Deep Face Analysis for Aesthetic Augmented Reality Applications

Yongzhe Yan

DOCTORAL THESIS

Deep Face Analysis for Aesthetic Augmented Reality Applications

Author:
Yongzhe YAN

Director of the thesis:
Thierry CHATEAU

Defended on:
19 June 2020

Composition of the jury:

Rapporteur: Chaabane DJERABA, Professeur, Univ. de Lille
Rapporteure: Catherine ACHARD, MCF-HDR, Sorbonne Université
Examinateur: Frédéric JURIE, Professeur, Univ. Caen Normandie
Examinateur: Vincent LEPETIT, Professeur, ENPC ParisTech

Encadrant: Thierry CHATEAU, Professeur, Univ. Clermont Auvergne
Co-encadrant: Stefan DUFFNER, MCF-HDR, Univ. de Lyon
Co-encadrant: Christophe BLANC, MCF, Univ. Clermont Auvergne
Co-encadrant: Christophe GARCIA, Professeur, Univ. de Lyon
Co-encadrant: Xavier NATUREL, Ingénieur Ph.D, Invité

A thesis submitted in fulfillment of the requirements
for the degree of Docteur de l'Université Clermont Auvergne
at the

Comsee - ISPR
Institut Pascal

UNIVERSITÉ CLERMONT AUVERGNE
Ecole Doctorale des Sciences Pour l'Ingénieur
Institut Pascal

Abstract

Deep Face Analysis for Aesthetic Augmented Reality Applications

by Yongzhe YAN

Precise and robust facial component detection is of great importance for the good user experience in aesthetic augmented reality applications such as virtual make-up and virtual hair dying. In this context, this thesis addresses the problem of facial component detection via facial landmark detection and face parsing. The scope of this thesis is limited to deep learning-based models.

The first part of this thesis addresses the problem of facial landmark detection. In this direction, we propose three contributions. For the first contribution, we aim at improving the precision of the detection. To improve the precision to pixel-level, we propose a coarse-to-fine framework which leverages the detail information on the low-level feature maps. We train different stages with different loss functions, among which we propose a boundary-aware loss that forces the predicted landmarks to stay on the boundary. For the second contribution in facial landmark detection, we improve the robustness of facial landmark detection. We propose 2D Wasserstein loss to integrate additional geometric information during training. Moreover, we propose several modifications to the conventional evaluation metrics for model robustness.

To provide a new perspective for facial landmark detection, we present a third contribution on exploring a novel tool to illustrate the relationship between the facial landmarks. We study the Canonical Correlation Analysis (CCA) of the landmark coordinates. Two applications are introduced based on this tool: (1) the interpretation of different facial landmark detection models (2) a novel weakly-supervised learning method that allows to considerably reduce the manual effort for dense landmark annotation.

The second part of this thesis tackles the problem of face parsing. We present two contributions in this part. For the first contribution, we present a framework for hair segmentation with a shape prior to enhance the robustness against the cluttered background. Additionally, we propose a spatial attention module attached to this framework, to improve the output of the hair boundary. For the second contribution in this part, we present a fast face parsing framework for mobile phones, which leverages temporal consistency to yield a more robust output mask. The implementation of this framework runs in real-time on an iPhone X.

Keywords: Facial Landmark Detection; Face Parsing; Hair Segmentation; Deep Learning

UNIVERSITÉ CLERMONT AUVERGNE
Ecole Doctorale des Sciences Pour l'Ingénieur
Institut Pascal

Résumé

Analyse du Visage pour les Applications de Réalité Augmentée Esthétique

par Yongzhe YAN

La détection précise et robuste des composants faciaux est d'une grande importance pour la bonne expérience utilisateur dans les applications de réalité augmentée à destination de l'industrie esthétique telles que le maquillage virtuel et la coloration virtuelle des cheveux. Dans ce contexte, cette thèse aborde le problème de la détection des composants faciaux via la détection des repères faciaux et la segmentation des composantes faciales. Cette thèse se concentre sur les modèles basés sur l'apprentissage profond.

La première partie de cette thèse aborde le problème de la détection des repères faciaux. Nous proposons trois contributions. Pour la première contribution de cette partie, nous visons à améliorer la précision de la détection. Afin d'améliorer la précision au niveau des pixels, nous proposons un framework grossier à fin qui exploite les informations détaillées sur les feature maps de bas niveau dans le modèle. Nous formons différentes étapes avec différentes fonctions de coût, parmi lesquelles nous proposons une fonction sensible aux contours qui force les points de repère estimés à rester sur le contour de composants faciaux. Dans la deuxième contribution de cette partie, nous améliorons la robustesse de la détection des repères faciaux. Nous proposons une fonction de coût, basée sur la distance Wasserstein, pour intégrer des informations géométriques supplémentaires lors de l'apprentissage. De plus, nous proposons plusieurs modifications aux métriques d'évaluation conventionnelles pour mieux appréhender la robustesse du modèle.

Pour fournir une nouvelle perspective sur la détection des repères faciaux, nous présentons une troisième contribution sur l'exploration d'un nouvel outil pour illustrer la relation entre les repères faciaux. Nous étudions l'analyse canonique de corrélation (CCA) des coordonnées du point de repère. Deux applications sont introduites avec cet outil: (1) l'interprétation de différents modèles pour la détection de points de repère (2) une nouvelle méthode d'apprentissage faiblement supervisé qui permet de réduire considérablement l'effort manuel pour l'annotation dense de points de repère.

La deuxième partie de cette thèse aborde le problème de la segmentation des composantes faciales. Nous proposons deux contributions. Dans la première contribution dans cette partie, nous présentons un framework pour la segmentation des cheveux, afin d'améliorer la robustesse sur les arrière-plans complexes. De plus, un module d'attention spatiale est attaché à ce framework pour améliorer les résultats sur le contour des cheveux. Dans la deuxième contribution de cette partie, nous présentons un framework rapide de segmentation des composantes faciales pour les téléphones mobiles, qui utilise la cohérence temporelle pour produire un masque de sortie plus robuste. L'implémentation de ce framework s'exécute en temps réel sur un iPhone X.

Mot clés : détection des repères faciaux; segmentation des composantes faciales; segmentation de cheveux; apprentissage profond

This thesis is dedicated to my parents.

Acknowledgements

First of all, I would like to thank professor Thierry Chateau, my main academic advisor, for his help, encouragement, enthusiasm, and patience in the last three years. This thesis would ever be possible without him. Having the opportunity to work with him, I was able to learn how to become a good scientific researcher. And this is the experience that I will cherish in my whole life.

My deep gratitude goes to my co-supervisor, Stefan Duffner, who inspires me a lot in the domain of deep learning. Besides, for countless times, he worked with me till late night to catch up with the conference deadline and I feel so lucky to have his support. I also want to thank my co-supervisors Christophe Gacia and Christophe Blanc, for their firm support when I met difficulties during this thesis.

I would thank Xavier Naturel, for his consistent help throughout the last three years, even after he had left Wisimage and did not have the obligation to do so.

Furthermore, I would like to express my sincere appreciation to Chaabane Djeraba and Catherine Achard who agreed to review this manuscript, as well as to the rest of the Ph.D. committee members, including Frédéric Jurie and Vincent Lepetit.

I would like to express my gratitude to Christian Wolf for generously granting access to the GPU cluster in the beginning of this thesis, and the Mésocentre Clermont Auvergne University for providing computing resources.

I would like to thank my fellow Ph.D. student Anthony Berthelier, it is a great pleasure to have you as a co-PhD since the D day. My gratitude also goes to the members of the ComSee group: Céline, Ruddy, Rémi, Yizhen, Antoine, Ala, Gauthier, and all others with whom I had lots of interesting discussions. I also want to thank my fellows in the Imagine Team in LIRIS: Yiqiang, Fabien, and Quentin.

My sincere gratitude goes to all of the (ex)members of Wisimage: Léo, Wendy, Benjamin, Aurélie, Arnaud, Issam, Mohammad, Zesheng, Alex, Gael. In particular, I would like to thank Priyanka Phutane for bringing some fresh air into this thesis since the 2nd year.

My thanks also go to my friends at Clermont-Ferrand, Chao Zhang, Jinpeng Wang, Chen Xu, Jiarui Xie for enriching my spare time.

Finally, I would like to express my infinite gratitude to my family and Pangpang, for their constant care and encouragement.

Contents

List of Figures

List of Tables

Glossary

AR Augmented Reality. xiii, 1, 2, 4–7, 9, 11, 22, 31, 33, 47, 55, 91, 93, 107, 108, 110, 114, 115

CCA Canonical Correlation Analysis. 8, 72, 115

CED Cumulative Error Distribution. xiii, 44, 46–48, 58, 71

CNN Convolutional Neural Network. xiii, 6, 8, 9, 11–19, 24, 25, 27, 31, 33, 35, 36, 41, 42, 49, 94, 95, 97, 107, 108, 113–115

FC Fully Connected. 11, 14, 33, 34, 36, 49, 50, 72

FCNN Fully Convolutional Neural Network. 8, 9, 14, 19, 72, 91, 94, 96, 100, 102

FR Failure Rate. xvii, 57–59, 63, 65–68, 70, 71

GAN Generative Adversarial Network. xiii, 2, 3, 6

IoU Intersection over Union. 25, 98, 99, 101

MLP Multilayer Perceptron. xiii, 11, 12, 17

NME Normalized Mean Error. xiv, xvii, 43–48, 58, 59, 62–68, 71, 72, 76, 82–87

RNN Recurrent Neural Network. 19, 20, 23–25

RoI Region of Interest. 5–7, 20, 25, 26, 107

Chapter 1

Introduction

1.1 General Context

1.1.1 A Brief History of Augmented Reality

The desire of experiencing alternative and reciprocative reality is deeply rooted in the nature of human beings. Figure 1.1 shows a brief timeline of the Augmented Reality (AR) history. In the late 1960s, I. Sutherland invented the first head-mounted three dimensional display device [Sutherland, 1968], called *The Sword of Damocles*, which laid the foundation for the AR that we use today. The user was able to visualize an alternate reality by this ceiling-hung device. The computer-generated images were able to interact with the user by detecting the head position.

In the early 1990s, the term of Augmented Reality was first introduced by T. Caudell [Caudell and Mizell, 1992]. Soon L. Rosenberg [Rosenberg, 1993] developed the first operational AR system, named *Virtual Fixtures*, which helps to improve the efficiency of the operators. Since the late 1990s, AR has started to be introduced in various applications such as space navigation, broadcasting of live sports events, battlefield simulation, etc. [Isberto, 2018].

Modern AR applications have been spread to wider domains. In the medical field, for example, AR has been used to guide the doctors during surgeries [Bichlmeier et al., 2007]. In the entertainment field, the phenomenal gaming application *Pokémon GO* enables more people to experience AR for the first time on their mobile phones. In the educational field, I. Radu [Radu, 2012] claimed that the AR-based learning materials help the student to keep the learning motivation, to increase the content understanding, to maintain a long-term memory retention, etc.

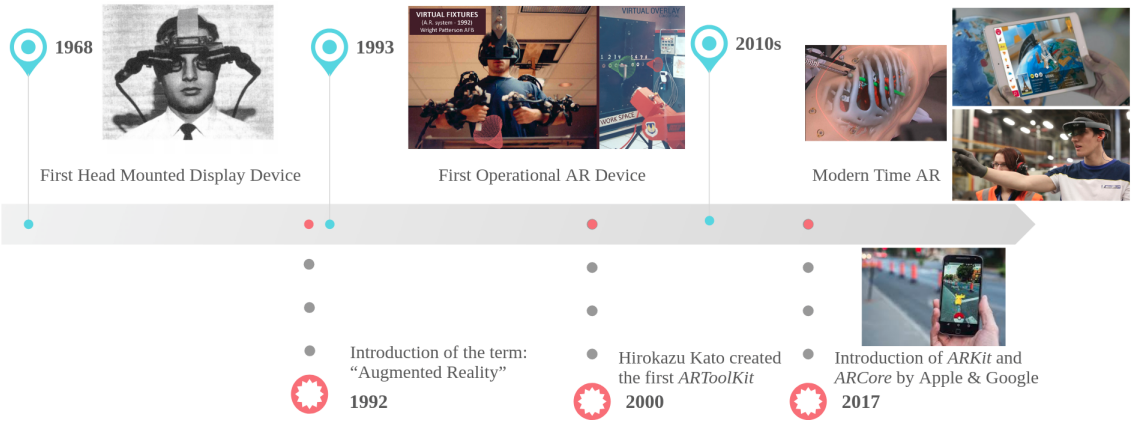
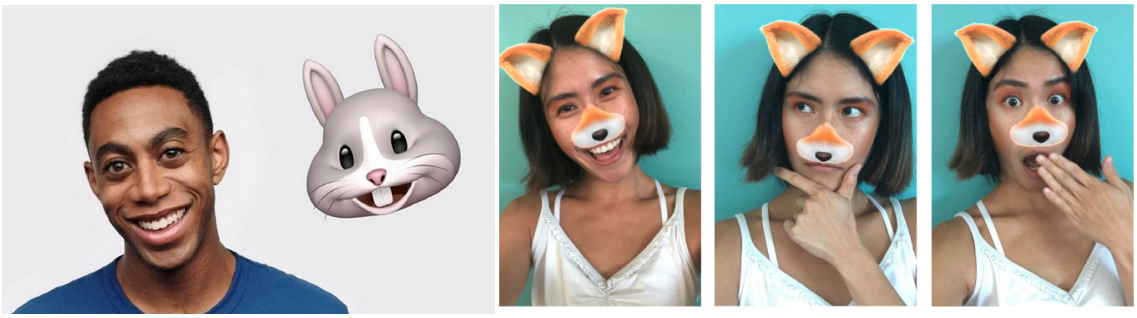


FIGURE 1.1: A brief timeline of the AR history.

In recent times, the development of AR profits from the popularity of the Internet and mobile phones. In 2017, Apple introduce *ARKit* on its iOS system and Google introduced *ARCore* on its Android system. Both of them significantly facilitate the development of AR applications on the mobile phones. From the users' point of view, the mobile phone drastically lowers the barrier of the device availability so that no supplementary equipment is required for an AR experience. Thus, the AR industry is now rapidly growing. According to a recent estimate by Goldman Sachs [The Goldman Sachs Group, 2016], the AR industry, along with the Virtual Reality industry, are expected to grow into a 95 billion dollar market by the year 2025.

1.1.2 Face AR Applications

Among all the AR applications, face AR application is a prominent and popular one. With the front camera of the mobile phones, thousands of AR applications are now easily accessible to the general public including face modelling, face animation, face swapping, face sticker etc. We show two examples of the face AR applications introduced by Google and Apple in figure 1.2.



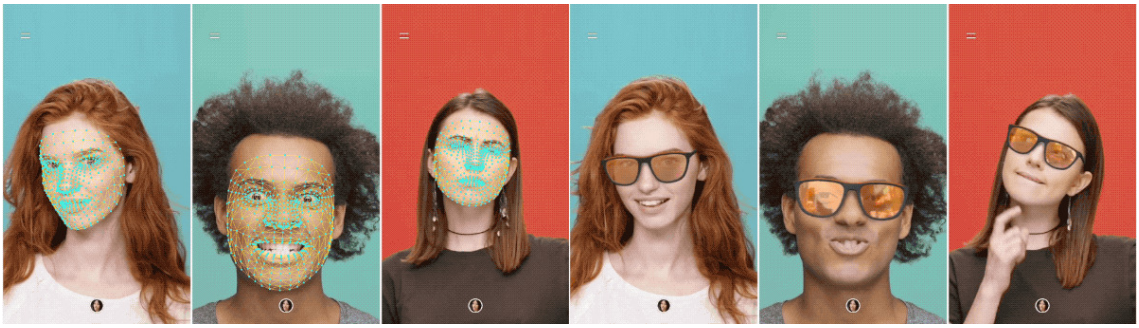
(A) *Animoji* introduced by Apple (B) *Augmented Faces* introduced in Google *ARCore*

FIGURE 1.2: Examples of face AR applications on mobile platforms.

In recent times, there are hundreds of applications developed for the usage of applying virtual effects on the faces, especially with the recent advances in Generative Adversarial Network (GAN) [Goodfellow et al., 2014] and style transfer [Gatys et al., 2016]. More specifically, we present three categories of techniques for facial photo editing.

- **Overlay-based face editing:** This is a traditional and intuitive approach for face photo editing. The main idea is simple: we first detect the position of the faces as well as the position of the facial components (such as eyes, nose and mouth), and then we overlay virtual effects on the corresponding places. We present an example in Fig. 1.3 to show how virtual glasses are put on the faces. Please note that the overlay is usually not a simple copy-and-paste operation but includes carefully designed post processing steps such as 3D deformation.

- **GAN-based face editing:** Unlike overlay-based photo editing, GAN-based face editing enables us to edit the photo by pure learning process without manual design of a virtual effect template. The principal advantage is that the virtual effect is more realistic as it can adapt to various lighting conditions and poses on the source image. We present an example of virtual makeup in Fig 1.4. We observe that compared to overlay-based method (mentioned as warping result), the GAN-based method retains more details on the source images such as the color of the eyes' iris, but is still capable of applying a similar virtual effect from the reference



(A) Detection of the face semantics (B) Overlaying of the visual effect

FIGURE 1.3: An example of overlay-based face photo editing. This figure is acquired from [Ablavatski and Grishchenko, 2019].

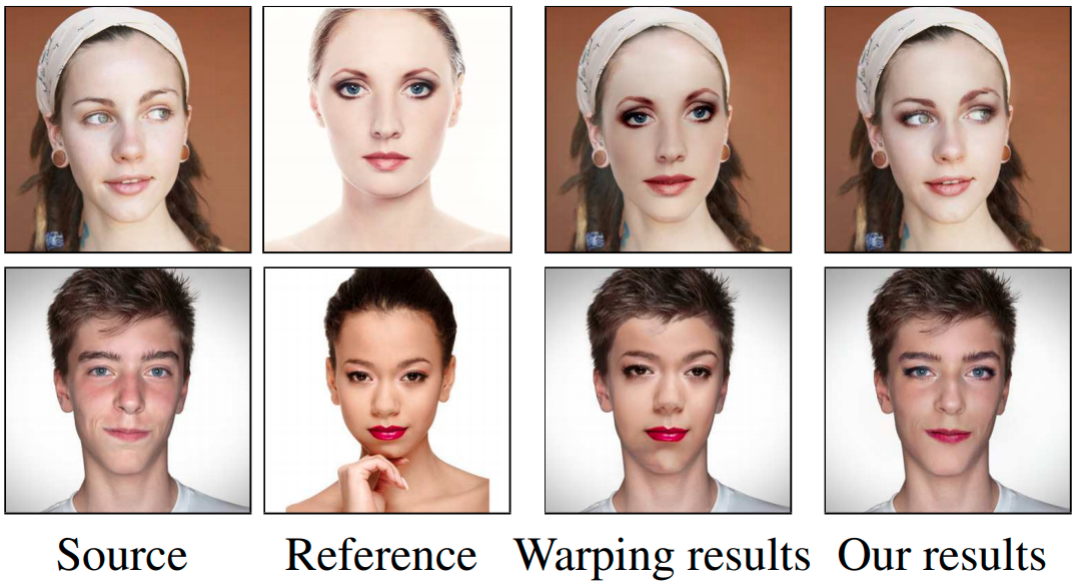


FIGURE 1.4: An example of GAN-based face photo editing. The makeup effect on the reference image is applied on the source image. Warping results denote previously mentioned overlay-based face editing. This figure is acquired from [Chang et al., 2018].

image. However, this method requires a large amount of images during training to achieve realistic effects.

- **Style transfer-based face editing:** Liu et al. [Liu et al., 2016] proposed to consider the facial virtual effect as an artistic style. Therefore, applying virtual facial effect can be achieved by using style transfer [Gatys et al., 2016, Liao et al., 2017]. An eye shadow effect transfer is shown in Fig. 1.5. Similar to GAN-based face photo editing, no template of virtual effect is required. However, there is no guarantee that the rendering will look natural and similar to the reference images.

In fact, all of the three categories mentioned above are performed locally and require precise and robust facial component detection, especially for overlay-based face editing. Therefore, in this thesis, we mainly focus on the detection and classification of semantic points or face regions for virtual facial makeup and virtual hair coloring.

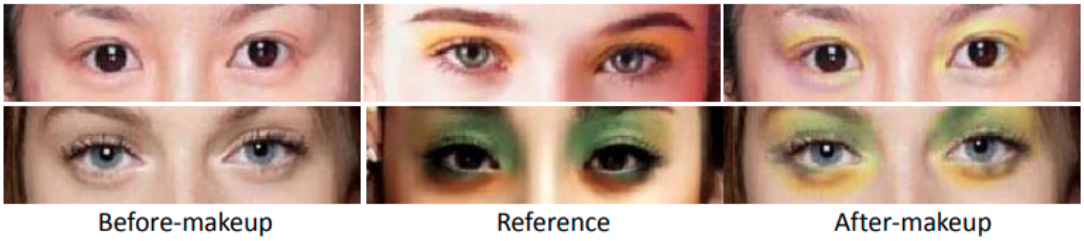


FIGURE 1.5: An example of style transfer-based face photo editing. The makeup effect on the reference image is applied on the source image. This figure is acquired from [Liu et al., 2016]

1.1.3 Industrial Context of This Thesis

This thesis has been conducted in collaboration with an industrial partner *Wisimage*, a start-up company which focuses on providing the face aesthetic AR applications for the beauty industry. Wisimage has conducted profound scientific research on human face analysis [Vu, 2010, Schwab, 2013] including face recognition and face tracking. Wisimage now manages a diversified product line (see Fig. 1.6) for cosmetic industry including virtual makeup simulation, skin diagnosis, cosmetic product database and cosmetic product recommendation. Several applications provided by *Wisimage* are illustrated in figure 1.7. The ongoing projects at Wisimage are (1) Software Development Kit for facial landmark detection (2) virtual colored lens simulation (3) virtual makeup simulation, including makeup recognition and makeup transfer.

Specifically in the virtual makeup simulation project (see Fig. 1.7 (A)), *Wisimage* provided this face AR applications on mobile phones, in real time. The algorithm required in this application for detecting the position of the facial components is closely related to the human face analysis problem in the computer vision research domain.

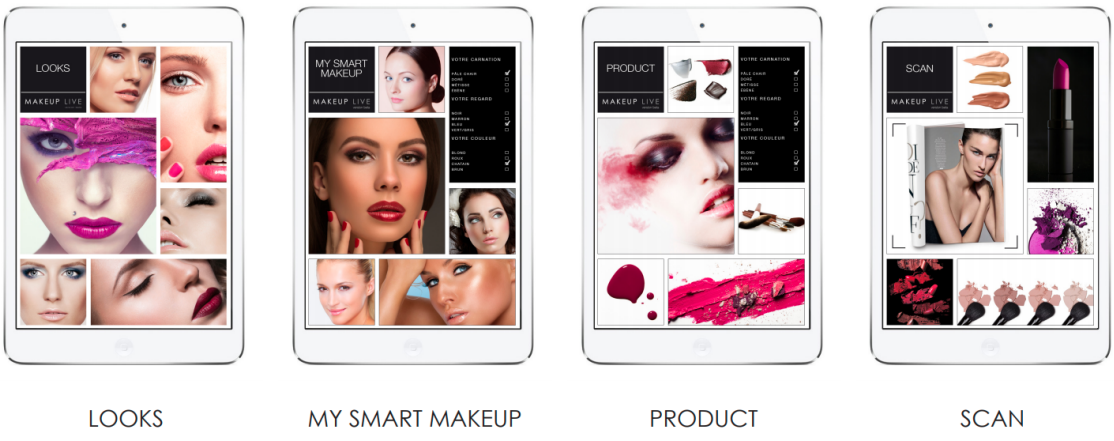
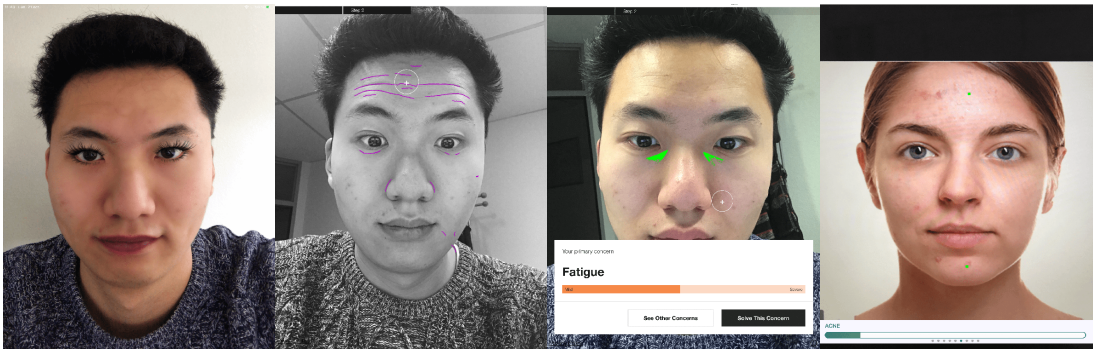


FIGURE 1.6: Product line of *Wisimage*.

1.2 Scientific Context

1.2.1 Standard Human Face Analysis Pipeline

One of the most important research topics which supports these face AR application is human face analysis. Human face analysis is a key subject of computer vision research



(A) Virtual makeup (B) Skin wrinkle detection (C) Skin fatigue detection (D) Skin acne detection

FIGURE 1.7: Several examples of face AR applications proposed by Wisimage.

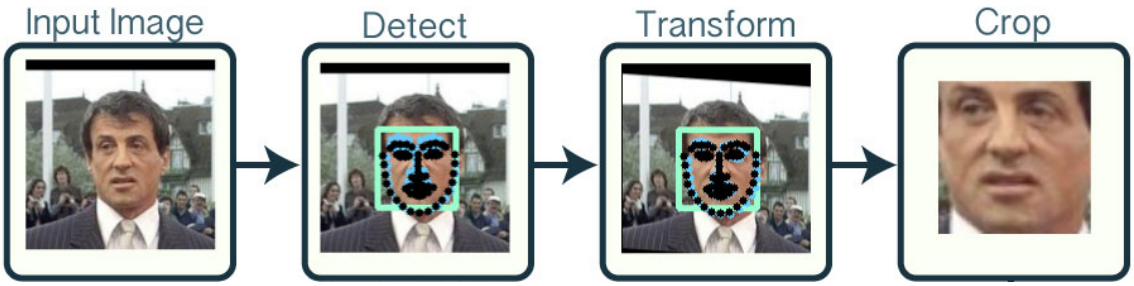


FIGURE 1.8: A standard pre-processing pipeline of face recognition. This figure is adopted from Open Face [Amos et al., 2016].

due to its wide-ranged applications. Traditional pipelines of face analysis usually comprise two prerequisite steps: face detection and facial landmark detection. A standard human face analysis pipeline for face recognition is presented in Fig. 1.8.

Face detection aims at finding all human faces presented in an image. This step provides a Region of Interest (RoI) for the consecutive analysis, so that the subsequent algorithm can focus on a single face. Viola-Jones face detector [Viola and Jones, 2001], introduced in 2001, is still a popular and effective detector under common circumstances. The recent research on face detection are principally focused on highly challenging situations [Yang et al., 2016]. In facial biometric AR applications, face detector rarely prevents the good user experiences.

Usually, facial landmark detection is a second step for human face analysis. Facial landmark detection aims at detecting the key landmarks around the facial components (such as eye contours, mouth contours etc). This step provides abundant information on face rotation, head pose and position of facial components, which is critical for the following steps. For example, in most face recognition pipelines, face images are first transformed to a canonical shape based on the output of face alignment (see Fig. 1.8). Therefore, facial landmark detection is also referenced as face alignment. Facial landmark detection is of great importance because it describes the detailed shape and position of the facial components.

1.2.2 Research Subjects

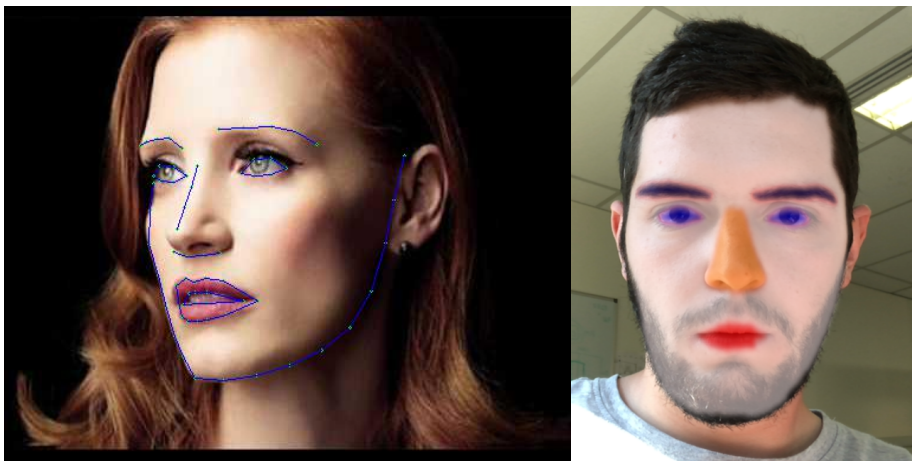
In this thesis, we focus on the problem of how to accurately find out the position of the facial components to improve the user experience for the face AR applications. To apply an artificial make-up effect on the human face, it is indispensable to correctly and precisely recognize the position and the exact shape of each facial components including

eyes, nose, lips and hair. Overall, we consider two approaches to detect the positions of the facial components:

- **Facial landmark detection**, as introduced previously, it aims at retrieving the position of facial feature points or fiducial facial points that are located at semantic boundaries such as eye contours, face contour and mouth contours (see Fig. 1.9 (A)).

  With the position of these landmarks, we are capable of warping the virtual AR make-up effect from a canonical shape onto the real shape on the face presented on the image for overlay-based face editing.

- **Facial parsing** aims at getting the pixel-wise label mask of the facial semantic components (see Fig. 1.9 (B)). The output mask naturally provides a region where we can determine the RoI for style transfer-based face editing and GAN-based face editing.



(A) Facial Landmark Detection (B) Face Parsing

FIGURE 1.9: Demonstration of the two main research subjects in this thesis.

Since 2011, deep learning methods have been profoundly influencing the area of computer vision and continue to do so. In many traditional tasks such as image classification [Krizhevsky et al., 2012], object detection [Ren et al., 2015] and semantic segmentation [Long et al., 2015], deep learning-based methods outperform the state-of-the-art. In fact, using deep learning (in particular Convolutional Neural Network (CNN) models) to analyze human faces can be dated back to the early 2000s [Garcia and Delakis, 2002, Garcia and Delakis, 2004, Osadchy et al., 2007, Duffner, 2008]. Recently, deep learning regained its dominating position on numerous face analysis tasks such as face recognition [Parkhi et al., 2015], face reconstruction [Dou et al., 2017], face detection [Ranjan et al., 2017], including the task of facial landmark detection and face parsing. Therefore, the main interest of this thesis lies in deep learning-based methods.

1.2.3 Challenges

We introduce the challenges of the two previously introduced research subjects in the context of face AR applications.

Challenges of facial landmark detection: Compared to the biometric applications such as face recognition, a significantly higher standard for facial landmark detection is required to ensure the good user experiences:

- **Precision:** The predicted landmarks need to be very accurate and aligned to the boundary. Users can easily notice a slight displacement of the overlying artificial effect if a predicted landmark misses the boundary. Several works [Feng et al., 2018b, Wang et al., 2019a] have been proposed to overcome this problem.

- **Robustness/stability:** The predicted landmarks need to be robust and stable to avoid the jittering of the artificial effect. In a recent work [Dong et al., 2018b], the authors found that although state-of-the-art landmark detection has achieved good precision, the jittering between the video frames still raises problems for practical AR applications.

- **Dense landmark detection:** Overlay-based face editing applications usually require a denser facial landmark prediction than biometric applications, so that the artificial effects can be morphed to a finer-grained shape and placed on the proper positions. Recently, a dataset [Wu et al., 2018b] annotated in a dense format of 98 landmarks was made publicly available. Some of these landmarks are difficult to localise on face images (even for humans) because they are not so well defined semantically (e.g. on the border of the chin).

Challenges of face parsing: face parsing is rarely used in biometric applications such as face recognition. In recent times, due to the great advance of the semantic segmentation [Long et al., 2015], it has raised more and more attention in the community but several challenges remain:

- **Speed:** Most of state-of-the-art semantic segmentation methods run hardly in real-time on mobile devices, which raises a great challenge for their use in AR applications, although real-time segmentation [Zhao et al., 2018, Yu et al., 2018] has attracted more and more attention lately.

- **Lack of annotated data:** The manual pixel-wise mask annotation for face parsing is time consuming and expensive. One of the challenges for learning-based face parsing is that the amount of publicly available annotated data is much lower than for facial landmark detection tasks.

- **No pre-defined RoI (in the wild):** Specifically for human hair segmentation, a detection is required from all views of the upper-body including the back view [Muhammad et al., 2018]. In this case, the RoI in the image is more variable and not clearly determined by the face position, which may include more noise on the background.

- **Refined boundary:** [Li et al., 2017] found that most of the pixels that are difficult to classify are located on the boundaries. A fine-grained boundary is essential for rendering a good virtual effect on the faces. If the mask is coarse, the virtual effect/makeup (for example the lipstick effect) may exceed the assigned region (the lip), which will degrade the final rendering and make it looks unnatural.

1.3 Contributions and Outline of the Thesis

The structure of this thesis is shown in Fig. 1.10. Corresponding to the two research subjects mentioned before, this thesis consists of two parts.

In part I, we present three contributions for facial landmark detection:

- **A Fine-grained Facial Landmark Detection Method:** as most facial landmarks are positioned on visible boundary lines, we present an approach that improves the detection precision by training a model that encourages the detected landmarks to stay on these boundaries. We propose a new CNN that effectively exploits lower-level feature maps containing abundant boundary information. We also introduce a novel robust spatial loss function based on pixel-wise differences between patches cropped from predicted and ground-truth positions. To further improve the landmark localisation, our framework uses several loss functions optimising the precision at several stages in different ways.

- **Rethinking Robust Facial Landmark Detection:** we argue that improving the robustness of facial landmark detection model requires rethinking many aspects, including the use of datasets, the format of landmark annotation, the evaluation metric as well as the training and detection algorithm itself. To this end, we propose a new method for robust facial landmark detection using a loss function based on the 2D Wasserstein distance combined with a new landmark coordinate sampling relying on the barycenter of the individual propability distributions. Further, with the large performance increase of state-of-the-art deep CNN models, we found that current evaluation metrics can no longer fully reflect the robustness of these models and we therefore propose several improvements on the standard evaluation protocol.

- **A Facial Landmark Correlation Analysis Method:** we present a facial landmark position correlation analysis as well as its applications. Although numerous facial landmark detection methods have been presented in the literature, few of them concern the intrinsic relationship among the landmarks. In order to reveal and interpret this relationship, we propose to analyze the facial landmark correlation by using Canonical Correlation Analysis (CCA). We experimentally show that dense facial landmark annotations in current benchmarks are strongly correlated, and we propose several applications based on this analysis.

  First, we give insights into the predictions from different facial landmark detection models (including cascaded random forests, cascaded CNN, heatmap regression models) and interpret how CNNs progressively learn to predict facial landmarks. Second, we propose a weakly-supervised learning method that allows to considerably reduce manual effort for dense landmark annotation. Unlike the previous methods, we mainly focus on how to find the most efficient sparse format to annotate. Overall, our correlation analysis provides new perspectives for the research on facial landmark detection.

In part II, we present two contributions for face parsing:

- **A Two-stage Human Hair Segmentation Method:** we propose a new hair segmentation approach integrating a deep shape prior into a carefully designed two-stage Fully Convolutional Neural Network (FCNN) pipeline. First, we utilize a FCNN with an Atrous Spatial Pyramid Pooling (ASPP) module to train a human hair shape prior based on a specific distance transform. In the second stage, we combine the hair shape prior and the original image to form the input of a symmetric encoder-decoder FCNN with a border refinement module to get the final hair segmentation output. We show that our method is more robust to cluttered background.

- **A Real-time Face Parsing Model for Mobile AR Applications:** we present a demonstration of face parsing for mobile platforms such as iPhone and Android. We design an efficient FCNN in an hourglass form that is adapted to live face parsing on mobile phones. The model is implemented on the iPhone with the CoreML framework. In order to visualize the output segmentation results, we superpose a mask with false colors such that users can have an instant AR experience.

1.4 Organization

The rest of this thesis is organized in 8 chapters:

- ***Chapter 2:*** A literature review concerning the recent advances of CNN facial landmark detection methods, face parsing methods and makeup transfer methods.
- **Part I *Chapter 3:*** Contributions to fine-grained facial landmark detection.
- **Part I *Chapter 4:*** Contributions to robust facial landmark detection.
- **Part I *Chapter 5:*** Contributions to facial landmark correlation analysis.
- **Part II *Chapter 6:*** Contributions to human hair detection in the wild.
- **Part II *Chapter 7:*** Contributions to real-time face parsing model for mobile AR applications.
- ***Chapter 8:*** Conclusions and perspectives for future work.

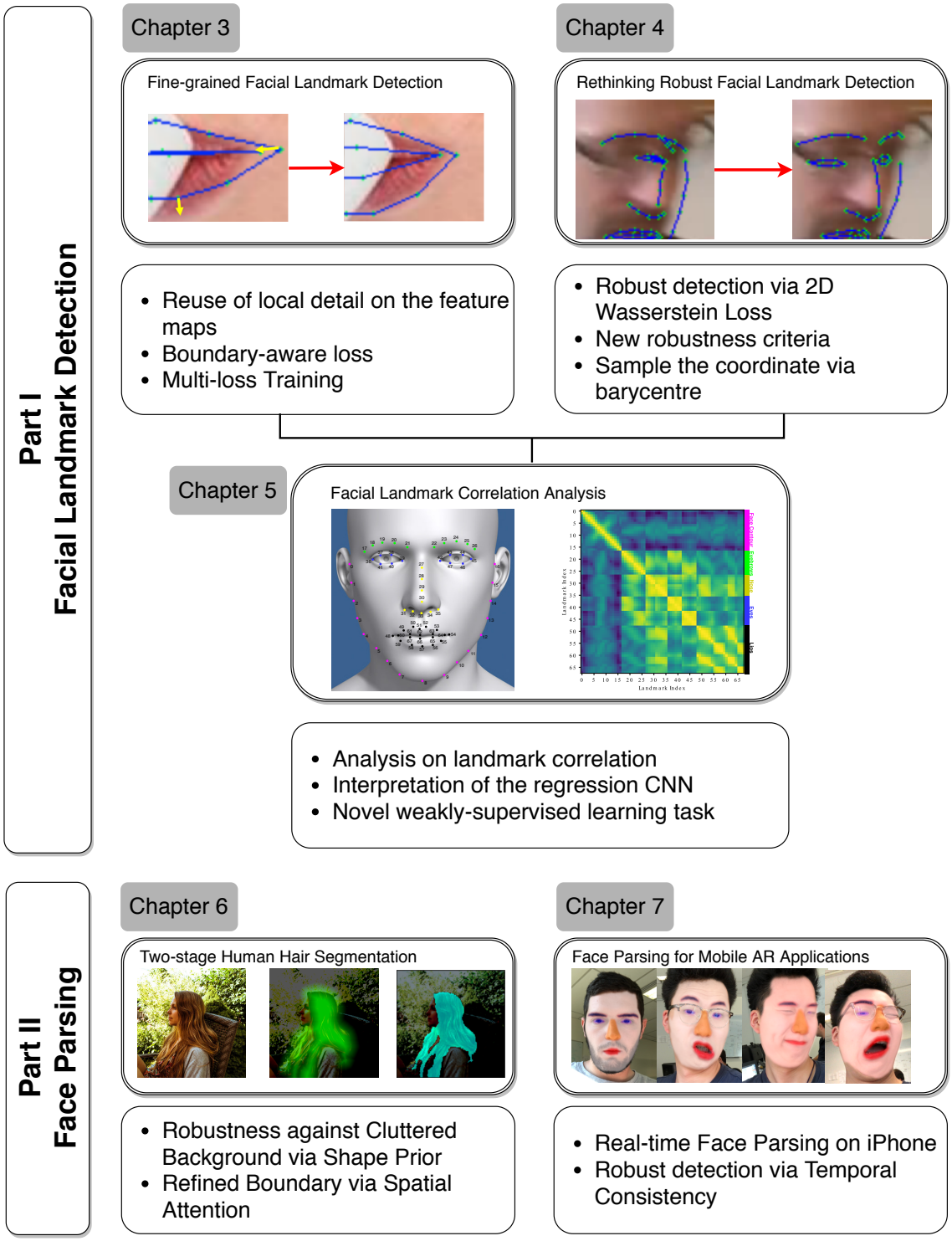


FIGURE 1.10: Graphical outline and contribution of this thesis. The contributions of each chapter are listed in bullet points. Specifically, the research in Chapter 5 is inspired by the problems raised in Chapter 3 and Chapter 4.

Chapter 2

Literature Review

This thesis is focused on deep CNN based face analysis methods, including facial landmark detection and face parsing for virtual makeup AR applications. This chapter presents a literature review of the four main subjects concerned in this thesis:

- ***Chapter 2.1:*** We present a literature review on the principles and the development of the CNN, which covers several ups and downs in its history.
- ***Chapter 2.2:*** We present a literature review on the recent advance of deep facial landmark detection.
- ***Chapter 2.3:*** We present a literature review on the recent advance of deep face parsing.
- ***Chapter 2.4:*** We present a literature review on the recent advance in makeup transfer and makeup recommendation.

2.1 Convolutional Neural Network

2.1.1 CNN before AlexNet

Multilayer Perceptron: A Multilayer Perceptron (MLP) [Rumelhart et al., 1988] is a feed-forward neural network composed of individual neurons called Perceptrons [Rosenblatt, 1961] (see Fig. 2.1). It is inspired by the structure of the human neural systems, where each neuron produces an output signal based on the signal received from other neurons. It is closely related to the connectionist approaches in cognitive science that try to explain mental phenomena using artificial neural networks. In an MLP, the neurons are organized in layers. Fig. 2.1 demonstrates a three-layer perceptron including one input layer, one hidden layer and one output layer. In each layer, all neurons are connected with all of the neurons in the previous layer. Therefore, the layer of this kind is also referred as Fully Connected (FC) layer.

In mathematical terms, we model each artificial neuron as follows (see Fig. 2.2):

$$f(\mathbf{x}) = \sigma\left(z(x)\right) = \sigma\left(\sum_i w_i x_i + b\right) = \sigma\left(w^T\mathbf{x} + b\right) \tag{2.1}$$

$\mathbf{x}$ is the input of the layer. w and b are the weight and bias parameter of this layer with i inputs. The activation function σ adds non-linearity on the output of each neuron. Several commonly used activation functions are Sigmoid function and Tanh function.

Network Training: Training artificial neural networks involves back-propagating the error from the output layer through intermediate layers. The non-linearity introduced by the activation function makes the optimization of the neural network non-convex. Therefore, the learning of the neural network is mainly based on the iterative gradient decent

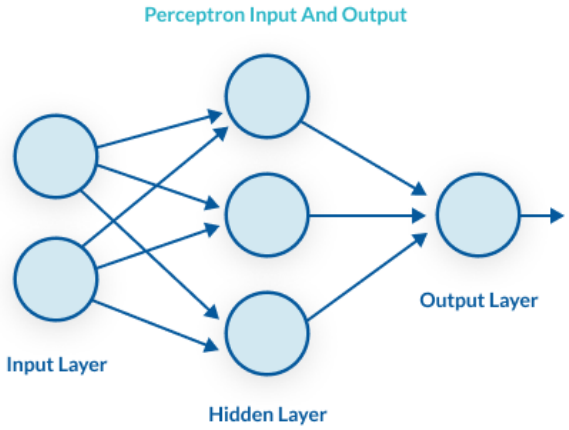


FIGURE 2.1: Structure of the Multilayer Perceptron (MLP).

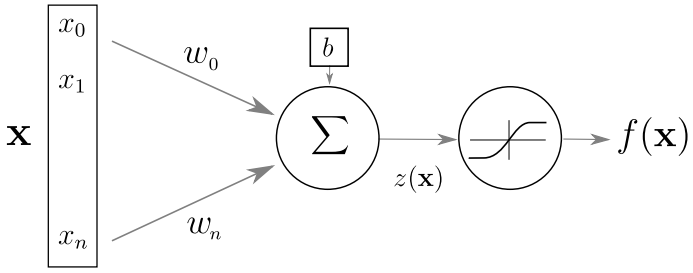


FIGURE 2.2: Mathematical model of an artificial neuron: the Perceptron

methods minimising the cost function [Rumelhart et al., 1988]. One of the most common optimization methods for neural network is Stochastic Gradient Descent [Bottou, 2010]. A simple update of stochastic gradient descent on the network (whose parameters are defined as $\boldsymbol{\theta}$, including both w and b) can be denoted as:

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \eta_k \nabla f\left(\boldsymbol{\theta}_k\right), \tag{2.2}$$

where η_k is the learning rate at step k. The parameters of the neural work can be learned after numerous iterations.

Convolutional Neural Network: An early prototype of CNN was first introduced by Fukushima [Fukushima, 1975]. In the 1990s, Lecun et al. proposed a Convolutional Neural Network (LeNet, see Fig. 2.3) [LeCun et al., 1989, LeCun et al., 1990, LeCun et al., 1998] for the recognition of hand-writing digits and trained it end to end using the gradient backpropagation algorithm, which is still commonly used nowadays.

Compared to the MLP, CNN involves convolutional layers. Convolutional layer enables the weight of the network to be shared across the spatial dimension of the input, which was proved to be beneficial for computer vision tasks due to its better generalization and efficiency. An illustration of the convolution layer is shown in Fig. 2.4. The convolution operation is executed on 2-dimensional inputs, which enables the network to learn meaningful visual features and maintain the spatial relation on the feature maps. The convolutional layers near the input learn the local details such as boundary, texture or colors. The convolutional layers near the output learn more general and abstractive information such as the presence of certain objects.

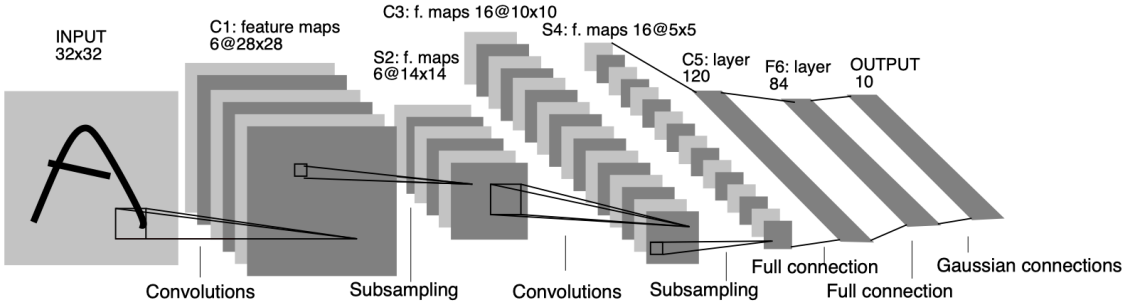
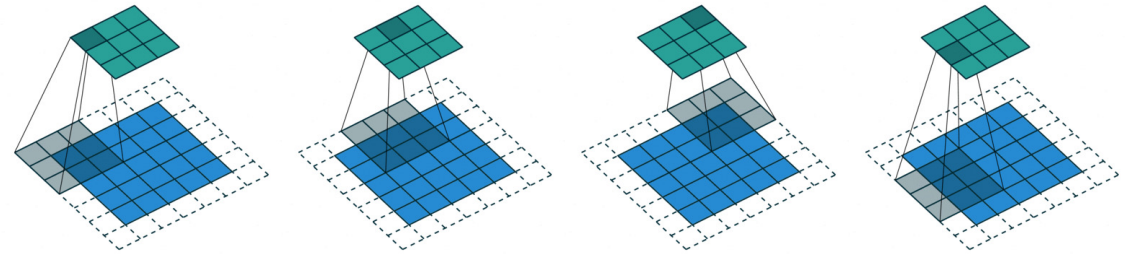


FIGURE 2.3: Structure of LeNet [LeCun et al., 1998]



(A) (B) (C) (D)

FIGURE 2.4: Demonstration of the convolution layer. Kernel size = 3, stride = 2, padding =1.

2.1.2 Early Face Analysis Methods based on CNN

Using neural networks for human face analysis can be dated back to the 1990s including face detection [Vaillant et al., 1994, Rowley et al., 1998, Feraund et al., 2001, Garcia and Delakis, 2002, Garcia and Delakis, 2004, Osadchy et al., 2007], face recognition [Lawrence et al., 1997, Duffner and Garcia, 2007], facial feature (landmark) detection [Reinders et al., 1996, Duffner and Garcia, 2005a, Duffner and Garcia, 2008] and face parsing [Low and Ibrahim, 1997].

