Robust and Reliable Facial Landmark Detection under Challenging Conditions

by **Jiahao Xia**

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Min Xu

School of Electrical and Data Engineering
Faculty of Engineering and IT
University of Technology Sydney
November 27, 2024

Certificate of Authorship / Originality

I, Jiahao Xia, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: November 27, 2024

Abstract

Facial landmark detection is crucial technology for many downstream tasks, such as talking head, face editing, facial emotion recognition and face recognition. Despite the recent progress brought by deep learning, there are still many challenges in this field that render the robustness of corresponding algorithms fragile in real-world scenarios. In this research, we primarily focus on improving the robustness for facial landmark detection algorithms from three different aspects. We first improve the robustness and reliability of the lightweight facial landmark detection model through the facial boundaries contained in low-level features. This approach ensures that facial landmark detection maintains competitive performance on platforms with limited computational ability. Additionally, by sharing features and employing a unique training strategy, the proposed method also demonstrates superior accuracy, even with limited parameters, in other face-related tasks, such as head pose estimation and face tracking. Then, we enhance the fragile robustness of facial landmark detection under heavy occlusion through inherent relation learning and uncertainty estimation. By learning a case-dependent inherent relation between landmarks, the proposed method can localize the occluded landmarks accurately based on the regular face shape and visible landmarks. Furthermore, we have evolved the method into a coarse-to-fine framework, which starts from a statistical mean shape to target shapes with multi stages. It also estimates the uncertainty for each landmark at each stage and adjusts the receptive field for the subsequent stage. The coarse-to-fine framework, adaptive inherent relation and dynamic receptive field yields highly competitive performance on extreme occlusion conditions. Finally, we achieve zero-shot facial landmark detection for the first time through a novel paradigm, significantly improving the robustness to locate landmarks that were unseen during training. Unlike previous works that set each landmark as an independent regression target, our approach utilizes labeled landmarks as anchors to learn a mapping from a plane to human faces. With the learned mapping, our method can localize any landmark, even those unseen in the training dataset. Additionally, because the paradigm unifies the learning targets of different facial landmark datasets, we can utilize multiple datasets with varying annotation formats to develop a unified large-scale model, which significantly enhances the robustness in various challenging conditions. Extensive experiments have been carried out, and the results show that our proposed methods significantly enhance the robustness and reliability of facial landmark detection under such challenging conditions.

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Min Xu, and Dr. Haimin Zhang, who have given me substantial support and allowed me sufficient space to reconsider the technical and mathematical details of existing machine learning and computer vision problems. I would also like to thank Prof. Jianguo Zhang and Dr. Wenjian Huang. Their guidance and encouragement played a crucial role throughout my entire Ph.D. studies.

I would like to acknowledge the University of Technology Sydney and the China Scholarship Council for their financial and technical support of my research. With these supports, I have been able to focus on my research without being overly concerned about my living expenses. As a result, I have been able to translate my ideas into practical applications in a short time.

Finally, I would like to express my appreciation to my wife, Yike Wu, my collaborators and my family. The Ph.D. period is a genuinely tough time for most people. However, their support and understanding have transformed this time into the happiest period of my entire study journey.

Jiahao Xia
November 27, 2024
Sydney, Australia

List of Publication

The contents of this thesis are based on the following papers that have been published or accepted, or preprints that have been under submission or submitted to peer-reviewed journals.

Publications

- Jiahao Xia, Haimin Zhang, Shiping Wen, Shuo Yang and Min Xu, “An efficient multi-task neural network for face alignment, head pose estimation and face tracking,” *Expert Systems with Applications*, vol. 205, p. 117 368, 2022, issn: 0957-4174
- Jiahao Xia, Weiwei Qu, Wenjian Huang, Jianguo Zhang, Xi Wang and Min Xu, “Sparse local patch transformer for robust face alignment and landmarks inherent relation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 4052–4061
- Jiahao Xia, Min Xu, Haimin Zhang, Jianguo Zhang, Wenjian Huang, Hu Cao and Shiping Wen, “Robust face alignment via inherent relation learning and uncertainty estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 358–10 375, 2023. doi: 10.1109/TPAMI.2023.3260926
- Jiahao Xia, Min Xu, Wenjian Huang, Jianguo Zhang, Haimin Zhang and Chunxia Xiao, “Task-agnostic unified face alignment via face structure prompts and semantic alignment,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, under review

Others

- Jiahao Xia, Wenjian Huang, Min Xu, Jianguo Zhang, Ziyu Sheng and Dong Xu, “Unsupervised part discovery via dual representation alignment,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 10597-10613, 2024. doi: 10.1109/T-PAMI.2024.3445582

- Haimin Zhang, Jiahao Xia, Guoqiang Zhang and Min Xu, “Learning graph representations through learning and propagating edge features,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022. doi: 10.1109/TNNLS.2022.3228102

- Wenjian Huang, Hao Wang, Jiahao Xia, Chengyan Wang and Jianguo Zhang, “Density-driven regularization for out-of-distribution detection,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 887–900

- Hu Cao, Guang Chen, Jiahao Xia, Genghang Zhuang and Alois Knoll, “Fusion-based feature attention gate component for vehicle detection based on event camera,” *IEEE Sensors Journal*, vol. 21, no. 21, pp. 24540–24548, 2021. doi: 10.1109/JSEN.2021.3115016

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Problem Statement

Facial landmark detection aims at predicting a group of pre-defined landmarks on face. This is the fundamental technique for many face related applications, such as talking head [1], face swapping [2], facial emotion recognition [3], [4], face recognition [5] and driver fatigue and distraction detection [6]. The robustness and reliability of facial landmark detection algorithms have significant influence to performance of these downstream tasks. These downstream applications are wildly used in daily life, having close tide with human safety and huge commercial values.

Although facial landmark detection has been investigated for over 20 years, many challenges remain in this field, with robustness and reliability being the most significant, especially in cases involving various challenging conditions. We have listed three main types of challenging conditions that are inevitable in practical applications:

1. **Platform with limited computational capability**: the mobile platform is the most widely used platform for computer vision applications. Considering manufacturing costs, existing mobile devices have very limited computational capabilities, which restricts the expressive ability of the network. Moreover, the working environment of these mobile devices is often outdoors with strong illumination variance. Therefore, improving the efficiency of the network while maintaining robustness using limited parameters is an urgent challenge that needs to be addressed.

2. **Faces with profile view and heavy occlusion**: the self-occlusion brought by profile view and the heavy external occlusion are inevitable in practical applications. For example, to identify driver distraction, facial landmarks must be accurately detected when the driver has a profile view. As a result, the driver head pose, which is an important clue for driver distraction detection, can be calculated based on the geometric relation between facial landmarks. Moreover, during the Covid-19 pandemic, wearing masks has also lead to significant performance degradation in many face-related applications, such as the Face ID of iPhone. Nowadays, researchers still keep seeking for a robust and reliable method for facial landmark detection with profile view and heavy occlusion.

3. **Landmarks not seen during training**: existing landmark detection method can only locate a group of predefined facial landmarks. For landmarks that were not seen during training, these methods often exhibit fragile robustness because they have different semantics compared to the landmarks used in training. However, re-labeling the training set is expensive and time-consuming. Therefore, transferring the knowledge to locate unseen landmarks with strong robustness and reliability is another challenge in the area of landmark detection.

Robustness and reliability remain challenges that continue to be investigated in the field of landmark detection. In the early stage, facial landmark detection is mainly based on PCA-based shape models, which can be further divided into three categories: Active Appearance Model (AAM), Active Shape Model (ASM) and Constrained Local Model (CLM). However, the limited expressive ability of features extracted by these PCA-based methods leads to that they are only effective in controlled scenarios with a frontal view. To implement facial landmark detection in the wild, cascaded shape regression (CSR) methods [7]–[16] are proposed. With carefully designed feature descriptors and multi-stage convergence, CSR methods successfully boost the performance in the wild. However, as for the cases in the wild with illumination variance, occlusion and profile view, these feature descriptors are still insufficient to describe their positions.

The recent progress of convolutional neural networks, which can be further divided into two categories: the coordinate regression method and the heatmap regression method, enables the model to employ more parameters for feature extraction. This significantly improves the expressive ability of the learned features, which promises competitive performance on the most conditions in the wild. Nevertheless, more parameters means larger computational complexity, resulting in worse real-time capability. While many existing applications, such as driver distraction monitoring and face recognition, are based on the mobile devices with very limited computational capability. Hence, there is a pressing need to address the challenge of developing lightweight models for these devices, all while ensuring their robustness and reliability, which is an extensively researched problem in the field of facial landmark detection and other face-related tasks.

Despite the huge progress brought by CNN, existing facial landmark detection models still suffer from significant performance degradation for the cases with heavy occlusion. Human tends to locate occluded landmarks through their relative position to the easily identified landmarks because human face has a regular shape. Unfortunately, existing methods fails to explicitly employ the clue for landmark localization. As for heatmap regression methods, their convolutional neural network kernels make them focus locally. As a result, they usually fail to capture the relations of landmarks farther away in a global manner. The coordinate regression methods directly flatten the feature map and regress the coordinates of the target landmarks via fully connected (FC) layers. However, the flattening process destroys the spatial information of the image, and the frozen weights of the FC layer in the testing phase prevent the model from learning case-dependent landmark relations. As a result, these coordinate regression methods cannot mimic human abilities in employing easily identifiable landmarks for the localization of occluded landmarks. For this reason, the robustness and reliability of existing facial landmark detection methods are compromised in cases with heavy occlusion.

Another problem that limits the performance on occluded samples is that existing methods are based on a simplified assumption: the variances of all facial landmark distributions are a constant value. Based on this assumption, heatmap regression methods generate heatmap with a fixed variance as the learning target during training; coordinate regression methods ignore the variance, only predicting the mean of distribution and constrainting the learning with L1 or L2 loss; patch-based regression methods [17]–[19] set the local patch of each landmark to a fixed size. However, through observation, we find that the easily identified landmark results in a smaller variance and the landmarks with high uncertainty always have a larger variance. Therefore, the assumption does not usually hold and using the patch with a fixed size may lead to performance degradation to face alignment. Unfortunately, the existing patch-based regression methods have not solved the problem yet.

Moreover, existing facial landmark detection models can only localize a group of pre-defined landmarks. Hence, other models must employ these pre-defined landmarks for downstream tasks, even though these landmarks are not always optimal for such tasks, thereby limiting their performance. To locate a group of unseen landmarks with different semantics compared to the landmarks in training, they must re-annotate the training samples and retrain the model. Nevertheless, re-annotating a large-scale facial landmark dataset is quite expensive and time-consuming. To address this problem, many methods [20], [21] transfer a model to a group of new landmarks with several annotated samples based on few-shot learning. Although these few-shot methods enables the models to achieve satisfactory performance for the easy cases. However, their robustness and reliability still significantly degrade in cases with heavy occlusion or a profile view, due to catastrophic forgetting. The main reason is that existing methods set each facial landmark as an independent regression target. As a result, it is difficult to transfer knowledge learned from a facial landmark dataset to a new set of landmarks with different semantics, even though the input images still depict human faces. Unifying the learning targets of different facial landmark datasets is a potential way to boost few-shot landmark detection. This approach also enables the model to utilize datasets with various annotation forms for training. Unfortunately, there is no existing research addressing this specific problem.

To completely eliminate the reliance on hand-craft annotation, unsupervised facial landmark detection [22]–[26] and self-supervised facial landmark detection [27] are wildly investigated in recent years. However, the semantics of the discovered landmarks cannot be controlled by humans during the training. The unpredictability makes it impossible to apply these methods to detect specific landmarks. In other fields, such as image classification [28] and object detection [29], zero-shot learning successfully bridges the gap between seen and unseen categories with the prompt of semantic information (ie. word vectors). As a result, the semantics of the detected categories become predictable. However, regarding facial landmark detection, no existing prompt can describe the semantics of different facial landmarks. Therefore, although zero-shot learning shows greater potential in eliminating reliance on annotations for facial landmark detection, research on this topic is still lacking.

1.2 Research Objectives

As stated in the research problem in the above section, the specific research objectives of this thesis can be summarized as follows:

1. to propose an efficient framework and training strategy for the mobile platform, enabling the model to retain strong robustness and reliability for facial landmark detection and other face-related tasks with very limited parameters and computational complexity.

2. to study the inherent relation between facial landmarks and propose an approach to encourage the model to explicitly learn a case-dependent inherent relation. The learned inherent relation is expected to drive the trained model to act as human, utilizing easily identified landmarks to localize those landmarks with heavy occlusion.

3. to address the limitation of existing patch-based methods, which assume a fixed patch size based on a simplified assumption that the variances of all landmark distributions are constant. The proposed method is expected to enable the trained model to obtain a dynamic patch size and receptive field that can adjust according to the landmark uncertainty.

4. to eliminate the negative influence of catastrophic forgetting in few-shot facial landmark detection and improve the robustness and reliability of existing few-shot landmark detection method. The optimized method is expected to effectively transfer a trained model to a set of unseen landmarks in the setting of few-shot learning.

5. to develop a zero-shot facial landmark detection method capable of localizing unseen landmarks with specific semantics without any re-training. Additionally, this method is expected to have competitive robustness and reliability in the zero-shot learning setting.

1.3 Contributions of the Research

The contributions of this thesis to the existing knowledge framework of facial landmark detection can be described as follows:

1. We propose a novel efficient training strategy and a lightweight facial landmark detection framework for mobile devices. By introducing the facial boundary information contained in low-level feature map into the prediction head through a shotcut, the robustness and reliability of facial landmark detection is boosted significantly with only few parameters

increasing. Moreover, we find that the feature for facial landmark detection can also improve the performance for other face-related tasks, such as head pose estimation and face tracking. Therefore, we propose a novel training strategy to further evolve the framework into a multi-task framework for facial landmark detection, head pose estimation and face tracking. Extensive experiments are carried out on several benchmarks and the results demonstrate the effectiveness of our proposed methods.

2. To encourage model to learn a case-dependent inherent relation for better performance on occlusion conditions, we propose a novel framework for facial landmark detection based on self and cross attention mechanism, named sparse local patch transformer (SLPT). SLPT explicitly generates representations for each facial landmark by a local patch cropped from the feature map. A series of learnable vectors, which is called landmark queries, aggregate these representations based on the attention mechanism. As a result, SLPT can model a case-dependent relation between facial landmarks to retain robustness and reliability for the cases with heavy occlusion. To further improve the performance of SLPT, we develop it to a coarse-to-fine framework, starting from a statistical mean shape to target shapes step by step. Extensive experiments and visualized attention maps illustrates that SLPT can maintain better robustness on extreme conditions compared to other state-of-the-art methods by learning a case-dependent inherent relation.

3. We propose the dynamic local patch and incorporate this algorithm with SLPT to evolve SLPT into a new model, dynamic sparse local patch transformer (DSLPT). Compared to previous facial landmark detection methods, DSLPT predicts the uncertainty for each landmark. Based on the predicted uncertainty, DSLPT can adaptively adjust the size of each local patch, which enables DSLPT to obtain a dynamic receptive field. With this, DSLPT can apply a larger patch size for the landmark with uncertainty, encouraging model to utilize more contexture information for landmark localization. For the landmark with low uncertainty, the smaller patch size promises higher feature resolution for higher accuracy. The experimental results show that DSLPT further improves the robustness and reliability without increasing computational completely.

4. We propose a task-agnostic unified face alignment (TUFA) framework to improve the performance of few-shot facial landmark detection and achieve zero-shot facial landmark detection for the first time. TUFA unifies the learning targets of multi datasets by learn-

ing a mapping between an interpretable plane to human faces, unlike previous methods that treats each facial landmark as an independent regression target. The unification of learning targets on different datasets can effectively address the catastrophic forgetting in few-shot facial landmark detection. Moreover, we present a novel encoding method to represent any coordinate on the interpretable plane as a high dimensional vector. This can serve as a kind of prompt to bridge the gap between seen and unseen facial landmarks in zero-shot facial landmark detection. As a result, TUFA successfully achieves zero-shot facial landmark detection for the first time. Experimental results show that TUFA outperforms previous few-shot methods with a large margin and the landmarks detected by TUFA in the zero-shot setting have better semantic consistency than the landmarks detected by previous unsupervised or self-supervised methods.

1.4 Thesis Outline

This thesis consists of seven chapters and their corresponding short descriptions are discussed as follows:

- In Chapter 1, we present the research background and discuss the limitations that have not been addressed in the field of facial landmark detection. Moreover, we also introduce the contributions and provide an outline of this thesis.
- Chapter 2 provides a comprehensive literature review for existing research about various facial landmark detection methods (PCA-based shape model, cascade shape regression methods, CNN-base methods) under different setting (many-shot learning, few-shot learning and zero-shot learning).
- Chapter 3 presents a efficient way that boosts the performance of facial landmark detection by introducing the facial boundary information contained in low-level feature map into prediction head. Besides, this chapter also introduces a novel strategy for training a multi task framework for facial landmark detection, head pose estimation and face tracking.

  This chapter resulted in the following publications: Jiahao Xia, Haimin Zhang, Shiping Wen, Shuo Yang and Min Xu, “An efficient multitask neural network for face alignment, head pose estimation and face tracking,” *Expert Systems with Applications*, vol. 205, p.

117 368, 2022, issn: 0957-4174

- Chatper 4 presents a novel facial landmark detection method based on attention mechanism to encourage model to learn a case-dependent inherent relation and a coarse-to-fine framework that enables a statistical mean face shape to converge to the target face shape step by step.

  This chapter resulted in the following publications: Jiahao Xia, Weiwei Qu, Wenjian Huang, Jianguo Zhang, Xi Wang and Min Xu, “Sparse local patch transformer for robust face alignment and landmarks inherent relation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 4052–4061

- Chatper 5 present a module that predicts uncertainty for each landmark and adjusts the patch size for patch-based methods to enable the model to obtain adaptive receptive. The module is further combined with SLPT to achieve better robustness and reliability.

  This chapter resulted in the following publications: Jiahao Xia, Min Xu, Haimin Zhang, Jianguo Zhang, Wenjian Huang, Hu Cao and Shiping Wen, “Robust face alignment via inherent relation learning and uncertainty estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 358–10 375, 2023. doi: 10.1109/TPAMI.2023.3260926

- Chapter 6 presents a task-agnostic unified facial landmark detection framework that successfully unifies the learning targets on different datasets for better performance on many-shot and few-shot facial landmark detection. Moreover, this chapter also introduces a structure prompt that bridges the gap between seen and unseen landmark and achieves zero-shot face landmark detection for the first time.

  This chapter resulted in the following publications: Jiahao Xia, Min Xu, Wenjian Huang, Jianguo Zhang, Haimin Zhang and Chunxia Xiao, “Task-agnostic unified face alignment via face structure prompts and semantic alignment,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, under review

- Chapter 7 concludes this thesis and suggests potential directions for future work in this field.

Chapter 2

Literature Review

To localize a group of pre-defined landmarks, existing training methods for facial landmark detection can be divided into three categories: many-shot learning, few-shot learning and zero-shot learning. The schematic diagram of them are shown in the Fig. 2.1. Many-shot learning is the most wildly used training method in facial landmark detection. It directly trains the model on a dataset annotated with these pre-defined landmarks. Most heavy weight and light weight facial landmark detection models in industry application are trained using many-shot learning. However, many-shot learning usually requires a large-scale dataset to learn a robust and reliable result, which significantly increases the cost of data labeling. Few-shot learning first pre-trains the model on a related task, whose knowledge can be shared with localizing target landmarks. Then, the model is fine-tuned on a small-scale dataset for learning to localize target landmarks. Zero-shot learning pre-trains a model on a very large-scale dataset for a related task, and then bridges the gap between the pre-trained task and target task with the semantic information. Therefore, zero-shot learning enables the model to localize target landmarks without re-training.

2.1 Many-shot Facial Landmark Detection

2.1.1 PCA-based method

PCA-based method dominants facial landmark detection at the very early stage. The core idea of PCA-based methods is to represent the variations of face shape using a certain number of parameters through Principal Component Analysis (PCA), and further aims to identify the most suitable parameters for describing the target face shape, based on image features.

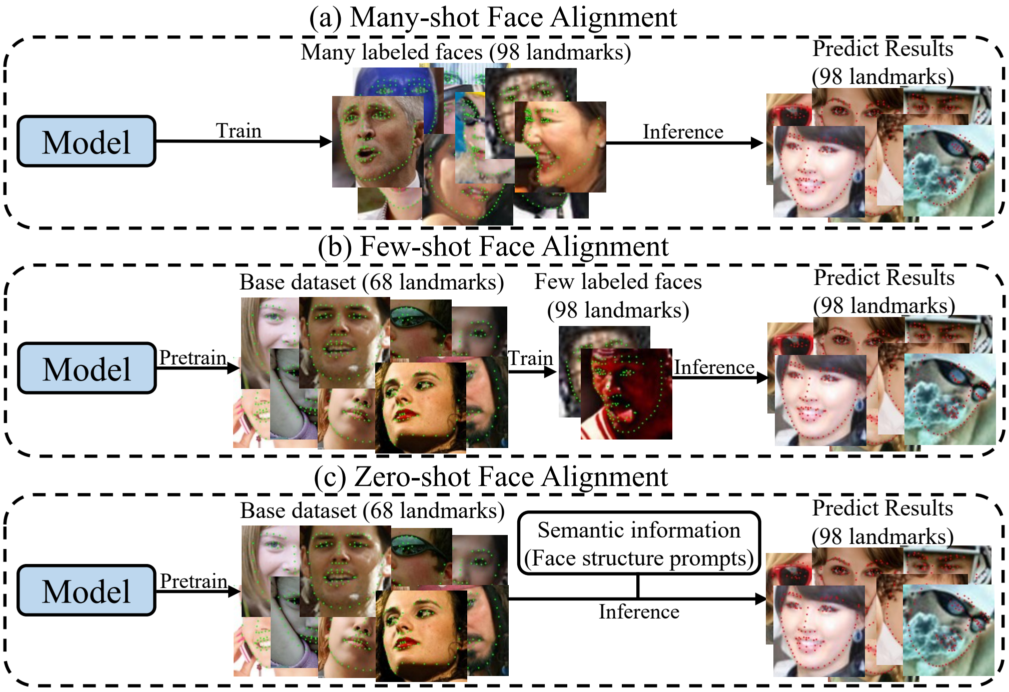


Figure 2.1: A schematic diagram of many-shot learning, few-shot learning and zero-shot learning in facial landmark detection.

This is why facial landmark detection was referred to as face alignment during that period. The PCA-based methods can be further classified into three categories: active shape model (ASM) [30], active appearance model (AAM) [31], [32] and constrained local model (CLM) [33]. ASM [30] proposes the PDM (point distribution model) to constraint face shape and further alignment the statistical shape to target face through facial boundaries. However, the reliance on boundary information leads to a very fragile robustness, even for the cases on controlled scenarios. Timothy et al. [31] further improves the performance of ASM by combining the shape variation with the texture variation to generate the statistical appearance model, which is referred as activate appearance model (AAM). Liu et al. [32] boost the generalization ability of AAM by introducing boosting-based classifier as an appearance model. To achieve better robustness in controlled scenarios, David et al. [33] propose the Constrained Local Model (CLM), which employs appearance model to generate likely feature templates, rather than approximating the image pixels. Although a lot of efforts are taken, PCA-based methods still degrades significantly in the wild. The main reason is that the variation of illumination leads to the variation of texture information around facial boundaries. The heavy reliance on texture of PCA-based methods make themselves not robust in the wild.

2.1.2 Cascade Shape Regression Method

Cascade shape regression (CSR) methods simplify the training process of facial landmark detection, enabling a statistical mean shape to converge to the target face shape using multi stages. Compared to PCA-based methods, CSR methods demonstrates much better robustness and reliability in the wild by using carefully designed feature descriotors (eg. histogram of oriented gradients (HOG) [34] and Scale-invariant Feature Transform (SIFT) [35]). CSR methods can also be roughly classified into two categories: tree-based regression methods [7], [10], [13], [16] and subsequent cascaded regression methods [8], [9], [11], [12], [14], [15].

To minimize the L2 distance between annotated face shape and predicted face shape, Cao et al. [7] introduce a two-level boosted regression method that regresses target face shape from the effective shape indexed features. Kazemi et al. [10] combine gradient boosting algorithm with a prior probability to explore relevant features efficiently. As a result, they successfully achieve millisecond face alignment with competitive performance. Ren et al. [16] further boost the efficiency of facial landmark detection, introducing a locality principle for learning a set of binary features. With this framework, facial landmark detection runs over 300 frames per second on mobile phones. Lindner et al. [13] employ random forest regression-voting to optimize CLM models, resulting in a efficient and accurate facial landmark detection model. Overall, the introducing of tree-based regression effectively simplifies the process of facial landmark detection, leading to a significant improvement in efficiency.

The subsequent cascaded regression methods extract features for indexing face shape through carefully designed descriptors and regressing the optimal position for each landmark via optimization-based methods. Xiong et al. [9] utilize SIFT descriptor for feature extraction and regressing the landmark coordinates from the features through a novel $2nd$ order descend methods. Similar to [9], Tzimiropoulos et al. [14] also propose a novel optimization method for facial landmark detection. They calculate Jacobian and Hessian matrices for each sample, finding the optimal solution via a method inspired by Gauss-Newton optimization. Feng et al. [15] further use multi domain-specific regressors to detect landmarks and fuse their results via fuzzy functions. This design effectively improves the fault tolerance ability of CSR methods. Despite the increased computational complexity introduced by these optimization methods, the enhanced fitting ability of subsequent cascaded regression methods, compared to that of tree-based regression methods, typically ensures better accuracy. To eliminate the negative influence of occluded landmark, Burgos-Artizzu et al. [8] employ the occlusion label for feature selection. As a result, the learned model can retain better robustness and reliability to the face with heavy occlusion. Inspired by the fact that human faces have a regular shape, Zhu et al. [12] search for a face shape closed to target face from the training set as the initial face shape, instead of using the statistical mean shape. With this prior knowledge, the robustness and reliability are further improved. Moreover, Asthana et al. [11] further propose a person-specific facial landmark detection setting for CSR-based methods based on incremental learning, which further enlarges the application range of facial landmark detection.

Despite the significant improvements brought about by carefully designed descriptors and regression methods, the expressive ability remains insufficient for cases under extreme conditions, such as illumination variation, profile view and heavy occlusion. Unfortunately, these conditions are inevitable in practical applications, and the failures of landmark detection under these conditions lead to severe landmark jitter in videos. Therefore, the robustness and reliability of CSR-based models remains to be further improved for downstream applications.

2.1.3 CNN-based method

The backward propagation of convolutional neural networks (CNNs) enable themselves to learn to extract features through end-to-end training. Compared to other classic descriptors, CNNs can employ much more parameters to extract the features with better expressive ability. As a result, the development of CNNs significantly boosts the performance of most computer vision tasks, including facial landmark detection. The existing CNN based methods can be roughly divided into two categories: coordinate regression methods [19], [36]–[54] and heatmap regression method [17], [20], [55]–[71].

Coordinate regression method

Coordinate regression methods extract feature map and project the feature into fully-connected layers to directly regress the coordinate of each facial landmark. Inspired by CSR-based methods [39], [42], [46], diverse cascaded CNN frameworks are proposed for coarse-to-fine facial landmark detection. Lv et al. [39] utilize a global stage and a local stage for coarse-to-fine facial landmark regression; Kowalski et al. [42] generate an attention map from the predicted result in the last stage to introduce attention mechanism for a more fine-grained result; Liu et al. [46] view the cascaded facial landmark detection as a Markov decision process, which interacts with a trajectory of state transitions, actions and rewards. Nevertheless, these multi stage facial landmark detection frameworks usually consists of several subnetworks, which significantly increases the number of parameters. To improve the efficiency, recurrent networks, which shares parameters in each stage, are proposed for facial landmark detection. Trigeorgis et al. [36] design a convolutional recurrent neural network that can maintain an internal memory for coarse-to-fine facial landmark detection; Xiao et al. [37] introduces an attentive-refinement mechanism, which identifies a reliable landmark as the center of the attention at each recurrent stage for better robustness to occluded landmarks. Both coarse-to-fine and recurrent facial landmark detection methods directly regress the landmark coordinates from global feature map. Whereas, the low resolution of global feature map leads to face details missing and inaccuracy of landmark localization. To generate more fine-grained features for each landmark, Zadeh et al. [17], Liu et al. [46] and Zhu et al. [19] further develop novel patch-based regression models. They crop local patches for landmarks and further regress facial landmarks from them. Despite the high resolution brought by the local patches, the losing of global features can also lead to a higher failure rate on the testing set. Generating local patches with high resolution while still retaining global information remains a challenge in this field.

Without explicitly identifying those nearly out-of-distribution landmarks (the landmarks with less significant features brought by occlusion and illumination variation), these outliers always lead to overfitting on the training set. To address this problem, wing loss [44], adaptive wing loss [61] and Softwing [52] are proposed. By adaptively applying larger weight to the in-domain landmarks compared to outliers, the performance is improved obviously, especially for the cases on extreme conditions. However, these optimized function still cannot completely eliminate the negative influence brought by these outliers. Human tends to locate these nearly out-of-distribution landmarks by their relative position to the easily identified landmarks because human face has a regular shape. By utilizing the landmarks with significant features for facial landmark localization, the negative influence of these outliers can be eliminated. Therefore, recent works try to model the inherent relation between facial landmarks for better performance. Lin et al. [52] retain the face structure information by an adjacency matrix defined by prior knowledge. Li et al. [50] further improve the performance by a learnable adjacency matrix so that the network can explore a task-specific structure. However, these models still fails to act as human because their learned inherent relation cannot be case-dependent. Therefore, learning a case-dependent inherent relation between landmarks is still a challenging but quite meaningful in facial landmark detection.

Moreover, with stronger expressive ability, many works [40], [41], [49], [51], [58] successfully achieve 3D facial landmark detection with CNN. Zhu et al. [40], [41] represent the 3D face shape with a set of parameters via PCA as ASM methods, and then regress these parameters from the input image based on CNN. Guo et al. [49] further decrease the number of parameters to 62 and optimize the learning process by meta-learning. Bulat et al. [58] introduce a network that projects 2D annotated landmarks to 3D landmarks and generates a large-scale dataset for 3D facial landmark detection. Wu et al. [51] introduce a synergy process, using the relations between landmarks and PCA-based parameters for better robustness and reliability. Despite extensive research in this field, the self-occlusion of faces in monocular images still makes it very challenging for existing methods.

Heatmap regression method

Heatmap regression methods generate a heatmap with gaussian distribution from the ground truth for each facial landmark and set regressing these heatmaps as the learning target. In the testing process, heatmap regression methods consider the pixel with the highest intensity as the optimal output. Therefore, the output coordinate can only be an integer that leads to a quantization error since the resolution of heatmap is always lower than the input image. To eliminate the error, various backbones [55]–[58] are proposed to predict heatmaps with high resolution. Lan et al. [69] adopt an additional decimal heatmap for subpixel estimation; Zhang et al. [66] utilize another network for subpixel offset estimation; Chen et al. [63], Tai et al. [65] and Kumar et al. [67] further predict landmark probability distribution on the heatmap for subpixel coordinate.

As mentioned in ASM [30] and AAM [31], [32], facial boundary information serves as a crucial clue in facial landmark detection. To maximize the use of boundaries information, Wu et al [43], Wang et al. [61], Huang et al. [70] and Zhou et al. [72] set facial boundary heatmap regression as an additional regressing objective and further fuse boundary heatmaps with feature map to introduce an attention mechanism for learning the relation between neighboring landmarks. Zou et al. [64] further project the output heatmaps into a graph network and model the holistic and local structure by clustering. However, the learned relation is fixed to all cases. An ideal inherent relation should be case dependent but there is no work yet on this topic unfortunately. Hence, we propose a method to fill this gap.

Recently, the development of Vision Transformer (ViT) [73] breaks the record of many computation vision tasks, such as image classification [73], [74], [75], object detection [76], [77] and semantic segmentation [78]. Although ViT models also break the record of a very similar task, human pose estimation [79], [80], directly applying ViT in face alignment does not promise an improvement because training ViT requires a large number of training samples. Lan et al. [69] generates decimal heatmaps by ViT. Unfortunately, the ViT based model fails to outperform CNN based model. Therefore, it is quite essential to optimize ViT for facial landmark detection, enabling it to work well with insufficient number of training samples.

2.1.4 Efficient CNN-based method

Despite the success of existing CNN-based methods, they significantly increases the number of parameters and computational complexity. Considering the manufacturing cost, mobile platform, which is the most widely used platform for computer vision application, always has very limited computational ability. As a results, it is hard for them to run many existing state-of-the-art facial landmark detection models. Nevertheless, their working environment is always in the wild, which may have larger illumination variation and external occlusion. Therefore, many existing studies work toward to efficient lightweight facial landmark detection models for mobile devices. The most intuitive method to maintain robustness of lightweight model is to utilize efficient CNN modules. Yu et al. [81] utilize a conditional channel weighting unit to replace pointwise convolutionas; Li et al. [82] propose an efficient multi-scale contextual information extraction for landmark detection; Wen et al. [83] utilize network architecture search to discover an efficient facial landmark detection framework. Because the reuse of parameters in recurrent networks can significantly reduce memory usage, employing recurrent networks is a potential direction for developing efficient facial landmark detection models. Micaelli et al. [84] and Gil et al. [85] encourage the recurrent network to learn to stop when the accuracy of landmarks satisfies a certain standard, so that redundant computation can be eliminated. Without modifying the structure of network, the robustness and reliability of lightweight models can also be improved from a data-driven aspect. Zhang et al. [86] employ multi-task learning to train a facial landmark detection model, enabling the model to use the knowledge learned from other face related tasks to improve the robustness and reliability; Bjorn et al. [20] set face reconstruction as a pretext task for pretraining, allowing the performance of a lightweight network, ResNet18 [87], to be significantly boosted in facial landmark detection; Qian et al. [88] expand the number of training samples using a generative model and further improves the robustness and reliability of lightweight models. Although many efforts have been made to improve the efficiency of facial landmark detection models, there remains a significant gap in the robustness and reliability between lightweight and heavyweight models.

