distributions. The constructed distribution provides more information and assigns a new relevance of a label to a particular instance based on global label correlations.

Temporal ensemble and consistency constraint

In order to slowly anneal clusters to learn a proper partition of the data, following [104, 178], we use temporal ensembling to improve the smoothness of the annealing process. The clustering model q computed at different epochs are aggregated by maintaining an exponentially moving average of the predictions within multiple previous training epochs. Hence, network predictions q are added to an ensemble prediction Q via

$$Q^t = \phi \cdot Q^{t-1} + (1-\phi) \cdot q_{ij}^{t-1}. \tag{5.4}$$

where ϕ is a factor to control the combinations of ensembles and the learning history information, and t indicates the iteration step. Mean squared error (MSE) is used as the consistency cost. The loss function in Eq. 5.2 now becomes

$$\begin{aligned} L_{cluster} =& \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij} log \frac{p_{ij}}{q_{ij}} \\ &+ \theta(t)\frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K} \|q_{ij} - q_{ij}^{'}\|^2. \end{aligned} \tag{5.5}$$

$q_{ij}^{'}$ represents the prediction of temporal ensembles. To initialize the process of the EMA, Q_t is re-scaled as $q_{ij}^t = \frac{1}{1-\beta^t} Q^t$. $\theta(t)$ is set to gradually increase the weights of constraints from 0 to 1. A consistency constraint enforces the predictions to be more stable among a large range of variations.

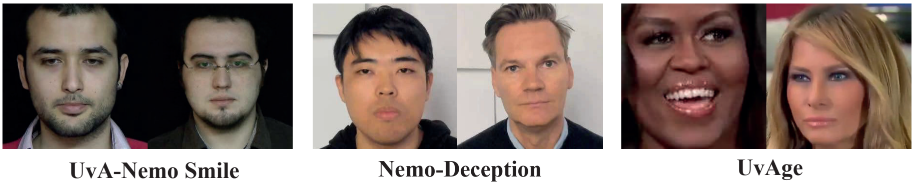


Figure 5.3: Data samples from UvA-NEMO Smile Database, NEMO-Deception, and UvAge dataset.

5.3.2 Age estimation module

We can further predict ages based on the learned embedding. For general softmax loss, given one positive class, all the negative classes are treated similarly when updating the network. Therefore, we choose a regression based model. Followed by the learned embedding, fully connected layers with ReLUs are used. Dropout is employed for regularization. Age estimation is supervised by the MAE loss function. MAE is defined as the mean absolute error between the predicted age and ground-truth age:

$$L_{age} = \frac{1}{N} \sum_{i=1}^{N} \left(y_{pred} - y_{true}\right)^2 , \tag{5.6}$$

y_{pred} and y_{true} are the prediction and ground truth label, respectively. The final loss function is as follows:

$$L = \lambda_1 L_{cluster} + \lambda_2 L_{age}. \tag{5.7}$$

λ_1, λ_2 are balancing factors.

5.4 Experiments

5.4.1 Implementation details

Dlib face detector is used to obtain the face bounding box for each frame. The side length of the bounding boxes is extended by 20% before inputting it to the network. Our pipeline is based on TensorFlow and Keras. All the experiments are based on subject exclusive protocol. The input image size is 224 * 224. Batch size is 8. Sequence length is 100. Ground truth of all the datasets are based on the biological age. For the clustering layer, α is set to 1. Balancing factors λ_1, λ_2 are set to be 2 and 0.1, respectively. We follow the official protocol to split each dataset for training and testing. Metrics for all the experiments are MAE (mean absolute error) and cumulative score (CS). CS measures the performance of age prediction accuracy based on a tolerance of absolute error.

5.4.2 Dataset

We evaluate the proposed method on three dataset: UvA-NEMO Smile Database [41], NEMO-Deception, and UvAge. Basic information about the three datasets are shown in

Dataset	Subjects	Videos	Age range
UvA-NEMO Smile	400	1240	8-76
NEMO-Deception	309	309	7-72
UvAge	516	6983	16-83

Table 5.1: Comparison between three age dataset

Table 5.1. The age distributions are presented in Figure 5.1. Visualizations of the data samples for each dataset are given in Figure 5.3.

UvA-NEMO Smile Database

UvA-NEMO Smile Database [41] has been collected to analyze the temporal dynamics of spontaneous/posed smiles for different ages. Videos were collected under constrained settings. Participants comprising the UvA-NEMO dataset contain relatively low inter-subject variations.

NEMO-Deception

NEMO-Deception dataset is created recently to analyze deceptive behavior manifested on human faces. Information (age, gender, and kin relationship) of participants is collected during the deception experiments. Participants are divided into different groups based on their family relationships. Participants in each group took turns to take the experiment as a test subject. Each session is recorded under constrained settings. The entire experiment was recorded by a web camera which is connected to a computer and an iPhone XS. The web camera records video information together with audio. The video has a resolution of 1920×1080 pixels at a speed of 60 frames per second. The audio codec format is MPEG-4 AAC with stereo channel, 48000 hz of sample rate and 320 kbps of Bit-rate. The recorded videos were manually labeled and split into separate clips according to the interviewing and answering information.

UvAge dataset

Today, UvAge is the largest video based face dataset with age annotation. It is composed of unconstrained video samples containing a large range of variations in face size, pose, illumination, and ethnicity. Subjects are celebrities. We split the dataset into 396, 65, and 55 subjects for training, validation, and testing respectively.

5.4.3 Age estimation

Age estimation experiments are conducted on UvA-NEMO Smile Database, NEMO-Deception, and UvAge, respectively. We compare our approach on age estimation task with the following methods:

Method	UvA-NEMO	UvA-Deception	UvAge
LSTM	5.80	6.81	7.35
GRU	5.47	7.31	6.85
LRCN [43]	5.51	7.26	7.56
DEX [153]	5.48	7.29	7.58
Focal Loss [117]	5.13	6.18	6.57
CBL-CE [33]	4.94	5.79	6.43
SSR-Net [205]	4.91	5.71	6.39
CBL-FL [33]	4.82	5.67	6.21
MV [139]	4.8	5.65	6.14
SIAM [145]	4.74	5.34	5.83
DB [199]	4.58	5.17	5.7
Ours	4.12	4.66	4.91

Table 5.2: Evaluations on multiple video datasets for face age estimation. Each dataset is trained and evaluated separately. “CBL-FL” and “CBL-CE” denote the class balanced focal loss and cross entropy loss, respectively.

Video-based age estimation methods

Since there are only a few methods on video-based age estimation, our baselines correspond to common recurrent architectures: LSTM, GRU and LRCN [43]. The input for LSTM and GRU are extracted from the pre-trained VGG16. Loss function is cross entropy. Focal loss [117] can be seen as a smoother version of hard example mining. It penalizes the examples from minority classes more than those of majority classes if the network is biased towards the majority classes during the training process. We apply the focal loss on LRCN. Moreover, we also consider the effect of Spatially-Indexed Attention Model (SIAM) [145]. SIAM is the latest end-to-end method for video-based age estimation. Both temporal and spatial attention modules are integrated into the model.

Image-based age estimation methods

Deep Expectation (DEX) [153] interprets age estimation as a classification problem followed by refinement module. Mean-Variance (MV) [139] incorporates the mean and variance loss to conduct distribution learning for age estimation. We modify the original architecture to handle video sequences. Stagewise Regression Network (SSR-Net) [205] uses a light-weighted model to avoid large amount of neurons. For SSR-Net, we calculate the average of age predictions using the frames from each video.

State-of-the-art deep imbalanced learning methods

Class Balance Loss (CBL) [33] re-weights the original losses by class-wise weights, which are based on effective numbers. We use the class-balanced cross-entropy loss and the focal loss. Distribution-Balanced Loss (DB) [199] re-balances the weights to

Group	LSTM	CBL-FL	MV	SIAM	Ours
20-29	5.65	6.92	3.89	3.95	3.16
30-39	8.47	7.83	5.68	4.91	4.09
40-49	8.80	10.92	9.14	8.95	8.90
50-59	10.12	9.23	10.08	10.37	8.71
60-69	12.57	14.92	12.39	11.08	9.69
70-79	14.85	13.50	13.58	11.68	9.67
80-89	31.26	25.15	26.93	23.34	16.26

Table 5.3: Mean absolute error of several age estimation methods on different age groups. Experiments are based on the UvAge dataset.

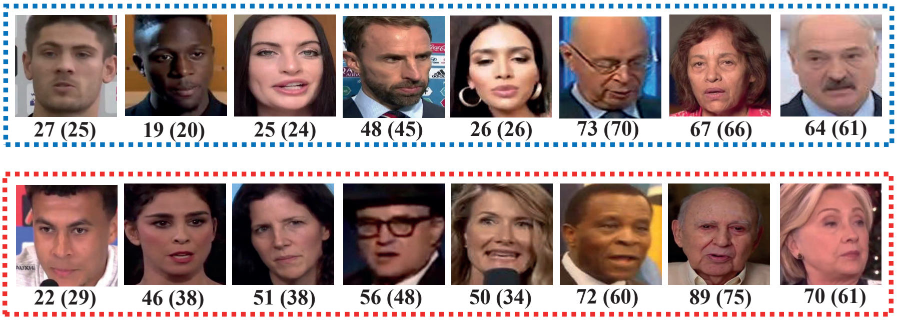


Figure 5.4: Visualization of the age estimation results produced by the proposed method on the UvAge dataset. The first row shows a number of successful age estimation examples. The second row shows a number of poor ones. The numbers below each image provide the ground-truth age and our prediction, i.e., ground-truth label (estimated age).

mitigate the impact from label co-occurrence and use a negative regularization to reduce the over-suppression of negative labels. For these deep imbalanced methods, we apply them on LRCN to conduct the experiments.