Despite the different ideas presented in these work, most of them are based on variants of the structure of LeNet. Due to the limited data and restricted model capacity, these applications can be hardly adapted to the challenging conditions in real world such as the faces in large pose and complex lighting condition.

2.1.3 CNN after AlexNet

Since 2012, we witnessed great improvements in the use of CNN after the appearance of AlexNet [Krizhevsky et al., 2012] and the term *deep learning* has been widely used since then. The main improvements in the last few years can be listed as follows:

Architectural design: As introduced before, the early work of computer vision research using CNN are mainly based on LeNet, whose layers are no more than 6 or 7 layers. With the introduction of GPU calculation, the first thing that was changed is that the networks are becoming much bigger and deeper. At the same time, the performance is largely improved as well. Fig. 2.5 shows a comparison of the capacities and performance of different deep CNNs on the ImageNet image classification benchmark. We observe that both the model size and the number of operation are greatly augmented as the recognition accuracy is improved accordingly.

Secondly, the design of the network follows no longer the layer-after-layer pattern. Several flexible structures are introduced such as skip connections [He et al., 2016] and

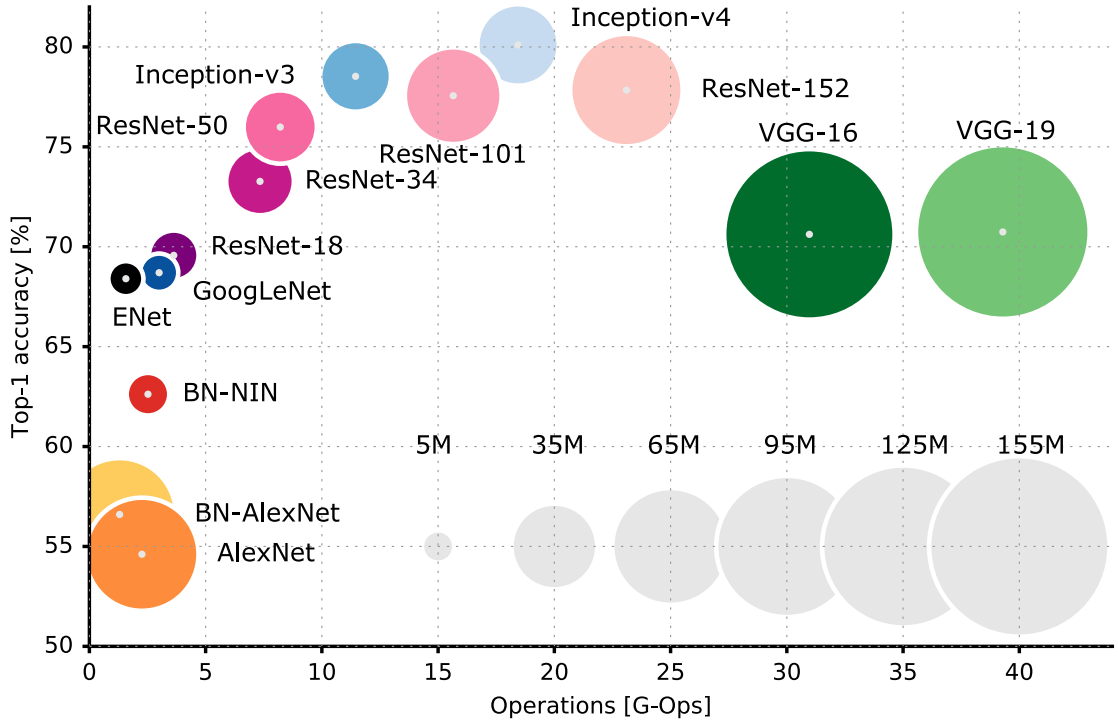


FIGURE 2.5: A comparison of the capacity and ImageNet performance of different deep CNNs.

dense connections [Huang et al., 2017] (see Fig. 2.6). Both of the design enable the gradient to flow through the network without the vanishing problem, therefore enables training very deep networks.

Thirdly, Fully Convolutional Neural Network (FCNN) was invented and has been expanded to more and more applications. FCNN does not involve the FC layers, which is able to retain spatial information on the feature maps and keep a relatively low capacity. Initially, FCNN [Long et al., 2015] was introduced for semantic segmentation. Afterwards, in the form of encoder-decoder, it has been widely and successfully applied in numerous subjects including image generation [Goodfellow et al., 2014], pose estimation [Wei et al., 2016], 3D reconstruction [Feng et al., 2018a], etc.

Regularization: More effective measures have been developed to regularize deep CNNs from over-fitting. Dropout [Srivastava et al., 2014] randomly drop units and their

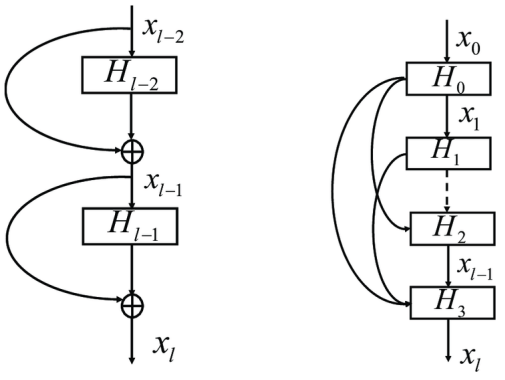


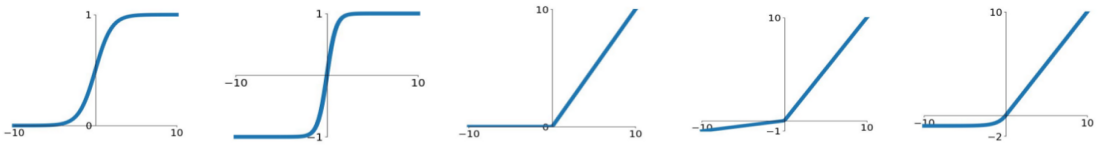
FIGURE 2.6: A demonstration of skip connection (left) and dense connection (right).

connections from the neural network during training, which prevents units from co-adapting too much. Batch Normalization [Ioffe and Szegedy, 2015] was introduced to stabilize, accelerate the training and regularize the networks by normalizing the input to the layer in each batch. Recently, Group Normalization [Wu and He, 2018] superseded the Batch Normalization when the batch size is small.

Optimizer: More efficient optimizers were invented to accelerate the training of the CNN by setting adaptive gradient such as Adam [Kingma and Ba, 2014] and RMSProp [Tieleman and Hinton, 2012]. These optimizers are able to significantly accelerate the training of the CNNs. However, it still remains controversial on the issues that such optimizers are able to always converge to similar solutions. Some of the papers [Wilson et al., 2017] claimed that the solutions found by adaptive methods generalize worse than traditional Stochastic Gradient Descent optimizer.

Activation function: Another important improvement in this field is the evolution of the activation functions. ReLU was used in the Alexnet [Krizhevsky et al., 2012] to solve the vanishing gradient problem usually confronted in the Sigmoid activated CNNs. However, since all the negative values are rectified to zero, the derivative of this function will be fixed to zero with minus zero input. The "dead ReLU problem" will appear in CNNs when some components of the network are most likely never updated to a new value. Leaky ReLU [He et al., 2015a] and ELU [Shah et al., 2016] were both introduced to alleviate this problem. From using these two activation functions, the values of the gradients are no longer stuck at zero for negative values. The mathematical formula of previously mentioned activation functions are listed as follows and visualized in Fig. 2.7.

- **Sigmoid:** $\text{Sigmoid}(z) = \dfrac{1}{1+e^{-z}}$
- **Tanh:** $\text{Tanh}(z) = \tanh(z)$
- **ReLU:** $\text{ReLU}(z) = \max\{0, z\}$
- **Leaky ReLU:** $\text{LReLU}(z) = \begin{cases} z & \text{if } z > 0 \\ \alpha z & \text{if } z \leq 0 \end{cases}$
- **ELU:** $\text{ELU}(z) = \begin{cases} z & \text{if } z > 0 \\ \alpha\left(e^{z}-1\right) & \text{if } z \leq 0 \end{cases}$



(A) Sigmoid (B) Tanh (C) ReLU (D) Leaky ReLU (E) ELU

FIGURE 2.7: Demonstration of different activation functions in deep CNNs.

2.2 Facial Landmark Detection

2.2.1 Overview

Facial landmarks, also known as facial feature points or fiducial facial points, play an important role in facial expression analysis [Martinez et al., 2017], 3D face reconstruction [Jackson et al., 2017], face recognition [Ding and Tao, 2016] and other related applications. The detection of facial landmarks is aimed to retrieve the coordinate of each facial

landmark given a face image. There already exists several comprehensive surveys for facial landmark detection [Çeliktutan et al., 2013, Yang et al., 2015b, Wang et al., 2017b, Jin and Tan, 2017, Yan et al., 2018, Wu and Ji, 2019] and facial landmark tracking [Chrysos et al., 2014]. Generally, facial landmark detection algorithms can be categorized into two types, generative model-based methods and discriminative model-based methods.

Generative models include part-based generative models such as ASM [Cootes et al., 1995, Cristinacce and Cootes, 2007] and holistic generative models such as AAM [Cootes et al., 2001, Edwards et al., 1998, Tzimiropoulos and Pantic, 2013, Alabort-i Medina and Zafeiriou, 2014]. Generative models represent the facial shape and facial texture as generative probabilistic distributions.

Discriminative models are far more common in the literature due to its capacity to model more complex distributions and their robustness to noise in unconstrained conditions. In the last decade, the discriminative cascaded regression model [Xiong and De la Torre, 2013, Cao et al., 2014, Asthana et al., 2014, Ren et al., 2014, Burgos-Artizzu et al., 2013, Zhu et al., 2015, Kazemi and Sullivan, 2014] has become popular thanks to its excellent performance and relatively low run-time. A cascaded regression model usually consists of three important parts, the initial shape s_0, the cascaded regressors R and the shape-indexed feature extraction function ϕ. To fit the model, the shape is iteratively updated stage by stage from the initial shape. At each stage t, the shape s^t is updated by:

$$s^t = s^{t-1} + R^t \phi(I, s^{t-1}) \tag{2.3}$$

where I is the input image. In Fig. 2.8, we show how the prediction is evolved stage by stage in the cascaded regression model.

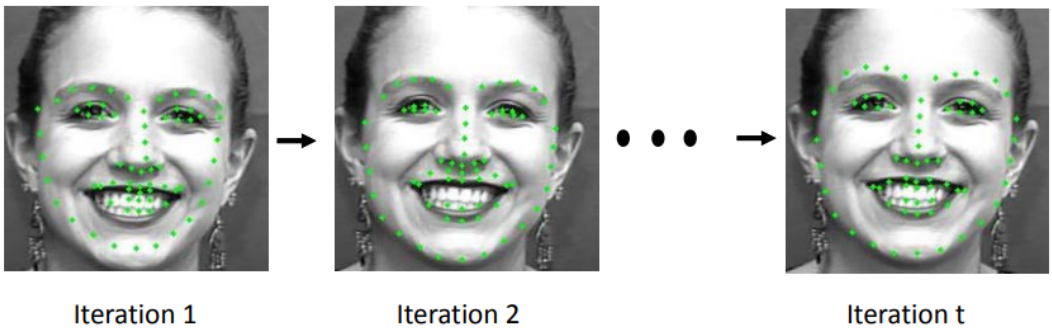


FIGURE 2.8: Shape evolution of the cascaded regression models in each stage. This figure is adopted from [Wu and Ji, 2019]

Apart from *2D facial landmark detection* on single images, more and more researchers have shown interests in closely-related subjects of *video face alignment* and *3D face alignment*. Video face alignment generally tackles the challenge of facial landmark detection in consecutive video frames of the same person by exploiting temporal continuity and identity-specific features. The objective of 3D face alignment is to predict facial landmarks in arbitrary poses, aiming to recover the projected 3D locations of invisible facial landmarks given a 2D image.

2.2.2 Deep 2D facial landmark detection

In this section, we present the recent advances on the deep CNN-based 2D facial landmark detection methods.

Sparse facial landmark detection: From 2013, there are several attempts to retrieve the positions of five or six key facial landmarks located on the eyes, nose and mouth by using deep CNN. Sun et al. [Sun et al., 2013] proposed to use a cascaded coarse-to-fine CNN to detect five essential facial landmarks. A 3-stage framework is adopted and several CNNs are included in each stage. The CNNs in the first stage estimate the rough positions of several different sets of landmarks. Each landmark is then separately refined by the CNNs in the following stages. Despite its innovation and high accuracy, this method is not completely end-to-end since the input of the following CNN depends on the local patches extracted from the previous one. The TCDCN method [Zhang et al., 2014b] adopted multi-task learning to optimize the performance of 5-point facial landmark detection. They proved that auxiliary facial attributes such as gender and pose can be helpful for the detection by providing additional information. Afterwards, they proposed to use representation transfer learning to predict a denser facial landmark format [Zhang et al., 2016c].

Kumar et al. [Kumar et al., 2016] showed that the local patch features extracted by a CNN work well with a linear regressor to give a five-point prediction. Zhang et al. [Zhang et al., 2016b] fine-tuned a pre-trained CNN to extract local facial patch features followed by a cascaded regressor predicting the facial landmarks. Both of them proved that a CNN can act as a good feature extractor in the conventional cascaded regression framework.

Dense facial landmark detection: The following works are focused on dense facial landmarks, i.e. landmarks that are not necessarily semantic but can also be part of a contour. Zhang et al. [Zhang et al., 2014a] proposed to use a coarse-to-fine approach for simultaneously detecting 68 facial points (see Fig. 2.9). They proposed a 4-stage cascaded encoder-decoder network with increasing input resolution at different stages, which processes progressively higher resolution images. The first auto-encoder is assigned to provide an initialization by taking the entire face image as input. The following auto-encoders are designed for refinement, which take the regional patches cropped by the output of the last auto-encoder to get a shape update. The landmark positions are updated at the end of each stage by the CNN output.

Afterwards, this method was further developped by adding an occlusion-recovering auto-encoder to reconstruct the occluded facial parts [Zhang et al., 2016a]. The occlusion-recovering auto-encoder network was designed to reconstruct the hidden genuine face appearance, and trained on a synthetic randomly occluded dataset.

Sun et al. [Sun et al., 2015] used a Multilayer Perceptron (MLP) as a graph transformer network to replace the regressors in a cascaded regression framework for the detection of the facial landmarks. They proved that this combination could be trained by back-propagation.

Wu and Ji [Wu and Ji, 2015a] used a 3-way factorized Restricted Boltzmann Machine [Hinton, 2002] to build a deep face shape model for the dense prediction of 68 facial landmarks. Given a facial image, they iteratively generated the measurements with the independent local point detectors and refine the measurements through inference in the face shape model. This method combines the top-down information from the embedded face shape patterns and the bottom-up measurements from the local point detectors in one unified model.

One disadvantage of using a single CNN to directly perform a dense prediction is that the neural network is trained to minimise its error on the global shape, which could possibly leave some large local imprecisions for specific facial points. A straight-forward idea is to refine different facial parts locally and independently in a post-processing step. Following the original idea of [Duffner and Garcia, 2005b], two approaches [Fan and Zhou,

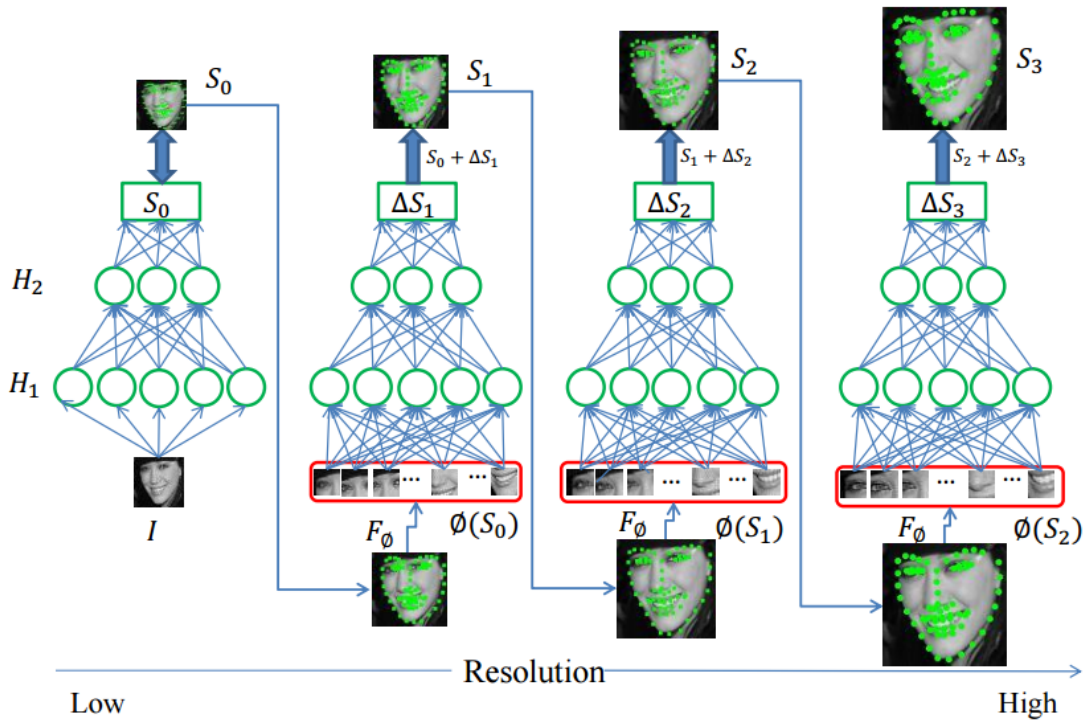


FIGURE 2.9: Network design of CFAN [Zhang et al., 2014a].

2016, Huang et al., 2015] are proposed to predict dense facial landmarks by first estimating rough positions using a global CNN followed by several small regional CNNs refining the estimates on different parts locally. This kind of structure is more time-consuming but can significantly improve the precision.

Another work by Lv et al. [Lv et al., 2017] proposed to use a two-stage re-initialization with a deep neural network regressor in each stage (See Figure 2.10). The framework consists of a global stage, where a coarse face landmark shape is predicted and a local stage, where landmarks of each facial part are estimated respectively and locally. One of the innovation is that the global/local transformation parameter is estimated by a CNN to reinitialize the facial region to a canonical shape, which largely improves the performance on large poses.

Despite the fact that deep learning-based methods are robust to different shapes or pose initializations, large head pose still remains a big challenge. Wu et al. [Wu et al., 2017b] proposed a specific structure added at the end of a vanilla neural network from TCDCN [Zhang et al., 2014b], where different branches are aimed at regressing shapes in different head poses. Kumar and Chellappa [Kumar and Chellappa, 2018] proposed to disentagle the 3D pose in the CNN to realize the 2D face alignment in unconstrained large poses. Valle et al. [Valle et al., 2018] propose to regularize the facial landmark detection under large pose by mapping on canonical 3D face models.

Several recent work also tried to boost the efficiency. Miao et al. [Miao et al., 2018] proposed to execute the landmark regression by Fourier feature pooling and low-rank learning layers to boost the detection speed. Yue et al. [Yue et al., 2018] proposed a similar structure with intermediate attention supervision. Feng et al. [Feng et al., 2018b] proposed a novel wing loss to train a light-weight 2-stage CNN to focus on the small errors.

We find interesting ideas in the three following work:

Trigeorgis et al. [Trigeorgis et al., 2016] adopted a cascaded regression-like method

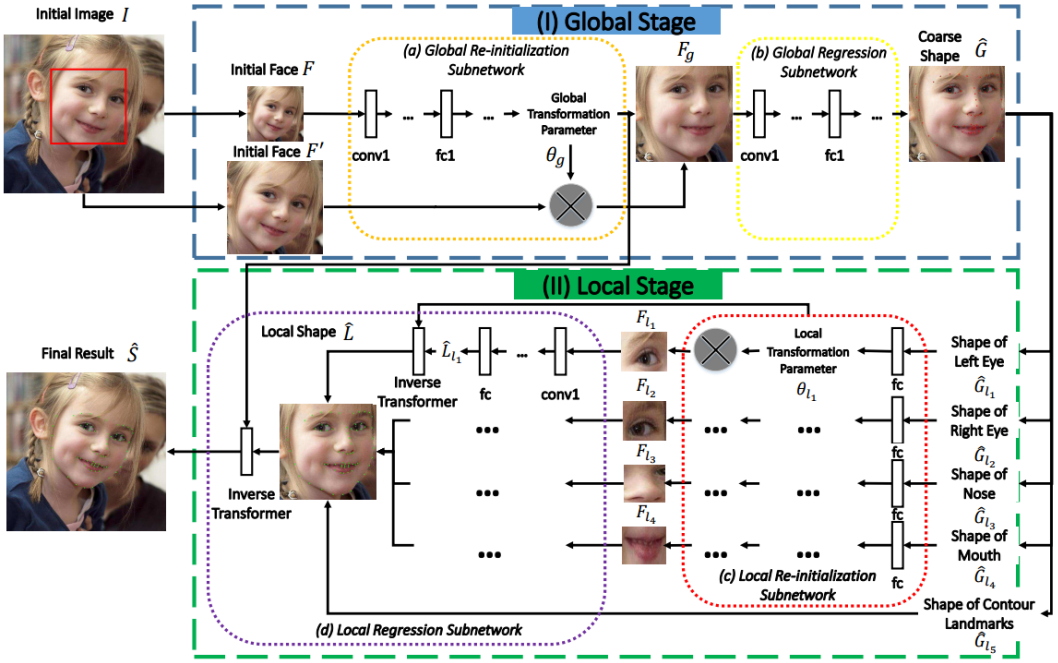


FIGURE 2.10: Network design of Lv et al. [Lv et al., 2017]. The transformation parameters are integrated into the framework in both global and local stage to reinitialized the region to a canonical shape.

with a Recurrent Neural Network (RNN) called Mnemonic Descent Method (MDM) (see Fig. 2.11). In the MDM network, the CNNs are used to extract the patch features instead of traditional hand-crafted feature extractors such as SIFT [Lowe, 2004] in SDM [Xiong and De la Torre, 2013]. In addition, they introduce RNNs as memory units to share information across the different cascade levels. The recurrent module facilitates the joint optimization of the regressors by assuming that the cascades form a non-linear dynamical system.

Shao et al. [Shao et al., 2016] proposed to use adaptive weights of different landmarks during different phases of the training. They gave a relatively bigger coefficient to some important points such as eye corners and mouth corners at the beginning of the training process and then reduced their weights later on. This operation enables the neural network to first learn a robust global shape, and locally-refined predictions afterwards.

Zhu et al. [Zhu et al., 2019a] proposed a deep network structure to alleviate the occlusion problem. They utilized adaptive weight on high-level features to reduce the impact of occlusion and obtain clean feature representations. The structure was coupled with a low-rank learning module. Being sensitive to the global geometry, their structure is able to provide a more robust detection.

Heatmap regression: Another fast-developing approach is training the CNN to predict heatmaps (also known as response maps, probability maps, voting maps or likelihood maps) as the network output by using FCNN. The value of each pixel on the likelihood maps could be represented as the probability of the existence of each facial landmark at the given position. This idea was first introduced by [Duffner and Garcia, 2005a] in 2005 (see Fig. 2.12).

In 2016, Zadeh et al. [Zadeh et al., 2016] proposed to use deep CNNs to produce a local response map and then fit a Constrained Local Model on it. Since the deep encoder-decoder could establish an image-to-image mapping, Lai et al. [Lai et al., 2016] used a fully convolutional neural network to predict an initial face shape instead of the mean

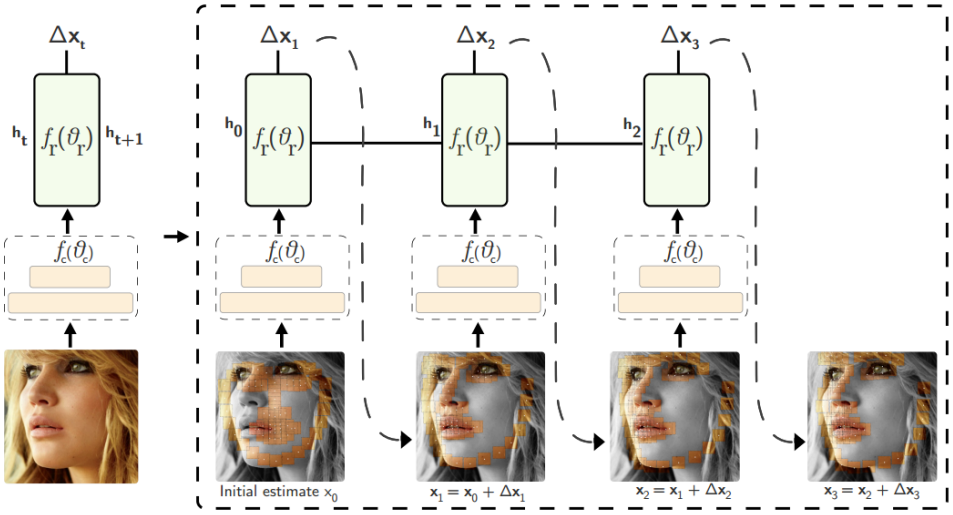


FIGURE 2.11: Network design of MDM [Trigeorgis et al., 2016]. The landmark positions are updated after each network cascade which is initialized as mean shape x_0 in the first stage. The input of feature extraction network $f_c(\theta_c)$ are the patches extracted based on the shape of last cascade $x_{n-1} + \Delta x_n$. The mnemonic module $f_r(\theta_r)$ is implemented as RNN which generates a new state h_{t+1} and a new set of descent directions Δx_{t+1} that indicates where the network should focus next. The information of the mnemonic modules are shared by hidden state h_t.

face shape which is commonly used in cascaded regression [Xiong and De la Torre, 2013, Cao et al., 2014]. They introduced "Shape-Indexed Pooling" as a feature mapping function to extract local patch features of each point, which is then given to the regressor. In their first version, they used a fully-connected layer to sequentially regress to the final shape while replacing it with a recurrent neural network in their second version inspired by MDM [Trigeorgis et al., 2016].

In the work of Xiao et al. [Xiao et al., 2016], an attention mechanism was used, where landmarks around the attention center are subjected to a specific refinement procedure. Wang et al. [Wang et al., 2017a] proposed an approach to detect multi-face landmarks by likelihood maps. With the help of an RoI pooling branch [Girshick, 2015], face detection is not required and non-face activation is eliminated in the global likelihood maps.

Another interesting work is proposed by Kowalski et al. [Kowalski et al., 2017] which predict the transformation to a canonical pose and a feature image simultaneously with global point likelihood maps, in a cascaded manner. The neural networks in different stages share the information by receiving the transformation parameters from the preceding stage.

In the past several years, the heatmap regression models have become a method that dominates the state-of-the-art performance. Bulat et al. [Bulat and Tzimiropoulos, 2017b, Bulat and Tzimiropoulos, 2017a] proposed to use the stacked Hourglass Model [Newell et al., 2016] as a heatmap regression model for facial landmark detection.

Dong et al. [Dong et al., 2018a] proposed a style-aggregated face generation module coupled with a heatmap regression model to predict robust results on large variance of image styles. The key idea is to develop an unsupervised data augmentation methods, which is able to apply distinct style (including gray scale/color, light/dark, intense/dull etc.) change on the training images.

Merget et al. [Merget et al., 2018] proposed a fully-convolutional local-global context network, which introduces a more global context in the heatmap regression model. One

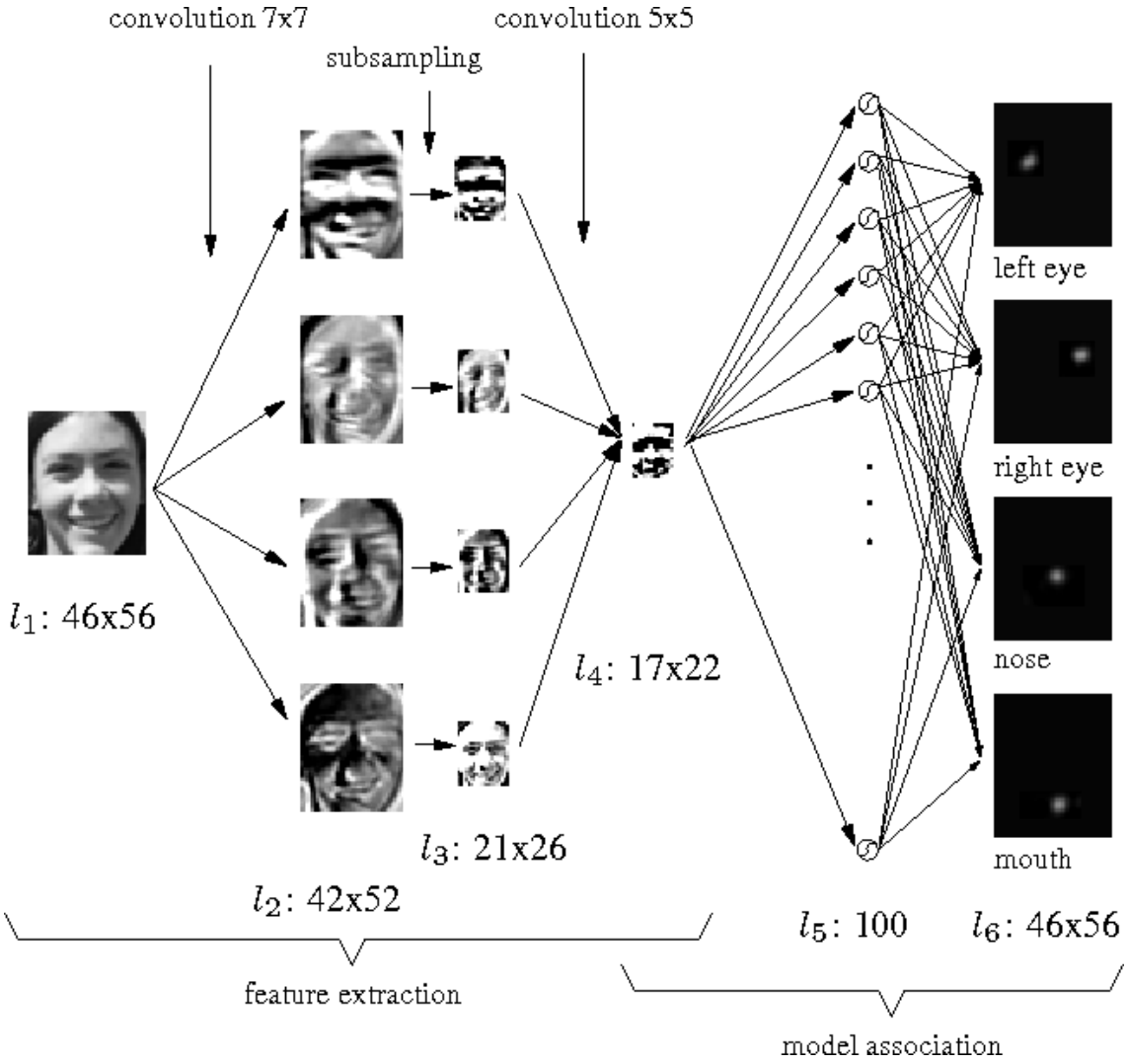


FIGURE 2.12: Network design of the first heatmap regression model for facial landmark regression from [Duffner and Garcia, 2005a]. The output of the network is modeled as a heatmap containing a Gaussian distribution on the landmark position.

advantage of this method is that this method does not require face detection as a pre-precessing step.

Tang et al. [Tang et al., 2018] proposed quantized densely-connected U-Nets to significantly accelerate the inference of the heatmap regression models. In their network, not only the parameters but also the gradients are quantized.

Wu et al. [Wu et al., 2018b] proposed to predict the boundary of the face and facial components on the heatmap rather than a Gaussian distribution of a landmark, which increases the model sensitivity to the boundary. Liu et al. [Liu et al., 2019] also proposed a method to improve accuracy of the detection by searching the real ground truth position along the boundary.

Compared to learning the boundary explicitly, Dapogny et al. [Dapogny et al., 2019] proposed to integrate landmark-wise attention maps with a cascaded heatmap regression model. The attention maps resembles the boundary map. Their method is able to learn the boundary in an end-to-end manner without explicitly training the boundary as a target.

[Chen et al., 2019b] and [Wang et al., 2019a] both concerned the uncertainty on the Gaussian distribution of each landmark. [Wang et al., 2019a] proposed a novel Weighted Loss Map, which assigns high attentions on the pixels around the center of the Gaussian distribution. It helps the training process to be more focused on the pixels that are crucial to landmark localization. [Chen et al., 2019b] introduced the Kernel Density Deep Neural Network that produces target probability map, without assuming a specific parametric distribution such as Gaussian distribution.

Zou et al. [Zou et al., 2019] concerned the structural information in the heatmap regression models. To obtain robust landmark prediction, Zou et al. [Zou et al., 2019] proposed to add a structural constraint based on Hierarchical Structured Landmark Ensemble.

Recently, HRNet [Sun et al., 2019b] superseded most of the state-of-the-art methods by addressing the importance of the high-resolution heatmap. The performance of the HRNet has "saturated" several commonly used benchmarks.

2.2.3 Video facial landmark detection

Video facial landmark detection, also referred to as sequential face alignment, aims at detecting the facical landmarks in a sequence of consecutive images by leveraging continuous information in the video. It is crucial for face AR applications due to the fact that most of the AR applications are running in real time.

Person-specific modeling is a direct idea for video face tracking since personal identity information remains unchanged. Incremental learning (also known as sequential learning or online learning) is used by Sung et al. [Sung and Kim, 2009] to track the facial landmarks with an AAM [Cootes et al., 2001] model. Chrysos et al. [Chrysos et al., 2015] proposed an off-line tracking pipeline to reinforce the tracking robustness in speech videos. Other than general face detection and alignment, they implemented a person-specific face detection using a Deformable Part Model [Felzenszwalb et al., 2008] and a person-specific generative landmark localizer. It iteratively updates the generic/person-specific appearance variations and shape/appearance parameters in turn. This method is also used to semi-automatically annotate the 300VW [Shen et al., 2015]. A pipeline of this method is shown in Fig. 2.13.

Asthana et al. [Asthana et al., 2014] reformulated the cascaded regression in a parallel form to enable fast and efficient learning for each cascade level. The information retained by the suceeding cascade level is derived from the statistical distribution from the preceding cascade regressor. In CCR [Sánchez-Lozano et al., 2016] and iCCR [Sánchez-Lozano

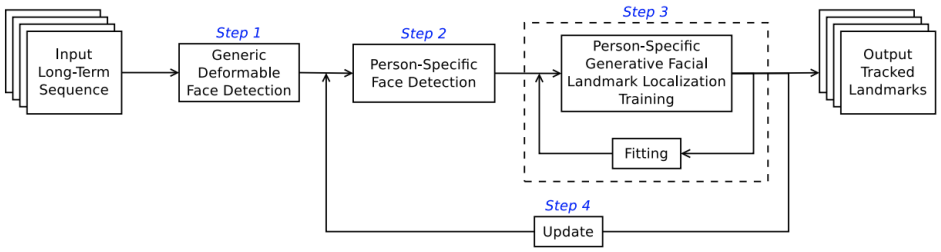


FIGURE 2.13: Pipeline of [Chrysos et al., 2015] for video face detection and facial landmark detection, which was further used to annotate the 300VW dataset [Shen et al., 2015] semi-automatically. Step 1: obtain an initial estimation of the shapes with minimal false positive detections; Step 2: train a person-specific face detector; Step 3: train a person-specific generative deformable model; Step 4: iterate Step 2 and Step 3 to gradually improve the results.

et al., 2017], Sanchez et al. implemented a continuous regression method while reformulating it such that it does not require sampling over the perturbed shapes (e.g. flipping, rotation, scaling etc.). As a result, the computational cost is largely reduced compared to the traditional cascaded regression-based methods.

Since Bayesian filters are popularly used in object tracking, a direct idea is to combine them with state-of-the-art landmark localizers. For example, Prabhu et al. [Prabhu et al., 2010] used a Kalman filter to track the facial landmarks by estimating the head position/orientation and the facial shapes in video sequences.

At the 300VW workshop [Shen et al., 2015], a challenging dataset specifically for tracking facial landmarks has been proposed. One of the best-performing methods, implemented a pose-specific cascaded regression method [Yang et al., 2015c] and another adopted a progressive initialization [Xiao et al., 2015] which aims to resolve the problem of bad initialization in large poses.

Additionally, Jeni et al. [Jeni et al., 2015] proposed a real-time 3D face alignment method on the video by dense cascaded regression. The points predicted are no longer feature landmarks but points scattered on the face. A 3D fitting is realized based on these dense points to reconstruct a 3D mesh of the face.

RNN is frequently used to model temporal information in the videos. Inspired by the incremental learning method [Peng et al., 2016b], RED-Net [Peng et al., 2016a] was proposed to optimize the performance of video face alignment by disentangling the identity information and pose/expression information. The identity information is considered to be invariant in the video while pose and expression information changes over time. The author proposed a dual-path neural network using point likelihood maps, which include one path to extract the identity information and the other path to learn the pose/expression information by using a RNN.

Gu et al. [Gu et al., 2017] proposed to integrate a one-layer RNN at the end of a VGG network to track facial landmarks as well as head poses. They proved that Bayesian filters could be formulated as a linearly-activated RNN without bias. According to their results, tracking with RNNs is more accurate and stable than both frame-by-frame detection and state-of-the-art landmark localizers combined with Kalman filter.

Another algorithm using RNNs called TSTN was proposed by Liu et al. [Liu et al., 2017a], which adopts two neural network branches: spatial and temporal. The spatial branch learns the face shape updates by local face patches, which are then used to refine the current facial shape based on the previous shape. The temporal branch is designed as a deep encoder-decoder with a two-layers RNN at the centre to capture facial dynamics across the temporal dimension. This branch takes consecutive frames as input and then

renders the temporal shape update. The final shape is determined by a weighted fusion of the shape updates from the two branches.

Hou et al. [Hou et al., 2018] used a Long-Short-Term Memory (LSTM) module to guide the spatial estimation for the next stage just as MDM and simultaneously helps the estimation in the following frame.

Dong et al. [Dong et al., 2018b] aimed at providing robust and consistent landmark prediction across the frames by leveraging optical flow. The main advantage of this method is that this method enables unsupervised learning on unlabeled videos.

Tai et al. [Tai et al., 2019] proposed Fractional Heatmap Regression to achieve extremely accurate facial landmark positions on the videos. Additionally, they also propose a novel loss to make the prediction between different frames more stable and consistent.

Belmonte et al. [Belmonte et al., 2019] proposed to include motion information to achieve more stable predictions over time and more robustness to variations. Unlike [Gu et al., 2017] and [Hou et al., 2018] who use RNN, they propose to replace the 2D convolution layers with 3D convolution layers, by processing an input of three consecutive frames.

Sun et al. [Sun et al., 2019a] propose a face deblurring network which cooperates with landmarks detector work as a virtuous circle to obtain robust landmark predictions, against the noise of motion-blur on the videos.

2.3 Face Parsing

Deep face parsing approaches emerged in 2012 from the work by Luo et al. [Luo et al., 2012], who proposed to detect the facial landmarks by using a deep belief network based on the face parsing segmentation results. Their hierarchical face parsing framework includes four parts, the face detector, the face part detector, the components detector and components segmentators. They model different layers under a Bayesian framework with a spatial consistency prior between the layers. Each layer is pre-trained in an unsupervised manner using layer-wise Restricted Boltzmann Machines and fine-tuned for classification using logistic regression. The segmentators are trained as deep auto-encoders to obtain face parsing results which is robust to the occlusion. This method can be used for facial landmark detection as well. Facial landmarks are obtained by fitting the mean shape in Fig. 2.14 to parsed label map.

Liu et al. [Liu et al., 2015] combine CNNs and Conditional Random Fields (CRF) for face parsing by exploiting rich features from CNNs and structured output from the CRF. They proposed two distinct loss functions: one encoding the unary label likelihoods and the other encoding the pairwise label dependencies. Given both of the unary and pairwise term, GraphCut algorithm is used for efficient inference.

Yamashita et al. [Yamashita et al., 2015] proposed to use a weighted loss function for certain classes, so that some small yet important regions such as eyes can be significantly improved.

Zhou et al. [Zhou et al., 2015] proposed Interlinked Convolutional Neural Networks for face parsing. Interlinked Convolutional Neural Networks utilize the multi-scale feature maps to more efficiently integrate local and contextual information. The method consists of two stages. The first stage gives a holistic prediction. The second stage is used for refined prediction based on the image patches.

Saito et al. [Saito et al., 2016] introduced a real-time face segmentation pipeline. They proposed a carefully designed datasets and several data augmentation strategies to handle challenging occlusions such as hands on faces. The efficiency of the segmentation

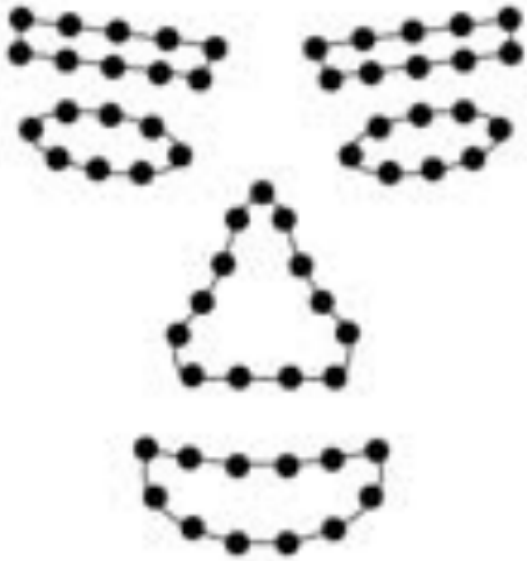


FIGURE 2.14: Face parsing regions in [Luo et al., 2012]. The region is used to (1) fit the parsed label map and (2) correspond the parsed label map with facial landmarks.

network is improved by using a two-stream deconvolution networks and a shared convolution network.

Liu et al. [Liu et al., 2017d] proposed a spatially variant RNN as a refinement module for face parsing. This lightweight RNN propagates on up/down/left/right directions to generate guidance of semantic edges, which significantly improves the output prediction.

Wei et al. [Wei et al., 2017] introduced a novel approach to regulate receptive field for face parsing. A novel affine transformation layer was proposed to enlarge or shrink feature maps by derivable interpolation algorithms. They demonstrated that by expanding the receptive field of the CNN, the face parsing prediction can be improved. A following work [Wei et al., 2019] further improves the performance by introducing concept of Normalized Receptive Field as well as a novel loss called Statistical Contextual Loss. Statistical Contextual Loss integrates richer contextual information and regularizes features during training.

Wang et al. [Wang et al., 2019b] proposed Convolutional LSTM (ConvLSTM) for face parsing on the video. By using LSTM, the model is capable of learning temporal consistency on a sequence of frames. Additionally, a segmentation loss was proposed to directly optimize the Intersection over Union (IoU) performances.

Recently, Lin et al. [Lin et al., 2019] proposed to use a novel RoI Tanh-warping operation to first align the whole image rather than simply crop it. For each sub-region, unlike the previous method, they directly crop on the feature maps in the CNN via RoI Align for further refinement. Finally, the output mask given by the CNN is warped back to the original image based on the inverse transform of the RoI Tanh-Warping transformation (see the whole pipeline in Fig. 2.15).

2.4 Makeup Transfer and Recommendation

Applying a specific makeup style automatically on a human face is an interesting and emerging subject in computer vision. The subject had been growing very rapidly in the last several years.

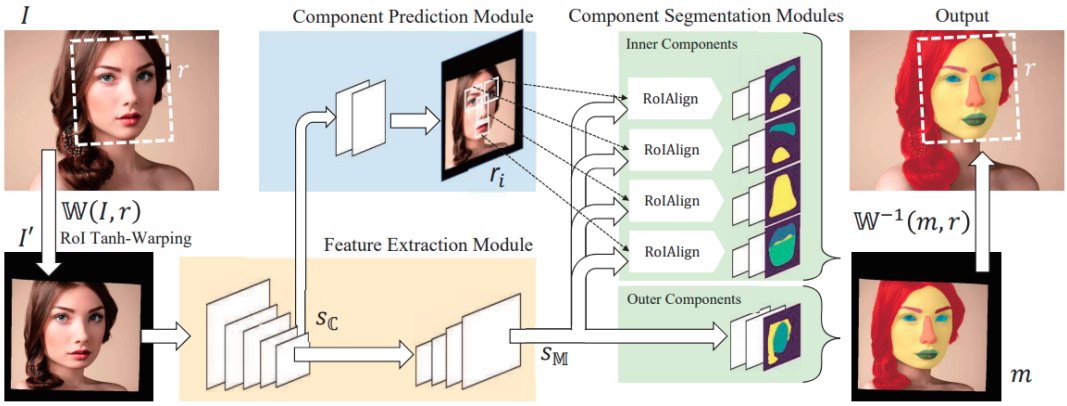


FIGURE 2.15: Pipeline of the face parsing method proposed in [Lin et al., 2019].

Tong et al. [Tong et al., 2007] first proposed to automatically apply the face makeup given an example image. For the first time, the concept of makeup transfer is introduced. This makeup transfer is useful as the following scenario happens frequently: a customer enters a beauty salon, selects an example image from a catalog and tells the makeup artist to apply the same makeup on her face. They propose to learn from an image pair of "before" makeup and "after" makeup. The makeup effect can be represented as the ratio of the intensity between the "before"/"after" pairs. To apply this makeup effect, this ratio is multiplied on the target face.

The first complete automatic facial makeup system including face parsing appeared in 2009 [Dhall et al., 2009]. They used fuzzy learning based skin segmentation and Haar based facial feature extraction to first define the RoI (the skin and the lip region). The digital makeup that they applied on the RoI is relatively simple. First, they applied a Gaussian smoothing and morphological dilation on the target image to remove the noises. Afterwards, they proposed to apply the foundation effect and the lipstick effect by changing the color of the RoI in RGB and HSV color spaces.

In [Guo and Sim, 2009], the authors decompose the face into three layers: face structure layer, skin detail layer and color layer for makeup transfer. They think that the makeup effect only exists in the color layer and skin detail layer. Therefore, the color and skin effect can be transferred accordingly while the general face structure on the target image remains unchanged.

Xu et al. [Xu et al., 2013] automatize the previous work of [Guo and Sim, 2009] by using ASM [Cootes et al., 1995] face landmark detection algorithm to locate key points and Gaussian mixture models to adjust skin areas.

Liu et al. [Liu et al., 2014] proposed an approach that not only simulates the makeup effect but also recommends the makeup style. This paper also concerns the simulation of the hair style as well. The synthesis of the makeup and the hair style are mainly achieved by aligning the virtual effect on the corresponding RoI with color modification in the CIELAB color space.

Li et al. [Li et al., 2015] addressed the problem of mistakenly transferring the lighting conditions and personal details on the example image as a makeup style. They proposed to separate the images into intrinsic image layers and transfer the makeup effect via adaptations of physical reflectance models. Using their method, the lighting effect and appearance characteristics on the target image is preserved, which gives a more realistic makeup transfer effect on the target image.

Wang and Fu [Wang and Fu, 2016] concerned the inverse problem of the makeup transfer: makeup removal. The makeup removal is useful for makeup transfer when

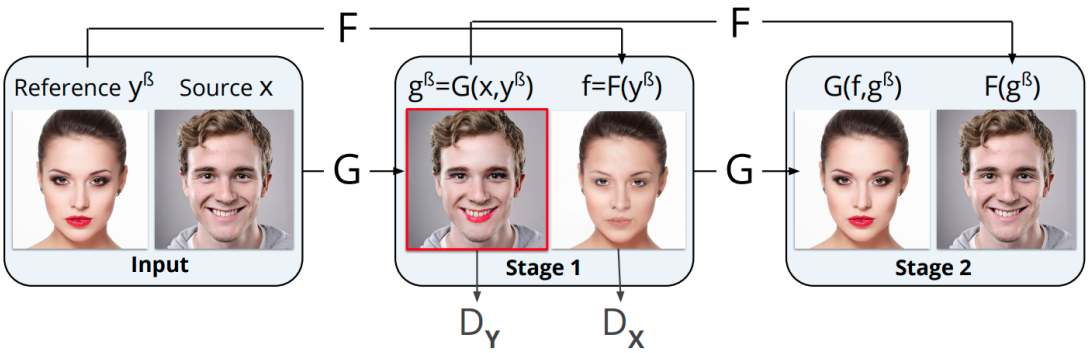


FIGURE 2.16: Pipeline of the method proposed in [Chang et al., 2018].

the client has already worn makeup. The authors first proposed a method to detect the makeup on the faces. For synthesizing the face with no makeup effect, they used a coupled locality-constrained dictionary learning framework to erase the makeup according to its style.