2.2 Few-shot Facial Landmark Detection

Few-shot learning aims to transfer a pre-trained model to a series of novel categories with only a few annotated samples, significantly reducing the cost of dataset annotation. Despite the success of few-shot learning in image classification [89], object detection [90] and semantic segmentation [91], it still cannot achieve competitive performance in facial landmark detection. The most intuitive solution for few-shot facial landmark detection is expanding the number of training samples. Therefore, Qian et al. [88] employ style transfer to fuse two face and further generate a set of synthetic faces with ground truth for training. Nevertheless, the synthetic faces generated by Generative adversarial network (GAN) has an obvious domain gap to real face, which leads to performance degradation in real scenarios. Zhang et al. [86] find that the features learned from other face related tasks can also help the learning of facial landmark detection. Therefore, Browatzki et al. [20] and He et al. [26] set image reconstruction as a pretext task. With the features learned from face reconstruction, these models can be transferred into facial landmark detection with a few annotated samples. Although the face reconstruction helps the model learn to localize facial landmarks, it has negative influence on the accuracy of landmark detection. The main reason is that face reconstruction focuses on the texture information whereas facial landmark focus more on geometric information of face. Therefore, we assume if the facial landmark detection can be encouraged to have a unified learning target among different datasets, the learned model can be transferred to a new dataset efficiently and effectively and the performance of few-shot facial landmark detection can be boosted significantly. Unfortunately, there is no existing research about this field yet.

2.3 Zero-shot Facial Landmark Detection

To eliminate the reliance on facial landmark annotations, various unsupervised landmark detection methods are proposed. Thewlis et al. [92] discover landmarks by introducing a novel equivalent constraint. However, with only an equivalent constraint, the model is easy to get into a trivial solution. To further improve the consistency of the discovered landmarks, Zhang et al. [23], Lorenz et al. [24] and Jin et al. [93] extract multiple descriptors from local feature and utilize these descriptor for image reconstruction. With setting the image reconstruction as the pretext task, the discovered landmarks have better semantic consistency compared to using a single equivalence constraint. As for unsupervised landmark discovery in video, the consecutive frames enables the model to utilize conditional generation for landmark discovery. Jakab et al. [25], Jakab et al. [94], Kim et al. [95] and Minderer et al. [96] reconstruct video frame using the discovered landmarks from the corresponding frame and the texture information from the other frame. He et al. [26] further introduce conditional generation to normal images for unsupervised landmark discovery. They set recovering the masked images as pretext task and employ the discovered landmarks as the clues. To discover more meaningful facial landmarks, Mallis et al [27] and Tourani [97] employ a keypoint detector [98] to discover potential facial landmarks and utilize the landmarks, which obtain consistent semantics across different faces, for self-supervised training. Although the semantics of the landmarks discovered by existing landmark detection methods across different images are highly consistent, they still cannot discover the landmarks with specific semantics. The semantics of the discovered landmarks are random, which makes it hard to apply them in downstream tasks.

Compared to unsupervised learning, zero-shot learning is a more practical way to eliminate the use of annotations because zero-shot leaning can detect the target with specific semantic by using semantic information as the prompt. Despite impressive progress in zero-shot classification [99], object detection [100], semantic segmentation [101] and action recognition [102]. Zhang et al. [103] align textual prompts with visual features to extend landmark detection model to unseen animal categories in zero-shot manner. However, textual prompts are difficult to describe the structure relations among landmarks, making it hard to extend the model to unseen landmarks. Therefore, zero-shot facial landmark detection for unseen landmarks has not been achieved yet. constructing a type of prompt to represent the semantics of facial landmarks and further using it to bridge the gap between the seen and unseen landmarks is quite challenging in zero-shot facial landmark detection.

Chapter 3

Robust and Reliable Facial Landmark Detection for Mobile Devices base on Multi-task Learning

3.1 Introduction

In this chapter, an efficient network, face alignment (facial landmark detection), tracking and pose estimation network (ATPN) is proposed for mobile devices. We find out that the shallow-layer features are highly correspond to facial boundaries and they contain the facial structural information(Fig. 3.1 $2th$ column). We give a shortcut to the features in shallow layers so that a light model can also explicitly employ the structural information for accurate face landmark detection. Then, a cheap heatmap (Fig. 3.1 $3rd$ column) is generated directly based on face landmark detection result and fused with the intermediate features. Based on the procedure, the ATPN can utilize both the appearance information of input image and the geometric information of the facial landmarks to achieve accurate and robust head pose estimation. It also provides the attention clues for face tracking task. Moreover, the face tracking task can save the face detection procedure to further accelerated facial landmark detection and head pose estimation in video-based processing.

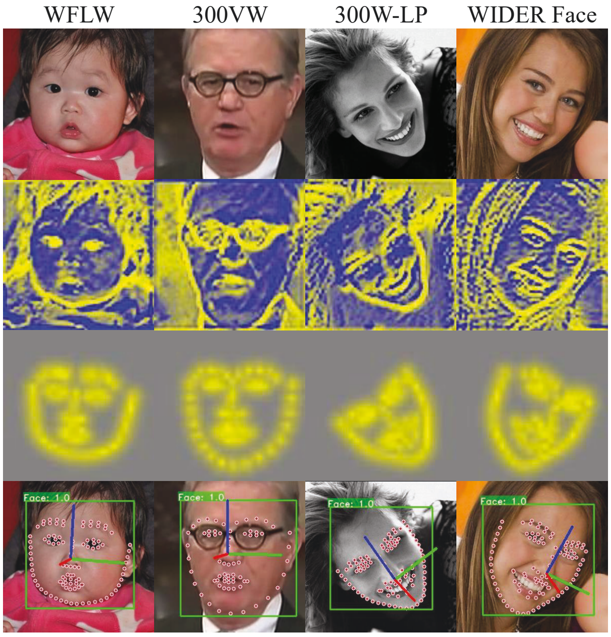


Figure 3.1: The input images (1^{st} row), low-level features (2^{nd} row), generated heat maps (3^{rd} row) and results (4^{th} row) on various benchmarks (WFLW, 300VW, 300W-LP and WIDER Face.)

In a nutshell, our main contributions of this chapter include:

- Proposing a light architecture that employs the structural information in low-level features for the face alignment. Compared to other light models, it achieves the best performance with the least parameters and lowest computational complexity.
- Providing the geometric information for head pose estimation and attention clues for face tracking by a heatmap generated from the face landmark detection results.
- Proposing a practical multitask framework for face alignment, head pose estimation and face tracking for video-based processing.
- Conducting extensive experiments and ablation studies on various datasets to prove the effectiveness of ATPN.

3.2 Method

In this section, we introduce an efficient multitask framework, ATPN, for face alignment (facial landmark detection), tracking and pose estimation. As illustrated on Fig. 3.2(a), the ATPN

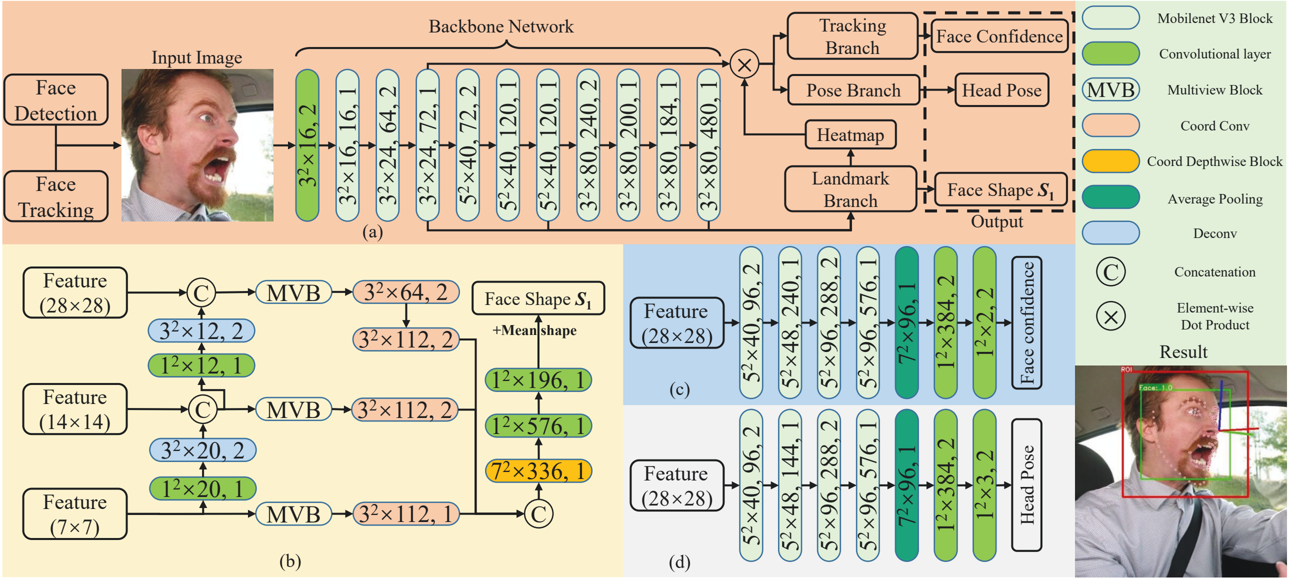


Figure 3.2: An overview of our multitask framework. (a) The overall framework. (b) The landmark branch. (c) The tracking branch. (d) The pose branch.

consists of four parts: the backbone network, landmark branch, pose branch and tracking branch. The input is the Region of Interest (ROI) of face that is acquired by a face detector or the minimum bounding rectangle of the face landmarks in previous frames (25% extension of each boundary). The backbone network firstly learns features from input image with Mobilenet-V3 block [104]. Then the landmark branch localize facial landmarks by the features at different levels (28×28, 14×14, 7×7). Based on the predicted facial landmarks, a heatmap is generated directly and fused with low-level features (28×28) to provide geometric information and attention clue for other tasks. Finally, the two branches regress the Euler angles and the confidence of the face in parallel. If the confidence is larger than a certain threshold (0.7 for ATPN), the ROI of next framework is calculated based on the predicted facial landmarks.

3.2.1 Landmark Branch

The structural information of the facial boundaries is crucial for facial alignment. Different from other works [43], [105] that utilize an additional network or branch to generate boundary heatmaps, we directly employ the structural information in low-level features, as shown in Fig. 3.2(b). Referred to feature pyramid network [106], we firstly fuse the low-level features with high-level features by 1×1 Conv and Deconv layers. Then, the features at different levels are fed into the multiview block [107]. As illustrated in Fig. 3.3, it consists of three CNN layers to

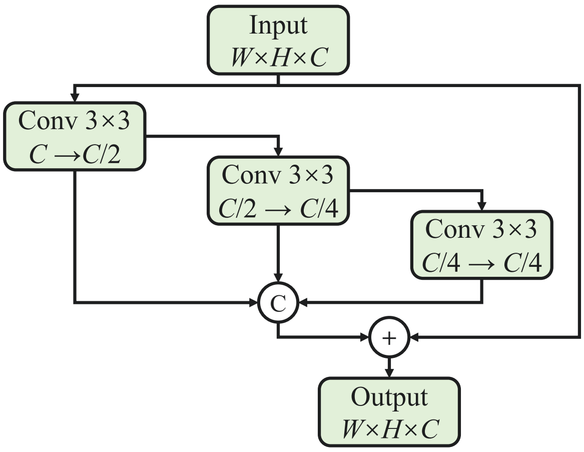


Figure 3.3: The structure of a multiview block.

create three different receptive fields. Besides, the residual connection of the multiview block also preserves the low-level features. The outputs of the multiview block are downsampled directly by CoordConv [108]. On the one hand, it shorten the information path between low-level featuress and output layers. On the other hand, it introduces a coordinate conception into CNN, which is significant to coordinate regression. Finally, we utilize a 7×7 Coord Depthwise Block to project the feature map into a vector. Compared to pooling or linear layers, the Coord Depthwise Block can preserve the spatial structure of the input image. Instead of predicting the facial landmarks directly, the landmark branch predicts the residual error between target face shape $\boldsymbol{S}_1$ and mean shape $\boldsymbol{S}_0$. The mean shape $\boldsymbol{S}_0$ provides a good initial face shape for the framework to make the result more stable.

3.2.2 Pose Branch and Tracking Branch

The tracking branch (Fig. 3.2(c)) and pose branch (Fig. 3.2(d)) mainly consist of MobileNet-V3 Block for better efficiency. In the pose branch, the output layer is activated by Tanh and then multiplied by π to normalize the output $\boldsymbol{P}_h$ into $[-\pi, +\pi]$. Hence, the predicted result can be more stable. In the tracking branch, the output is activated by Softmax to normalize face confidence C_f in [0, 1].

3.2.3 Heatmap

Different from the heatmap in other works [43], [60], [105], the heatmap in ATPN is generated directly based on the predicted landmarks by a formula. Therefore, the parameters and computational complexity can be reduced significantly. The formula can be written as:

$$H\left(x,y\right)=\frac{1}{\sqrt{1+\min_{(x_i',y_i')\in\boldsymbol{S}_1}\left\|(x,y)-(x_i',y_i')\right\|}},\tag{3.1}$$

where $H(x,y)$ is the intensity of point (x,y). (x_i', y_i') indicates the coordinate for the i-th landmark of the predicted shape $\boldsymbol{S'}$. To avoid the problem that the CNN completely ignore the feature far from facial landmarks, $H(x,y)$ is set to 0.5 if the value is less than 0.5. The heatmap is then fused with features using element-wise multiplication, as shown below:

$$\boldsymbol{F}_O=\boldsymbol{F}_I\otimes\boldsymbol{H},\tag{3.2}$$

where $\boldsymbol{F_O}$ indicates the output features, $\boldsymbol{H}$ and $\boldsymbol{F_I}$ indicates the heatmap and the features learned by backbone network. By fusing the heatmap with the intermediate features, The output features contain both the appearance information of the input images and the geometric information of the facial landmarks. Moreover, the heatmap also provides the attention clue for ATPN to eliminate the interference of background and improve the performance of face tracking.

3.2.4 Training Strategy

We train the ATPN in three stages to mitigate the problem of underfitting for difficult tasks while avoiding overfitting in simpler tasks.

In the first stage, we train the backbone network and face landmark detection branch together. We calculate the point-to-point Euclidean and normalize it with the Inter-ocular distance [109], which can be written as:

$$L_{\mathrm{A}}=\frac{\left\|\boldsymbol{S}'-\boldsymbol{S}^{\mathrm{gt}}\right\|_2}{\boldsymbol{d}_{\mathrm{ION}}},\tag{3.3}$$

where L^{Landmark} indicates the normlizaed error of face alignment. $\boldsymbol{d}_{\mathrm{ION}}$ is the Inter-ocular distance, $\boldsymbol{S'}$ and $\boldsymbol{S}^{\mathrm{gt}}$ are the predicted and annotated landmarks respectively. Then, the loss function is formulated as:

$$\min\left(\frac{1}{N_1}\left(\sum_{i=1}^{N_1}L_i^{\mathrm{Landmark}}\right)+w_1l_2^{\mathrm{Landmark}}\right),\tag{3.4}$$

where N_1 indicates the number of samples which are used in stage 1. l_2^{Landmark} is the L2-regularization loss of backbone network and face landmark detection branch, and w_1 is the weight of l_2^{Landmark}.

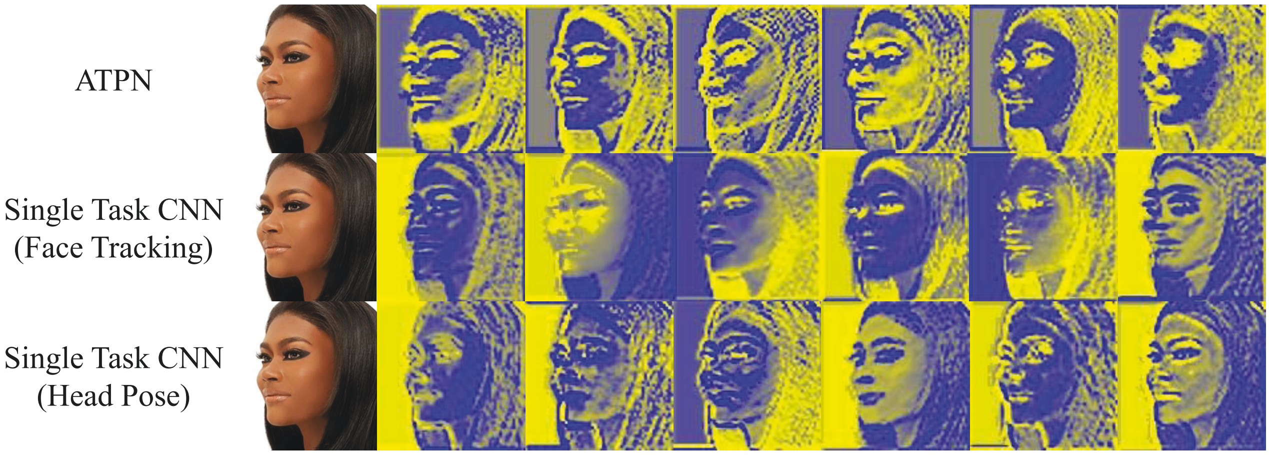


Figure 3.4: Column 1 shows the input image and column 2-7 are the low-level feature maps in the ATPN and two single task CNNs for head pose estimation and face tracking respectively. ATPN and the two single task CNNs utilize the same backbone for feature extraction. All visualized feature maps are from the 4-th MobileNet V3 block, and their resolution is 28×28. The pixel intensity of these feature maps represents the response strength of the corresponding pixel to the target tasks.

Then, we freeze the weights of the backbone and train the tracking branch and pose branch in stage 2 and stage 3 respectively. Since branches 2 and 3 are trained on different datasets for different tasks, they are trained individually, even though they run in parallel during inference. The face tracking error L^{Track} can be written as:

$$L^{\text{Track}} = -\left(y^{\text{Track}} \log\left(p^{\text{Track}}\right) + \left(1 - y^{\text{Track}}\right)\left(1 - \log\left(p^{\text{Track}}\right)\right)\right), \tag{3.5}$$

where p^{Track} is the predicted face confidence and $y^{\text{Track}} \in [0, 1]$ is the annotation (0: background, 1: face). The learning object of the second stage can be written as:

$$\min\left(\frac{1}{N_2}\left(\sum_{i=1}^{N_2} L_i^{\text{Track}}\right) + w_2 l_2^{\text{Track}}\right), \tag{3.6}$$

where N_2 indicates the number of samples used in the second stage. l_2^{Track} is the L2-regularization loss of the tracking branch and w_2 is the weight of l_2^{Track}. The error of head pose estimation L^{Pose} can be written as:

$$L^{\text{Pose}} = \frac{\sqrt{\|\boldsymbol{P}_{\text{head}} - \boldsymbol{P}_{\text{head}}{}'\|}}{3}, \tag{3.7}$$

where $\boldsymbol{P}_{\text{head}}$ and $\boldsymbol{P}'_{\text{head}}$ are the predicted and annotated head pose respectively. Finally, the learning object of third stage can be written as:

$$\min\left(\frac{1}{N_3}\left(\sum_{i=1}^{N_3} L_i^{\text{Pose}}\right) + w_3 l_2^{\text{Pose}}\right), \tag{3.8}$$

where N_3 indicates the number of samples used in the third stage. l_2^{Pose} is the L2-regularization loss of pose branch and w_3 is the weight of l_2^{Pose}.

Based on the training strategy, head pose estimation and face tracking can take advantages of a part of knowledge of face alignment. We visualize low-level features in the ATPN and two single task CNNs with same layers, as shown in Fig. 3.4. Compared to single task CNNs, the low-level features in the ATPN focus more on the edge of face rather than background. Therefore, sharing low-level features makes CNNs learn more effective features.

3.3 Experiments

3.3.1 Datasets

- **WFLW** [43] contains 10,000 faces (7,500 for training and 2,500 for testing) with 98 landmarks. Besides, each face also has attributes annotation in terms of large pose, expression, illumination, make-up, occlusion and blur.
- **300W-LP** [40] is a synthetic dataset. It fits every sample of 300W [109] to 3D models by 3DDFA [40] and rotates 3D model to generate 122450 samples with head pose label.
- **WIDER Face** [110] is a face detection benchmark dataset, including 32,203 images and 393,703 labeled faces with a high degree of variability in scale, pose and occlusion.
- **300VW** [111] is a video face landmark detection dataset, containing 50 videos for training and 64 videos for testing. The testing set is divided into three subsets (category A, B and C).

3.3.2 Implementation Detail

We utilize the samples with facial landmark annotation for the first stage training. The learning rate is set to 0.001 and reduced by a factor of 0.03 after every 4 epochs. In the second stage, tracking branch is trained with positive and negative samples created with WIDER Face. We only utilize the face area whose margin is more than 60 for positive samples. For each positive

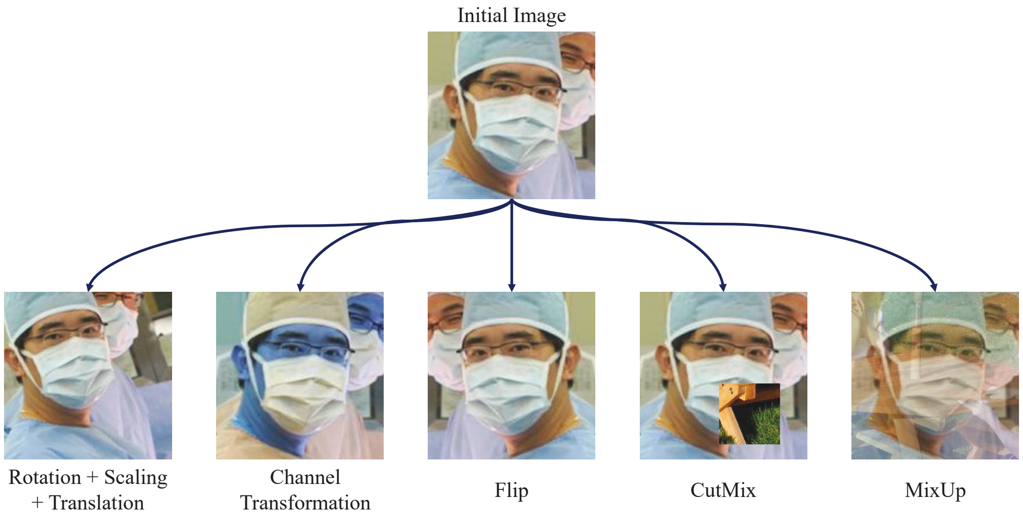


Figure 3.5: Data augmentations method used in our method.

sample generation, 10 negative samples whose IoU is smaller than 0.3 are generated. The learning rate is set to 0.001 and decayed by a factor 0.04 after every 2 epochs. In the third stage, we only use the samples of 300-LP without 3DDFA [40] data augmentation (3148 for training, 689 for testing) for training. It can keep domain consistency. The method used for data augmentation is the same as that used in the first stage. The learning rate setting is the same as the second stages.

Moreover, each sample randomly translates, rotates, flips, transforms channels or mixes up with the background image created by WIDER to augment the training data, as shown in Fig. 3.5. we utilize Adam [112] optimizer with a β_1 of 0.9 and a β_2 of 0.999 in the three stages. The batch size is set to 128.

3.3.3 Evaluation Metrics

Face Alignment

We evaluate the proposed method with Normalized Mean Error (NME). In WFLW, we use the inter-ocular distance [109] as the normalization factor. In 300VW, we also use the inter-pupil distance [43] as the normalization factor to compare ATPN with other methods. Besides, we also report the Failure Rate (FR) for a maximum error of 0.1 and Area Under Curve (AUC) derived from Cumulative Errors Distribution (CED) curve by plotting the curve from 0 to 0.1.

Head Pose Estimation

Following the works of [43], [105], we divide the test set into two subsets (common subset and challenging subset) and report the Mean Average Error (MAE), as given below:

$$MAE = \frac{1}{N_{\text{test}}} \sum_{l=1}^{N_{\text{test}}} \frac{|\alpha_l - \alpha_l'| + |\beta_l - \beta_l'| + |\gamma_l - \gamma_l'|}{3}, \tag{3.9}$$

where (α, β, γ) and $(\alpha', \beta', \gamma')$ are respectively the predicted and annotated roll, pitch, yaw angle of head. N_{test} is the total number of testing samples. Besides, we also report the FR in for a maximum error of 10° and CED curve.

Face Tracking

We evaluate ATPN with the NME and tracking failure rate. The threshold of failure rate is set to 0.7. Besides, we also report the Precision-Recall (PR) curve and Average Precision (AP) on the test set of WIDER Face.

3.3.4 Comparisons with State-of-the-Art Methods

WFLW

Table 3.1 compares the ATPN to the methods based on structural information and the methods for mobile devices respectively. Table 3.2 illustrates number of parameters and computational complexity of the methods. Despite the PropNet achieves best performance of 4.05% in NME, the computational complexity is much more than the methods for mobile devices (more than 150×). By taking advantages of the structural information at low-level features, ATPN achieves best performance in NME among the methods for mobile devices. Compared to $G\&LSR_{\omega}$, we observe an improvement of 2.47% in NME with comparable computational complexity. Although LAB adopts GAN to generate a boundary heatmap, ATPN still outperforms it with only 0.3% computational complexity and 8.9% parameters. Therefore, ATPN is much more efficient than the state-of-the-art methods.

Metric	Method	Full set	Pose	Expression	Illumination	Make-up	Occlusion	Blur
NME(%)↓	LAB† [43]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	SAN† [47]	5.22	10.39	5.71	5.19	5.49	6.83	5.8
	2 Hourglass† [56]	5.19	9.03	5.53	5.08	4.97	6.45	5.93
	DeCaFA† [60]	**4.62**	**8.11**	**4.65**	**4.41**	**4.63**	**5.74**	**5.38**
	PropNet† [105]	**4.05**	**6.92**	**3.87**	**4.07**	**3.76**	**4.58**	**4.36**
	MuSiCa98‡ [85]	7.90	15.80	8.52	7.49	8.56	10.04	8.92
	3FabRec‡ [20]	5.62	10.23	6.09	5.55	5.68	**5.92**	6.38
	$G\&LSR_\omega{}^\ddagger$[53]	5.26	-	-	-	-	-	-
	Res18+AVS‡ [88]	**5.25**	**9.10**	**5.83**	**4.93**	**5.47**	**6.26**	**5.86**
	ATPN	**5.13**	**8.97**	**5.49**	**4.95**	**4.94**	6.30	**5.78**
$\mathrm{FR}_{0.1}$(%)↓	LAB†	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	SAN†	6.32	27.91	7.01	4.87	6.31	11.28	**6.60**
	2 Hourglass†	6.04	25.46	5.41	5.59	**5.34**	12.5	8.40
	DeCaFA†	**4.84**	**21.4**	**3.73**	**3.22**	6.15	**9.26**	6.61
	PropNet†	**2.96**	**12.58**	**2.55**	**2.44**	**1.46**	**5.16**	**3.75**
	MuSiCa98‡	-	-	-	-	-	-	-
	3FabRec‡	8.28	34.35	**8.28**	6.73	10.19	15.08	9.44
	$G\&LSR_\omega{}^\ddagger$	**5.72**	-	-	-	-	-	-
	Res18+AVS‡	7.44	**32.52**	8.3	**4.3**	**8.25**	**12.77**	**9.06**
	ATPN	**6.27**	**26.99**	**6.05**	**4.72**	**7.28**	**12.22**	**7.89**
$\mathrm{AUC}_{0.1}$↑	LAB†	0.532	0.235	0.495	0.543	0.539	0.449	0.463
	SAN†	0.536	0.236	0.462	0.555	0.552	0.456	0.493
	2 Hourglass†	0.531	**0.337**	0.509	0.540	0.543	0.475	0.488
	DeCaFA†	**0.563**	0.292	**0.546**	**0.579**	**0.575**	**0.485**	**0.494**
	PropNet†	**0.615**	**0.382**	**0.628**	**0.616**	**0.638**	**0.572**	**0.583**
	MuSiCa98‡	-	-	-	-	-	-	-
	3FabRec‡	0.484	0.192	0.448	0.496	0.473	0.398	0.434
	$G\&LSR_\omega{}^\ddagger$	0.493	-	-	-	-	-	-
	Res18+AVS‡	**0.503**	**0.229**	**0.453**	**0.525**	**0.484**	**0.431**	**0.453**
	ATPN	**0.557**	**0.337**	**0.528**	**0.568**	**0.565**	**0.495**	**0.516**

Table 3.1: Performance comparison of the ATPN and the state-of-the-art methods on WFLW and its subsets. Key: [**Best**, **Second Best**, ↓=the lower the better, ↑=the larger the better, †=the method is based on structural information or semi-supervised learning, ‡=the method is for mobile devices]

Method	Params(M)↓	FLOPS(G)↓
PropNet [105]	36.3	42.83
LAB[43]	12.3	18.85
DeCaFA [60]	≈10	≈30
2 Hourglass [56]	6.30	8.00
MuSiCa98‡ [85]	≈3	**≈0.25**
$G\&LSR_\omega$ [53]	**1.83**	**0.06**
ATPN (Backbone + landmark Branch)	**1.1**	**0.06**

Table 3.2: Parameters and computational complexity for ATPN and other the state-of-the-art methods. Key: [**Best**, **Second Best**, ↓=the lower the better]

Method	Cat A	Cat B	Cat C
inter-ocular distance normalization (%)↓			
TSCN [113]	12.54	7.25	13.13
CFSS [12]	7.68	6.42	13.67
SDM [9]	7.41	6.08	14.03
TSTN [114]	5.21	4.23	10.11
ADC [115]	4.17	3.89	7.28
DeCaFA [60]	3.82	3.63	6.67
FAB [116]	3.56	3.88	5.02
ATPN (Tracking)	**3.52**	**3.64**	**4.99**
ATPN (Detection)	**3.49**	**3.57**	**4.89**
inter-pupil distance normalization (%)↓			
4S-HG [56]	6.54	5.65	8.13
4S-HG+UFLD [117]	6.09	5.34	7.76
ATPN (Tracking)	**4.83**	**5.04**	**7.75**
ATPN (Detection)	**4.79**	**4.95**	**7.61**

Table 3.3: NME for the ATPN in tracking and detection mode compared with previous methods on Category A, Category B and Category C of 300VW. Key: [**Best**, **Second Best**, ↓=the lower the better]

	Cat A	Cat B	Cat C
Sequences	146	39	188
Detection frame	233	39	210
Tracking frame	61902	32766	26128
Failure frame	87	0	22
Failure rate	0.14%	0.00%	0.08%

Table 3.4: Tracking details on 300VW. **Note**: 300VW fails to label some frames in several videos, resulting in a video being divided into dozens of sequences.

Method	Params(M)↓	FLOPS(G)↓
MTCNN [86]	0.47	0.7∼1.4
Tracking Branch	**0.19**	**0.017**

Table 3.5: Parameters and computational complexity for tracking branch and MTCNN. Key: [**Best**, ↓=the lower the better].

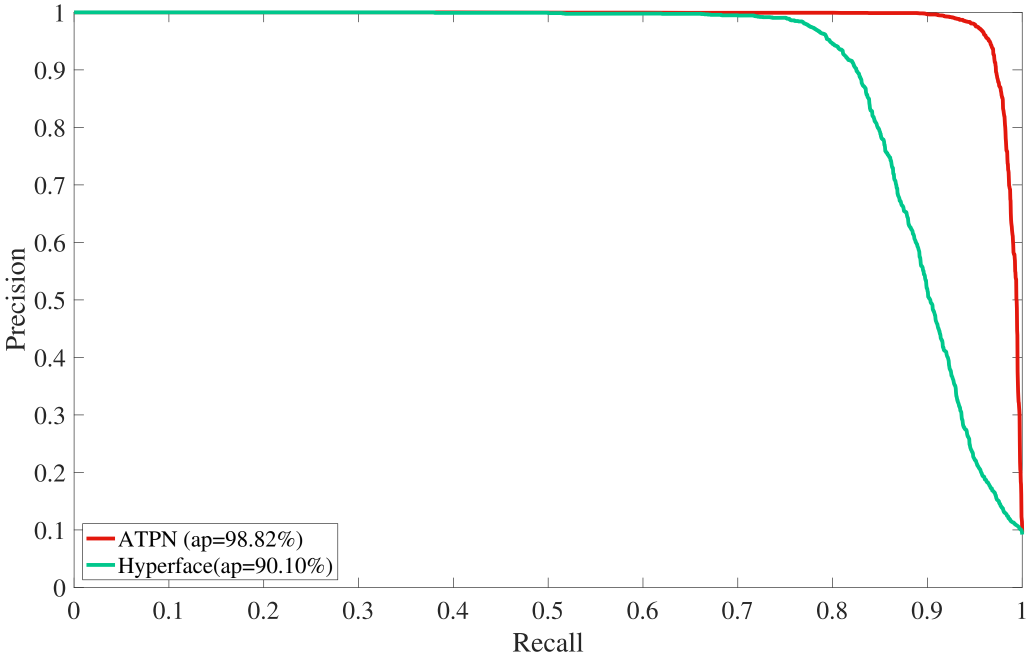


Figure 3.6: Evaluation on the validation set of WIDER FACE. The number following the method indicates the average accuracy of face tracking.