Table 5.2 shows that our method outperforms other methods on all three datasets. We also report the performance of several typical methods in different age groups in Table 5.3. The proposed method achieves better performances on minority age groups.

5.4.4 Influence of variation on age estimation

Influence of expression variation

The Facial Action Coding System (FACS) is used to describe facial expressions by action units. We use [156] to detect the intensity of action units. The estimated intensity stays within the range of 0-5 where 5 represents the maximum level of expression. An example of the detected action units heatmap is shown in Figure 5.5. The correspondence of actions units are: Cheek raiser (AU6), Upper lip raiser (AU10), Lip corner puller

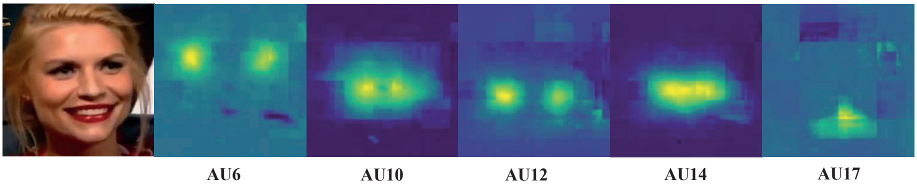


Figure 5.5: Visualization of detected action unit heatmap from [156].

Intensity	DEX [153]	MV [139]	SIAM [145]	Ours
1	7.83	6.41	6.02	5.93
2	8.34	6.25	6.16	5.79
3	8.81	7.66	7.53	7.25
4	9.11	6.96	6.63	6.44

Table 5.4: Mean absolute error of several age estimation methods applied on videos of faces with different intensity thresholds for the action units. The performance is obtained on different subsets of the UvAge dataset. If the intensity of one action unit is larger than the threshold, the corresponding face is included in the test set.

(AU12), Dimpler (AU14), and Chin raiser (AU17). The performance of our method versus others are compared for different intensity thresholds of action units in Table 5.4. Different subsets of data are selected for different thresholds. Overall, prediction become less accurate with expression intensity. Our method outperforms other methods for all levels of expression.

Influence of gender

To evaluate the effect of gender on the accuracy of the system, we also consider a gender-specific age estimation system. In the gender-specific system, the proposed method is trained and tested for both male and female subjects, respectively. The MAEs for both gender-specific and general training are given in Table 5.5.

Dataset	Male	Female	All
UvA-NEMO Smile	3.95	4.18	4.12
NEMO Deception	4.48	4.75	4.66
UvAge	4.86	5.08	4.91

Table 5.5: Comparison of the gender-specific method with the general method for age estimation. Experiments are conducted separately.

Yaw	CBL-FL	MV [139]	SIAM [145]	DB	Ours
30	6.81	6.13	5.87	5.72	5.08
45	6.84	6.34	5.9	5.76	5.14
60	6.85	6.41	5.96	5.83	5.22

Table 5.6: Mean absolute error of several age estimation methods on videos of faces with different yaw angle degrees. Experiments are based on the UvAge dataset. Each method is trained and evaluated separately.

Modules	M	K	K+S	K+S+C	M+K+S+C
UvAge	7.63	6.17	5.41	5.26	4.91

Table 5.7: Performance comparison with different components of our method on UvAge dataset. “M” means that the method uses the age estimation module only with the MAE loss. “K” means that the method contains the clustering layer only with the KL loss. “S” denotes soft label assignment. “C” represents temporal ensemble and consistency constraint.

Influence of pose variation

We evaluate the effect of head pose on the accuracy of age estimation. Since faces from UvA-NEMO Smile and NEMO-Deception are mainly frontal, we focus on the videos of the UvAge dataset. Using PRNet [48], the pose is computed for each video. We notice that yaw angles change predominantly. Therefore, videos are selected with relatively large pose variations (pitch and roll $\geq$ 30 degrees) as the test set, in comparison with other methods in Table 5.6. In general, the performance degrades with larger pose variation.

5.4.5 Ablation study

To investigate the effect of each component, we evaluate the performance of the proposed method with different settings, as shown in Table 5.7. We start with the baseline architecture and add different components separately.

Model effectiveness in mitigating data imbalance

In order to systematically investigate the effect of our model on imbalanced data, we remove the age estimation module and compare its relative performance improvement of other imbalanced data learning models. To do so, we change the imbalanced ratio for different age groups to quantitatively measure the performance on imbalanced data. Following [33], we define the imbalance ratio as the class size of the first head class divided by the size of the last tail class. We use class-imbalanced cumulative score (CS) as metric [49, 83]. Accuracy is calculated by $0.5(t_p/N_p + t_n/N_n)$, where N_p and N_n are the numbers of positive and negative samples, while t_p and t_n are the numbers of true positive and true negative. Performance comparisons can be found in Figure 5.6.

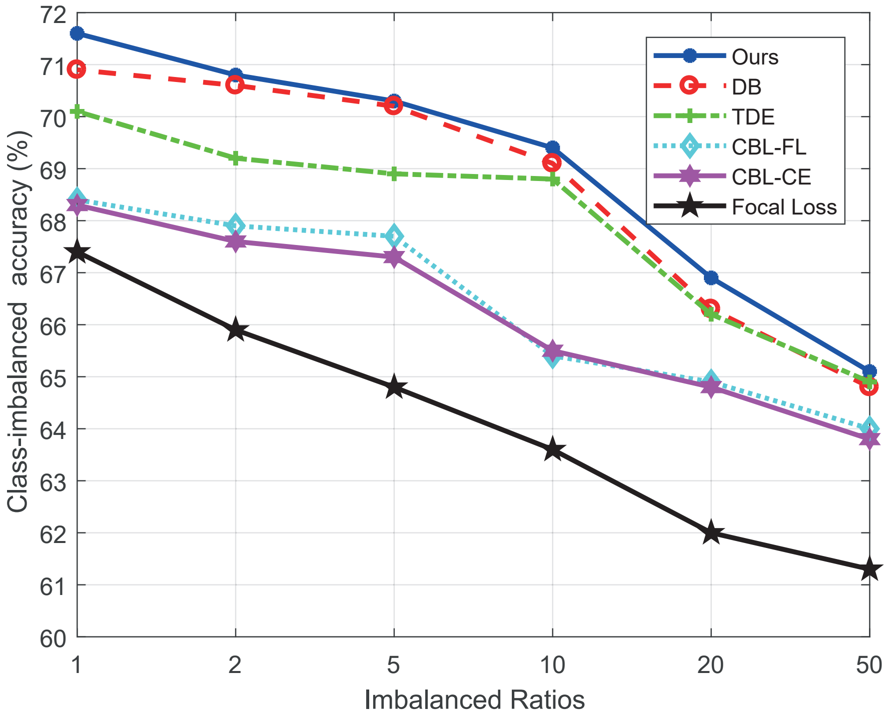


Figure 5.6: Class-imbalanced accuracy of several deep imbalanced learning on different imbalanced ratios. Experiments are based on the UvAge dataset. The class-imbalanced accuracy is calculated given a tolerance of absolute error less than 5 (θ=5).

Evaluation of different similarity metrics

The clustering layer plays a crucial role in our pipeline. To test the influence of different clustering metrics, we replace the default similarity function with other metrics, like Correlation based Similarity (COR) [60] and Auto Correlation based Similarity (ACF) [50]. COR is computed by $\sqrt{2(1-\rho)}$, where ρ denotes the pearson's correlation, given by

$$\rho_{x,y} = cov(x,y)/\left(\sigma_x \sigma_y\right). \tag{5.8}$$

where cov is the covariance. As for ACF, we use auto-correlation coefficients to compute the similarity between features and centers. Then, we compute the weighted Euclidean distance between the auto-correlation coefficients. The results are shown in Table 5.8. The default Euclidean distance is a better option for the similarity function.

Comparisons of different losses

To validate the effectiveness of our loss functions, we compare our setting with two widely used losses in age estimation (i.e., softmax loss and MAE loss) by performing

Metrics	COR [60]	ACF [50]	Ours
UvA-NEMO Smile	4.87	4.83	4.12
NEMO-Deception	4.95	4.96	4.66
UvAge	5.3	5.17	4.91

Table 5.8: Evaluation on different similarity metrics for clustering layer. Each dataset is trained and evaluated separately.

Losses	M	S	M+S	M+K
UvAge	6.57	6.25	5.4	4.91

Table 5.9: Evaluation on different combinations of losses with our method. "M" denotes the mean absolute error loss. "S" means the softmax loss. "K" represents the Kullback–Leibler loss.

age estimation on UvAge dataset. The results are shown in Table 5.9. It can be found that using a combination of loss functions leads to a better performance than using either of them individually. These results show that learning a meaningful distribution is better than single label learning for age estimation task.

5.4.6 Analysis

A number of successful and poor predictions of our method are shown in Figure 5.4. The network tends to generate younger predictions than their actual ages. In general, the inaccurate predictions are from two categories: feature variations and personal factors. Challenging variations like large changes in pose or expression can degrade the performance of our method. As for personal factors, whereas human faces share some common features across different ages, the aging process for each person may differ.

5.5 Conclusion

In this paper, we proposed an end-to-end method to predict age from videos. To resolve the imbalanced issue in existing datasets, we used a clustering layer to jointly cluster data and extract a better feature representation. This can leverage knowledge from related categories to improve performance of minority groups.