The use of deep CNN in this domain started from 2016. Liu et al. [Liu et al., 2016] proposed an end-to-end makeup transfer method as well as a makeup recommendation system based on deep CNN. They consider the makeup effect as an artistic style, which can be transferred by CNN based Style Transfer [Gatys et al., 2016] method. The applied makeup effect can be generated arbitrarily from light to heavy by changing the weight of the style coefficient.

Alashkar et al. [Alashkar et al., 2017] proposed a CNN based system for makeup recommendation. In order to validate their approach, they proposed a dataset including the knowledge rules given by professional makeup artists. Nguyen et al. [Nguyen and Liu, 2017] also proposed a similar recommendation system nonetheless based on latent SVM.

Chen et al. [Chen et al., 2017b] proposed a novel task to restore the original portrait image from an image with makeup effect already applied, without any prior knowledge on the details of the beautification operation. They mainly focus on two kinds of makeup effect: skin color change and skin smoothing. Their proposed Component Regression Network is capable of restoring smoothed wrinkle and very subtle freckle.

A critical difficulty in the dataset construction for make up transfer is the lack of paired samples. From the Internet, it is easy to retrieve large amount of images with makeup or without makeup. However, obtaining "before"/"after" makeup image pair on the same person under constrained lighting conditions usually requires expensive and time-consuming process. Chang et al. [Chang et al., 2018] proposed a novel GAN-based unsupervised framework (see Fig. 2.16). In their framework, the makeup transfer operation G and the makeup removal operation F are learnt altogether. In the first stage, G first applies the makeup effect on the source image and F removes the makeup effect on the reference image. In the second stage, F removes the makeup effect on the source image which is previously applied by G. Similarly, G applies the makeup effect on the reference image whose makeup effect was removed by F in the first stage. The output from the second stage can be used to ensure the identity preservation and style consistency on the input images.

Li et al. [Li et al., 2018] proposed a similar framework with [Chang et al., 2018]. Additionally, they proposed a local instance-level loss based on histogram matching to ensure

that the makeup styles applied between the source image and the target image are similar.

Instead of using GAN based generative model, Chen at al. [Chen et al., 2019a] proposed to use Glow [Kingma and Dhariwal, 2018] model to decompose the image into two latent features, the face feature and the makeup feature. Upon reconstruction, the weight of the makeup feature can be adjusted to change the intensity of the makeup effect.

Ren et al. [Ren et al., 2019] proposed a method to transfer different makeup styles to the same target and get the appropriate makeup lightness of the makeup effect.

Jin et al. [Jin et al., 2019] proposed to introduce illumination transfer in the makeup transfer so that the dark and white facial makeup could be effectively transferred. The transfer process is efficient, which can be finished within 1 second.

Gu et al. [Gu et al., 2019] leveraged additional multiple overlapping local discriminators to transfer dramatic makeup, i.e. complex makeup styles with high-frequency details.

Part I

Facial Landmark Detection

In the first part, we focus on two challenges of facial landmark detection:

- **Local precision in Chapter 3**: as mentioned in the Chapter 1, the local precision is of great importance for overlay-based face AR applications. Even a slight displacement of the virtual effect is able to largely degrade the user experience. To this end, we propose a novel CNN model that effectively exploits abundant boundary information on the low-level feature maps, as well as a novel robust spatial loss to force the predicted landmarks to stay on the semantic facial boundary.
- **Robustness in Chapter 4**: the robustness of facial landmark detection guarantees the stability of face AR applications. To this end, we propose a novel loss function based on the 2D Wasserstein distance combined with a new landmark coordinate sampling method. We also propose several improvements on the standard evaluation protocol to better reflect the robustness of our model.

Based on the previous discussion, we present **a novel analytical tool in Chapter 5**. We find that it is necessary to develop a novel tool to directly interpret and quantify if the prediction lacks local precision or robustness. Therefore, we present a facial landmark position correlation analysis. With this analysis, we gain insights on the predictions of different facial landmark detection models (including cascaded random forests, cascaded CNN, heatmap regression models) and on how CNNs progressively learn to predict facial landmarks. In addition, we propose a weakly-supervised learning method that allows to considerably reduce the manual effort for dense landmark annotation.

Chapter 3

Fine-grained facial landmark detection exploiting intermediate feature representations

The first subject that we are interested in is the local precision of the landmark position, which is critical for AR applications. We present an approach for *Fine-grained Facial Landmark Detection* (FFLD). Our framework employs a coarse-to-fine structure, through the exploitation of intermediate feature representations. The proposed method outperforms state-of-the art methods in terms of local precision, and is capable of alleviating the random annotation error along the semantic boundaries of facial components.

3.1 Introduction

Precision of facial landmark position is of extreme importance for AR applications such as face modeling, virtual make-up, and in tasks that require pixel-level accuracy for aesthetic AR applications. In these applications, slight displacements of the estimated landmark positions significantly deteriorate the user experience. In existing methods from the literature, this refinement is usually performed in a post-processing step [Zeng et al., 2015, Wang et al., 2017c]. In this chapter, the proposed method is integrated to the learning of CNN.

In chapter 2.2, we have categorized deep CNN facial landmark detection models into two different types according to their network output: *Coordinate Regression Models* and *Heatmap Regression Models* [Wu and Ji, 2019, Yan et al., 2018, Nibali et al., 2018, Merget et al., 2018]. Most of the Coordinate Regression Models end with a Fully Connected (FC) layer to directly predict the numeric coordinate values. On the other hand, Heatmap Regression Models generally adopt a Fully Convolutional Neural Network that provides one "heatmap" per landmark as output. Each pixel value of a particular heatmap represents the conditional probability of the landmark being present at that point given the pixel position. Heatmap Regression Models are an alternative to Coordinate Regression Models and has become popular in recent years due to their strong capacity of handling large head poses. However, both of them cannot directly provide highly accurate FFLD.

Local imprecision problem of Coordinate Regression Models: A single-stage Coordinate Regression Model usually suffers from local imprecision. This is likely due to the decrease of local detail through successive feature map down-sampling. Hence, numerous coarse-to-fine methods [Sun et al., 2013, Zhou et al., 2013a, Zhang et al., 2014a, Trigeorgis et al., 2016, Fan and Zhou, 2016, Kowalski et al., 2017, Chen et al., 2017a, He et al., 2017b, Lv et al., 2017] have been proposed to cope with this issue. In most of them, the refinement is performed in a cascade that sequentially processes local image patches to recover the local detail information.

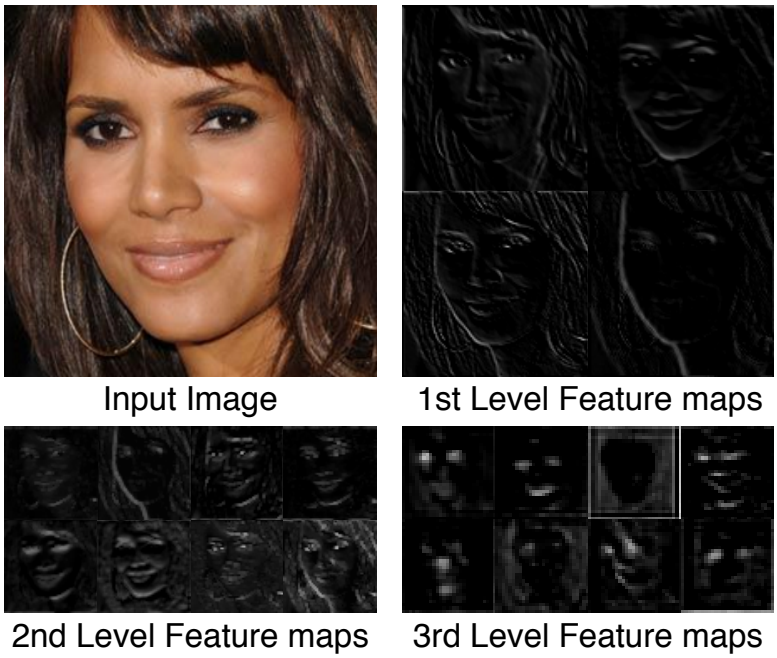


FIGURE 3.1: A visualization of the feature maps in different levels of ResNet18 trained for facial landmark detection. We can observe that low level feature maps retain abundant visual boundary information. Higher-level feature maps present more general information compared to lower-level ones. (The spatial dimension of the third-level feature maps is increased and interpolated for better visibility.)

Local imprecision problem of Heatmap Regression Models: The imprecision of Heatmap Regression Models is mainly due to the quantization error on the output heatmaps. In most of these models, the final landmark prediction is obtained from the position of the maximum value on the output heatmap, thus an integer value. Furthermore, the predicted heatmap is typically around four times smaller than the input image. These two factors provoke considerable quantization errors for Heatmap Regression Models.

To provide FFLD, we propose to combine the advantages of both Coordinate Regression Models and Heatmap Regression Models by exploiting the intermediate low-level feature maps in Coordinate Regression Models and reusing them in an additional processing step that is integrated in the model. This reuse of low-level feature maps is inspired by skip connections that are widely used in Heatmap Regression Models with encoder-decoder structures. It enables information on local detail to be directly used by higher-level processing stages in the neural network. We found that the boundaries of facial components still remain clear on the low-level feature maps even if the Coordinate Regression Model output suffers from local imprecision problem (see Fig. 3.1).

This shows the potential that our method leverages by using this information in the output layers. On the other hand, because we adopt a Coordinate Regression Model framework, the output predicted by the final FC layer is a vector of real (floating-point) numbers, which avoids the quantization errors inherent in Heatmap Regression Models.

Besides, the $L2$ loss used in Heatmap Regression Models [Newell et al., 2016, Bulat and Tzimiropoulos, 2017b] differs from the traditional one commonly used for Coordinate Regression Models. Heatmap $L2$ loss computes pixel-wise $L2$ distance between the predicted heatmaps and target heatmaps. It is robust to outliers, as its value saturates when two Gaussian distribution (representing the ground-truth and predicted landmark positions) do not overlap regardless of the distance between two landmarks. Robust loss

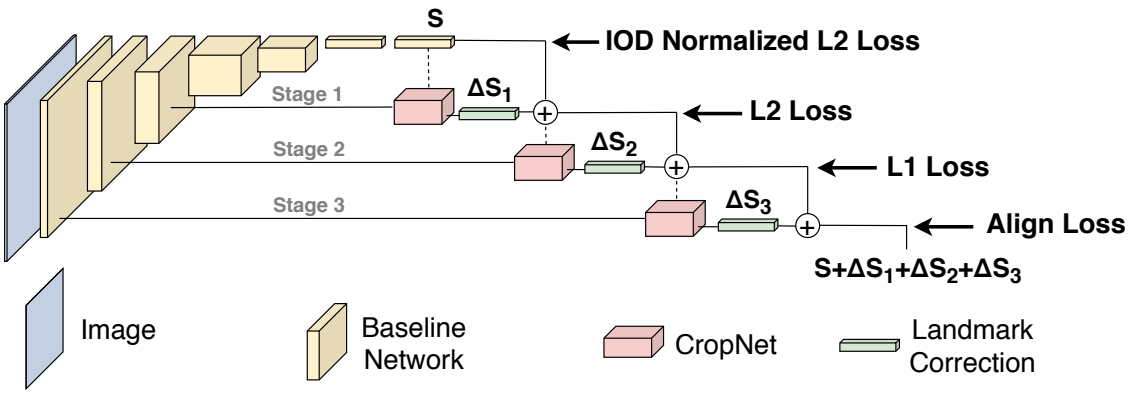


FIGURE 3.2: Overview of our 3-stage coarse-to-fine coordinate regression framework. It contains skip connections between the intermediate feature maps and the main network output S. In each stage, a CropNet (described in Sect. 3.3.2) refines the landmark location (ΔS) based on the patches cropped from lower-level feature maps. The refined landmark locations are passed to the next stage (dotted lines) for cropping. Different loss functions (IOD Normalized $L2$ Loss, $L2$ Loss, $L1$ Loss and Align Loss described in Sect. 3.4.3) are used to train different refinement stages in an end-to-end manner (cf. details of gradient back-propagation in Sect. 3.5).

functions have proved to be helpful for CNNs to focus on small-range errors [Belagiannis et al., 2015, Feng et al., 2018b], and we will show that this is beneficial for FFLD.

Generally speaking, our Coordinate Regression Model-based coarse-to-fine framework (see Fig. 3.2) is inspired by the use of skip connections and the robust spatial loss function used in Heatmap Regression Models. The main contributions in this chapter are:

- A novel *feature map patch alignment* method (Sect. 3.3), to establish skip connections in coordinate regression models, where a small subsidiary network *CropNet* cooperates with the main CNN (baseline network) to provide refinement corrections. CropNets leverage local detail information on the patches of low-level feature maps. The refined correction is based on a direct measure of the crop misalignment. Unlike the previous coarse-to-fine methods, our baseline network can be jointly learned with refinement since CropNets enable the gradient to be back-propagated directly through low-level feature maps.
- A novel robust loss function named *Align Loss* (Sect. 3.4.3), which measures the minor but important misalignment between the patches cropped by the ground truth and predicted localization. This loss calculates the pixel-wise value differences between two patches. Compared to standard $L2$ Loss, our Align Loss forces the predicted landmarks to stay on the boundary lines and is thus able to improve the precision to the pixel level.
- A *multi-loss training scheme* (Sect. 3.4), where different loss functions in different refinement stages are employed. To achieve extreme localization precision, loss functions sensitive to big errors are assigned to coarser stages, and loss functions more sensitive to small errors are assigned to finer stages.

3.2 Related Work to Robust Regression

Robust training is critical to enhance the landmark accuracy especially for small errors. Previous work on robust loss function for deep model regression is mainly inspired by

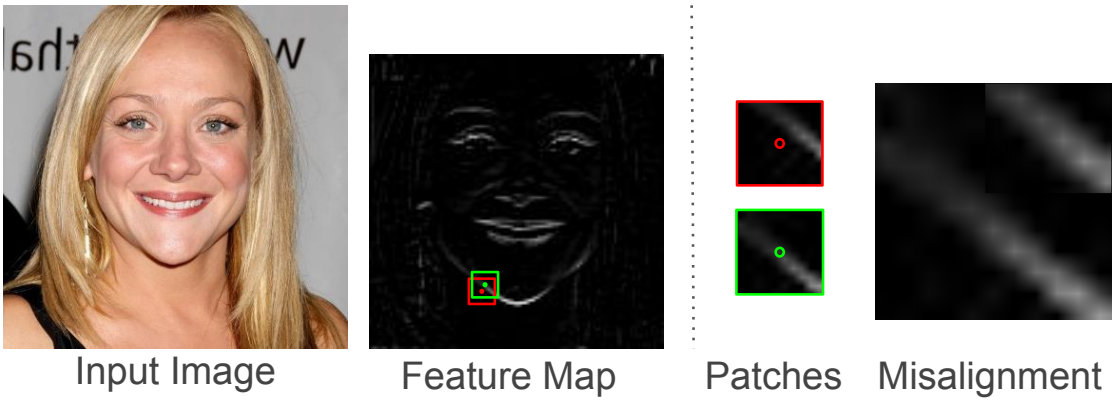


FIGURE 3.3: The main idea of feature map patch alignment: A misplaced landmark (in red) leads to a boundary misalignment on the cropped feature map patch compared to the ground truth (in green). Our model measures this patch misalignment to estimate a refined correction to coarse landmark prediction. Note that our patch alignment approach is different from existing image alignment methods which take pairs of input images [Chang et al., 2017]: it only uses misaligned patches as input and learns misalignment for each landmark based on the statistics.

the use of the M-estimator in robust statistics. The primary goal is to attenuate the impact of outliers on the overall loss. [Belagiannis et al., 2015] proposed to use Tukey's biweight loss function for human pose estimation. Their loss function saturates with large residuals. They showed that their loss function helps the deep regression model to converge both faster and better compared to the traditional $L2$ loss function. [Feng et al., 2018b] proposed a novel wing loss for deep robust facial landmark detection, which behaves like the logarithmic function for small errors and like the $L1$ loss function for large errors. They emphasized the importance of small residuals during the calculation of the loss. Recently, [Lathuilière et al., 2018] combined a robust mixture modeling to deep CNN regression models which adapts to an evolving outlier distribution without setting a manual threshold.

3.3 Feature Map Patch Alignment

In Heatmap Regression Models, the skip connections are intuitive due to the similar spatial dimension shared between input layers and output layers in each stage. However, in Coordinate Regression Models, the spatial resolutions do not match between the input an the output. The dimension of low-level feature maps are too large for the output FC layer. [Miao et al., 2018] and [Yue et al., 2018] used a non-linear embedding layer to reduce the dimension of feature maps for the succeeding FC layer. In contrast, we propose to reduce the feature map dimension by cropping patches around each landmark. Therefore, we developed a feature map patch alignment method (see Fig. 3.3) for FFLD described in the following sections.

3.3.1 Baseline Network

The aim of facial landmark coordinate regression is to establish a non-linear mapping between an image $\mathcal{I} \in \mathbb{R}^{h \times w}$ and landmark Cartesian coordinates $\mathcal{S} \in \mathbb{R}^{2N}$, where N represents the number of landmarks. We use ResNet [He et al., 2016] as our baseline network but reduce 75% of its feature map channels in each layer. Despite a fairly good

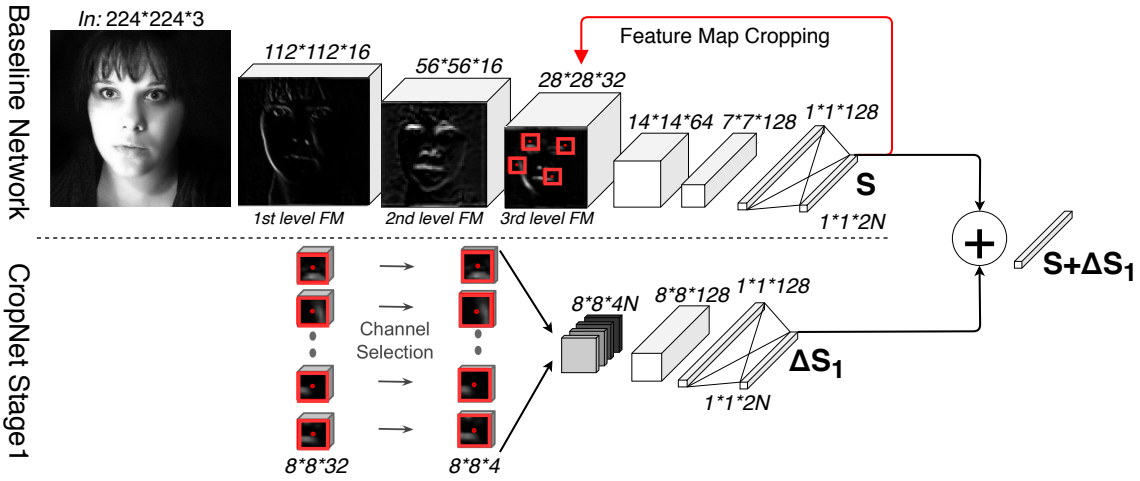


FIGURE 3.4: An illustration of our feature map patch alignment method in the first stage. Small patches are cropped from the 3rd-level feature maps based on the coarse landmark detection S given by the baseline network. The number of channels is reduced by a linear 1×1 convolutional layer. The selected feature maps are then concatenated as input to the CropNet, which predicts the correction of landmark localization ΔS_1. Finally, $S + \Delta S_1$ is passed as coarse prediction to crop patches from the 2nd-level feature maps in the next stage.

overall prediction performance of this model, it lacks some local precision. Following refinement is then performed based on this baseline network.

3.3.2 CropNet

The objective of the CropNet modules is to find a correction to the coarse prediction based on the low-level feature maps from the main network. Figure 3.4 illustrates how we utilize CropNet to refine the facial landmark localization in the first refinement stage. In order to process detailed information on high-dimensional low-level feature maps with low computational complexity, the input dimension of the refinement network is reduced. Given low-level feature maps of dimension (C, H, W) as input, we propose to reduce the dimension of (H, W) by cropping feature maps and to reduce the number of channels C by learning a linear channel reduction explained in the following.

Feature map cropping: Similar to previous coarse-to-fine frameworks [He et al., 2017b], we perform a central crop according to the coarse prediction from the previous stage. As shown in Fig. 3.4, the spatial dimension is reduced from 28×28 to 8×8 on the 3rd level feature maps. The crop on the 3rd level feature maps is performed at first in our coarse-to-fine framework, therefore it is indicated as the 1st stage. Similarly, we also crop patches of 8×8 from the 2nd level feature maps (sized 56×56) in the 2nd stage and from the 1st level feature maps (sized 112×112) in the 3rd stage. In different stages, due to identical patch sizes but different feature map sizes, the patches have different support on the input image. Thus, patches cropped from the 3rd level feature maps have bigger support but contain coarser information while the patches cropped from the 1st level feature maps have smaller support and contain more details. This corresponds well to a coarse-to-fine strategy, where relatively larger errors are corrected in stage 1 and detailed, pixel-level errors in stage 3.

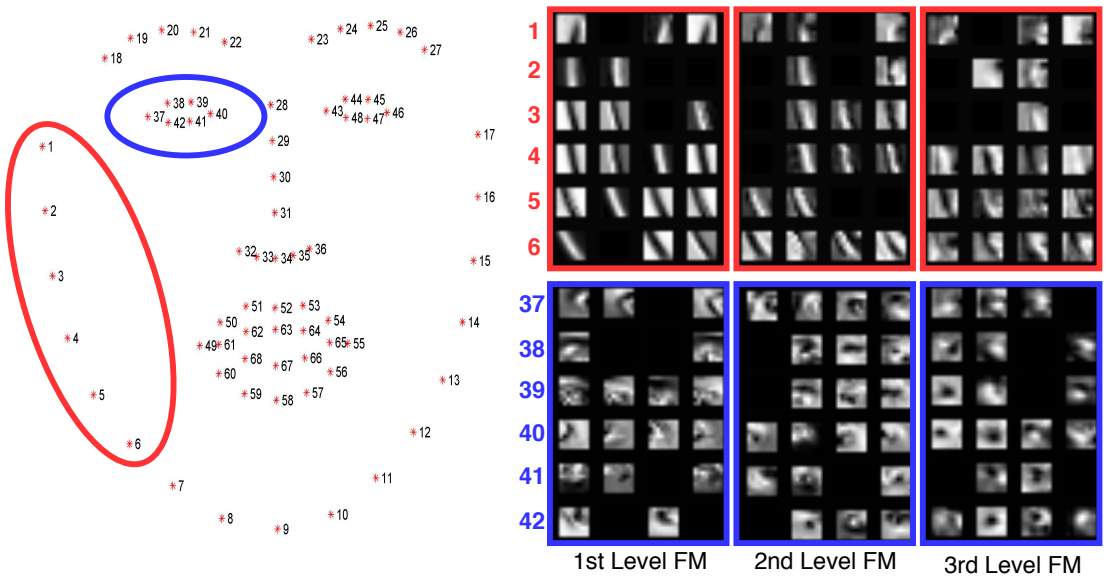


FIGURE 3.5: Examples of the patches around landmarks on the face contour (red, top row) and the eye contour (blue, bottom row) after channel reduction. Four channels (columns) are selected per landmark (lines).

To avoid introducing additional quantization error, the pixels on cropped patches are resampled by bilinear interpolation from original feature maps (as in the Mask-RCNN [He et al., 2017a], for example). This enables us to crop a feature map even on a non-integer position.

Channel reduction: Without the reduction of channels, the input channel dimension to CropNet would lead to high computational complexity. Therefore, we use a linear 1×1 convolutional layer to select and combine the most useful channels (here: 4) from the original ones, especially the ones containing boundary information. As shown in Fig. 3.4, the channel dimension is reduced from $32N$ to $4N$ in stage 1, where N represents the number of the landmarks. We visualize several output examples from channel reduction in Fig. 3.5. We observe that most of the selected channels contain important boundary information.

3.4 Multi-loss Training Scheme

Most previous coarse-to-fine methods use the traditional $L2$ loss function for all refinement stages. Here, we propose to use a set of different loss functions to train different stages. Our motivation is that some loss functions are more adapted to optimize the bigger errors for coarse detection yet other loss functions are more sensitive to the smaller errors for fine detection.

3.4.1 IOD Normalized $L2$ Loss

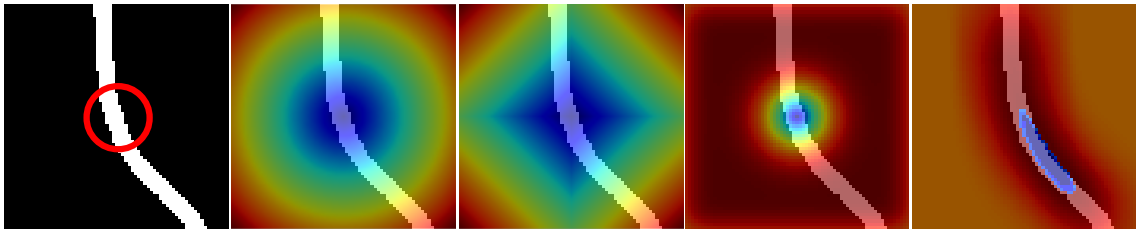
To train our baseline network, we use the $L2$ loss, i.e. the squared error of landmark positions, normalized by the Inter-Occular Distance, like [Lv et al., 2017] and [Kowalski et al., 2017]:

$$L = \frac{\|S_{GT} - S_{Pred}\|_2^2}{d}, \tag{3.1}$$

where S_{GT} and S_{Pred} denote the ground-truth positions (shape) and predicted positions respectively. d denotes the Inter-Ocular Distance. The $L2$ penalizes big errors, especially those occurring in hard examples with large head pose.

3.4.2 $L2$ & $L1$ Loss

For illustration, in Fig. 3.6 we visualize the values of different loss functions on a synthetic example.



(A) Patch (B) $L2$ (C) $L1$ (D) HM $L2$ (E) Align

FIGURE 3.6: A synthetic example to illustrate different loss functions for facial landmark detection. We simulate an artificial feature map patch as a crop on a landmark localized on the face contour. (a) is the feature map patch center cropped by the ground-truth landmark location. The red circle indicates the ground-truth landmark location. Each pixel value on (b), (c), (d), (e) represents the loss value when the prediction is positioned on this pixel. Blue indicates lower loss values while red indicates higher values. (b), (c), (d), (e) represent respectively $L2$ loss, $L1$ loss, heat-map regression $L2$ Loss and our Align Loss. Note that the loss values are normalized on each image.

The traditional $L2$ loss (Fig. 3.6 (b)) used in Coordinate Regression Models calculates the Euclidean distance between the Cartesian coordinates of prediction and ground truth. The loss values grow infinitely when the predictions get further away from the ground truth. For the heatmap $L2$ loss (HM $L2$, Fig. 3.6 (d)), which calculates the $L2$ distance between the predicted and ground truth heatmaps, the loss values saturate when they are far away from the ground truth. Hence, compared to the heatmap $L2$ loss function, the standard $L2$ loss is more suitable for minimizing relatively big errors since the large errors do not saturate and thus do not vanish in the gradient descent optimization. Therefore, we use standard $L2$ loss in the first refinement stage.

[Feng et al., 2018b] showed that the $L1$ loss function (Fig. 3.6 (c)) performs better than $L2$ as it focuses on middle and small-ranged errors. Thus, we use the $L1$ loss function in the second refinement stage.

3.4.3 Align Loss

To provide pixel-level precision and force the predicted landmarks to stay on the boundary lines, we propose a novel loss function, called "Align Loss". The Align Loss is used to train our CropNet in the last (3rd) refinement stage, i.e. the one with the most detailed feature maps. The Align Loss operation is intuitive when applied on small patches since it measures patch misalignment by simply calculating the pixel value difference between the patches cropped by the ground truth and by the prediction.

Our Align Loss is inspired by the robust spatial $L2$ loss used in Heatmap Regression Model-based approaches. It calculates the pixel-wise squared difference of values between the patches cropped using the ground truth location and the patches cropped using the predicted location. To facilitate the network convergence, we apply a Gaussian kernel G to filter both of the patches prior to the loss calculation. This essentially smooths

the gradient spatially over the patch region and helps the gradient descent algorithm to converge to the optimal solution.

Our loss function L for a given stage can be represented as:

$$L = \sum_{m=1}^{|P_{GT}|} \|(G * P_{GT})_m - (G * P_{S+\Delta S})_m\|_2^2, \tag{3.2}$$

where P_{GT} indicates the patches cropped by ground-truth locations (with m being the pixel index), and $P_{S+\Delta S}$ indicates the patches cropped by the CropNet prediction. The $*$ refers to a convolution operation.

Our Align Loss bears three advantages:

- It improves landmark precision to the pixel level because even a small patch misalignment may result in large loss values, which is beneficial to FFLD.
- It forces the landmark refined by our CropNet to stay on visual boundary lines. Align loss is more sensitive to the misalignment in the orthogonal direction to the boundary than those along the boundary direction (see 3.6 (e)). We believe that if a landmark is misaligned along the boundary (e.g. on the chin or cheek contour), it is visually more acceptable than a landmark misaligned in the orthogonal direction even though the error remains the same. This is supported by the work of [Dong et al., 2018b], where it has been observed that there exists a random error of manual annotation along the boundary direction.
- It is robust to outliers as they have less influence during the training stage, even if they are out of the patch scope.

To summarize, based on the different characteristics of the loss functions introduced before, we assign them as follows: IOD Normalized $L2$ loss for the baseline network (big errors on hard examples); standard $L2$ for the first refinement stage (relatively big errors); $L1$ loss for the second refinement stage (middle and small-range errors) and Align Loss for the last refinement stage (pixel-level errors).

3.5 Gradient Backpropagation

Similar to the Heatmap Regression Models, our framework enables the gradient to be back-propagated to the low-level feature maps as well as the crop locations given by the last stage (see Fig. 3.7). Deriving the crop operation gives us the gradient with respect to: low-level feature maps (blue arrow) and crop location (red dotted arrow). In this section we will show the composition of the gradient ∇I at the low level feature map. Following [Jaderberg et al., 2015] and [He et al., 2017b], we demonstrate an example on the first refinement stage, but it is similar in all other refinement stages.

3.5.1 Preliminaries

First, we compute the gradient of the $L2$ loss at the first refinement stage named L_1 with respect to the output of the feature map patch V that is then back-propagated through CropNet:

$$\begin{aligned} \frac{\partial L_1}{\partial V} &= \frac{\partial L_1}{\partial (S_0 + \Delta S)} \cdot \frac{\partial (S_0 + \Delta S)}{\partial \Delta S} \cdot \frac{\partial \Delta S}{\partial V} \\ &= \frac{\partial L_1}{\partial (S_0 + \Delta S)} \cdot \frac{\partial \Delta S}{\partial V}, \end{aligned} \tag{3.3}$$

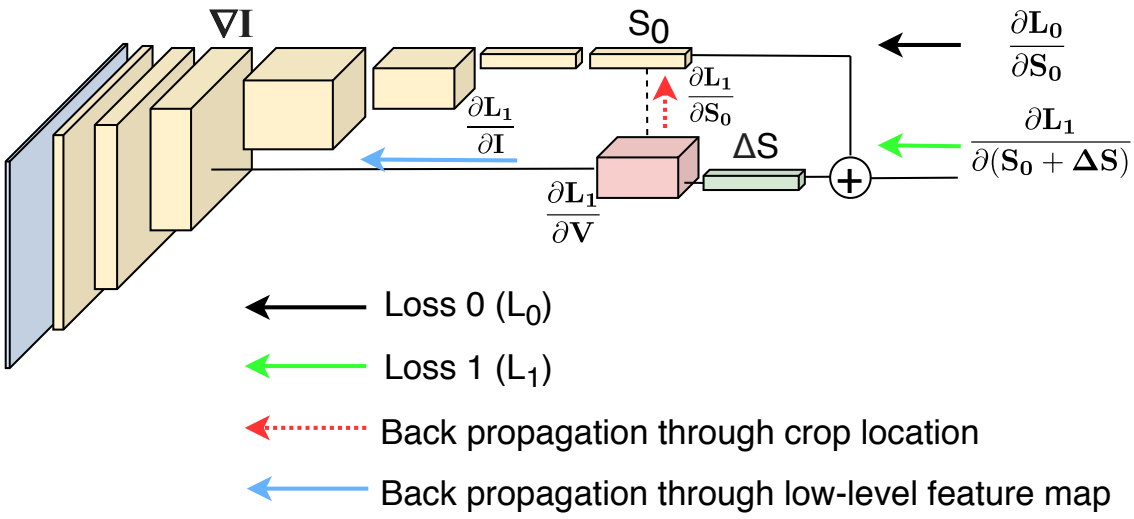


FIGURE 3.7: An illustration of gradient back-propagation in the first refinement stage of our framework.

where $\frac{\partial L_1}{\partial(S_0+\Delta S)}$ is the gradient of the loss L_1 w.r.t. the output coordinates of the first refinement stage, and $\frac{\partial \Delta S}{\partial V}$ is the gradient of the CropNet output w.r.t. its input (standard CNN back-propagation).

Given $\frac{\partial L_1}{\partial V}$, we derive the gradient through our crop operation Γ which can be defined as:

$$V = \Gamma(I, S_0), \tag{3.4}$$

where I is the low-level feature map and S_0 is the crop location obtained by the baseline network. We will derive the gradient w.r.t. both I : $\frac{\partial L_1}{\partial I}$ (blue arrow in Fig. 3.7) and S_0 : $\frac{\partial L_1}{\partial S_0}$ (red dotted arrow in Fig. 3.7) in the following subsections.

Considering that bilinear interpolation is used when we crop the patches, the pixel positioned at (q, p) of V is obtained as:

$$V_{qp} = \sum_{n=0}^{H-1} \sum_{m=0}^{W-1} I_{nm} \max\left(0, 1 - \left|y_q - n\right|\right) \max\left(0, 1 - \left|x_p - m\right|\right) \tag{3.5}$$

$$y_q = y + q - (h-1)/2 \tag{3.6}$$

$$x_p = x + p - (w-1)/2, \tag{3.7}$$

where $(y_q, x_p) \in \mathbb{R}^2$ is the corresponding position of V_{qp} w.r.t the entire feature map I, and (y, x) (components of S_0) is the crop location of each landmark. h and w represent the height and the width of the patch V. H and W represent the height and the width of the feature map I. I_{nm} is the value of the pixel positioned at (n, m) on I.

3.5.2 Gradient w.r.t. low-level feature maps (blue arrow)

We derive the gradient w.r.t the low-level feature maps.

$$\nabla_{low\_level} = \frac{\partial L_1}{\partial I} = \frac{\partial L_1}{\partial V} \cdot \frac{\partial V}{\partial I}. \tag{3.8}$$

Thus, for each (n, m) on I and using Eq. 3.5, we have:

$$\frac{\partial L_1}{\partial I_{nm}} = \sum_{p,q} \frac{\partial L_1}{\partial V_{qp}} \cdot \frac{\partial V_{qp}}{\partial I_{nm}}$$
$$= \sum_{p,q} \frac{\partial L_1}{\partial V_{qp}} \max\left(0, 1 - \left|y_q - n\right|\right) \max\left(0, 1 - \left|x_p - m\right|\right). \quad (3.9)$$

The value of y_q and x_p can be obtained from S_0, h and w in Eq. 3.6 and Eq. 3.7. We can therefore calculate $\frac{\partial L_1}{\partial I}$ in Eq. 3.8 since $\frac{\partial L_1}{\partial V_{qp}}$ has been obtained in Eq. 3.3. In fact, this step can be intuitively interpreted as a reprojection of the gradient on V back to the corresponding position on I based on the crop location S_0.

3.5.3 Gradient w.r.t. crop locations (red dotted arrow)

We now derive the gradient of the loss function w.r.t. a crop location $\frac{\partial L_1}{\partial S_0}$:

$$\frac{\partial L_1}{\partial S_0} = \frac{\partial L_1}{\partial (S_0 + \Delta S)} \cdot \frac{\partial (S_0 + \Delta S)}{\partial S_0} = \frac{\partial L_1}{\partial (S_0 + \Delta S)} \cdot \left(1 + \frac{\partial \Delta S}{\partial S_0}\right)$$
$$= \frac{\partial L_1}{\partial (S_0 + \Delta S)} \cdot \left(1 + \frac{\partial \Delta S}{\partial V} \frac{\partial V}{\partial S_0}\right) \quad (3.10)$$

The terms $\frac{\partial L_1}{\partial (S_0 + \Delta S)}$ and $\frac{\partial \Delta S}{\partial V}$ have already been computed in Eq. 3.3. We can separately consider each coordinate x and y of each landmark, i.e. each component of S_0. Thus, for $\frac{\partial V}{\partial S_0}$ and a landmark coordinate x, we have $\frac{\partial V}{\partial x} = \frac{\partial V}{\partial x_p}$ And we can compute $\frac{\partial V}{\partial x_p}$ for each position (q, p) of V deriving Eq. 3.5:

$$\frac{\partial V_{qp}}{\partial x_p} = \sum_{n=0}^{H-1} \sum_{m=0}^{W-1} I_{nm} \max\left(0, 1 - \left|y_q - n\right|\right) \begin{cases} 0, & \left|m - x_p\right| \geq 1; \\ 1, & m \geq x_p; \\ -1, & m < x_p. \end{cases} \quad (3.11)$$

When applying the bilinear interpolation, we can consider only the four neighbouring pixels of (y_q, x_p), therefore the above equation can be simplified to:

$$\frac{\partial V_{qp}}{\partial x_p} = -I_{\lfloor y_q \rfloor \lfloor x_p \rfloor} y_d + I_{\lfloor y_q \rfloor \lceil x_p \rceil} y_d - I_{\lceil y_q \rceil \lfloor x_p \rfloor} y_u + I_{\lceil y_q \rceil \lceil x_p \rceil} y_u, \quad (3.12)$$

where

$$y_d = 1 - (y_q - \lfloor y_q \rfloor) \quad (3.13)$$
$$y_u = 1 - (\lceil y_q \rceil - y_q), \quad (3.14)$$

and $\lfloor\ \rfloor$ and $\lceil\ \rceil$ are the floor and ceiling function respectively. A similar simplification can be applied to Eq. 3.9.

Hence, we obtain the gradient through the crop location on x coordinate $\frac{\partial V_{qp}}{\partial x_p}$, and analogously for the y coordinate:

$$\frac{\partial V_{qp}}{\partial S_0} = \left(\frac{\partial V_{qp}}{\partial x_p^0}, \frac{\partial V_{qp}}{\partial y_q^0}, ..., \frac{\partial V_{qp}}{\partial x_p^k}, \frac{\partial V_{qp}}{\partial y_q^k}, ...\right), \quad (3.15)$$

where k indicates the index of each landmark. Now back-propagating gradient of Eq. 3.10 further until the feature map I gives:

$$\nabla_{crop\_location} = \frac{\partial L_1}{\partial S_0} \cdot \frac{\partial S_0}{\partial I}, \quad (3.16)$$

where $\frac{\partial S_0}{\partial I}$ can be obtained by standard gradient back-propagation through the main CNN.

Method	Common	Challenge	Full
Inter-Pupil Distance NME (%)			
ESR [Cao et al., 2014]	5.28	17.00	7.58
SDM [Xiong and De la Torre, 2013]	5.57	15.40	7.52
LBF [Ren et al., 2014]	4.95	11.98	6.32
TCDCN [Zhang et al., 2014b]	4.80	8.60	5.54
CFSS [Zhu et al., 2015]	4.73	9.98	5.76
MDM [Trigeorgis et al., 2016]	4.83	10.14	5.88
[Lv et al., 2017]	4.36	**7.56**	**4.99**
AAN [Yue et al., 2018]	4.38	9.44	5.39
DSRN [Miao et al., 2018]	**4.12**	9.68	5.21
ResNet18* (baseline)	6.37	11.32	7.34
ResNet18-FFLD	4.41	8.14	5.14
ResNet50 (baseline)	6.02	10.65	6.94
ResNet50-FFLD	**4.25**	**7.85**	**4.92**
Inter-Eye Corner Distance NME (%)			
PCD-DCNN [Kumar and Chellappa, 2018]	3.67	7.62	4.44
SAN [Dong et al., 2018a]	3.34	**6.60**	**3.98**
Reg + SBR [Dong et al., 2018b]	7.93	15.98	9.46
CPM + SBR [Dong et al., 2018b]	**3.28**	7.58	4.10
ODN [Zhu et al., 2019a]	3.56	6.67	4.17
ResNet18 (baseline)	4.38	7.46	4.98
ResNet18-FFLD	3.18	5.64	3.66
ResNet50 (baseline)	4.12	7.05	4.73
ResNet50-FFLD	**3.06**	**5.44**	**3.50**

TABLE 3.1: NME comparision on 300W dataset of ResNet with our Fine-grained Facial Landmark Detection (ResNet18/50-FFLD) framework and other approaches. * Note that all ResNet18/50 used here are simplified version with only 25% channels compared to the original version.

3.5.4 Summary

Consider the standard gradient back-propagated through the baseline network (black arrow in Fig. 3.7):

$$\nabla_{standard} = \frac{\partial L_0}{\partial I} = \frac{\partial L_0}{\partial S_0} \cdot \frac{\partial S_0}{\partial I}. \tag{3.17}$$

$\frac{\partial L_0}{\partial S_0}$ is calculated through deriving the loss function and $\frac{\partial S_0}{\partial I}$ can be obtained by standard gradient back-propagation through the baseline network.

The overall gradient arriving at I through back-propagation is the sum of the three gradients described before:

$$\nabla I = \nabla_{low\_level} + \nabla_{crop\_location} + \nabla_{standard}. \tag{3.18}$$

3.6 Experiments

3.6.1 Datasets

300W dataset: 300W dataset [Sagonas et al., 2013] involves five facial landmark datasets: HELEN [Le et al., 2012], LFPW [Belhumeur et al., 2013], AFW [Zhu and Ramanan, 2012],

XM2VTS [Messer et al., 1999] and IBUG. We follow [Ren et al., 2014] to use the training set of 3148 images which includes the entire AFW dataset, HELEN training sets and LFPW training sets. The test set of 689 images in total is divided into (i) common subset and (ii) challenging subset. (i) consists of 554 images from the test set of LFPW and HELEN and (ii) consists of 135 images from the IBUG dataset.

300VW dataset: 300VW [Shen et al., 2015] is a video-based facial landmark detection dataset which is annotated in the same manner as 300W. It provides 114 videos in total including 64 videos for validation. The test subset is further categorized into 3 categories based on the level of unconstrained conditions.

AFLW dataset: AFLW [Koestinger et al., 2011] is a large-scale dataset which contains 24386 faces with large pose variations of +/- 90 degree in yaw. All of the images in the dataset are annotated with up to 21 points depending on the landmark visibility. We adopt two protocols from [Zhu et al., 2016a]. In the *AFLW-Full* protocol, the entire dataset is split into 20,000 images for training and 4,386 images for test. In *AFLW-Frontal* protocol, a subset of 1,314 frontal faces are selected from the entire 4,386 images for frontal evaluation. Note that the landmark format is changed to 19 points excluding the landmarks on both ears in these two protocols.

3.6.2 Evaluation Metrics

We evaluate our method by measuring the Normalized Mean Error (NME) between our model prediction and the ground truth. On 300W and 300VW datasets, the errors are normalized by the inter-ocular distance (eye-corners or pupils) as in most of the recent comparisons. On the AFLW dataset, due to the large pose variations, we use face size to normalize our mean errors as in [Lv et al., 2017]. Additionally, the Cumulative Error Distribution (CED) curve is used for evaluation.

3.6.3 Comparison with State-of-the-art Methods

Results on 300W: A comparison of our methods with other facial landmark detection algorithms is presented in Table 3.1. In [Dong et al., 2018b], we note that the Coordinate Regression Models (Reg+SBR) shows inferior performance compared to the Heatmap Regression Models (CPM+SBR). By integrating our Fine-grained Facial Landmark Detection (FFLD) framework, our Coordinate Regression Model based model has a comparable performance to state-of-the-art Heatmap Regression Models [Kumar and Chellappa, 2018, Dong et al., 2018b, Dong et al., 2018a, Tai et al., 2019]. Refined prediction has been significantly improved by nearly 25% compared to the baseline output. We show our CED curve compared to 3DDFA [Zhu et al., 2016b], DRMF [Asthana et al., 2013], CFSS [Zhu et al., 2015] and TCDCN [Zhang et al., 2014b] in Fig. 3.8 (b). We observed that our coarse-to-fine FFLD framework (gray) is able to provide precise fine-grained correction to the coarse prediction given by the baseline network (pink). Several qualitative results are presented in Fig. 3.9. Specifically, by using our proposed Align Loss, we found that the landmarks are more likely to be aligned on the boundaries of the facial components. Additional visual comparisons are presented in Fig. 3.14 and Fig. 3.15.

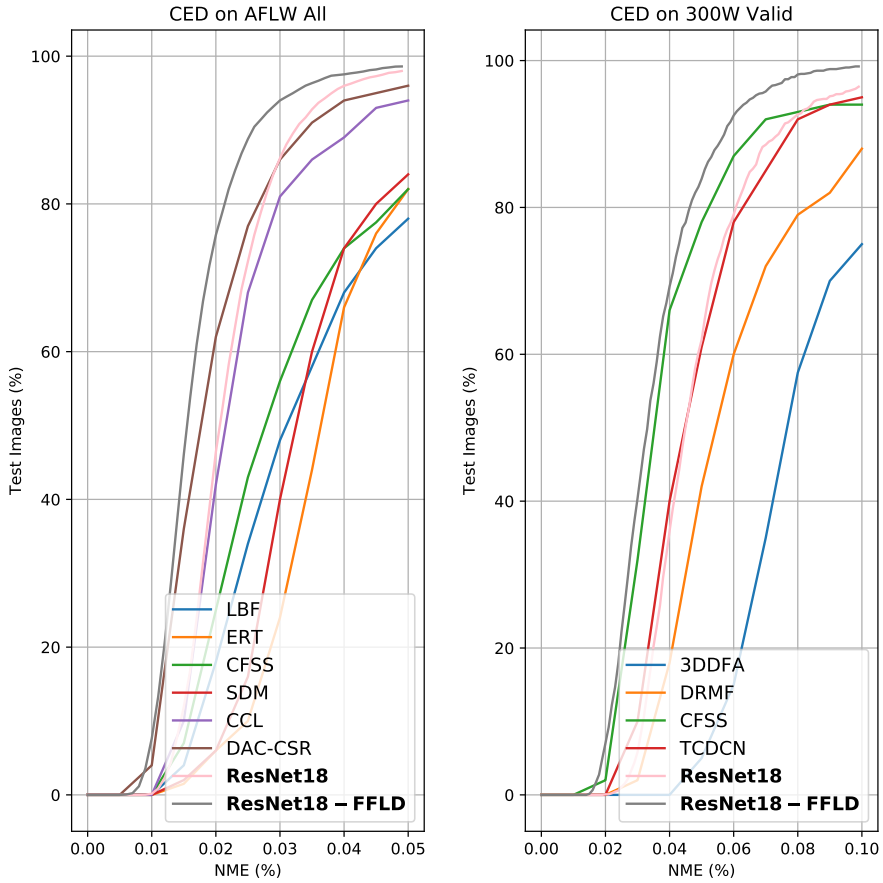
Results on AFLW: We show the performance comparison on the AFLW dataset in Table 3.2. Compared to our baseline, the precision of ResNet18-FFLD is significantly improved by a large margin of 25%. We compare our methods with LBF [Ren et al., 2014], ERT [Kazemi and Sullivan, 2014], CFSS [Zhu et al., 2015], SDM [Xiong and De la Torre, 2013], CCL [Zhu et al., 2016a], DAC-CSR [Feng et al., 2017b] in CED curve shown in Fig. 3.8 (a). We visually compare the prediction between our approach and the baseline in Fig. 3.16.

Method	AFLW-Full	AFLW-Front
SDM	4.05	2.94
ERT	4.35	2.75
LBF	4.25	2.74
CFSS	3.92	2.68
CCL [Zhu et al., 2016a]	2.72	2.17
DAC-CSR [Feng et al., 2017b]	2.27	1.81
Reg + SBR	4.77	-
CPM + SBR	2.14	-
SAN	1.91	1.85
DSRN	1.86	-
ODN	**1.63**	**1.38**
ResNet18*	2.30	1.99
ResNet18-FFLD	1.75	1.52
ResNet50	2.13	1.86
ResNet50-FFLD	**1.62**	**1.42**

TABLE 3.2: NME (%) comparision on AFLW dataset of ResNet18/50-FFLD and other approaches. * Note that all ResNet18/50 used here are simplified version with only 25% channels compared to the original version.

Method	Cat. 1	Cat. 2	Cat. 3
TCDCN	7.66	6.77	15.00
CFSS	7.68	6.42	13.70
HG [Newell et al., 2016]	5.44	4.71	7.92
TSTN [Liu et al., 2017a]	5.36	4.51	12.80
DSRN	5.33	4.92	8.85
FHR [Tai et al., 2019]	**5.07**	**4.34**	**7.36**
ResNet18	5.86	5.12	9.14
ResNet18-FFLD	**4.85**	**4.24**	**7.62**

TABLE 3.3: NME (%) Comparison on 300VW dataset of ResNet18-FFLD and other approaches.