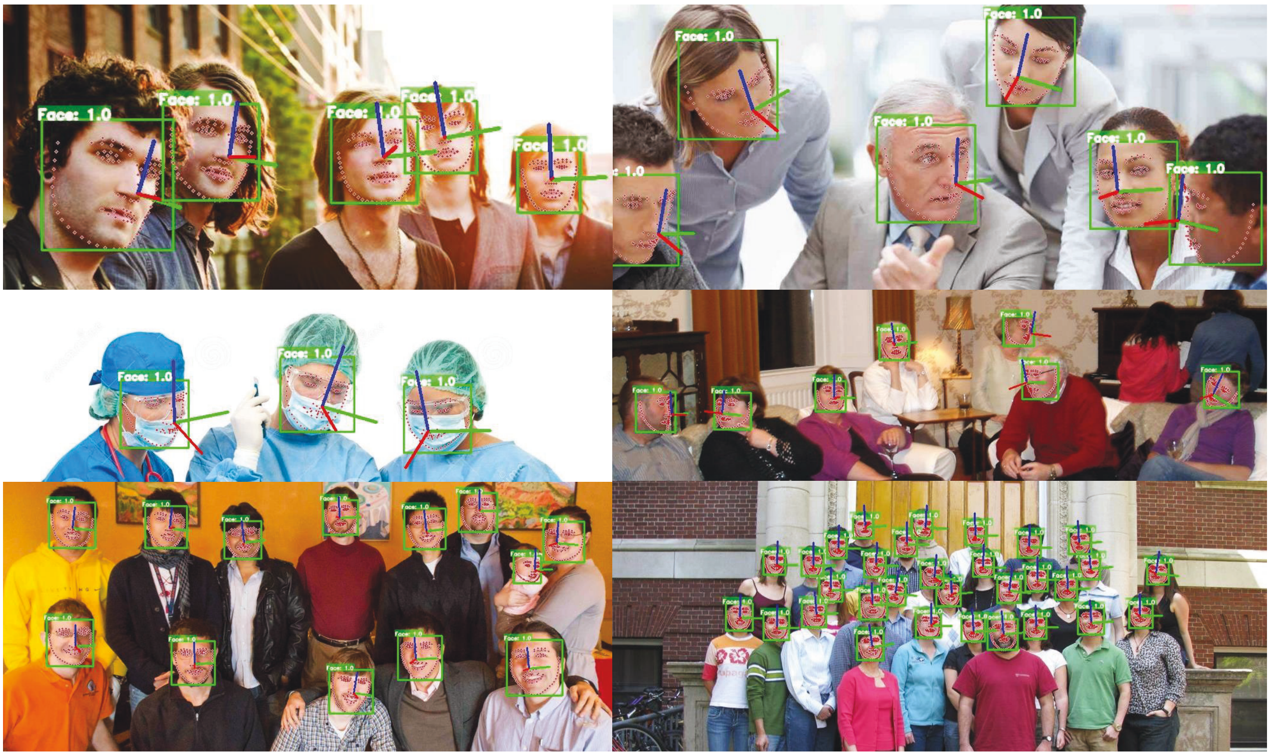


Figure 3.7: Sample results on WIDER Face validation set (the number above the bounding box indicates the confidence of face, the red axis points towards the front of the face, blue pointing upward and green pointing left side).

300VW

We pretrain the ATPN on WFLW and then retrain it with the training set of 300VW. We carry out two different experiments on three subsets of 300VW. In the frist one, we evaluate ATPN with detection mode. In this mode, the input image is cropped based on the ground truth. In the second experiment, the input image is cropped based on the face landmark detection result in the last frame for tracking. The comparison result are illustrated in Table 3.3. The most samples in 300VW are without occlusion, which enables the low-level features to produce more complete facial boundaries. Therefore, the improvement of ATPN is more significantly compared to other state-of-the-art methods. For example, although DeCaFA achieves excellent performance on WFLW, ATPN still achieves an impressive improvement of 8.64%, 1.66% and 26.69% in NME respectively on the three subsets.

Moreover, as shown in Table 3.4, the tracking failure rates are only 0.14%, 0.00% and 0.08% on the three subsets, which means the tracking branch is with satisfactory robustness. As a result, the performance of tracking mode only degrades a little compared to detection mode.

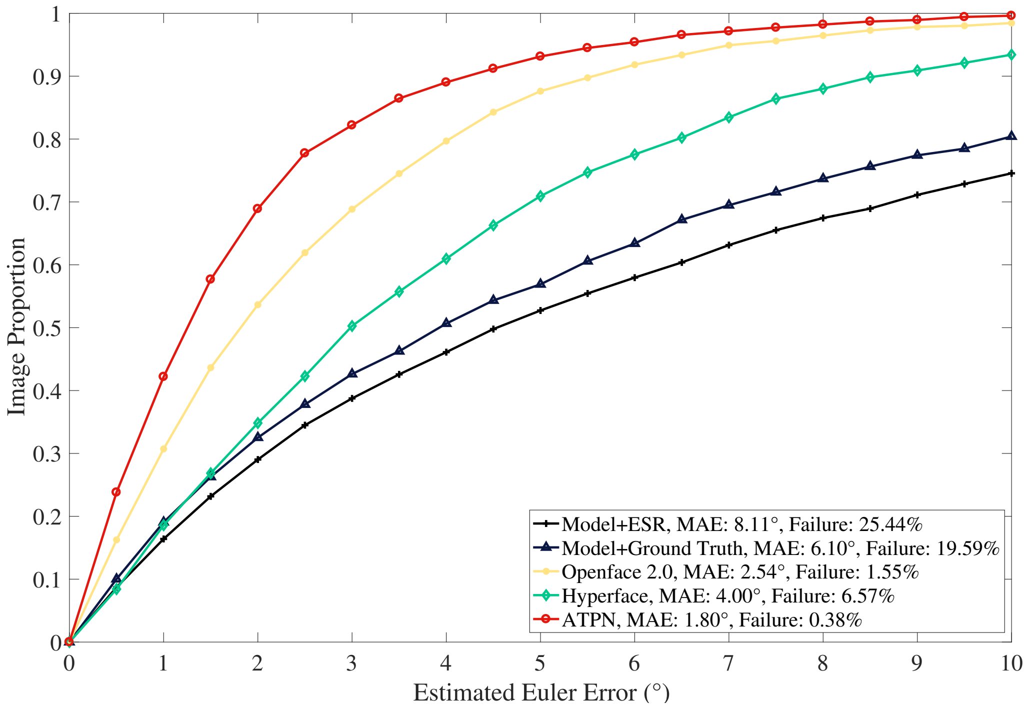


Figure 3.8: CED for the full test set of 300W-LP. MAE and Failure Rate are also reported.

Besides, in Fig. 3.6, we also demonstrate the PR curves of the ATPN and Hyperface [118] on the validation set of **WIDER Face** [110]. The average precision of ATPN reaches at 98.82% that outperforms Hyperface a lot. Some predicted results in the validation set can be viewed in Fig. 3.7.

The Table 3.5 shows the parameters and computational complexity of the tracking branch and a commonly used face detection framework (MTCNN). The MTCNN is with 10× higher computational complexity than ATPN. Hence, the real-time capability of ATPN will degrades dramatically if the face detection framework is employed in each frame. By taking advantages of tracking branch, the processing time can be accelerated more than 50 times.

300W-LP

We compare the pose branch to other state-of-the-arts methods on both the common and challenging subset, as shown in Table 3.6. Besides, Fig. 3.8 also illustrates the CED curves of the ATPN and other methods on the full set. It is difficult for 5-landmarks model-based method to achieve accurate head pose estimation even if it utilizes the annotated landmarks. Hyperface and FSA-Net utilize the appearance features of the input image. The performance is improved significantly because the appearance features contains more information compared

Method	Common Subset					Challenging Subset				
	Yaw↓	Pitch↓	Roll↓	MAE↓	FR_{10°↓	Yaw↓	Pitch↓	Roll↓	MAE↓	FR_{10°↓
ESR [7]*	8.54°	7.47°	2.61°	6.21°	19.31%	24.52°	12.70°	10.64°	15.95°	50.62%
Annotated landmarks*	7.32°	6.99°	1.88°	5.40°	16.37%	11.64°	9.96°	5.40°	9.00°	32.84%
Hyperface [118]	3.46°	3.87°	2.92°	3.42°	3.37%	6.30°	5.57°	7.34°	6.40°	19.76%
Openface 2.0 [119]	2.69°	3.56°	**1.10°**	2.45°	**1.72%**	**3.23°**	**3.56°**	**1.93°**	**2.91°**	**1.51%**
FSA-Net [120]	**1.85°**	**2.84°**	1.12°	**1.94°**	**2.35%**	4.51°	5.29°	**2.27°**	4.12°	19.26%
ATPN	**1.31°**	**2.38°**	**0.97°**	**1.55°**	**0.12%**	**2.62°**	**3.97°**	**1.93°**	**2.84°**	**1.49%**

Table 3.6: Mean Average Error (MAE) in degrees and FR_{10° on the common and challenging subset. Key: [**Best**, **Second Best**, ↓=the lower the better, *=5-landmarks model-based method (eyes corner, mouth corner, nose tip and chin)]

to 5 landmarks. Nevertheless, it degrades significantly for the images with extreme conditions (challenging subset) because of the fragile robustness. OpenPose 2.0 adopt 3D geometric information to estimate head pose and it can maintain excellent performance on the challenging subset. ATPN estimates head pose based on both geometric and appearance features, which significantly reduces the MAE by 36.73%, from 2.45° to 1.55° on the common subset compared to Openface 2.0. Besides, the geometric features enables ATPN to maintain comparable performance.

3.3.5 Ablation Study

Landmark Branch

The landmark branch consists of several pivotal modules: the structural information at low-level, multiview block and CoordConv. We compare a baseline with several models that consist of these modules. The baseline is set as a model that regresses facial landmarks directly from the last layer of the backbone. The evaluation results on WFLW are shown in Table 3.7.

We observe 3.74% and 17.51% improvement respectively in NME and $\mathrm{FR}_{0.1}$ by introducing the structural information into CNN. It illustrates that the structural information of low-level features is essential to face alignment. Then, CoordConv also improves the performance of face landmark detection significantly by providing spatial information for ATPN. Without adding much computational complexity, NME and FR are reduced by 3.57% and 16.62% respectively. Finally, the improvement brought by multiview block is not as significant as others, although it obtains the largest number of parameters. It suggests that increasing parameters in network cannot improve the performance effectively.

Feature Sharing and Heatmap

Feature sharing and heatmap also play an important role in the tracking and pose branch. To further investigate whether the heatmap can introduce the geometric information into the network, we carry out an additional experiment on background images and some results are shown in Fig. 3.9. Despite the input images do not include any face, the predicted head pose of the model with heatmap is still highly correspond to the predicted facial landmarks, while the results of the model without heatmap are not. It illustrates that when the appearance features are not reliable, the model with heatmap can utilize the geometric information for head pose

Method	Modules			NME(%)↓							$\mathrm{FR}_{0.1}$(%)↓
	low-level Features	Multiview Block	CoordConv	Full set	Pose	Expression	Illumination	Make-up	Occlusion	Blur	Testset
Baseline				5.61	9.90	6.11	5.36	5.47	6.77	6.38	8.68
Model1	✓			5.40	9.22	5.92	5.30	5.30	6.43	6.00	7.16
Model2	✓	✓		5.33	9.15	5.75	5.16	5.33	6.36	5.95	7.52
Model3	✓	✓	✓	**5.13**	**8.97**	**5.49**	**4.95**	**4.94**	**6.30**	**5.78**	**6.27**

Table 3.7: The contributions of different modules, ↓=the lower the better]

Method	Common Subset					Challenging Subset				
	Yaw↓	Pitch↓	Roll↓	MAE↓	FR_{10°↓	Yaw↓	Pitch↓	Roll↓	MAE↓	FR_{10°↓
Baseline	1.47°	2.53°	1.03°	1.68°	0.12%	2.93°	3.98°	2.44°	3.12°	4.70%
Feature sharing	1.42°	2.60°	1.11°	1.71°	0.24%	2.96°	4.17°	2.38°	3.17°	3.46%
Feature sharing+Heatmap	**1.31°**	**2.38°**	**0.97°**	**1.55°**	**0.12%**	**2.62°**	**3.97°**	**1.93°**	**2.84°**	**1.49%**

Table 3.8: The effectiveness of feature sharing and heatmap on face tracking on common subset and challenging subset. Key: [**Best**, ↓=the lower the better]

Method	Baseline	Feature Sharing	Feature Sharing + Heatmap
Average Precision ↑	98.57%	98.72%	**98.82%**

Table 3.9: The influence of feature sharing and heatmap on the face tracking on WIDER Face. Key: [**Best**, ↑=the larger the better]

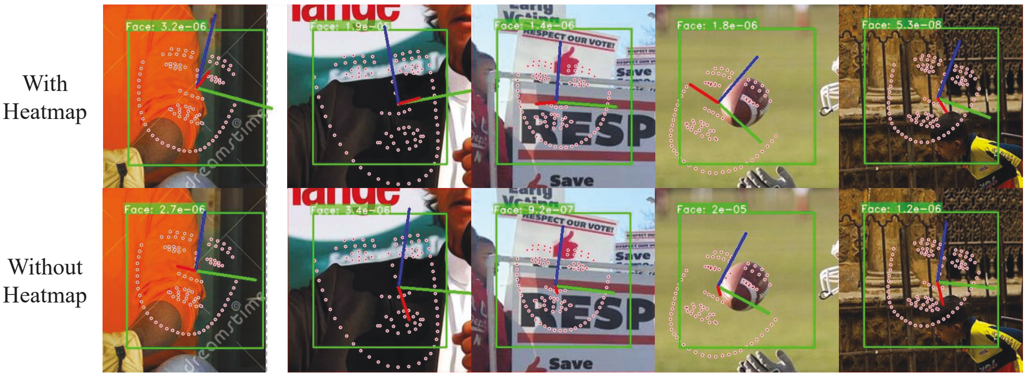


Figure 3.9: Estimation results of the models with/without heatmap on the background images (the **red axis** points towards the front of the face, blue pointing upward and green pointing left side).

estimation.

To study the influence of feature sharing and heatmap on two tasks, we set the baseline as a single task CNN with the same layers. In feature sharing model, the low-level layers are shared with face landmark detection and only high-level layers are trained.

The results of head pose estimation on 300W-LP are shown in Table 3.8. Compared to the baseline, the feature sharing model exhibits the comparable performance on the both common and challenging subset with much less parameters. With the geometric information of the heatmap, the MAE is significantly improved by 10.41% and 9.36% respectively on the common and challenging subset. Therefore, sharing features can boost the real-time capability and the geometric information of the heatmap is crucial to head pose estimation.

In terms of face tracking, the experiment is carried out on WIDER Face. The influence of feature sharing and heatmap is shown in Table 3.9. Although the trainable parameters in features sharing model is much less than the baseline, the average precision is still improved from 98.57%

to 98.72%. By introducing the attention clue into network with heatmap, the performance is further boosted to 98.82%. It indicates that the features learned by face landmark detection task and the attention clue of heatmap can improve the performance of face tracking.

3.4 Conclusion

In this chapter, we present the Alignment & Tracking & Pose Network (ATPN), an efficient multitask network for facial landmark localization, face tracking and head pose estimation. Different from other state-of-the-art works that generates the structural information by an additional network or branch, we directly utilize the structural information in the low-level features. By taking advantage of the structural information, our method retains strong robustness and reliability in facial landmark detection under extreme conditions, with only 1/300 of the computational complexity of other methods utilizing the same information. Moreover, it also achieves the best performance with the least parameters and FLOPS compared to other methods for mobile devices. On the video dataset, the improvement of real-time capability is more significant because the tracking branch eliminates the face detection process. Different from other multitask frameworks that only share features, ATPN also generates a heatmap by face landmark detection result to provide the geometric information for head pose estimation and the attention clues for face tracking. The experimental results on existing benchmarks further show that the proposed method successfully retains strong robustness and reliability under extreme conditions with much less computational complexity. The ablation studies demonstrate the effectiveness of our heatmap attention mechanism and feature sharing strategy.

Chapter 4

Robust and Reliable Facial Landmark Detection for Heavy Occluded Faces via Inherent Relation Learning

4.1 Introduction

In this chapter, we propose a Sparse Local Patch Transformer (SLPT) to learn the inherent relations among facial landmarks. Human tends to locate the landmarks with heavy occlusion or illumination variation by their relative position to the easily identified landmark. We define this clue as the inherent relation and the relation is case dependent. Based on these relations, the proposed method can utilize visible landmarks to infer the locations of occluded landmarks, which significantly improves robustness and reliability in cases of heavy occlusion.

To learn the inherent relations, instead of predicting the coordinates from the full feature map like DETR [76], the SLPT firstly generates the representation for each landmark from a local patch, , as shown in Fig. 4.1. Then, a series of learnable queries, which are called *landmark queries*, are used to aggregate the representations. Based on the cross-attention mechanism of transformer, the SPLT learns an adaptive adjacency matrix in each layer, which represents the landmark inherent relation. Finally, the subpixel coordinate of each landmark in their corresponding patch is predicted independently by a multilayer perceptron (MLP). Due to the use of sparse local patches, the number of the input token decreases significantly compared to other vision transformer[73], [76].

To further improve the performance, a coarse-to-fine framework is introduced to incorporate with the SLPT. Similar to cascaded shape regression method [12], [15], [42], the proposed framework optimizes a group of initial landmarks to the target landmarks by several stages. The local patches in each stage are cropped based on the initial landmarks or the landmarks predicted in the former stage, and the patch size for a specific stage is 1/2 of its former stage. As a result, the local patches evolve in a pyramidal form and get closer to the target landmarks for the fine-grained local feature.

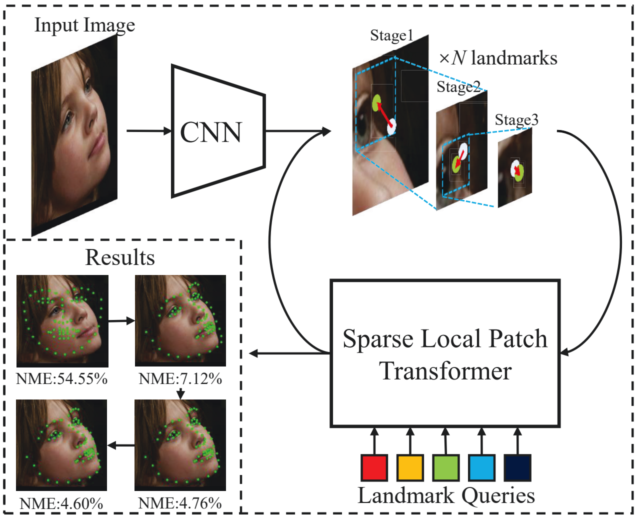


Figure 4.1: The proposed coarse-to-fine framework leverages the sparse local patches for robust face alignment. The sparse local patches are cropped according to the landmarks in the previous stage and fed into the same SLPT to predict the facial landmarks. Moreover, the patch size narrows down with the increasing of stages to enable the local features to evolve into a pyramidal form.

4.2 Method

4.2.1 Sparse Local Patch Transformer

As shown in Fig. 4.2, Sparse Local Patch Transformer (SLPT) consists of three parts, the patch embedding & structure encoding, inherent relation layers and prediction heads.

Patch embedding & structure encoding

ViT [73] divides an image or a feature map $\boldsymbol{I} \in \mathbb{R}^{H_I \times W_I \times C}$ into a grid of $\frac{H_I}{P_h} \times \frac{W_I}{P_w}$ with each patch of size $P_h \times P_w$ and maps it into a d-dimension vector as the input. Different from ViT,

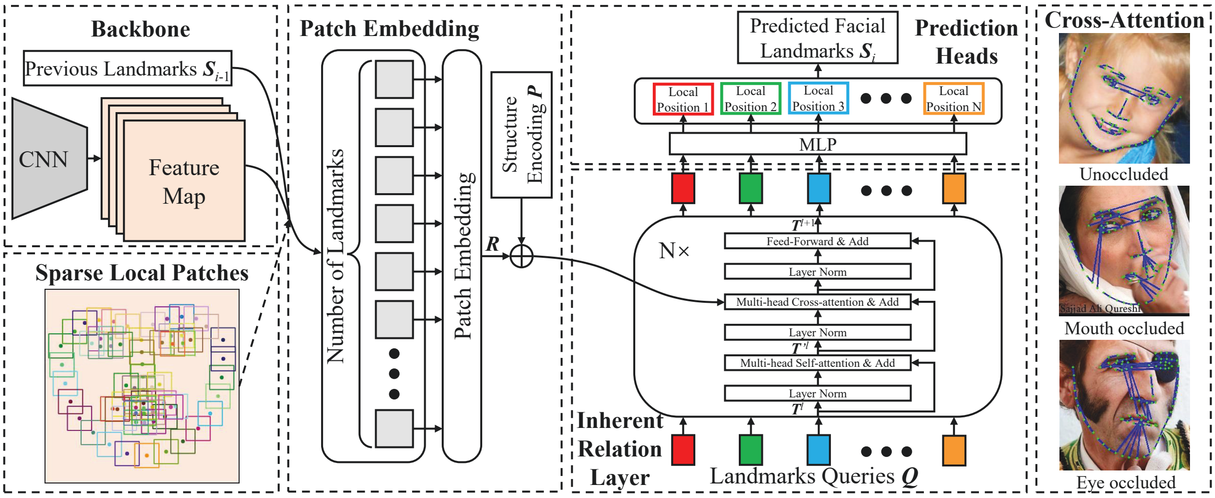


Figure 4.2: An overview of the SLPT.

for each landmark, the SLPT crops a local patch with the fixed size (P_h, P_w) from the feature map as its supporting patch, whose center is located at the landmark. Then, the patches are resized to $K \times K$ by linear interpolation and mapped into a series of vectors by a CNN layer. Hence, each vector can be viewed as the representation of the corresponding landmark. Besides, to retain the relative position of landmarks in a regular face shape (structure information), we supplement the representations with a series of learnable parameters called *structure encoding*. As shown in Fig. 4.3, the SLPT learns to encode the distance between landmarks within the regular facial structure in the similarity of encodings. Each encoding has high similarity with the encoding of neighboring landmark (eg. left eye and right eye).

Inherent relation layer

Inspired by Transformer [121], we propose inherent relation layers to model the relation between landmarks. Each layer consists of three blocks, multi-head self-attention (MSA) block, multi-head cross-attention (MCA) block, and multilayer perceptron (MLP) block, and an additional Layernorm (LN) is applied before every block. Based on the self-attention mechanism in MSA block, the information of queries interact adaptively for learning a $query - query$ inherent relation. Supposing the l-th MSA block obtains H heads, the input T^l and landmark queries Q with C_I-dimension are divided into H sequences equally (T^l is a zero matrix in 1st layer). The self-attention weight of the h-th head $\boldsymbol{A}_h$ is calculated by:

$$\boldsymbol{A}_h = softmax\left(\frac{\left(\boldsymbol{T}_h^l + \boldsymbol{Q}_h\right)\boldsymbol{W}_h^q\left(\left(\boldsymbol{T}_h^l + \boldsymbol{Q}_h\right)\boldsymbol{W}_h^k\right)^T}{\sqrt{C_h}}\right), \tag{4.1}$$

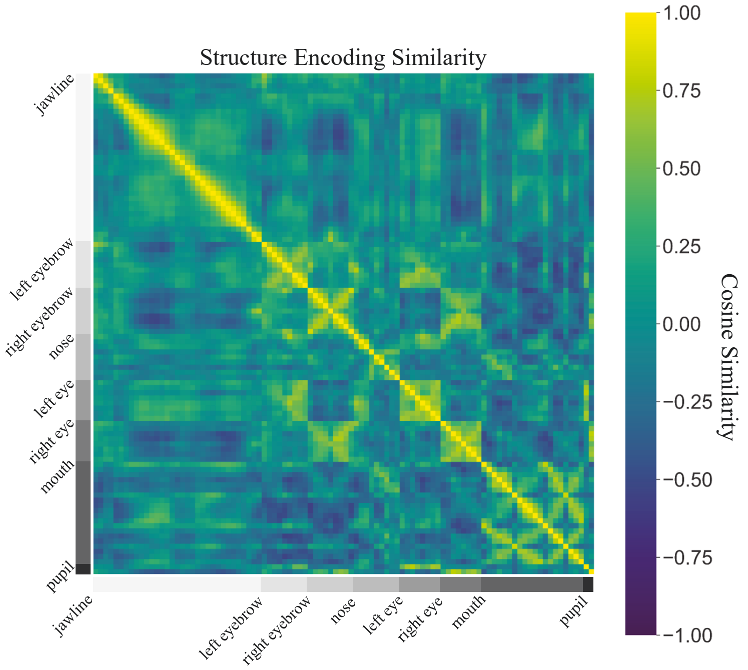


Figure 4.3: Cosine similarity for structure encodings of SLPT learned from a dataset with 98 landmark annotations. High cosine similarities are observed for the corresponding points which are close in the regular face structure.

where $\boldsymbol{W}_h^q$ and $\boldsymbol{W}_h^k \in \mathbb{R}^{C_h \times C_h}$ are the learnable parameters of two linear layers. $\boldsymbol{T}_h^l \in \mathbb{R}^{N \times C_h}$ and $\boldsymbol{Q}_h \in \mathbb{R}^{N \times C_h}$ are the input and landmark queries respectively of the h-th head with the dimension $C_h = C_I / H$. Then, MSA block can be formulated as:

$$MSA\left(\boldsymbol{T}^l\right) = \left[\boldsymbol{A}_1 \boldsymbol{T}_1^l \boldsymbol{W}_1^v; ...; \boldsymbol{A}_H \boldsymbol{T}_H^l \boldsymbol{W}_H^v\right] \boldsymbol{W}_P, \tag{4.2}$$

where $\boldsymbol{W}_h^v \in \mathbb{R}^{C_h \times C_h}$ and $\boldsymbol{W}_P \in \mathbb{R}^{C_I \times C_I}$ are also the learnable parameters of linear layers.

The MCA block aggregates the representations of facial landmarks based on the cross-attention mechanism for learning an adaptive $representation-query$ relation. As shown in the rightmost images of Fig. 4.2, by taking advantage of the cross attention, each landmark can employ neighboring landmarks for coherent prediction and the occluded landmark can be predicted according to the representations of visible landmarks. Similar to MSA, MCA also has H heads and the attention weight in the h-th head $\boldsymbol{A}_h'$ can be calculated by:

$$\boldsymbol{A}_h' = softmax\left(\frac{\left(\boldsymbol{T}_h'^l + \boldsymbol{Q}_h\right) \boldsymbol{W}_h'^q \left(\left(\boldsymbol{R}_h + \boldsymbol{P}_h\right) \boldsymbol{W}_h'^k\right)^T}{\sqrt{C_h}}\right). \tag{4.3}$$

Where $\boldsymbol{W}_h'^q$ and $\boldsymbol{W}_h'^k \in \mathbb{R}^{C_h \times C_h}$ are learnable parameters of two linear layers in the h-th head. $\boldsymbol{T}_h'^l \in \mathbb{R}^{N \times C_h}$ is the input l-th MCA block; $\boldsymbol{P}_h \in \mathbb{R}^{N \times C_h}$ is the structure encodings; $\boldsymbol{R}_h \in \mathbb{R}^{N \times C_h}$

is the landmark representations. MCA block can be formulated as:

$$MCA\left(\boldsymbol{T}^{\prime l}\right)=\left[\boldsymbol{A}_1^{\prime}\boldsymbol{T}_1^{\prime l}\boldsymbol{W}_1^{\prime v};...;\boldsymbol{A}_H^{\prime}\boldsymbol{T}_H^{\prime l}\boldsymbol{W}_H^{\prime v}\right]\boldsymbol{W}_P^{\prime}, \tag{4.4}$$

where $\boldsymbol{W}_h^{\prime v}\in\mathbb{R}^{C_h\times C_h}$ and $\boldsymbol{W}_P^{\prime}\in\mathbb{R}^{C_I\times C_I}$ are also the learnable parameters of linear layers in MCA block.

Supposing predicting N pre-defined landmarks, the computational complexity of the MCA that employ sparse local patches $\Omega(S)$ and full feature map $\Omega(F)$ is:

$$\Omega(S)=4HNC_h^2+2HN^2C_h, \tag{4.5}$$

$$\Omega(F)=\left(2N+2\frac{W_IH_I}{P_wP_h}\right)HC_h^2+2NH\frac{W_IH_I}{P_wP_h}C_h. \tag{4.6}$$

Compared to using the full feature map, the number of representations decreases from $\frac{H_I}{P_h}\times\frac{W_I}{P_w}$ to N (with the same input size, $\frac{H_I}{P_h}\times\frac{W_I}{P_w}$ is 16×16 in the related framework [76]), which decreases the computational complexity significantly. For a 29 landmark dataset [8], $\Omega(S)$ is only 1/5 of $\Omega(F)$ ($H=8$ and $C_h=32$ in the experiment).

Prediction head: the prediction head consists of a layernorm to normalize the input and a MLP layer to predict the result. The output of the inherent relation layer is the local position of the landmark with respect to its supporting patch. Based on the local position on the i-th patch $\left(t_x^i,t_y^i\right)$, the global coordinate of the i-th landmark (x^i,y^i) can be calculated by:

$$\begin{aligned}x^i&=x_{lt}^i+w^it_x^i,\\y^i&=y_{lt}^i+h^it_y^i,\end{aligned} \tag{4.7}$$

where (w^i,h^i) is the size of the supporting patch.

4.2.2 Coarse-to-fine locating

To further improve the performance and robustness of SLPT, we introduce a coarse-to-fine framework trained in an end-to-end method to incorporate with the SLPT. The pseudo-code in **Algorithm 1** shows the training pipeline of the framework. It enables a group of initial facial landmarks $\boldsymbol{S}_0$ calculated from the mean face in the training set to converge to the target facial landmarks gradually with several stages. Each stage takes the previous landmarks as center to crop a series of patches. Then, the patches are resized into a fixed size $K\times K$ and fed into the SLPT to predict the local point on the supporting patches. Large patch size in the initial stage enables the SLPT to obtain a large receptive filed that prevents the patch from deviating from the target landmark. Then, the patch size in the following stages is 1/2 of its former stage, which enables the local patches to extract fine-grained features and evolve into a pyramidal form. By taking advantage of the pyramidal form, we can observe a significant improvement for SLPT.

Algorithm 1 Training pipeline of the coarse-to-fine framework

```
Require: Training image $\boldsymbol{I}$, initial landmarks $\boldsymbol{S}_0$, backbone network $B$, SLPT $T$, loss function
    $L$, ground truth $\boldsymbol{S}_{gt}$, Stage number $N_{stage}$
while the training epoch is less than a specific number do
    Forward $B$ for feature map by $\boldsymbol{F} = B(I)$;
    Initialize the local patch size $(P_w, P_h) \leftarrow \left(\frac{W}{4}, \frac{H}{4}\right)$
    for $i \leftarrow 1$ to $N_{stage}$ do
        Crop local pactes $\boldsymbol{P}$ from $\boldsymbol{F}$ according to former landmarks $\boldsymbol{S}_{i-1}$;
        Resize patches from $(P_w, P_h)$ to $K \times K$;
        Forward $T$ for landmarks by $\boldsymbol{S}_i = T(\boldsymbol{P})$;
        Reduce the patch size $(P_w, P_h)$ by half;
    end for
    Minimize $L\left(\boldsymbol{S}_{gt}, \boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_{N_{stage}}\right)$
end while
```

4.2.3 Loss Function

We employ the normalized L2 loss to provide the supervision for stages of the coarse-to-fine framework. Moreover, similar to other works [56], [57], providing additional supervision for the intermediate output during the training is also helpful. Therefore, we feed the intermediate output of each inherent relation layer into a shared prediction head. The loss function is written as:

$$L = \frac{1}{SDN} \sum_{i=1}^{S} \sum_{j=1}^{D} \sum_{k=1}^{N} \frac{\left\| \left(x_{gt}^{k}, y_{gt}^{k}\right) - \left(x^{ijk}, y^{ijk}\right) \right\|_2}{d}, \tag{4.8}$$

where S and D indicate the number of coarse-to-fine stage and inherent relation layer respectively. $\left(x_{gt}^{k}, y_{gt}^{k}\right)$ is the labeled coordinate of the k-th point. $\left(x^{ijk}, y^{ijk}\right)$ is the coordinate of k-th point predicted by j-th inherent relation layer in i-th stage. d is the distance between outer eye corners that acts as a normalization factor.

4.3 Experiment

4.3.1 Datasets

Experiments are conducted on three popular benchmarks, including WFLW [43], 300W[109] and COFW[8].

WFLW dataset is a very challenging dataset that consists of 10,000 images, 7,500 for training and 2,500 for testing. It provides 98 manually annotated landmarks and rich attribute labels, such as profile face, heavy occlusion, make-up and illumination.

300W is the most commonly used dataset that includes 3,148 images for training and 689 images for testing. The training set consists of the fullset of AFW [122], the training subset of HELEN [123] and LFPW [124]. The test set is further divided into a challenging subset that includes 135 images (IBUG fullset [109]) and a common subset that consists of 554 images (test subset of HELEN and LFPW). Each image in 300W is annotated with 68 facial landmarks.

COFW mainly consists of the samples with heavy occlusion and profile face. The training set includes 1,345 images and each image is provided with 29 annotated landmarks. The test set has two variants. One variant presents 29 landmarks annotation per face image (COFW), The other is provided with 68 annotated landmarks per face image (COFW68 [125]). Both contains 507 images. We employ the COFW68 set for *cross*-dataset validation.