Through extensive experiments, we showed that our method substantially outperformed other methods on both constrained (UvA-Nemo, Nemo-Deception) and unconstrained datasets (UvAge). Moreover, the performance on minority age groups has also been improved.

6 Summary and Conclusion

This chapter concludes this dissertation by revisiting our research questions from Chapter 1, discussing our main findings, and sketching directions for future research. We focus on the main findings and general lessons. Additional detailed findings are in the conclusion sections of the individual chapters.

6.1 Summary

The individual conclusions for each chapter are presented as follow:

6.1.1 Chapter 2: Analysis for object features and face detection performance

In this chapter, we proposed an experimental comparison of main characteristics that influence face detection performance. A synthetic data generator is proposed to synthesize 2D faces based on 3D face models. We customized the synthetic dataset to address specific types of features (scale, pose, occlusion, blur, etc.). Then we select three representative face detectors to systematically investigate the influence of different features on face detection performance. Our results show that synthetic data can be a good complementary source for real datasets. The performance of face detectors can also be improved through various types of synthesized variations. Through our analyses, we also identified some potential deficiencies of the current face detection architectures. To conclude, there are often challenging features in real-world face detection. By providing an overview of the relationship between object features and face detection performances, we hope to assist researchers to choose more appropriate synthetic data when addressing challenging real-life variations.

6.1.2 Chapter 3: Pose invariant age estimation of face images in the wild

In this chapter, we focus on tackling the negative effect of head pose in age estimation tasks. First, we introduced the largest in-the-wild video dataset for age estimation. It contains unconstrained videos from celebrities in different events. To make age prediction more robust against pose variation, we reconstruct face uv textures from the original 2D frames of videos. Parts of the reconstructed uv textures are missing because of the self-occlusion effect of head pose. In our task, the size and shape of missing regions are highly irregular. We provide a Wasserstein GAN based approach (AgeGAN) to simultaneously estimate the real age and complete the partial uv textures. Our method can force the network to retrieve the missing regions with more meaningful features for age estimation. To demonstrate the effectiveness, we compare our face completion method with other advanced inpainting methods. We also systematically evaluate our age estimation method on other datasets.

6.1.3 Chapter 4: Discriminative learning for multi-domain face forgery detection

This chapter focused on the challenge of multi-domain face forgery detection. The major challenges of detecting face forgery are: 1) The difference between pristine and fake samples is much smaller than the difference among pristine examples. 2) The artifacts from imaging features and face characteristics do not persist across all generated results for the same generative method. We proposed an end-to-end deep network based architecture. Inspired from the applications of neural style transfer, we want our network to focus on more discriminative features instead of over-fitting to manipulation-specific artifacts. Maximum mean discrepancy can help us align features from different distributions. MMD loss is used to learn a more generalized feature space for multiple domains of manipulation results. Furthermore, triplet constraint is incorporated to minimize the intra-distances and maximize the inter-distances. Center loss has been integrated to provide a discriminative embedding for forensics detection.

Our proposed method achieved the best overall performance on UADFV, DF-TIMIT, Celeb-DF, and FaceForensics++. Moreover, we provided a detailed analysis of each component in our framework and considered the performance of other distribution alignment methods. Extensive experiments showed that our algorithm has high capacity and accuracy in forensically sound detection of deep fakes.

6.1.4 Chapter 5: Deep imbalanced learning for age estimation from videos

In this chapter, we aim at predicting ages from videos. Datasets for face related research often exhibit highly-skewed class distributions. Most data belong to only a few majority categories, while the minority classes include much fewer instances. we proposed an end-to-end method to predict age from videos. To resolve the imbalanced issue in existing datasets, we used a clustering module to jointly cluster data and extract a better feature representation. This can leverage knowledge from related categories to improve performance of minority groups. To mitigate the impact of imbalanced age samples in the training process, the target distribution is constructed by a linear combination of different age predictions. The proposed method does not incorporate hard constraints on the margin of the age distribution. In addition, temporal ensemble and consistency constraints are added to further improve the smoothness of the annealing process.

Through extensive experiments, we showed that our method substantially outperformed other methods on both constrained (UvA-Nemo, Nemo-Deception) and unconstrained datasets (UvAge). Moreover, the performance on minority age groups has also been improved.

6.2 General Discussion

In this section, we will discuss two main limitations existing in our work and possible future directions.

Sythetic data plays a crucial role in our Chapter 2. The variations from datasets have provided both opportunities and challenges for the deep learning system. The final performance of the face related applications are heavily relying on the variations of datasets. Synthetic data can be an effective way to enrich the variations of real datasets. The advantage of using synthetic data is that the we can have more understanding about how the performance of learning system is contingent on different types of synthesized variations. With all the benefits from synthetic data, admittedly, there is always an unavoidable domain gap between synthetic data and real data. Future work could focus on how to fill or reduce the gap.

In Chapter 3, we use GAN to recover the missing regions of face uv texture. Results from generative methods are becoming more and more realistic. They can provide additional variations for the learning system. However, it is difficult to fully control the generating process. Currently, we cannot fully explain the results of generative methods. Also, there is no common metrics to measure the performance of generative methods. For future research, we can investigate how to really control the behaviour of generative methods. It is also better if we can remove the correlations between some facial attributes like male and beard in the generation process. For most face modification methods, they have been widely used before being investigated about their social impact. User study would be helpful to understand how humans understand the synthesized faces.

The last decade has witnessed an explosive advancement of face studies and extended real-life applications. In contrast, research on their societal, especially ethical, implications has been largely overlooked. The accuracy and range of face recognition have grown drastically, and they sometimes become intrusive forces to people's daily life and even privacy domains. We suggest that research on face algorithms and their social implications should develop hand in hand, such that the face generative methods evolve in a way that is seen as credible and trustworthy in the eyes of the general public.

6.3 Conclusion

We conclude the thesis by revisiting the questions posed in the Chapter 1.

Question 1: Can we systematically manipulate variations in synthetic data to complement the real dataset and further achieve better performance on face detection?

The short answer is yes. The majority of data from existing datasets normally belongs to a limited range of variations. The faces did not sufficiently represent extreme poses, scale, or heavy occlusion, to train a robust detector against all potential variations. We use synthetic data as data augmentation to compensate the insufficient variation of real datasets. First, we design a novel synthetic face data generator with full controlled variations (on pose, scale, background, illumination, and occlusion). Compared to existing face manipulation methods, we have more control over the synthesized variations. Then we are able to study the influence of synthetic data under different configurations. We compile various synthetic datasets based on the variations from real datasets. Although deep learning based face detectors have various types, they share some underlying architectures. We choose three different detectors as representatives. Synthetic data are customized based on the configuration of the datasets and face detectors. In our experiment results, we demonstrate that synthetic data can be a good complementary source to real data, to make face detectors more robust against extreme variations. Our analysis provides an example to evaluate the performance of face detectors towards different variations.

Question 2: How can we alleviate the negative influence of pose variations when predicting age?

To overcome the impact of extreme head pose, our method is based on a pose invariant representation. Face uv texture representation is reconstructed from the original video frames. uv texture assigns 3D texture into 2D space with universal per-pixel alignment for all textures. Each vertex in a 3D shape has a corresponding 2D texture coordinate. The reconstructed uv texture in a pose invariant fashion contains the estimated frontal view of the face. Normally, parts of the uv texture are missing due to the influence of pose. The challenge lies in the completion of uv texture because the missing regions are highly irregular. Wasserstein based GAN is used to recover the full face region. For age estimation, besides cross entropy loss, we also consider to penalize the mean and variance of predicted age distribution. This combination of loss functions can force the network to predict a more concentrated age distribution towards the ground truth. Age loss is added to the generator part of the total loss function. Our method is able to complete the face uv texture and predict age simultaneously. In order to train our method, we collect the largest unconstrained face videos dataset (UvAge) with age labels. Compared to existing datasets, UvAge dataset has much larger inter-subject variations. Our dataset can provide researchers with more robust age prediction methods in the future. Through extensive experiments, it was shown that our method improved accuracy in age estimation. A completed and frontal uv texture can help the network to retrieve the most influential features for age estimation.

Question 3: Can we find a robust method to distinguish deep fake data from multiple domains?

The most challenging part of detecting deep fake data are: 1) Pristine and modified samples of the same subject look more similar than pristine examples from other subjects. 2) Different generative methods induce various artifacts from imaging features and face characteristics in the modified samples. We propose a deep network based on a joint supervision framework to detect manipulated face images. Maximum mean discrepancy loss has been used to learn a more generalized feature space for multiple domains of manipulation results. In order to maximize the intra-distances between positive samples and minimize their inter-distances, triplet constraint is used to penalize triplet relationship among batches. Furthermore, center loss has been integrated to provide a more discriminative embedding for forensics detection. Our proposed method achieved the best overall performance on UADFV, DF-TIMIT, Celeb-DF and FaceForensics++.

Our method is not only robust against one certain manipulation method but also multi-domain deep fake datasets. Moreover, we provided a detailed analysis about other distribution alignment methods. Extensive experiments showed that our algorithm has high accuracy in detecting face forensics.

Question 4: How can we mitigate the influence of imbalanced distribution and improve the performance of video based age predictions?