(A) AFLW (B) 300W

FIGURE 3.8: CED curves of our method on AFLW and 300W dataset.

Results on 300VW: A comparison of ResNet18-FFLD with other methods on the 300VW dataset is shown in Table 3.3. The 300VW dataset contains frames with large poses. We apply facial landmark detection in a frame-by-frame manner without exploiting any temporal information. Compared to the baseline network, the performance is improved by >20% with our framework.

Method	Baseline (w/o Ref)	1 Stage Ref		2 Stages Ref	
Loss	Norm $L2$	$L2$	$L1$	$L2/L2$	$L2/L1$
NME	4.98	4.03	3.96	3.85	3.77
Method	3 Stages Ref				
Loss	$L2/L2/L2$	$L2/L1/L1$		$L2/L1/AL$	
NME	3.81	3.72		**3.66**	

TABLE 3.4: Multi-loss ablation study on 300W with ResNet-18 as baseline network. Ref - Refinement. AL - Align Loss.

3.6.4 Ablation Studies

Multi-stage & multi-loss comparison: The improvement of our method originates from two aspects: (1) the use of CropNet (multi-stage refinement) and (2) the use of different loss functions in different stages. In Table 3.4, we compare the results of different

FIGURE 3.9: Qualitative results of our approach on the 300W dataset. The prediction by the baseline network ResNet18 (the 1st row). The prediction by ResNet18 with our Fine-grained Facial Landmark Detection (ResNet18-FFLD) framework **without Align Loss** (the 2nd row). The prediction by ResNet18 with FFLD framework **with Align Loss** (the 3rd row). To better visualize the small error, we provide the zoomed image aside. More examples are provided in Fig. 3.14.

loss functions and different number of refinement stages on the 300W dataset based on the ResNet-18 baseline. With more refinement stages, the precision is progressively improved. We find that the 1st stage, the 2nd stage and the 3rd stage contribute respectively 77%, 14% and 8% of the total improvement on NME.

Specifically, when using 3 refinement stages, we test different combinations of the loss functions. By assigning the loss functions that are more sensitive to the small errors in the last stages, the overall performance is further improved. Compared to using $L2$ loss for all stages, using our combination can additionally improve the NME by 0.15. Although this improvement is numerically incremental, it is nonetheless critical for many aesthetic AR applications. We visually illustrate this improvement in Fig. 3.9 and Fig. 3.14.

Gradient backpropagation: To demonstrate that the gradient ∇_{low_level} and $\nabla_{crop\_location}$ do help to improve the refinement. We show the performance by disabling gradient back-propagation on the baseline network feature maps in Table 3.5. We found that the NME on 300W is further improved from 3.78% to 3.66% by allowing both ∇_{low_level} and $\nabla_{crop\_location}$ to be back-propagated on the low-level feature maps of the baseline network. Though both gradient back-propagation achieves incremental improvement, we observe that ∇_{low_level} acts a more important role compared to $\nabla_{crop\_location}$.

∇_{low_level}	$\nabla_{crop\_location}$	NME
✗	✗	3.78
✗	✓	3.74
✓	✗	3.69
✓	✓	**3.66**

TABLE 3.5: Gradient back-propagation ablation study on 300W with ResNet-18.

Effectiveness of our method on small errors: We now focus on the small errors to validate our FFLD method. To further prove the effectiveness of using different losses and back propagation strategies on small errors, we show a landmark-wise CED in Fig. 3.10. In this figure, we focus on the small landmark-wise NME from 1.0 to 3.0. We also sample the difference of the models at NME=2.75. We observe that by using a unique $L2$ loss for all refinement stages, the improvement is limited when the third refinement stage is

added (green & red). When multi-loss scheme is applied (blue), the performance can be improved by a large margin. We also observe that by enabling the ∇_{low_level} and $\nabla_{crop\_location}$ on the low-level feature maps, the performance on small errors can be further improved.

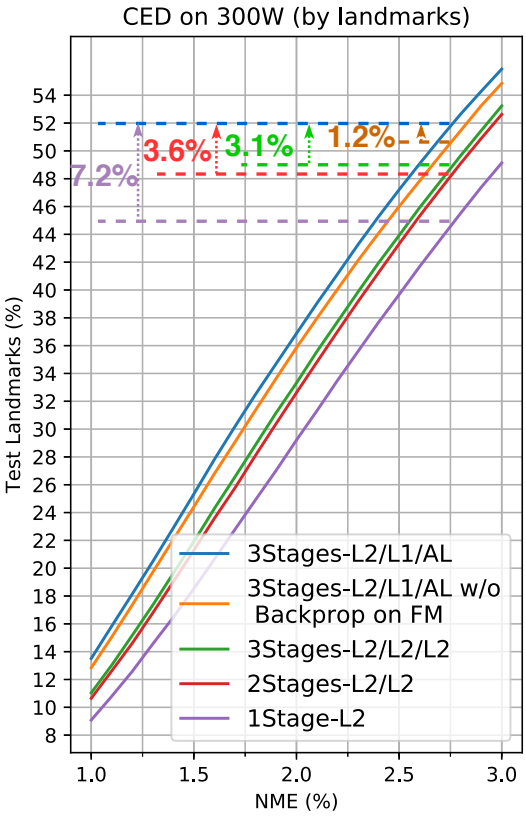


FIGURE 3.10: Landmark-wise CED focused on small errors. AL-Align Loss. w/o Backprop on FM: neither ∇_{low_level} nor $\nabla_{crop\_location}$ is back-propagated on the low-level feature maps. All models are based on the ResNet-18 baseline and tested on 300W.

IOD normalized $L2$ loss: We also compared the results of our baseline ResNet network between using IOD normalized $L2$ loss and standard $L2$ loss. By training with IOD Normalized $L2$ loss, the inter eye-corner distance NME on 300W is improved from 5.14% to 4.98%.

3.6.5 Discussions

Run time and model size: Without any speed optimization such as MobileNet blocks [Sandler et al., 2018], our 3-stage ResNet18-FFLD model runs at 130 fps on a NVIDIA 1080Ti GPU. Our model contains 1.46M parameters, including the baseline network. With an input size of 224 $\times$ 224 and a batch size of 64, our networks require less than 2GB GPU memory during training compared to the heatmap regression model in [Wu et al., 2018b] which require 4 NVIDIA Titan X GPUs to train with a batch size of 8.

Robustness: While being highly dependent on the boundary information on the low-level feature map, we found that our CropNet is robust to partial occlusions (see Fig. 3.11). In Fig. 3.12, we show a histogram of the CropNet output ΔS in each stage. We

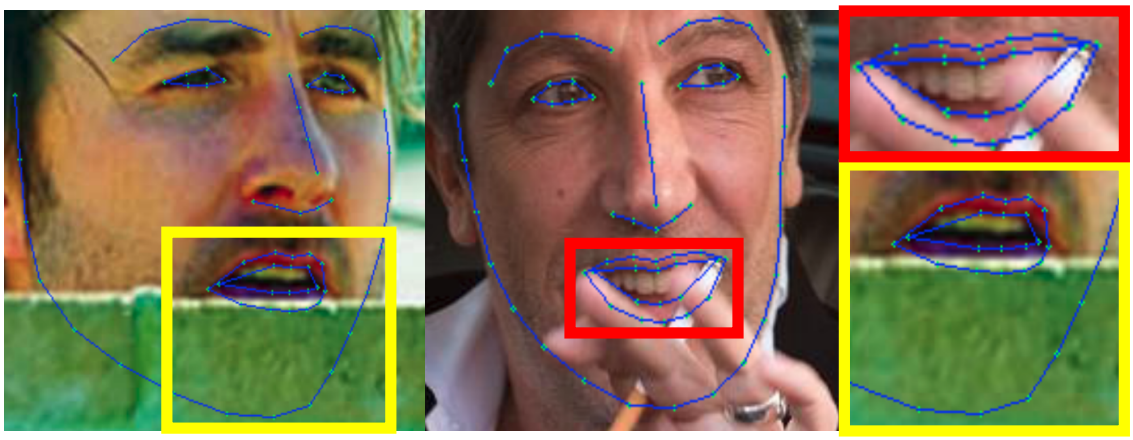


FIGURE 3.11: Examples of our detection on partially occluded images. We can observe that our detection is robust to the occlusions on the face images. That means that even when the boundary information is not given, CropNet still provides a reasonable shape as output.

found that the ΔS forms a stable distribution without long tail, which proves the robustness of our model. We think that ΔS is always regularized by the overall shape because the ΔS for each landmark are predicted by the same FC layer.

Failure cases: We show three worst cases of our detection. All of them are on rather low resolution test images (see Fig. 3.13). In this case, CropNet has difficulties in capturing enough meaningful details (e.g. boundaries) from the indistinct low-level feature maps.

Connection with cascaded CNNs: Most of the cascaded patch-based structures [Chen et al., 2017a] depend on the image patches, which require the refinement stage to learn from RGB information. Our approach is the first to focus on feature map patches. Our patch-based method stands out for the following reasons: (a) The gradient from the refinement stage can be back-propagated directly on the feature maps of the main network. Refinement networks are learned end-to-end, jointly with the main network. (b) It is easier to learn with boundary information from the feature map patches than RGB information from the image patches. (c) With identical patch support, the spatial dimension of feature map patches is much smaller than image patches, which is computationally less expensive.

Connection with Heatmap Regression Models: Both our skip connection and spatial robust loss function are inspired by the popular heatmap regression models. Our method and Heatmap Regression Models can be both trained end-to-end. Moreover, our approach is more memory-efficient than existing heatmap regression models thanks to the smaller spatial dimension in skip connections.

3.6.6 Implementation Details

For CropNet, we propose a relatively simple structure: one batch normalization layer, one ResNet block, one max-pooling layer and one FC layer (see Table 3.6).

For the baseline CNN, we used ResNet18/50 [He et al., 2016] but reduced 75% of its feature map channels in each layer. One important challenge for facial landmark detection is to correctly detect the facial landmarks on extreme poses [Sagonas et al., 2013,Zafeiriou et al., 2017]. In this context, [Feng et al., 2018b] argued that this issue is due to the imbalanced data distribution. To balance the examples in different poses and construct a pose-robust model for landmark refinement on 300W and 300VW, we followed

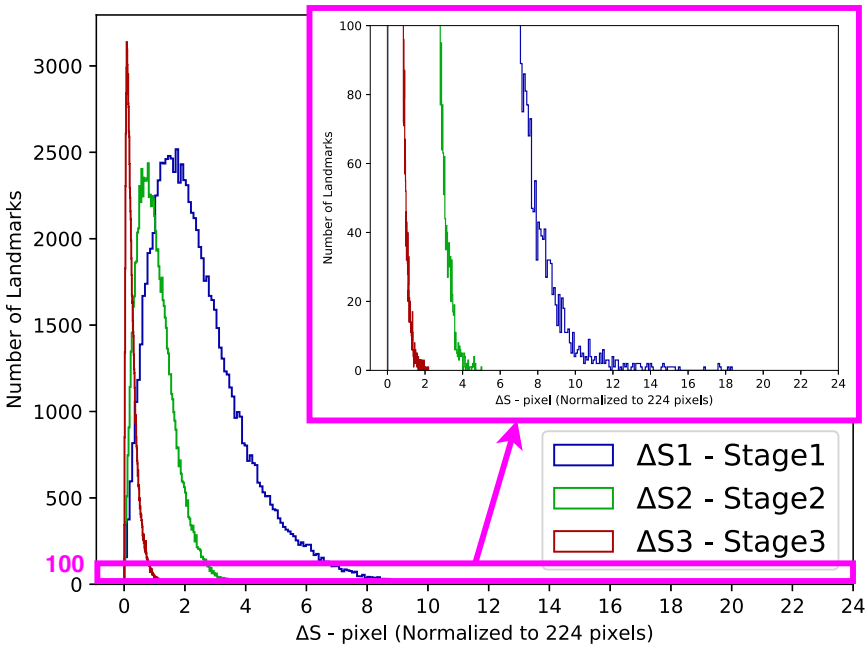


FIGURE 3.12: Histogram of ΔS in each refinement stage (by landmark). The ΔS of each CropNet stage forms a stable distribution without long tail.

Layer	(In_channels, Out_channels, Stride)
Batch Norm	$(4 \times N, 4 \times N, 1)$
ResNet Block	$(4 \times N, 128, 1)$
Max-pooling	(128, 128, 8)
FC	$(128, 2 \times N, -)$

TABLE 3.6: The structure of proposed CropNet. The right column shows the parameters for each layer/block including input channels, output channels and stride. N denotes the number of facial landmarks, which can vary for different datasets.

the training strategy in [Bulat and Tzimiropoulos, 2017b]. We first pre-trained the model on a large synthetic dataset 300W-LP [Zhu et al., 2016b] (LP means Large Pose) with a learning rate of 0.0002 for 80 epochs. 300W-LP expands the 300W dataset by synthesizing large-pose face appearances with a 3D face model and rendering them in different poses (no extra faces). However, the 2D annotation in 300W-LP is not compatible with the original 300W annotation. We then trained our model on 300W dataset for another 350 epochs. The learning rate starts from 0.0001 with a decay of 0.3 for each 50 epochs. On AFLW, we used the PDB strategy from [Feng et al., 2018b] to overcome the imbalanced data distribution problem. We trained our model with a learning rate of 0.0001 for 56 epochs. The learning rate is decayed by 0.3 every 8 epochs.

Afterwards, we trained our 3-stage ResNet-FFLD with three different losses for 400 epochs. The learning rate is initialized to 0.0005 and decayed by 0.3 for each 80 epochs. We initialized the weights of FC layers in CropNets to zero. For the Align Loss, the initial Gaussian kernel size of the convolution kernel is 3 and the sigma is 1. In order to achieve extreme precision, this operation is removed once the loss stops going down.

All experiments are conducted with PyTorch. We used a batch size of 64, Adam as

FIGURE 3.13: Failure cases on low resolution images.

optimizer and 0.0005 as weight decay for all of the training. We further applied data augmentation of +/- 20% on scale, +/- 10% on vertical/horizontal translation and +/- 20% of rotation.

3.7 Conclusion

We presented an novel effective end-to-end framework for Fine-grained Facial Landmark Detection based on coordinate regression deep neural network models. We showed that low-level feature maps contain important contour information, that is useful for refining the landmark positions. By establishing skip connections, the localization accuracy of a coordinate regression model can be significantly improved and achieves comparable performance to the state-of-the-art heatmap regression models. In addition, training different refinement stages with different loss functions, including the proposed Align Loss which forces the landmark to learn extreme accurate prediction, can further increase the localization precision.

The contributions in this chapter led to an article published at *Computer Vision and Image Understanding*.

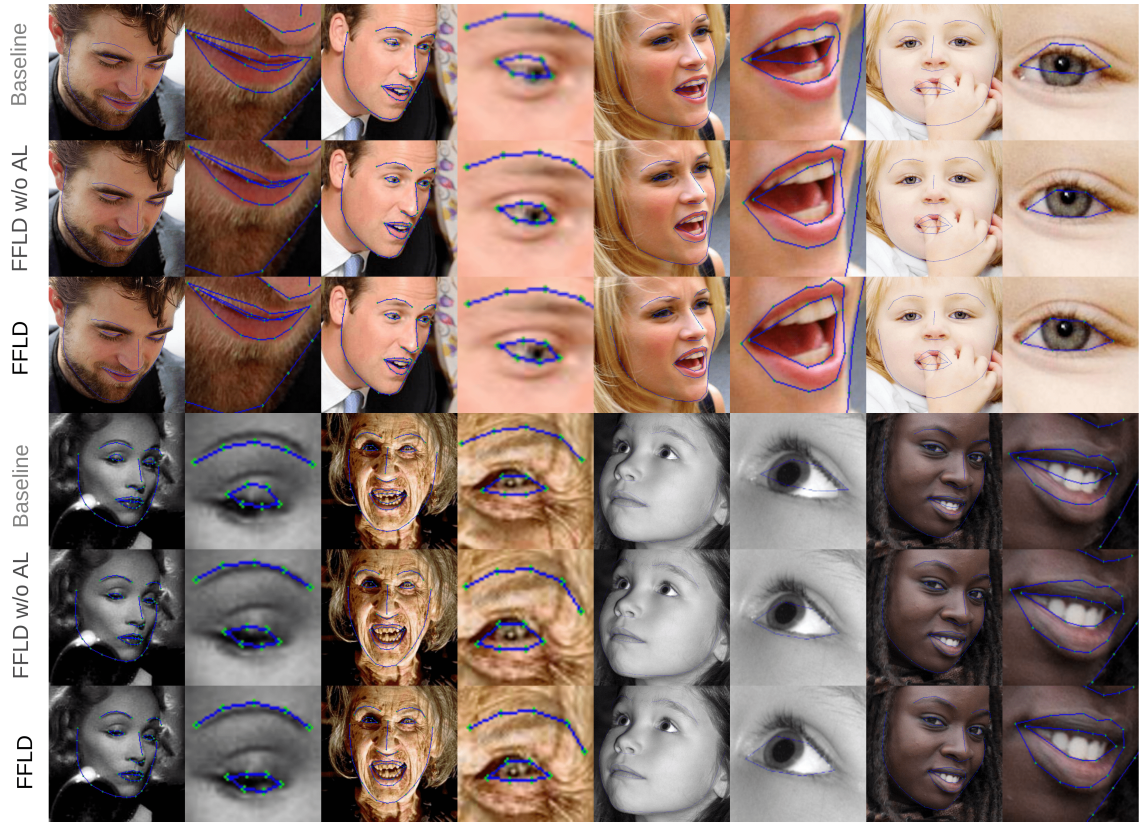


FIGURE 3.14: Qualitative results of our approach on the 300W dataset. The prediction by the baseline network ResNet18 (the 1st row). The prediction by ResNet18 with our Fine-grained Facial Landmark Detection (ResNet18-FFLD) framework **without Align Loss** (the 2nd row). The prediction by ResNet18 with FFLD framework **with Align Loss** (the 3rd row). To better visualize the small error, we provide the zoomed image aside.

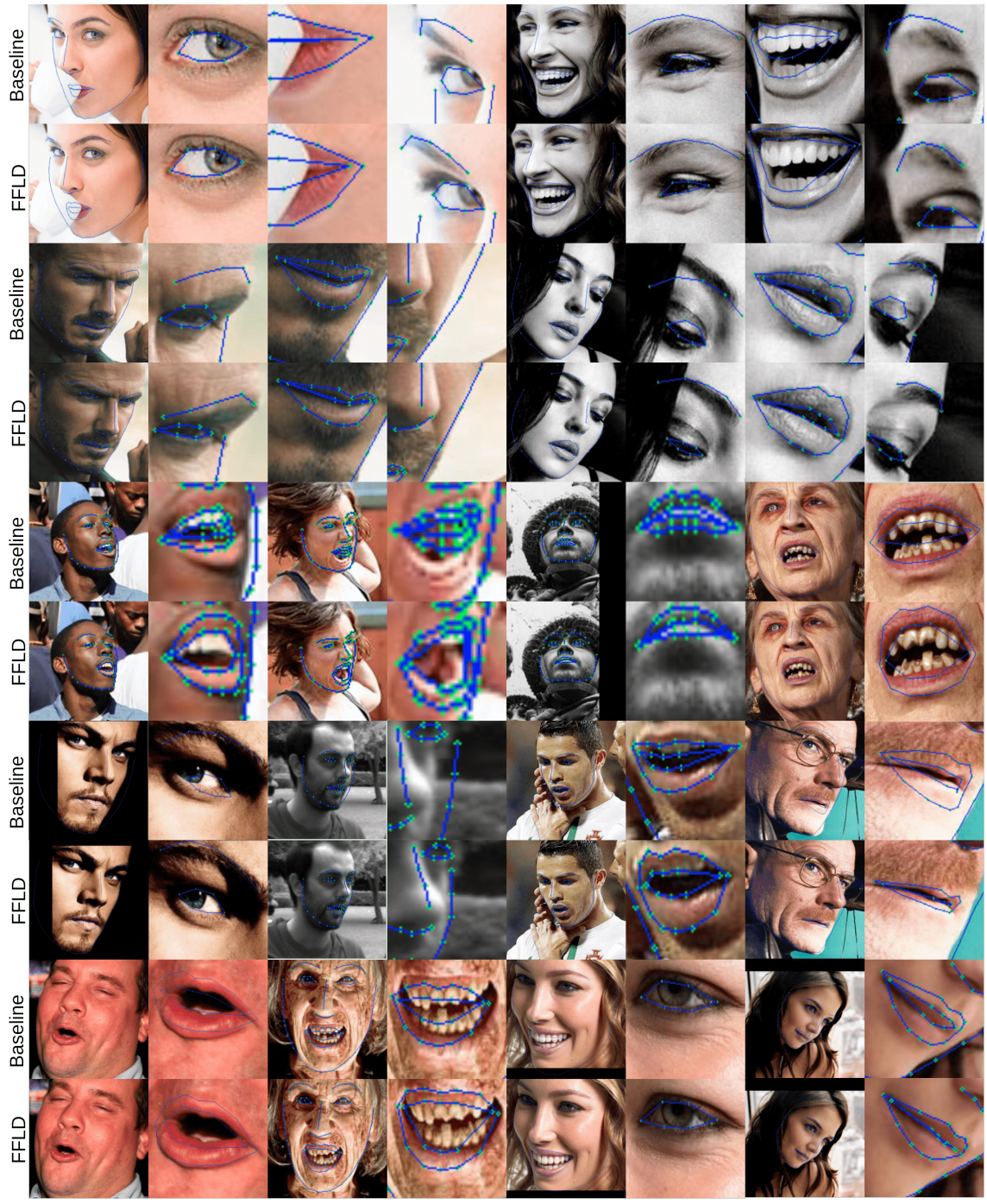


FIGURE 3.15: Qualitative results of our approach on the 300W dataset. The prediction by the baseline network ResNet18 (first row). The prediction by ResNet18 with our Fine-grained Facial Landmark Detection (ResNet18-FFLD) framework (second row). To better visualize the small error, we provide the zoomed image aside.

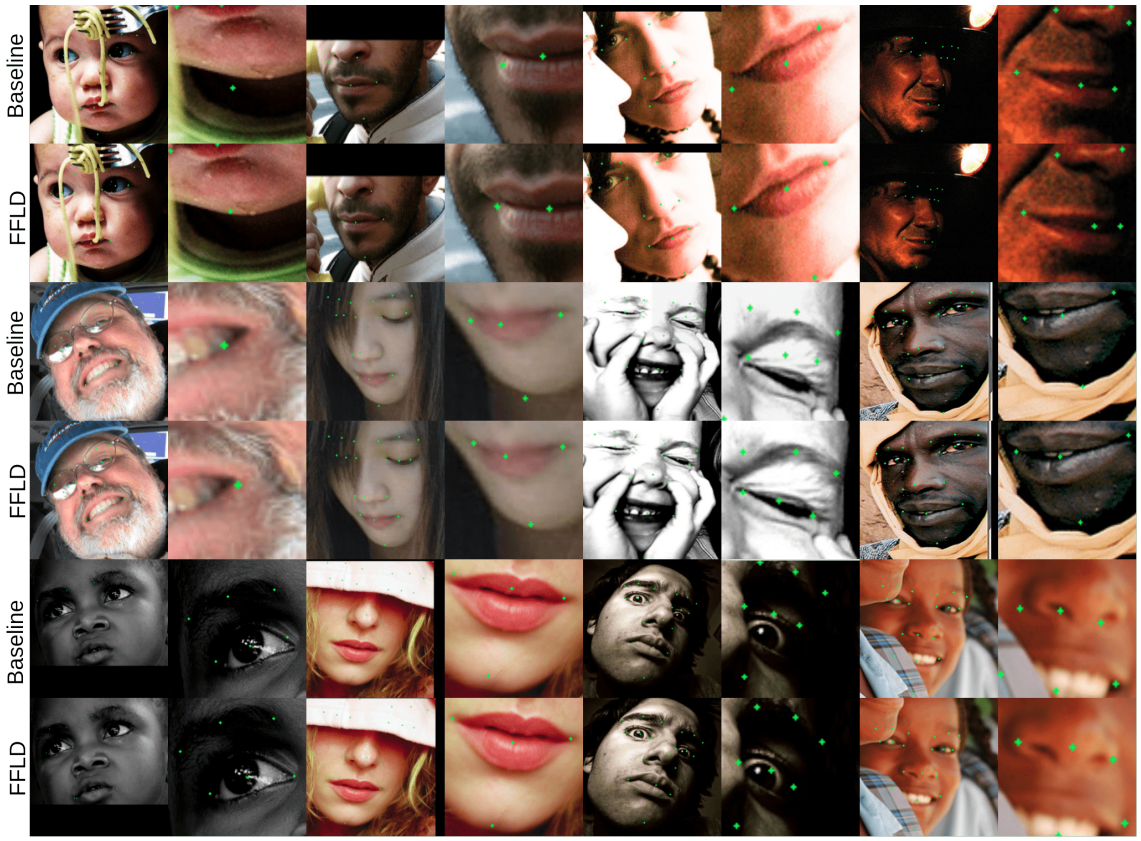


FIGURE 3.16: Qualitative results of our approach on the AFLW dataset. The prediction by the baseline network ResNet18 (first row). The prediction by ResNet18 with our Fine-grained Facial Landmark Detection (ResNet18-FFLD) framework (second row). To better visualize the small error, we provide the zoomed image aside.

Chapter 4

Rethinking Robust Facial Landmark Detection

The second difficulty of facial landmark detection that we want to concentrate on in this chapter is the model robustness, especially for challenging scenarios. For example, motion blur is a typical noise in videos. As the real-time AR applications mainly work on videos, ensuring the robustness against motion blur in addition to other types of noise, is of great importance. In this work, we propose a novel loss function to regularize the output of heatmap regression models by imposing geometric information. In addition, we also discuss the problem of the existing evaluation metric for model robustness and consequently propose several modifications to improve it.

Specifically, we want to make clear that the *robust detection* that we will mention in this chapter is different from the *robust regression* that we were dealing with in the last chapter. *Robust regression*, analogous to robust statistics, aims at alleviating the influence of outliers (due to noise) in the datasets on the estimation. However, *robust detection* here aims at avoiding to make predictions for outliers, e.g. landmarks that are too far away or hidden.

4.1 Introduction

Recently, Heatmap Regression Models have brought the performance on current benchmarks to a very high level. However, maintaining robustness is still challenging in the practical use, especially with video streams that involve motion blur, self-occlusions, changing lighting conditions, etc.

We think that the use of geometric information is the key to further improve the robustness. As faces are 3D objects bound to some physical constraints, there exists a natural correlation between landmark positions in the 2D images. This correlation contains important but implicit geometric information. However, the $L2$ loss that is comonly used to train state-of-the-art Heatmap Regression Models is not able to exploit this geometric information. Hence, we propose a new loss function based on the 2D Wasserstein distance (loss).

The Wasserstein distance, a.k.a. Earth Mover's Distance, is a widely used metric in Optimal Transport Theory [Villani, 2008]. It measures the distance between two probability distributions and has an intuitive interpretation. If we consider each probability distribution as a pile of earth, this distance represents the minimum effort to move the earth from one pile to the other. Unlike other measurements such as $L2$, Kullback-Leibler divergence and Jensen-Shannon divergence, the most appealing property of the Wasserstein distance is its sensitivity to the geometry (see Fig. 4.1).

The contribution of this chapter is two-fold:

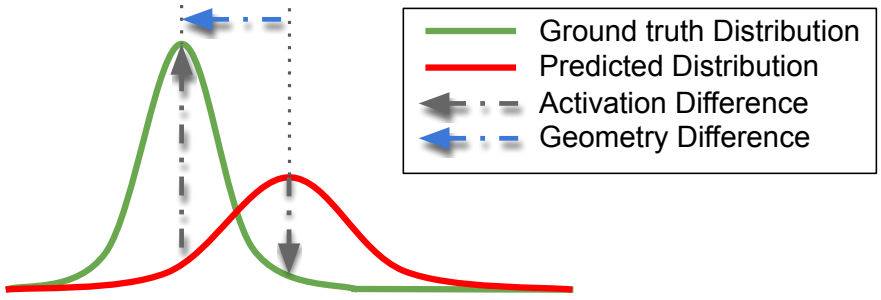


FIGURE 4.1: An illustration of the Wasserstein loss between two 1D distributions. Standard $L2$ loss only considers the "activation" difference (point-wise value difference, vertical gray arrows), whereas the Wasserstein loss takes into account both the activation and the geometry differences (distance between points, horizontal blue arrow).

- We propose a novel method based on the Wasserstein loss to significantly improve the robustness of facial landmark detection.
- We propose several modifications to the current evaluation metrics to reflect the robustness of the state-of-the-art methods more effectively.

4.2 Context & Motivation

Related work to robust facial landmark detection: Robust facial landmark detection in images is a long-standing research topic. Numerous works [Burgos-Artizzu et al., 2013, Smith et al., 2014, Zhao et al., 2013, Yu et al., 2014, Wu et al., 2017a, Zhou et al., 2013b, Feng et al., 2017a, Yang et al., 2015a, Wu and Ji, 2015b, Baltrusaitis et al., 2013] propose methods to improve the overall detection robustness, notably on Active Appearance Models [Cootes et al., 2001], Constrained Local Models [Cristinacce and Cootes, 2006, Asthana et al., 2013], Exemplars-based Models [Belhumeur et al., 2013] and Cascaded Regression Models [Dollár et al., 2010]. These approaches have been superseded more recently with the advent of very powerful deep neural network models. In this context, several works have been proposed for robust facial landmark detection [Zhu et al., 2019a, Xiao et al., 2016, Kowalski et al., 2017, Merget et al., 2018, Yang et al., 2017, Feng et al., 2018b, Wang et al., 2019a] by carefully designing CNN models, by balancing the data distribution and other specific techniques.

Robustness problem of Heatmap Regression Models: Figure 4.2 shows some example results of the state-of-the-art method HRNet [Sun et al., 2019b]. HRNet can handle most of the challenging situations (e.g. Fig. 4.2 (a)). However, we observed that a well-trained HRNet still has difficulties in the practical use when facing extreme poses (Fig. 4.2 (b)(d)(e)(f)), heavy occlusions (Fig. 4.2 (b)(c)(d)(e)) and motion blur (Fig. 4.2 (g)(h)).

These observed robustness issues are rather specific to Heatmap Regression Models. When using Cascaded Regression Models or Coordinate Regression CNNs, even if the prediction is poor, the output still forms a plausible shape. On the contrary, with Heatmap Regression Models, there may be only one or several landmarks that are not robustly detected whereas the others are. In addition, they may be located at completely unreasonable positions according to the general morphology of the face.

This is a well-known problem. Tai et al. [Tai et al., 2019] proposed to improve the robustness by enforcing some temporal consistency. And the approach of Liu et al. [Liu et al., 2019] tries to correct the outliers by integrating a Coordinate Regression CNN at the end. These two methods either add complexity to the models or require learning

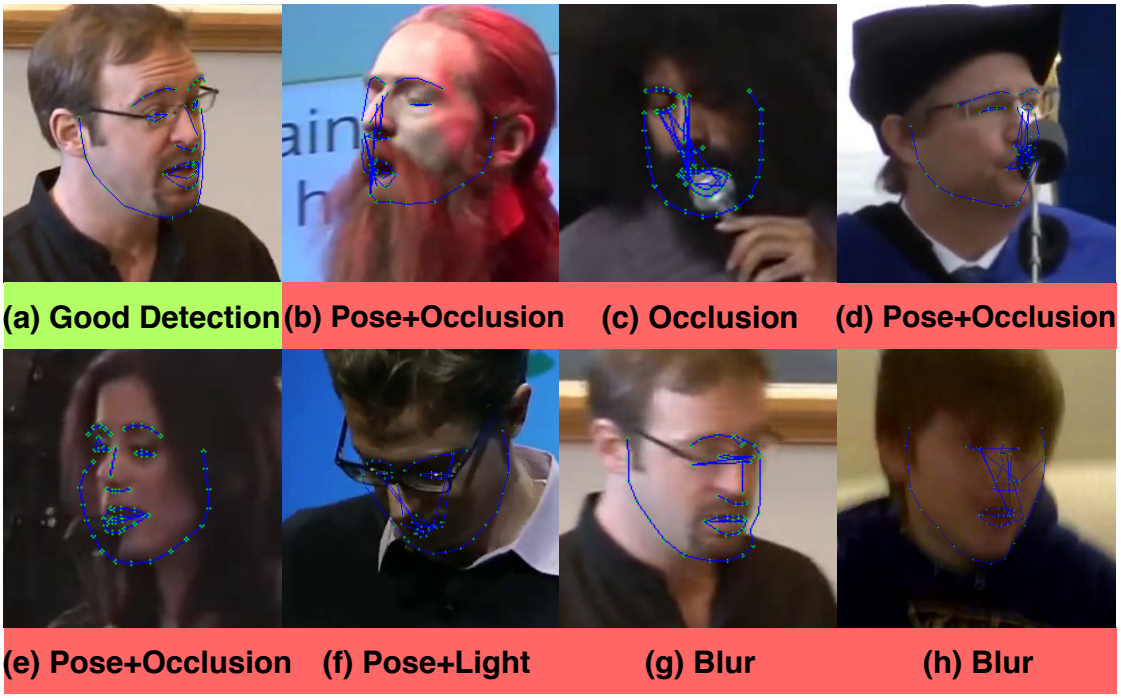


FIGURE 4.2: Examples of HRNet detection on 300VW-S3.

	COFW	300W	300W-Test	AFLW	WFLW
Num. Landmarks	29	68	68	19	98
Num. Train Images	1,345	3,148	/	20,000	7,500
Num. Valid Images	507	689	600	4,386	2,500
HRNet $FR_{0.1}$ (%)	**0.19**	**0.44**	**0.33**	**0.046**	3.12
FR (%) per Image	**0.19**	**0.15**	**0.33**	**0.023**	0.040
FR (%) per Landmark	0.0068	0.0021	0.0025	0.0012	0.00041

TABLE 4.1: Numerical details of the facial landmark datasets and the FR of HRNet on each dataset.

on a video stream. We propose a more general approach regularizing the output shape of Heatmap Regression Models by imposing additional geometric and global contextual constraints during training, directly integrated into the loss function. This adds no complexity during inference and can be trained on both image and video datasets.

Problem of current evaluation metrics for robustness: The most common metric for robustness is Failure Rate (FR). It measures the proportion of images in a (validation) set whose error is greater than a threshold. Table 4.1 shows the FR with an error threshold of 0.1 ($FR_{0.1}$) of HRNet. We can see that this widely used $FR_{0.1}$ measure is almost "saturated" on several benchmarks such as COFW [Burgos-Artizzu et al., 2013], 300W [Sagonas et al., 2013], 300W-Test and AFLW [Koestinger et al., 2011]. That is, there are only 1 , 3 , 1 and 2 failure images respectively (bold numbers in Tab. 4.1). This means that there are only very few challenging images for the state-of-the-art model HRNet in these datasets. At this level, this indicator is saturated and becomes difficult to interpret when comparing the robustness of different methods as it is sensitive to random statistical variations. Therefore, it becomes necessary to modifiy the current evaluation metrics on these datasets and to find more challenging evaluation protocols to further decrease the gap with real-world application settings.

4.3 Proposed evaluation metrics

Dataset: The dataset is crucial to evaluate the robustness of the model. The most common robustness issues treated in the literature concern partial occlusions and large pose variations. COFW [Burgos-Artizzu et al., 2013] is one of the first datasets that aims at benchmarking the performance of facial landmark detection under partial occlusion. 300W [Sagonas et al., 2013] comprises a challenging validation subset with face images with large head pose variations, heavy occlusion, low resolution and complex lighting conditions. AFLW [Koestinger et al., 2011] is a large-scale dataset including face images in extreme poses. WFLW [Wu et al., 2018b] is a recently released dataset with even more challenging images. All the images are annotated in a dense format (98 points). The validation set of WFLW is further divided into 6 subsets based on the different difficulties such as occlusion, large pose or extreme expressions. 300VW [Shen et al., 2015] is a video dataset annotated in the same format as 300W. The validation dataset is split into three scenarios, where the third one (300VW-S3) contains the videos in highly challenging conditions.

Current Evaluation metrics: The main performance indicator for facial landmark detection is the Normalized Mean Error: $\text{NME} = \frac{1}{N}\sum_i \text{NME}_i$, an average over all N images of a validation set, where for one image i the error is averaged over all M landmarks:

$$\text{NME}_i = \frac{1}{M}\sum_j \text{NME}_{i,j} \,, \tag{4.1}$$

and for each landmark j:

$$\text{NME}_{i,j} = \frac{\left\| \mathbf{S_{i,j}} - \mathbf{S^*_{i,j}} \right\|_2}{d_i} \,, \tag{4.2}$$

where $\mathbf{S_{i,j}}, \mathbf{S^*_{i,j}} \in \mathbb{R}^2$ denote the j-th predicted and the ground truth landmarks respectively. For each image, we consider the inter-occular distance as normalization distance d_i for 300W, 300VW, COFW, WFLW and the face bounding box width for AFLW.

As mentioned before, Failure Rate FR_θ measures the proportion of the images in the validation set whose NME_i is greater than a threshold θ. We will denote this classical failure rate: FR^I in the following. In the literature, $\text{FR}^I_{0.1}$ and $\text{FR}^I_{0.08}$ are the principle metrics to measure the prediction robustness as they focus on rather large errors (i.e. 8%/10% of the normalization distance).

It is also very common to compute the FR^I_θ over the entire range of θ, called the Cumulative Error Distribution (CED), which gives an overall idea on the distribution of errors over a given dataset. Finally, for easier quantitative comparison of the performance of different models the total area under the CED distribution can be computed, which is usually denoted as the Area Under Curve (AUC).

We propose three modifications to these measures:

Landmark-wise FR: Instead of computing the average failure rate per image: FR^I, we propose to compute this measure *per landmark*. That is, for each landmark j, the proportion of images with an $\text{NME}_{i,j}$ larger than a threshold is determined. Finally, an average over all landmarks is computed, called FR^L in the following. There are two advantages of computing the failure rate in this way: (1) With Heatmap Regression Models, it happens that only one or few landmarks are not detected well (outliers). However, the NME_i *per image* may still be small because the rest of the landmarks are predicted with high precision and an average is computed per image. Thus, possible robustness problems of some individual landmarks are not revealed by the FR^I measure. (2) FR^L can provide a finer granularity for model comparison, which is notably beneficial when the state-of-the-art

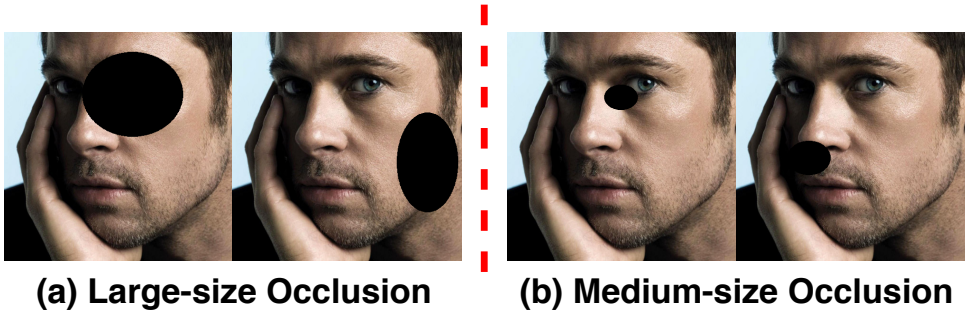


FIGURE 4.3: An illustration of proposed synthetic occlusion protocol.

methods have an FR^I that is very close and almost zero on several benchmark datasets (see Tab. 4.1).

Cross-dataset validation: Leveraging several datasets simultaneously is not new and has already been adopted by some previous works [Smith and Zhang, 2014, Zhu et al., 2014, Zhang et al., 2015, Wu and Yang, 2017, Wu et al., 2018b]. Most of them focus on unifying the different semantic meanings among different annotation formats. In [Zhu et al., 2019a], the authors validated the robustness of their model by training on 300W and validating on the COFW dataset.

We assume the reason why the performance of HRNet has "saturated" on several datasets is that the data distributions in the training and validation subsets are very close. Therefore, to effectively validate the robustness of a model, we propose to train it on a small dataset and test on a different dataset with more images to avoid any over-fitting to a specific dataset distribution. Thus, two important aspects of robustness are better evaluated in this way: firstly, the *number* of possible test cases, which reduces the possibility to "miss out" more rare real-world situations. And secondly, the generalisation capacity to different data distributions, for example corresponding to varying application scenarios, acquisition settings etc.

We propose four cross-dataset validation protocols: COFW→AFLW (trained on COFW training set, validated on AFLW validation set with 19 landmarks), 300W→300VW, 300W →WFLW and WFLW→300VW. The annotation of 300W and 300VW has identical semantic meaning. On the other three protocols, we only measure the errors on the common landmarks between two formats. There are indeed slight semantic differences on certain landmarks. However, in our comparing study this effect is negligible because: (1) We mainly focus on the large errors when validating the robustness. That is, these differences are too small to influence the used indicators such as $FR^L_{0.1}$. (2) When applying the same protocol for each compared model, this systematic error is roughly the same for all models.

Synthetic occlusion: Occlusion is a big challenge for robust facial landmark detection. However, annotating the ground truth positions of occluded facial landmarks is very difficult in practice. To further evaluate the robustness of the model against occlusions, we thus propose to apply synthetic occlusions on the validation images. More specifically, a black ellipse of random size is superposed on each image at random positions. We adopt two protocols: large-size occlusion and medium-size occlusion, illustrated in Fig. 4.3. Obviously, the landmark detection performance of a model is deteriorated by such synthetic occlusions. But more robust models should be resilient to this type of noise by leveraging contextual information, and the growth of NME and FR should be less significant.

4.4 Proposed method

We propose to add geometric and global constraints during the training of Heatmap Regression Models. Our method consists of the following three parts:

2D Wasserstein Loss: Sun et al. [Sun et al., 2018] discussed the use of different loss functions for Heatmap Regression Model. The most widely used loss function is heatmap $L2$ loss. It simply calculates the $L2$ norm of the pixel-wise value difference between the ground truth heatmap and the predicted heatmap.

We propose to train Heatmap Regression Models using a loss function based on the Wasserstein distance. Given two distributions u and v defined on M, the first Wasserstein distance between u and v is defined as:

$$l_1(u,v) = \inf_{\pi \in \Gamma(u,v)} \int_{M \times M} |x-y| \mathrm{d}\pi(x,y), \tag{4.3}$$

where $\Gamma(u,v)$ denotes the set of all joint distributions on $M \times M$ whose marginals are u and v. The set $\Gamma(u,v)$ is also called the set of all couplings of u and v. Each coupling $\pi(x,y)$ indicates how much "mass" must be transported from the position x to the position y in order to transform the distributions u into the distribution v.

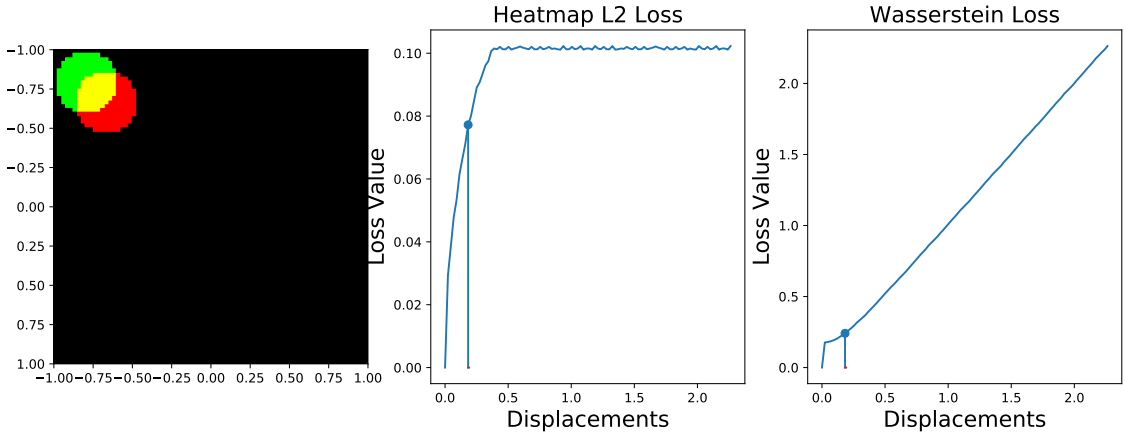
Intuitively, the Wasserstein distance can be seen as the minimum amount of "work" required to transform u into v, where "work" is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved. This notion of distance provides additional geometric information that cannot be expressed with the point-wise $L2$ distance (see Fig. 4.1).

To define our Wasserstein loss function for heatmap regression, we formulate the continuous first Wasserstein metric for two discrete 2D distributions u', v' representing a predicted and ground truth heatmap respectively:

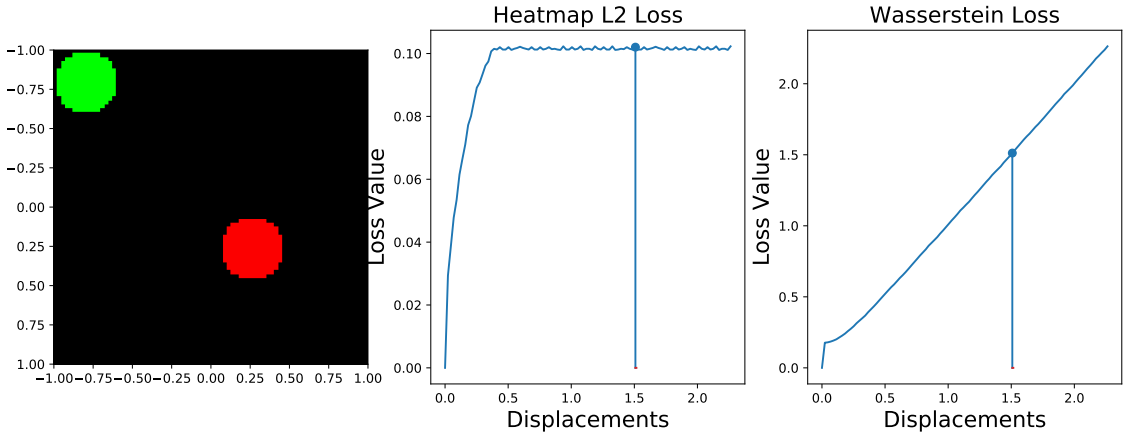
$$L_W(u,v) = \min_{\pi' \in \Gamma'(u,v)} \sum_{x,y} |x-y|_2 \, \pi'(x,y) \tag{4.4}$$

where $\Gamma'(u,v)$ is the set of all possible 4D distributions whose 2D marginals are our heatmaps u and v, and $|\cdot|_2$ is the Euclidean distance. The calculation of the Wasserstein distance is usually solved by linear programming and considered as NP-hard. However, Cuturi [Cuturi, 2013] proposed to add an entropic regularization and calculate an approximation of the loss by Sinkhorn iteration. This drastically accelerates the calculation and enables the gradient back-propagation through the loss calculation. Further, in our case, having discrete 2D distributions of size 64^2 leading to a joint size of $64^4 \approx 1.6710^7$ (for "weights" and distances) as well as existing GPU implementations [Viehmann, 2019, Daza, 2019] make the computation tractable. A visual comparison of Wasserstein Loss and heatmap $L2$ loss on 2D distribution is presented in Fig. 4.4.

Using Wasserstein loss for Heatmap Regression Model has two advantages: (1) It makes the regression sensitive to the global geometry, thus effectively penalizing predicted activations that appear far away from the ground truth position. (2) When training with the $L2$ loss, the heatmap is not strictly considered as a distribution as no normalisation applied over the map. When training with the Wasserstein loss, the heatmaps are first passed through a softmax function. That means the sum of all pixel values of an output heatmap is normalized to 1, which is statistically more meaningful as each normalised value represents the probability of a landmark being at the given position. Moreover, when passed through a softmax function, the pixel values on a heatmap are projected to the e-polynomial space. This highlights the largest pixel value and suppresses other pixels whose values are inferior.



(A) Ground truth (green) and predicted (red) distributions **overlap**.



(B) Ground truth and predicted distributions **do not overlap**.

FIGURE 4.4: Comparison of heatmap $L2$ loss and Wasserstein loss on 2D distributions. We observe that the value of $L2$ loss saturates when the two distributions do not overlap. However, the value of Wasserstein Loss continues to increase. The Wasserstein loss is able to better integrate the global geometry on the overall heatmap and the gradient becoms more geometrically meaningful. (Figure taken from [Tralie, 2018] with slight modifications.)