4.3.2 Evaluation Metrics

Referring to other related work [52], [61], [67], we evaluate the proposed methods with standard metrics, Normalized Mean Error (NME), Failure Rate (FR) and Area Under Curve (AUC). **NME** is defined as:

$$NME\left(\boldsymbol{S}, \boldsymbol{S}_{\text{gt}}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\boldsymbol{p}^{i}-\boldsymbol{p}_{\text{gt}}^{i}\right\|_{2}}{d} \times 100\%, \tag{4.9}$$

where $\boldsymbol{S}$ and $\boldsymbol{S}_{\text{gt}}$ denote the predicted and annotated coordinates of landmarks respectively. $\boldsymbol{p}^i$ and $\boldsymbol{p}_{\text{gt}}^i$ indicate the coordinate of i-th landmark in $\boldsymbol{S}$ and $\boldsymbol{S}_{\text{gt}}$. N is the number of landmarks, d is the reference distance to normalize the error. d could be the distance between outer eye corners (inter-ocular) or the distance between pupil centers (inter-pupils). **FR** indicates the percentage of images in the test set whose NME is higher than a certain threshold. **AUC** is calculated based on Cumulative Error Distribution (CED) curve. It indicates the fraction of

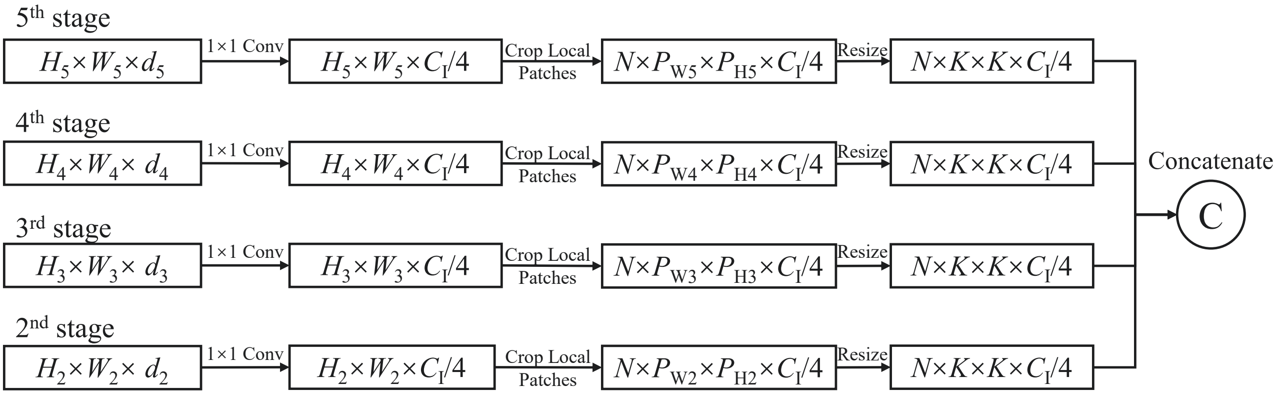

5th stage
$H_5 \times W_5 \times d_5$
1×1 Conv
$H_5 \times W_5 \times C_I/4$
Crop Local Patches
$N \times P_{W5} \times P_{H5} \times C_I/4$
Resize
$N \times K \times K \times C_I/4$
4th stage
$H_4 \times W_4 \times d_4$
1×1 Conv
$H_4 \times W_4 \times C_I/4$
Crop Local Patches
$N \times P_{W4} \times P_{H4} \times C_I/4$
Resize
$N \times K \times K \times C_I/4$
Concatenate
C
3rd stage
$H_3 \times W_3 \times d_3$
1×1 Conv
$H_3 \times W_3 \times C_I/4$
Crop Local Patches
$N \times P_{W3} \times P_{H3} \times C_I/4$
Resize
$N \times K \times K \times C_I/4$
2nd stage
$H_2 \times W_2 \times d_2$
1×1 Conv
$H_2 \times W_2 \times C_I/4$
Crop Local Patches
$N \times P_{W2} \times P_{H2} \times C_I/4$
Resize
$N \times K \times K \times C_I/4$

Figure 4.4: Constructing multi-level feature maps for SLPT

test images whose NME(%) is less or equal to the value on the horizontal axis. AUC is the area under CED curve, from zero to the threshold for FR.

4.3.3 Implementation Details

Each input image is cropped and resized to 256×256 pixels. We train the proposed framework with Adam [112], setting the initial learning rate to 1×10^{-3}. Without specifications, the size of the resized patch is set to 7×7 and the framework has 6 inherent relation layers and 3 coarse-to-fine stages. Besides, we augment the training set with random horizontal flipping (50%), gray (20%), occlusion (33%), scaling (±5%), rotation (±30°), translation ($\pm 10px$). We implement our method with two different backbone: a light HRNetW18C [55] (the modularized block number in each stage is set to 1) and Resnet34 [87]. For the HRNetW18C-lite, the resolution of feature map is 64×64. For the ResNet34, we extract representations from the output feature maps of stages C2 through C5, as shown in Fig. 4.4. Supposing the feature map size of k-th stage in ResNet34 is $W_k \times H_k \times d_k$, we firstly adopt a 1×1 CNN layer to reduce the channels from d_k to $C_I/4$. Then, the SLPT crops N patches whose size is $P_{Wk} \times P_{Hk}$ from each level and resizes these patches to $K \times K$. Note that $P_{Wk} \times P_{Hk}$ is $W_k/4 \times H_k/4$ in the initial coarse-to-fine stage and is reduced by half in each following stage. Finally, the resized patches from different levels are concatenated on the channel dimension which is C_I. As the result, the SLPT can utilize both high level and low level features for facial landmark detection.

Metric	Method	Full	Pose	Expression	Illumination	Make-up	Occlusion	Blur
NME(%)↓	LAB [43]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	SAN [47]	5.22	10.39	5.71	5.19	5.49	6.83	5.80
	Coord* [55]	4.76	8.48	4.98	4.65	4.84	5.83	5.49
	DETR† [76]	4.71	7.91	4.99	4.60	4.52	5.73	5.33
	Heatmap* [55]	4.60	7.94	4.85	4.55	4.29	5.44	5.42
	AVS + SAN [88]	4.39	8.42	4.68	4.24	4.37	5.60	4.86
	LUVLi [67]	4.37	7.56	4.77	4.30	4.33	5.29	4.94
	AWing [122]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	SDFL* [52]	4.35	7.42	4.63	4.29	4.22	5.19	5.08
	SDL* [50]	4.21	7.36	4.49	4.12	4.05	**4.98**	4.82
	HIH [69]	**4.18**	7.20	**4.19**	4.45	**3.97**	**5.00**	**4.81**
	ADNet [70]	**4.14**	**6.96**	**4.38**	4.09	4.05	5.06	**4.79**
	SLPT‡	4.20	**7.18**	4.52	**4.07**	4.17	5.01	4.85
	SLPT†	**4.14**	**6.96**	4.45	**4.05**	**4.00**	5.06	**4.79**
FR$_{0.1}$(%)↓	LAB	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	SAN	6.32	27.91	7.01	4.87	6.31	11.28	6.60
	Coord*	5.04	23.31	4.14	3.87	5.83	9.78	7.37
	DETR†	5.00	21.16	5.73	4.44	4.85	9.78	6.08
	Heatmap*	4.64	23.01	3.50	4.72	2.43	8.29	6.34
	AVS + SAN	4.08	18.10	4.46	2.72	4.37	7.74	4.40
	LUVLi	3.12	15.95	3.18	**2.15**	3.40	6.39	**3.23**
	AWing	2.84	13.50	2.23	2.58	2.91	5.98	3.75
	SDFL*	**2.72**	12.88	**1.59**	2.58	2.43	**5.71**	3.62
	SDL*	3.04	15.95	2.86	2.72	**1.45**	**5.29**	4.01
	HIH	2.96	15.03	**1.59**	2.58	**1.46**	6.11	**3.49**
	ADNet	**2.72**	**12.72**	**2.15**	**2.44**	1.94	5.79	3.54
	SLPT‡	3.04	15.95	2.86	**1.86**	3.40	6.25	4.01
	SLPT†	**2.76**	**12.27**	2.23	**1.86**	3.40	5.98	3.88
AUC$_{0.1}$ ↑	LAB	0.532	0.235	0.495	0.543	0.539	0.449	0.463
	SAN	0.536	0.236	0.462	0.555	0.522	0.456	0.493
	Coord*	0.549	0.262	0.524	0.559	0.555	0.472	0.491
	DETR†	0.552	0.285	0.520	0.558	0.563	0.471	0.497
	Heatmap*	0.524	0.251	0.510	0.533	0.545	0.459	0.452
	AVS + SAN	0.591	0.311	0.549	**0.609**	0.581	0.516	**0.551**
	LUVLi	0.557	0.310	0.549	0.584	0.588	0.505	0.525
	AWing	0.572	0.312	0.515	0.578	0.572	0.502	0.512
	SDFL*	0.576	0.315	0.550	0.585	0.583	0.504	0.515
	SDL*	0.589	0.315	0.566	0.595	**0.604**	0.524	0.533
	HIH	**0.597**	0.342	**0.590**	**0.606**	**0.604**	**0.527**	**0.549**
	ADNet	**0.602**	**0.344**	0.523	0.580	0.601	**0.530**	0.548
	SLPT‡	0.588	0.327	0.563	0.596	0.595	0.514	0.528
	SLPT†	0.595	**0.348**	**0.574**	0.601	**0.605**	0.515	0.535

Table 4.1: Performance comparison of the SLPT and the state-of-the-art methods on WFLW and its subsets. The normalization factor is inter-ocular and the threshold for FR is set to 0.1. Key: [**Best**, **Second Best**, *=HRNetW18C, †=HRNetW18C-lite, ‡=ResNet34]

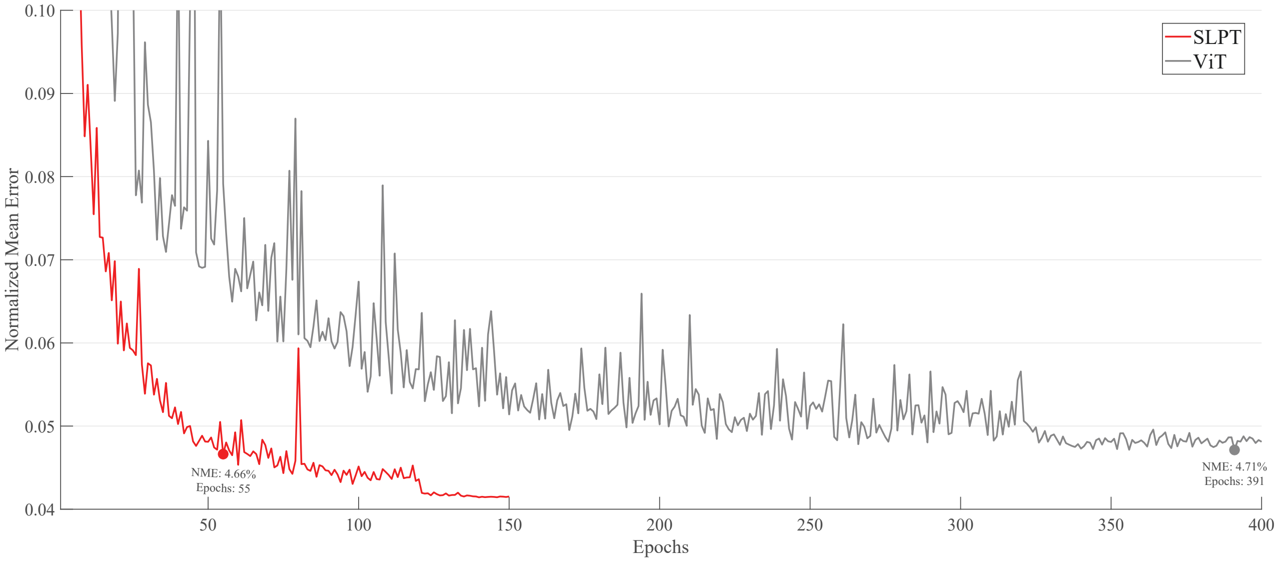


Figure 4.5: Convergence curves of SLPT and DETR on WFLW test set. The learning rate of SLPT is reduced at 120 and 140 epochs; the learning rate of DETR is reduced at 320 and 360 epochs.

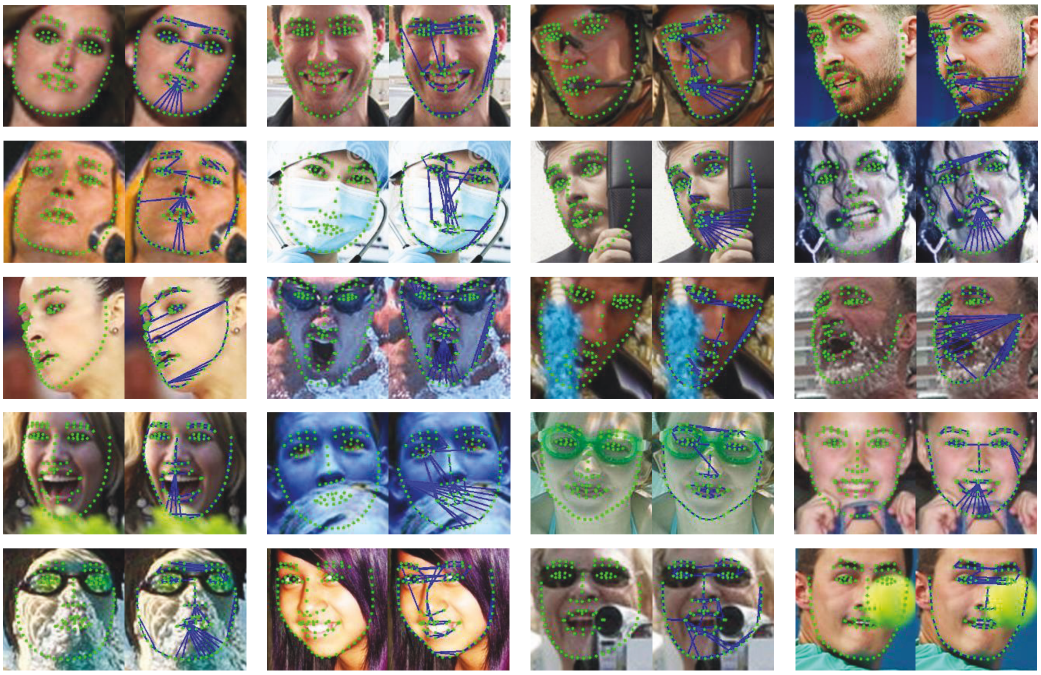


Figure 4.6: Predicted results and learned inherent relations on WFLW. We connect each point to the point with highest cross-attention weight in the first inherent relation layer using line.

4.3.4 Comparison with State-of-the-Art Method

WFLW

As tabulated in Table 4.1, SLPT demonstrates impressive performance. With the increasing of inherent layers, the performance of SLPT can be further improved and outperforms the ADNet (see Table. 4.9). Referring to DETR, we also implement a Transformer based method that employs the full feature map for facial landmark detection. The number of the input tokens is 16×16. Using multi-scale local patches in SLPT for coarse-to-fine facial landmark regression also serves as a form of regularization, which prevents the model from overfitting and accelerates the convergence speed. Therefore, with the same backbone (HRNetW18C-lite), we observe an improvement of 12.10% in NME, and the number of training epoch is $8\times$ less than the DETR, as demonstrated in Fig. 4.5. Moreover, the SLPT also outperforms the coordinate regreesion and heatmap regression methods significantly. Some qualitative results are shown in Fig. 4.6. It is evident that our method could localize the landmarks robustly using the case-dependent inherent relation between landmarks, in particular for face images with blur, profile view and heavy occlusion.

300W

The comparison result is shown in Table 4.2. Compared to the coordinate and heatmap regression methods (HRNetW18C [55]), SLPT still achieves an impressive improvement of 9.69% and 4.52% respectively in NME on the fullset. However, the improvement on 300W is not as significant as WFLW since learning an adaptive inherent relation requires a large number of annotated samples. With limited training samples, the methods with prior knowledge, such as facial boundaries (Awing and ADNet) and affined mean shape (SDL), always achieve better performance. Some qualitative results on 300W are shown in Fig. 4.7.

COFW

We conduct two experiments on COFW for comparsion, the *within*-dataset validation and *cross*-dataset validation. For the *within*-dataset validation, the model is trained with 1,345 images and validated with 507 images on COFW. The inter-ocular and inter-pupil NME of SLPT and the state-of-the-art methods are reported in Table 4.3 respectively. In this experiment, the number of training sample is quite small, which leads to the significant degradation of

Method	Inter-Ocular NME (%) ↓		
	Common	Challenging	Fullset
SAN [47]	3.34	6.60	3.98
Coord* [55]	3.05	5.39	3.51
LAB [43]	2.98	5.19	3.49
DeCaFA [60]	2.93	5.26	3.39
HIH [69]	2.93	5.00	3.33
Heatmap* [55]	2.87	5.15	3.32
SDFL* [52]	2.88	4.93	3.28
HG-HSLE [64]	2.85	5.03	3.28
LUVLi [67]	2.76	5.16	3.23
AWing [61]	2.72	**4.53**	3.07
SDL* [50]	**2.62**	4.77	**3.04**
ADNet [70]	**2.53**	**4.58**	**2.93**
SLPT‡	2.78	4.93	3.20
SLPT†	2.75	4.90	3.17

Table 4.2: Performance comparison for SLPT and the state-of-the-art methods on 300W common subset, challenging subset and fullset. Key: [**Best**, **Second Best**, *=HRNetW18C, †=HRNetW18C-lite, ‡=ResNet34]

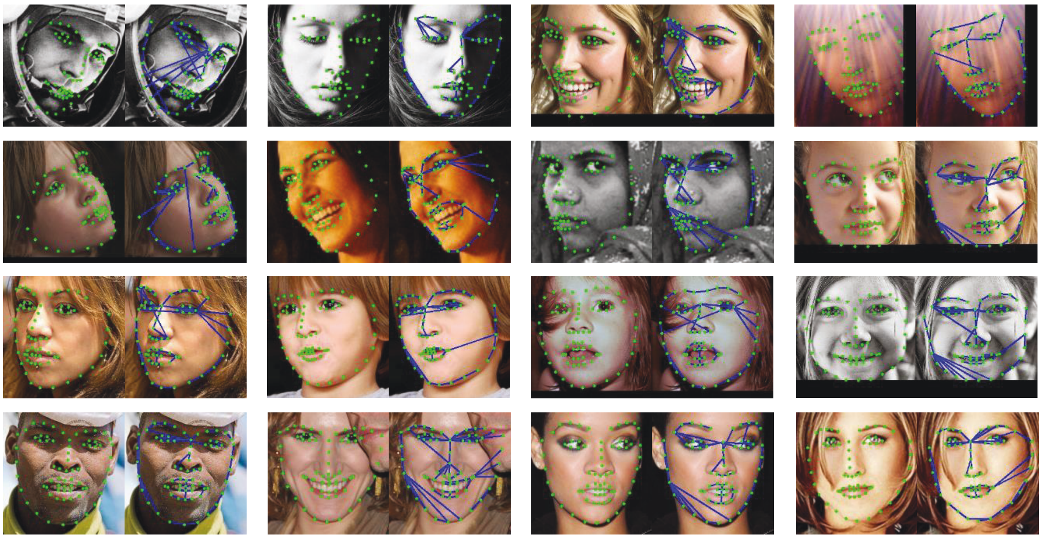


Figure 4.7: Predicted results and learned inherent relations on 300W. We connect each point to the point with highest cross-attention weight in the first inherent relation layer using **line**.

Method	Inter-Ocular		Inter-Pupil	
	NME(%)↓	FR$_{0.1}$(%)↓	NME(%)↓	FR$_{0.1}$(%)↓
DAC-CSR [15]	6.03	4.73	-	-
LAB [43]	3.92	0.39	-	-
Coord* [55]	3.73	0.39	-	-
SDFL* [52]	3.63	**0.00**	-	-
Heatmap* [55]	3.45	**0.20**	-	-
Human [8]	-	-	5.60	-
TCDCN [45]	-	-	8.05	-
Wing [44]	-	-	5.44	3.75
DCFE [38]	-	-	5.27	7.29
AWing [61]	-	-	4.94	**0.99**
ADNet [70]	-	-	**4.68**	**0.59**
SLPT‡	**3.36**	0.59	4.85	1.18
SLPT†	**3.32**	**0.00**	**4.79**	1.18

Table 4.3: NME and FR$_{0.1}$ comparisons under Inter-Ocular normalization and Inter-Pupil normalization on *within*-dataset validation. Key: [**Best**, **Second Best**, *=HRNetW18C, †=HRNetW18C-lite, ‡=ResNet34]

Method	Inter-Pupil NME(%)↓	FR$_{0.1}$(%)↓
CFSS [12]	6.28	9.07
ODN [48]	5.30	-
AVS+SAN [88]	4.43	2.82
LAB [43]	4.62	2.17
SDL* [50]	4.22	**0.39**
SDFL* [52]	4.18	**0.00**
SLPT‡	**4.11**	0.59
SLPT†	**4.10**	0.59

Table 4.4: Inter-ocular NME and FR$_{0.1}$ comparisons on 300W-COFW68 *cross*-dataset evaluation. Key: [**Best**, **Second Best**, *=HRNetW18C, †=HRNetW18C-lite, ‡=ResNet34]

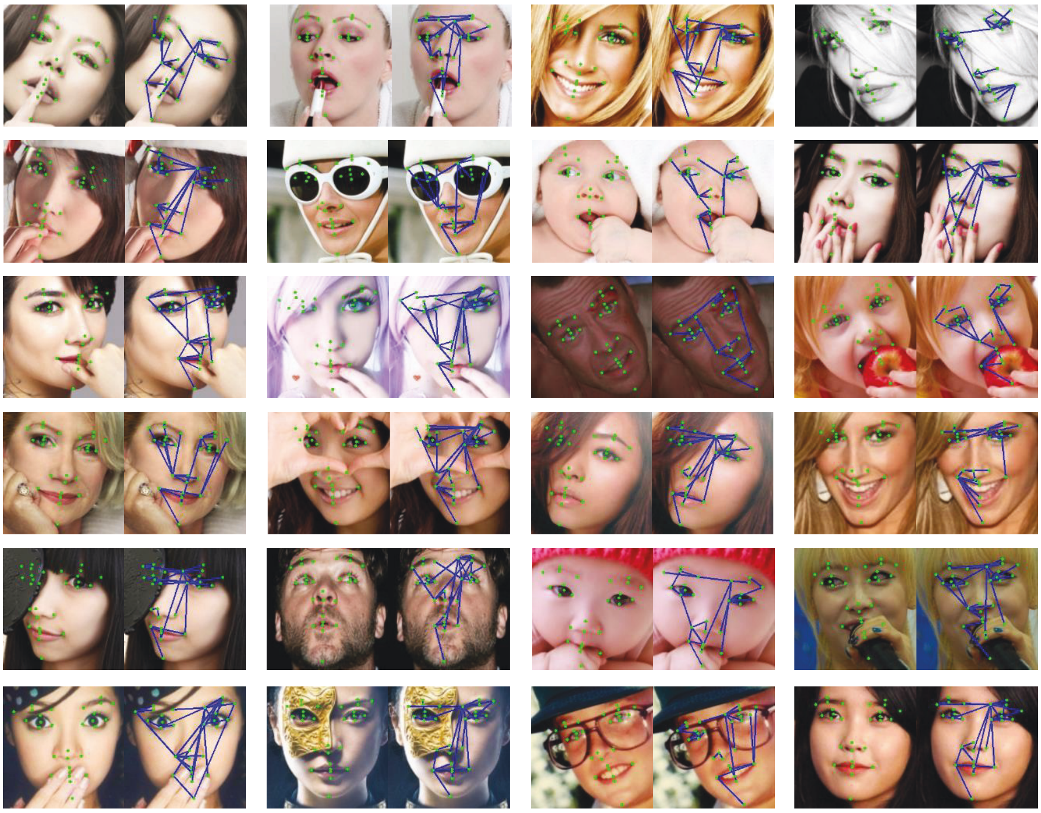


Figure 4.8: Predicted results and learned inherent relations on COFW. We connect each point to the point with highest cross-attention weight in the first inherent relation layer using line.

the coordinate regression methods, such as SDFL, LAB. Nevertheless, SLPT still maintains excellent performance and yields the second best performance. It improves the metric by 3.77% and 11.00% in NME over the heatmap regression and coordinate regression methods respectively. Some qualitative results on COFW are shown in Fig. 4.8.

For the *cross*-dataset validation, the training set includes the complete 300W dataset (3,837 images) and the test set is COFW68 (507 images with 68 landmark annotation). Most samples of COFW68 are under heavy occlusion. The inter-ocular NME and FR of SLPT and the state-of-the-art methods are reported in Table 4.4. Compared to the methods based on GCN (SDL and SDFL), the SLPT (HRNet) achieves impressive result, as low as 4.10% in NME. The result illustrates that the adaptive inherent relation of SLPT works better than the fixed adjacency matrix of GCN for robust facial landmark detection, especially for the condition of heavy occlusion.

Model	Intermediate Stage											
	1st stage			2rd stage			3rd stage			4th stage		
	NME	$FR_{0.1}$	$AUC_{0.1}$	NME	$FR_{0.1}$	$AUC_{0.1}$	NME	$FR_{0.1}$	$AUC_{0.1}$	NME	$FR_{0.1}$	$AUC_{0.1}$
Model† with 1 stage	4.79%	5.08%	0.538	-	-	-	-	-	-	-	-	-
Model† with 2 stages	4.52%	4.24%	0.563	4.27%	3.40%	0.585	-	-	-	-	-	-
Model† with 3 stages	4.38%	3.60%	0.574	4.16%	2.80%	0.594	**4.14%**	**2.76%**	**0.595**	-	-	-
Model† with 4 stages	4.47%	4.00%	0.567	4.26%	3.40%	0.586	4.24%	3.36%	0.588	4.24%	3.32%	0.587

Table 4.5: Performance comparison of the SLPT with different number of coarse-to-fine stages on WFLW. Key: [**Best**, †=HRNetW18C-lite]

Method	MSA	MCA	NME	FR$_{0.1}$	AUC$_{0.1}$
Model† 1	w/o	w/o	4.48%	4.32%	0.566
Model† 2	w/	w/o	4.20%	3.08%	0.590
Model† 3	w/o	w/	4.17%	2.84%	0.593
Model† 4	w/	w/	**4.14%**	**2.76%**	**0.595**

Table 4.6: NME(↓), FR$_{0.1}$(↓) and AUC$_{0.1}$(↑) with/without Encoder and Decoder. Key: [**Best**, †=HRNetW18C-lite]

4.3.5 Ablation Study

Evaluation on different numbers of coarse-to-fine stages

To explore the contribution of the coarse-to-fine framework, we train the SLPT with different number of coarse-to-fine stages on the WFLW dataset. The NME, AUC$_{0.1}$ and FR$_{0.1}$ of each intermediate stage and the final stage are shown in Table 4.5. Compared to the model with only one stage, the local patches in multi-stages model evolve into a pyramidal form, which improves the performance of intermediate stages and final stage significantly. When the stage increases from 1 to 3, the NME of the first stage decreases dramatically from 4.79% to 4.38%. When the number of stages is more than 3, the performance converges and additional stages cannot bring any improvement to the model.

Evaluation on MSA and MCA block

To explore the influence of *query-query* inter relation and *representation-query* inter relation created by MSA and MCA blocks, we implement four different models with/without MSA and MCA, ranging from 1 to 4. For the models without MCA block, we utilize the landmark representations as the queries input. The performance of the four models are tabulated in Table 4.6. Without MSA and MCA, each landmark is regressed merely based on the feature of the supporting patches in model 1. Nevertheless, it still outperforms other coordinate regression methods because of the coarse-to-fine framework. When self-attention or cross-attention is introduced into the model, the performance is boosted significantly, reaching at 4.20% and 4.17% respectively in terms of NME. Moreover, the self-attention and cross-attention can be combined to improve the performance of model further.

Method	NME	$\text{FR}_{0.1}$	$\text{AUC}_{0.1}$
w/o structure encoding†	4.16%	2.84%	0.593
w structure encoding†	**4.14%**	**2.76%**	**0.595**

Table 4.7: NME(↓), $\text{FR}_{0.1}$(↓) and $\text{AUC}_{0.1}$(↑) with/without structure encoding. Key: [**Best**, , †=HRNetW18C-lite]

Patch size	NME(%)	$\text{FR}_{0.1}$(%)	$\text{AUC}_{0.1}$
5×5	4.17%	**2.76%**	0.593
7×7	**4.14%**	**2.76%**	**0.595**
9×9	4.16%	2.84%	0.594

Table 4.8: NME(↓), $\text{FR}_{0.1}$(↓) and $\text{AUC}_{0.1}$(↑) with different patch sizes $K \times K$ on WFLW test set. Key: [**Best**]

Evaluation on structure encoding

We implement two models with/without structure encoding to explore the influence of structural information. With structural information, the performance of SLPT is improved, as shown in Table 4.7.

Evaluation on the input patch size

Each local patch is resized to $K \times K$ and then projected into a vector by a CNN layer with $K \times K$ kernel size. In this section, we explore the influence of the patch size on WFLW test set, as tabulated in Table 4.8. Compared to 7×7 patches, the 5×5 patches lose more information because of the lower resolution, which leads to degradation of the performance. When the patch size is extended from 7×7 to 9×9, the parameters of the CNN layer is doubled, which leads to the overfitting on the training set. Therefore, we can also observe a slight degradation with 9×9 patch size, from 4.14% to 4.16% in NME.

Evaluation on the number of inherent relation layers

Table 4.9 demonstrates the influence of inherent relation layer number. The performance of SLPT relies on the inherent relation layer heavily. When the number of inherent relation layers increases from 2 to 12, We can observe a significant improvement, from 4.19% to 4.12%

Layer number	NME(%)	$\mathrm{FR}_{0.1}$(%)	$\mathrm{AUC}_{0.1}$
2	4.19%	2.88%	0.592
4	4.17%	2.84%	0.593
6	4.14%	2.76%	0.595
12	**4.12%**	**2.72%**	**0.596**

Table 4.9: NME(↓), $\mathrm{FR}_{0.1}$(↓) and $\mathrm{AUC}_{0.1}$(↑) with different patch sizes $K \times K$ on WFLW test set. Key: [**Best**]

Method	FLOPs(G)	Params(M)
HRNet* [55]	4.75	9.66
LAB [43]	18.85	12.29
AVS + SAN [88]	33.87	35.02
AWing [61]	26.8	24.15
DETR† (98 landmarks) [76]	4.26	11.00
DETR† (68 landmarks) [76]	4.06	11.00
DETR† (29 landmarks) [76]	3.80	10.99
SLPT† (98 landmarks)	6.12	13.19
SLPT† (68 landmarks)	5.17	13.18
SLPT† (29 landmarks)	3.99	13.16

Table 4.10: Computational complexity and parameters of SLPT and SOTA methods. All SLPT models are implemented with three coarse-to-fine stages. Key: [*=HRNetW18C, †=HRNetW18C-lite]]

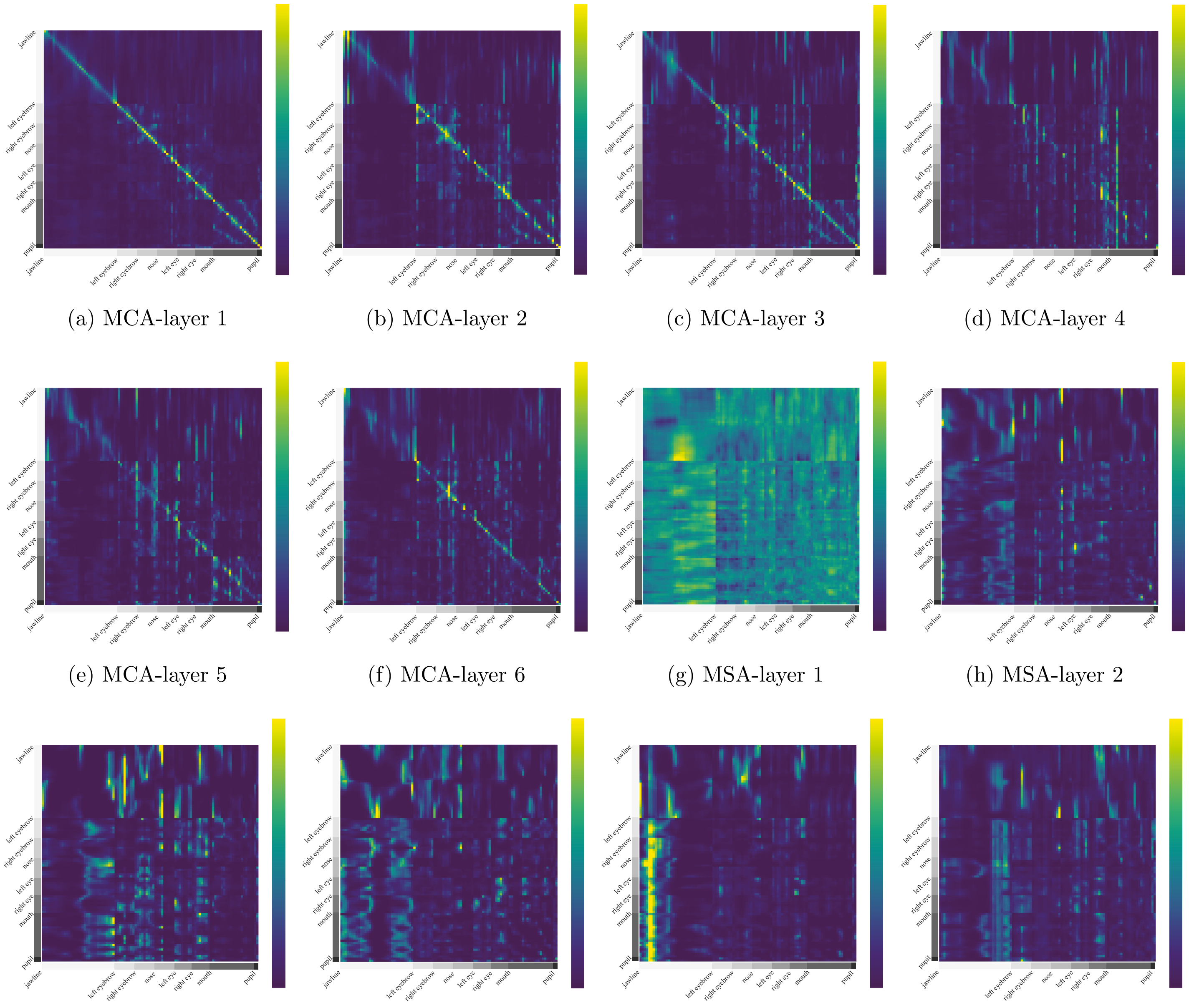
in NME. Nevertheless, too many inherent relation layers also increase the parameters and computational complexity dramatically. Considering the real-time capability, we choose the model with 6 inherent relation layers as the optimal model.