Most age datasets have an imbalanced and long-tailed distribution. The imbalanced issue has considerably degraded the performance of video based age estimation. We used a transfer learning module to mitigate the imbalanced issue in existing age datasets. Our transfer learning module is essentially based on a deep embedded clustering (DEC) module to jointly cluster data and learn a better feature representation. This can transfer knowledge from majority categories to improve performance on minority groups. To alleviate the impact of imbalanced age samples in the training process, the target age distribution is built on a linear combination of different age predictions. Our method does not rely on hard constraints of the age distribution. Through extensive experiments, we showed that our method substantially outperformed other age estimation methods on both constrained (UvA-Nemo, Nemo-Deception) and unconstrained datasets (UvAge). The performance on minority age groups has also been improved. At last, we examine the influence of different variations on age estimation. The results show that our method is robust against different type of variations.

Bibliography

[1] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey. Using synthetic data to improve facial expression analysis with 3d convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1609–1618, 2017. (Cited on page 13.)

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. (Cited on pages 47, 49, and 50.)

[3] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–45, 2019. (Cited on pages 4, 41, and 43.)

[4] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. (Cited on page 43.)

[5] R. Angulu, J. R. Tapamo, and A. O. Adewumi. Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, 2018(1):42, 2018. (Cited on pages 29 and 55.)

[6] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international conference on image processing (ICIP)*, pages 2089–2093. IEEE, 2017. (Cited on page 5.)

[7] H. Bal, D. Epema, C. de Laat, R. van Nieuwpoort, J. Romein, F. Seinstra, C. Snoek, and H. Wijshoff. A medium-scale distributed system for computer science research: Infrastructure for the long term. *Computer*, 49(5):54–63, 2016. (Cited on page 19.)

[8] B. Bayar and M. C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10, 2016. (Cited on pages 49 and 50.)

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19 (7):711–720, 1997. (Cited on page 3.)

[10] M. Blackall. Channel 4 under fire for deepfake queen's christmas message. URL `https://www.theguardian.com/technology/2020/dec/24/channel-4-under-fire-for-deepfake-queen-christmas-message`. (Cited on page 4.)

[11] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. (Cited on pages 3 and 13.)

[12] K. Bowyer and S. Sarkar. Usf humanid 3d face dataset. 2001. (Cited on page 3.)

[13] J. F. Boylan. Will deep-fake technology destroy democracy. URL `https://www.nytimes.com/2018/10/17/opinion/deep-fake-technology-democracy.html`. (Cited on page 4.)

[14] N. I. Brown. Deepfakes and the weaponization of disinformation. *Va. JL & Tech.*, 23:1, 2020. (Cited on page 4.)

[15] M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. (Cited on page 57.)

[16] J. Buolamwini. Gender shades. URL `http://gendershades.org/overview.html`. (Cited on page 6.)

[17] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019. (Cited on page 57.)

[18] T. Carvalho, F. A. Faria, H. Pedrini, R. d. S. Torres, and A. Rocha. Illuminant-based transformed spaces for image forensics. *IEEE transactions on information forensics and security*, 11(4):720–733, 2015. (Cited on page 43.)

[19] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. (Cited on page 18.)

[20] K.-Y. Chang and C.-S. Chen. A learning framework for age rank estimation based on face images with scattering transform. *IEEE Transactions on Image Processing*, 24(3):785–798, 2015. (Cited on page 31.)

[21] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR 2011*, pages 585–592. IEEE, 2011. (Cited on page 31.)

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. (Cited on page 57.)

[23] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015. (Cited on pages 30 and 37.)

[24] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5183–5192, 2017. (Cited on pages 31 and 55.)

[25] Z. Chen, S. Huang, and D. Tao. Context refinement for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–86, 2018. (Cited on page 14.)

[26] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. (Cited on pages 3, 5, and 13.)

[27] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. (Cited on pages 47, 49, and 50.)

[28] V. Conotter, E. Bodnari, G. Boato, and H. Farid. Physiologically-based detection of computer generated faces in video. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 248–252. IEEE, 2014. (Cited on page 43.)

[29] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. (Cited on page 3.)

[30] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *European conference on computer vision*, pages 484–498. Springer, 1998. (Cited on page 3.)

[31] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. (Cited on pages 5 and 43.)

[32] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018. (Cited on page 58.)

[33] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019. (Cited on pages 6, 57, 63, and 66.)

[34] cydonia999. A pytorch implementation of detectron. `https://github.com/roytseng-tw/Detectron.pytorch`, 2017. (Cited on page 19.)

[35] S. Dack. Deep fakes, fake news, and what comes next, 2019. (Cited on page 4.)

[36] D.-T. Dang-Nguyen, G. Boato, and F. G. De Natale. Identify computer generated characters by analysing facial expressions variation. In *2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 252–257. IEEE, 2012. (Cited on page 43.)

[37] T. J. De Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7):1182–1194, 2013. (Cited on page 43.)

[38] Deepfakes. deepfakes faceswap. `https://github.com/deepfakes/faceswap`, 2018. (Cited on page 5.)

[39] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2018. (Cited on pages 3, 32, and 34.)

[40] H. Dibeklioğlu, T. Gevers, A. A. Salah, and R. Valenti. A smile can reveal your age: Enabling facial dynamics in age estimation. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 209–218. ACM, 2012. (Cited on pages 29, 31, and 56.)

[41] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, pages 525–538. Springer, 2012. (Cited on pages 29, 31, 33, 61, and 62.)

[42] H. Dibeklioğlu, F. Alnajar, A. A. Salah, and T. Gevers. Combining facial dynamics with appearance for age estimation. *IEEE Transactions on Image Processing*, 24(6):1928–1943, 2015. (Cited on pages 29, 31, and 56.)

[43] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634,

2015. (Cited on pages 59 and 63.)

[44] Q. Dong, S. Gong, and X. Zhu. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1851–1860, 2017. (Cited on page 57.)

[45] C. N. Duong, K. Luu, K. G. Quach, N. Nguyen, E. Patterson, T. D. Bui, and N. Le. Automatic face aging in videos via deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10013–10022, 2019. (Cited on pages 3, 4, 31, 32, and 33.)

[46] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. (Cited on page 30.)

[47] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016. (Cited on page 30.)

[48] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. (Cited on pages 29, 34, and 66.)

[49] F. Fernández-Navarro, C. Hervás-Martínez, and P. A. Gutiérrez. A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recognition*, 44(8):1821–1833, 2011. (Cited on page 66.)

[50] P. Galeano and D. Peña. Multivariate analysis in vector time series. 2001. (Cited on pages 67 and 68.)

[51] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 256–263. IEEE, 2009. (Cited on pages 4, 30, 31, and 56.)

[52] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. (Cited on pages 5 and 42.)

[53] S. Ge, J. Li, Q. Ye, and Z. Luo. Detecting masked faces in the wild with lle-cnns. In *IEEE CVPR*, 2017. (Cited on pages 2, 11, 13, 14, and 16.)

[54] Y. Geifman and R. El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017. (Cited on page 57.)

[55] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412, 2013. (Cited on pages 31 and 56.)

[56] X. Geng, Q. Wang, and Y. Xia. Facial age estimation by adaptive label distribution learning. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4465–4470. IEEE, 2014. (Cited on pages 31 and 56.)

[57] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. (Cited on page 3.)

[58] P. Ghosh, P. S. Gupta, R. Uziel, A. Ranjan, M. Black, and T. Bolkart. Gif: Generative interpretable faces. *arXiv preprint arXiv:2009.00149*, 2020. (Cited on page 3.)

[59] R. Goh, L. Liu, X. Liu, and T. Chen. The cmu face in action (fia) database. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 255–263. Springer, 2005. (Cited on pages 31 and 33.)

[60] X. Golay, S. Kollias, G. Stoll, D. Meier, A. Valavanis, and P. Boesiger. A new correlation-based fuzzy logic clustering algorithm for fmri. *Magnetic Resonance in Medicine*, 40(2):249–260, 1998. (Cited on pages 67 and 68.)

[61] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. (Cited on page 3.)

[62] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012. (Cited on page 43.)

[63] P. Grother, M. Ngan, and K. Hanaoka. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects.* National Institute of Standards and Technology, 2019. (Cited on page 6.)

[64] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017. (Cited on page 35.)

[65] G. Guo and N. Zhang. A survey on deep learning based face recognition. *Computer Vision and Image*

Understanding, 189:102805, 2019. (Cited on page 1.)

[66] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119. IEEE, 2009. (Cited on pages 29 and 55.)

[67] Hadid and Abdenour. Analyzing facial behavioral features from videos. In *International Workshop on Human Behavior Understanding*, pages 52–61. Springer, 2011. (Cited on pages 31 and 56.)

[68] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *2013 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2013. (Cited on page 29.)

[69] J. Han and T. Gevers. Mmd based discriminative learning for face forgery detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[70] J. Han, W. Wang, and T. Gevers. Deep imbalanced learning for age estimation from videos. *Under review at Computer Vision and Image Understanding*, .

[71] J. Han, W. Wang, S. Karaoglu, W. Zeng, and T. Gevers. Pose invariant age estimation of face images in the wild. *Computer Vision and Image Understanding*, 202:103123, .

[72] J. Han, W. Wang, S. Karaoglu, W. Zeng, and T. Gevers. Pose invariant age estimation of face images in the wild. *Computer Vision and Image Understanding*, page 103123, 2020. (Cited on page 4.)

[73] J. Han, S. Karaoglu, H.-A. Le, and T. Gevers. Object features and face detection performance: Analyses with 3d-rendered synthetic data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9959–9966. IEEE, 2021.

[74] K. Han, A. Vedaldi, and A. Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8401–8409, 2019. (Cited on page 58.)

[75] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu. Scale-aware face detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017. (Cited on page 13.)

[76] D. Harwell. Scarlett johansson on fake ai-generated sex videos: nothing can stop someone from cutting and pasting my image. (Cited on page 4.)