Smoother target heatmaps: To improve convergence and robustness, the values of the ground truth heatmaps of Heatmap Regression Models for facial landmark detection are generally defined by 2D Gaussian functions, where the parameter σ is commonly set to 1 or 1.5 (see Fig. 4.5).

Intuitively, enlarging σ will implicitly force the Heatmap Regression Model to consider a larger local neighborhood in the visual support throughout the different CNN layers. Therefore, when confronting partial interferences (e.g. occlusion, bad lighting conditions), the model should consider a larger context and thus be more robust to these types of noise. Nonetheless, the Gaussian distribution should not be too spread out to ensure some precision and to avoid touching the map boundaries.

Figure 4.6 shows an example comparing the output heatmaps from a vanilla HRNet (trained with $L2$ loss, $\sigma = 1$) and our HRNet (trained with Wasserstein loss, $\sigma = 3$). We observe that our training strategy effectively removes the spurious activation on the unrelated regions, so that the prediction will be more robust. We empirically found that $\sigma = 3$ is an appropriate setting for facial landmark detection. In our experiments, we systematically demonstrate the effectiveness of using $\sigma = 3$ compared to $\sigma = 1$ or $\sigma = 1.5$ for robust landmark detection under challenging conditions.

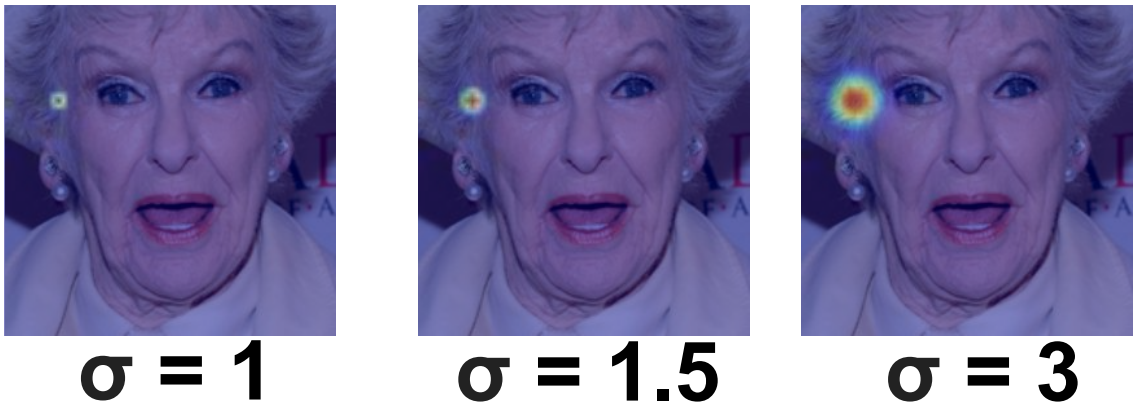


FIGURE 4.5: Illustration of ground truth target heatmaps defined by Gaussian functions with different σ.

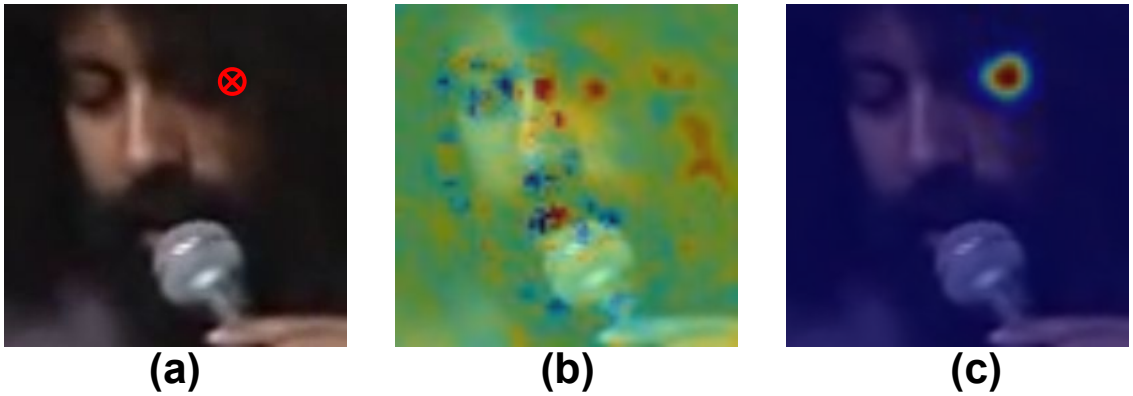


FIGURE 4.6: Comparison of the output heatmaps under challenging conditions. (a) The input image (frame No. 35, video No. 533 of 300VW dataset) with partial occlusion on the right eye. We visualize the output heatmap of the 46th landmark (outer corner of the right eye, marked in red). (b) Output heatmap given by vanilla HRNet (trained with $L2$ loss & $\sigma = 1$). (c) Output heatmap given by our HRNet (trained with Wasserstein loss & $\sigma = 3$).

Predicted landmark sampling: In the early work of Heatmap Regression Model [Newell et al., 2016, Bulat and Tzimiropoulos, 2017b], the position of a predicted landmark p is sampled directly at the position of the maximum value of the given heatmap H:

$$(p_x, p_y) = \arg\max_p(H). \tag{4.5}$$

However, this inevitably leads to considerable quantization error because the size of the heatmap is generally smaller than the original image (usually around 4 times). An improvement is to use interpolation and resample the numerical coordinates using 4 neighbouring pixel (bilinear interpolation). We denote this method as "GET_MAX".

Liu et al. discussed in [Liu et al., 2019] that using a target Gaussian distribution with bigger σ decreases the overall NME. Indeed, using bigger σ flattens the output distribution and therefore obfuscates the position of the peak value. As a result, the predictions are locally less precise.

To compensate this local imprecision when using bigger σ, we propose another approach to sample numerical coordinates from the heatmap. Inspired by [Sun et al., 2018],

we propose to use the spatial barycenter of the heatmap:

$$(p_x, p_y) = \int_{q \in \Omega} q \cdot H(q) \ , \tag{4.6}$$

where Ω denotes the set of pixel positions on the heatmap. We denote this method as "GET_BC" (BaryCenter).

GET_BC enables sub-pixel prediction, which effectively improves the local precision of the model trained with Wasserstein loss and big σ. On the other hand, GET_BC considers the entire heatmap and thus involves a global context for a more robust final detection.

4.5 Experiments

In this section, we compare our method with other state-of-the-art methods and realize ablation studies using both traditional evaluation metrics and proposed evaluation metrics. We also apply our method on various Heatmap Regression Models to demonstrate that our method can be directly used for any model structure without any further adjustments.

Effectiveness of barycenter sampling: The GET_BC method for estimating the predicted landmark coordinates is able to significantly improve the precision of the model trained with Wasserstein loss and larger σ (see Tab. 4.2).

σ	Loss	Method	NME (%)	$FR^L_{0.05}$ (%)
1	Heatmap $L2$	GET_MAX	3.34	18.33
		GET_BC	**20.15**	**93.70**
3	Wasserstein	GET_MAX	4.00	24.69
		GET_BC	**3.46**	**19.42**

TABLE 4.2: Performance of HRNet on 300W validation set when using different coordinate sampling methods. GET_BC improves the local precision (see $FR^L_{0.05}$) of the model trained with Wasserstein loss and large σ. However, it harms the performance of the model trained with $L2$ loss.

In contrast, GET_BC is not compatible with the output trained with heatmap $L2$ loss due to two reasons: (1) No normalization is applied on the heatmap when training with $L2$ loss (2) Training with $L2$ is less robust and generally leads to spurious activations far away from the ground truth position (as illustrated in Fig. 4.6), which prevents GET_BC from estimating good positions. Therefore, in the following experiments, we will use GET_MAX for models trained with the $L2$ loss and GET_BC for models trained with the Wasserstein loss.

Comparison with the state-of-the-art methods: We performed an ablation study using a "vanilla" HRNet (trained with heatmap $L2$ loss and $\sigma = 1$) as our baseline. First, we benchmark our method with standard evaluation metrics NME on 300VW in Tab. 4.3, 300W in Tab. 4.4 and WFLW in Tab. 4.5. More results on AFLW and COFW are presented in the supplementary material. Additionally, we also tested a recent method called CoordConv (CC) [Liu et al., 2018b] to integrate geometric information to the CNN. To this end, we replaced all the convolutional layers by CoordConv layers.

On 300VW, our method shows promising performance, especially under challenging conditions on S3. Our method outperforms the state-of-the-art method FHR+STA [Tai et al., 2019] by almost 1% point on scenario 3. Using the Wasserstein loss combined with a larger σ, our method outperforms the vanilla HRNet by a significant margin of 0.39%, 0.15% and 0.5% points on scenario 1, 2 and 3 respectively.

Method	Scenario 1	Scenario 2	Scenario 3
TSTN [Liu et al., 2017a]	5.36	4.51	12.84
DSRN [Miao et al., 2018]	5.33	4.92	8.85
FHR+STA [Tai et al., 2019]	4.42	4.18	**5.98**
SA [Liu et al., 2019]	**3.85**	**3.46**	7.51
HRNet, $\sigma = 1$, $L2$	3.74	3.73	5.49
$\sigma = 3$, $L2$	3.42	**3.58**	5.12
$\sigma = 1$, W Loss	3.41	3.66	5.01
$\sigma = 3$, W Loss	3.39	3.64	**4.99**
$\sigma = 1$, W Loss, CC	3.45	3.61	5.21
$\sigma = 3$, W Loss, CC	**3.35**	3.61	5.05

TABLE 4.3: NME (%) comparision on 300VW. W Loss - Wasserstein Loss. CC - CoordConv.

Method	Common	Challenge	Full
PCD-CNN [Kumar and Chellappa, 2018]	3.67	7.62	4.44
CPM+SBR [Dong et al., 2018b]	3.28	7.58	4.10
SAN [Dong et al., 2018a]	3.34	6.60	3.98
DAN [Kowalski et al., 2017]	3.19	5.24	3.59
LAB [Wu et al., 2018b]	2.98	**5.19**	3.49
DCFE [Valle et al., 2018]	**2.76**	5.22	**3.24**
HRNet, $\sigma = 1$, $L2$	2.91	5.11	3.34
$\sigma = 3$, $L2$	3.05	5.28	3.49
$\sigma = 1$, W Loss	2.85	5.13	3.29
$\sigma = 3$, W Loss	3.01	5.30	3.46
$\sigma = 1$, W Loss, CC	**2.81**	**5.08**	**3.26**
$\sigma = 3$, W Loss, CC	2.95	5.22	3.39

TABLE 4.4: NME (%) comparison on 300W validation set. W Loss - Wasserstein Loss. CC - CoordConv.

On 300W, our model shows comparable performance to the state-of-the-art methods. Here, using the Wasserstein loss only achieves a marginal improvement. And using a larger σ even slightly decreases the NME performance. As discussed in Sect. 4.2, the performance of vanilla HRNet has already reached a high level on this dataset. Thus, there are only very few challenging validation images for HRNet. Here, the NME is dominated by a large amount of small errors, which is the disadvantage of using a larger σ, and it can thus no longer reflect the robustness of the models. We will demonstrate the robustness of these models by using cross-dataset validation in the following experiments.

On WFLW, our method outperforms other state-of-the-art methods by using a strong baseline. Nonetheless, our method only achieves marginal improvement compared to the vanilla HRNet. We think that it is because the predictions are already "regularized" by the dense annotation of WFLW. We will analyze this issue in detail in Sect. 4.6.

Cross-dataset validation: We use cross-dataset validation to measure the robustness of HRNet trained on 300W. The landmark-wise CEDs with protocol 300W→WFLW are shown in Fig. 4.7. Results of protocol 300W→300VW is shown in Tab. 4.6. Note that the models we evaluate in Fig. 4.7 and Tab. 4.6 are exactly the same models in Tab. 4.4.

On 300W validation set, as discussed before, our method achieves only marginal improvement. However, when cross-validated on another dataset, the advantage of using Wasserstein loss becomes significant. However, when GET_BC is used, a larger σ still slightly decreases the local precision. As a result, on the less challenging datasets such as

Method	Full	Pose	Expr.	Ilm.	Mkup	Occ.	Blur
ESR [Cao et al., 2014]	11.13	25.88	11.47	10.49	11.05	13.75	12.20
SDM [Xiong and De la Torre, 2013]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
CFSS [Zhu et al., 2015]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
DVLN [Wu and Yang, 2017]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
Wing-Loss [Feng et al., 2018b]	**4.99**	**8.43**	**5.21**	**4.88**	5.26	**6.21**	**5.81**
LAB [Wu et al., 2018b]	5.27	10.24	5.51	5.23	**5.15**	6.79	6.32
HRNet, $\sigma = 1.5$, $L2$	4.60	7.86	4.78	4.57	4.26	5.42	5.36
$\sigma = 3$, $L2$	4.73	7.99	4.97	4.62	4.50	5.51	5.39
$\sigma = 1.5$, W Loss	4.57	7.76	4.80	4.45	4.37	5.38	**5.24**
$\sigma = 3$, W Loss	4.76	8.01	5.08	4.68	4.61	5.56	5.42
$\sigma = 1.5$, W Loss, CC	**4.52**	**7.65**	**4.72**	**4.33**	4.26	**5.27**	5.28
$\sigma = 3$, W Loss, CC	4.82	8.16	5.11	4.68	4.67	5.57	5.45

TABLE 4.5: NME (%) comparison on WFLW. W Loss - Wasserstein Loss. CC - CoordConv.

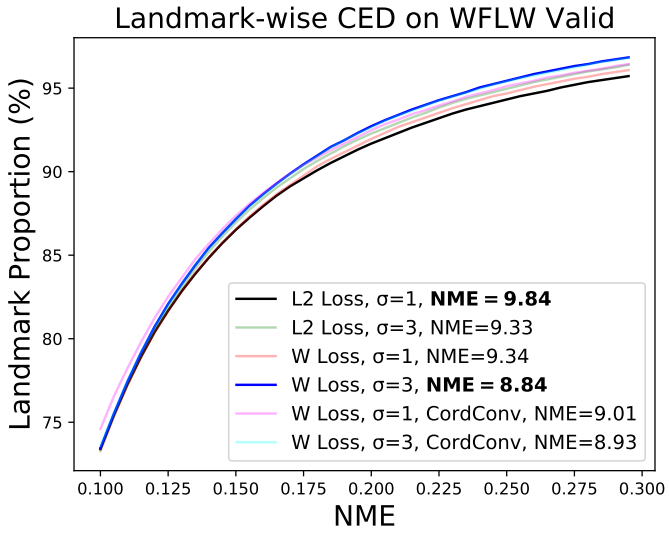


FIGURE 4.7: Landmark-wise CED of 300W→WFLW cross-dataset validation using HRNet.

300VW-S1 and 300VW-S2, we found that the best performance can be obtained by using a combination of small σ, Wasserstein loss and CoordConv. On more challenging datasets such as WFLW and 300VW-S3, the best performance is obtained by using a combination of the Wasserstein loss and a larger σ.

For protocol COFW→AFLW (see Fig. 4.8), our method achieves a bigger improvement on AFLW-All compared to AFLW-Frontal. Note that AFLW-All contains non-frontal images, which is more challenging than AFLW-Frontal.

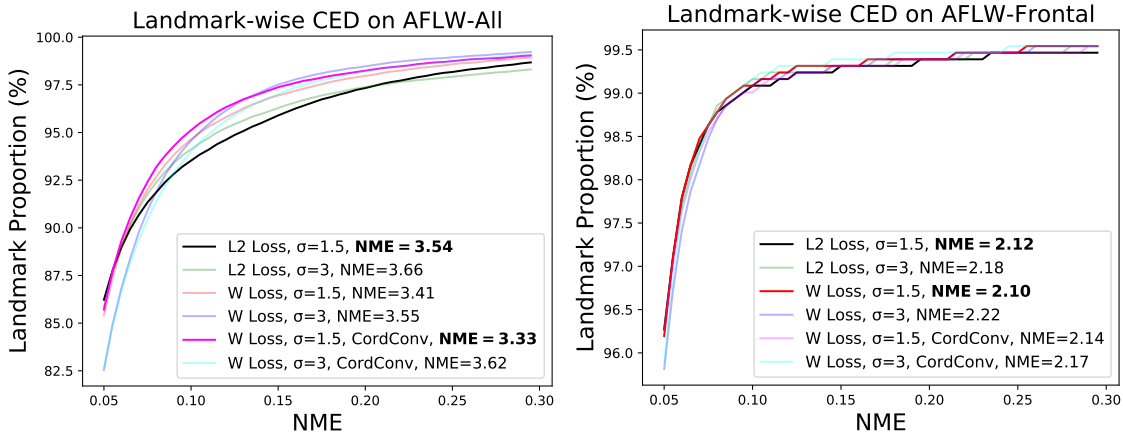
The cross-dataset validation with protocol WFLW→300VW is shown in Fig. 4.9. We observe that by using Wasserstein Loss and CordConv, the HRNet trained on WFLW can be much better generalized on the 300VW dataset.

Synthetic occlusions: We further evaluate the robustness against synthetic occlusion that we described in Sect. 4.3. The increase of NME, FR^I and FR^L on 300W is shown in Tab. 4.7. The model is more robust to occlusion by using a larger σ and Wasserstein loss (except when combining it with CoordConv).

Comparison with other loss functions: We compare our 2D Wasserstein loss function (marked as W) with other loss functions in Tab. 4.8, including standard Heatmap $L2$ loss

Method	Scenario 1		Scenario 2		Scenario 3	
	NME	$FR^L_{0.1}$	NME	$FR^L_{0.1}$	NME	$FR^L_{0.1}$
$\sigma = 1$, $L2$	4.44	5.02	4.37	4.86	6.67	11.65
$\sigma = 3$, $L2$	4.36	4.89	4.38	4.83	6.35	10.97
$\sigma = 1$, W	4.16	4.68	4.21	4.67	6.51	11.08
$\sigma = 3$, W	4.17	4.84	4.16	4.47	**6.01**	9.91
$\sigma = 1$, W, CC	**4.05**	**4.22**	**4.11**	**4.26**	6.32	10.61
$\sigma = 3$, W, CC	4.21	4.78	4.24	4.61	6.02	**9.58**

TABLE 4.6: NME (%) and $FR^L_{0.1}$ (%) comparison of 300W→300VW cross-dataset validation using HRNet. W - Wasserstein Loss. CC - CoordConv.



(A) COFW→AFLW-All (B) COFW→AFLW-Frontal

FIGURE 4.8: Landmark-wise CED of COFW→AFLW cross-dataset validation using HRNet.

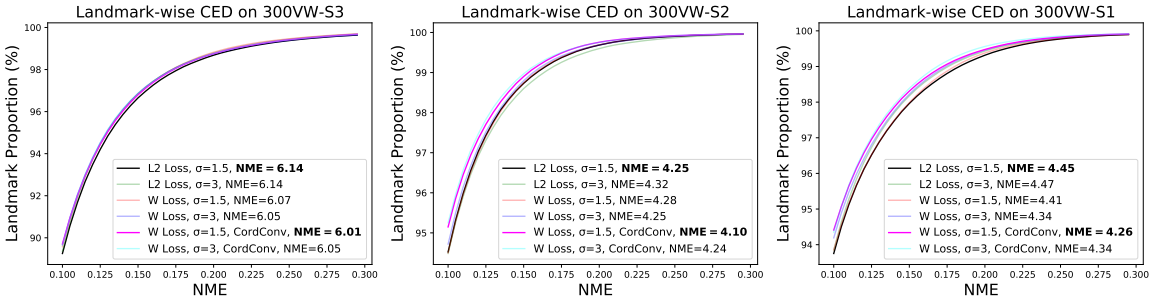
(marked as HM $L2$), Jensen-Shannon divergence Loss (marked as JS), Soft ArgMax loss (marked as Soft AM).

Jensen-Shannon divergence is frequently used to measure the distance between two probabilistic distributions. However, the global geometry information is not explicitly integrated in the calculation of Jensen-Shannon divergence. From Tab. 4.8, we found that the HRNet trained with Wasserstein Loss is more robust than the HRNet trained with Jensen-Shannon divergence loss, especially when cross-validated on 300VW and WFLW dataset (see the FR^I and FR^L).

Soft ArgMax loss function was concurrently proposed by Luvizon et al. [Luvizon et al., 2018] and Sun et al. [Sun et al., 2018] for human pose estimation. The main advantage of this loss is that it makes the ArgMax operation differentiable, so that the numerical coordinates can be directly trained on Heatmap Regression Models. From Tab. 4.8, we oberve that the HRNet trained with Soft ArgMax loss function is slightly more robust than the HRNet trained with Jensen-Shannon divergence loss.

The model trained with Wasserstein Loss is more robust than the models trained with all other loss functions.

Different models: To demonstrate that our method can be used on different Heatmap Regression Models regardless of the model structure, we test our method on three popular Heatmap Regression Models: HourGlass [Newell et al., 2016], CPN [Chen et al., 2018c] and SimpleBaselines [Xiao et al., 2018]. In Fig. 4.10 we can see that all of the three models benefit from our method. This indicates that our approach is quite general and can be applied to most existing Heatmap Regression Models.



(A) WFLW→300VW-S3 (B) WFLW→300VW-S2 (C) WFLW→300VW-S1

FIGURE 4.9: Cross-dataset validation of HRNet trained on WFLW (WFLW→300VW).

Protocol	σ	Loss	CC	NME	$FR^I_{0.08}$	$FR^I_{0.1}$	$FR^L_{0.05}$	$FR^L_{0.08}$	$FR^L_{0.1}$	$FR^L_{0.15}$	$FR^L_{0.2}$
Large	1	L2	✗	1.26	6.71	3.73	8.29	6.59	5.35	3.27	2.03
	3	L2	✗	**1.00**	**4.58**	**2.49**	**7.91**	**5.93**	**4.70**	2.56	1.36
	1	W	✗	1.09	5.10	2.72	8.49	6.22	4.83	2.56	1.36
	3	W	✗	1.02	5.34	2.66	8.84	6.33	4.86	**2.36**	**1.18**
	1	W	✓	1.38	7.64	3.96	9.40	7.36	6.14	3.70	2.24
	3	W	✓	1.21	5.98	3.50	9.85	7.10	5.58	2.87	1.52
Medium	1	L2	✗	0.20	0.82	0.39	1.86	1.21	0.82	0.38	0.20
	3	L2	✗	0.17	0.47	0.34	**1.68**	**0.96**	0.67	0.30	0.11
	1	W	✗	0.18	0.41	**0.11**	1.76	0.99	**0.63**	0.27	0.10
	3	W	✗	**0.16**	0.59	0.17	1.76	0.97	0.70	**0.22**	**0.09**
	1	W	✓	0.23	0.97	0.42	2.09	1.18	0.93	0.41	0.21
	3	W	✓	0.21	**0.38**	0.29	2.20	1.13	0.80	0.26	0.10

TABLE 4.7: Results of the HRNet on 300W validation set with synthetic occlusion. We report the increase (Δ performance) of each indicator compared to non-occluded images. Δ performance is the average value based on the inference run 50 times on the entire validation set.

Visual comparison: We visually compare the predictions from vanilla HRNet and our HRNet on a challenging video clip in Fig. 4.11. Our HRNet gives a more robust detection when confronted to extreme poses and motion blur. By using the Wasserstein loss, a larger σ and GET_BC, the predicted landmarks are more regularized by the global geometry compared to the prediction from the vanilla HRNet.

4.6 Discussions

Does dense annotation naturally ensure the robustness? We find that our method shows less significant improvement on the model trained on WFLW. Intuitively, we presume that by training with a dense annotation (98 landmarks), the model predictions are somewhat regularized by the correlation between neighbouring landmarks. In Tab. 4.9, we compare the models trained with different number of landmarks. The 68 landmark format is a subset of the original 98 landmark format, which is similar to the 300W annotation. The 17 landmark format is a subset of the 68 landmark format, which is similar to the AFLW annotation (except the eye centers). We found that the prediction is naturally more robust by training with denser annotation formats. Therefore, compared to the model trained with sparse annotation, our method achieves less important improvement on the model trained with dense annotation.

Dataset	Loss	Sampling	NME	$FR^{I}_{0.08}$	$FR^{I}_{0.1}$	$FR^{L}_{0.05}$	$FR^{L}_{0.08}$	$FR^{L}_{0.1}$	$FR^{L}_{0.15}$	$FR^{L}_{0.2}$
300W	HM $L2$ †	GET_MAX	3.34	1.60	0.44	18.33	7.60	4.49	1.49	0.58
	JS†	GET_MAX	3.48	1.60	0.44	19.18	8.00	4.80	1.58	0.58
	JS†	GET_BC	3.36	1.74	0.44	18.40	7.79	4.68	1.57	0.57
	Soft AM †	GET_BC	3.69	1.60	0.44	21.66	8.36	4.94	1.54	0.45
	W†	GET_BC	**3.29**	1.60	0.44	**17.89**	7.43	4.47	1.49	0.43
	W*	GET_BC	3.39	1.60	**0.29**	18.49	**7.40**	**4.31**	**1.39**	**0.41**
VW-S3	HM $L2$ †	GET_MAX	6.67	11.72	4.66	44.77	19.26	11.65	4.13	1.85
	JS†	GET_MAX	6.65	11.07	5.14	43.11	19.27	11.97	4.45	2.06
	JS†	GET_BC	6.96	10.87	5.34	42.03	18.93	11.92	4.66	2.28
	Soft AM†	GET_BC	6.46	11.08	5.61	43.71	19.00	11.59	4.45	2.09
	W†	GET_BC	6.51	9.72	3.73	41.38	17.96	11.08	4.15	1.79
	W*	GET_BC	**6.02**	**7.52**	**2.96**	**38.99**	**16.03**	**9.58**	**3.39**	**1.46**
WFLW	HM $L2$ †	GET_MAX	9.84	41.08	25.76	64.08	37.42	26.56	13.47	8.31
	JS†	GET_MAX	9.77	40.48	26.24	63.70	37.60	26.89	13.72	8.37
	JS†	GET_BC	9.93	40.44	26.12	**63.34**	37.49	27.17	14.37	9.04
	Soft AM†	GET_BC	9.34	42.60	26.40	66.09	40.37	29.23	14.86	8.51
	W†	GET_BC	9.34	**39.24**	25.12	63.71	**37.27**	26.61	13.42	8.03
	W*	GET_BC	**8.93**	39.96	**23.64**	64.01	37.42	**26.39**	**12.81**	**7.32**

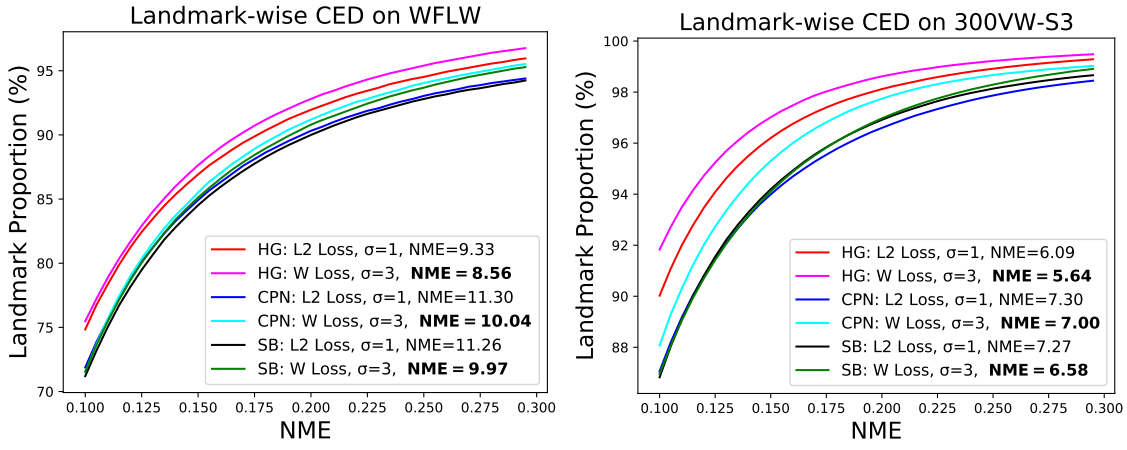
TABLE 4.8: Validation (300W) and cross-dataset validation (300W→300VW-S3 & 300W→WFLW) of the HRNet using different loss functions. †: Trained with Gaussian Distribution $\sigma = 1$ without CoordConv. $*$: Trained with Gaussian Distribution $\sigma = 3$ with CoordConv.

Recommended settings: We recommend to use the Wasserstein loss and GET_BC to improve the robustness of the model in all cases. Using a larger σ will significantly improve the robustness under challenging conditions. Nonetheless, it deteriorates the local precision at the same time. Therefore, we recommend to use a larger σ only when confronting crucial circumstances. When facing less challenging conditions, we recommend to use a combination of Wasserstein loss and small σ. Complementing CoordConv with Wasserstein loss and small σ will further improve the NME performance. However, it adds slight computational complexity to the Heatmap Regression Models. Specifically, when using small σ, the models with CoordConv are less robust against the occlusions compared to those without CoordConv.

Disadvantages: The main disadvantage of using Wasserstein loss is that the loss calculation is relatively time-consuming, even with GPU. We also tested our method for the task of human pose estimation, we do not observe improvement on the MPII dataset [Andriluka et al., 2014]. It is probably due to the fact that human joints have more articulations and left/right confusions than facial landmarks, thus involving limited geometric information and global context.

4.7 Conclusions

In this chapter, we studied the problem of robust facial landmark detection regarding several aspects such as the use of datasets, evaluation metrics and methodology. Due to the performance saturation, we found that the widely used FR and NME measures can no longer effectively reflect the robustness of a model on several popular benchmarks. Therefore, we proposed several modifications to the current evaluation metrics and a novel method to make Heatmap Regression Models more robust. Our approach is based



(A) 300W→WFLW. (B) 300W→300VW-S3.

FIGURE 4.10: Cross-dataset validation of HourGlass(HG) [Newell et al., 2016], CPN [Chen et al., 2018c] and SimpleBaselines(SB) [Xiao et al., 2018].

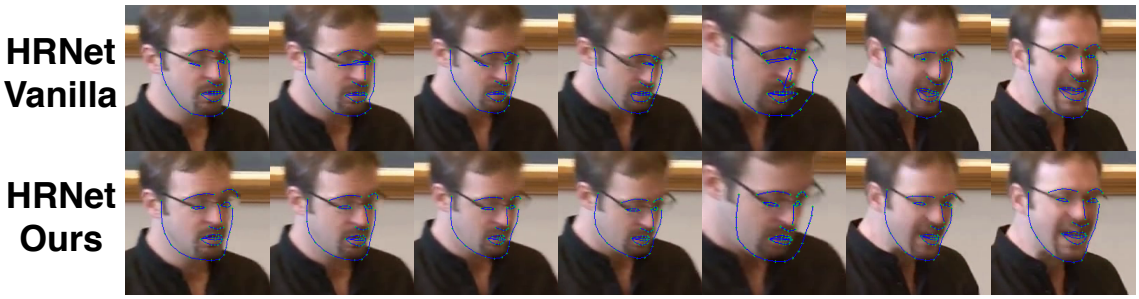


FIGURE 4.11: Visual comparison of vanilla HRNet ($L2$ Loss and $\sigma = 1$) and our HRNet (Wasserstein loss and $\sigma = 3$).

on the Wasserstein loss and involves training with smoother target heatmaps as well as a more precise coordinate sampling method using the barycenter of the output heatmaps.

N. Landmarks	σ	Loss	$FR^L_{0.15}$	$FR^L_{0.2}$
17	1	L2	2.79	1.60
	3	W	**2.68**	**1.29**
68	1	L2	0.65	0.37
	3	W	**0.62**	**0.33**
98	1	L2	0.44	0.25
	3	W	**0.43**	**0.22**

TABLE 4.9: Comparison of the HRNet trained with different number of landmarks on WFLW. To ensure the fair comparison, though trained with different number of landmarks, all the models listed are tested on the common 17 landmarks.

Chapter 5

Facial Landmark Correlation Analysis

After examining two important difficulties of the facial landmark detection in the last two chapters, we find that it is still difficult to quantitatively describe the problem of precision and robustness. The standard evaluation metric for facial landmark detection is Normalized Mean Error (NME). However, it is not able to describe the relationship among the landmarks. Therefore, we are not able to measure if a prediction is overly regularized (such as the locally imprecise output from Coordinate Regression Models, discussed in chapter 3) or insufficiently regularized (such as the unstable output from Heatmap Regression Models, discussed in chapter 4). In this chapter, we present a novel tool to statistically describe the relationship among the landmarks, so that the correlation of predicted landmarks could be quantified. With this tool, we obtain several interesting insights on three important facial landmark detection models, and propose a weakly-supervised learning method to save laborious effort for manual landmark annotation.

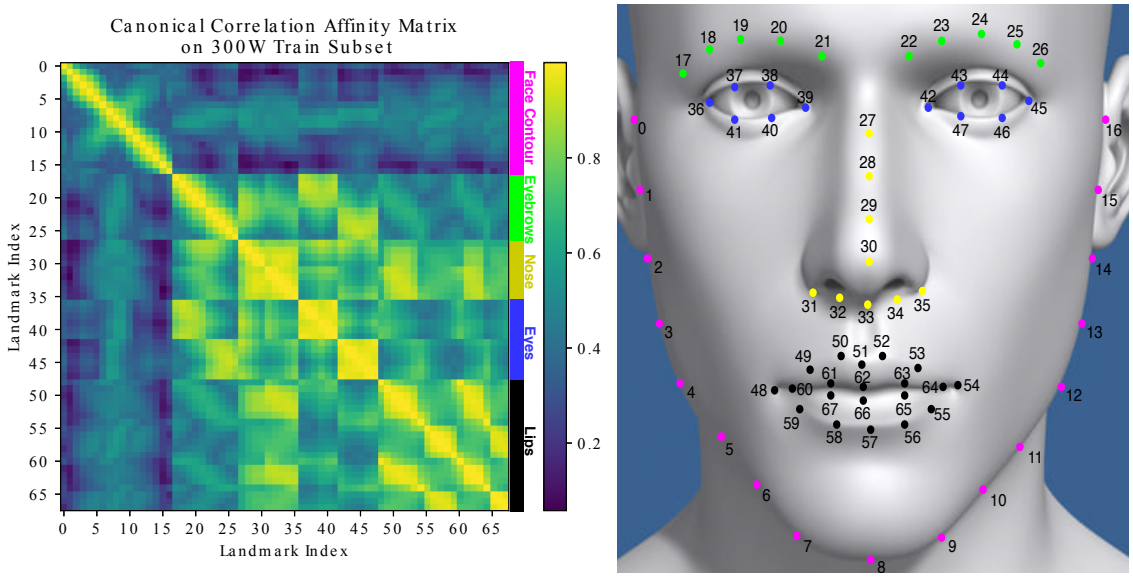
5.1 Introduction

What is facial landmark correlation? Due to the shape and motion of real 3D objects, there exists a natural correlation between landmarks positioned on these objects (e.g. faces, human body, hands or other objects), also in corresponding 2D projections. Especially for faces, the correlation among landmarks is very strong due to the following two reasons: First, the human face is more rigid than the entire human body or hands which have more articulations and may be observed from any point of view and under severe rotations or deformations. Second, recently-released facial landmark datasets are densely annotated with up to 98 landmarks [Wu et al., 2018b], exhibiting an even stronger correlation. Therefore, we focus on the correlation of densely annotated facial landmarks in 300W dataset [Sagonas et al., 2013] (68 landmarks, see Fig. 5.1) and WFLW dataset [Wu et al., 2018b] (98 landmarks).

Motivation of this analysis: The standard evaluation metric for facial landmark detection is the NME. NME is the averaged Euclidean distance between each predicted landmark and ground truth, normalized by the inter-ocular distance. A smaller NME indicates a more precise prediction and vice-versa. Other commonly used metrics, including Failure Rate (FR), Cumulative Error Distribution (CED), and Area Under Curve (AUC), are all based on NME.

However, we think that the NME can not describe all aspects of the model prediction. A large NME can signify that the prediction is not precise, but it can not reflect how the prediction is mistaken. We will illustrate this in the following example.

As stated in Chapter 2.2, current deep learning-based state-of-the-art methods can be categorized into two types: Coordinate Regression Models and Heatmap Regression



(A) CCA Affinity Matrix (B) Landmark Index (300W)

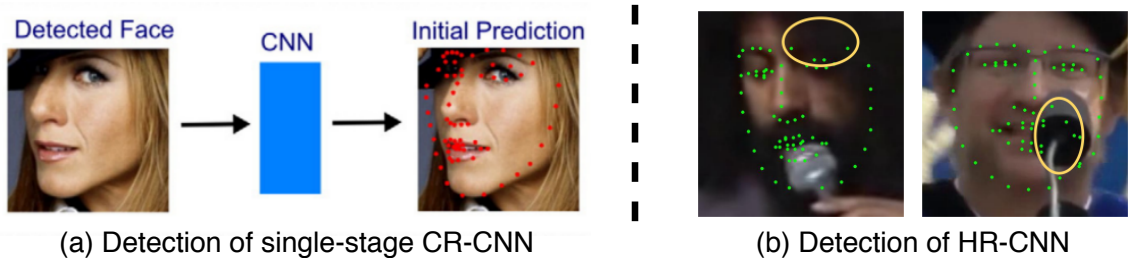
FIGURE 5.1: Facial landmark correlation analysis on the ground truth of 300W train subset [Sagonas et al., 2013]. (a) Canonical Correlation Analysis (CCA) affinity matrix. Bright yellow colored points indicate that the two respective landmarks are highly correlated, and dark blue color indicates low correlation. (b) Illustration of the annotated landmark indices for the 300W dataset. Best viewed in color.

Models [Yan et al., 2018, Wu et al., 2017b]. Coordinate Regression Models predict the numeric X and Y coordinate values of each landmark in the last Fully Connected (FC) layer. Heatmap Regression Models adopt Fully Convolutional Neural Network (FCNN) [Long et al., 2015] architectures that estimate a spatial probability map for each landmark. That is, the value of each pixel on the heatmap represents the presence probability of the landmark at this pixel [Wei et al., 2016].

Each model has its strengths and weaknesses. Heatmap Regression Models show a strong capability of handling complex pose variations. However they globally lack robustness, and in failure cases, landmarks are predicted at unreasonable positions which are far away from the ground truth (see Fig. 5.2 (b)). On the other hand, Coordinate Regression Models are generally more efficient in terms of computation and memory usage but also locally less precise. The prediction of single-stage Coordinate Regression Models is usually constrained in a reasonable shape similar to the ground truth, being not extremely precise (see Fig. 5.2 (a)). This investigation can be confirmed by the current research trend. Most of the latest Heatmap Regression Models aim at reinforcing the robustness of the detection by introducing global constraints [Valle et al., 2018, Liu et al., 2019, Merget et al., 2018] or temporal consistency [Tai et al., 2019]. However, the recent Coordinate Regression Models enhance local precision using coarse-to-fine frameworks [Trigeorgis et al., 2016, Fan and Zhou, 2016, Lv et al., 2017, Chen et al., 2017a, Kowalski et al., 2017, He et al., 2017b, Feng et al., 2018b].

For instance, the models whose results are illustrated in Fig. 5.2 (a) and (b) may have similar NME. Nevertheless, these two models have distinct characteristics.

In this case, we think that landmark correlation can be the key to explain and, furthermore, to quantify the weaknesses of the two models. We assume that the local imprecision problem of Coordinate Regression Models is due to the fact that the predicted landmark positions are too much correlated (or regularized). In contrast, for Heatmap Regression Models, the "outliers" predicted in unreasonable positions can be considered



(a) Detection of single-stage CR-CNN (b) Detection of HR-CNN

FIGURE 5.2: Illustration of the weaknesses of single-stage Coordinate Regression Models and Heatmap Regression Models. Figure (a) is taken from [Fan and Zhou, 2016] and figure (b) is taken from [Liu et al., 2019].

as a violation of the natural landmark correlation.

We want to make clear that landmark correlation can not be used as a stand-alone evaluation metric, though it provides a new perspective to interpret the model prediction. Similar correlation compared to the ground truth is a necessary condition but not sufficient to precise prediction. Even identical correlation does not ensure precise prediction. However, big correlation difference between prediction and ground truth can conclude that the prediction is not precise. Our contributions in this chapter can be summarized as follows:

- We present a CCA-based correlation analysis as a novel tool to interpret and quantify the relationship among a set of landmarks (section 5.3).
- We use this model-agnostic correlation analysis to interpret the three most popular facial landmark detection models in the last decade, including cascaded random forest, Coordinate Regression Models and Heatmap Regression Models (section 5.4).
- We propose a weakly-supervised learning method to reduce the effort of manual annotation of dense landmarks with the help of the landmark correlation (section 5.5). By analyzing the landmark correlation in the dense formats, we are able to form a sparse format by selecting a set of landmarks which are most correlated to the rest of them. We propose to learn dense facial landmark predictions by the images annotated with the sparse format, which requires less annotation cost. Our method shows two advantages: **(i)** Compared to existing methods which use existing sparse formats [Belhumeur et al., 2013, Zhang et al., 2016c, Van Gool, 2012, Burgos-Artizzu et al., 2013, Koestinger et al., 2011], the selection of our sparse format is purely data-driven. **(ii)** The number of sparse landmarks can be arbitrarily chosen depending on the minimum correlation required between the selected landmarks and the rest.

5.2 Related Work

Facial landmark detection in the last decade: A detailed literature review is provided in Chapter 2.2. In this section, we only present a brief review. In 2010, Dollár *et al.* proposed Cascaded Pose Regression [Dollár et al., 2010], which laid the foundation for several well-known cascaded regression methods including SDM [Xiong and De la Torre, 2013], ESR [Cao et al., 2014] and ERT [Kazemi and Sullivan, 2014]. In the deep learning era, cascaded Coordinate Regression Models [Sun et al., 2013, Zhang et al., 2014a, Trigeorgis et al., 2016] continue to follow its general coarse-to-fine structure. Heatmap Regression Models [Wei et al., 2016, Newell et al., 2016, Bulat and Tzimiropoulos, 2017b],

originally introduced in 2005 [Duffner and Garcia, 2005a], gained much popularity in recent years. In this chapter, we propose facial landmark correlation analysis to take a closer look into three of the most important models in the last decade: Cascaded Random Forest model [Kazemi and Sullivan, 2014], Coordinate Regression Model [Fan and Zhou, 2016] and Heatmap Regression Model [Bulat and Tzimiropoulos, 2017b].

Component analysis in facial landmark detection: The use of Principal Component Analysis (PCA), especially the 3DMM model [Blanz et al., 1999], is of great importance in the current research of face analysis. PCA has been used for facial landmark detection since 1995 [Cootes et al., 1995] in the Point Distribution Model. PCA is used to analyze the shape variance with respect to the mean shape, including face rotation, facial expressions and identity variance. The biggest difference between the PCA and our CCA study is that our CCA study analyzes the relationship among individual facial landmarks while PCA focuses on the global face shape.

CNN Interpretation via CCA: Lately, using CCA to interpret CNN representations [Raghu et al., 2017, Morcos et al., 2018, Kornblith et al., 2019] is an emerging subject. They used CCA to analyze the representations of two different CNNs and gained some insights on the learning process. They mainly focused on attenuating the noise in the CNN representation, which is brought by different initializations. However, as we will show in Sect. 5.4.4, when we use CCA to analyze the Coordinate Regression Models, we analyse the correlation between different neurons in the same layer. As being trained altogether, no such noise will be involved. Therefore, we do not apply any pre-preprocessing steps such as Singular Value Decomposition (SVD) as in [Raghu et al., 2017].

Weakly-supervised learning for facial landmark detection: Weakly supervised learning, or few-shot learning, is now attracting increasing attention in the community. A recent work [Dong and Yang, 2019] proposed a mechanism to enable the training on fewer labeled images. Differently, we focus on how to learn with fewer *landmarks* rather than fewer *images*.

We assume that a landmark can be easily transferred from another landmark that is highly correlated. This is not new and has already been proved in several work which focus on transferring the data between two annotations with different semantic meanings [Smith and Zhang, 2014, Zhu et al., 2014, Zhang et al., 2015]. We also find similar ideas in some existing coarse-to-fine approaches [Lv et al., 2017, Chen et al., 2017a, Shao et al., 2016, Shao et al., 2019], where the entire set of landmarks is divided into several partitions inside which the authors assume a strong correlation. Specifically, Tan *et al.* [Tan et al., 2017] proposed a weakly-supervised learning method to reconstruct the global shape from a sparse landmark format. DeCaFA [Dapogny et al., 2019] can be trained with coarsely annotated examples by exploiting landmark-wise attention. However, in the above works, they mainly focused on how to improve the model performance given the pre-defined sparse format. The choice of the sparse format and their partitioning are heuristic. In contrast, we focus on how to find the best sparse format that will most benefit the weakly-supervised learning. Our selection of the sparse landmark format is entirely based on the statistics of the underlying data. Our approach is inspired by the work on multi-task learning [Zamir et al., 2018, Li et al., 2019].

5.3 Facial Landmark Correlation Analysis

5.3.1 Canonical Correlation Analysis [Hotelling, 1936]

Given a p-dimensional random variable $\mathbf{U} \in \mathbb{R}^p$ and a q-dimensional variable $\mathbf{V} \in \mathbb{R}^q$, CCA aims to find the best linear transformation $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ that maximize the

correlation:

$$\mathrm{Cor}(\mathbf{U}, \mathbf{V}) = \frac{\mathbf{a}^T \sum_{\mathbf{UV}} \mathbf{b}}{\sqrt{\mathbf{a}^T \sum_{\mathbf{UU}} \mathbf{a}} \cdot \sqrt{\mathbf{b}^T \sum_{\mathbf{VV}} \mathbf{b}}}, \tag{5.1}$$

where

$$\sum_{\mathbf{UV}} = \mathrm{Cov}(\mathbf{U}, \mathbf{V}) = \mathrm{E}[(\mathbf{U} - \mathrm{E}[\mathbf{U}])(\mathbf{V} - \mathrm{E}[\mathbf{V}])]. \tag{5.2}$$

The operator E denotes the expected value of its argument. This problem can be solved by SVD after basis change. This gives $\min(p, q)$ correlation coefficients sorted from the most correlated to the least correlated canonical directions. We consider the mean value $\overline{\mathrm{Cor}(\mathbf{U}, \mathbf{V})}$ of the correlation coefficients as an overall measure [Raghu et al., 2017].

5.3.2 Facial Landmark Correlation

To focus on the variance of the face shape, we apply an important pre-processing step. We crop, center all the faces and then further normalize their sizes. We consider the 2D Cartesian coordinates as a two-dimensional variable. Specifically, we calculate the absolute value of the correlation coefficients (ranged from -1 to 1) as we are interested in the magnitude of the correlation between two landmarks but not their directions.

To be clear, the canonical correlation between the i-th and the j -th landmark in the annotation format can be found at the i-th row and the j-th column on the affinity matrix $\mathbf{A}$:

$$\mathbf{A}_{i,j} = \left| \overline{\mathrm{Cor}(\mathbf{L_i}, \mathbf{L_j})} \right|, \tag{5.3}$$

where $\mathbf{L_i}, \mathbf{L_j} \in \mathbb{R}^2$ indicate the annotation of the i-th and the j-th landmark on the entire dataset.

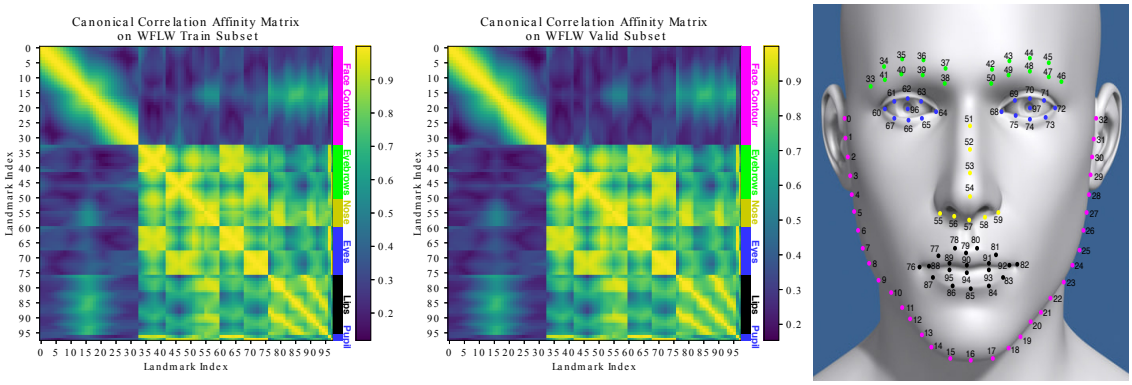
The correlation affinity matrix on the 300W train subset [Sagonas et al., 2013] and WFLW dataset [Wu et al., 2018b] is shown in Fig. 5.1 and Fig. 5.3. We draw several conclusions from the affinity matrices in the two previous figures: **(i)** The correlation among the landmarks belonging to the same facial component is generally more significant than the others. **(ii)** Some landmarks from the same component are less correlated (such as upper-lip and lower-lip). This is due to the shape variance e.g. different facial expressions. **(iii)** Certain facial components from different facial components are strongly correlated, such as eyebrows and eyes, the outer and inner contour of lips, which is plausible. **(iv)** we find that the landmark correlation on the WFLW train subset and the WFLW valid subset are almost the same. It confirms the universality of facial landmark correlation, which means that our analysis is statistically meaningful.