Evaluation on computational complexity

The computational complexity and parameters of SLPT and other SOTA methods are shown in Table 4.10. The computational complexity of SLPT is only 1/8 to 1/5 FLOPs of the previous SOTA methods (AVS and AWing), demonstrating that learning inherent relation is more efficient than other methods. Although SLPT runs three times for coarse-to-fine localization, patch embedding and linear interpolation procedures, we do not observe a significant increasing of computational complexity, especially for 29 landmarks, because the sparse local patches lead to less tokens.

4.3.6 Visualization

We calculate the mean attention weight of each MCA and MSA block on the WFLW test set, as shown in Fig. 4.9. We find out that the MCA block tends to aggregate the representation of the supporting and neighboring patches to generate the local feature, while MSA block tends to pay attention to the landmarks with a long distance to create the global feature. That is why the MCA block can incorporate with the MSA block for better performance.



(a) MCA-layer 1

(b) MCA-layer 2

(c) MCA-layer 3

(d) MCA-layer 4

(e) MCA-layer 5

(f) MCA-layer 6

(g) MSA-layer 1

(h) MSA-layer 2

(i) MSA-layer 3

(j) MSA-layer 4

(k) MSA-layer 5

(l) MSA-layer 6

Figure 4.9: The statistical attention interactions of MCA and MSA in the final stage on the WFLW test set. Each row indicates the attention weight of the landmark.

4.3.7 Limitation

The learning of adaptive inherent relations requires a certain number of annotated faces. For COFW, which consists of only 1,345 training samples, the inherent relations learned by SLPT are not as reliable as those learned on other datasets, consequently resulting in overfitting, as shown in Fig. 4.10. Therefore, transferring knowledge learned from large-scale datasets to maintain robustness on small-scale datasets is a potential direction for future work.

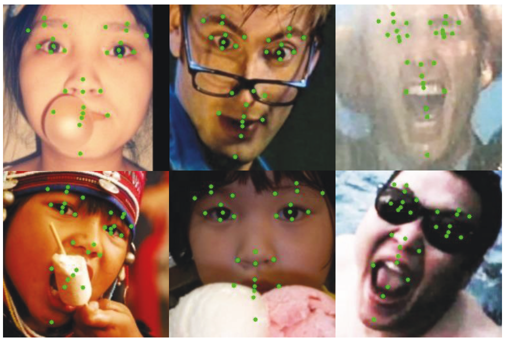


Figure 4.10: Some bad cases predicted by SLPT on COFW dataset. The green points represent the predicted landmarks.

4.4 Conclusion

In this chapter, we find out that the case-dependent inherent relation between landmarks is significant to the performance of facial landmark detection, especially for the cases with heavy occlusion. However, this is always ignored by the most state-of-the-art methods. To address the problem, we propose a sparse local patch transformer to learn a *query-query* and a *representation-query* relation for better performance. Moreover, a coarse-to-fine framework that enables the local patches to evolve into pyramidal former is proposed to further improve the performance of SLPT. With the adaptive inherent relation learned by SLPT, our method achieves robust facial landmark detection, especially for the faces with blur, heavy occlusion and profile view, and outperforms the state-of-the-art methods significantly with much less computational complexity. Ablation studies verify the effectiveness of the proposed method. In future work, the inherent relation learning will be studied further and extended to other tasks.

Chapter 5

Robust and Reliable Facial Landmark Detection for Heavy Occluded Faces via Uncertainty Estimation

5.1 Introduction

The performance of facial landmark detection is hindered by a common limitation in existing methods: they assume a constant variance in the probability distribution across all landmarks. For instance, heatmap regression methods generate heatmaps with a fixed variance as the learning objective; coordinate regression methods utilize L1 or L2 loss to constrain the model learning; patch-based regression methods [17], [46], [19] set the local patch of each landmark to a fixed size. However, through observation, we find that the easily identified landmark results in a smaller variance and the landmarks with high uncertainty always have a larger variance. Therefore, the assumption does not usually hold and using the patch with a fixed size may lead to performance degradation to facial landmark detection, especially for the cases with heavy occlusion. Unfortunately, the existing patch-based regression methods have not solved the problem yet.

In this chapter, we propose a novel framework, Dynamic Sparse Local Patch Transformer (DSLPT) to solve the this problems. As shown in Fig. 5.1, similar to SLPT, DSLPT first crops a local patch for each landmark according to an initial mean face calculated from training samples [42] and then embeds it into a vector. Each vector can be regarded as a rough representation of the corresponding landmark. Then, the landmark representations are added with the proposed structure encoding to retain the structure information of a regular face. Subsequently, a series of landmark queries adaptively aggregate the representations based on the attention mechanism, which enables DSLPT to learn a case dependent inherent relation.

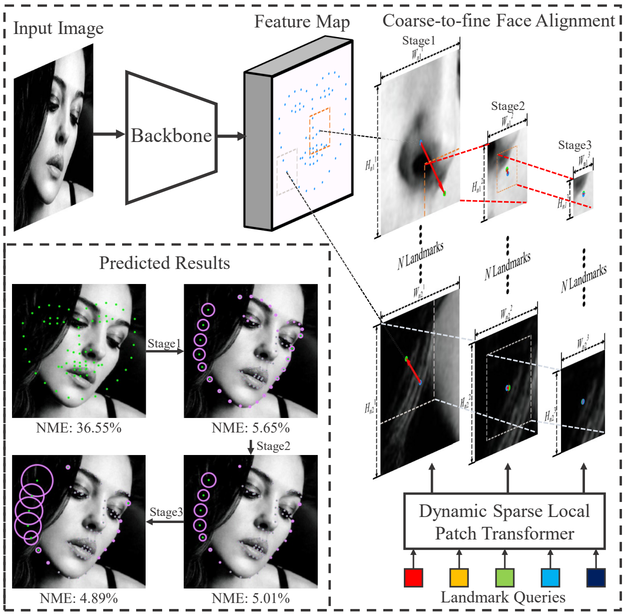


Figure 5.1: Proposed coarse-to-fine framework leverages the dynamic patches for robust face alignment. The initial local patches are cropped according to a meanface. Then, the size and position of each patch are adjusted dynamically according to the location and uncertainty predicted in the previous stage for fine-grained representation. The **blue** and **green** point indicate the initial and predicted landmark respectively. The uncertainty of each landmark is shown by **pink** circle.

We also introduce a coarse-to-fine framework to incorporate with DSLPT, as SLPT, to enable a rough predicted result to converge to the target facial landmarks gradually. However, instead of using the patch with a fixed size like other patch-based regression methods, the position and size of each patch in this coarse-to-fine framework are determined by the predicted position and uncertainty of the corresponding landmark in the previous stage. A larger patch size commonly leads to more contextual information but lower feature resolution, and vice versa. The dynamic patch applies a relatively large patch size to the landmarks with high uncertainty, such as heavily occluded landmarks, for more contextual information and a relatively small patch size to the landmarks with low uncertainty for high feature resolution. This dynamic patch size of DSLPT enables the model to obtain the advantages of large patch size and small

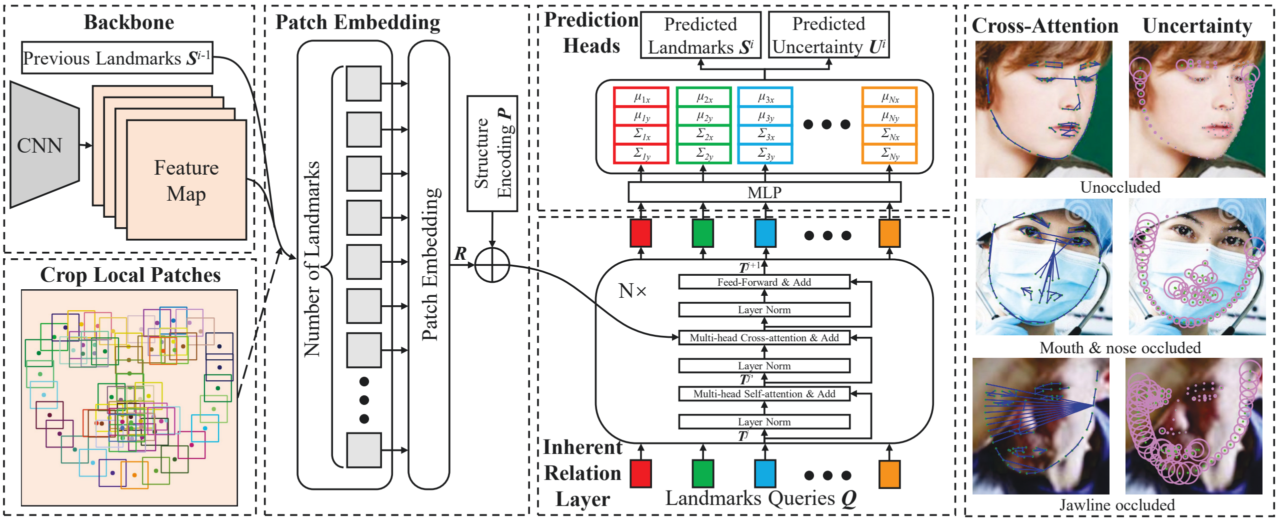


Figure 5.2: A pipeline of the DSLPT.

patch size simultaneously.

5.2 Method

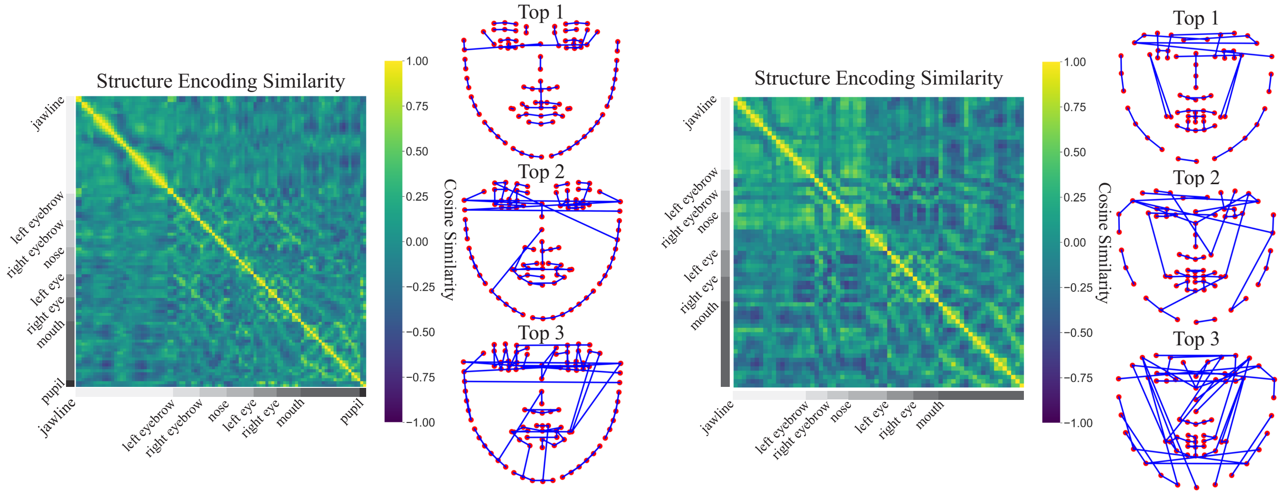
The proposed method mainly consists of three parts: the Dynamic Sparse Local Patch Transformer for adaptive inherent relation learning, the distribution estimation part for patch size and localization adjustment, and the coarse-to-fine framework for fine-grained result. Each of these parts plays an important role in facial landmark detection and we will describe them in the following subsections.

5.2.1 Dynamic Sparse Local Patch Transformer

As shown in Fig. 5.2, the DSLPT consists of three complementary components: patch embedding & structure encoding, inherent relation layers and prediction heads.

Similar to SLPT, DSLPT crops the sparse local patches with size (W_n^i, H_n^i) (i and n are the index of stage and landmark respectively) from the feature map $\boldsymbol{F}$ according to the landmarks $\boldsymbol{S}^{i-1}$ predicted in the previous stage ($\boldsymbol{S}^0$ is a mean shape calculated from the training set). Each patch can be regarded as the supporting patch of the corresponding landmark. Then, the patches with different sizes are resized to (P_w, P_h) by linear interpolation and embedded into a d-dimension representation by a CNN layer with a kernel size of (P_w, P_h).

Human face has a regular shape and the relative position of the landmarks in the shape is defined



(a) 98 landmarks

(b) 68 landmarks

Figure 5.3: **Left**: cosine similarity of the structure encodings learned from 98 landmarks and 68 landmarks datasets. **Right**: to better visualize, we connect each landmark to the landmark with the highest, second highest and third highest similarity respectively.

as the structure information in many works [52], [50]. Nevertheless, the structure information is missing in the sparse local patches. ViT retains the spatial information of patches with a 1D or 2D position encoding generated by cosine & sine function [121]. Unfortunately, face shape is hard to be represented by a 1D or 2D encoding. To retain the structure information, we propose the structure encodings $\boldsymbol{P} \in \mathbb{R}^{N\times d}$ (N is the landmark number and d is the dimension of landmark representation), which are learnable vectors and updated by back propagation. We then add them to the landmark representations. The neighboring and symmetrical patches commonly have high similarity in appearance, and the principle can be used for describing the face structure. The structure encodings learn the similarity during the training procedure. As a result, they encode the relative position of facial landmarks into the cosine similarity and further retain the structure information. As shown in Fig. 5.3, the structure encoding tends to have high cosine similarity with the structure encoding of the neighboring and symmetrical landmarks. Besides, the cosine similarity map of 98 landmarks is similar to the adjacency matrix generated by prior knowledge in [52], which means the structure information learned by unsupervised learning in DSLPT is quite close to human prior knowledge.

Inherent Relation Layer

Similar to SLPT, each inherent relation layer mainly consists of three blocks: a multi-head self-attention (MSA) block, a multi-head cross-attention (MCA) block and a multilayer perceptron

(MLP) block. Moreover, an additional Layernorm (LN) is applied before every block. The MSA block learns an inter *query-query* relation based on the self-attention mechanism. The self-attention weight of the m-th head $\boldsymbol{A}_m$ can be formulated as:

$$\boldsymbol{A}_m = softmax\left(\frac{\left(\boldsymbol{T}_m^j + \boldsymbol{Q}_m\right)\boldsymbol{W}_m^q\left(\left(\boldsymbol{T}_m^j + \boldsymbol{Q}_m\right)\boldsymbol{W}_m^k\right)^T}{\sqrt{d_m}}\right), \tag{5.1}$$

where j and m are the index of inherent relation layers and attention heads respectively. M is the number of attention heads and the input dimension of m-th head d_m can be written as $d_m = d/M$. $\boldsymbol{W}_m^q \in \mathbb{R}^{N \times d_m}$ and $\boldsymbol{W}_m^k \in \mathbb{R}^{N \times d_m}$ are the weights of FC layers. $\boldsymbol{T}_m^j \in \mathbb{R}^{N \times d_m}$ is the input of the m-th head in the j-th MSA block ($\boldsymbol{T}_m^j$ is a zero matrix in the first layer). The output of the MSA block can be written as:

$$MSA\left(\boldsymbol{T}^j\right) = \left[\boldsymbol{A}_1\boldsymbol{T}_1^j\boldsymbol{W}_1^v; ...; \boldsymbol{A}_M\boldsymbol{T}_M^j\boldsymbol{W}_M^v\right]\boldsymbol{W}^p, \tag{5.2}$$

where $\boldsymbol{W}_m^v \in \mathbb{R}^{N \times d_m}$ and $\boldsymbol{W}^p \in \mathbb{R}^{N \times d}$ are the weights of FC layers. Subsequently, the MCA block aggregates the landmark representations by an inter *representation-query* relation. As shown in the right of Fig. 5.2, we connect each landmark to the landmark with the highest cross-attention weight in the first inherent relation layer. The model tends to localize the occluded landmark according to the easily identified landmarks. As for other landmarks, their localization accuracy can be further improved with the representation of neighboring landmarks. The cross-attention weight of m-th head $\boldsymbol{A}'_m$ can be formulated as:

$$\boldsymbol{A}'_m = softmax\left(\frac{\left(\boldsymbol{T}_m^{j\prime} + \boldsymbol{Q}_m\right)\boldsymbol{W}_m^{q\prime}\left(\left(\boldsymbol{R}_m + \boldsymbol{P}_m\right)\boldsymbol{W}_m^{k\prime}\right)^T}{\sqrt{d_m}}\right), \tag{5.3}$$

where $\boldsymbol{T}_m^{j\prime} \in \mathbb{R}^{N \times d_m}$ is the input of the m-th head in the j-th MCA block; $\boldsymbol{R}_m \in \mathbb{R}^{N \times d_m}$ and $\boldsymbol{P}_m \in \mathbb{R}^{N \times d_m}$ are the layer representations and structure encoding respectively in the m-th head; $\boldsymbol{W}_m^{q\prime} \in \mathbb{R}^{N \times d_m}$ and $\boldsymbol{W}_m^{k\prime} \in \mathbb{R}^{N \times d_m}$ are the weights of FC layers. The output of MCA block can be written as:

$$MCA\left(\boldsymbol{T}^{j\prime}\right) = \left[\boldsymbol{A}'_1\boldsymbol{T}_1^{j\prime}\boldsymbol{W}_1^{v\prime}; ...; \boldsymbol{A}'_M\boldsymbol{T}_M^{j\prime}\boldsymbol{W}_M^{v\prime}\right]\boldsymbol{W}'_P, \tag{5.4}$$

where $\boldsymbol{W}_m^{v\prime} \in \mathbb{R}^{N \times d_m}$ and $\boldsymbol{W}^{p\prime} \in \mathbb{R}^{N \times d}$ are the weights of FC layers.

Similar to SLPT, the dynamic sparse local patches significantly decrease the token number of the MCA blocks compared to dividing the feature map in a dense grid manner adopted in other Transformers [73], [76], [77]. As a result, the sparse patch leads to much lower computational complexity.

Prediction Heads

The heads predict the parameters of a Gaussian or Laplace distribution $(\mu^i_{x_n}, \Sigma^i_{x_n}, \mu^i_{y_n}, \Sigma^i_{y_n})$ for each landmark, where $\mu^i_{x_n}$ and $\mu^i_{y_n}$ are the distribution mean of n-th landmark in i-th stage on the X axis and Y axis respectively. $\Sigma^i_{x_n}$ and $\Sigma^i_{y_n}$ are the distribution variances. Note that the probability distribution is predicted in patch coordinate system (the origin is set to the top left corner of the patch and the patch size (W^i_n, H^i_n) is normalized in $[0,1]$). Therefore, DSLPT does not require positional encoding to retain a global spatial information. The global coordinate (x^i_n, y^i_n) and uncertainty $(U^i_{x_n}, U^i_{y_n})$ of each landmark can be calculated as follows:

$$\begin{aligned} x^i_n &= x^i_{\mathrm{lt}_n} + W^i_n \mu^i_{n_x}, \\ y^i_n &= y^i_{\mathrm{lt}_n} + H^i_n \mu^i_{n_y}, \end{aligned} \tag{5.5}$$

$$\begin{aligned} U^i_{x_n} &= W^i_n \Sigma^i_{x_n}, \\ U^i_{y_n} &= H^i_n \Sigma^i_{y_n}, \end{aligned} \tag{5.6}$$

where $(x^i_{\mathrm{lt}_n}, y^i_{\mathrm{lt}_n})$ is the global coordinate of the top left point of the n-th patch in the i-th stage, and (W^i_n, H^i_n) is the corresponding patch size.

5.2.2 Distribution Estimation

Maximum Log-likelihood Estimation

The L1 loss and L2 loss, which are widely used in facial landmark detection [44], [52], is a degenerated case of probability distribution estimation. We assume the probability distribution of each landmark on the X axis and Y axis is a Gaussian distribution respectively. Then, the density function of each landmark on the X and Y axes can be written as:

$$P_\Theta(z|\boldsymbol{I}) = \frac{1}{\Sigma\sqrt{2\pi}} \exp(-\frac{1}{2}\left(\frac{z-\mu}{\Sigma}\right)^2), \tag{5.7}$$

where Θ is the parameters of the model and $\boldsymbol{I}$ is the input image. To estimate the distribution, the model maximizes the likelihood of the annotated label μ^g with the negative log-likelihood function, which can be formulated as:

$$\mathcal{L} = -\log P_\Theta(z|\boldsymbol{I})|_{z=\mu^g} \propto \log\Sigma + \frac{(\mu^g - \mu)^2}{2\Sigma^2}, \tag{5.8}$$

As mentioned by [67], if Σ is set to 1 and all landmarks are assumed to be visible, then $\mathcal{L} \propto (\mu^g - \mu)^2$, which degrades to the L2 loss function. Similarly, if we assume the distribution

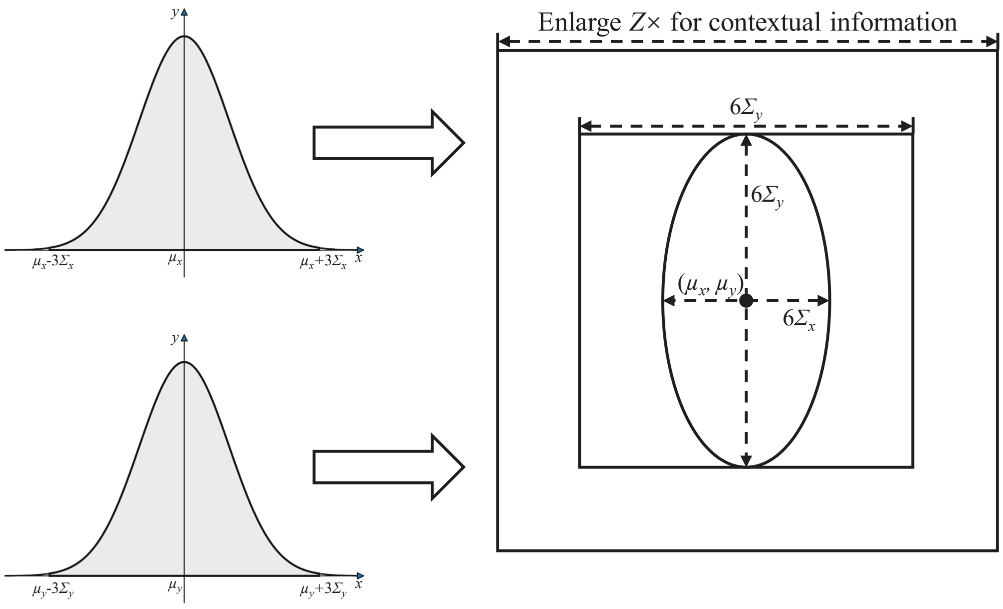


Figure 5.4: An intuitive example of how the predicted probability distribution determines the patch size.

is a 1D Laplace distribution and set the Σ to 1, then $\mathcal{L} \propto |\mu^g - \mu|$, which degrades to the L1 loss function. Obviously, Σ should not be 1 in most conditions. For the landmarks in different conditions, they are commonly with different Σ. Therefore, the DSLPT predicts both μ and Σ with the negative log-likelihood function for a more coherent result.

Previous patch based regression methods [17], [46] predict rough landmarks from the global feature and utilize the patches with a fixed size for fine-grained locating. The landmark with a large Σ is usually under occlusion. To improve the robustness for locating these landmarks, a larger patch size should be applied for more contextual information. However, a larger patch size usually leads to a lower feature resolution, resulting in performance degradation for the landmark with a small Σ. Therefore, we propose the dynamic local patch whose size can adjust according to Σ so that an adaptive patch size is applied to each landmark.

As shown in Fig. 5.4, the DSLPT dynamically adjusts the patch size according to the predicted distribution. It takes $[\mu - 3\Sigma, \mu + 3\Sigma]$ as the confidence interval. For Gaussian and Laplace distribution, the probability that the landmarks are within the interval is more than 95% theoretically. Then, the region of interest (ROI) size of the landmark in the patch coordinate system can be written as $\max(6\Sigma_x, 6\Sigma_y)$. In the global coordinate system, the size can be written as $\max(6W_n^i\Sigma_x, 6H_n^i\Sigma_y)$. Finally, the ROI size is enlarged by Z as the final patch size

(Z is set to 2 in DSLPT) for contextual information. Therefore, the patch size of the following stage can be written as: $W_n^{i+1} = H_n^{i+1} = \max(6ZW_n^i\Sigma_x, 6ZH_n^i\Sigma_y)$

Moreover, the patch size of n-th landmark in (i+1)-th stage should also be limited in [$L_{\text{down}}W_n^i$, $L_{\text{up}}W_n^i$] and [$L_{\text{down}}H_n^i$, $L_{\text{up}}H_n^i$]. Both too small or too large patch size can lead to performance degradation. A too small patch size cannot provide sufficient contextual information for the inherent relation learning though it can ensure a higher feature resolution. And a too large patch size leads to a high patch size variance in the same stage, causing a domain gap.

5.2.3 Coarse-to-fine localization

Inherent relation learning heavily relies on accurate landmark representations. Therefore, we incorporate DSLPT with a coarse-to-fine framework so that a rough landmark representation can converge to an optimal one gradually. The training pipeline of the framework is shown in **Algorithm 2** with pseudo-code. In the first stage, DSLPT crops local patches according to a mean face shape to generate a rough representation for each landmark. In the following stages, both position and size of the local patch are determined by the predicted probability distribution of the corresponding landmark for a fine-grained representation. DSLPT takes the output of the last stage as the final result. Despite the variance of local patch size in different stages, the inherent relation keeps consistent for the same sample. Therefore, the DSLP can be shared in each stage for less parameters. Besides, the patches with different scales augment the training data significantly, which enables DSLPT to be trained with very limited samples.

5.2.4 Auxiliary Inherent Relation Loss

Similar to other facial landmark detection methods [56], [57], we apply an auxiliary loss to provide supervision to the intermediate layers for learning a more coherent inherent relation. The output of each inherent relation layer is fed to a Layernorm layer, followed by a shared prediction head to estimate the probability distribution of landmarks. For the prediction results of the intermediate layers, we also apply the negative log-likelihood function to constrain the model learning. Then total loss $\mathcal{L}_t$ can be calculated as follows:

$$\mathcal{L}_t = \sum_{i=1}^{N_S}\sum_{j=1}^{N_I}\sum_{n=1}^{N} \frac{\mathcal{L}\left(\mu_{n_x}^{g_i}, \mu_{jn_x}^{i}\Sigma_{jn_x}^{i}\right) + \mathcal{L}\left(\mu_{n_y}^{g_i}, \mu_{jn_y}^{i}, \Sigma_{jn_y}^{i}\right)}{2}, \tag{5.9}$$

Algorithm 2 Training pipeline of the DSLPT coarse-to-fine framework

```
Require: Input image $\boldsymbol{I}$, initial mean shape $\boldsymbol{S}^0$, backbone $B$, DSLPT $D$, negative log-likelihood function $\mathcal{L}$, Annotated observed landmark mean $\left(\mu^{g}_{x_n}\right)$, the number of stage $N_S$
while the training epoch is less than a specific number do
    Forward $B$ for feature map by $\boldsymbol{F} = B\left(\boldsymbol{I}\right)$;
    Initialize the local patch size $(W^1_n, H^1_n) \leftarrow \left(\frac{W}{4}, \frac{H}{4}\right)$
    for $i \leftarrow 1$ to $N_S$ do
        Crop local pactes with size $(W^i_n, H^i_n)$ according to previous landmarks $\boldsymbol{S}^{i-1}$;
        Resize local patches to $(P_w, P_h)$;
        embed local patches into representations $\boldsymbol{R}$;
        estimate distribution parameters for each landmark by $\left(\mu^i_{n_x}, \mu^i_{n_y}, \Sigma^i_{n_x}, \Sigma^i_{n_y}\right) = D(\boldsymbol{R})$;
        Adjust $(W^{i+1}_n, H^{i+1}_n)$ according to $\left(\Sigma^i_{n_x}, \Sigma^i_{n_y}\right)$;
        Calculate negative log-likelihood on X and Y axes for each landmark;
    end for
    Minimize $\sum_{i=1}^{N_S} \sum_{n=1}^{N} \frac{\mathcal{L}\left(\mu^{g_i}_{n_x}, \mu^i_{n_x}, \Sigma^i_{n_x}\right) + \mathcal{L}\left(\mu^{g_i}_{n_y}, \mu^i_{n_y}, \Sigma^i_{n_y}\right)}{2}$
end while
```

where $\mathcal{L}$ is the negative log-likelihood function as Eq. 10. $(\mu^{g_i}_{n_x}, \mu^{g_i}_{n_y})$ is the annotated patch coordinate of n-th landmark in i-th stage, $(\mu^i_{jn_x}, \mu^i_{jn_y}, \Sigma^i_{jn_x}, \Sigma^i_{jn_y})$ are the distribution parameters of n-th landmark predicted by j-th head in i-th stage. The $(\mu^{g_i}_{n_x}, \mu^{g_i}_{n_y})$ can be calculated from the annotated global coordinate (x^g_n, y^g_n) as follows:

$$\begin{aligned} \mu^{g_i}_{n_x} &= \frac{x^g_n - x^i_{\mathrm{lt}_n}}{W^i_n}, \\ \mu^{g_i}_{n_y} &= \frac{y^g_n - y^i_{\mathrm{lt}_n}}{H^i_n}. \end{aligned} \tag{5.10}$$

5.3 Experiments

In this section, we evaluate the proposed facial landmark detection method on eight popular benchmarks and carry out extensive experiments to verify the effectiveness. Specifically, we first introduce the eight popular facial landmark detection benchmarks in detail. Then, we describe the metrics for evaluation and the implementation details of the proposed method. Finally, we compare the proposed method to other state-of-the-art methods and conduct extensive ablation studies to study the influence of each component quantitatively.

5.3.1 Benchmarks

- **WFLW** [43]: the WFLW is a very challenging facial landmark detection dataset with significant variations in occlusion, illumination, expression and head pose. It consists of 10,000 faces, including 7,500 for training and 2,500 for testing. Each face is manually labeled with 98 landmarks and rich attributes.

- **300W** [109]: 300W includes 3,148 faces for training and 689 faces for testing. The faces in the training set come from the fullset of AFW [122] and the training subset of HELEN [123] and LFPW [124]. The testing set can be further divided into two subsets: the common subset that includes 554 faces (the test set of HELEN and LFPW) and the challenging subset which consists of 135 faces (the full set of IBUG [109]). Moreover, 300W also annotates additional 600 face images with 68 landmarks to form the 300W-private subset.

- **COFW** [8]: COFW mainly consists of the face with heavy occlusion and profile view, including 1,345 faces for training and 507 faces for testing. Each face in the training set is labeled with 29 landmarks. The annotations of test set have two variants. One variant presents 29 landmark annotations and the other variant is provided with 68 landmarks for each face image (COFW68 [125]).

- **Menpo** [126] [127]: Menpo annotates 11,988 frontal or near frontal faces with 68 landmarks (6,653 faces for training and 5,335 faces for testing) and 4,236 profile faces with 39 landmarks (2,290 faces for training and 1,946 faces for testing).

- **AFLW-19** [128]: AFLW-19 consists of 24,386 faces from AFLW [129], including 20,000 faces for training and 4,836 for testing. It manually annotates each face with 19 landmarks. The testing set has two variants: 1) **Full**: all 4,836 faces for testing; 2) **Front**: 1,314 faces with frontal view are selected from the 4,836 faces for testing.

- **MERL-RAV** [67]: MERL-RAV re-annotates 19,314 faces from AFLW [129] with 68 landmarks manually (15,449 for training and 3,865 for testing). Unlike other datasets, the annotated landmarks of MERL-RAV can be further divided into three categories: unoccluded, externally occluded and self-occluded landmark. Only unoccluded and external occluded landmarks are labeled with location information.

- **Masked 300W** [68]: Masked 300W synthesizes 689 masked faces from the test set of

300W [109]. The average occluded area in Masked 300W is over 50% of the face area.

- **300W-LP & AFLW2000-3D** [40]: 300W-LP synthesizes 122,450 samples from 300W [109] via face profiling. Each sample is annotated with 68 3D landmarks. AFLW2000-3D selects 2,000 faces from AFLW [129] and each face is also labeled with 68 3D landmarks.

5.3.2 Evaluation Metrics

Referring to related work [59], [70], [43], we employ three metrics: **Normalized Mean Error** (NME), **Failure Rate** (FR) and **Area Under the Curve** (AUC) for a fair comparison. The NME is defined as:

$$NME = \frac{1}{N} \sum_{n=1}^{N} \frac{||\,(x_n^g, y_n^g) - (x_n, y_n)||\,}{d_{\text{norm}}}, \tag{5.11}$$

where d_{norm} is the normalized factor. The d_{norm} is the **inter-pupil distance** (the distance between pupil centers) or the **inter-ocular distance** (the distance between outer eye corners) on the WFLW, 300W, Masked 300W and COFW. The d_{norm} is the geometric mean of the annotated bounding box size ($\sqrt{H_{\text{box}} \times W_{\text{box}}}$) or the diagonal of annotated bounding box ($\sqrt{H_{\text{box}}^2 + W_{\text{box}}^2}$) on 300W-private, Menpo, COFW68, AFLW-19, MERL-RAV and AFLW2000-3D, where H_{box} and W_{box} are the height and width respectively of the face bounding box. FR$_\alpha$ indicates the percentage of the testing samples whose NME is higher than a certain threshold α. The AUC is calculated based on the Cumulative Errors Distribution (CED) curve. CED curve indicates a cumulative distribution function $f(\epsilon)$ of the NME and the AUC can be calculated by $\int_0^\alpha f(\epsilon)\, d\epsilon$, where α is the threshold of FR$_\alpha$.