[77] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4295–4304, 2015. (Cited on page 13.)

[78] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini. The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233, 2000. (Cited on page 1.)

[79] Z. He, X. Li, Z. Zhang, F. Wu, X. Geng, Y. Zhang, M.-H. Yang, and Y. Zhuang. Data-dependent label distribution learning for age estimation. *IEEE Transactions on Image processing*, 26(8):3846–3858, 2017. (Cited on page 56.)

[80] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012. (Cited on page 12.)

[81] P. Hu and D. Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 951–959, 2017. (Cited on pages 13 and 14.)

[82] T. Hu, J. Xu, C. Huang, H. Qi, Q. Huang, and Y. Lu. Weakly supervised bilinear attention network for fine-grained visual classification. *arXiv preprint arXiv:1808.02152*, 2018. (Cited on page 46.)

[83] C. Huang, Y. Li, C. C. Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. (Cited on pages 6, 57, and 66.)

[84] C. Huang, Y. Li, C. L. Chen, and X. Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2019. (Cited on pages 2, 55, and 57.)

[85] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, 2017. (Cited on pages 11 and 12.)

[86] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2439–2448, 2017. (Cited on pages 3 and 5.)

[87] Y. Huang, F. Juefei-Xu, R. Wang, X. Xie, L. Ma, J. Li, W. Miao, Y. Liu, and G. Pu. Fakelocator: Robust localization of gan-based face manipulations via semantic segmentation networks with bells and whistles. *arXiv preprint arXiv:2001.09598*, 2020. (Cited on page 43.)

[88] B. L. Hughes, N. P. Camp, J. Gomez, V. S. Natu, K. Grill-Spector, and J. L. Eberhardt. Neural adaptation to faces reveals racial outgroup homogeneity effects in early perception. *Proceedings of the National Academy of Sciences*, 116(29):14532–14537, 2019. (Cited on page 1.)

[89] Z. Huo, X. Yang, C. Xing, Y. Zhou, P. Hou, J. Lv, and X. Geng. Deep age distribution learning for apparent age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 17–24, 2016. (Cited on page 55.)

[90] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. (Cited on page 3.)

[91] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010. (Cited on page 11.)

[92] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. (Cited on page 34.)

[93] V. A. Jones. *Artificial Intelligence Enabled Deepfake Technology: The Emergence of a New Threat*. PhD thesis, Utica College, 2020. (Cited on page 4.)

[94] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. (Cited on pages 3 and 4.)

[95] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. (Cited on pages 3 and 4.)

[96] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. (Cited on page 3.)

[97] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017. (Cited on pages 6 and 57.)

[98] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. (Cited on page 4.)

[99] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. (Cited on pages 42, 47, 48, and 49.)

[100] A. Kortylewski, A. Schneider, T. Gerig, B. Egger, A. Morel-Forster, and T. Vetter. Training deep face recognition systems with synthetic data. *arXiv preprint arXiv:1802.05891*, 2018. (Cited on page 13.)

[101] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372. IEEE, 2009. (Cited on page 30.)

[102] P. Kumar, M. Vatsa, and R. Singh. Detecting face2face facial reenactment in videos. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2589–2597, 2020. (Cited on pages 4, 5, and 43.)

[103] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *Computer vision and image understanding*, 74(1):1–21, 1999. (Cited on pages 29 and 55.)

[104] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. (Cited on pages 58 and 60.)

[105] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1): 621–628, 2004. (Cited on pages 31 and 56.)

[106] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017. (Cited on page 43.)

[107] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. (Cited on page 43.)

[108] L. Li, Y. Peng, G. Qiu, Z. Sun, and S. Liu. A survey of virtual sample generation technology for face recognition. *Artificial Intelligence Review*, 50(1):1–20, 2018. (Cited on page 13.)

[109] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. *arXiv preprint arXiv:1912.13458*, 2019. (Cited on page 43.)

[110] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian. Bridgenet: A continuity-aware probabilistic network for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1145–1154, 2019. (Cited on page 31.)

[111] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint*

arXiv:1811.00656, 2, 2018. (Cited on pages 42, 43, 47, 48, and 49.)

[112] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016. (Cited on page 52.)

[113] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3911–3919, 2017. (Cited on page 31.)

[114] Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017. (Cited on pages 42, 43, and 45.)

[115] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-df: A new dataset for deepfake forensics. *ArXiv*, 2019. (Cited on pages 42, 47, 48, and 49.)

[116] H. Liao, Y. Yan, W. Dai, and P. Fan. Age estimation of face images based on cnn and divide-and-rule strategy. *Mathematical Problems in Engineering*, 2018, 2018. (Cited on pages 31 and 56.)

[117] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. (Cited on page 63.)

[118] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. (Cited on page 58.)

[119] Y. Lu, Y.-W. Tai, and C.-K. Tang. Conditional cyclegan for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*, 2, 2017. (Cited on page 5.)

[120] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou. Data augmentation for face recognition. *Neurocomputing*, 230:184–196, 2017. (Cited on pages 2, 11, and 13.)

[121] N. S. Madiraju, S. M. Sadat, D. Fisher, and H. Karimabadi. Deep temporal clustering: Fully unsupervised learning of time-domain features. *arXiv preprint arXiv:1802.01059*, 2018. (Cited on page 58.)

[122] MarekKowalski. Faceswap. `https://github.com/MarekKowalski/FaceSwap`, 2015. (Cited on page 5.)

[123] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596. Springer, 2016. (Cited on pages 2 and 13.)

[124] I. Masi, Y. Wu, T. Hassner, and P. Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018. (Cited on page 1.)

[125] I. Masi, A. T. Tran, T. Hassner, G. Sahin, and G. Medioni. Face-specific data augmentation for unconstrained face recognition. *International Journal of Computer Vision*, 127(6-7):642–667, 2019. (Cited on pages 2 and 11.)

[126] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019. (Cited on page 43.)

[127] I. Matthews, J. Xiao, and S. Baker. 2d vs. 3d deformable face models: Representational power, construction, and real-time fitting. *International journal of computer vision*, 75(1):93–113, 2007. (Cited on page 3.)

[128] G. Mazaheri, N. C. Mithun, J. H. Bappy, and A. K. Roy-Chowdhury. A skip connection architecture for localization of image manipulations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 119–129, 2019. (Cited on page 43.)

[129] M. Minear and D. C. Park. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4):630–633, 2004. (Cited on page 30.)

[130] H. Nada, V. A. Sindagi, H. Zhang, and V. M. Patel. Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. *arXiv preprint arXiv:1804.10275*, 2018. (Cited on pages 2 and 17.)

[131] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis. Ssh: Single stage headless face detector. In *IEEE ICCV*, 2017. (Cited on pages 13 and 14.)

[132] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019. (Cited on page 43.)

[133] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. IEEE, 2019. (Cited on pages 43, 47, and 49.)

[134] C. Nhan Duong, K. Luu, K. Gia Quach, and T. D. Bui. Longitudinal face modeling via temporal deep restricted boltzmann machines. In *Proceedings of the IEEE Conference on Computer Vision and*

Pattern Recognition, pages 5772–5780, 2016. (Cited on page 31.)

[135] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. (Cited on page 31.)

[136] M. Osadchy, Y. Wang, O. Dunkelman, S. Gibson, J. Hernandez-Castro, and C. Solomon. Genface: Improving cyber security using realistic synthetic face generation. In *International Conference on Cyber Security Cryptography and Machine Learning*, pages 19–33. Springer, 2017. (Cited on page 13.)

[137] W. Ouyang, X. Wang, C. Zhang, and X. Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 864–873, 2016. (Cited on page 58.)

[138] S. Palsson, E. Agustsson, R. Timofte, and L. Van Gool. Generative adversarial style transfer networks for face aging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2084–2092, 2018. (Cited on pages 4 and 32.)

[139] H. Pan, H. Han, S. Shan, and X. Chen. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5285–5294, 2018. (Cited on pages 31, 34, 36, 56, 63, 65, and 66.)

[140] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the fg-net ageing database. *Iet Biometrics*, 5(2):37–46, 2016. (Cited on pages 2, 4, 30, 31, and 56.)

[141] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015. (Cited on page 45.)

[142] K. Paul. California makes 'deepfake' videos illegal, but law may be hard to enforce, 2019. URL https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce. (Cited on page 4.)

[143] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. IEEE, 2009. (Cited on page 3.)

[144] W. Pei, H. Dibeklioğlu, T. Baltrušaitis, and D. M. Tax. Attended end-to-end architecture for age estimation from facial expression videos. *arXiv preprint arXiv:1711.08690*, 2017. (Cited on page 31.)

[145] W. Pei, H. Dibeklioğlu, T. Baltrušaitis, and D. M. Tax. Attended end-to-end architecture for age estimation from facial expression videos. *IEEE Transactions on Image Processing*, 29:1972–1984, 2019. (Cited on pages 56, 63, 65, and 66.)

[146] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018. (Cited on pages 4 and 13.)

[147] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017. (Cited on pages 43, 49, and 50.)

[148] N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131–144, 2009. (Cited on pages 31 and 56.)

[149] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. (Cited on pages 2 and 14.)

[150] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE, 2006. (Cited on page 30.)

[151] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv*, 2018. (Cited on pages 5 and 49.)

[152] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. (Cited on pages 41, 42, 47, and 49.)

[153] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018. (Cited on pages 30, 36, 63, and 65.)

[154] N. O. Rule and N. Ambady. Face and fortune: Inferences of personality from managing partners' faces predict their law firms' financial success. *The Leadership Quarterly*, 22(4):690–696, 2011. (Cited on page 1.)