5.4 Facial Landmark Model Interpretation

We now use the proposed landmark correlation analysis to interpret three important facial landmark detection models: cascaded random forest [Kazemi and Sullivan, 2014], cascaded Coordinate Regression Model [Fan and Zhou, 2016] and Heatmap Regression Model [Bulat and Tzimiropoulos, 2017b]. We will focus on three aspects. **(i)** What are the characteristics of the final prediction from each model? **(ii)** Are there any meaningful differences between cascading and stacking? **(iii)** Can we interpret the learning dynamics of the CNN models for landmark detection?

5.4.1 Model Settings

All of the analyzed models are trained on the 300W train subset and analyzed on 300W validation subset.



(A) CCA Affinity Matrix on WFLW Train (B) CCA Affinity Matrix on WFLW Valid (C) Landmark Index (WFLW)

FIGURE 5.3: Facial landmark correlation analysis on the ground truth of WFLW train subset and valid subset. Bright yellow colored points indicate that the two respective landmarks are highly correlated, and dark blue color indicates low correlation. Best viewed in color.

Cascaded Random Forest: ERT [Kazemi and Sullivan, 2014] consists of 10 cascaded random forest regressors. Each regressor comprises 500 trees and the depth of the trees is 5. We use the implementation from [Xiao, 2019]. The initialized shape is the mean shape of the train subset. The NME of this model is 6.18% for the validation set.

Cascaded Coordinate Regression Model: We reproduced the model of Fan *et al.* [Fan and Zhou, 2016]. However, we added two additional stages to further boost its performance. Therefore, the overall structure has four stages. The main network in the first stage is ResNet18 and the sub-networks in the following stages consist of a single ResNet block and a single FC layer. The NME of this model is 3.66%.

Stacked Heatmap Regression Model: We used the official implementation of [Bulat and Tzimiropoulos, 2017b]. The hourglass models are stacked in 4 stages. The NME of this model is 3.52%.

5.4.2 Characteristics of the Prediction

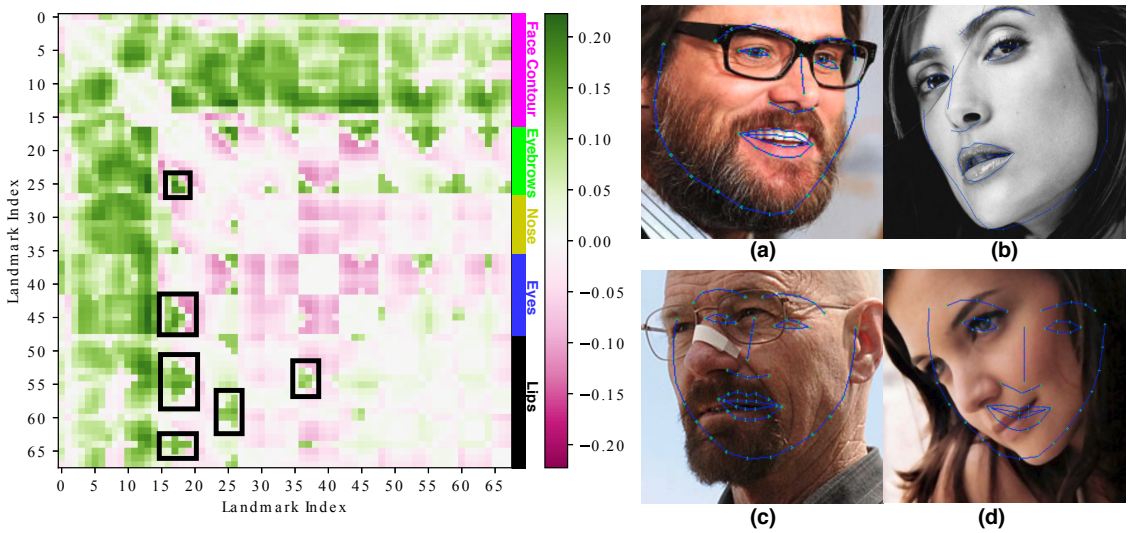
In this section, we visualize the affinity matrix error $\mathbf{A}_{pred} - \mathbf{A}_{GT}$, where $\mathbf{A}_{pred}$ is the CCA affinity matrix calculated on the output of each model and $\mathbf{A}_{GT}$ is the CCA affinity matrix calculated on the ground truth.

Cascaded random forest ERT: In Fig. 5.4, we show the final prediction of ERT through landmark correlation. We can see that the overall correlation of the prediction is higher than the ground truth. There are more green parts than red parts and the shades of the green parts are higher than the shades of red parts. We observe two important points from the CCA matrix error.

(i) The landmarks on the face contour are generally more correlated to the other facial components. It means that the predicted face contour from ERT is too regularized.

(ii) Some landmarks on the right are over-correlated with other landmarks on the left (marked in the black rectangles in Fig. 5.4 (a)). For example, the correlation between the left tip of the left eyebrow (landmark index 17) and the right tip of the lip (landmark index 54) is significantly bigger than the ground truth. It statistically signifies that the prediction of the ERT does not have enough horizontal variance compared to the ground truth, probably due to the failures confronting extreme head poses.

The shown visual examples (Fig. 5.4 (b)) confirm the above investigations on face contours ((a)(b)) and large poses ((c)(d)). Further, our observations are consistent with



(A) Affinity Matrix Error (B) Visual Examples

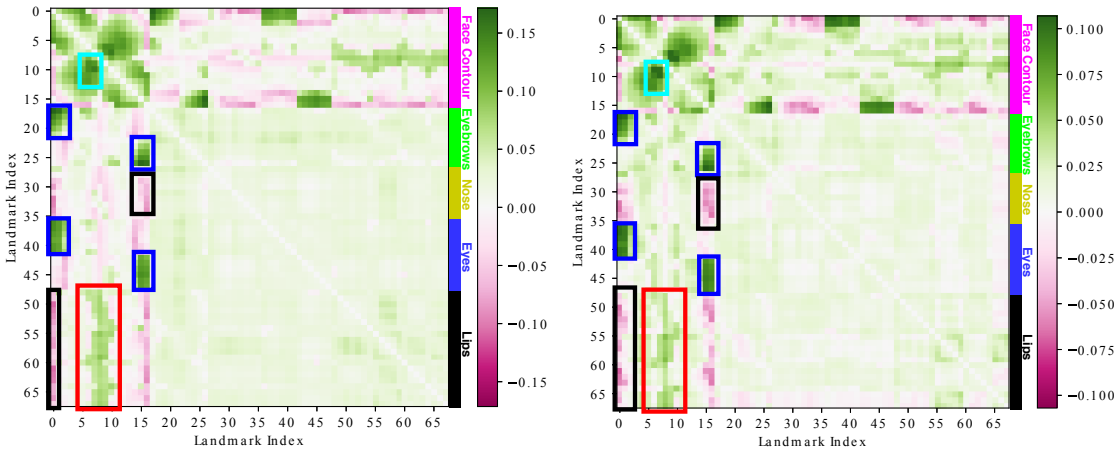
FIGURE 5.4: Facial landmark correlation analysis on the final prediction of ERT.

the major concern about ERT expressed in the literature [Zhu et al., 2016a], which is the poor robustness to pose variations.

Cascaded Coordinate Regression Model & stacked Heatmap Regression Model: In Fig. 5.5, we show the correlation matrix error of Coordinate Regression Model and Heatmap Regression Model. Overall, the prediction of stacked Heatmap Regression Model has a lower correlation error compared to cascaded Coordinate Regression Model. And both of them show a smaller correlation error than ERT (see the scale of colorbar on the right). Also note that the error mainly exists on the face contour.

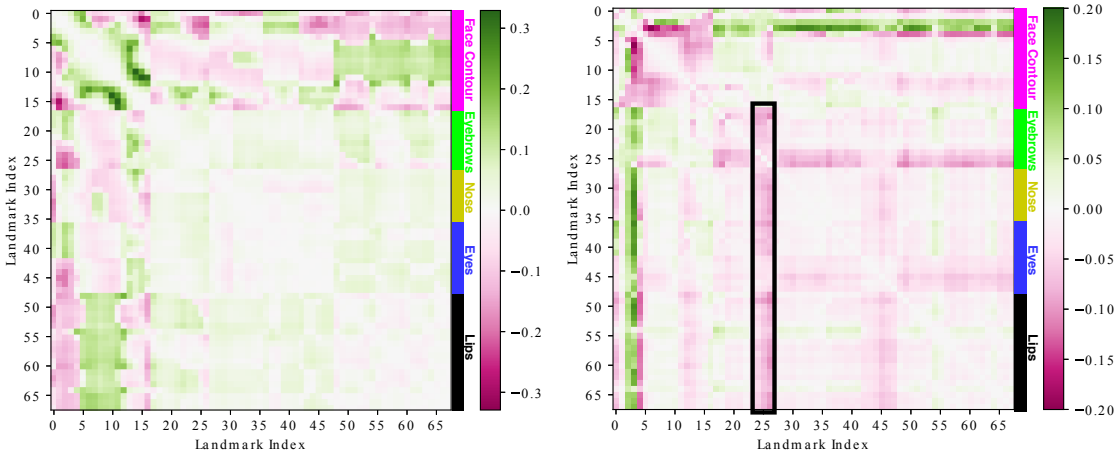
CNN tends to correlate adjacent landmarks. Both of the CNN based methods share this important characteristic. This is probably due to the convolution operation used in the CNN, which excessively exploits local semantic information. For example, in Fig. 5.5, the correlation between the left tip of the left eyebrows/eyes and the upper-left face contour (blue rectangles), the correlation between the lip and the bottom face contour (red rectangles) and the correlation among the landmarks on the bottom face contour (cyan rectangles) are significantly higher than the ground truth correlation. Some landmarks that are over-correlated to their adjacent landmarks, show inferior correlation with the more distant landmarks (the correlation between upper-left face contour and lips, black rectangles).

Heatmap Regression Model is more likely to violate landmark correlation than Coordinate Regression Model under challenging conditions. In Fig. 5.6 (b), we can see that the correlation between the inner facial components on 300VW Scenario3 is weaker than the ground truth, especially on the right eyes/eyebrows (black rectangle). This is consistent with the weakness of Heatmap Regression Model that we mentioned in Fig. 5.2 (b). If we compare Fig. 5.6 (b) and Fig. 5.5 (b), we observed that this problem only happens on 300VW S3, which involves challenging conditions such as occlusions, motion blurs, complex lighting conditions, etc. However, if we compare Fig. 5.6 (a) and Fig. 5.6 (b), we find that Coordinate Regression Model is still robust under these challenging conditions, especially on inner facial components.



(A) Cascaded Coordinate Regression Model (B) Stacked Heatmap Regression Model

FIGURE 5.5: The affinity matrix error of cascaded Coordinate Regression Model and stacked Heatmap Regression Model on 300W valid.



(A) Cascaded Coordinate Regression Model (B) Stacked Heatmap Regression Model

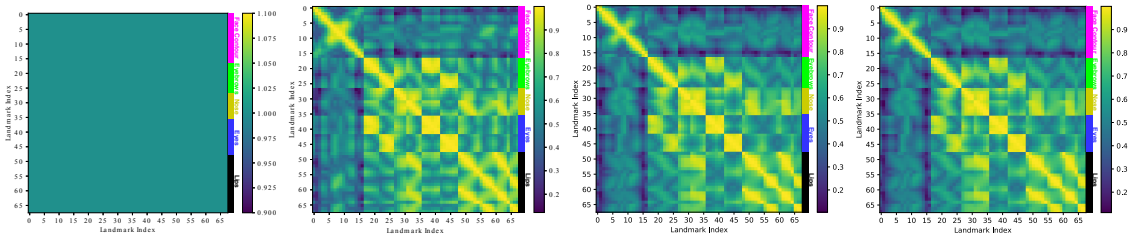
FIGURE 5.6: The affinity matrix error of cascaded Coordinate Regression Model and stacked Heatmap Regression Model on 300VW Scenario3.

5.4.3 Cascading and Stacking

In this section, we provide detailed analysis about the differences between the cascading and the stacking that are used in Cascaded Random Forest (ERT) [Kazemi and Sullivan, 2014], Cascaded Coordinate Regression Model [Fan and Zhou, 2016] and Stacked Heatmap Regression Model [Bulat and Tzimiropoulos, 2017b].

Cascaded Random Forest: In Fig. 5.7, we demonstrate the CCA affinity matrices on the output of the 1st/6th/10th stage. The landmark correlation on the output of the 1st stage is generally higher than the landmark correlation on the output of the 6th and 10th stage.

To further study the use of each stage, we calculate the correlation affinity matrix difference between the input and the output of each stage (see Fig. 5.8). We observe that in most of the stages, there are more red parts than green parts and the shade of the red parts is more intense than the shade of the green parts. It indicates that the cascading



(A) Mean Shape (B) Stage 1 (C) Stage 6 (D) Stage 10

FIGURE 5.7: CCA affinity matrices of the mean shape and the outputs in the 1st/6th/10th stage of Cascaded Random Forest [Kazemi and Sullivan, 2014].

continuously reduces the landmark correlation. Cascading helps to learn the shape variances, which transforms a more regularized shape into a finer shape. For example, from Fig. 5.8 (e), we find that the correlation between the upper lip and lower lip is significantly reduced (black rectangles), indicating that the model is focused on learning the shape variance of mouth open/close in stage 5.

Cascaded Coordinate Regression Model: We illustrate the correlation affinity matrix error between the input and the output of each stage in Fig. 5.9. Similar to the cascading in ERT, the cascading continues to de-correlate a regularized shape to a finer shape, especially on the adjacent landmarks on the face contour (cyan rectangles), lips and bottom face contour (blue rectangles), upper and lower lip (black rectangles).

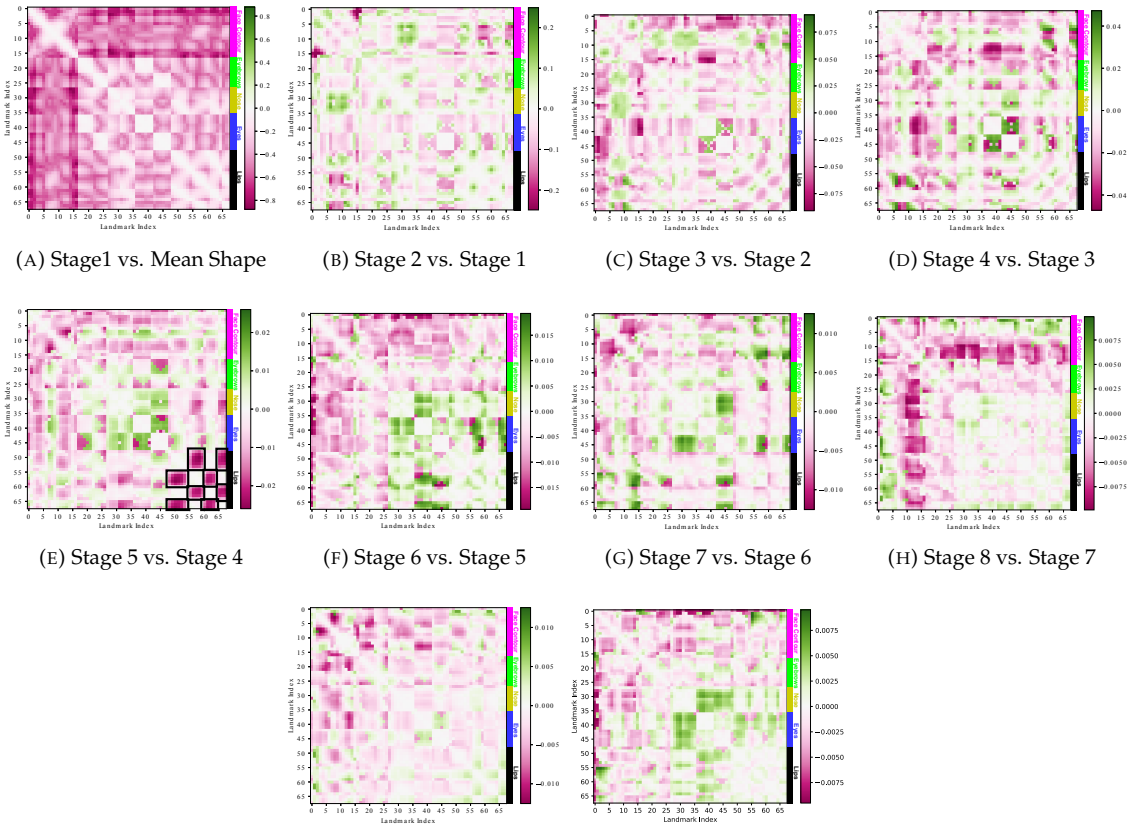
Stacked Heatmap Regression Model: We illustrate the correlation affinity matrix difference between the input and the output of each stage in Fig. 5.10. On the contrary, we do not observe similar correlation reduction as in Cascaded Random Forest and Cascaded Coordinate Regression Model. Furthermore, unlike the cascading, the evolution of the landmark correlation is not grouped in blocks, which means that the evolution of landmark correlation is no longer shared among neighbouring landmarks.

Differences between cascading and stacking: We observe that a single-stage Coordinate Regression Model suffered from more severe over-correlation problem compared to the cascaded Coordinate Regression Model. Therefore, the coarse prediction of the first stage in cascaded Coordinate Regression Model is indeed over-regularized. We note that in Coordinate Regression Models, the output is linearly connected with the previous FC layer. This may explain why there are always excessive correlations present in the output of Coordinate Regression Model.

More generally, we think that the cascading, which is widely used in Cascaded Random Forest and Cascaded Coordinate Regression Model, is served for learning the shape variances step by step. The shape is evolved from a regularized shape (or a mean shape) to a more dynamic shape. The stacking, which is frequently used in Heatmap Regression Model, does not learn further shape variances in the following stages. To delve deeper into this aspect, we believe that it is necessary to develop a two-dimensional CCA study, which is able to analyze the spatial correlation directly on the output heatmap.

When "mouth open/close" is learned in each model? To further investigate how the shape variance is learned in these models, we raise an example on an important shape variance, "mouth open/close". We plot the evolution of the correlation between the 62rd and the 66th landmark (center upper lip & center lower lip) on the output of each stage in all of the three models (see Fig. 5.11).

On the Cascaded Random Forest model (the blue curve), the correlation is no longer reduced since the 5th stage. We believe that the variance "mouth open/close" has been roughly learnt since the 5th stage.



(A) Stage1 vs. Mean Shape (B) Stage 2 vs. Stage 1 (C) Stage 3 vs. Stage 2 (D) Stage 4 vs. Stage 3

(E) Stage 5 vs. Stage 4 (F) Stage 6 vs. Stage 5 (G) Stage 7 vs. Stage 6 (H) Stage 8 vs. Stage 7

(I) Stage 9 vs. Stage 8 (J) Stage 10 vs. Stage 9

FIGURE 5.8: CCA affinity matrix difference on the input and the output of each stage in Cascaded Random Forest [Kazemi and Sullivan, 2014].

On the Cascaded Coordinate Regression Model model (the green curve), the correlation is reduced continuously until the end. In addition, the final correlation is still higher than the ground truth correlation.

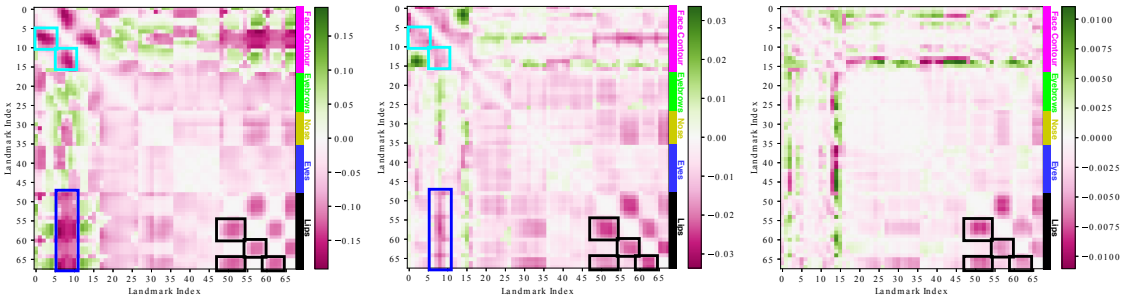
On the Stacked Heatmap Regression Model model (the red curve), the correlation on the output of the first stage is already very close to the ground truth correlation. We believe that the variance "mouth open/close" has been learnt since the first stage of the Stacked Heatmap Regression Model.

5.4.4 Coordinate Regression Model Learning Dynamics

In this section, we study how the Coordinate Regression Model progressively learns from the beginning. To this end, we plot the evolution of the Coordinate Regression Model output correlations during training. We do not analyze the learning dynamic of Heatmap Regression Model due to the operation of taking the maximum value on the final heatmap. We think that it is necessary to develop a 2D CCA method in the future to analyze the output heatmap directly.

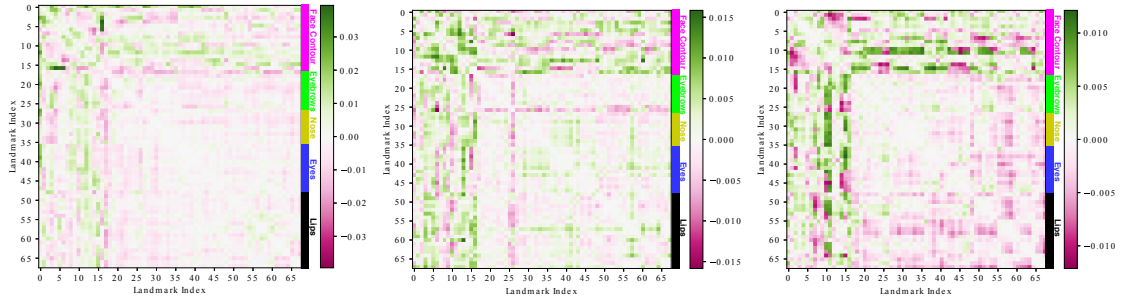
We trained a one-staged ResNet-18 on the 300W dataset. Both the convolutional layers and the FC layers are initialized from a normal distribution [He et al., 2015a]. The Coordinate Regression Model is trained for 350 epochs with the learning rate decayed by 0.3 for each 70 epochs. We observe the following phases during the first 70 epochs:

Phase 1 Group Inner Facial Components: More rigid parts learn first. The first thing that CNN starts to learn is to group the inner facial components. We can observe in



(A) Stage 2 vs. Stage 1 (B) Stage 3 vs. Stage 2 (C) Stage 4 vs. Stage 3

FIGURE 5.9: CCA affinity matrix difference on the input and the output of each stage in Cascaded Coordinate Regression Model [Fan and Zhou, 2016]



(A) Stage 2 vs. Stage 1 (B) Stage 3 vs. Stage 2 (C) Stage 4 vs. Stage 3

FIGURE 5.10: CCA affinity matrix difference on the input and the output of each stage in Stacked Heatmap Regression Model [Bulat and Tzimiropoulos, 2017b]

Fig. 5.12 (b) that CNN firstly learns a relatively strong correlation among the landmarks on the inner facial components (eyebrows, eyes, noses) and separate them from the other landmarks on face contours.

Phase 2 Recognize Each Facial Component: Next, the CNN starts to gradually identify the facial components (eyebrows, nose, mouths, etc.). In this phase, the correlation among the landmarks which belong to the same facial component grow stronger (see Fig. 5.12 (c)). The CNN recognizes the eyes, nose and lips almost simultaneously.

Phase 3 Refine the Prediction: The CNN learns to refine the prediction in two aspects: **(i)** enforce the correlation inside each facial component, especially the neighbouring landmarks; **(ii)** reduce some excessive correlations (e.g. the correlation between lower face contour and lips is reduced, see Fig. 5.12 (d)).

The evolution of the affinity matrices after 70 epochs is difficult to visualize as the evolution of the correlation value is small. To this end, we calculate the standard deviation (Std) of the affinity matrices in different stages (see Fig. 5.13). We observe that after 70 epoch, the variation of the correlation value related to the landmarks on the face contour is significantly higher than the others, indicating that the model struggles learning to refine the face contour.

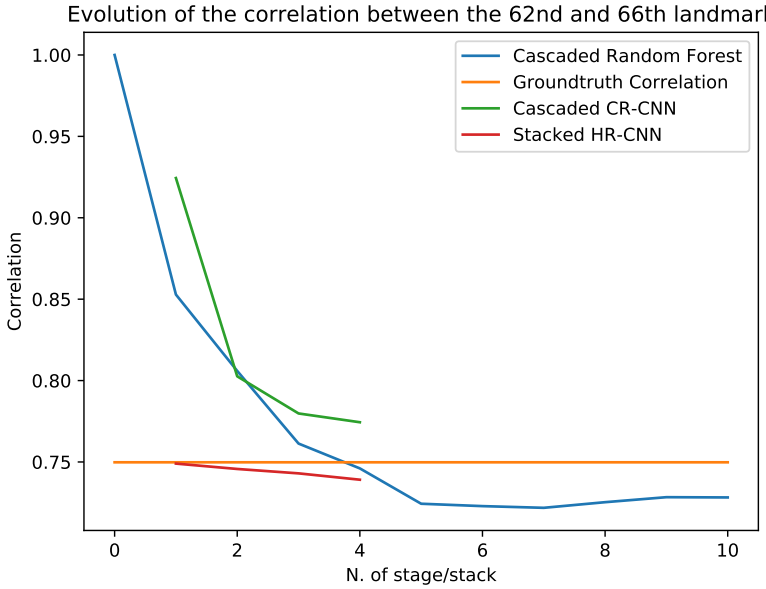
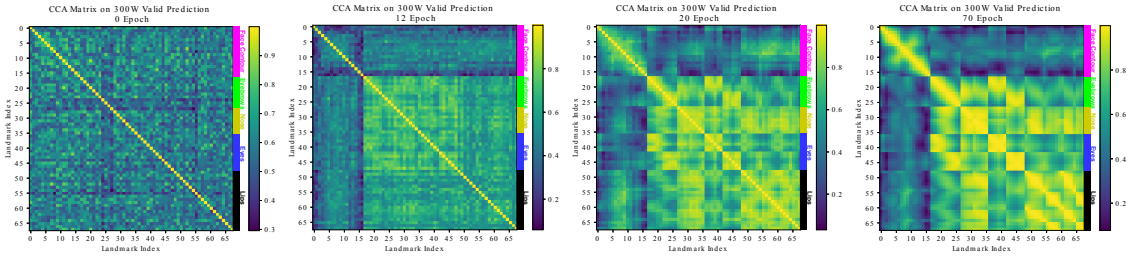


FIGURE 5.11: Evolution of the landmark correlation between the 62nd and the 66th landmark in different stages.



(A) Epoch 0: 170.1% (B) Epoch 12: 15.5% (C) Epoch 20: 10.0% (D) Epoch 70: 6.6%

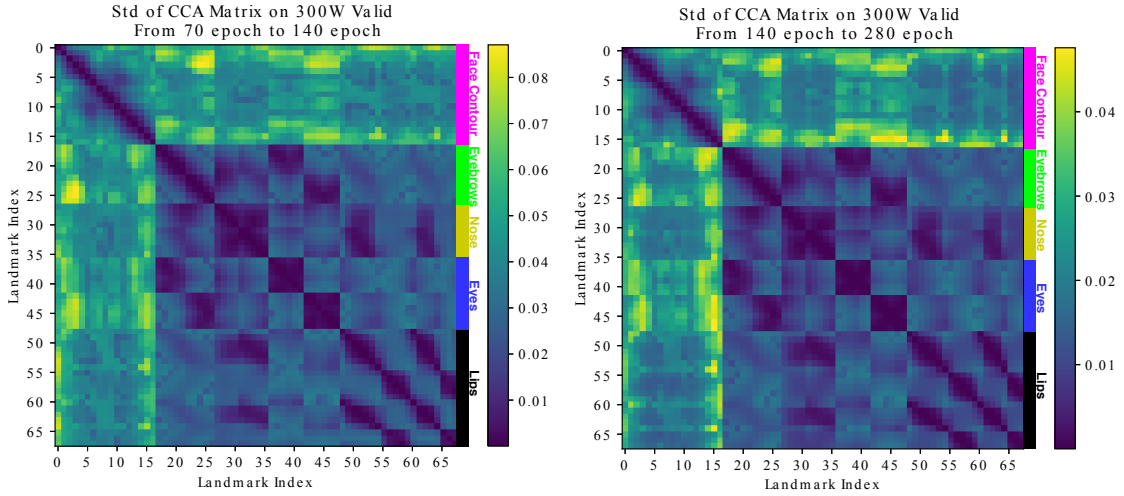
FIGURE 5.12: CCA affinity matrices on the prediction of Coordinate Regression Model in different training epochs. The percentage shown under each figure caption refers to NME.

5.5 Weakly-supervised learning

5.5.1 Motivation

As the size of datasets grows larger and the landmark format becomes denser [Wu et al., 2018b], it is time-consuming to densely annotate each landmark on all of the images. Weakly-supervised learning has attracted increasing attention in the community. Due to the presence of strong landmark correlation in the dense format, we believe that it is not cost-effective to annotate every landmark, especially when the budget for manual annotation is limited.

Our weakly-supervised learning method is also useful in the following situation: we want to extend an existing format (e.g. 300W format) for a specific use (e.g. detect the landmarks on the wing of the nose or the face contour around the forehead) with limited budget. With this correlation analysis, we are able to find out how the landmarks we want to extend are correlated to the landmarks already annotated and find an efficient strategy for manual annotation.



(A) Epoch 71-140: 5.7%

(B) Epoch 141-280: 4.8%

FIGURE 5.13: Standard deviation (Std) of CCA affinity matrices on the prediction of Coordinate Regression Model in different training epochs. The percentage shown in each caption refers to NME at 140th/280th epoch respectively.

5.5.2 Workflow

We propose a weakly-supervised regression method to find the most cost-effective landmarks to annotate (see Fig. 5.14).

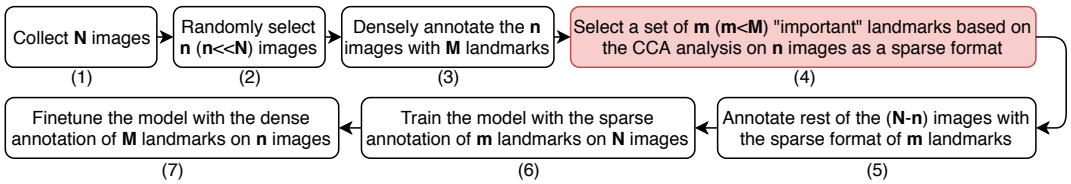


FIGURE 5.14: The workflow of our weakly-supervised learning method. The value of **M** and **N** indicates respectively the total number of the landmarks in the dense annotation format and the total number of the images collected. The value of **m** and **n** can be arbitrarily chosen depending on the requirements. **m** can be considered as annotation budget. We save the time to annotate **M-m** landmarks on **N-n** images.

We assume that a landmark can be easily transferred from another landmark that is highly correlated. As a result, to find the most "important" landmarks facilitating the learning of the others, we search for a set of landmarks that has maximum correlation with the rest of the landmarks. The problem to solve in Fig. 5.14 step (4) can be described as:

Find a set of landmarks indexed by m, which maximize the minimum correlation $\mathbf{c}$ with rest of the landmarks indexed by $\mathcal{M} - m$:

$$m = \arg\max_{m \subset \mathcal{M}} (\mathbf{c}_m), \tag{5.4}$$

$$\mathbf{c}_m = \min_{j \in (\mathcal{M}-m)} (\max_{i \in m} (\mathbf{A}_{i,j})). \tag{5.5}$$

$\mathcal{M}$ denotes the complete set of landmark index in the dense format. $\mathbf{A}_{i,j}$ denotes the i-th row and j-th column of the correlation affinity matrix $\mathbf{A}$ analyzed on $\mathbf{n}$ images. $\mathbf{c}$

can be considered as a criterion of the sparse format selection m. Maximized minimum correlation $\hat{\mathbf{c}} = \max_{m \subset \mathcal{M}}(\mathbf{c}_m)$.

This problem resembles K-center facility problem [Hochbaum and Shmoys, 1985]. A classical K-center problem can be described as: Given a city with **M** locations, find the best k locations to build k facilities, so that the farthest distance from location to its nearest facility has to be as small as possible.

In our problem, the locations in the city can be considered as all the landmarks in the dense annotation format $\mathcal{M}$. The k facilities can be considered as the landmarks selected in our sparse format m. The distance between the landmark i and j can be considered as $1 - \mathbf{A}_{i,j}$. In fact, a high correlation between two landmarks signifies that the distance between two landmarks is small.

K-center problem is NP-hard. Fortunately, this problem can be efficiently solved by mixed-integer programming using Gurobi [Gurobi, 2019], a powerful mathematical optimization solver. We present the canonical form of this problem:

Minimize z, with subject to:

$$\begin{array}{ll} \sum_j x_{ij} = 1 & \forall i \\ \sum_j y_j = \mathbf{m} & \\ x_{ij} \leq y_j & \forall i, j \\ (1 - \mathbf{A}_{i,j}) x_{ij} \leq z & \forall i, j \\ x_{ij} \in \{0, 1\} & \forall i, j \\ y_j \in \{0, 1\} & \forall j. \end{array} \tag{5.6}$$

$x_{ij} = 1$ indicates that landmark i is inferred from the position of landmark j. $y_j = 1$ indicates that the landmark j is selected in the sparse format. $\sum_j x_{ij} = 1$ ensures that all the landmarks are inferred from another landmark. $\sum_j y_j = \mathbf{m}$ ensures that there are **m** landmarks selected in the sparse format. $x_{ij} \leq y_j$ ensures that landmark i can be inferred from landmark j only when landmark j is selected in the sparse format. Finally, the maximized minimum correlation can be obtained by $\hat{\mathbf{c}} = 1 - z$. This optimization can be finished in just several seconds on a normal PC.

5.5.3 Experiments

We demonstrate several sparse formats searched by our method on the dense formats of 300W [Sagonas et al., 2013], 300W inner (exclude the face contour and eyebrows) and WFLW [Wu et al., 2018b] (see Fig. 5.15). Note that the searched formats can be different each time depending on the data (**n** images) sampled from the entire dataset (**N** images). We also list some existing sparse formats: MAFL [Zhang et al., 2016c], LFW [Van Gool, 2012], AFLW [Koestinger et al., 2011] and COFW [Burgos-Artizzu et al., 2013] with same annotation budget **m** as comparison. The advantage of our method is that we are able to distribute the annotation budget (**m** landmarks to annotate) evenly on each component based on the difficulty to learn. Compared to the heuristic choices made by common knowledge, our choice is completely data-driven.

In Tab. 5.1, we present the performance comparison on this task. We find that our sparse format achieves comparable performance compared to MAFL format and LFW format on 300W Inner. When the landmarks on the face contour are included in the learning (on 300W Full & WFLW Full), our format demonstrates more significant improvement compared to AFLW format and COFW format. We also noticed that our searched format is more advantageous with fewer densely annotated images. NME difference between our format and pre-defined format is larger when trained with ratio of 5% and 10%.

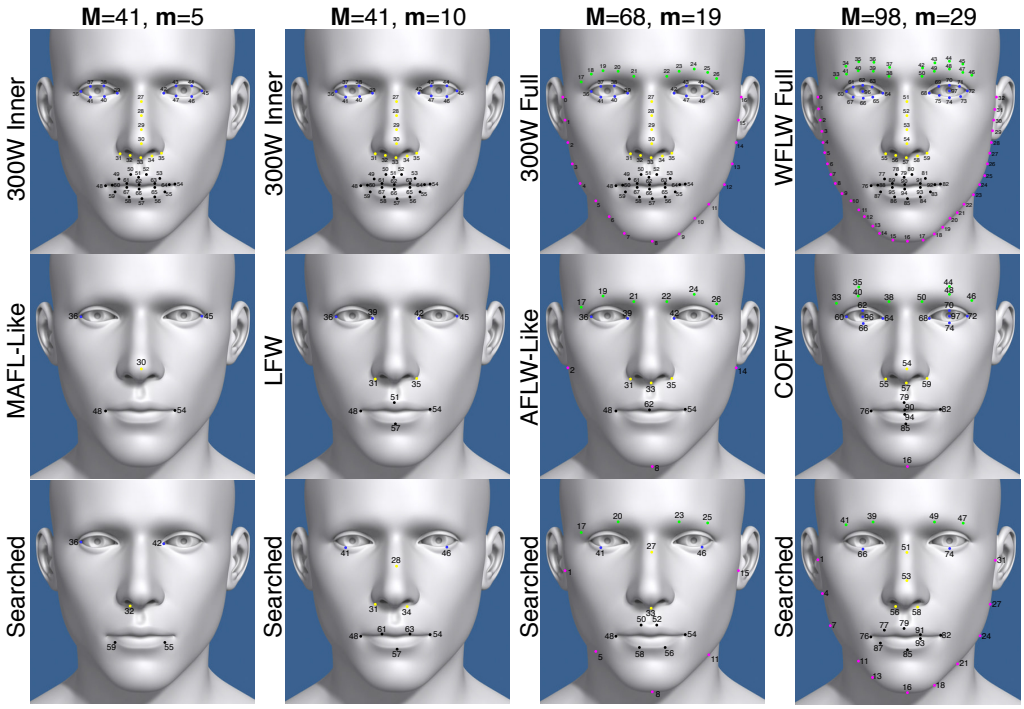


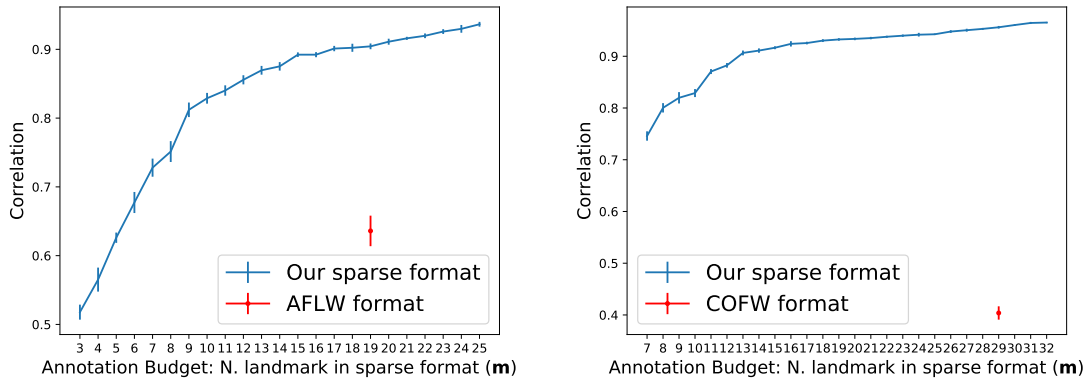
FIGURE 5.15: Examples of the sparse formats obtained by our methods. The first row shows the dense format with **M** landmarks. The second row shows an existing sparse format with **m** landmarks. The third row shows one of the sparse formats searched by our method with **m** landmarks. The total cost **M** and annotation budget **m** are indicated on the top of each column.

	M=41, **m**=5		**M**=41, **m**=10		**M**=68, **m**=19		**M**=98, **m**=29	
Ratio (**n**/**N**)	MAFL	Ours	LFW	Ours	AFLW	Ours	COFW	Ours
5%	4.24	**4.17**	**3.80**	3.83	5.32	**5.17**	7.27	**6.99**
10%	3.92	**3.86**	**3.59**	3.61	5.08	**4.94**	7.03	**6.70**
25%	3.74	**3.66**	3.43	3.43	4.85	**4.84**	6.62	**6.42**

TABLE 5.1: NME(%) performance comparison of the weakly-supervised learning task in Fig. 5.14 by using existing formats and searched sparse format (denoted as ours). The settings of **M** and **m** is consistent with the columns in Fig. 5.15. Ratio represents the percentage of densely annotated images, which is the value of **n**/**N**.

To further investigate the relationship between the annotation budget **m** and the maximized minimum correlation $\hat{\mathbf{c}}$ on different dense formats, we run our sparse format searching method on each dataset with incremental **m**. The relationship between $\hat{\mathbf{c}}$ and **m** is shown in Fig. 5.16. We also demonstrate the values of **c** when using existing sparse formats. Our search method is able to find a bigger **c** compared to the existing ones, which can result in better performance on the weakly-supervised learning task. This figure is useful for us to choose an appropriate annotation budget **m**. For example, on the 300W full format, we find a significant improvement on $\hat{\mathbf{c}}$ by including 9 landmarks in the sparse format. It indicates that it is more advantageous to set the annotation budget **m** to 9 than 8 because the performance can probably be largely improved by adding only 1 annotation budget.

In Fig. 5.17, we plot the relationship between the performance of our weakly-supervised learning tasks (NME %) and the maximized minimum correlation $\hat{\mathbf{c}}$ with different **m** in our sparse format. We find that as the $\hat{\mathbf{c}}$ goes up, the NME is decreased accordingly. It confirms our assumption that the performance of this weakly-supervised learning task is



(A) 300W full: 68 landmarks

(B) WFLW full: 98 landmarks

FIGURE 5.16: Relationship between annotation budget $\mathbf{m}$ and maximized minimum correlation $\hat{\mathbf{c}}$. For each $\mathbf{m}$, we run our searching method 10 times and plot the mean and variance of $\hat{\mathbf{c}}$.

strongly related to the $\hat{\mathbf{c}}$ when using our sparse format.

5.5.4 Implementation details

We use non-pretrained ResNet-18 as our base network. The network disposes two output branches, one branch for the landmarks chosen in the sparse format (main output branch) and the other (secondary output branch) for the rest of the landmarks. We use $L2$ loss and Stochastic Gradient Descent (SGD) optimizer.

Overall, our training consists of three stages:

Stage 1 Train the network with sparsely annotated images: We first train our network with the sparsely annotated images for 1500 epochs. The loss is only applied on the main output branch. The learning rate is initialized to 0.01 and decayed by 0.1 once the loss stops going down for 200 epochs.

Stage 2 Train the secondary output branch with densely annotated images: We freeze the parameters of the entire network except the secondary output branch. In this stage, the loss is only applied on the secondary output branch. We train our network for 3000 epochs in this stage. The learning rate setting is similar to the stage 1.

Stage 3 Finetune the entire network with densely annotated images: Finally, we finetune the entire network, including the backbone as well as both of the branches, for 3000 epochs. The initial learning rate in this stage is set to 0.001. The learning rate decay is similar to the previous stages.

5.6 Conclusions

We propose a correlation analysis as a simple yet effective tool to interpret the relationship among facial landmarks. Our analysis provides a new perspective which is completely different to the commonly used metric NME. Conducting this analysis on the output prediction, we gain some interesting insights on the three most important models in the last decade. We also propose a weakly-supervised learning method to drastically reduce the cost of laborious manual dense annotation. Our methodology on the coordinate correlation can be further extended to 3D facial landmarks, hand/body pose, object landmarks and even the bounding boxes of object detection.

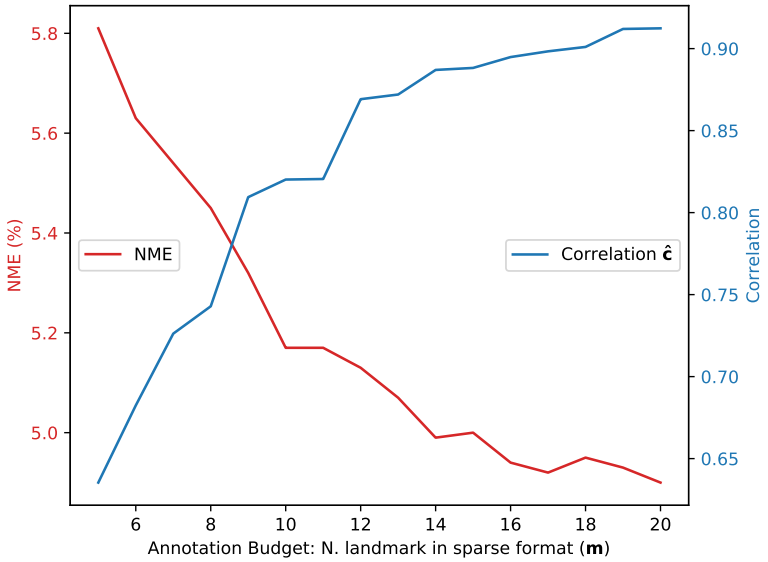


FIGURE 5.17: NME and maximized minimum correlation $\hat{\mathbf{c}}$ with different annotation budget using our sparse format. Tested on 300W full format, $\mathbf{M}$=68, $\mathbf{m}$=5-20.

Part II

Face Parsing

In the second part, we focus on two aspects of face parsing:

- **Hair segmentation in Chapter 6**: Human hair is the most challenging part to locate and recognize on human faces. Unlike the eyes, nose and mouth, human hair cannot be localized by facial landmarks due to the large variance of its shape and texture. The only way to recognize human hair is by face parsing.

  However, the manual annotation for hair segmentation is expensive. With limited training data, the noise on the background frequently perturbs the segmentation output. To this end, we present a method that is more resistant to the cluttered background and gives more consistent hair boundaries by introducing a deep shape prior and a border refinement module.

- **Inference speed in Chapter 7**: To bridge the gap between the research and the application, we present a real-time demonstration of face parsing on mobile platforms such as iPhone and Android. We design an efficient FCNN in an hourglass form that is adapted to live face parsing on mobile phones. The model is deployed on iPhone with CoreML framework. This demonstration proves the feasibility of the face parsing based face AR applications.

Chapter 6

Two-stage Human Hair Segmentation In the Wild

In this chapter, we present a method for robust human hair segmentation. Human hair is the most difficult category in human face parsing. Our proposed method demonstrates superior robustness against cluttered background compared to the state-of-the-art methods and delivers visually more consistent hair boundaries.

6.1 Introduction

Human hair contains rich color, shape and textural information. At the same time, it is generally related to gender, culture and appearance. [Muhammad et al., 2018] argued that hair detection and segmentation is important in two domains. Firstly, hair segmentation is an essential prior step for 3D hair modeling from a single portrait image as well as for some AR applications such as hair dying and facial animation. Secondly, it can be used in biometric recognition applications such as human presence detection from the back view or gender and face recognition.

Hair segmentation *in the wild* consists in performing hair segmentation in an unconstrained view without any explicit prior face or head-shoulder detection [Muhammad et al., 2018]. We address this problem as a semantic segmentation problem by taking texture and shape constraints into account. Hair segmentation, especially under such unconstrained conditions, is challenging for the four following reasons:

- **Cluttered background**: textures in the background can be similar to human hair, which introduce significant difficulties for hair segmentation *in the wild*.
- **Lack of rigid and consistent form**: the form of hair can be totally different according to the head pose, different points of view and ambient environment such as wind. However, we believe that human hair, although in different situations, share implicit shape constraints.
- **Hair style/color variation**: There are numerous appearance variations in terms of hair style such as straight hair, curly hair, braided hair, short hair, etc. Hair colors are divergent from person to person and can be easily biased by different environment and cameras.
- **Complex lighting conditions**: Under complex lighting conditions, hair texture information is usually distorted. It is even difficult for a human to figure out the exact boundary of human hair in extreme lighting situations for example in shadow or backlight.

In this chapter, we aim at improving hair segmentation *in the wild* by correctly distinguishing hair texture from similar texture in the background as well as estimating refined hair borders. Previous CNN-based methods [Levinshtein et al., 2017, Liu et al., 2017d] generally adopt a single stage, which we think is insufficient under such extreme conditions. We propose a two-stage pipeline (see Fig. 6.1.) consisting of a shape prior detection stage and a hair segmentation stage. Our contributions in this chapter can be summarized as follows:

1. Before segmentation, we propose to first detect a hair shape prior which is based on a specific distance transform map. The results show that it helps to improve the robustness against cluttered background.

2. In the segmentation stage, we propose a border refinement module along with a symmetric encoder-decoder FCNN to obtain a more precise segmentation output.