5.3.3 Implementation Details

The proposed facial landmark detection framework is implemented in Pytorch [130], and we employ four different networks as the backbone: ResNet34 [87], ResNet50 [87], HRNetW18C [55], HRNetW18C-lite (the modularized block number in each stage is set to 1). Each backbone is pre-trained on the ImageNet dataset [131] as the related works [55], [50]. For ResNet34 and ResNet50, we employ multi-level feature maps for facial landmark detection, as Fig. 4.4. Supposing the feature map size in the k-th CNN stage is $(H_{\text{stage}k}, W_{\text{stage}k}, d_{\text{stage}k})$, the initial patch size (H_{nk}^1, W_{nk}^1) is $\left(\frac{H_{\text{stage}k}}{4}, \frac{W_{\text{stage}k}}{4}\right)$. The patch size of following stages (H_{nk}^i, W_{nk}^i) is calculated using **Algorithm 2**. For HRNetW18C and HRNet18C-lite, we only utilize a single level feature map following the heatmap regression method [55].

We employ AdamW [132] as the optimizer, and the model is trained for 100 epochs with a batch size of 16 (64 for the model initialized from scratch). The initial learning rate is set to 0.0005 for HRNetW18C and HRNetW18C-lite and is set to 0.0004 for ResNet34 and ResNet50. Moreover, the learning rate is reduced by 1/10 at epoch 80 and 90. Each face image is cropped and resized to 256×256 as the input. For the training samples, we apply augmentation techniques, including random horizontal flipping (50%), shearing (33%), gray (20%), occlusion (50%), brightness adjustment ($50\%, \pm 0.3$), rotation ($\pm 30°$), translation ($\pm 10px$), scaling ($\pm 5\%$). **Without specifications**, the size of the resized local patch (P_w, P_h) is set to $(7, 7)$; the number of stage N_S is set to 3; the number of inherent relation layer N_I is set to 6; the up threshold L_{up} and down threshold L_{down} of dynamic patch size are set to 0.7 and 0.5 respectively; the probability distribution of each landmark is assumed as a Gaussian distribution.

5.3.4 Comparison with State-of-the-art Methods

To demonstrate the effectiveness of DSLPT quantitatively, we carry out eight experiments on eight popular benchmarks and compare the performance of the proposed DSLPT with the state-of-the-art methods.

WFLW

The performance of DSLPT and other state-of-the-art methods on WFLW are reported in Table 5.1 and Table 5.2. Compared to SLPT, the dynamic patch further improves the performance, especially on the occlusion and illumination subset, yielding the best performance in NME and AUC. The results illustrate that the dynamic patch significantly improves the locating accuracy for the cases with high uncertainty. In term of computational complexity, DSLPT only uses one additional FC layer to predict the variance of landmark probability distribution. Besides, we also optimize the bilinear interpolation procedure with a more efficient implementation. Therefore, with HRNetW18C-lite as the backbone, the computational complexity of DSLPT (**6.06G Flops**) is even lower than SLPT (**6.12G Flops**) slightly. Moreover, we also implement a DETR (ResNet50-DC5 [76]) with 6 encoders and decoders to estimate the probability distribution of each landmark. The token number of the DETR is $\mathbf{16 \times 16}$. Compared to predicting landmark coordinates from dense patches as DETR, the inherent relation learning of DSLPT is more efficient, achieving much better performance with only 98 tokens. We also implement a DSLPT initialized from scratch to study the effectiveness of the pretraining on ImageNet.

Method	Backbone	type	ImageNet pretraining	Flops ↓	Params ↓	Inter-Ocular NME (%) ↓						
						Full	Pose	Exp.	Ill.	Mu.	Occ.	Blur
HRNet [55]	HRNetW18C	heatmap	Y	4.75G	9.66M	4.60	7.94	4.85	4.55	4.29	5.44	5.42
LUVLi [67]	8 DU-Net	heatmap	N	-	-	4.37	7.56	4.77	4.30	4.33	5.29	4.94
AWing [61]	4 Hourglass	heatmap	N	26.8G	24.15M	4.21	7.21	4.46	4.23	4.02	4.99	4.82
HIH [69]	2 Hourglass	heatmap	N	10.38G	14.47M	4.18	7.20	**4.19**	4.45	3.97	5.00	4.81
ADNet [70]	4 Hourglass	heatmap	N	17.04G	13.37M	4.14	6.96	4.38	4.09	4.05	5.06	4.79
SDFL [52]	ResNet34	coordinate	N	-	-	4.55	-	-	-	-	-	-
AV w. SAN [88]	ResNet152	coordinate	Y	33.87G	35.02M	4.39	8.42	4.68	4.24	4.37	5.60	4.86
DETR (R50) [76]	ResNet50	coordinate	Y	10.62G	35.25M	4.32	7.64	4.67	4.24	4.19	5.12	4.90
SDL [50]	HRNetW18C	coordinate	Y	-	-	4.21	7.36	4.49	4.12	4.05	4.98	4.82
SLPT [133]	HRNetW18C-lite	coordinate	Y	6.12G	13.19M	4.14	6.96	4.45	4.05	4.00	5.06	4.79
DSLPT*	ResNet34	coordinate	N	8.04G	31.06M	4.37	7.58	4.76	4.30	4.37	5.33	4.97
DSLPT	ResNet34	coordinate	Y	8.04G	31.06M	4.14	7.13	4.40	4.12	3.98	5.05	4.81
DSLPT	ResNet50	coordinate	Y	8.71G	33.47M	4.11	7.17	4.44	4.06	**3.96**	4.96	**4.78**
DSLPT	HRNetW18C-lite	coordinate	Y	6.06G	13.25M	**4.02**	**6.92**	4.42	**3.95**	3.97	**4.83**	**4.66**
DSLPT	HRNetW18C	coordinate	Y	7.83G	19.35M	**4.01**	**6.87**	**4.29**	**3.99**	**3.86**	**4.79**	**4.66**

Table 5.1: Performance comparisons with heatmap regression and coordinate regression methods on WFLW full set and its subsets. Key: [**Best**, **Second Best**, *=initialized from scratch]

Metric	Method	Full	Pose	Exp.	Ill.	Mu.	Occ.	Blur
$\mathrm{FR}_{0.1}(\%)\downarrow$	HRNet	4.64	23.01	3.50	4.72	2.43	8.29	6.34
	LUVLi	3.12	15.95	3.18	2.15	3.40	6.39	**3.23**
	AWing	**2.04**	**9.20**	**1.27**	**2.01**	**0.97**	**4.21**	**2.72**
	HIH	2.96	15.03	**1.59**	2.58	1.46	6.11	3.49
	ADNet	2.72	12.72	2.15	2.44	1.94	5.79	3.54
	AV w. SAN	4.08	18.10	4.46	2.72	4.37	7.74	4.40
	DETR (R50)	3.60	18.71	3.18	3.30	2.91	5.43	4.53
	SDL	3.04	15.95	2.86	2.72	**1.45**	5.29	4.01
	SLPT (W18C-l)	2.76	**12.27**	2.23	**1.86**	3.40	5.98	3.88
	DSLPT (R34)*	3.64	16.56	3.50	3.58	2.91	8.02	4.91
	DSLPT (R34)	2.72	13.80	1.91	2.87	2.43	5.57	3.62
	DSLPT (R50)	3.08	16.26	3.18	2.29	2.43	5.84	4.27
	DSLPT (W18C-l)	**2.40**	13.19	2.55	**2.01**	2.43	**4.34**	3.62
	DSLPT (W18C)	2.52	13.19	2.23	2.44	**0.97**	4.89	3.49
$\mathrm{AUC}_{0.1}\uparrow$	HRNet	0.524	0.251	0.510	0.533	0.545	0.459	0.452
	LUVLi	0.557	0.310	0.549	0.584	0.588	0.505	0.525
	AWing	0.590	0.334	0.572	0.596	0.602	0.528	0.539
	HIH	0.597	0.342	**0.590**	**0.606**	0.604	0.527	0.549
	ADNet	**0.602**	0.344	0.523	0.580	0.601	0.530	0.548
	AV w. SAN	0.591	0.311	0.549	0.609	0.581	0.516	**0.551**
	DETR (R50)	0.579	0.298	0.548	0.589	0.583	0.510	0.527
	SDL	0.589	0.315	0.566	0.595	0.604	0.524	0.533
	SLPT (W18C-l)	0.595	0.348	0.574	0.601	0.605	0.515	0.535
	DSLPT (R34)*	0.575	0.304	0.544	0.583	0.581	0.496	0.522
	DSLPT (R34)	0.597	0.336	0.569	0.600	0.614	0.519	0.538
	DSLPT (R50)	0.599	0.336	0.573	0.605	0.609	0.524	0.540
	DSLPT (W18C-l)	**0.607**	**0.351**	0.580	**0.616**	**0.616**	**0.534**	**0.550**
	DSLPT (W18C)	**0.607**	**0.353**	**0.586**	**0.614**	**0.623**	**0.535**	0.549

Table 5.2: Performance comparisons with state-of-the-art methods in $\mathrm{FR}_{0.1}$ and $\mathrm{AUC}_{0.1}$ on WFLW. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C=HRNetW18C, W18C-l=HRNetW18C-lite, *=initialized from scratch]

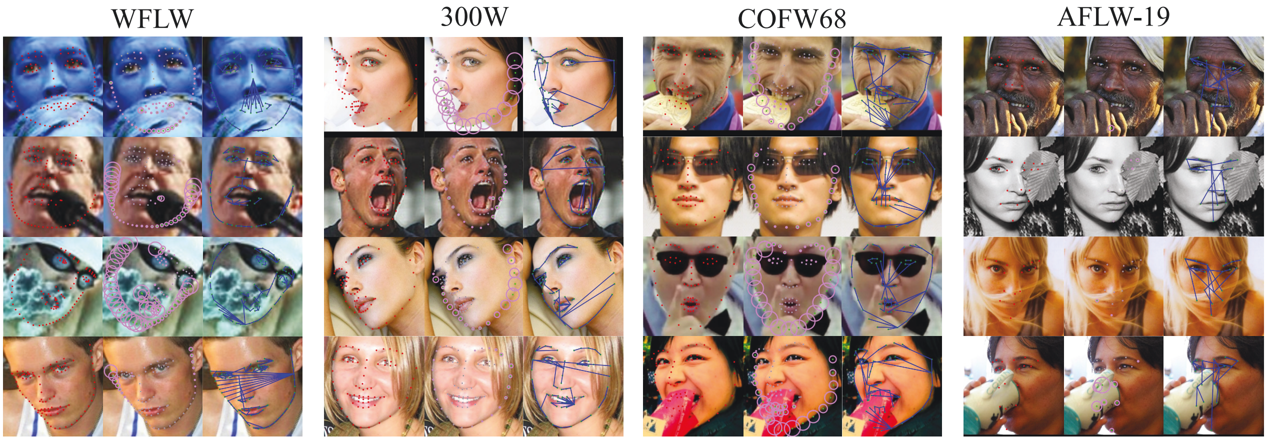


Figure 5.5: Visualized results on WFLW, 300W, COFW68, AFLW-19 testset. The **red** point and **green** point indicate the ground truth and the predicted landmark respectively. The uncertainty of each landmark is shown by **pink** circle. Each landmark is connected with the landmark with highest cross attention weigh by **blue** line.

With a light backbone (ResNet34), the DSLPT initialized from scratch still achieves a comparable performance to LUVLi. Besides, the DSLPT also improves the metric by 4.00% in NME compared to SDFL with the same backbone. Some qualitative results are demonstrated in Fig. 5.5.

300W

As shown in Table 5.3, DSLPT achieves an impressive improvement of 6.55% and 2.24% in NME on the common and challenging subset respectively compared to SLPT. It also demonstrates the effectiveness of the proposed dynamic pacth. Besides, DSLPT is the only coordinate regression method whose NME is smaller than 3.00% on the 300W full set. ADNet and Awing set facial boundary heatmap as an additional regression target to utilize the extra boundary information for better performance. However, DSLPT achieves a comparable performance without any extra information. With ResNet50 as the backbone, DSLPT even improves the metric by 2.89% in NME over ADNet. Therefore, DSLPT sets a remarkable milestone for coordinate regression methods, outperforming the heatmap regression method on 300W for the first time.

COFW

We carry out a *within*-dataset validation on COFW, employing the training subset (1345 images) for training and the test subset of COFW (507 images) for testing. The comparison

Method	Inter-Ocular NME (%) ↓		
	Common	Challenging	Fullset
MHHN [62]	3.18	6.01	3.74
LAB [43]	2.98	5.19	3.49
DeCaFA [60]	2.93	5.26	3.39
HIH [69]	2.93	5.00	3.33
HRNet [55]	2.87	5.15	3.32
SDFL (W18C) [52]	2.88	4.93	3.28
HG-HSLE [64]	2.85	5.03	3.28
DETR (R50) [76]	2.86	4.96	3.27
LUVLi [67]	2.76	5.16	3.23
SHN-GCN [66]	2.73	4.64	3.10
AWing (4HG) [61]	2.72	**4.53**	3.07
SDL [50]	2.62	4.77	3.04
ADNet (R50) [70]	-	-	3.11
ADNet (4HG) [70]	**2.53**	**4.58**	**2.93**
SLPT (W18C-l) [133]	2.75	4.90	3.17
DSLPT (R34)*	2.80	4.94	3.21
DSLPT (R34)	2.62	4.73	3.04
DSLPT (R50)	2.58	4.81	3.02
DSLPT (W18C-l)	**2.57**	4.79	3.00
DSLPT (W18C)	**2.57**	4.69	**2.98**

Table 5.3: Performance comparisons with the state-of-the-art methods on 300W. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C, 4HG=4 hourglass module, *=initialized from scratch]

results are shown in Table 5.4. With very limited number of training samples, this experiment is quite challenging for coordinate regression methods. Many coordinate regression methods, such as SDFL and DETR, degrade significantly. Compared to SLPT, the proposed dynamic patch cannot promise a significant improvement on this condition because it requires more training samples to learn regressing landmarks from different scales of patches. Besides, a deeper backbone commonly leads to more severe overfitting on this condition. Compared to coordinate

Method	Inter-Ocular		Inter-Pupil	
	NME(%)↓	FR(%)↓	NME(%)↓	FR(%)↓
MHHN [62]	4.95	1.78	-	-
LAB [43]	3.92	0.39	-	-
SDFL (W18C) [52]	3.63	**0.00**	-	-
HRNet [55]	3.45	**0.20**	-	-
TCDCN [45]	-	-	8.05	-
SHN-GCN [66]	-	-	5.67	2.36
DETR (R50) [76]	3.79	0.59	5.46	2.37
Wing [44]	-	-	5.44	3.75
DCFE [38]	-	-	5.27	7.29
AWing [61]	-	-	4.94	0.99
ADNet [70]	-	-	**4.68**	**0.59**
SLPT (W18C-l)[133]	**3.32**	**0.00**	4.79	1.18
DSLPT (R34)	3.34	0.39	4.81	0.98
DSLPT (R50)	3.34	**0.00**	4.81	1.18
DSLPT (W18C-l)	**3.31**	**0.00**	**4.77**	**0.79**
DSLPT (W18C)	3.33	**0.20**	4.79	1.36

Table 5.4: Performance comparisons on *within*-dataset validation. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C]

Method	$\text{NME}_{\text{box}}(\%)\downarrow$			$\text{AUC}^{0.07}_{\text{box}}(\%)\uparrow$		
	Menpo	300W-p	COFW68	Menpo	300W-p	COFW68
SAN† [47]	2.95	2.86	3.50	61.9	59.7	51.9
2D-FAN† [58]	2.16	2.32	2.95	69.0	66.5	57.5
KDN [63]	2.26	2.49	-	68.2	67.3	-
Softlabel† [63]	2.27	2.32	2.92	67.4	66.6	57.9
KDN† [63]	2.01	2.21	2.73	71.1	68.3	60.1
LUVLI [67]	2.18	2.24	2.75	70.1	68.3	60.8
LUVLI† [67]	2.04	2.10	**2.57**	71.9	70.2	**63.4**
DSLPT (R34)⋆	1.98	2.23	2.70	73.4	67.9	61.5
DSLPT (R34)†	**1.89**	2.13	**2.57**	**74.5**	69.3	63.3
DSLPT (R34)	1.96	2.11	**2.57**	73.6	69.8	**63.4**
DSLPT (R50)	1.95	2.09	**2.56**	73.7	70.2	**63.5**
DSLPT (W18C-l)	1.93	**2.07**	2.59	74.0	**70.4**	63.3
DSLPT (W18C)	**1.92**	**2.04**	**2.57**	**74.2**	**70.9**	63.1

Table 5.5: Performance comparisons on Menpo, 300W-private and COFW68. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C, ⋆=initialized from scratch, †=Pretrained on 300W-LP-2D]

regression methods, heatmap regression methods naturally exhibit better performance because the semantic landmark localization can avoid overfitting to a certain extent. Nevertheless, DSLPT still yields the second best performance in NME and FR.

Menpo, COFW68, 300W-private

To better verify the generalization ability of DSLPT, we carry out three *cross*-dataset validations as [67]. DSLPT employs the full set of 300W (3,837 images) as the training set, and is then evaluated on 6,653 near-frontal training faces of *Menpo*, 600 faces of *300W-private* and 507 faces of *COFW68* respectively. We report the NME_{box} (set d_{norm} to the geometric mean of bounding box size) and $\text{AUC}^{0.07}_{\text{box}}$ on the three test sets in Table 5.5. Without pretraining, even the lightest DSLPT (ResNet34) can outperform LUVLi, improving the metric by 9.17%, 0.44% and 1.81% in NME_{box} on Menpo, 300W-private and COFW68 respectively. The improvement is more significant when we compare DSLPT to KDN. With sufficient samples, the results il-

Method	Unocculded		Externally Occluded	
	NME$_{\text{box}}$ ↓	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$	NME$_{\text{box}}$ ↓	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$
Softlabel [63]†	2.30	5.99	5.01	7.32
KDN [63]†	2.34	1.63	4.03	11.62
LUVLi [67]†	2.15	9.37	**4.00**	32.49
SLPT (W18C-l) [133]	**2.09**	-	4.33	-
DSLPT (R34)*	2.22	1.02	4.32	2.96
DSLPT (R34)†	2.19	1.30	**3.86**	3.91
DSLPT (R34)	**2.09**	0.98	4.18	2.99
DSLPT (R50)	**2.06**	1.00	4.22	3.21
DSLPT (W18C-l)	2.12	1.12	4.14	3.81
DSLPT (W18C)	2.10	0.83	4.12	3.03

Table 5.6: NME$_{\text{box}}$ and $|\mathbf{\Sigma}|^{\frac{1}{2}}$ on the externally occluded and unoccluded landmarks of COFW68. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C, *=initialized from scratch, †=Pretrained on 300W-LP-2D]

lustrate that DSLPT has a very competitive generalization ability. For a fair comparison, we also implement a model initialized from scratch and pretrain it on 300W-LP-2D [40] with 20 epochs. 300W-LP-2D consists of a large number of samples with various views. Therefore, the pretraining on 300W-LP-2D can encourage DSLPT to learn a more coherent inherent relation for better robustness compared to the model without pretraining, which effectively improves the performance on the cases with occlusion. Therefore, compared to the DSLPT without pretraining, we can observe an improvement of 4.54%, 4.48% and 4.81% in NME$_{\text{box}}$ on Menpo, 300W-private and COFW68 respectively.

To further explore the influence of pretraining and the dynamic patches, we tabulate the NME$_{\text{box}}$ and square root of the determinant of uncertainty (SQDU) on the externally occluded and unoccluded landmarks of COFW68 in Table 5.6. For the ease of comparisons, we restore the predicted SQDU of DSLPT from the normalized patch coordinate system to the unnormalized global coordinate system. The value of SQDU $|\mathbf{\Sigma}|^{\frac{1}{2}}$ is calculated and reported using the unnormalized global coordinates. Similar to Softlabel, KDN and LUVLi, the SQDU on

Method	Inter-Ocular NME(%)↓	FR$_{0.1}$(%)↓
TCDCN [45]	7.66	16.17
CFSS [12]	6.28	9.07
ODN [48]	5.30	-
AV w. SAN [88]	4.43	2.82
LAB [43]	4.62	2.17
SDL [50]	4.22	0.39
SDFL (W18C) [52]	4.18	**0.00**
DETR (R50) [76]	4.15	0.59
SLPT (W18C-l) [133]	4.10	0.59
DSLPT (R34)*	4.13	0.59
DSLPT (R34)†	**4.04**	**0.00**
DSLPT (R34)	4.05	0.39
DSLPT (R50)	**4.03**	0.59
DSLPT (W18C-l)	4.05	**0.20**
DSLPT (W18C)	**4.03**	**0.20**

Table 5.7: Performance comparisons under Inter-Ocular normalization on *cross*-dataset validation. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C, *=initialized from scratch, †=Pretrained on 300W-LP-2D]

Method	$\text{NME}_{\text{diag}} \downarrow$		$\text{NME}_{\text{box}} \downarrow$	$\text{AUC}^{0.07}_{\text{box}} \uparrow$
	Full	Frontal	Full	Full
DeepReg [39]	2.12%	-	-	-
RND [46]	2.06%	-	-	-
SAN [47]	1.91%	1.85%	-	-
Wing [44]	-	-	3.56%	0.535
KDN [63]	-	-	2.80%	0.603
ODN [48]	1.63%	1.38%	-	-
HRNet [55]	1.57%	1.46%	-	-
LUVLi [67]	1.39%	1.19%	2.28%	0.680
MHHN [62]	1.38%	1.19%	-	-
SHN-GCN [66]	-	-	2.15%	-
LAB [43]	1.25%	1.14%	-	-
DETR (R50) [76]	**0.970%**	0.838%	1.372%	0.806
DSLPT (R34)*	1.029%	0.870%	1.455%	0.794
DSLPT (R34)	0.974%	0.834%	1.376%	0.805
DSLPT (R50)	**0.967%**	**0.826%**	**1.368%**	**0.807**
DSLPT (W18C-l)	**0.967%**	**0.822%**	**1.367%**	**0.807**
DSLPT (W18C)	**0.967%**	0.837%	**1.367%**	**0.808**

Table 5.8: Performance comparisons on the full set and frontal subset of AFLW-19. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C, *=initialized from scratch]

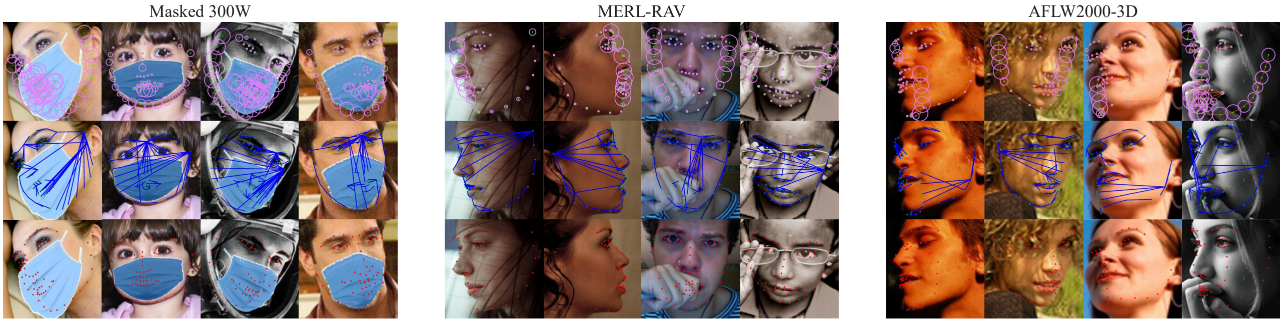


Figure 5.6: Visualized results of the externallly occluded and self-occluded cases on Masked 300W, MERL-RAV and AFLW2000-3D. The **red** point and **green** point indicate the ground truth and the predicted landmark respectively. The uncertainty of each landmark is shown by **pink** circle. Each landmark is connected with the landmark with highest cross attention weigh by **blue** line. The **Orange** lines represent the 3D facial boundaries.

unoccluded landmarks predicted by DSLPT is $1/4 \sim 1/3$ of the SQDU on occluded landmarks, as shown in Table 5.6. It demonstrates that the predicted uncertainty of DSLPT can reflect the landmark occlusion properly. Since SLPT cannot predict the uncertainty of landmarks, we did not report the SQDU of SLPT. With the same backbone, DSLPT achieves an improvement of 4.4% in NME_{box} on the occluded landmarks of COFW68 compared to SLPT, which illustrates the adaptive receptive field of the dynamic patch can improve the robustness on occluded landmarks effectively. Besides, with the same setting (pretrained on 300W-LP-2D), the lightest DSLPT (ResNet34) also improves NME_{box} by 22.95%, 4.22% and 3.50% on externally occluded landmarks respectively compared to Softlabel, KDN and LUVLi.

Moreover, we also report the Inter-Ocular NME on COFW68 in Table 5.7 to compare DSLPT to other state-of-the-art methods.

AFLW-19

We report the NME_{box} (set d_{norm} to the geometric mean of bounding box size) and NME_{diag} (set d_{norm} to the diagonal of the bounding box) of DSLPT on full set and frontal subset, and compare them to other state-of-the-art methods, as shown in Table 5.8. With sufficient training samples, both DETR and DSLPT, including the model initialized from scratch, outperform other heatmap regression methods by a large margin. It illustrates the performance of coordinate regression methods heavily relies on the scale of dataset.

Method	Inter-Ocular NME (%) ↓		
	Common	Challenging	Fullset
CFSS [12]	11.73	19.98	13.35
Hourglass [56]	8.17	13.52	9.22
MDM [36]	7.66	11.67	8.44
FAN [58]	7.36	10.81	8.02
LAB [43]	6.07	9.59	6.76
SAAT [68]	5.42	11.36	6.58
GlomFace [19]	5.29	8.81	5.98
DSLPT (R34)*	4.95	8.10	5.56
DSLPT (R34)	**4.66**	**7.49**	**5.22**
DSLPT (R50)	**4.51**	**7.67**	**5.13**
DSLPT (W18C-l)	4.86	8.03	5.48
DSLPT (W18C)	4.78	8.10	5.42

Table 5.9: Performance comparisons with the state-of-the-art methods on Masked 300W common subset, challenging subset and fullset. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C, *=initialized from scratch]

Metrics (%)	Method	Full	Frontal	Half-Profile	Profile
$\text{NME}_{\text{box}} \downarrow$	DU-Net [57]	1.99	1.89	2.50	1.92
	LUVLi [67]	1.61	1.74	1.79	1.25
	SLPT [133]	1.51	1.62	1.68	1.21
	DSLPT (R34)⋆	1.64	1.76	1.69	1.30
	DSLPT (R34)	1.52	1.63	1.69	1.19
	DSLPT (R50)	**1.50**	1.62	**1.67**	**1.18**
	DSLPT (W18C-l)	**1.48**	**1.59**	**1.64**	**1.16**
	DSLPT (W18C)	**1.48**	**1.60**	**1.64**	**1.16**
$\text{AUC}_{\text{box}}^{0.07} \uparrow$	DU-Net	71.80	73.25	64.78	72.79
	LUVLi	77.08	75.33	74.69	82.10
	SLPT	78.33	76.82	76.01	82.74
	DSLPT (R34)⋆	76.58	74.89	75.90	81.47
	DSLPT (R34)	78.29	76.67	75.90	82.94
	DSLPT (R50)	78.55	76.93	76.17	83.21
	DSLPT (W18C-l)	**78.85**	**77.28**	**76.51**	**83.40**
	DSLPT (W18C)	**78.87**	**77.24**	**76.58**	**83.46**

Table 5.10: Performance comparisons on MERL-RAV. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C, ⋆=initialized from scratch]

Masked 300W

Following [19], we use the training set of 300W [109] to train the proposed model and each image is randomly occluded by five blocks with different sizes for data augmentation. The Masked 300W is only used for evaluation and the results are tabulated in Table 5.9. GlomFace is also a patch based regression method and designed for the cases with heavy occlusion. With the dynamic patches and the case dependent inherent relation learning, DSLPT further achieves an impressive improvement of 14.21% in NME on the fullset compared to GlomFace. It illustrates both the dynamic patches and inherent relation learning can significantly improve the robustness of patch based methods, especially for the cases with heavy occlusion. Some visualized results are shown in Fig. 5.6.

Method	Self-occluded		Unoccluded		Externally Occluded	
	$\text{NME}_{\text{box}} \downarrow$	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$	$\text{NME}_{\text{box}} \downarrow$	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$	$\text{NME}_{\text{box}} \downarrow$	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$
LUVLI [67]	-	-	1.60%	9.28	3.53%	34.41
SLPT [133]	-	-	**1.50%**	-	3.33%	-
DSLPT (R34)*	-	4.47	1.64%	0.72	3.55%	2.52
DSLPT (R34)	-	2.55	1.51%	0.67	3.34%	2.53
DSLPT (R50)	-	3.65	**1.50%**	0.67	3.29%	2.62
DSLPT (W18C-l)	-	3.12	**1.48%**	0.71	**3.25%**	2.83
DSLPT (W18C)	-	2.82	**1.48%**	0.68	**3.26%**	2.69

Table 5.11: NME_{box} and $\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$ on self-occluded, externally occluded and unoccluded landmarks of MERL. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C, *=initialized from scratch]

MERL-RAV

As shown in Table 5.10, DSLPT improves the metric by 8.07% and 25.63% in NME_{box} over LUVLi and DU-Net respectively. Some cases with external occlusion and self-occlusion are demonstrated in Fig. 5.6. Although MERL-RAV does not provide the coordinate annotation for the self-occluded landmark, DSLPT can still localize them properly in the testing phase. The main reason is that the learned inherent relation enables the model to locate the self-occluded landmarks with the annotated landmarks. Moreover, the dynamic patch provides large receptive field for the self-occluded landmarks because of their high uncertainty. Therefore, DSLPT outperforms other state-of-the-art methods significantly and obtains much stronger robustness. We also report the NME_{box} and SQDU on three types of landmarks of MERL in Table 5.11. The SQDU on unoccluded landmarks predicted by DSLPT is also $1/4 \sim 1/3$ of the SQDU on externally occluded landmarks and $1/6 \sim 1/4$ of the SQDU on self-occluded landmarks. The fact that the self-occluded landmarks have larger uncertainty than the externally occluded landmark is also consistent with human perception: human labelers are generally very bad at localizing self-occluded landmarks [67]. As a result, the adaptive receptive field brings an improvement of 2.38% and 8.38% in NME_{box} on the externally occluded landmarks compared to SLPT and LUVLi.

Method	NME$_{\text{box}}$ (%)↓			
	$[0°, 30°]$	$[30°, 60°]$	$[60°, 90°]$	Mean
SDM [9]	3.67	4.94	9.67	6.12
3DDFA [40]	3.78	4.54	7.93	5.42
3DDFA+SDM [40]	3.43	4.24	7.17	4.94
3D-FAN [58]	3.15	3.53	4.60	3.76
3DDFA-TPAMI [41]	2.84	3.57	4.96	3.79
3DDFAV2 (MR) [49]	2.75	3.49	4.53	3.59
3DDFAV2 (MRS) [49]	2.63	3.42	4.48	3.51
SynergyNet [51]	2.65	**3.30**	**4.27**	**3.41**
DSLPT (R34)*	**2.51**	**3.40**	**4.32**	**3.41**
DSLPT (R50)*	**2.54**	3.61	4.37	**3.50**

Table 5.12: Performance comparisons with the state-of-the-art methods on AFLW2000-3D. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, $\star$=initialized from scratch]

300W-LP & AFLW2000-3D

To evaluate the performance of DSLPT on extremely self-occluded conditions, we carry out experiments on AFLW2000-3D to predict the 2D projection of 3D faces. Following [40], we use the 300W-LP samples synthesized from the training set of LFPW, HELEN and the whole AFW for training. As shown in Table 5.12, DSLPT outperforms the state-of-the-art methods with a large margin for the cases whose absolute yaw angles are within 30°. As shown in Fig. 5.6, DSLPT locates the extremely self-occluded landmarks via the visible landmarks. Therefore, DSLPT still yields the second best performance for the cases with large absolute yaw angle. Besides, the augmentation technique used by [40] leads to a domain gap between 300W-LP and AFLW2000-3D. A deeper backbone will fit the training domain better but performs worse in the testing domain. Therefore, the DSLPT with ResNet34 outperforms the DSLPT with ResNet50.