[155] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3:1, 2019. (Cited on page 43.)

[156] E. Sánchez-Lozano, G. Tzimiropoulos, and M. Valstar. Joint action unit localisation and intensity estimation through heatmap regression. *arXiv preprint arXiv:1805.03487*, 2018. (Cited on pages 64 and 65.)

[157] C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *International conference on biometrics*, pages 199–208. Springer, 2009. (Cited on page 47.)

[158] K. Scherbaum, J. Petterson, R. S. Feris, V. Blanz, and H.-P. Seidel. Fast face detector training using tailored views. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2848–2855, 2013. (Cited on page 13.)

[159] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. (Cited on pages 45 and 47.)

[160] R. Shao, X. Lan, J. Li, and P. C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019. (Cited on page 43.)

[161] L. Shen, Z. Lin, and Q. Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016. (Cited on page 57.)

[162] W. Shen and R. Liu. Learning residual images for face attribute manipulation. *arXiv preprint arXiv:1612.05363*, 2016. (Cited on page 3.)

[163] W. Shen and R. Liu. Learning residual images for face attribute manipulation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1225–1233. IEEE, 2017. (Cited on pages 5 and 13.)

[164] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille. Deep regression forests for age estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2304–2313. IEEE, 2018. (Cited on page 36.)

[165] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille. Deep differentiable random forests for age estimation. *arXiv preprint arXiv:1907.10665*, 2019. (Cited on page 36.)

[166] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen. Real-time rotation-invariant face detection with progressive calibration networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2295–2303, 2018. (Cited on pages 2 and 13.)

[167] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5444–5453. IEEE, 2017. (Cited on page 13.)

[168] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. (Cited on page 34.)

[169] B. Singh and L. S. Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3578–3587, 2018. (Cited on page 14.)

[170] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Josa a*, 4(3):519–524, 1987. (Cited on page 3.)

[171] J. Snow. Amazon's face recognition falsely matched 28 members of congress with mugshots. URL `https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28`. (Cited on page 6.)

[172] G. Somanath, M. Rohith, and C. Kambhamettu. Vadana: A dense dataset for facial image analysis. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, pages 2175–2182. IEEE, 2011. (Cited on page 30.)

[173] K. Songsri-in and S. Zafeiriou. Complement face forensic detection and localization with faciallandmarks. *arXiv preprint arXiv:1910.05455*, 2019. (Cited on page 43.)

[174] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. (Cited on page 4.)

[175] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. (Cited on page 47.)

[176] K. Tang, J. Huang, and H. Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv preprint arXiv:2009.12991*, 2020. (Cited on page 57.)

[177] X. Tang, D. K. Du, Z. He, and J. Liu. Pyramidbox: A context-assisted single shot face detector. In

Proceedings of the European Conference on Computer Vision (ECCV), pages 797–813, 2018. (Cited on page 13.)

[178] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. (Cited on pages 58 and 60.)

[179] L. Taylor and G. Nitschke. Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547. IEEE, 2018. (Cited on pages 2 and 11.)

[180] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. (Cited on pages 4, 13, 43, and 49.)

[181] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv preprint arXiv:1904.12356*, 2019. (Cited on pages 43 and 49.)

[182] K. M. Ting. A comparative study of cost-sensitive boosting algorithms. In *In Proceedings of the 17th International Conference on Machine Learning*. Citeseer, 2000. (Cited on pages 6 and 57.)

[183] A. Todorov, A. N. Mandisodza, A. Goren, and C. C. Hall. Inferences of competence from faces predict election outcomes. *Science*, 308(5728):1623–1626, 2005. (Cited on page 1.)

[184] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020. (Cited on page 43.)

[185] M. Turk, A. Pentland, P. Belhumeur, and J. Hespanha. Eigenfaces for recognition: Journal of cognitive neurosicence. 1991. (Cited on page 3.)

[186] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. (Cited on page 43.)

[187] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65, 2010. (Cited on page 33.)

[188] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. (Cited on pages 59 and 60.)

[189] J. Wang and L. Perez. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, page 11, 2017. (Cited on pages 2 and 11.)

[190] J. Wang, Y. Yuan, and G. Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017. (Cited on pages 2 and 13.)

[191] R. Wang, L. Ma, F. Juefei-Xu, X. Xie, J. Wang, and Y. Liu. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019. (Cited on pages 41 and 43.)

[192] S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros. Detecting photoshopped faces by scripting photoshop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10072–10081, 2019. (Cited on page 43.)

[193] X. Wang, K. Wang, and S. Lian. A survey on face data augmentation. *arXiv preprint arXiv:1904.11685*, 2019. (Cited on page 13.)

[194] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1968–1984, 2009. (Cited on page 13.)

[195] Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017. (Cited on page 57.)

[196] Z. Wang, X. Tang, W. Luo, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7939–7947, 2018. (Cited on pages 3, 4, and 32.)

[197] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. (Cited on page 46.)

[198] J. Willis and A. Todorov. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598, 2006. (Cited on page 1.)

[199] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. *arXiv preprint arXiv:2007.09654*, 2020. (Cited on pages 57 and 63.)

[200] H. C. W.W. Bledsoe. A man–machine facial recognition system—some preliminary results. *Panoramic Research*, 1965. (Cited on page 2.)

[201] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016. (Cited on pages 58 and 60.)

[202] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu, and K. Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018. (Cited on pages 41 and 45.)

[203] H. Yang, D. Huang, Y. Wang, and A. K. Jain. Learning face age progression: A pyramid architecture of gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 31–39, 2018. (Cited on pages 3, 4, and 32.)

[204] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016. (Cited on pages 11 and 17.)

[205] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang. Ssr-net: A compact soft stagewise regression network for age estimation. In *IJCAI*, volume 5, page 7, 2018. (Cited on pages 36 and 63.)

[206] X. Yang, X. Geng, and D. Zhou. Sparsity conditional energy label distribution learning for age estimation. In *IJCAI*, pages 2259–2265, 2016. (Cited on pages 31 and 56.)

[207] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. (Cited on page 43.)

[208] Z. Yang and R. Nevatia. A multi-scale cascade fully convolutional network face detector. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 633–638. IEEE, 2016. (Cited on page 13.)

[209] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–684, 2015. (Cited on pages 2 and 13.)

[210] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, 2019. (Cited on pages 6 and 58.)

[211] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. (Cited on page 34.)

[212] N. Yu, L. S. Davis, and M. Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7556–7566, 2019. (Cited on pages 41 and 43.)

[213] D. Yudin, M. Shchendrygin, and A. Dolzhenko. Age and gender recognition on imbalanced dataset of face images with deep learning. In *International Conference on Intelligent Information Technologies for Industry*, pages 30–40. Springer, 2019. (Cited on page 58.)

[214] L. ZEBROWITZ. *READING FACES: Window to the Soul?*. ROUTLEDGE, 2019. (Cited on page 1.)

[215] L. A. Zebrowitz, R. Wang, P. M. Bronstad, D. Eisenberg, E. Undurraga, V. Reyes-García, and R. Godoy. First impressions from faces among us and culturally isolated tsimane'people in the bolivian rainforest. *Journal of Cross-Cultural Psychology*, 43(1):119–134, 2012. (Cited on page 1.)

[216] C. Zhang, S. Liu, X. Xu, and C. Zhu. C3ae: Exploring the limits of compact model for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12587–12596, 2019. (Cited on page 31.)

[217] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. (Cited on page 13.)

[218] S. Zhang, L. Wen, H. Shi, Z. Lei, S. Lyu, and S. Z. Li. Single-shot scale-aware network for real-time face detection. *International Journal of Computer Vision*, pages 1–23, 2019. (Cited on page 13.)

[219] Y. Zhang, L. Liu, C. Li, et al. Quantifying facial age by posterior of age comparisons. *arXiv preprint arXiv:1708.09687*, 2017. (Cited on page 30.)

[220] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020. (Cited on page 57.)

[221] M. Zhou, H. Lin, S. S. Young, and J. Yu. Hybrid sensing face detection and registration for low-light and unconstrained conditions. *Applied optics*, 57(1):69–78, 2018. (Cited on page 13.)

[222] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839. IEEE, 2017. (Cited on page 43.)

[223] Y. Zhou, D. Liu, and T. Huang. Survey of face detection on low-quality images. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 769–773. IEEE, 2018. (Cited on page 14.)

[224] Z.-H. Zhou and X.-Y. Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77, 2005. (Cited on pages 6 and 57.)

[225] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015. (Cited on page 13.)

[226] Y. Zhu, J.-Y. Zhu, and W.-S. Zheng. Part-based convolutional network for imbalanced age estimation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 946–951. IEEE, 2019. (Cited on page 58.)

[227] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. (Cited on page 57.)