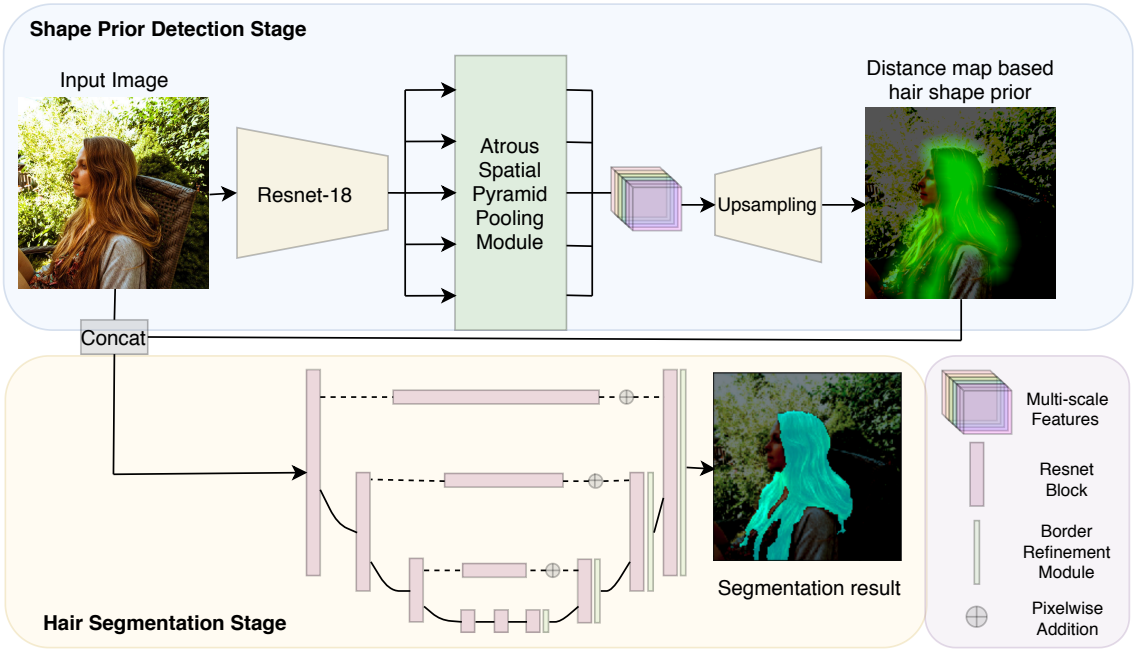


FIGURE 6.1: Our two-stage human hair segmentation pipeline.

6.2 Related Work to Various Subjects of Segmentation

6.2.1 Semantic Segmentation

Hair segmentation can be considered as a type of semantic segmentation, a problem for which FCNN have achieved remarkable results in the past few years. According to [Chen et al., 2018b], there are mainly two types of FCNN models: the encoder-decoder structures which privilege refined boundaries [Badrinarayanan et al., 2017, Ronneberger et al., 2015], and structures integrating a spatial pyramid pooling module [Chen et al., 2018a, He et al., 2015b, Zhao et al., 2017], which gather rich multi-scale contextual information. The former ones normally adopt a symmetric structure with skip connections which enable low-level information to flow from the encoder directly to the decoder. This is now widely used in various applications such as image matting [Xu et al., 2017], landmark detection [Newell et al., 2016, Bulat and Tzimiropoulos, 2017b] etc. The latter

ones employ "atrous" convolutions [Holschneider et al., 1990] at different rates to capture features in arbitrary resolutions and show excellent performance on large-scale semantic segmentation datasets [Everingham et al., 2015, Cordts et al., 2016, Zhou et al., 2017].

6.2.2 Texture Recognition and Segmentation

The most characteristic feature of human hair is the texture. Texture recognition is usually considered as a basic image processing problem without taking semantic information into account. As a result, many approaches are based on the Bag of Words model [Leung and Malik, 2001] to obtain spatially invariant features for texture representation [Liu et al., 2018a]. Recently, CNN-based methods with orderless feature pooling [Gong et al., 2014, Cimpoi et al., 2015, Zhang et al., 2017] have shown good performance on texture recognition, which was later proved to be beneficial for semantic segmentation [Zhang et al., 2018]. In terms of texture segmentation, many approaches are based on active contours and integrate different texture features [Wu et al., 2015, Reska et al., 2015, Varnosfaderani and Moallem, 2017, Liu et al., 2017b, Yuan et al., 2015, Gao et al., 2016]. [Cimpoi et al., 2015] proposed to use object detection-like region proposal classification to assign the texture/object labels to each pixel.

6.2.3 Coarse-to-fine Segmentation

Recently, there are several works concerning the refinement for the CNN-based semantic segmentation. Chen et al. [Chen et al., 2015] used Conditional Random Field (CRF) to establish an additional pair-wise supervision between the pixels. Wu et al. [Wu et al., 2018a] proposed a guided image filter, which is designed to generate a high-resolution output and from a low-resolution input given a guidance input. Liu et al. [Liu et al., 2017c] proposed to construct a linear propagation model, which constitutes a spatial affinity matrix that models dense, global pairwise relationships of an image. Similarly, Jiang et al. [Jiang et al., 2018] proposed DifNet, which models the pairwise information by cascaded random walks. Our method uses spatial attention as additional supervision for boundary refinement. Compared to the previous methods, our method bears two advantages: (1) Unlike CRF and DifNet, our approach does not require iterative operations. (2) Our border refinement module can be easily integrated with the feature maps of different scales in most of the layers.

Li et al. [Li et al., 2017] found that most of the difficult pixels are located on the boundaries. Therefore they proposed a cascaded scheme to process all the pixels step-by-step from easy (center pixels) to hard (boundary pixels). Zhu et al. [Zhu et al., 2019b] proposed a boundary label relaxation strategy to alleviate the influence of the hard pixels on the boundary on the overall score. In our approach, the first stage is trained to learn a general shape without detailed boundaries, which prevents to mistaking the noises on the cluttered background.

6.2.4 Human Hair Segmentation

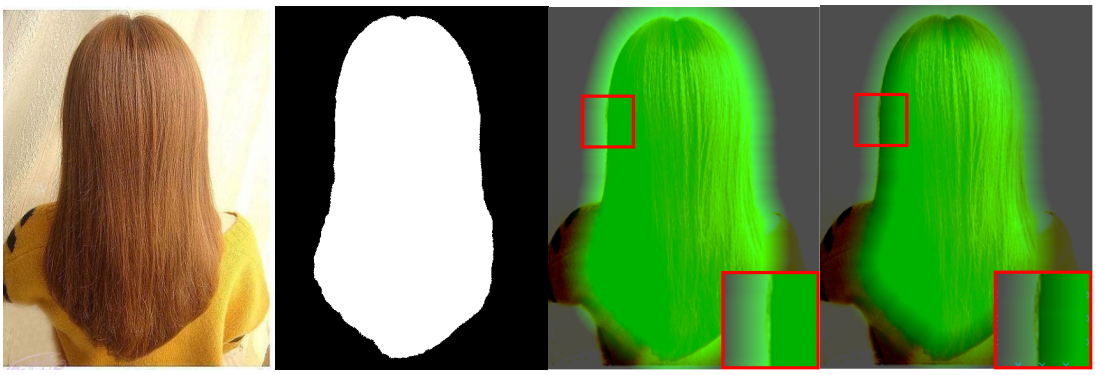
Early methods proposed to segment human hair by modeling color, location and frequency information [Yacoob and Davis, 2006, Lee et al., 2008, Rousset and Coulon, 2008]. [Wang et al., 2010, Wang et al., 2012] decompose the hair segmentation into local parts. Several other approaches [Wang et al., 2009, Wang et al., 2011, Wang et al., 2013] use region growing followed by refining regression on the coarse mask. Recent work [Chai et al.,

2016,Qin et al., 2017,Guo and Aarabi, 2018,Levinshtein et al., 2017] based on FCNN models achieved good performance for practical applications. However, most of the methods only focus on the cases under constrained conditions, such as head-shoulder images.

Muhammad et al. [Muhammad et al., 2018] proposed a challenging hair analysis dataset along with a method to realize hair detection, segmentation and style classification. Their method renders quite good detection. However, they perform a sliding window texture recognition operation on the whole image, which is computationally very expensive.

6.3 Proposed Approach

We decouple the hair segmentation task into two important steps: (a) find the general hair shape prior and (b) find the refined border of the hair. In the hair shape detection stage, inspired by the hair occurrence probability mask used in previous methods [Wang et al., 2011, Wang et al., 2012, Muhammad et al., 2018] and soft segmentation, we aim at finding a coarse hair mask that indicates the hair texture presence (a coarse hair shape prior) regardless of the exact border. In the hair segmentation stage, we aim at identifying the exact hair border by integrating a border refinement module in a symmetric encoder-decoder FCNN.



(A) (B) (C) (D)

FIGURE 6.2: An illustration of our distance map transformation. From left to right: (A) Original image (B) Ground truth hair mask (C) Clipped distance transform map overlaid on original image (D) Clipped distance transform map with "erosion" overlaid on the original image. With "erosion", an uncertain region is created on both sides of the hair boundary.

6.3.1 Hair Shape Prior Detection

Distance map regression. As stated before, two significant challenges for hair segmentation *in the wild* are: distinguishing hair appearance from similar background texture and learning challenging hair shape geometries. In most of the previously proposed FCNN models for semantic segmentation, object shape constraints are not explicitly imposed. We propose to introduce a coarse mask without precise boundary as a shape prior for hair segmentation. We transform the binary ground truth hair mask to a boundary-less coarse hair mask by using a specific type of distance transform.

An illustration of our distance map transform is shown in Fig. 6.2. Consider a binary ground truth hair mask $I(x,y)$ in Fig. 6.2 (B). Hair pixels and non-hair pixels can be denoted respectively as $I^+ = \{I(x,y) = 1\}$ and $I^- = \{I(x,y) = 0\}$. We define a clipped

distance transform map dt_{mask} on the image positions $p(x, y)$ as:

$$dt_{mask}(p) = d_{max} - \min(d_{max}, \min_{p^+ \in I^+} \|p^+ - p\|) \tag{6.1}$$

where d_{max} denotes the maximum clipping threshold for distance values (see Fig. 6.2 (c)). And, similarly, we define a clipped inverse distance transform map with respect to the background pixels:

$$dt_{inv}(p) = e_{max} - \min(e_{max}, \min_{p^- \in I^-} \|p^- - p\|) \tag{6.2}$$

where $e_{max} (< d_{max})$ denotes the second clipping threshold. Then, the final distance transform map dt is obtained by:

$$dt = dt_{mask} - dt_{inv} \tag{6.3}$$

which is then normalized between -1 and +1 to be formed as a regression target (see Fig. 6.2 (d)). The use of dt_{inv} "erodes" the initial distance transform dt_{mask} and produces an uncertain hair boundary region for the target image. e_{max} can be considered as the magnitude of "erosion". We do not use traditional morphological erosion to do this because some small hair regions on the binary mask might be ignored while small holes might be filled. We use "HardTanh" as final activation function, defined as:

$$\text{HardTanh}(x) = \begin{cases} 1 & \text{if } x > 1 \\ -1 & \text{if } x < -1 \\ x & \text{otherwise.} \end{cases} \tag{6.4}$$

This activation function is a linear approximation of Tanh function and clipped from -1 to +1, which naturally fits the range of our shape prior regression target. We use L1 loss to train our distance map regression. In our implementation, we empirically set d_{max} to 25 and e_{max} to 10.

Atrous Spatial Pyramid Pooling (ASPP) encoder. Although texture is considered as very local information, in the setting of hair segmentation *in the wild*, the scale of the hair region varies considerably. ASPP with different atrous rates effectively captures multi-scale information to learn the presence of hair texture. We use DeeplabV3 [Chen et al., 2018a] structure with Resnet18 [He et al., 2016] pre-trained on ImageNet as backbone encoder in our hair detection network. Finally we upsample the multi-scale feature map to obtain the final distance transform map at the original image size.

6.3.2 Refined Hair Segmentation

Symmetric encoder-decoder. In hair segmentation stage, we implement a symmetric encoder-decoder structure with skip connections. At each level, we use a ResNet block in both the encoder and the decoder part. Additionally, as in [Newell et al., 2016], we add a ResNet block in the skip connections to process the low-level information transferred from the decoder.

Border refinement module. The boundary of the human hair is difficult to detect due to the presence of tiny details. These tiny details only concern limited number of pixels. However, the final rendering might be visually unsatisfying if they are not well treated. To refine the hair boundary, we propose to use spatial attention, which will help the CNN to focus on the pixels around the hair boundary.

In [Zhang et al., 2018], the authors implemented a squeeze-and-excitation channel-wise attention [Hu et al., 2018] module for semantic segmentation. Here in refinement segmentation stage, we are more interested in pixel-wise attention to recover refined hair

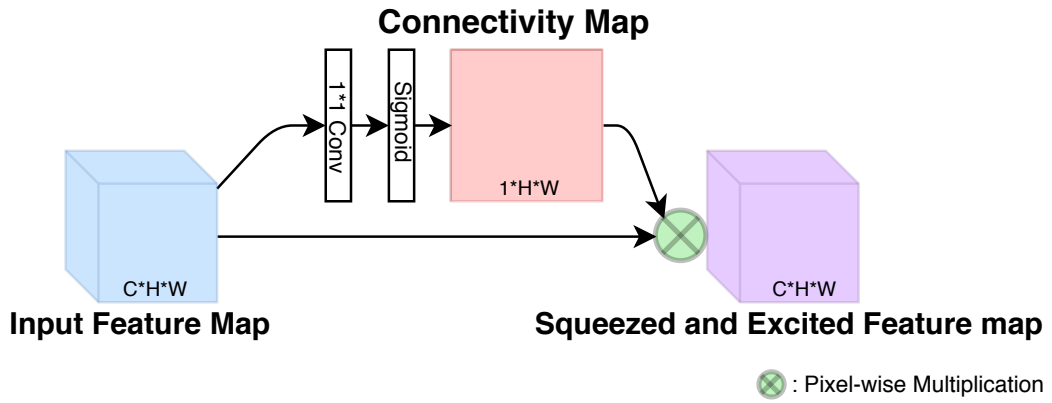


FIGURE 6.3: An illustration of our proposed border refinement module. The input feature map is transformed into a **single channel** Connectivity Map by an 1×1 convolutional layer with sigmoid activation. The Connectivity Map is then multiplied with each channel in the input feature map before output.

border. Inspired by this work, we propose a refinement module which generates spatial attention. The input feature map is passed through a 1×1 kernel convolutional layer with a sigmoid activation function to a single-channel feature map. We call it connectivity map. It is multiplied by each channel of the input feature maps afterwards to obtain the "squeezed and excited" output feature map. The module is illustrated in Fig. 6.3. We place this module at each level of the decoder part before upsampling. We noticed that this module helps to improve the performance, smooth the boundary and get better visual result.

6.4 Experiments

6.4.1 Datasets

We conducted our experiments on LFW-Part dataset [Kae et al., 2013] and the newly-released Figaro-1k [Muhammad et al., 2018] dataset. The LFW-Part dataset is a face parsing dataset with hair annotation which consists of 2927 images. To the best of our knowledge, Figaro-1k is the only hair analysis dataset *in the wild* with precise hair annotation. It consists of 1050 images (210 for validation) and manually annotated ground truth hair masks, which varies in seven hair styles, different hair colors, length and levels of background complexity.

6.4.2 Experimental Settings

For quantitative evaluation, we adopted several standard measures e.g. mean Intersection over Union (mIoU), accuracy and F1-score. The images are resized to 256×256 for training. The evaluation is performed at their original size. For data augmentation, due to limited number of images in the Figaro1k dataset, we apply various data augmentation on the input images during training: (1) a random resize of ± 20 % on image width and height (2) a random translation of ± 60 % in horizon and ± 30 % in vertical (3) a random crop/pad of ± 20 % on image width and height (4) a random horizontal flip (5) a color jitter (on brightness, contrast and saturation) of ± 30 % (6) a random gaussian lighting noise based on ImageNet PCA analysis.

6.4.3 Hyperparameter settings

Distance transform map regression in the hair detection stage is trained by using L1 loss while final refinement segmentation in the second stage is trained by using standard softmax loss. We use RMSprop to train the networks in both stages at the same time for 190 epochs with a initial learning rate of 0.0005 and batch size of 6. The learning rate is decayed by 0.3 for the first 30 epochs and then decayed in the same manner each 40 epochs. We use PyTorch to implement the training on a single NVIDIA GTX 1080Ti. The training stage finishes in around 13 hours on Figaro1k dataset. Each inference takes around 15ms compared to 1.79s in [Muhammad et al., 2018] and 3.3ms in [Liu et al., 2017d].

6.4.4 Quantitative Comparison

On Figaro-1k dataset, we compared our method with the encoder-decoder fully convolutional neural network U-Net [Ronneberger et al., 2015], the state-of-the-art semantic segmentation approach DeeplabV3+ [Chen et al., 2018b] and the previous work on hair analysis *in the wild* [Muhammad et al., 2018]. The result is reported in Table 6.1. Our approach outperforms all the previous methods for hair segmentation *in the wild*. By adding a detection stage, a gain of more than 1% point on IoU and F1-score can be achieved. The larger improvement on precision shows that our method is effective for removing false positives on the background. With the border refinement module, the performance is additionally improved by only a small margin but gives better visual results. On the LFW-Part dataset, by adding a hair detection stage, our method outperforms other methods by more than 1% point on the hair F1-score (see Table. 6.2).

Method	Precision(%)	F1(%)	mIoU(%)	Accuracy(%)
U-Net [Ronneberger et al., 2015]	95.63	94.39	89.69	96.36
DeeplabV3+ [Chen et al., 2018b]	96.86	95.05	91.11	97.07
[Muhammad et al., 2018]	-	84.90	-	91.50
Only Seg Stage	95.64	94.53	89.91	96.56
(Det + Seg) Stage	97.25	95.09	91.15	97.20
(Det + Seg + Refine) Stage	**97.33**	**95.15**	**91.25**	**97.23**

TABLE 6.1: Comparison of Hair Segmentation Results on Figaro1k.

Method	Precision(%)	F1-hair(%)	mIoU(%)	Accuracy(%)
U-Net	89.11	87.66	88.33	96.58
DeeplabV3+	91.66	88.36	**90.64**	96.82
[Liu et al., 2017d]	-	83.43	-	95.46
Only Seg Stage	89.13	88.07	90.12	96.71
(Det + Seg) Stage	98.24	88.94	90.53	96.76
(Det + Seg + Refine) Stage	**98.34**	**89.42**	90.56	**97.05**

TABLE 6.2: Comparison of Hair Segmentation Results on LFW-Part.

6.4.5 Qualitative Comparison

Ablation study: Necessity of using hair shape prior. Fig. 6.4 shows several challenging images where the hair segmentation fails without shape prior. In these images, there are either similar textures or complicated lighting conditions present. We notice that (a) false positive segmentation on similar texture in the background is rectified and (b) tiny isolated false positive hair parts are suppressed. We think that the improvement originates from (1) the ImageNet pre-trained features, (2) our trained hair shape prior. To ablate the influence from (1), we compare our results with the ImageNet pre-trained DeepLabV3+. We find that our shape prior-integrated approach is more robust to cluttered background and renders more reasonable hair shapes in complex situations.

To investigate how the FCNN in the segmentation stage processes the shape constraint prior from the detection stage, we give an illustration in Fig. 6.5. With the help of the eroded distance map, the two small hair regions in the red rectangles are assigned weaker values. We note that both of them are eliminated by the FCNN in the segmentation stage, which removes one false positive but creates one false negative as well. Intuitively, this amounts to a learnable "thresholding".

FIGURE 6.4: Comparison on challenging examples in Figaro1k. First row: Input image. Second row: Segmentation results by our model without detection stage. Third row: Segmentation results by ImageNet pre-trained DeepLabV3+. Fourth row: Segmentation results by our two-stage model. Many tiny isolated false positives can still be observed on the man's shirt in the first image of DeepLabV3+ results.

Ablation study: Necessity of using border refinement module. In order to disconnect the influence of different shape priors from the detection stage, we pre-trained the detection network and fixed the weights during the training of the segmentation network to explore this necessity. Even though we remark only a slight quantitative improvement on both datasets, we find that the results are visually better because of the smooth boundaries as shown in Fig. 6.6. It provides more consistent hair regions and eliminates spurious small detections in the background thanks to the use of the pixel connectivity

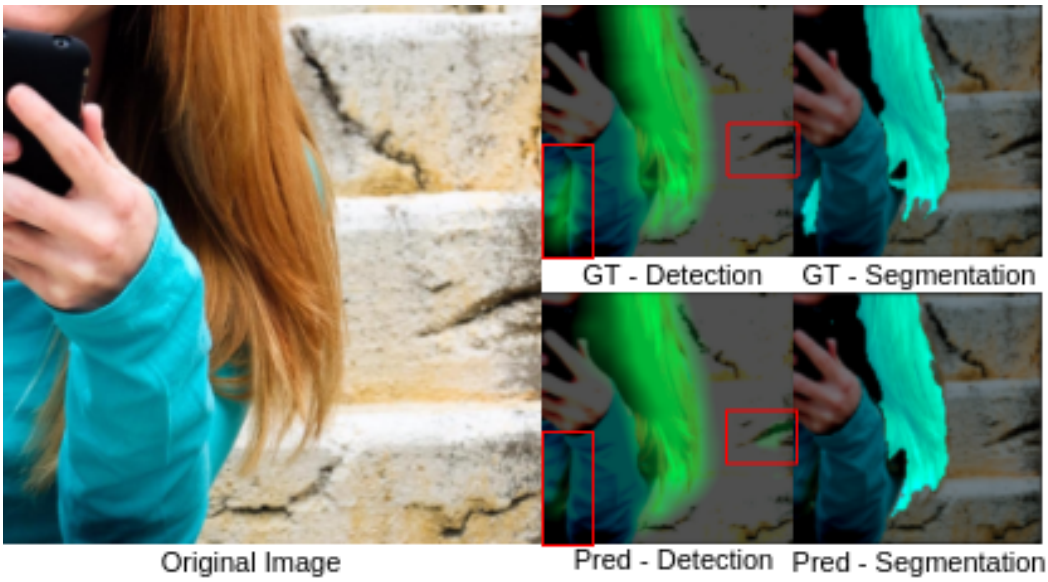


FIGURE 6.5: An illustration of the relation between shape prior and final segmentation. GT refers to ground truth and Pred denotes our two-stage network prediction. Weaker values are observed inside the red rectangles on the detection prediction.

map. Smoothed boundaries do not necessarily translate into a better mIoU, but are visually more appealing, even compared to the ground truth. In fact, even for humans, it is challenging to annotate the boundary in a very precise way by using only a binary mask. A promising future work to further improve the hair boundary could be image matting [Xu et al., 2017, Levinshtein et al., 2017].

Visual Comparison with Stae-of-the-art Methods. We visualize the output of our method on Figaro dataset in Fig. 6.7. We compare our method with the state-of-the-art method [Muhammad et al., 2018]. We observe that our detection, especially the hair boundary, is much finer compared to [Muhammad et al., 2018], which is more appealing for the practical applications such as virtual hair coloring.

We show some challenging examples on LFW-Part dataset in Fig. 6.8. We compare the visual results from (1) our model without detection stage, (2) DeepLabV3+ [5] with ImageNet pretrained Resnet18 as backbone and (3) our model with both detection and segmentation stage. We observed that our method outperforms the others by suppressing the spurious detection on the cluttered background. It shows the importance of performing shape prior detection before segmentation stage.

Failure cases. In Fig 6.9, we provide several examples where our model fails. Our approach still cannot completely ensure the identification of the correct textures on the cluttered background especially when they are very close to the hair or has a similar form for example in (a) and (b). Furthermore, complex lighting conditions in (c) and irregular upside down pose in (d) introduce big challenges for our methods. Nonetheless, our method still suppressed more false positive detection than the segmentation-only-models such as ImageNet pre-trained DeepLabV3+. From (e), (f) and (g), we observe that our method is less sensitive to weakly-textured and small hair regions compared to other methods.

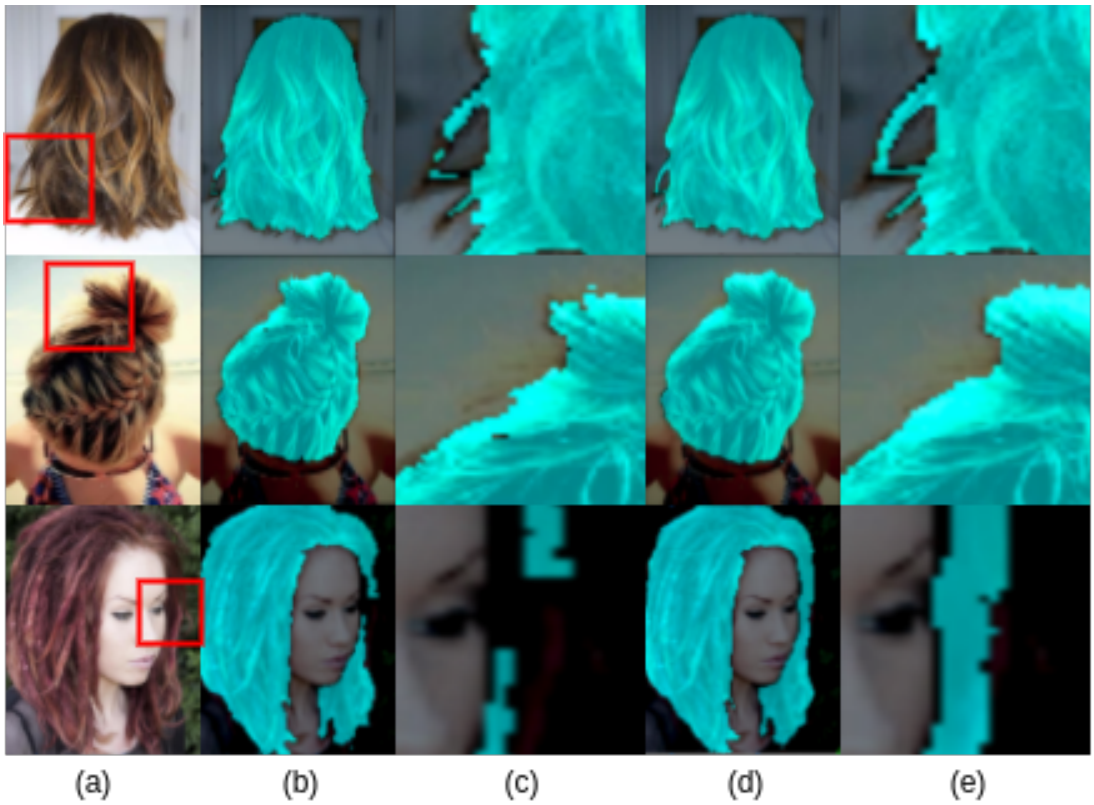


FIGURE 6.6: Impact of the border refinement module: (a) Original image (b)(c) Global and zoomed-in segmentation result w/o refinement module (d)(e) Global and zoomed-in segmentation results with border refinement module.

6.5 Conclusions

In this chapter, we presented a two-stage pipeline for hair segmentation *in the wild*. We train a distance map-based hair shape prior from data, and then estimate the final segmentation by a symmetric FCNN using a border refinement module. Our approach outperforms previous state-of-the-art methods, being more robust to cluttered background and giving visually more consistent hair borders. Our approach can be further extended to textured object segmentation with difficult boundaries such as clothes parsing [Yamaguchi et al., 2012] and road scene parsing [Fritsch et al., 2013]. Nonetheless, these segmentation models for face parsing are very time-consuming, especially for mobile phones. In the next chapter, we present a face parsing demo that is able to run in real-time on mobile platforms.

The contributions in this chapter led to two publication at *CVPR workshop* and *Pattern Recognition Letters*.

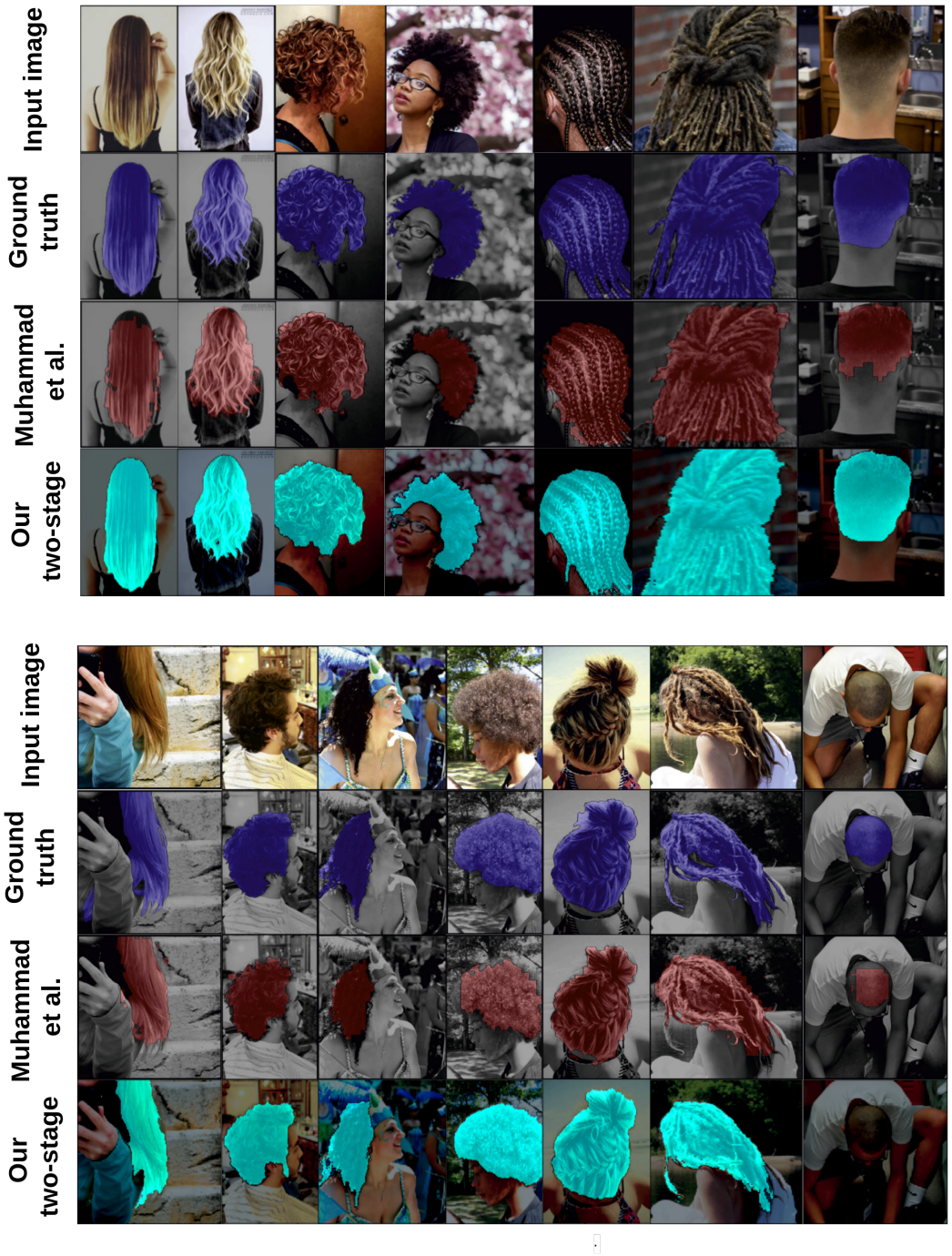


FIGURE 6.7: Visual results compared to [Muhammad et al., 2018]. First row: Input image. Second row: Groundtruth. Third row: Results from [Muhammad et al., 2018]. Fourth row: Results from our two-stage human hair segmentation model. Best viewed in color.

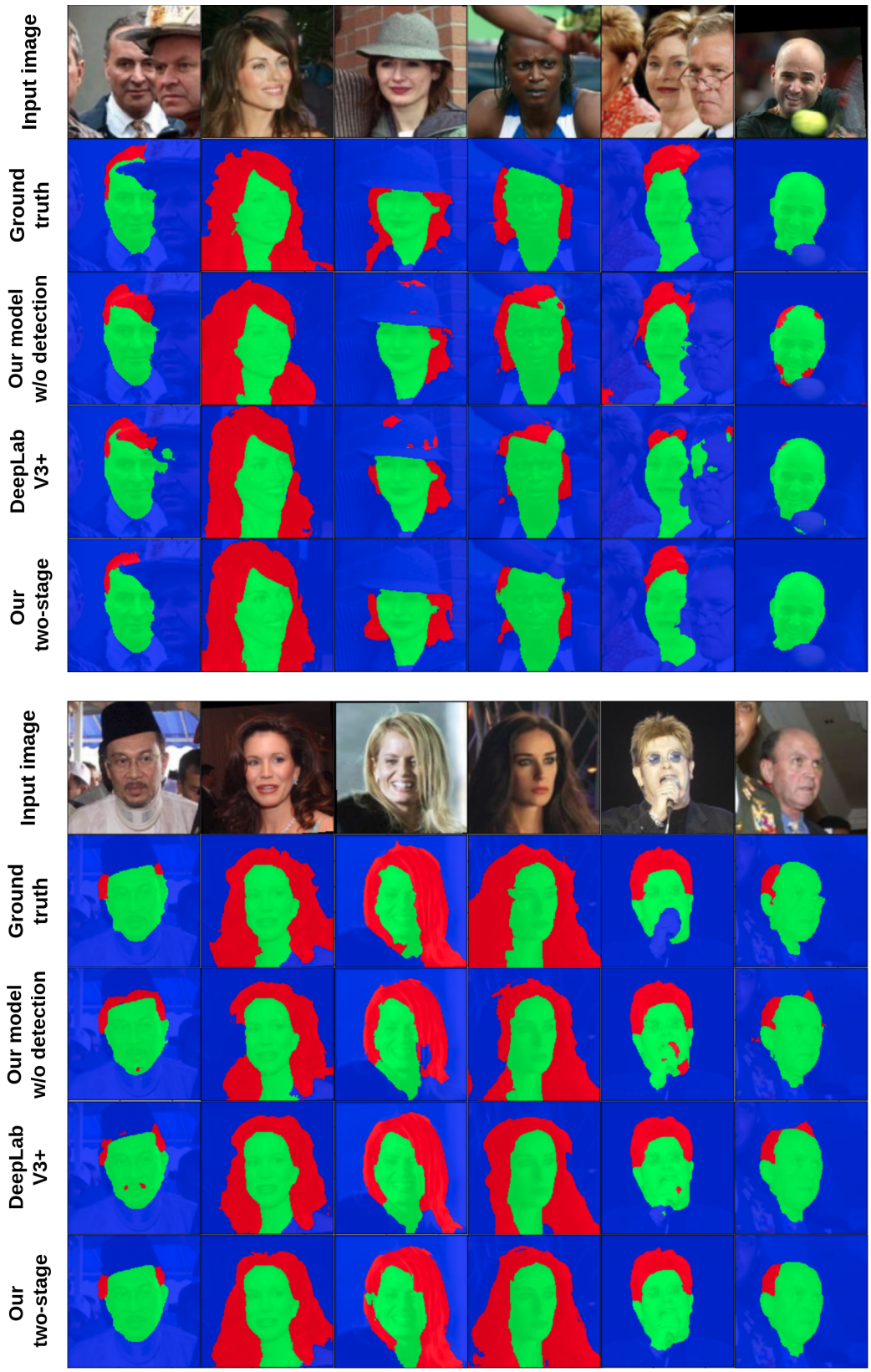


FIGURE 6.8: Challenging examples on LFW-Part dataset. First row: Input image. Second row: Groundtruth. Third row: Results from our model without detection stage. Fourth row: Results from ImageNet pre-trained DeepLabV3+. Fifth row: Results from our model with both detection and segmentation stage. Red-Hair, Green-Face, Blue-Background. Best viewed in color.

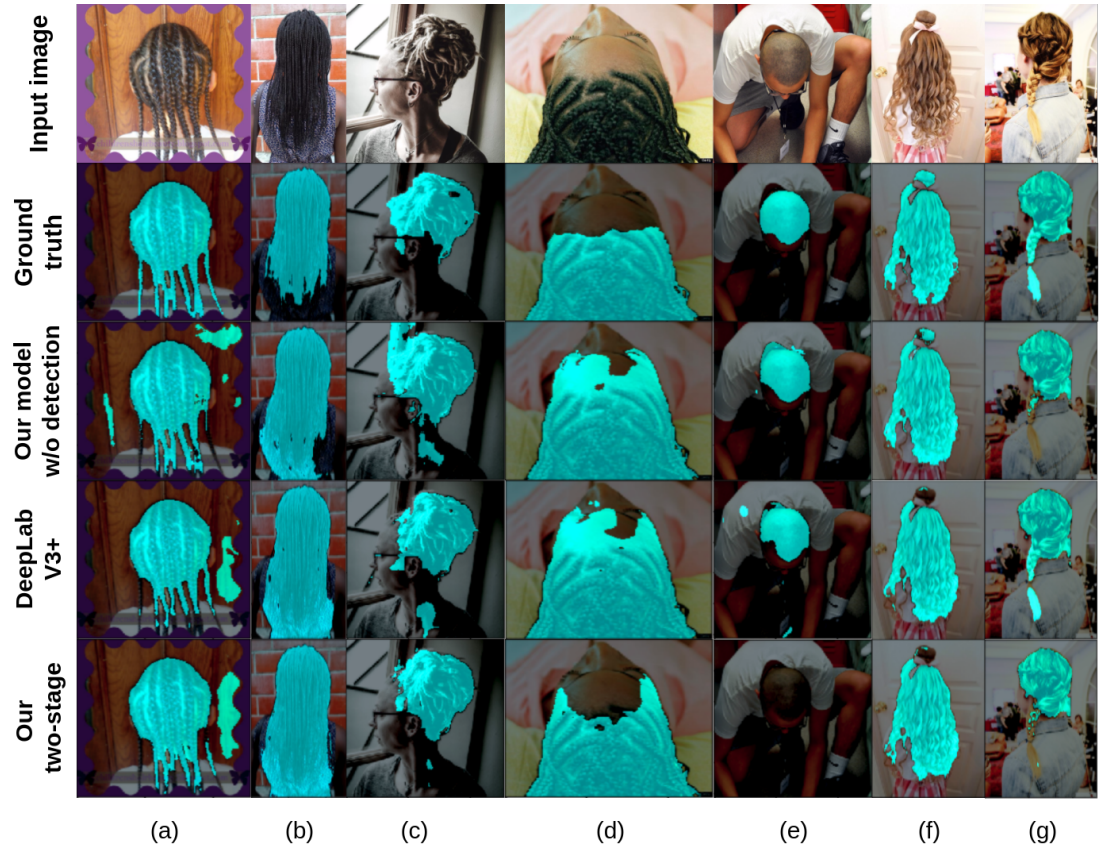


FIGURE 6.9: Failure examples when using our methods. First row: Input image. Second row: Groundtruth. Third row: Results from our model without detection stage. Fourth row: Results from ImageNet pre-trained DeepLabV3+. Fifth row: Results from our model with both detection and segmentation stage. Best viewed in color.

Chapter 7

Face Parsing for Mobile AR Applications

In this chapter, we present a demonstration to prove the feasibility of the deep face parsing based AR application. Our demonstration runs on iPhone X in real-time, and delivers more consistent results across frames.

7.1 Introduction

Detecting different facial components is of great interest for a lot of AR applications such as facial image beautification and facial image editing. For example, given the lip area, we can apply virtual lipstick-wearing effect on it by colorizing the region with proper colors. Here, we focus on designing an efficient face parsing methods by neural networks since lots of these applications are aimed at the mobile platforms such as iOS and Android.

Deep CNNs have been proved to be the leading methods for lots of computer vision tasks especially semantic segmentation [Badrinarayanan et al., 2017, Ronneberger et al., 2015]. Nonetheless, these methods are either time-consuming or over-sized due to the excessive amount of parameters and computation. Most of the semantic segmentation methods are designed for general complex scenes or street scenes. However, face parsing is a quite different task for the following reasons.

- Face parsing is usually done based on an RoI given by preliminary face detection compared to the general semantic segmentation which is generally performed on the entire image.
- Sharp boundaries are demanded for facial AR applications in order to render better visual effects.
- Facial components have more deformable variance but less position and size variance compared to semantic segmentation.

In this chapter, we present a real-time AR face parsing demonstration on iPhone with an efficient deep convolutional neural network. Unlike the preceding face parsing methods, we consider the adaption of neural networks on video in order to provide a fluid and temporally consistent rendering. The users are able to visualize the segmentation results by a mask which indicates different facial regions. A rendering result is shown in figure 7.1. Various AR applications can be realized based on our results.

7.2 Related Work to Semantic Segmentation

A commonly shared consensus in deep semantic segmentation area is that there exists two kinds of mainstream methods [Chen et al., 2018b], the spatial pyramid pooling

FIGURE 7.1: The visual results of our method on iPhone. Our method is robust to extreme expressions and poses.

(SPP) module based structures and encoder-decoder structures [Badrinarayanan et al., 2017, Ronneberger et al., 2015]. Encoder-decoder structures adopt progressive upsampling with skip connection to reconstruct the object boundary as sharp as possible. Skip connections play an important role in the network structure so that the CNN can transfer the low-level detailed information to the output layers.

7.3 Related Work to Network Acceleration

To ensure the best user experience in AR applications, short inference time and low latency are required. Some researchers proposed to use optical flow and reuse the feature map of the past frames if the static scene persists. Another way to accelerate the inference time is to use network acceleration techniques like quantification or pruning. Recent works such as Mobile-net [Sandler et al., 2018] and Shuffle-net proposed to use depth-wise separable convolution to reduce computational complexity of CNN with almost the same performance.

7.4 Mobile Face Parsing Demo Description

In this part, we present the design of our segmentation network, how we adapt it to the video as well as some implementation details.

7.4.1 Mobile Hourglass Network

We follow the design of the hourglass model in human pose estimation [Newell et al., 2016]. Our network structure is presented in Figure 7.2. The Hourglass model is a symmetric encoder-decoder fully convolutional network with a depth of 4, which means the encoder downsamples the input for 4 times and the decoder upsamples the feature map for 4 times to reconstruct the output. Each yellow block represents a network block which enables the CNN to learn the information flow at each stage and skip connection.

We drop several convolutional and max-pooling layers at the beginning of the network and augment the size of the output map to 256, which is the same size as the input image. This will make the boundaries of facial objects sharper but increase the computational complexity as well. In order to accelerate the inference, we replace all of the ResNet blocks in the hourglass network by MobileNet [Sandler et al., 2018] inverted bottleneck block. Expansion factor τ is an important hyper parameter which indicates how many times the number of feature channels are expanded in its blocks. We chose the number of features f as 32 and the expansion factor τ as 6 by finding the best compromise between

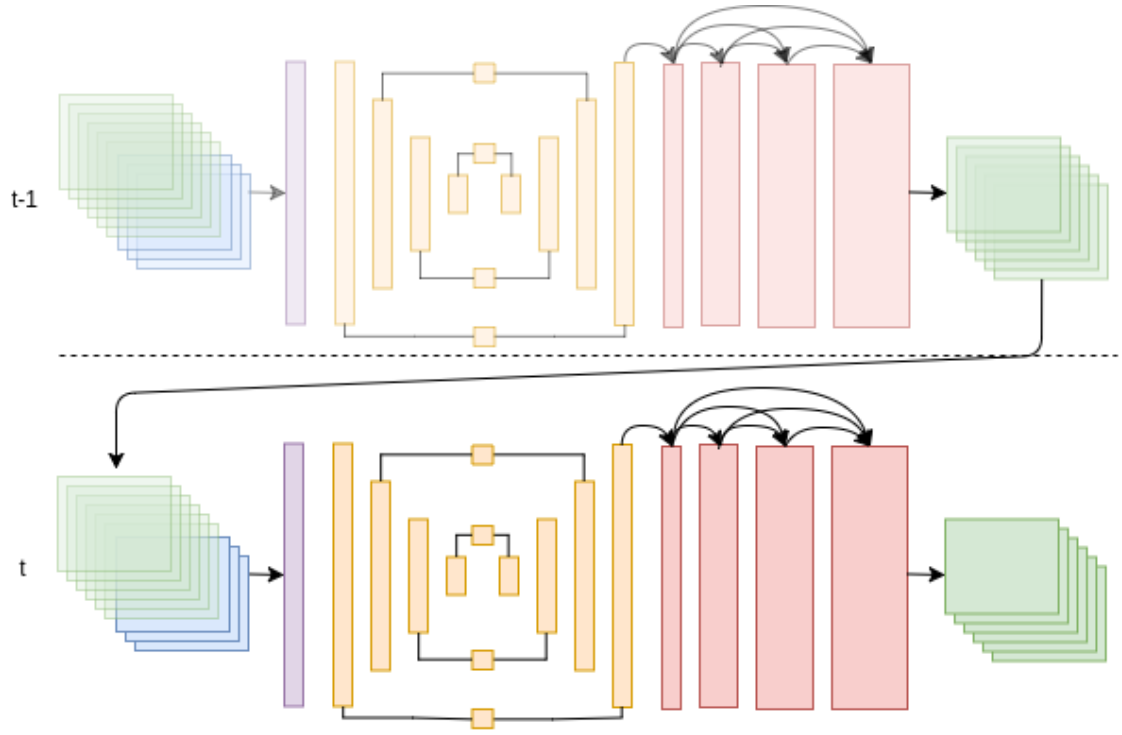


FIGURE 7.2: Overview of our method for video face parsing. Blue channels: RGB facial image, Green channels: Mask predictions, More transparency signifies earlier predictions in time dimension. The output predictions of $t-1$ is reinjected to t for robustness. Purple block: convolutional layer + batch norm layer. Yellow blocks: The Hourglass model composed of Mobile-net Blocks. Red blocks: Dense [Huang et al., 2017] Blocks.

the speed and segmentation quality. A comparison of the models using different f and τ is shown in Tab. 7.1.

7.4.2 Video Adaption

Inspired by this blog [Bazarevsky and Tkachenka, 2018], we implemented two strategies to adapt our model to video face parsing.

For inference, we take the mask of the frame $t-1$ as input to the frame t to stabilize the segmentation. Due to the lack of video dataset, we apply a randomly transformed mask as a fake previous mask during the training. According to our experiments, this will eliminate the random segmentation noise which is present without taking the previous frame mask as input.

We add four dense layers [Huang et al., 2017] with a growth rate of 8 at the end of the network for more robust rendering.

7.5 Experiments

We train our model on the Helen dataset [Smith et al., 2013] which contains 2330 images manually labeled in 11 classes including background, skin, hair nose, left/right eyebrows, left/right eyes, upper/bottom lips as well as inner-mouth. The models are trained on all of the labels except the hair because the hair annotations are not precise. The images are cropped with a margin of 30%-70% of bounding box size according to the facial landmark annotations.

Model	Overall F-score	Num. of parameters
SegNet [Badrinarayanan et al., 2017]	92.90	29.45M
Unet [Ronneberger et al., 2015]	93.72	13.40M
Mobile-Hourglass-f16-τ3	93.00	0.09M
Mobile-Hourglass-f16-τ6	93.08	0.12M
Mobile-Hourglass-f32-τ3	93.18	0.17M
Mobile-Hourglass-f32-τ6	93.55	0.27M

TABLE 7.1: Quantitative evaluation of face parsing results on Helen dataset.

We use the RMSprop as optimizer and softmax cross-entropy function as loss function. We apply an initial learning rate of 0.0005 with decay of 0.1 for each 40 training epochs until total 190 epochs are finished. We use ONNX as an intermediate format to transform our Pytorch model to CoreML model, which is optimized for iOS devices. We adopt the Vision framework for face detection that is anterior to face parsing .

We compare our methods with several well-known segmentation networks [Ronneberger et al., 2015, Badrinarayanan et al., 2017]. The results are measured in F1-score in Table 7.1. The number of parameters are also listed aside to provide more information about the model size, which is critic for mobile platform.

We measure the runtime of different MobileNet block settings by changing the number of channels f and expansion factor τ. We also provide a profiling time analysis on the face detection, array transformation and colorization in Tab. 7.2. We find that the face parsing is able to run in real-time on mobile phones.

7.6 Conclusion

In this chapter, we presented a real-time encoder-decoder video face parsing mobile AR demonstration. Using a specific neural architecture that takes into account the estimation of the previous time step and reducing the computational efficiency by depth-wise factorised convolutions (MobileNet), we were able to show the feasibility of mobile face parsing in real-time. Furthermore:

- Face parsing is crucial and feasible for numerous facial editing applications for example virtual make-up, hair dying, skin analysis, face morphing, reenactment etc.
- Our method is not only limited to facial AR applications but also interesting for other fine-grained segmentation based AR applications, such as foreground/background extraction.

The contribution in this chapter led to a demonstration paper published at *ISMAR*.

Model	Face Detection	Inference	Colorization	Total
Unet [Ronneberger et al., 2015]	7	203	54	269
Mobile-Hourglass-f16-τ3	8	77	18	106
Mobile-Hourglass-f16-τ6	7	75	18	103
Mobile-Hourglass-f32-τ3	7	76	18	104
Mobile-Hourglass-f32-τ6	7	78	18	106
Dense Mobile-Hourglass-f32-τ6	7	75	18	110

TABLE 7.2: Run-time (in ms) profiling on iPhone X.