5.3.5 Ablation Study

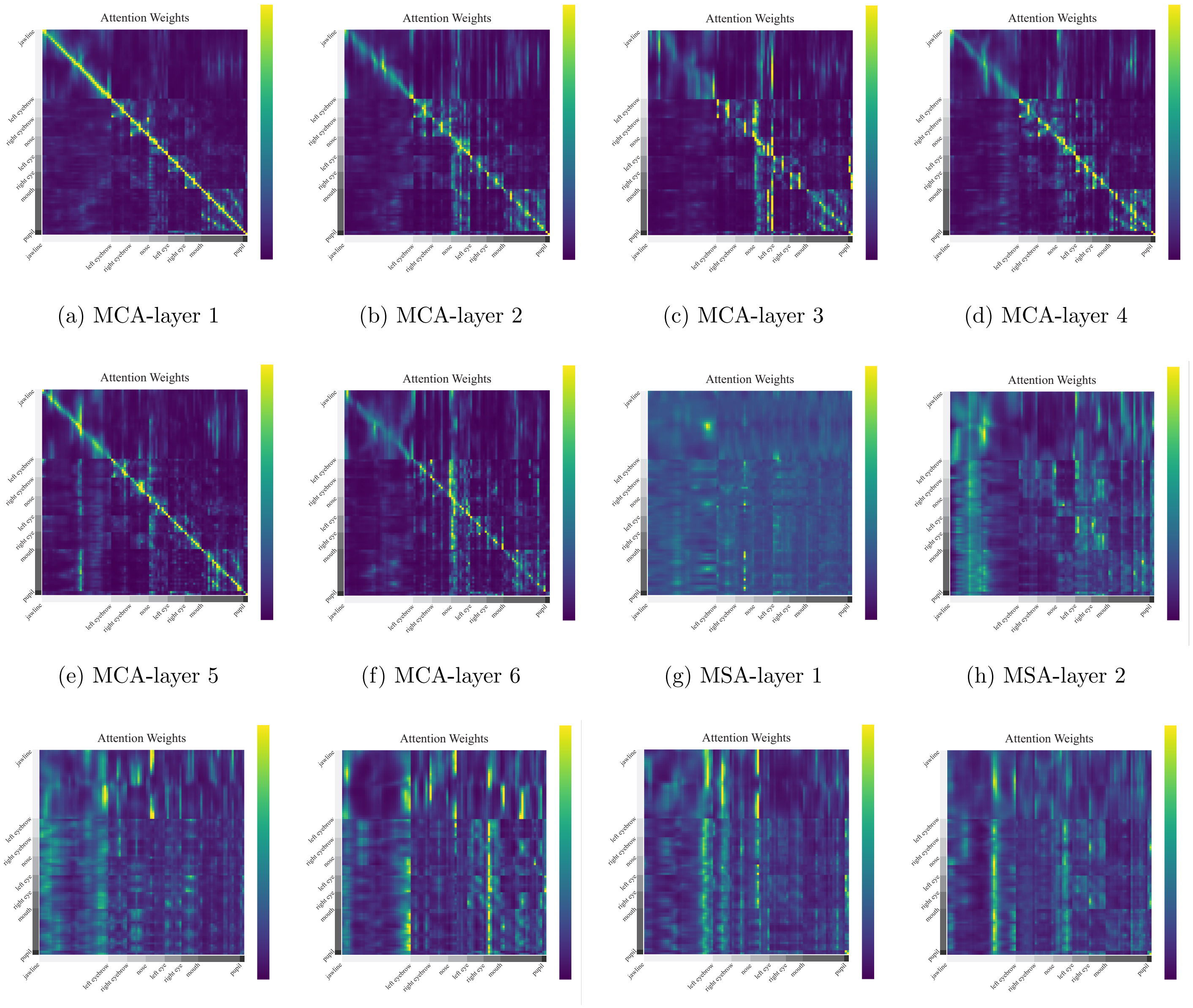
In this section, we explore how the key components of the proposed DSLPT influence the final performance by performing extensive ablation studies on the most challenging dataset, WFLW.

Model	Intermediate Stage											
	1st stage			2rd stage			3rd stage			4th stage		
	NME↓	FR$_{0.1}$↓	AUC↑	NME↓	FR$_{0.1}$↓	AUC↑	NME↓	FR$_{0.1}$↓	AUC↑	NME↓	FR$_{0.1}$↓	AUC↑
DSLPT (W18C-l) with 1 stage	4.34%	3.72%	0.580	-	-	-	-	-	-	-	-	-
DSLPT (W18C-l) with 2 stages	4.25%	3.04%	0.586	4.05%	2.56%	0.603	-	-	-	-	-	-
DSLPT (W18C-l) with 3 stages	4.25%	3.24%	0.586	4.03%	2.52%	0.606	**4.02%**	**2.40%**	**0.607**	-	-	-
DSLPT (W18C-l) with 4 stages	4.43%	3.92%	0.574	4.12%	2.92%	0.600	4.11%	3.00%	0.601	4.11%	2.92%	0.601

Table 5.13: The influence of coarse-to-fine framework when using different number of stages. Key: [**Best**, W18C-l=HRNetW18C-lite]

Influence of coarse-to-fine framework

We demonstrate the performance of the final stage and intermediate stages of the DSLPT with different coarse-to-fine stage numbers, as shown in Table 5.13. Compared to the DSLPT with a single stage, the DSLPT with 3 stages improves the metric by 7.37%, 35.5% and 4.44% in NME, $\mathrm{FR}_{0.1}$ and AUC respectively. It's worth mentioning that the coarse-to-fine framework can also improve the performance of the intermediate stage. The main reason is that the DSLPT is shared in each stage and the variance of patch size in different stages significantly augments the training data. As a result, the inherent relation learned by DSLPT becomes more coherent, promising a better performance to the intermediate stage. However, the performance converges when the stage number is more than 3 since the too large variance of patch size leads to a domain gap. Besides, the patches in the 4th stage are too small and they cannot serve as the contextual information for other landmarks.



(a) MCA-layer 1 (b) MCA-layer 2 (c) MCA-layer 3 (d) MCA-layer 4

(e) MCA-layer 5 (f) MCA-layer 6 (g) MSA-layer 1 (h) MSA-layer 2

(i) MSA-layer 3 (j) MSA-layer 4 (k) MSA-layer 5 (l) MSA-layer 6

Figure 5.7: The statistical attention interactions of MCA and MSA in the final stage on the WFLW test set.

Method	MSA	MCA	NME ↓	$\text{FR}_{0.1}$ ↓	$\text{AUC}_{0.1}$ ↑
DSLPT (W18C-l)	w/o	w/o	4.27%	3.48%	0.587
DSLPT (W18C-l)	w/	w/o	4.07%	2.76%	0.604
DSLPT (W18C-l)	w/o	w/	4.08%	2.92%	0.603
DSLPT (W18C-l)	w/	w/	**4.02%**	**2.40%**	**0.607**

Table 5.14: Influence of MSA and MCA block on WFLW test. Key: [**Best**, W18C-l=HRNetW18C-lite]

Influence of MCA and MSA block

To verify the effectiveness of **inherent relation learning**, we implement four models with-/without MSA and MCA block and their performance on WFLW testset is reported in Table 5.14. For the model without MCA block, we replace landmark queries with landmark representations as the input of Transformer directly. Without MSA and MSA block, each landmark is predicted merely based on its supporting patch. However, it still outperforms most coordinate regression methods because the coarse-to-fine framework and dynamic patches enable the model to generate a more fine-grained representation for each landmark, promising an accurate localization. The inter *representation-query* relation learned by MCA block and the inter *query-query* relation learned by MSA block significantly boost the performance, reaching at 4.07% and 4.08% in NME respectively. We visualize the mean attention weights in the 3^{rd} stage on the WFLW testset, as shown in Fig. 5.7. The MCA blocks tend to aggregate the representation of the corresponding and neighboring landmark to generate a local feature, while the MSA blocks pay more attention to the landmark with a long distance for a global feature. Therefore, MSA and MCA can incorporate with each other for better performance.

Influence of structure encoding

To explore the influence of the structure encoding, we implement different models with 1D positional encoding or with/without structure encoding, ranging from 1 to 3. The 1D positional encoding is generated by the cosine and sine function [121]. Their performance on WFLW testset is reported in Table 5.15. Both structure encoding and positional encoding can improve the performance of DSLPT. However, the improvement brought by 1D positional encoding is not as significant as the structure encoding. The main reason is that the face structure is hard

Method	Encoding	NME↓	$\text{FR}_{0.1}\downarrow$	$\text{AUC}_{0.1}\uparrow$
Model 1 (W18C-l)	N/A	4.08%	2.64%	0.603
Model 2 (W18C-l)	Positional encoding	4.04%	2.56%	0.606
Model 3 (W18C-l)	Structure encoding	**4.02%**	**2.40%**	**0.607**

Table 5.15: Influence of different kinds of encodings. Key: [**Best**, W18C-l=HRNetW18C-lite]

Method	Layer number	Flops	Params	NME↓	$\text{FR}_{0.1}\downarrow$	$\text{AUC}_{0.1}\uparrow$
Model 1 (W18C)	2	6.32G	15.1M	4.05%	2.48%	0.604
Model 2 (W18C)	4	7.08G	17.2M	4.02%	2.52%	0.607
Model 3 (W18C)	6	7.83G	19.3M	4.01%	2.52%	0.607
Model 4 (W18C)	12	10.1G	25.7M	**3.98%**	**2.44%**	**0.609**

Table 5.16: Computational complexity, parameters and performance of the DSLPT with different inherent relation layer number on WFLW testset. Key: [**Best**, W18C=HRNetW18C]

to be represented by a 1D shape.

Influence of inherent relation layer number

To further explore the influence of inherent relation layer number, we implement four DSLPT models with 2, 4, 6, and 12 inherent relation layers respectively, ranging from 1 to 4. As shown in Table 5.16, the improvement brought more inherent relation layers is more significant than a deeper backbone. Replacing HRNetW18C-lite with HRNetW18C increases parameters from 13.3M to 19.3M and Flops from 6.06G to 7.83G while it only improves the metrics by 0.25% in NME. Increasing the inherent relation layer number from 4 to 6 promises a similar improvement in NME. Nevertheless, it only leads to an improvement of 10.6% and 12.2% in Flops and parameters respectively. Therefore, learning inherent relation with DSLPT is more efficient than learning a simple feature map with a CNN network. With 12 inherent relation layers, the performance of DSLPT can be further improved, reaching at 3.98%, 2.44% and 0.609 in NME, FR and AUC respectively.

Influence of inherent relation layer number

To further explore the influence of inherent relation layer number, we implement four DSLPT models with 2, 4, 6, and 12 inherent relation layers respectively, ranging from 1 to 4. As shown in Table 5.16, the improvement brought more inherent relation layers is more significant than a deeper backbone. Replacing HRNetW18C-lite with HRNetW18C increases parameters from 13.3M to 19.3M and Flops from 6.06G to 7.83G while it only improves the metrics by 0.25% in NME. Increasing the inherent relation layer number from 4 to 6 promises a similar improvement in NME. Nevertheless, it only leads to an improvement of 10.6% and 12.2% in Flops and parameters respectively. Therefore, learning inherent relation with DSLPT is more efficient than learning a simple feature map with a CNN network. With 12 inherent relation layers, the performance of DSLPT can be further improved, reaching at 3.98%, 2.44% and 0.609 in NME, FR and AUC respectively.

Influence of dynamic patches and probability distribution estimation

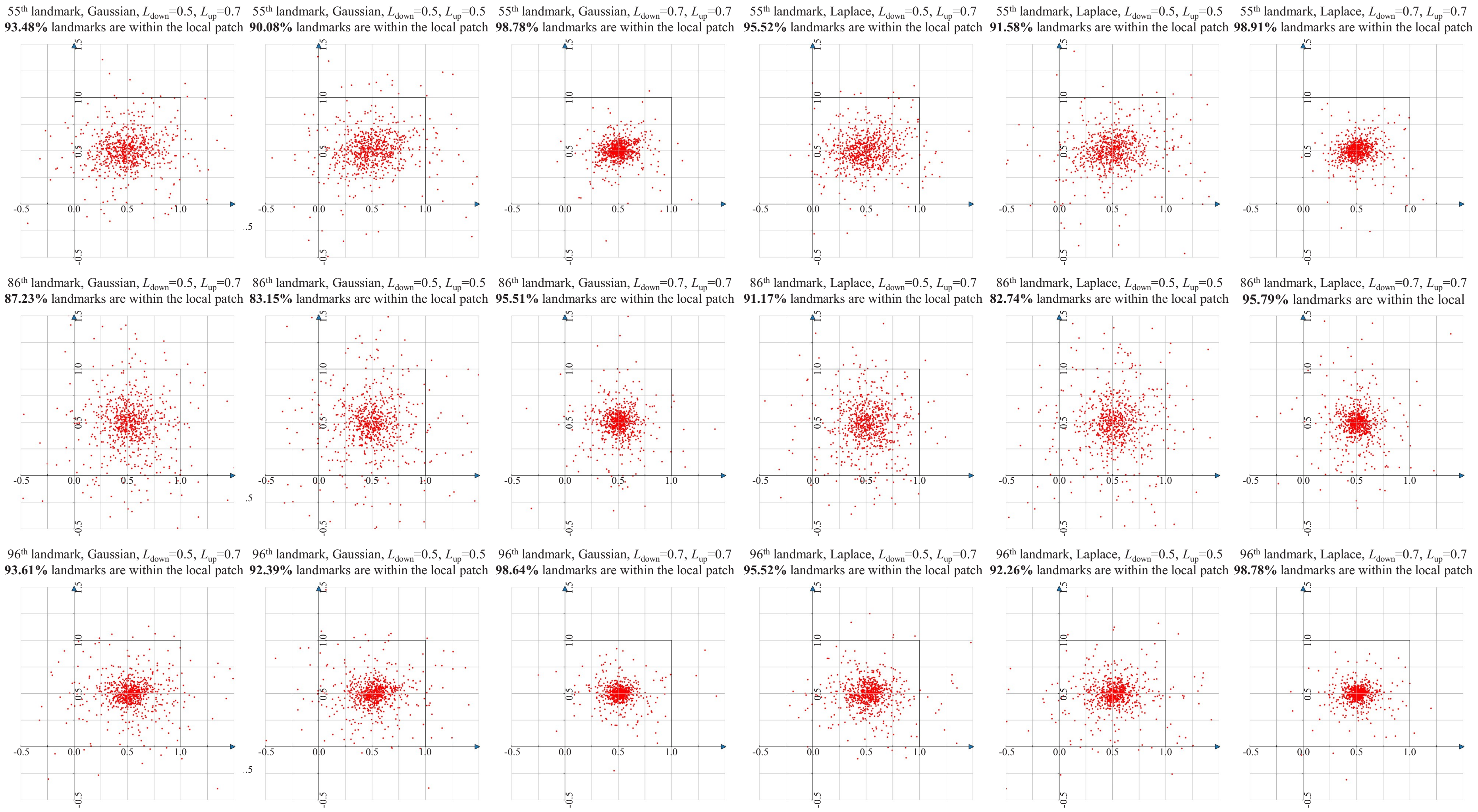
We report the performance of different loss functions as well as the model with/without the dynamic patch in Table 5.17. When we set L_{down} and L_{up} to 0.5 and constrain the model learning with the L2 function, the DSLPT downgrades to our **original SLPT** [133]. In the same condition, replacing the L1 or L2 loss function with the Laplace or Gaussian negative log-likelihood function leads to a slight improvement. Unlike [67], both the Laplace and Gaussian negative log-likelihood function demonstrate comparable performance in DSLPT. The main reason is that [67] predicts heatmap and covariance matrix for each landmark from a sharing global feature. The Gaussian likelihood is the probabilistic analog of the L2 loss, which is sensitive to outliers. The negative influence brought by outliers propagates to each landmark through the sharing global feature. However, the prediction heads of DSLPT predict each landmark independently from its corresponding feature. As a result, it is less sensitive to outliers. Besides, the Gaussian negative log-likelihood function drives DSLPT to focus on the challenging samples so that it performs better in FR and AUC.

Compared to different loss functions, the proposed dynamic patch leads to a more significant improvement. We visualize the annotated landmark position distribution in the patch coordinate system on the occlusion subset of WFLW, as shown in Fig.5.8. A smaller fixed patch size (L_{down} and L_{up} are set to 0.5) ensures higher feature resolution. But when it comes to the

Loss function	L_{down}	L_{up}	NME↓	$\text{FR}_{0.1}\downarrow$	$\text{AUC}_{0.1}\uparrow$
Gaussian	0.5	0.7	4.020%	**2.40%**	**0.607**
Gaussian	0.5	0.5	4.064%	2.68%	0.604
Gaussian	0.7	0.7	4.084%	2.92%	0.603
Laplace	0.5	0.7	**4.018%**	2.68%	0.606
Laplace	0.5	0.5	4.059%	2.76%	0.604
Laplace	0.7	0.7	4.070%	2.88%	0.604
L1	0.5	0.5	4.076%	2.68%	0.601
L2	0.5	0.5	4.083%	2.60%	0.603

Table 5.17: Performance of the DSLPT with different loss function on WFLW testset. Each model is with HRNetW18C-lite as the backbone. Key: [**Best**, Gaussian=Gaussian negative log-likelihood function, Laplace=Laplace negative log-likelihood function]

landmark with high uncertainty, the patch is usually with limited contextual information and the ground truth of the landmarks with high uncertainty deviates from the patch area, resulting in an inaccurate representation. And the lower feature resolution brought by the larger patch (L_{down} and L_{up} are set to 0.7) also leads to performance degradation. For the dynamic patch (L_{down} is set to 0.5 and L_{up} is set to 0.7), the distribution density in the patch center is very similar to the distribution of the small patch size while the distance of the sample with high uncertainty to the center of the local patch is shortened effectively. The results demonstrate that the dynamic patch applies smaller size to most landmarks for higher feature resolution and larger size to the landmark with high uncertainty for more contextual information.


55th landmark, Gaussian, L_{down}=0.5, L_{up}=0.7 **93.48%** landmarks are within the local patch

55th landmark, Gaussian, L_{down}=0.5, L_{up}=0.5 **90.08%** landmarks are within the local patch

55th landmark, Gaussian, L_{down}=0.7, L_{up}=0.7 **98.78%** landmarks are within the local patch

55th landmark, Laplace, L_{down}=0.5, L_{up}=0.7 **95.52%** landmarks are within the local patch

55th landmark, Laplace, L_{down}=0.5, L_{up}=0.5 **91.58%** landmarks are within the local patch

55th landmark, Laplace, L_{down}=0.7, L_{up}=0.7 **98.91%** landmarks are within the local patch

86th landmark, Gaussian, L_{down}=0.5, L_{up}=0.7 **87.23%** landmarks are within the local patch

86th landmark, Gaussian, L_{down}=0.5, L_{up}=0.5 **83.15%** landmarks are within the local patch

86th landmark, Gaussian, L_{down}=0.7, L_{up}=0.7 **95.51%** landmarks are within the local patch

86th landmark, Laplace, L_{down}=0.5, L_{up}=0.7 **91.17%** landmarks are within the local patch

86th landmark, Laplace, L_{down}=0.5, L_{up}=0.5 **82.74%** landmarks are within the local patch

86th landmark, Laplace, L_{down}=0.7, L_{up}=0.7 **95.79%** landmarks are within the local

96th landmark, Gaussian, L_{down}=0.5, L_{up}=0.7 **93.61%** landmarks are within the local patch

96th landmark, Gaussian, L_{down}=0.5, L_{up}=0.5 **92.39%** landmarks are within the local patch

96th landmark, Gaussian, L_{down}=0.7, L_{up}=0.7 **98.64%** landmarks are within the local patch

96th landmark, Laplace, L_{down}=0.5, L_{up}=0.7 **95.52%** landmarks are within the local patch

96th landmark, Laplace, L_{down}=0.5, L_{up}=0.5 **92.26%** landmarks are within the local patch

96th landmark, Laplace, L_{down}=0.7, L_{up}=0.7 **98.78%** landmarks are within the local patch

Figure 5.8: The distribution of 55th, 86th and 96th landmark in patch coordinate system on the occlusion subset of WFLW. The percentage of the landmarks that are within the local patch is also reported in captions.

	Inter-ocular NME(%) ↓				
	L_{up}=0.5	L_{up}=0.6	L_{up}=0.7	L_{up}=0.8	L_{up}=0.9
L_{down}=0.4	4.097	4.061	4.050	4.061	4.077
L_{down}=0.5	4.064	4.047	**4.020**	4.041	4.063

Table 5.18: Performance of the DSLPT with different L_{down} and L_{up} on WFLW testset. Key: [**Best**]

Influence of the threshold of patch size

We implement DSLPT with different L_{down} and L_{up} on WFLW testset to study the influence of patch size threshold. As shown in Table 5.18, both low L_{down} and L_{up} lead to performance degradation since they result in contextual information missing. And, high L_{up} leads to a high variance in patch size, causing a domain gap in the same coarse-to-fine stage. The domain gap commonly has a negative influence on the inherent relation learning. The experiment results demonstrate that DSLPT exhibits the best performance when L_{down} and L_{up} are set to 0.5 and 0.7 respectively.

Computational complexity and parameters: although DSLPT is trained with 3 stages, we can still use the intermediate result as the final output and do not run the following stages to further reduce computational complexity. It does not require any modification to the weights since the DSLPT is shared in each stage. Therefore, DSLPT is quite flexible to fit the devices with different computational capacities. DETR cannot be implemented in a coarse-to-fine manner so it has only one stage. We report the performance, computational complexity and parameters of DETR and the DSLPT with different stage numbers in Table 5.19. Although DSLPT runs three times for coarse-to-fine landmark localization, its computational complexity is still much lower than DETR because the sparse local patch significantly decreases the token number. Most state-of-the-art methods adopt a very deep backbone, such as 8 DU-Net and ResNet152, to extract landmark representation from global feature. Besides, heatmap regression methods require an additional post-processing procedure to transfer the heatmaps into landmark coordinates, which makes them less efficient. With the proposed local sparse patches, DSLPT explicitly produces the representation for each landmark and directly regresses the landmark coordinates from the representations. Therefore, DSLPT can achieve better performance with a lighter backbone. As shown in Table 1, DSLPT achieves a comparable performance with

Method	Landmark number	1 stage			2 stages			3 stages		
		NME↓	Flops	Params	NME↓	Flops	Params	NME↓	Flops	Params
DSLPT (ResNet34)	98 Landmarks	4.337%	5.90G	31.06M	4.148%	6.97G	31.06M	4.138%	8.04G	31.06M
	68 Landmarks	3.239%	5.56G	31.05M	3.043%	6.29G	31.05M	3.035%	7.02G	31.05M
	29 Landmarks	3.675%	5.13G	31.03M	3.356%	5.44G	31.03M	3.338%	5.74G	31.03M
	19 Landmarks	1.420%	5.02G	31.02M	1.378%	5.22G	31.02M	1.377%	5.42G	31.02M
DSLPT (ResNet50)	98 Landmarks	4.325%	6.57G	33.47M	4.122%	7.67G	33.47M	4.112%	8.71G	33.47M
	68 Landmarks	3.224%	6.22G	33.46M	3.026%	6.96G	33.46M	3.019%	7.69G	33.46M
	29 Landmarks	3.664%	5.80G	33.44M	3.348%	6.10G	33.44M	3.338%	6.41G	33.44M
	19 Landmarks	1.408%	5.69G	33.43M	1.370%	5.89G	33.43M	1.368%	6.09G	33.43M
DSLPT (HRNetW18C -lite)	98 Landmarks	4.252%	3.91G	13.25M	4.031%	4.99G	13.25M	4.020%	6.06G	13.25M
	68 Landmarks	3.208%	3.57G	13.24M	3.014%	4.30G	13.24M	3.002%	5.04G	13.24M
	29 Landmarks	3.575%	3.15G	13.22M	3.320%	3.45G	13.22M	3.314%	3.76G	13.22M
	19 Landmarks	1.410%	3.04G	13.21M	1.368%	3.24G	13.21M	1.367%	3.44G	13.21M
DSLPT (HRNetW18C)	98 Landmarks	4.181%	5.69G	19.35M	4.015%	6.76G	19.35M	4.008%	7.83G	19.35M
	68 Landmarks	3.225%	5.35G	19.33M	2.988%	6.08G	19.33M	2.982%	6.81G	19.33M
	29 Landmarks	3.542%	4.92G	19.31M	3.305%	5.23G	19.31M	3.328%	5.53G	19.31M
	19 Landmarks	1.403%	4.82G	19.31M	1.368%	5.01G	19.31M	1.367%	5.21G	19.31M
DETR-DC5 [76] (ResNet50)	98 Landmarks	4.316%	10.62G	35.25M	-	-	-	-	-	-
	68 Landmarks	3.269%	10.39G	35.24M	-	-	-	-	-	-
	29 Landmarks	3.788%	10.10G	35.23M	-	-	-	-	-	-
	19 Landmarks	1.372%	10.03G	35.23M	-	-	-	-	-	-

Table 5.19: Performance, computational complexity and parameters of the DSLPT with different stages and backbone (all models are trained with 3 stages). The NME on WFLW (98 landmarks), 300W (68 landmarks) and COFW (29 landmarks) is normalized by inter-ocular distance. The NME on AFLW (19 landmarks) is normalized by $\sqrt{W_{\text{box}} \times H_{\text{box}}}$.

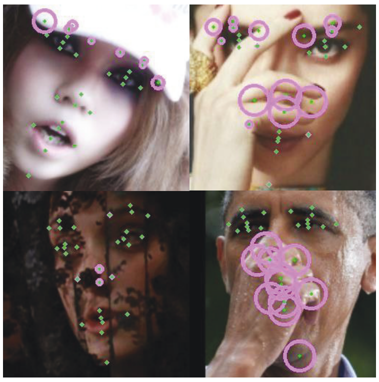


Figure 5.9: Some bad cases predicted by DSLPT on COFW dataset. The green points represent the predicted landmarks and the pink circles represent the predicted uncertainty.

only 1/5 ∼ 1/2 computational complexity compared to the state-of-the-art methods (Awing, ADNet, AVS+SAN and HIH).

5.3.6 Limitation

Compared to SLPT, DSLPT is capable of predicting additional uncertainties for facial landmarks without requiring any annotations. While the predicted uncertainties can significantly enhance the performance of facial landmark detection, they also demand a larger number of training samples. Therefore, as shown in Fig. 5.9, the DSLPT trained with a limited dataset, such as COFW, does not perform as well as a model trained using sufficient samples.

To address this limitation, we attempted to transfer the knowledge learned from WFLW to COFW. However, the landmark queries are highly dependent on the semantics of their corresponding landmarks. As a result, directly transferring the learned model leads to catastrophic forgetting. Therefore, bridging the gap between landmarks seen during pretraining and unseen landmarks is a promising direction for future research and could effectively overcome this challenge.

5.4 Conclusion

In this chapter, we propose the Dynamic Sparse Local Patch Transformer to address two main issues in facial landmark detection for better robustness and reliability in the cases with heavy occlusion: ignoring the landmark inherent relation and assuming the variance of a landmark probability distribution is a constant number. DSLPT generates representation for each landmark from the local patch and learns an inter *query-query* and inter *representation-query* relation in inherent relation layers. The learned case dependent inherent relation enables DSLPT to locate the landmarks with heavy occlusion by their relative position to the easily identified landmarks for better robustness. The model learning is constrained by a negative log-likelihood function rather than the L1 or L2 loss. Therefore, DSLPT predicts the probability distribution rather than a numerical coordinate. Moreover, we further incorporate DSLPT with a coarse-to-fine framework and the predicted distribution determines the size and position of the patches in the following coarse-to-fine stages. The variance of the predicted distribution enables DSLPT to apply a larger patch to the landmark with high uncertainty for more contextual information and a smaller patch to the landmark with low uncertainty for the higher resolution feature. Therefore, the dynamic patch ensures a more fine-grained landmark representation for the next stage and an initial face can converge to the target face gradually in the coarse-to-fine framework. The experiment results demonstrate that DSLPT successfully addresses the limitations in facial landmark detection and outperforms other methods with much less computational complexity.

Chapter 6

Robust and Reliable Facial Landmark Detection for Unseen Landmarks via Structure Prompts and Semantic Alignment

6.1 Introduction

In this chapter, we propose a **t**ask-agnostic **u**nified **f**ace **a**lignment (facial landmark detection) framework, named TUFA, that tackles the challenge of zero-shot facial landmark detection for the first time and improves performance in few-shot facial landmark detection. Instead of viewing each facial landmark as an independent regression target, TUFA employ labeled facial landmarks to learn a mapping between a plane and target faces. With this mapping, TUFA is **task-agnostic**, predicting not only seen landmarks, but also unseen landmarks. Specifically, we first encode the 2D coordinates of the mean face shape, which serves as the anchors of the plane, into a series of high-dimensional vectors named face structure prompts. The structure prompts then serve as queries to aggregate the image features produced by an encoder adaptively in a transformer-based decoder [121]. Finally, a multilayer perceptron (MLP) block is used to regress target landmarks from the decoder output, enabling the learning of the mapping. By employing the mean face shape as anchors, the plane explicitly represents the face structure in an interpretable manner. Therefore, the structure prompts, which bridge the gap between

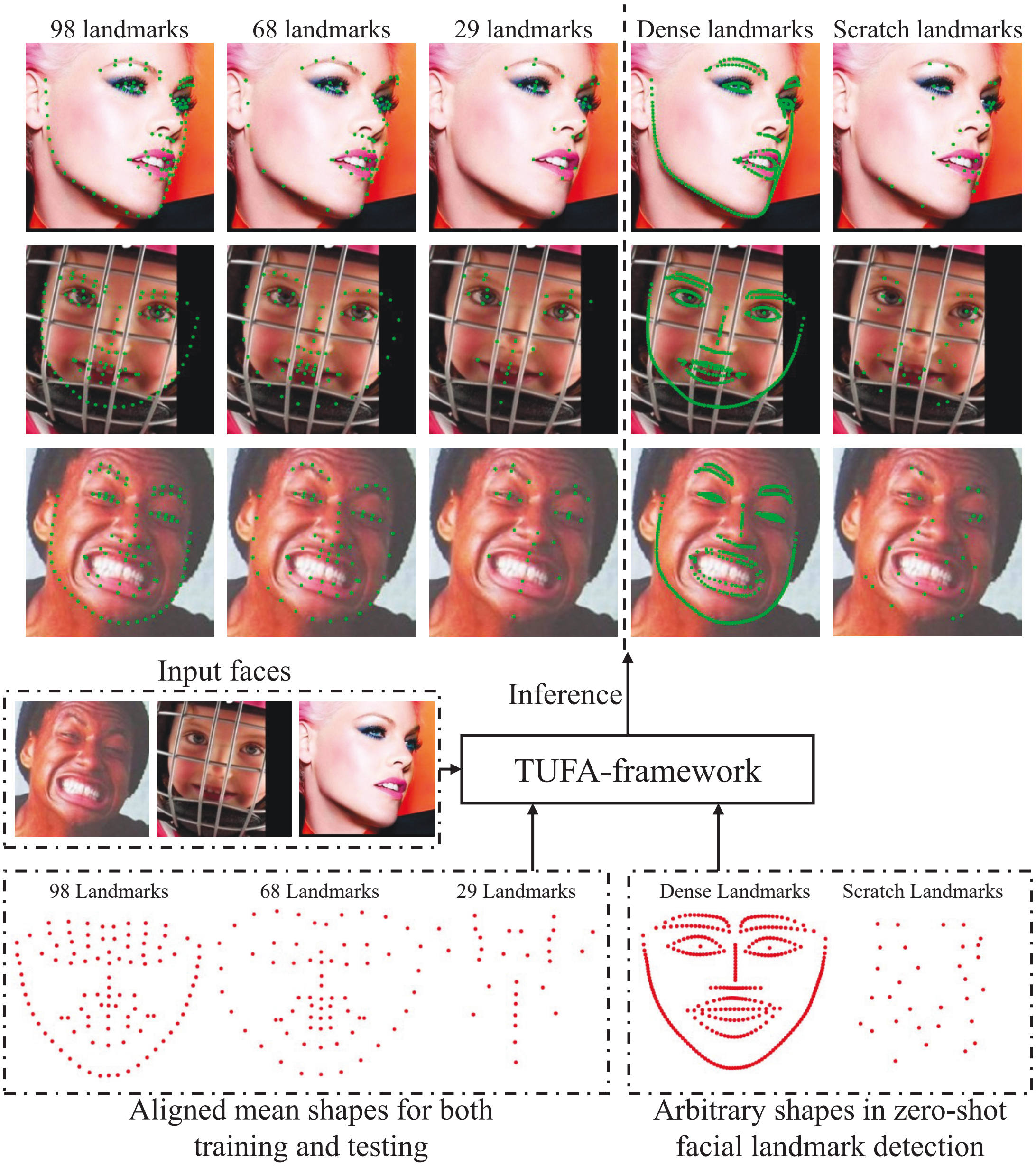


Figure 6.1: TUFA aims at learning a mapping from a plane to target faces, instead of setting each landmark as an independent regression target. The mapping enables both seen and unseen facial landmarks to have corresponding 2D positions on the plane. TUFA locates unseen landmarks in a zero-shot manner by encoding their 2D positions into face structure prompts. Moreover, the mean shapes from multiple datasets with aligned semantics enable TUFA to utilize the labeled landmarks from different datasets to learn a more robust result.

seen and unseen landmarks in zero-shot facial landmark , can be easily edited according to the geometric relationships of their corresponding positions on the plane. As shown in Fig. 6.1, by encoding various 2D positions on the plane into structure prompts, TUFA can locate arbitrary number of seen and unseen landmarks.

Moreover, despite the diverse semantic definitions of the labeled landmarks across multiple datasets, their learning targets can be unified in TUFA by aligning their semantic definitions on the same plane. To do so, we adjust each mean shape, which actually represents the semantic definitions of the labeled landmarks from its corresponding dataset, using a group of learnable 2D offsets. These offsets are updated during training, finding optimal positions on the plane to align the semantic definitions of the labeled landmarks. After the mean shapes are aligned, even the slight semantic differences between the similar landmarks from different datasets can be explicitly represented through the geometric relationships on the plane. Compared to learning from a single dataset, the denser anchors and larger number of labeled samples provided by multiple datasets help TUFA learn a result with very strong generalization ability. Besides, the unified learning target of TUFA across various datasets enables it to be easily transferred to a group of newly defined landmarks in a few-shot manner.