Samenvatting

Overzicht

Dit hoofdstuk besluit dit proefschrift door onze onderzoeksvragen uit Hoofdstuk 1 opnieuw te bekijken, onze belangrijkste bevindingen te bespreken en richtingen voor toekomstig onderzoek te schetsen. We concentreren ons op de belangrijkste bevindingen en algemene lessen; aanvullende gedetailleerde bevindingen staan in de slotparagrafen van de afzonderlijke hoofdstukken. De individuele conclusies voor elk hoofdstuk worden als volgt gepresenteerd:

Hoofdstuk 2: Analyse van object kenmerken en kwaliteit van gezichtsdetectie
In dit hoofdstuk hebben we een experimentele vergelijking voorgesteld van de belangrijkste kenmerken die de kwaliteit van gezichtsdetectie beïnvloeden. Er wordt een synthetische generator van data voorgesteld om 2D-gezichten te synthetiseren op basis van 3D-gezichtsmodellen. We hebben de synthetische dataset aangepast om specifieke soorten kenmerken nader te analyseren (schaal, houding, occlusie, vervaging, enz.). Vervolgens selecteren we drie representatieve gezichtsdetectoren om systematisch de invloed van verschillende functies op de prestaties van gezichtsdetectie te onderzoeken. Onze resultaten laten zien dat synthetische data een goede aanvullende bron kunnen zijn voor echte datasets. De prestaties van gezichtsdetectoren kunnen ook worden verbeterd door verschillende soorten gesynthetiseerde variaties. Door middel van onze analyses hebben we ook enkele mogelijke tekortkomingen van de huidige architecturen voor gezichtsdetectie geïdentificeerd. Tot slot zijn er vaak uitdagende complicaties bij gezichtsherkenning in de echte wereld. Door een overzicht te geven van de relatie tussen de objectkenmerken en de kwaliteit van gezichtsdetectie, hopen we onderzoekers te helpen bij het kiezen van meer geschikte synthetische data bij het adresseren van uitdagende variaties in het echte leven.

Hoofdstuk 3: Houding-invariante leeftijd schatting van gezichtsafbeeldingen in het wild
In dit hoofdstuk richten we ons op het aanpakken van het negatieve effect van de houding van het hoofd bij het inschatten van leeftijd. Ten eerste hebben we de grootste video dataset in het wild geïntroduceerd voor het schatten van de leeftijd. Het bevat onbeperkte video's van beroemdheden in verschillende evenementen. Om de voorspelling van de leeftijd meer robuust te maken tegen houding-variatie, reconstrueren we gezicht uv texturen van de originele 2D-frames van video's. Delen van de gereconstrueerde uv -structuren ontbreken vanwege het zelf-occlusie-effect van de houding van het hoofd. In deze taak zijn de grootte en vorm van ontbrekende gebieden van het gezicht zeer onregelmatig. We bieden een op-Wasserstein-GAN gebaseerde benadering (AgeGAN) om tegelijkertijd de werkelijke leeftijd te schatten en de gedeeltelijke uv texturen te voltooien. Onze methode kan het netwerk dwingen om de ontbrekende regio's op te halen met meer betekenisvolle functies voor het schatten van de leeftijd. Om de effectiviteit aan te tonen, vergelijken we onze methode voor het aanvullen van gezichten met andere geavanceerde invulmethoden. Ook evalueren we systematisch onze methoden om leeftijd mee te schatten op andere datasets.

Hoofdstuk 4: Discriminatief leren voor detectie van gezichtsvervalsing in meerdere domeinen

Dit hoofdstuk concentreerde zich op de uitdaging van detectie van gezichtsvervalsing in meerdere domeinen. De belangrijkste uitdagingen bij het opsporen van gezichtsvervalsing zijn: 1) Het verschil tussen ongerepte en neppe voorbeelden is veel kleiner dan het verschil tussen ongerepte voorbeelden. 2) De artefacten van beeldkenmerken en gezichtskenmerken blijven niet bestaan in alle gegenereerde resultaten voor dezelfde generatieve methode. We hebben een end-to-end diepe netwerk gebaseerde architectuur voorgesteld. Geïnspireerd door de toepassingen van neurale stijloverdracht, willen we dat ons netwerk zich concentreert op meer onderscheidende kenmerken in plaats van over-aanpassing aan manipulatie-specifieke artefacten. Maximale gemiddelde discrepantie (MMD) kan ons helpen kenmerken van verschillende distributies op elkaar af te stemmen. De MMD kost functie wordt gebruikt om een meer algemene representatie te leren voor meerdere domeinen van manipulatie-resultaten. Bovendien is de triplet-beperking opgenomen om de intra-afstanden te minimaliseren en de inter-afstanden te maximaliseren. De Centrum kost functie is geïntegreerd om een onderscheidende representatie voor forensische detectie te bieden.

Onze voorgestelde methode behaalde de beste algehele prestaties op UADFV, DF-TIMIT, Celeb-DF en FaceForensics++. Bovendien gaven we een gedetailleerde analyse van elk onderdeel in ons raamwerk en hielden we rekening met de prestaties van andere distributiemethoden. Uitgebreide experimenten hebben aangetoond dat ons algoritme een hoge capaciteit en nauwkeurigheid heeft bij het detecteren van forensische aspecten gezichten.

Hoofdstuk 5: Diep onevenwichtig leren voor leeftijdsschatting op basis van video's

In dit hoofdstuk proberen we leeftijden te voorspellen op basis van video's. Datasets voor gezicht-gerelateerd onderzoek vertonen vaak zeer scheve klasse-verdelingen. De meeste gegevens behoren tot slechts een paar meerderheidscategorieën, terwijl de minderheidsklassen veel minder gevallen bevatten. We hebben een end-to-end-methode voorgesteld om leeftijd te voorspellen op basis van video's. Om het onevenwichtige probleem in bestaande datasets op te lossen, hebben we een clustermodule gebruikt om gezamenlijk gegevens te clusteren en een betere weergave van functies te extraheren. Dit kan kennis uit verwante categorieën gebruiken om de prestaties van minderheidsgroepen te verbeteren. Om de impact van onevenwichtige leeftijdssteekproeven in het trainingsproces te verzachten, wordt de echte data verdeling geconstrueerd door een lineaire combinatie van verschillende leeftijdsvoorspellingen. De voorgestelde methode houdt geen rekening met stricte beperkingen op de marge van de leeftijdsverdeling. Bovendien worden tijdelijke ensemble- en consistentiebeperkingen toegevoegd om de gladheid van het gloeiproces verder te verbeteren.

Door uitgebreide experimenten hebben we aangetoond dat onze methode substantieel beter presteerde dan andere methoden op zowel beperkte (UvA-Nemo, Nemo-Deception) als niet-beperkte datasets (UvAge). Bovendien zijn de prestaties op de leeftijdsgroepen van minderheden ook verbeterd.

Algemene discussie

In deze sectie bespreken we twee belangrijke beperkingen in ons werk en mogelijke toekomstige richtingen.

Synthetische data spelen een cruciale rol in ons Hoofdstuk 2. De variaties van datasets hebben zowel kansen als uitdagingen geboden voor het deep learning-systeem. De uiteindelijke prestaties van de gezicht-gerelateerde applicaties zijn sterk afhankelijk van de variaties in datasets. Synthetische data kunnen een effectieve manier zijn om de variaties van echte datasets te verrijken. Het voordeel van het gebruik van synthetische gegevens is dat we meer inzicht krijgen in hoe de prestaties van het leersysteem afhankelijk zijn van verschillende soorten gesynthetiseerde variaties. Met alle voordelen van synthetische data is er weliswaar altijd een onvermijdelijke kloof tussen synthetische data en echte data. Toekomstig werk zou zich kunnen concentreren op het opvullen of verkleinen van de kloof.

In Hoofdstuk 3 gebruiken we een GAN om de ontbrekende gebieden van de uv-textuur van het gezicht te herstellen. Resultaten van generatieve methoden worden steeds realistischer. Ze kunnen aanvullende variaties voor het leersysteem bieden. Het is echter moeilijk om de generator volledig te beheersen tijdens het leerproces. Momenteel kunnen we de resultaten van generatieve methoden niet volledig verklaren. Er zijn ook geen gemeenschappelijke statistieken om de prestaties van generatieve methoden te meten. Voor toekomstig onderzoek kunnen we onderzoeken hoe we het gedrag van generatieve methoden echt kunnen beheersen. Het is ook beter als we de correlaties tussen sommige gezichtskenmerken zoals mannelijkheid en het hebben van een baard tijdens het leerproces kunnen verwijderen. De meeste methoden voor gezichtsmodificatie worden al op grote schaal gebruikt voordat er wordt onderzocht wat de sociale impact van de methoden is. Gebruikersonderzoek zou nuttig zijn om te begrijpen hoe mensen de gesynthetiseerde gezichten begrijpen.

Het afgelopen decennium is getuige geweest van een explosieve vooruitgang van gezichtsonderzoek en uitgebreide toepassingen in de praktijk. Daarentegen is onderzoek naar hun maatschappelijke, vooral ethische, implicaties grotendeels over het hoofd gezien. De nauwkeurigheid en het bereik van gezichtsherkenning zijn drastisch gegroeid en worden soms opdringerige krachten in het dagelijkse leven van mensen en zelfs in privacydomeinen. We stellen voor dat onderzoek naar gezichtsalgoritmen en hun sociale implicaties hand in hand moet gaan, zodat de gezichtsgeneratieve methoden evolueren op een manier die in de ogen van het grote publiek als geloofwaardig en betrouwbaar wordt beschouwd.

Conclusie

We sluiten het proefschrift af door de vragen in Hoofdstuk 1 opnieuw te bekijken.
Vraag 1: Kunnen we systematisch variaties in synthetische data manipuleren om de echte dataset aan te vullen en verder betere prestaties te behalen bij gezichtsdetectie?

Het korte antwoord is ja. Het merendeel van de gegevens uit bestaande datasets behoort normaal gesproken tot een beperkt aantal variaties. De gezichten vertegenwoordigden niet voldoende extreme houding, schaal of zware occlusie om een robuuste detector te trainen tegen alle mogelijke variaties. We gebruiken synthetische data als vergroting van de data om de onvoldoende variatie van echte datasets te compenseren.