Chapter 8

Conclusion

8.1 Summary of Contributions

In this thesis work, we have pushed further two aspects of deep face analysis, namely facial landmark detection and face parsing. We have introduced several methods to improve both of the tasks in terms of precision, robustness and speed. Live application of aesthetic augmented reality applications could benefit from the better facial component detection introduced in this thesis. Moreover, we also take a closer look and gained important insights into the deep CNN models which are usually considered as black-boxes. A summary of the contributions in each chapter is listed below:

Facial landmark detection: In chapter 3, we tackled the problem of the imprecise landmark detection, especially for coordinate regression models. Our contribution is three-fold: (1) we proposed CropNet to learn the local misalignment on extracted patches in a coarse-to-fine manner (2) we proposed a loss function that is sensitive to semantic boundaries, which forces the predicted landmarks to stay on these boundaries (3) we proposed to use different loss functions at different stages, such that both big errors and small errors can be reduced accordingly in different places.

In chapter 4, we presented several contributions to increasing the robustness of facial landmark detection and the way to measure it. We demonstrated that the current metrics for robustness can no longer effectively benchmark the state-of-the-art methods, due to the performance saturation. Based on this observation, we proposed several modifications to the current metrics, including the landmark-wise failure rate, cross dataset validation and synthetic occlusions. Then we proposed a novel 2D Wasserstein loss function for the heatmap regression models, which significantly improves the robustness of the model. The proposed solution can be easily generalized to most of the heatmap regression models without any computational overhead during the inference.

Based on the analysis in chapters 3 and 4, we proposed to differentiate the characteristics of the models via quantitative values. In chapter 5, we developed a new tool for advancing the research in facial landmark detection. The proposed approach statistically analyzes the correlation among the annotated (or predicted) facial landmarks on large-scale datasets. We confirmed the universality of landmark correlation on several datasets and train/validation subsets. We made several important observations through this analysis, including (1) CNN tend to correlate adjacent landmarks. (2) Heatmap regression models are more likely to violate landmark correlation than coordinate regression models under challenging conditions. (3) Cascading and stacking behave quite differently in facial landmark detection models. We also studied the learning dynamics of the coordinate regression models. In addition, by exploiting these results and determining the most correlated landmarks, we developed a weakly-supervised learning method to search for an efficient sparse annotation format, which leads to better results compared to existing sparse formats. Our weakly-supervised learning method significantly reduces the cost of the laborious manual dense annotation.

Face parsing: In chapter 6, we presented a method for the segmentation of a challenging facial component: human hair, especially under difficult and unconstrained conditions ("*in the wild*"). We demonstrated that one of the major difficulties of human hair segmentation lies in the interference under noise from cluttered background. To alleviate this problem, our proposed deep shape prior provides an attention mechanism to guide the segmentation network to deliver improved results with cluttered background. In addition, we integrated a border refinement module into our model to provide a more consistent hair boundary, which is visually more pertinent and appealing.

In chapter 7, we proposed a real-time face parsing demo working on iPhone. By establishing a more efficient pipeline with MobileNet [Sandler et al., 2018], our demo confirms the feasibility of the aesthetic AR application based on face parsing. In our proposed model and implementation, we also integrated the dense layers and the re-injection of the output from the last frame to increase the consistency across different frames.

8.2 Future Work

In this section, we present several future directions of research that could come out from this thesis.

8.2.1 Short-term Future Work

Facial landmark detection: Compared to human body or human hand, human face presents a more rigid shape. Nonetheless, it still involves plenty of shape variances. The noise (such as occlusion and blur) also introduces significant perturbation for the detection. Therefore, for this task, it is always difficult to balance the trade-off between precision and robustness. A model achieves more locally precise results may lack robustness under challenging conditions. However, a model which is conditioned with stronger shape constraints, is less likely to learn complex shape variances. We think that it may be useful to develop a mechanism that determines if the input face image has been taken under challenging conditions or not. In this case, the compromise between the precision and the robustness can be arranged accordingly.

Another promising research direction of facial landmark detection is unsupervised/weakly-supervised learning. At present, it is easy to collect billions of face images on the Internet. However, the manual annotation for facial landmark is expensive. Most of the recent research of unsupervised/weakly-supervised facial landmark detection is still limited on the annotated datasets, which is not able to utilize the massive unlabeled data available on the Internet.

Hair segmentation: Human hair segmentation is still a challenging problem for the practical use of AR applications. First, the size of the annotated datasets is quite limited, especially under difficult (*in the wild*) conditions. Nonetheless, obtaining ground truth for segmentation is both expensive and time-consuming. Unsupervised or semi-supervised learning could be an interesting and promising direction.

Second, in the current research of semantic segmentation, the texture of the object is rarely considered in an explicit manner. For deep learning-based models, we leave the CNN to learn the texture automatically. However, human hair presents unique textural information, which could be more explicitly exploited to improve the performance.

Face parsing: Compared to facial landmark detection, face parsing is a subject which is less studied in the literature. Most of the research of this subject is closely related to the general semantic segmentation. Therefore, the shape constraints on human face is rarely

integrated into the learning process. We think that introducing these constraints into the face parsing models will significantly regularize the face parsing result, which could be helpful for AR applications.

Video: Most of the discussions in this thesis, except Chapter 7, are focused on images rather than videos. However, it is obvious that most of the AR applications are based on video. The question of how to efficiently exploit the temporal information for facial landmark detection and hair segmentation is of great importance for the future research on these subjects.

8.2.2 Long-term Future Work

We think the most important subject in the longer term is to delve deeper into the interpretation and the mechanism of CNN. A better understanding of the CNN could subsequently guide us to find better methods and models for deep facial landmark detection and deep face parsing.

Inter-landmark relationship and CNN interpretation: To perceive a further understanding on the working mechanism of the CNN, our CCA analysis could be further extended to human pose, general object landmarks and the bounding box regression (in object detection). Moreover, this analysis could be generalized to a broader usage. For example, in the task of image recognition, we could consider the prediction of each category as a separate task. In this way, we are able to figure out the correlation among different categories. We think that it could provide another perspective to interpret the behavior of a CNN for a more general purpose.

How the geometric information is encoded in CNNs: Another prominent long-term future direction is to understand how geometric information is encoded in CNNs. The convolution operation is known as translation-invariant or translation-equivariant, which means that it is not sensitive to the absolute geometric location on the input image. On contrary, we find that the CNN performs quite well on the facial landmark detection task. It still remains difficult to explain how CNNs can predict the position of the landmarks in an end-to-end manner. A recent work argued that the *padding* operation in the convolutional layers plays an important role in the learning of the geometric information [Islam et al., 2020]. We believe that further investigation on this issue could guide the design of the facial landmark detection and face parsing models.

Bibliography

[Ablavatski and Grishchenko, 2019] Ablavatski, A. and Grishchenko, I. (2019). Real-time ar self-expression with machine learning. https://ai.googleblog.com/2019/03/real-time-ar-self-expression-with.html.

[Alabort-i Medina and Zafeiriou, 2014] Alabort-i Medina, J. and Zafeiriou, S. (2014). Bayesian active appearance models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3445.

[Alashkar et al., 2017] Alashkar, T., Jiang, S., Wang, S., and Fu, Y. (2017). Examples-rules guided deep neural network for makeup recommendation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[Amos et al., 2016] Amos, B., Ludwiczuk, B., and Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

[Andriluka et al., 2014] Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Asthana et al., 2013] Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3451.

[Asthana et al., 2014] Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M. (2014). Incremental face alignment in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1859–1866.

[Badrinarayanan et al., 2017] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(12):2481–2495.

[Baltrusaitis et al., 2013] Baltrusaitis, T., Robinson, P., and Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*.

[Bazarevsky and Tkachenka, 2018] Bazarevsky, V. and Tkachenka, A. (2018). Mobile real-time video segmentation. https://ai.googleblog.com/2018/03/mobile-real-time-video-segmentation.html.

[Belagiannis et al., 2015] Belagiannis, V., Rupprecht, C., Carneiro, G., and Navab, N. (2015). Robust optimization for deep regression. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2830–2838.

[Belhumeur et al., 2013] Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2930–2940.

[Belmonte et al., 2019] Belmonte, R., Ihaddadene, N., Tirilly, P., Bilasco, I. M., and Djeraba, C. (2019). Video-based face alignment with local motion modeling. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2106–2115. IEEE.

[Bichlmeier et al., 2007] Bichlmeier, C., Wimmer, F., Heining, S. M., and Navab, N. (2007). Contextual anatomic mimesis hybrid in-situ visualization method for improving multi-sensory depth perception in medical augmented reality. In *IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*.

[Blanz et al., 1999] Blanz, V., Vetter, T., et al. (1999). A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, volume 99, pages 187–194.

[Bottou, 2010] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Computational Statistics*, pages 177–186. Springer.

[Bulat and Tzimiropoulos, 2017a] Bulat, A. and Tzimiropoulos, G. (2017a). Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *IEEE International Conference on Computer Vision (ICCV)*.

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

[Burgos-Artizzu et al., 2013] Burgos-Artizzu, X. P., Perona, P., and Dollár, P. (2013). Robust face landmark estimation under occlusion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1513–1520.

[Cao et al., 2014] Cao, X., Wei, Y., Wen, F., and Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision (IJCV)*, 107(2):177–190.

[Caudell and Mizell, 1992] Caudell, T. P. and Mizell, D. W. (1992). Augmented reality: an application of heads-up display technology to manual manufacturing processes. In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, volume ii, pages 659–669 vol.2.

[Çeliktutan et al., 2013] Çeliktutan, O., Ulukaya, S., and Sankur, B. (2013). A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1):13.

[Chai et al., 2016] Chai, M., Shao, T., Wu, H., Weng, Y., and Zhou, K. (2016). Autohair: fully automatic hair modeling from a single image. *ACM Transactions on Graphics (ToG)*, 35(4).

[Chang et al., 2017] Chang, C.-H., Chou, C.-N., and Chang, E. Y. (2017). Clkn: Cascaded lucas-kanade networks for image alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2213–2221.

[Chang et al., 2018] Chang, H., Lu, J., Yu, F., and Finkelstein, A. (2018). Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40–48.

[Chen et al., 2019a] Chen, H.-J., Hui, K.-M., Wang, S.-Y., Tsao, L.-W., Shuai, H.-H., and Cheng, W.-H. (2019a). Beautyglow: On-demand makeup transfer framework with reversible generative network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10042–10050.

[Chen et al., 2019b] Chen, L., Su, H., and Ji, Q. (2019b). Face alignment with kernel density deep neural network. In *IEEE International Conference on Computer Vision (ICCV)*.

[Chen et al., 2015] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2015). Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International Conference on Learning Representations (ICLR)*.

[Chen et al., 2018a] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848.

[Chen et al., 2018b] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*.

[Chen et al., 2017a] Chen, X., Zhou, E., Mo, Y., Liu, J., and Cao, Z. (2017a). Delving deep into coarse-to-fine framework for facial landmark localization. In *IEEE Conference on Computer Vision and Pattern Recognition Worshop (CVPRW)*, pages 2088–2095.

[Chen et al., 2018c] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., and Sun, J. (2018c). Cascaded pyramid network for multi-person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Chen et al., 2017b] Chen, Y.-C., Shen, X., and Jia, J. (2017b). Makeup-go: Blind reversion of portrait edit. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4501–4509.

[Chrysos et al., 2014] Chrysos, G. G., Antonakos, E., Snape, P., Asthana, A., and Zafeiriou, S. (2014). A comprehensive performance evaluation of deformable face tracking "in-the-wild". *International Journal of Computer Vision (IJCV)*, pages 1–35.

[Chrysos et al., 2015] Chrysos, G. G., Antonakos, E., Zafeiriou, S., and Snape, P. (2015). Offline Deformable Face Tracking in Arbitrary Videos. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 954–962.

[Cimpoi et al., 2015] Cimpoi, M., Maji, S., and Vedaldi, A. (2015). Deep filter banks for texture recognition and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3828–3836.

[Cootes et al., 2001] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(6):681–685.

[Cootes et al., 1995] Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding (CVIU)*, 61(1):38–59.

[Cordts et al., 2016] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Cristinacce and Cootes, 2006] Cristinacce, D. and Cootes, T. F. (2006). Feature detection and tracking with constrained local models. In *The British Machine Vision Conference (BMVC)*.

[Cristinacce and Cootes, 2007] Cristinacce, D. and Cootes, T. F. (2007). Boosted regression active shape models. In *The British Machine Vision Conference (BMVC)*, volume 2, pages 880–889.

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Dapogny et al., 2019] Dapogny, A., Bailly, K., and Cord, M. (2019). Decafa: Deep convolutional cascade for face alignment in the wild. In *IEEE International Conference on Computer Vision (ICCV)*.

[Daza, 2019] Daza, D. (2019). Approximating wasserstein distances with pytorch. https://dfdazac.github.io/sinkhorn.html.

[Dhall et al., 2009] Dhall, A., Sharma, G., Bhatt, R., and Khan, G. M. (2009). Adaptive digital makeup. In *International Symposium on Visual Computing*, pages 728–736. Springer.

[Ding and Tao, 2016] Ding, C. and Tao, D. (2016). A comprehensive survey on pose-invariant face recognition. *ACM Transactions on Intelligent Systems and Technology*, 7(3):37:1–37:42.

[Dollár et al., 2010] Dollár, P., Welinder, P., and Perona, P. (2010). Cascaded pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Dong et al., 2018a] Dong, X., Yan, Y., Ouyang, W., and Yang, Y. (2018a). Style aggregated network for facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Dong and Yang, 2019] Dong, X. and Yang, Y. (2019). Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *IEEE International Conference on Computer Vision (ICCV)*.

[Dong et al., 2018b] Dong, X., Yu, S.-I., Weng, X., Wei, S.-E., Yang, Y., and Sheikh, Y. (2018b). Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Dou et al., 2017] Dou, P., Shah, S. K., and Kakadiaris, I. A. (2017). End-to-end 3d face reconstruction with deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5908–5917.

[Duffner, 2008] Duffner, S. (2008). *Face image analysis with convolutional neural networks*. PhD thesis, University of Freiburg.

[Duffner and Garcia, 2005a] Duffner, S. and Garcia, C. (2005a). A connexionist approach for robust and precise facial feature detection in complex scenes. In *International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 316–321, Zagreb, Croatia.

[Duffner and Garcia, 2005b] Duffner, S. and Garcia, C. (2005b). A hierarchical approach for precise facial feature detection. In *Compression et Représentation des Signaux Audiovisuels (CORESA)*, pages 29–34, Rennes, France.

[Duffner and Garcia, 2007] Duffner, S. and Garcia, C. (2007). Face recognition using non-linear image reconstruction. In *International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 459–464, London, UK.

[Duffner and Garcia, 2008] Duffner, S. and Garcia, C. (2008). Robust face alignment using convolutional neural networks. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, Funchal, Portugal.

[Edwards et al., 1998] Edwards, G. J., Taylor, C. J., and Cootes, T. F. (1998). Interpreting face images using active appearance models. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 300–305. IEEE.

[Everingham et al., 2015] Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136.

[Fan and Zhou, 2016] Fan, H. and Zhou, E. (2016). Approaching human level facial landmark localization by deep learning. *Image and Vision Computing (IVC)*, 47:27–35.

[Felzenszwalb et al., 2008] Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[Feng et al., 2018a] Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2018a). Joint 3d face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision (ECCV)*, pages 534–551.

[Feng et al., 2017a] Feng, Z.-H., Kittler, J., Awais, M., Huber, P., and Wu, X.-J. (2017a). Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition Worshop (CVPRW)*.

[Feng et al., 2018b] Feng, Z.-H., Kittler, J., Awais, M., Huber, P., and Wu, X.-J. (2018b). Wing loss for robust facial landmark localisation with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Feng et al., 2017b] Feng, Z.-H., Kittler, J., Christmas, W., Huber, P., and Wu, X.-J. (2017b). Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3681–3690.

[Feraund et al., 2001] Feraund, R., Bernier, O. J., Viallet, J.-E., and Collobert, M. (2001). A fast and accurate face detector based on neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(1):42–53.

[Fritsch et al., 2013] Fritsch, J., Kuehnl, T., and Geiger, A. (2013). A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*.

[Fukushima, 1975] Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, 20(3-4):121–136.

[Gao et al., 2016] Gao, M., Chen, H., Zheng, S., and Fang, B. (2016). A factorization based active contour model for texture segmentation. In *International Conference on Image Processing (ICIP)*, pages 4309–4313. IEEE.

[Garcia and Delakis, 2002] Garcia, C. and Delakis, M. (2002). A neural architecture for fast and robust face detection. In *International Conference on Pattern Recognition (ICPR)*, volume 2, pages 44–47. IEEE.

[Garcia and Delakis, 2004] Garcia, C. and Delakis, M. (2004). Convolutional face finder: A neural architecture for fast and robust face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(11):1408–1423.

[Gatys et al., 2016] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423.

[Girshick, 2015] Girshick, R. (2015). Fast r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*.

[Gong et al., 2014] Gong, Y., Wang, L., Guo, R., and Lazebnik, S. (2014). Multi-scale orderless pooling of deep convolutional activation features. In *European Conference on Computer Vision (ECCV)*, pages 392–407. Springer.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680.

[Gu et al., 2017] Gu, J., Yang, X., De Mello, S., and Kautz, J. (2017). Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Gu et al., 2019] Gu, Q., Wang, G., Chiu, M. T., Tai, Y.-W., and Tang, C.-K. (2019). Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10481–10490.

[Guo and Sim, 2009] Guo, D. and Sim, T. (2009). Digital face makeup by example. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 73–79. IEEE.

[Guo and Aarabi, 2018] Guo, W. and Aarabi, P. (2018). Hair segmentation using heuristically-trained neural networks. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 29(1):25–36.

[Gurobi, 2019] Gurobi (2019). Gurobi optimizer reference manual.

[He et al., 2017a] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017a). Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

[He et al., 2015a] He, K., Zhang, X., Ren, S., and Sun, J. (2015a). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

[He et al., 2015b] He, K., Zhang, X., Ren, S., and Sun, J. (2015b). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(9):1904–1916.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[He et al., 2017b] He, Z., Kan, M., Zhang, J., Chen, X., and Shan, S. (2017b). A fully end-to-end cascaded cnn for facial landmark detection. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 200–207.

[Hinton, 2002] Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, pages 1771–1800.

[Hochbaum and Shmoys, 1985] Hochbaum, D. S. and Shmoys, D. B. (1985). A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184.

[Holschneider et al., 1990] Holschneider, M., Kronland-Martinet, R., Morlet, J., and Tchamitchian, P. (1990). A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, pages 286–297. Springer.

[Hotelling, 1936] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

[Hou et al., 2018] Hou, Q., Wang, J., Bai, R., Zhou, S., and Gong, Y. (2018). Face alignment recurrent network. *Pattern Recognition (PR)*, 74:448–458.

[Hu et al., 2018] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Huang et al., 2017] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Huang et al., 2015] Huang, Z., Zhou, E., and Cao, Z. (2015). Coarse-to-fine face alignment with multi-scale local patch regression. *arXiv preprint arXiv:1511.04901*.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456.

[Isberto, 2018] Isberto, M. (2018). The history of augmented reality. https://www.colocationamerica.com/blog/history-of-augmented-reality. Accessed: 2019-10-20.

[Islam et al., 2020] Islam, M. A., Jia, S., and Bruce, N. D. B. (2020). How much position information do convolutional neural networks encode? In *International Conference on Learning Representations (ICLR)*.

[Jackson et al., 2017] Jackson, A. S., Bulat, A., Argyriou, V., and Tzimiropoulos, G. (2017). Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *IEEE International Conference on Computer Vision (ICCV)*.

[Jaderberg et al., 2015] Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2017–2025.

[Jeni et al., 2015] Jeni, L. A., Cohn, J. F., and Kanade, T. (2015). Dense 3D face alignment from 2D videos in real-time. *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*.

[Jiang et al., 2018] Jiang, P., Gu, F., Wang, Y., Tu, C., and Chen, B. (2018). Difnet: Semantic segmentation by diffusion networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1630–1639.

[Jin et al., 2019] Jin, X., Han, R., Ning, N., Li, X., and Zhang, X. (2019). Facial makeup transfer combining illumination transfer. *IEEE Access*, 7:80928–80936.

[Jin and Tan, 2017] Jin, X. and Tan, X. (2017). Face alignment in-the-wild: A survey. *Computer Vision and Image Understanding (CVIU)*, 162:1 – 22.

[Kae et al., 2013] Kae, A., Sohn, K., Lee, H., and Learned-Miller, E. (2013). Augmenting CRFs with Boltzmann machine shape priors for image labeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Kazemi and Sullivan, 2014] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Kingma and Dhariwal, 2018] Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10215–10224.

[Koestinger et al., 2011] Koestinger, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 2144–2151.

[Kornblith et al., 2019] Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*.

[Kowalski et al., 2017] Kowalski, M., Naruniec, J., and Trzcinski, T. (2017). Deep alignment network: A convolutional neural network for robust face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition Worshop (CVPRW)*, pages 88–97.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105.

[Kumar and Chellappa, 2018] Kumar, A. and Chellappa, R. (2018). Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 430–439.

[Kumar et al., 2016] Kumar, A., Ranjan, R., Patel, V., and Chellappa, R. (2016). Face alignment by local deep descriptor regression. *arXiv preprint arXiv:1601.07950*.

[Lai et al., 2016] Lai, H., Xiao, S., Pan, Y., Cui, Z., Feng, J., Xu, C., Yin, J., and Yan, S. (2016). Deep recurrent regression for facial landmark detection. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 28(5):1144–1157.

[Lathuilière et al., 2018] Lathuilière, S., Mesejo, P., Alameda-Pineda, X., and Horaud, R. (2018). Deepgum: Learning deep robust regression with a gaussian-uniform mixture model. In *European Conference on Computer Vision (ECCV)*, pages 205–221.

[Lawrence et al., 1997] Lawrence, S., Giles, C. L., Tsoi, A. C., and Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113.

[Le et al., 2012] Le, V., Brandt, J., Lin, Z., Bourdev, L., and Huang, T. S. (2012). Interactive facial feature localization. In *European Conference on Computer Vision (ECCV)*, pages 679–692.

[LeCun et al., 1990] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 396–404.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[LeCun et al., 1989] LeCun, Y. et al. (1989). Generalization and network design strategies. In *Connectionism in perspective*, volume 19. Citeseer.

[Lee et al., 2008] Lee, K.-c., Anguelov, D., Sumengen, B., and Gokturk, S. B. (2008). Markov random field models for hair and face segmentation. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE.

[Leung and Malik, 2001] Leung, T. and Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision (IJCV)*, 43(1):29–44.

[Levinshtein et al., 2017] Levinshtein, A., Chang, C., Phung, E., Kezele, I., Guo, W., and Aarabi, P. (2017). Real-time deep hair matting on mobile devices. *arXiv preprint arXiv:1712.07168*.

[Li et al., 2015] Li, C., Zhou, K., and Lin, S. (2015). Simulating makeup through physics-based manipulation of intrinsic image layers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4621–4629.

[Li et al., 2019] Li, J., Zhou, P., Chen, Y., Zhao, J., Roy, S., Shuicheng, Y., Feng, J., and Sim, T. (2019). Task relation networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[Li et al., 2018] Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., and Lin, L. (2018). Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *ACM International Conference on Multimedia (ACM MM)*, pages 645–653.

[Li et al., 2017] Li, X., Liu, Z., Luo, P., Chen, C. L., and Tang, X. (2017). Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Liao et al., 2017] Liao, J., Yao, Y., Yuan, L., Hua, G., and Kang, S. B. (2017). Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics (TOG)*, 36(4):1–15.

[Lin et al., 2019] Lin, J., Yang, H., Chen, D., Zeng, M., Wen, F., and Yuan, L. (2019). Face parsing with roi tanh-warping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5654–5663.

[Liu et al., 2017a] Liu, H., Lu, J., Feng, J., and Zhou, J. (2017a). Two-stream transformer networks for video-based face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

[Liu et al., 2018a] Liu, L., Chen, J., Fieguth, P., Zhao, G., Chellappa, R., and Pietikainen, M. (2018a). A survey of recent advances in texture representation. *arXiv preprint arXiv:1801.10324*.

[Liu et al., 2017b] Liu, L., Fan, S., Ning, X., and Liao, L. (2017b). An efficient level set model with self-similarity for texture segmentation. *Neurocomputing*, 266:150–164.

[Liu et al., 2014] Liu, L., Xing, J., Liu, S., Xu, H., Zhou, X., and Yan, S. (2014). Wow! you are so beautiful today! *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(1s):1–22.

[Liu et al., 2018b] Liu, R., Lehman, J., Molino, P., Such, F. P., Frank, E., Sergeev, A., and Yosinski, J. (2018b). An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Liu et al., 2017c] Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.-H., and Kautz, J. (2017c). Learning affinity via spatial propagation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1520–1530.

[Liu et al., 2016] Liu, S., Ou, X., Qian, R., Wang, W., and Cao, X. (2016). Makeup like a superstar: deep localized makeup transfer network. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2568–2575.

[Liu et al., 2017d] Liu, S., Shi, J., Liang, J., and Yang, M.-H. (2017d). Face parsing via recurrent propagation. In *The British Machine Vision Conference (BMVC)*.

[Liu et al., 2015] Liu, S., Yang, J., Huang, C., and Yang, M.-H. (2015). Multi-objective convolutional learning for face labeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3451–3459.

[Liu et al., 2019] Liu, Z., Zhu, X., Hu, G., Guo, H., Tang, M., Lei, Z., Robertson, N. M., and Wang, J. (2019). Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Long et al., 2015] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440.

[Low and Ibrahim, 1997] Low, B. and Ibrahim, M. (1997). A fast and accurate algorithm for facial feature segmentation. In *International Conference on Image Processing (ICIP)*, volume 2, pages 518–521. IEEE.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, pages 91–110.

[Luo et al., 2012] Luo, P., Wang, X., and Tang, X. (2012). Hierarchical face parsing via deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2480–2487. IEEE.

[Luvizon et al., 2018] Luvizon, D. C., Picard, D., and Tabia, H. (2018). 2d/3d pose estimation and action recognition using multitask deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Lv et al., 2017] Lv, J., Shao, X., Xing, J., Cheng, C., and Zhou, X. (2017). A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3317–3326.

[Martinez et al., 2017] Martinez, B., Valstar, M. F., Jiang, B., and Pantic, M. (2017). Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*.

[Merget et al., 2018] Merget, D., Rock, M., and Rigoll, G. (2018). Robust facial landmark detection via a fully-convolutional local-global context network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Messer et al., 1999] Messer, K., Matas, J., Kittler, J., Luettin, J., and Maitre, G. (1999). Xm2vtsdb: The extended m2vts database. In *International Conference on Audio and Video-based Biometric Person Authentication*, volume 964, pages 965–966.

[Miao et al., 2018] Miao, X., Zhen, X., Liu, X., Deng, C., Athitsos, V., and Huang, H. (2018). Direct shape regression networks for end-to-end face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Morcos et al., 2018] Morcos, A., Raghu, M., and Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5727–5736.

[Muhammad et al., 2018] Muhammad, U. R., Svanera, M., Leonardi, R., and Benini, S. (2018). Hair detection, segmentation, and hairstyle classification in the wild. *Image and Vision Computing (IVC)*, 71:25–37.

[Newell et al., 2016] Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*.

[Nguyen and Liu, 2017] Nguyen, T. V. and Liu, L. (2017). Smart mirror: Intelligent makeup recommendation and synthesis. In *ACM International Conference on Multimedia (ACM MM)*, pages 1253–1254.

[Nibali et al., 2018] Nibali, A., He, Z., Morgan, S., and Prendergast, L. (2018). Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*.

[Osadchy et al., 2007] Osadchy, M., Cun, Y. L., and Miller, M. L. (2007). Synergistic face detection and pose estimation with energy-based models. *Journal of Machine Learning Research (JMLR)*, 8(May):1197–1215.

[Parkhi et al., 2015] Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition. In *The British Machine Vision Conference (BMVC)*.

[Peng et al., 2016a] Peng, X., Feris, R. S., Wang, X., and Metaxas, D. N. (2016a). A recurrent encoder-decoder network for sequential face alignment. In *European Conference on Computer Vision (ECCV)*, pages 38–56.

[Peng et al., 2016b] Peng, X., Huang, J., and Metaxas, D. N. (2016b). Sequential face alignment via person-specific modeling in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition Worshop (CVPRW)*, pages 107–116.

[Prabhu et al., 2010] Prabhu, U., Seshadri, K., and Savvides, M. (2010). Automatic facial landmark tracking in video sequences using kalman filter assisted active shape models. In *European Conference on Computer Vision (ECCV)*, pages 86–99.

[Qin et al., 2017] Qin, S., Kim, S., and Manduchi, R. (2017). Automatic skin and hair masking using fully convolutional networks. In *International Conference on Multimedia and Expo (ICME)*, pages 103–108. IEEE.

[Radu, 2012] Radu, I. (2012). Why should my students use ar? a comparative review of the educational impacts of augmented-reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*.

[Raghu et al., 2017] Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6076–6085.

[Ranjan et al., 2017] Ranjan, R., Patel, V. M., and Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(1):121–135.

[Reinders et al., 1996] Reinders, M. J., Koch, R., and Gerbrands, J. J. (1996). Locating facial features in image sequences using neural networks. In *International conference on Automatic Face and Gesture Recognition (FG)*, pages 230–235. IEEE.

[Ren et al., 2014] Ren, S., Cao, X., Wei, Y., and Sun, J. (2014). Face alignment at 3000 fps via regressing local binary features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1685–1692.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99.

[Ren et al., 2019] Ren, Y., Sun, Y., Jing, X., Cui, Z., and Shi, Z. (2019). Adaptive makeup transfer via bat algorithm. *Mathematics*, 7(3):273.

[Reska et al., 2015] Reska, D., Boldak, C., and Kretowski, M. (2015). A texture-based energy for active contour image segmentation. In *Image Processing & Communications Challenges 6*, pages 187–194. Springer.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer.

[Rosenberg, 1993] Rosenberg, L. B. (1993). Virtual fixtures: Perceptual tools for telerobotic manipulation. In *IEEE Virtual Reality Annual International Symposium (VR)*.

[Rosenblatt, 1961] Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY.

[Rousset and Coulon, 2008] Rousset, C. and Coulon, P.-Y. (2008). Frequential and color analysis for hair mask segmentation. In *International Conference on Image Processing (ICIP)*, pages 2276–2279. IEEE.

[Rowley et al., 1998] Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20(1):23–38.

[Rumelhart et al., 1988] Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, 5(3):1.

[Sagonas et al., 2013] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE Conference on Computer Vision and Pattern Recognition Worshop (CVPRW)*, pages 397–403.

[Saito et al., 2016] Saito, S., Li, T., and Li, H. (2016). Real-time facial segmentation and performance capture from rgb input. In *European Conference on Computer Vision (ECCV)*, pages 244–261. Springer.

[Sánchez-Lozano et al., 2016] Sánchez-Lozano, E., Martinez, B., Tzimiropoulos, G., and Valstar, M. (2016). Cascaded continuous regression for real-time incremental face tracking. In *European Conference on Computer Vision (ECCV)*, pages 645–661.

[Sánchez-Lozano et al., 2017] Sánchez-Lozano, E., Tzimiropoulos, G., Martinez, B., De la Torre, F., and Valstar, M. (2017). A functional regression approach to facial landmark tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(9):2037–2050.

[Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520.

[Schwab, 2013] Schwab, S. (2013). *Suivi visuel multi-cibles par partitionnement de détections: application à la construction d'albums de visages*. PhD thesis, Clermont-Ferrand 2.

[Shah et al., 2016] Shah, A., Kadam, E., Shah, H., Shinde, S., and Shingade, S. (2016). Deep residual networks with exponential linear unit. In *International Symposium on Computer Vision and the Internet*, pages 59–65. ACM.

[Shao et al., 2016] Shao, Z., Ding, S., Zhao, Y., Zhang, Q., and Ma, L. (2016). Learning deep representation from coarse to fine for face alignment. In *IEEE International Conference on Multimedia and Expo (ICME)*.

[Shao et al., 2019] Shao, Z., Zhu, H., Tan, X., Hao, Y., and Ma, L. (2019). Deep multi-center learning for face alignment. *Neurocomputing*.

[Shen et al., 2015] Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., and Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 50–58.

[Smith et al., 2014] Smith, B. M., Brandt, J., Lin, Z., and Zhang, L. (2014). Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Smith and Zhang, 2014] Smith, B. M. and Zhang, L. (2014). Collaborative facial landmark localization for transferring annotations across datasets. In *European Conference on Computer Vision (ECCV)*.

[Smith et al., 2013] Smith, B. M., Zhang, L., Brandt, J., Lin, Z., and Yang, J. (2013). Exemplar-based face parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3484–3491.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958.

[Sun et al., 2019a] Sun, K., Wu, W., Liu, T., Yang, S., Wang, Q., Zhou, Q., Ye, Z., and Qian, C. (2019a). Fab: A robust facial landmark detection framework for motion-blurred videos. In *IEEE International Conference on Computer Vision (ICCV)*.

[Sun et al., 2019b] Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., and Wang, J. (2019b). High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*.

[Sun et al., 2015] Sun, P., Min, J. K., and Xiong, G. (2015). Globally tuned cascade pose regression via back propagation with application in 2d face pose estimation and heart segmentation in 3d ct images. In *IEEE International Conference on Computer and Information Technology Workshops*.

[Sun et al., 2018] Sun, X., Xiao, B., Wei, F., Liang, S., and Wei, Y. (2018). Integral human pose regression. In *European Conference on Computer Vision (ECCV)*, pages 529–545.

[Sun et al., 2013] Sun, Y., Wang, X., and Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3476–3483.

[Sung and Kim, 2009] Sung, J. and Kim, D. (2009). Adaptive active appearance model with incremental learning. *Pattern Recognition Letters (PRL)*, pages 359–367.

[Sutherland, 1968] Sutherland, I. E. (1968). A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, AFIPS '68 (Fall, part I), pages 757–764, New York, NY, USA. ACM.

[Tai et al., 2019] Tai, Y., Liang, Y., Liu, X., Duan, L., Li, J., Wang, C., Huang, F., and Chen, Y. (2019). Towards highly accurate and stable face alignment for high-resolution videos. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[Tan et al., 2017] Tan, S., Chen, D., Guo, C., and Huang, Z. (2017). Extra facial landmark localization via global shape reconstruction. *Computational intelligence and neuroscience*, 2017.

[Tang et al., 2018] Tang, Z., Peng, X., Geng, S., Wu, L., Zhang, S., and Metaxas, D. (2018). Quantized densely connected u-nets for efficient landmark localization. In *European Conference on Computer Vision (ECCV)*.

[The Goldman Sachs Group, 2016] The Goldman Sachs Group (2016). Virtual and augmented reality: understanding the race for the next computing platform. https://www.goldmansachs.com/insights/pages/technology-driving-innovation-folder/virtual-and-augmented-reality/report.pdf.

[Tieleman and Hinton, 2012] Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

[Tong et al., 2007] Tong, W.-S., Tang, C.-K., Brown, M. S., and Xu, Y.-Q. (2007). Example-based cosmetic transfer. In *Pacific Conference on Computer Graphics and Applications (PG)*, pages 211–218. IEEE.

[Tralie, 2018] Tralie, C. (2018). 2d histogram wasserstein distance via pot library. https://gist.github.com/ctralie/66352ae6ab06c009f02c705385a446f3.

[Trigeorgis et al., 2016] Trigeorgis, G., Snape, P., Nicolaou, M. A., Antonakos, E., and Zafeiriou, S. (2016). Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4177–4187.

[Tzimiropoulos and Pantic, 2013] Tzimiropoulos, G. and Pantic, M. (2013). Optimization problems for fast aam fitting in-the-wild. In *IEEE International Conference on Computer Vision (ICCV)*, pages 593–600.

[Vaillant et al., 1994] Vaillant, R., Monrocq, C., and Le Cun, Y. (1994). Original approach for the localisation of objects in images. *IEEE Proceedings-Vision, Image and Signal Processing*, 141(4):245–250.

[Valle et al., 2018] Valle, R., Buenaposada, J. M., Valdés, A., and Baumela, L. (2018). A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *European Conference on Computer Vision (ECCV)*.

[Van Gool, 2012] Van Gool, L. (2012). Real-time facial feature detection using conditional regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Varnosfaderani and Moallem, 2017] Varnosfaderani, A. A. and Moallem, P. (2017). Texture images segmentation by geometric active contour models based on color and gabor features. *International Journal of Tomography & Simulation*, 30(1):105–117.

[Viehmann, 2019] Viehmann, T. (2019). Implementation of batched sinkhorn iterations for entropy-regularized wasserstein loss. *arXiv preprint arXiv:1907.01729*.

[Villani, 2008] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Vu, 2010] Vu, N.-S. (2010). *Towards unconstrained face recognition from one sample*. PhD thesis, Institut National Polytechnique de Grenoble-INPG.

[Wang et al., 2011] Wang, D., Chai, X., Zhang, H., Chang, H., Zeng, W., and Shan, S. (2011). A novel coarse-to-fine hair segmentation method. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 233–238. IEEE.

[Wang et al., 2009] Wang, D., Shan, S., Zeng, W., Zhang, H., and Chen, X. (2009). A novel two-tier bayesian based method for hair segmentation. In *International Conference on Image Processing (ICIP)*, pages 2401–2404. IEEE.

[Wang et al., 2013] Wang, D., Shan, S., Zhang, H., Zeng, W., and Chen, X. (2013). Isomorphic manifold inference for hair segmentation. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE.

[Wang et al., 2017a] Wang, L., Yu, X., and Metaxas, D. (2017a). A coupled encoder-decoder network for joint face detection and landmark localization. In *International Conference on Automatic Face and Gesture Recognition (FG)*.

[Wang et al., 2010] Wang, N., Ai, H., and Lao, S. (2010). A compositional exemplar-based model for hair segmentation. In *Asian Conference on Computer Vision (ACCV)*, pages 171–184. Springer.

[Wang et al., 2012] Wang, N., Ai, H., and Tang, F. (2012). What are good parts for hair shape modeling? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 662–669. IEEE.

[Wang et al., 2017b] Wang, N., Gao, X., Tao, D., Yang, H., and Li, X. (2017b). Facial feature point detection: A comprehensive survey. *Neurocomputing*.

[Wang and Fu, 2016] Wang, S. and Fu, Y. (2016). Face behind makeup. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[Wang et al., 2019a] Wang, X., Bo, L., and Fuxin, L. (2019a). Adaptive wing loss for robust face alignment via heatmap regression. In *IEEE International Conference on Computer Vision (ICCV)*.

[Wang et al., 2019b] Wang, Y., Luo, B., Shen, J., and Pantic, M. (2019b). Face mask extraction in video sequence. *International Journal of Computer Vision (IJCV)*, 127(6-7):625–641.

[Wang et al., 2017c] Wang, Z., Liu, S., Hu, W., Tong, R., Yang, X., and Zhang, J. J. (2017c). Face alignment refinement via exploiting low-rank property and temporal stability. In *International Conference on Computer Animation and Social Agents (CASA)*, volume 17.

[Wei et al., 2016] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732.

[Wei et al., 2019] Wei, Z., Liu, S., Sun, Y., and Ling, H. (2019). Accurate facial image parsing at real-time speed. *IEEE Transactions on Image Processing (TIP)*, 28(9):4659–4670.

[Wei et al., 2017] Wei, Z., Sun, Y., Wang, J., Lai, H., and Liu, S. (2017). Learning adaptive receptive fields for deep image parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR*.

[Wilson et al., 2017] Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. (2017). The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4148–4158.

[Wu et al., 2018a] Wu, H., Zheng, S., Zhang, J., and Huang, K. (2018a). Fast end-to-end trainable guided filter. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1838–1847.

[Wu et al., 2015] Wu, Q., Gan, Y., Lin, B., Zhang, Q., and Chang, H. (2015). An active contour model based on fused texture features for image segmentation. *Neurocomputing*, 151:1133–1141.

[Wu et al., 2018b] Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., and Zhou, Q. (2018b). Look at boundary: A boundary-aware face alignment algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Wu and Yang, 2017] Wu, W. and Yang, S. (2017). Leveraging intra and inter-dataset variations for robust face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition Worshop (CVPRW)*.

[Wu et al., 2017a] Wu, Y., Gou, C., and Ji, Q. (2017a). Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Wu et al., 2017b] Wu, Y., Hassner, T., Kim, K., Medioni, G., and Natarajan, P. (2017b). Facial landmark detection with tweaked convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

[Wu and He, 2018] Wu, Y. and He, K. (2018). Group normalization. In *European Conference on Computer Vision (ECCV)*, pages 3–19.

[Wu and Ji, 2015a] Wu, Y. and Ji, Q. (2015a). Discriminative deep face shape model for facial point detection. *International Journal of Computer Vision (IJCV)*, pages 37–53.

[Wu and Ji, 2015b] Wu, Y. and Ji, Q. (2015b). Robust facial landmark detection under significant head poses and occlusion. In *IEEE International Conference on Computer Vision (ICCV)*.

[Wu and Ji, 2019] Wu, Y. and Ji, Q. (2019). Facial landmark detection: A literature survey. *International Journal of Computer Vision (IJCV)*, 127(2):115–142.

[Xiao et al., 2018] Xiao, B., Wu, H., and Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*.

[Xiao, 2019] Xiao, J. (2019). Face alignment with an ensemble of gradient boosting trees. `https://github.com/JiaShun-Xiao/face-alignment-ert-2D`.

[Xiao et al., 2016] Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., and Kassim, A. (2016). Robust facial landmark detection via recurrent attentive-refinement networks. In *European Conference on Computer Vision (ECCV)*, pages 57–72. Springer.

[Xiao et al., 2015] Xiao, S., Yan, S., and Kassim, A. A. (2015). Facial Landmark Detection via Progressive Initialization. *IEEE International Conference on Computer Vision Workshop (ICCVW)*.

[Xiong and De la Torre, 2013] Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539.

[Xu et al., 2013] Xu, L., Du, Y., and Zhang, Y. (2013). An automatic framework for example-based virtual makeup. In *IEEE International Conference on Image Processing (ICIP)*, pages 3206–3210. IEEE.

[Xu et al., 2017] Xu, N., Price, B., Cohen, S., and Huang, T. (2017). Deep image matting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 311–320. IEEE.

[Yacoob and Davis, 2006] Yacoob, Y. and Davis, L. S. (2006). Detection and analysis of hair. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

[Yamaguchi et al., 2012] Yamaguchi, K., Kiapour, M. H., Ortiz, L. E., and Berg, T. L. (2012). Parsing clothing in fashion photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3577. IEEE.

[Yamashita et al., 2015] Yamashita, T., Nakamura, T., Fukui, H., Yamauchi, Y., and Fujiyoshi, H. (2015). Cost-alleviative learning for deep convolutional neural network-based facial part labeling. *IPSJ Transactions on Computer Vision and Applications*, 7:99–103.

[Yan et al., 2018] Yan, Y., Naturel, X., Chateau, T., Duffner, S., Garcia, C., and Blanc, C. (2018). A survey of deep facial landmark detection. In *Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP)*.

[Yang et al., 2015a] Yang, H., He, X., Jia, X., and Patras, I. (2015a). Robust face alignment under occlusion via regional predictive power estimation. *IEEE Transactions on Image Processing (TIP)*, 24(8).

[Yang et al., 2015b] Yang, H., Jia, X., Loy, C. C., and Robinson, P. (2015b). An empirical study of recent face alignment methods. *arXiv preprint arXiv:1511.05049*.

[Yang et al., 2015c] Yang, J., Deng, J., Zhang, K., and Liu, Q. (2015c). Facial Shape Tracking via Spatio-Temporal Cascade Shape Regression. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 994–1002.

[Yang et al., 2017] Yang, J., Liu, Q., and Zhang, K. (2017). Stacked hourglass network for robust facial landmark localisation. In *IEEE Conference on Computer Vision and Pattern Recognition Worshop (CVPRW)*.

[Yang et al., 2016] Yang, S., Luo, P., Loy, C.-C., and Tang, X. (2016). Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533.

[Yu et al., 2018] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision (ECCV)*.

[Yu et al., 2014] Yu, X., Lin, Z., Brandt, J., and Metaxas, D. N. (2014). Consensus of regression for occlusion-robust facial feature localization. In *European Conference on Computer Vision (ECCV)*.

[Yuan et al., 2015] Yuan, J., Wang, D., and Cheriyadat, A. M. (2015). Factorization-based texture segmentation. *IEEE Transactions on Image Processing (TIP)*, 24(11):3488–3497.

[Yue et al., 2018] Yue, L., Miao, X., Wang, P., Zhang, B., Zhen, X., and Cao, X. (2018). Attentional alignment networks. In *The British Machine Vision Conference (BMVC)*, page 208.

[Zadeh et al., 2016] Zadeh, A., Baltrušaitis, T., and Morency, L.-P. (2016). Deep constrained local models for facial landmark detection. *arXiv preprint arXiv:1611.08657*, 3(5):6.

[Zafeiriou et al., 2017] Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., and Shen, J. (2017). The menpo facial landmark localisation challenge: A step towards the solution. In *IEEE Conference on Computer Vision and Pattern Recognition Worshop (CVPRW)*, volume 1, page 2.

[Zamir et al., 2018] Zamir, A. R., Sax, A., Shen, W., Guibas, L., Malik, J., and Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zeng et al., 2015] Zeng, A., Boddeti, V. N., Kitani, K. M., and Kanade, T. (2015). Face alignment refinement. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 162–169.

[Zhang et al., 2018] Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., and Agrawal, A. (2018). Context encoding for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zhang et al., 2017] Zhang, H., Xue, J., and Dana, K. (2017). Deep ten: Texture encoding network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2896–2905. IEEE.

[Zhang et al., 2015] Zhang, J., Kan, M., Shan, S., and Chen, X. (2015). Leveraging datasets with varying annotations for face alignment via deep regression network. In *IEEE International Conference on Computer Vision (ICCV)*.

[Zhang et al., 2016a] Zhang, J., Kan, M., Shan, S., and Chen, X. (2016a). Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zhang et al., 2014a] Zhang, J., Shan, S., Kan, M., and Chen, X. (2014a). Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision (ECCV)*, pages 1–16. Springer.

[Zhang et al., 2016b] Zhang, S., Yang, H., and Yin, Z.-P. (2016b). Transferred deep convolutional neural network features for extensive facial landmark localization. *IEEE Signal Processing Letters*, 23(4):478–482.

[Zhang et al., 2014b] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014b). Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision (ECCV)*, pages 94–108. Springer.

[Zhang et al., 2016c] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2016c). Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(5):918–930.

[Zhao et al., 2018] Zhao, H., Qi, X., Shen, X., Shi, J., and Jia, J. (2018). Icnet for real-time semantic segmentation on high-resolution images. In *European Conference on Computer Vision (ECCV)*.

[Zhao et al., 2017] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890.

[Zhao et al., 2013] Zhao, X., Shan, S., Chai, X., and Chen, X. (2013). Cascaded shape space pruning for robust facial landmark detection. In *IEEE International Conference on Computer Vision (ICCV)*.

[Zhou et al., 2017] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2017). Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zhou et al., 2013a] Zhou, E., Fan, H., Cao, Z., Jiang, Y., and Yin, Q. (2013a). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 386–391.

[Zhou et al., 2013b] Zhou, F., Brandt, J., and Lin, Z. (2013b). Exemplar-based graph matching for robust facial landmark localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zhou et al., 2015] Zhou, Y., Hu, X., and Zhang, B. (2015). Interlinked convolutional neural networks for face parsing. In *Advances in Neural Networks (ISNN)*. Springer.

[Zhu et al., 2019a] Zhu, M., Shi, D., Zheng, M., and Sadiq, M. (2019a). Robust facial landmark detection via occlusion-adaptive deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3486–3496.

[Zhu et al., 2015] Zhu, S., Li, C., Change Loy, C., and Tang, X. (2015). Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006.

[Zhu et al., 2014] Zhu, S., Li, C., Loy, C. C., and Tang, X. (2014). Transferring landmark annotations for cross-dataset face alignment. *European Conference on Computer Vision (ECCV)*.

[Zhu et al., 2016a] Zhu, S., Li, C., Loy, C.-C., and Tang, X. (2016a). Unconstrained face alignment via cascaded compositional learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3409–3417.

[Zhu et al., 2016b] Zhu, X., Lei, Z., Liu, X., Shi, H., and Li, S. Z. (2016b). Face alignment across large poses: A 3d solution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155.

[Zhu and Ramanan, 2012] Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886.

[Zhu et al., 2019b] Zhu, Y., Sapra, K., Reda, F. A., Shih, K. J., Newsam, S., Tao, A., and Catanzaro, B. (2019b). Improving semantic segmentation via video propagation and label relaxation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8856–8865.

[Zou et al., 2019] Zou, X., Zhong, S., Yan, L., Zhao, X., Zhou, J., and Wu, Y. (2019). Learning robust facial landmark detection via hierarchical structured ensemble. In *IEEE International Conference on Computer Vision (ICCV)*, pages 141–150.