Eerst ontwerpen we een nieuwe synthetische gezichtsgegevensgenerator met volledig gecontroleerde variaties (op houding, schaal, achtergrond, verlichting en occlusie). In vergelijking met bestaande methoden voor gezichtsmanipulatie hebben we meer controle over de gesynthetiseerde variaties. Dan zijn we in staat om de invloed van synthetische data onder verschillende configuraties te bestuderen. We stellen verschillende synthetische datasets samen op basis van de variaties van echte datasets. Hoewel op deep learning gebaseerde gezichtsdetectoren verschillende typen hebben, delen ze een aantal onderliggende architecturen. We kiezen drie verschillende detectoren als vertegenwoordigers. Synthetische gegevens worden aangepast op basis van de configuratie van de datasets en gezichtsdetectoren. In onze experimenten laten we zien dat synthetische gegevens een goede aanvullende bron kunnen zijn voor echte gegevens, om gezichtsdetectoren robuuster te maken tegen extreme variaties. Onze analyse biedt een voorbeeld om de prestaties van gezichtsdetectoren voor verschillende variaties te evalueren.

Vraag 2: Hoe kunnen we de negatieve invloed van houding-variaties bij het voorspellen van leeftijd verminderen?

Om de impact van een extreme hoofdhouding te overwinnen, is onze methode gebaseerd op een houding-invariante representatie. De structuurweergave van het gezicht uv wordt gereconstrueerd op basis van de originele videobeelden. uv texture wijst 3D-textuur toe aan 2D-ruimte met universele uitlijning per pixel voor alle texturen. Elk hoekpunt in 3D-vorm heeft een corresponderende 2D-textuurcoördinaat. De gereconstrueerde uv-textuur op een houding-invariante manier bevat het geschatte vooraanzicht van het gezicht. Normaal gesproken ontbreken delen van de uv-textuur vanwege de invloed van houding. De uitdaging ligt in de voltooiing van uv textuur omdat de ontbrekende gebieden zeer onregelmatig zijn. Een Wasserstein GAN wordt gebruikt om het volledige gezichtsgebied te herstellen. Voor het inschatten van de leeftijd overwegen we naast de cross-entropie ook het gemiddelde en de variantie van de voorspelde leeftijdsverdeling te bestraffen. Deze combinatie van kost functies kan het netwerk dwingen een meer geconcentreerde leeftijdsverdeling naar de grondwaarheid te voorspellen. Ouderdomsverlies wordt toegevoegd aan het generatorgedeelte van de totale kost functie. Onze methode is in staat om de face uv textuur te voltooien en tegelijkertijd de leeftijd te voorspellen. Om onze methode te trainen, verzamelen we de grootste dataset met ongedwongen gezichtsvideo's (UvAge) met leeftijdslabels. In vergelijking met bestaande datasets heeft de UvAge dataset veel grotere interpersoonlijke variaties. Onze dataset kan onderzoekers in de toekomst robuustere methoden voor leeftijdsvoorspelling bieden. Door uitgebreide experimenten werd aangetoond dat onze methode de nauwkeurigheid bij het inschatten van leeftijd verbeterde. Een voltooide en frontale uv-textuur kan het netwerk helpen om de meest invloedrijke functies voor het schatten van de leeftijd op te halen.

Vraag 3: Kunnen we een robuuste methode vinden om diepe nep gegevens van meerdere domeinen te onderscheiden?

Het meest uitdagende deel van het detecteren van diepe nep gegevens zijn: 1) Ongerepte en gemodificeerde voorbeelden van hetzelfde onderwerp lijken meer op elkaar dan ongerepte voorbeelden van andere onderwerpen. 2) Verschillende generatieve methoden wekken verschillende artefacten op van beeldkenmerken en gezichtskenmerken in de gemodificeerde voorbeelden. We stellen een diep netwerk voor op basis van een gezamenlijk toezichtkader om gemanipuleerde gezichtsbeelden te detecteren. Het maximale gemiddelde discrepantieverlies is gebruikt om een meer algemene feature-ruimte te leren voor meerdere domeinen van manipulatieresultaten. Om intra-afstanden tussen positieve voorbeelden te maximaliseren en inter-afstanden te minimaliseren, wordt een triplet-beperking gebruikt om de triplet-relatie tussen batches te bestraffen. Bovendien is centrumverlies geïntegreerd om een meer onderscheidende representatie voor forensische detectie te bieden. Onze voorgestelde methode behaalde de beste algehele prestaties op UADFV, DF-TIMIT, Celeb-DF en FaceForensics ++. Onze methode is niet alleen robuust tegen één bepaalde manipulatiemethode, maar ook tegen multi-domein deep fake datasets. Bovendien hebben we een gedetailleerde analyse gegeven van andere methoden voor distributie-uitlijning. Uitgebreide experimenten hebben aangetoond dat ons algoritme een hoge nauwkeurigheid heeft bij het detecteren van gezichtsforensisch onderzoek.

Vraag 4: Hoe kunnen we de invloed van onevenwichtige distributie verminderen en de prestaties van op video gebaseerde leeftijdsvoorspellingen verbeteren?

De meeste ouderdomsdatasets hebben een onevenwichtige en langdurige verdeling. Het onevenwichtige probleem heeft de prestaties van op video gebaseerde leeftijdsschatting aanzienlijk verslechterd. We hebben een leermodule voor overdracht gebruikt om het onevenwichtige probleem in bestaande leeftijdsdatasets te verminderen. Onze transfer learning-module is in wezen gebaseerd op een deep embedded clustering (DEC)-module om gezamenlijk gegevens te clusteren en een betere weergave van functies te leren. Dit kan kennis van meerderheidscategorieën overdragen om de prestaties van minderheidsgroepen te verbeteren. Om de impact van onevenwichtige leeftijdssteekproeven in het trainingsproces te verminderen, is de beoogde leeftijdsverdeling gebaseerd op een lineaire combinatie van verschillende leeftijdsvoorspellingen. Onze methode is niet gebaseerd op harde beperkingen van de leeftijdsverdeling. Door middel van uitgebreide experimenten hebben we aangetoond dat onze methode substantieel beter presteerde dan andere leeftijdsschattingsmethoden op zowel beperkte (UvA-Nemo, Nemo-Deception) als niet-beperkte datasets (UvAge). De prestaties op de leeftijdsgroepen van minderheden zijn ook verbeterd. Ten slotte onderzoeken we de invloed van verschillende variaties op leeftijdsschatting. De resultaten laten zien dat onze methode robuust is tegen verschillende soorten variaties.

Acknowledgement

Pursuing the Ph.D. degree has been a truly challenging journey. Throughout this journey I have received a great amount of support and help. I am so grateful to all the people who have inspired and helped me throughout these years. Without their help and care, my thesis would not have been possible.

I would like first to thank my supervisor and promoter, Theo Gevers, for his guidance, motivation, and immense knowledge. His guidance helped in all the time of my Ph.D. study. His expertise was invaluable in formulating the research questions and methodology. We had many discussions to clarify my research interests and what I can do in this journey. In these years, I have learned from Theo about how to make plans, how to approach a problem as a scientist, how to think solution-oriented, and how to set priorities right, and so many things I could not possibly mention.

I would like to also thank my co-promoter Sezer Karaoglu. Whenever I had a problem, I could always go to him and ask for advice. I received enormous advice and support from him during my study. This inspired me a lot for my research. It is a great pleasure to work with him. His advice and optimism have helped me tremendously.

I would like to thank my committee, Prof. dr. Dong Shen, Prof. dr. Albert Salah, Prof. dr. Marcel Worring, Dr. Shaodi You, and Dr. Arnoud Visser for honoring me with reading my thesis and participating in my defense. I am fortunate to have many wonderful people who helped me over the past years. I would like to thank all of them: Berkay, Chang, Dan, Dennis, Emrah, Erik, Fares, Finde, Gjorgji, Hamdi, Jacqueline, Jan-Mark, Jiaojiao, Juan, Kirill, Lei, Leo, Lorena, Mert, Morris, Nicole, Nour, Ozzy, Pascal, Partha, Peter, Qi Bi, Qi Wang, Que, Ran, Rick, Ruihong, Shaodi, Shuai, Shihan, Shuo, Shuxian, Thomas, Tingting, Tom, Wenyang, Weijie, William, Xiaotian, Xiaoyue, Xiaowei, Yang, Yi Ding, Yi Liu, Yunlu, Zenglin, Zhenyang, Zhiwei. It is a great pleasure of being part of this big family, and having you in my life. I am so grateful for having such a nice time in the past few years.

My sincere gratitude goes to my great colleagues who share the office and memories for last five years, Anil Baslamisli, Hoang-An Le and Hanan ElNaghy, Minh Ngo, Wei Zeng. Wei Zeng, we share a lot about our life. No matter what happens, I know you are always there to support me. Anil, An, Hanan and Minh, you were always kind, supportive, and willing to lend an ear when asked. I dont't know how many times you helped me. I learned a lot from all of you. Thanks for your company and friendship. Without all of you, I am not going to work through my research. I will always remember our conversations and laughter. I also want to thank Wei Wang and Yahui Zhang. Although we only work together for a relatively short time, it is always a pleasure to work and discuss with you. I wish both of you have a successful career. A big thank you goes to Ziming Li as a loyal friend who supports me throughout my research. At last, I want to thank my girlfriend Mengchen Dong. This is the second diploma I get since we were together seven years ago. Without you, I would not even start this advanture. I feel fortunate that you always stand by me for all these years. I cannot forget to thank my parents and friends for all the unconditional and unlimited support and love in these five years.