6.2 Methodology

In this section, we first introduce the structure of TUFA and explain its approach to achieving task-agnostic facial landmark detection. Next, we describe the proposed method for unifying the learning targets of the datasets with various annotation forms, as well as the loss function of TUFA. Finally, we discuss the strategy for achieving few-shot and zero-shot facial landmark detection using TUFA.

6.2.1 Task-agnostic Unified Facial landmark detection

The overall training pipeline of TUFA is demonstrated in Fig. 6.2. It mainly consists of three parts: a ViT-based encoder for image feature extraction, a structure prompt encoder that generates structure prompts from the 2D coordinates of the input anchors, and a decoder aimed at mapping the input shapes to target faces.

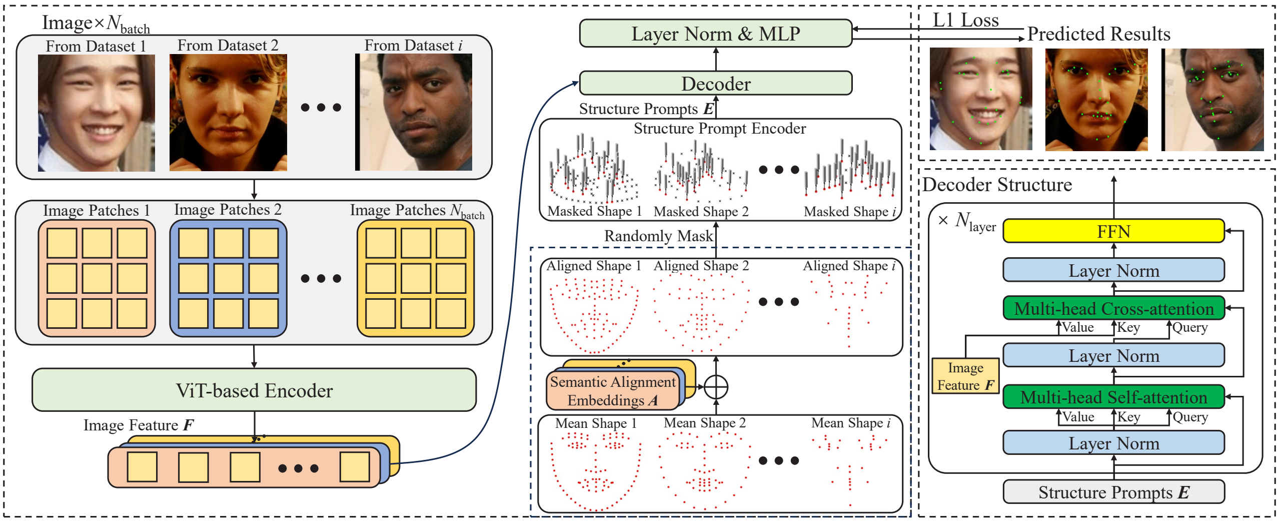


Figure 6.2: Overall training pipeline of TUFA. It mainly consists of three parts: a ViT-based encoder, a structure prompt encoder and a decoder.

ViT-based Encoder

We employ a ViT-based encoder to extract an image features $\boldsymbol{F}$ that later serve as essential clues in constructing the mapping between the pre-defined plane and various target faces. Given an input image $\boldsymbol{I} \in \mathbb{R}^{H_{\text{I}} \times W_{\text{I}} \times 3}$, the ViT-based encoder first splits it into regular patches with size $(H_{\text{P}}, W_{\text{P}})$. Then, each patch is projected into a vector with C dimensions by a CNN layer and is further added with learnable positional embeddings $\boldsymbol{P} \in \mathbb{R}^{L \times C}$ to retain spatial information, where $L = \frac{H_{\text{I}}}{H_{\text{P}}} \times \frac{W_{\text{I}}}{W_{\text{P}}}$. Finally, the encoder learns the image feature $\boldsymbol{F} \in \mathbb{R}^{L \times C}$ via the attention mechanism. $\boldsymbol{F}$ is further fed into the cross-attention block of the decoder.

Structure Prompt Encoder

A human face can be represented on a 2D plane because it is the surface of the human head and has a regular shape. During single dataset learning, to constrain the representation of the human face to the 2D plane, we calculate a statistical mean shape from the training samples, and the coordinates of the mean shape are normalized in the range of $[-1, 1]$. The mean shape, which explicitly represents the regular structure of faces, later serves as the anchors of this plane. By regressing target landmarks based on the structure prompts of these anchors, we can establish a mapping between the 2D plane and various faces. Moreover, the mean shape also ensures that the face structure representation on the plane is easily understood by humans, allowing the structure prompts to be readily edited. During multi-dataset learning, the additional semantic alignment embeddings $\boldsymbol{A}$ should be incorporated into the mean shapes for semantic alignment. The details of the semantic alignment embeddings $\boldsymbol{A}$ will be discussed in Section 6.2.2.

Before feeding the mean shapes into the structure prompt encoder, TUFA randomly masks part of these shapes during each iteration, retaining only N_{a} landmarks as the input anchors (N_{a} is set to 24, 25% of 98 landmarks). The use of the mean shape masking is threefold: 1) the masked mean shape is much sparser, encouraging TUFA to learn the mapping based on long-term landmark relation rather than the relation between neighbouring landmarks. It can effectively prevent overfitting and facilitate a coherent mapping for improved performance. 2) The computational complexity of the decoder during the training phase is significantly reduced with fewer structure prompts, which enhances training speed and reduces GPU memory consumption. 3) For multi-dataset learning, it ensures the consistency in the input anchors numbers across different datasets, allowing them to form a training batch.

To encode the 2D coordinates of masked shapes into high-dimensional vectors while retaining their geometric relations, we utilize cosine and sine functions. The high-dimensional vector can be written as:

$$\begin{aligned} E^{\mathrm{X}}_{(x,2c)} &= \sin(x/\tau^{(2c/0.5C)}), \\ E^{\mathrm{X}}_{(x,2c+1)} &= \cos(x/\tau^{(2c/0.5C)}), \end{aligned} \tag{6.1}$$

$$\begin{aligned} E^{\mathrm{Y}}_{(y,2c)} &= \sin(y/\tau^{(2c/0.5C)}), \\ E^{\mathrm{Y}}_{(y,2c+1)} &= \cos(y/\tau^{(2c/0.5C)}), \end{aligned} \tag{6.2}$$

where $\boldsymbol{E}^{\mathrm{X}}_{x}$ and $\boldsymbol{E}^{\mathrm{Y}}_{y}$ are two vectors with $0.5C$ dimensions, representing the landmark coordinate (x, y) on X-axis and Y-axis respectively. $c \in [0, C/4)$ is the index of dimension and τ is a hyperparameter that determines the wavelengths (τ is set to 10000 in this chapter). The final structure prompt $\boldsymbol{E}$ can be formulated as:

$$\boldsymbol{E}_{(x,y)} = \mathrm{Concat}(\boldsymbol{E}^{\mathrm{X}}_{x}; \boldsymbol{E}^{\mathrm{Y}}_{y}), \tag{6.3}$$

where Concat means concatenation process. For any fixed offset $({}_{\vartriangle}x, {}_{\vartriangle}y)$, the vectors $\boldsymbol{E}^{\mathrm{X}}_{x+{}_{\vartriangle}x}$ and $\boldsymbol{E}^{\mathrm{Y}}_{y+{}_{\vartriangle}y}$ can be written as:

$$\begin{bmatrix} E^{\mathrm{X}}_{(x+{}_{\vartriangle}x,2c)} \\ E^{\mathrm{X}}_{(x+{}_{\vartriangle}x,2c+1)} \end{bmatrix} = \begin{bmatrix} \cos(\frac{{}_{\vartriangle}x}{\tau^{2c/0.5C}}) & \sin(\frac{{}_{\vartriangle}x}{\tau^{2c/0.5C}}) \\ -\sin(\frac{{}_{\vartriangle}x}{\tau^{2c/0.5C}}) & \cos(\frac{{}_{\vartriangle}x}{\tau^{2c/0.5C}}) \end{bmatrix} \begin{bmatrix} E^{\mathrm{X}}_{(x,2c)} \\ E^{\mathrm{X}}_{(x,2c+1)} \end{bmatrix}, \tag{6.4}$$

$$\begin{bmatrix} E^{\mathrm{Y}}_{(y+\triangle y,2c)} \\ E^{\mathrm{Y}}_{(y+\triangle y,2c+1)} \end{bmatrix} = \begin{bmatrix} \cos(\frac{\triangle y}{\tau^{2c/0.5C}}) & \sin(\frac{\triangle y}{\tau^{2c/0.5C}}) \\ -\sin(\frac{\triangle y}{\tau^{2c/0.5C}}) & \cos(\frac{\triangle y}{\tau^{2c/0.5C}}) \end{bmatrix} \begin{bmatrix} E^{\mathrm{X}}_{(y,2c)} \\ E^{\mathrm{X}}_{(y,2c+1)} \end{bmatrix}. \tag{6.5}$$

Thus, $\boldsymbol{E}^{\mathrm{X}}_{x+\triangle x}$ and $\boldsymbol{E}^{\mathrm{Y}}_{y+\triangle y}$ can be viewed as a linear function of $\boldsymbol{E}^{\mathrm{X}}_{x}$ and $\boldsymbol{E}^{\mathrm{Y}}_{y}$ respectively, and the transformation matrix is determined by the corresponding offset. This property enables the 2D geometric relationships of the mean shape landmarks to be well retained in the high-dimensional vector for the learning of the mapping.

Decoder

The decoder is the key component for learning the mapping between the 2D plane and the target face. As shown at the bottom left in Fig. 3, the decoder mainly consists of three blocks: a multi-head self-attention (MSA) block, a multi-head cross-attention (MCA) block, and a feed-forward network (FFN). Besides, there is an extra LayerNorm **LN** before each block.

The mapping between the 2D plane and target faces should be determined by multiple anchors, not just a single one. Therefore, the MSA block is crucial because it enables the structure prompts to share their geometric information for the mapping determination. The key ($\boldsymbol{K}_z$), query ($\boldsymbol{Q}_z$), and value ($\boldsymbol{V}_z$) of the z-th head in MSA block can be calculated as:

$$\boldsymbol{K}_z = (\boldsymbol{T}_z + \boldsymbol{E}_z)\boldsymbol{W}^{\mathrm{k}}_z, \boldsymbol{Q}_z = (\boldsymbol{T}_z + \boldsymbol{E}_z)\boldsymbol{W}^{\mathrm{q}}_z, \boldsymbol{V}_z = \boldsymbol{T}_z\boldsymbol{W}^{\mathrm{v}}_z, \tag{6.6}$$

where $\boldsymbol{E} \in \mathbb{R}^{N_{\mathrm{a}} \times C}$ and $\boldsymbol{T} \in \mathbb{R}^{N_{\mathrm{a}} \times C}$ are the face structure prompts and the input to the MSA block respectively. $\boldsymbol{E}$ and $\boldsymbol{T}$ are further divided into N_{h} (N_{h} is the number of heads) sequences equally with $C_{\mathrm{h}} = C/N_{\mathrm{h}}$ dimensions. $\boldsymbol{W}^{\mathrm{k}}_z \in \mathbb{R}^{C_{\mathrm{h}} \times C_{\mathrm{h}}}$, $\boldsymbol{W}^{\mathrm{q}}_z \in \mathbb{R}^{C_{\mathrm{h}} \times C_{\mathrm{h}}}$, and $\boldsymbol{W}^{\mathrm{v}}_z \in \mathbb{R}^{C_{\mathrm{h}} \times C_{\mathrm{h}}}$ are three learnable matrices. The output of z-th head $\boldsymbol{H}_z$ can be calculated as:

$$\boldsymbol{H}_z = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_z\boldsymbol{K}^T_z}{\sqrt{C_{\mathrm{h}}}}\right)\boldsymbol{V}_z. \tag{6.7}$$

The final output of MSA block can be written as:

$$\mathcal{F}_{\mathrm{MSA}}(\boldsymbol{T}) = \mathrm{Concat}(\boldsymbol{H}_1; ...; \boldsymbol{H}_{N_{\mathrm{h}}})\boldsymbol{W}_{\mathrm{MSA}}, \tag{6.8}$$

where $\boldsymbol{W}_{\mathrm{MSA}} \in \mathbb{R}^{C \times C}$ is also a learnable matrix for linear projection.

The mapping also depends on the input face. Therefore, the MCA block is essential since it aggregates the image feature $\boldsymbol{F}$ learned by the ViT-based encoder based on the structure prompts, enabling the learned mapping to be case-dependent. The key ($\boldsymbol{K}_z'$), query ($\boldsymbol{V}_z'$) and value ($\boldsymbol{V}_z'$) of the z-th head in MCA block can be formulated as:

$$\boldsymbol{K}_z' = (\boldsymbol{F}_z + \boldsymbol{P}_z)\boldsymbol{W}_z^{\mathrm{k}\prime}, \boldsymbol{Q}_z' = (\boldsymbol{T}_z' + \boldsymbol{E}_z)\boldsymbol{W}_z^{\mathrm{q}\prime}, \boldsymbol{V}_z = \boldsymbol{F}_z\boldsymbol{W}_z^{\mathrm{v}\prime}, \tag{6.9}$$

where $\boldsymbol{T}' \in \mathbb{R}^{N_{\mathrm{a}} \times C}$ and $\boldsymbol{F} \in \mathbb{R}^{L \times C}$ are the input to MCA block and image feature respectively. We reuse the learnable positional embeddings $\boldsymbol{P} \in \mathbb{R}^{L \times C}$ in the ViT-based encoder to retain the spatial information of image. Similar to the MSA block, $\boldsymbol{T}'$, $\boldsymbol{F}$ and $\boldsymbol{P}$ are further divided into N_{h} sequences equally with $C_{\mathrm{h}} = C/N_{\mathrm{h}}$ dimensions. $\boldsymbol{W}_z^{\mathrm{k}\prime} \in \mathbb{R}^{C_{\mathrm{h}} \times C_{\mathrm{h}}}$, $\boldsymbol{W}_z^{\mathrm{q}\prime} \in \mathbb{R}^{C_{\mathrm{h}} \times C_{\mathrm{h}}}$, and $\boldsymbol{W}_z^{\mathrm{v}\prime} \in \mathbb{R}^{C_{\mathrm{h}} \times C_{\mathrm{h}}}$ are three learnable matrices. The output of z-th head $\boldsymbol{H}_z'$ can be written as:

$$\boldsymbol{H}_z' = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_z'\boldsymbol{K}_z'^{T}}{\sqrt{C_{\mathrm{h}}}}\right)\boldsymbol{V}_z'. \tag{6.10}$$

The final output of MCA block can be formulated as:

$$\mathcal{F}_{\mathrm{MCA}}(\boldsymbol{T}) = \mathrm{Concat}(\boldsymbol{H}_1'; ...; \boldsymbol{H}_{N_{\mathrm{h}}}')\boldsymbol{W}_{\mathrm{MCA}}, \tag{6.11}$$

where $\boldsymbol{W}_{\mathrm{MCA}} \in \mathbb{R}^{C \times C}$ is also a learnable matrix.

Moreover, a FFN block fuses the features aggregated from $\boldsymbol{F}$ in the channel-wise, further enhancing the expressive ability of the network. It allows the decoder to model a more fine-grained mapping between the target face and 2D plane.

Finally, the output features of the decoder are fed into an MLP for regressing the coordinates of the mapped anchors.

6.2.2 Semantic Alignment Embedding

The calculated mean shape reflects the statistically geometric relationship between the pre-defined landmark of the corresponding dataset, actually representing the semantic definitions of the landmarks. Based on the fact that human faces have a regular shape, the semantics of pre-defined landmarks across different datasets can be aligned on the same 2D plane, regardless of the variations in landmark definition and number. Although some landmarks from different datasets have similar definitions, there is still semantic variance between these landmarks because of varying annotation methods. The semantic variance can be represented as a geometric offset on the plane, but this offset is hard to be calculated manually.

Therefore, we introduce extra learnable semantic embeddings $\boldsymbol{A}_i \in \mathbb{R}^{N_{\mathrm{D}}^i \times 2}$ to the mean shape of each dataset, where i is the index of dataset and N_{D}^i is the number of the pre-defined landmarks

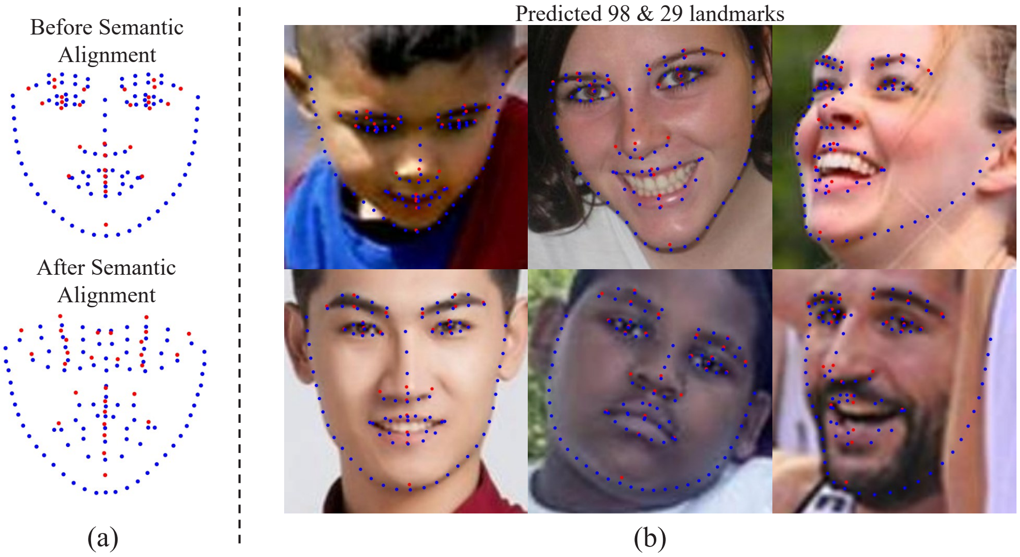


Figure 6.3: (a) the mean shapes of 29 & 98 landmarks before/after adding semantic alignment embeddings $\boldsymbol{A}$. (b) the predicted 29 & 98 landmarks on the test image. The **red** and **blue** points indicate the 29 landmarks and 98 landmarks respectively.

in the i-th dataset. $\boldsymbol{A}$ ensures that each landmark of the mean shapes has a corresponding learnable 2D offset. As shown in Eq. 6.4 and Eq. 6.5, adding $\boldsymbol{A}$ to the meanshape equals to applying a linear transformation to the initial structure prompts. Because the coordinates of mean shapes are normalized in the range of $[-1, 1]$, the transformation matrices are continuous and unique for any offset. It ensures that the structure prompts after linear transformation are also continuous and unique. As a result, it is possible for TUFA to find an optimal offset for the semantic alignment.

As shown in Fig. 6.3 (a), after adding the semantic alignment embeddings, the mean shapes of both the 98 and 29 landmarks still retain a clear face structure, which illustrates that these semantic alignment embeddings specifically incorporate the face structure information. The predicted results shown in Fig. 6.3 (b) demonstrate that even the landmarks from different datasets, defined similarly, still exhibit semantic variance due to different annotation methods. For instance, the eye corners in the 29 landmarks are always located above the eye corners in the 98 landmarks. After semantic alignment, the geometric relationship between the predicted 29 & 98 landmarks is consistent with the geometric relationships between the mean shapes of 29 & 98 landmarks. Therefore, with the semantic alignment embeddings, the semantics of the landmarks from different datasets have been aligned on a plane successfully.

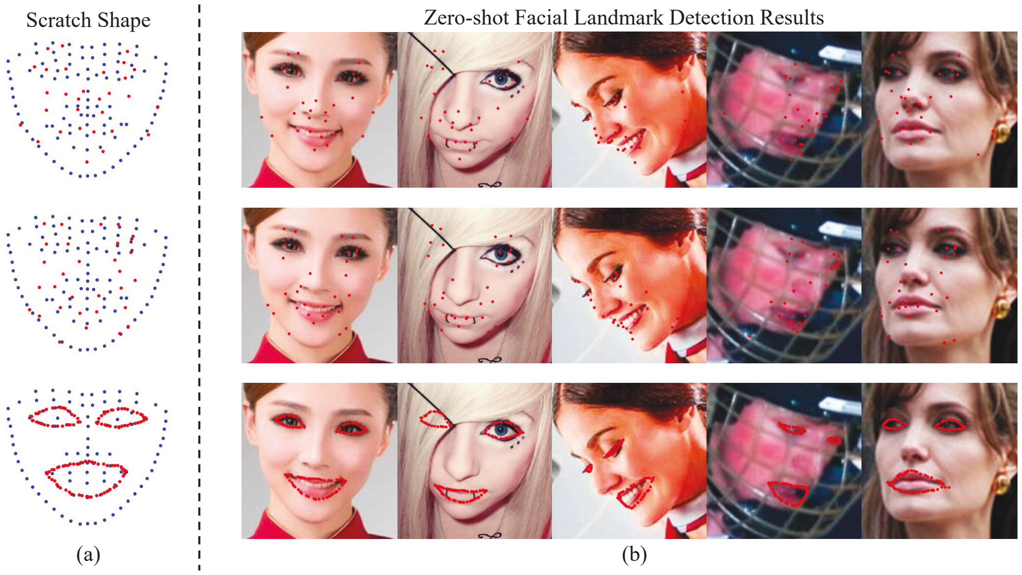


Figure 6.4: (a) the **red** points indicate the randomly generated scratch shape, and the **blue** points represent the mean shape of 98 landmarks, which serves as anchors during training. (b) the **red** points represent the zero-shot facial landmark detection results under various conditions.

6.2.3 Loss Function

When the input anchors are dense enough, TUFA can learn a mapping from a 2D plane to a very complex surface by minimizing the distance between the mapped anchors and the corresponding labeled landmarks. We measure the distance using L1 loss as follows:

$$\mathcal{L} = \frac{1}{N_{\text{batch}} N_{\text{a}}} \sum_{j=1}^{N_{\text{batch}}} \sum_{k=1}^{N_{\text{a}}} \left| (x_{\text{map}}^{jk}, y_{\text{map}}^{jk}) - (x_{\text{gt}}^{jk}, y_{\text{gt}}^{jk}) \right|, \tag{6.12}$$

where N_{batch} is the batch size, $(x_{\text{map}}^{jk}, y_{\text{map}}^{jk})$ is the position of the k-th mapped anchor in j-th sample, and $(x_{\text{gt}}^{jk}, y_{\text{gt}}^{jk})$ is the corresponding labeled position.

6.2.4 Zero-shot & Few-shot Facial Landmark Detection

The transformer-based decoder enables TUFA to accept an arbitrary number of structure prompts as the input, and the interpretable plane, learned based on the statistical mean shape, allows for easy editing of these structure prompts. Leveraging the structure prompts, the gap between the seen and unseen landmarks can be bridged. To locate an unseen landmark, we can determine its corresponding position on the plane, based on its geometric relationship with the anchors used during training. By encoding this 2D coordinate into a structure prompt, TUFA can locate the coordinate of the unseen landmark across different faces. As shown in Fig. 6.4, we randomly generate three scratch shapes and locate the corresponding landmarks on the target faces. We can clearly observe that the geometric relationship between the scratch shapes and the training anchors on the plane remains consistent with that geometric relationship on the target faces. Moreover, the semantics of the located landmarks also keep consistent across different faces. This demonstrates the editability of the predicted target in TUFA, as well as the successful implementation of zero-shot facial landmark detection.

The unified learning target also enables TUFA to be easily transferred to a group of newly defined facial landmarks in a few-shot manner. We first calculate a mean shape from the very few training samples. This mean shape is then added with a group of newly defined learnable semantic alignment embeddings. Finally, the pretrained TUFA is fine-tuned end-to-end on the new dataset without any changes of the structure. Despite the semantic difference in the newly defined landmarks, TUFA can still be efficiently transferred to the dataset. This is because TUFA explicitly learns the face structure, unifying the learning target across different datasets. Therefore, TUFA can easily inherit the knowledge learned from pre-trained datasets in few-shot learning, regardless of the differences in landmark semantics. Even with fewer training samples, TUFA significantly outperforms other state-of-the-art few-shot facial landmark detection methods.

6.3 Experiments

6.3.1 Datasets

- **WFLW** [43]: WFLW consists of 10,000 faces (7,500 for training and 2,500 for testing) from WIDER Face [110]. Each face is fully manual annotated with 98 landmarks and attribute labels. Compared to other datasets, the WFLW is more challenging because it contains a large number of samples under extreme conditions, such as heavy occlusion and profile view.

- **300W** [109]: 300W consists of 3,837 faces (3,148 for training and 689 for testing) from AFW [122], HELEN [123], LFPW [124] and IBUG [109]. Each face is annotated with 68 landmarks using a semi-automatic methodology [134]. The testing set can be further divided into a challenging subset (135 faces) and a common subset (554 faces). Moreover, 300W also provides additional 600 faces named as **300W-private** to test the generalization ability of the trained model.

- **Masked 300W** [68]: Masked 300W is a variant of 300W [109] used exclusively for testing the performance of facial landmark detection under heavy occlusion. It synthesizes 689 masked faces from the testing set of 300W, and the landmark annotation remains consistent with the original 300W.

- **MERL-RAV** [67]: MERL-RAV manually re-annotates 19,314 faces from AFLW [129] with 68 landmarks, providing 15,449 samples for training and 3,865 samples for testing. MERL-RAV further categorizes facial landmarks into unoccluded, externally occluded and self-occluded landmarks. Only unoccluded and externally occluded landmarks are provided with location information.

- **COFW** [8]: COFW collects 1,007 challenging faces with an average occlusion of over 23% from a variety of sources. It uses 500 of these faces and 845 samples from LFPW [124] training set for training, and tests on the remaining 507 faces. Each face is annotated with 29 landmarks.

- **COFW68** [125]: COFW68, a variant of COFW [8], is used for cross-dataset validation. The testing set of COFW is manually re-annotated with 68 landmarks in this variant.

- **CelebA-aligned** [135]: CelebA-aligned contains 10,000 identities with 200,000 face images. Each face is labeled with 5 landmarks and aligned to center. As suggested by [92], CelebA can be subdivided into three subsets: CelebA training set without MAFL (160,000 images), MAFL training set (19,000 images) and MAFL testing set (1,000 images).

6.3.2 Evaluation Metrics

Following previous works [43], [70], [133], [136], we use Normalized Mean Error (NME), Failure Rate (FR) and Area Under Curve (AUC) to quantitatively measure the performance of TUFA. NME is the mean of L2 distance between the predicted landmarks and the annotated landmarks. The mean distance is then normalized by a factor, denoted as d_{norm}. For *inter-pupil* NME, d_{norm} represents the distance between pupil centers, while for *inter-ocular* NME, it stands for the distance between outer eye corners. In the case of NME_{box}, d_{norm} is defined as the geometric mean of the labeled box, calculated as $d_{\text{norm}} = \sqrt{W_{\text{box}} \times H_{\text{box}}}$. FR_{α} represents the percentage of the testing cases in which the NME exceeds a certain threshold, denoted as α. AUC indicates the area beneath the Cumulative Errors Distribution (CED) curve from 0 to the threshold of FR α, which can be formulated as $\int_0^{\alpha} f(\epsilon)\, d\epsilon$.

6.3.3 Implementation Details

The input image is the face region cropped from the initial image and is then resized to a fixed size (256 × 256). For training data, we apply augmentation techniques, which include random translation (±10 pixels), random rotation (±30°), random scaling (±5%), random horizontal flipping (50%), random gray (20%), random brightness adjustment (50%, ±0.3), random occlusion (50%), random shearing (33%). In TUFA, We employ two types of encoders : ViT-S/16 and ViT-S/8 [73], both of which are pretrained on ImageNet [131] with DINO [138]. The parameters of the decoder and semantic alignment embeddings are initialized from scratch. By default, the batch size is set to 16 and the layer number of the decoder is set to 6. TUFA is trained with AdamW [132], setting the initial learning rate to 1×10^{-4} for ViT-S/8 and 5×10^{-5} for ViT-S/16. For multi-dataset learning, we employ the training sets of WLFW, 300W, MERL-RAV and COFW for training. The model is trained for 100 epochs, with the learning rate decaying by a factor of 0.1 at the 80th and 90th epochs respectively. The learned result is then tested on the WFLW, 300W, 300W-private, masked 300W, MERL-RAV, COFW and COFW68 **simultaneously**. For single-dataset learning, due to the smaller number of training samples, the model is trained for 140 epochs, with the learning rate decaying by a factor of 0.1 at the 120th and 130th epochs respectively.

6.3.4 Comparisons in Within-dataset Validation

WFLW

We report the inter-ocular NME, parameter size and flops of TUFA and other state-of-the-art methods in Table 6.1. $\text{FR}_{0.1}$ and $\text{AUC}_{0.1}$ are tabulated in Table 6.2. The TUFA learned from multiple datasets yields best performance in terms of NME on the full set and all subsets. The unique semantic alignment enables TUFA to maximize the utilization of extra face images. Even though TUFA employs fewer extra samples compared to other methods trained with multiple datasets, we still observe a significant improvement of 7.09% in NME compared to the TUFA trained with a single dataset. This demonstrates that TUFA can maximize the

Method	Backbone	type	Extra face images	Flops ↓	Params ↓	Inter-Ocular NME (%) ↓						
						Full	Pose	Exp.	Ill.	Mu.	Occ.	Blur
3FabRec [20]*	ResNet18	heatmap	2.1 millions	-	-	5.62	10.23	6.09	5.55	5.68	6.92	6.38
HRNet [55]	HRNetW18C	heatmap	N/A	4.75G	9.66M	4.60	7.94	4.85	4.55	4.29	5.44	5.42
ATF [71]*	HRNetW18C	heatmap	20,000	4.75G	9.66M	4.50	7.54	4.63	4.45	4.20	5.30	5.19
LUVLi [67]	8 DU-Net	heatmap	N/A	-	-	4.37	7.56	4.77	4.30	4.33	5.29	4.94
AWing [61]	4 Hourglass	heatmap	N/A	26.8G	24.15M	4.21	7.21	4.46	4.23	4.02	4.99	4.82
HIH [69]	2 Hourglass	heatmap	N/A	10.38G	14.47M	4.18	7.20	**4.19**	4.45	3.97	5.00	4.81
ADNet [70]	4 Hourglass	heatmap	N/A	17.04G	13.37M	4.14	6.96	4.38	4.09	4.05	5.06	4.79
FaRL [137]*	ViT-B/16	heatmap	20 millions	-	-	4.03	6.81	4.32	3.92	3.87	**4.70**	**4.54**
AV w. SAN [88]*	ResNet152	coordinate	120,000	33.87G	35.02M	4.39	8.42	4.68	4.24	4.37	5.60	4.86
SDFL [52]	HRNetW18C	coordinate	N/A	5.17G	-	4.35	7.42	4.63	4.29	4.22	5.19	5.08
SDL [50]	HRNetW18C	coordinate	N/A	-	-	4.21	7.36	4.49	4.12	4.05	4.98	4.82
SLPT [133]	HRNetW18C-lite	coordinate	N/A	6.12G	13.19M	4.14	6.96	4.45	4.05	4.00	5.06	4.79
SPIGA [54]	4 Hourglass	coordinate	N/A	-	-	4.06	7.14	4.46	4.00	**3.81**	4.95	4.65
DSLPT [136]	HRNetW18C	coordinate	N/A	7.83G	19.35M	4.01	6.87	4.29	3.99	**3.86**	4.79	4.66
TUFA	ViT-S/8	coordinate	N/A	35.17G	36.03M	4.23	7.23	4.53	4.17	4.12	5.05	4.78
TUFA*	ViT-S/16	coordinate	19,942	8.043G	36.00M	**4.00**	**6.57**	4.20	**3.90**	**3.86**	4.73	4.55
TUFA*	ViT-S/8	coordinate	19,942	35.17G	36.03M	**3.93**	**6.48**	**4.11**	**3.82**	**3.81**	**4.68**	**4.53**

Table 6.1: Performance comparison against state-of-the-art heatmap regression and coordinate regression methods on WFLW and its subsets. Corresponding parameter size, flops are also reported. Key: [**Best**, **Second Best**, *=trained with multiple datasets]

Metric	Method	Full	Pose	Exp.	Ill.	Mu.	Occ.	Blur
$FR_{0.1}$(%)↓	3FabRec*	8.28	34.35	8.28	6.73	10.19	15.08	9.44
	HRNet	4.64	23.01	3.50	4.72	2.43	8.29	6.34
	ATF*	2.52	13.19	2.23	2.44	**0.49**	5.03	3.88
	LUVLi	3.12	15.95	3.18	2.15	3.40	6.39	3.23
	AWing	2.04	9.20	**1.27**	2.01	**0.97**	4.21	2.72
	HIH	2.96	15.03	**1.59**	2.58	1.46	6.11	3.49
	ADNet	2.72	12.72	2.15	2.44	1.94	5.79	3.54
	FaRL*	1.76	-	-	-	-	-	-
	AV w. SAN*	4.08	18.10	4.46	2.72	4.37	7.74	4.40
	SDFL	2.72	12.88	**1.59**	2.58	2.43	5.71	3.62
	SDL	3.04	15.95	2.86	2.72	1.46	5.29	4.01
	SLPT	2.76	12.27	2.23	1.86	3.40	5.98	3.88
	SPIGA	2.08	11.66	2.23	1.58	1.46	4.48	**2.20**
	DSLPT	2.52	13.19	2.23	2.44	**0.97**	4.89	3.49
	TUFA (ViT-S/8)	2.44	12.27	1.91	2.15	1.46	5.30	3.49
	TUFA (ViT-S/16)*	**1.72**	**8.28**	**1.59**	**1.29**	1.46	**3.53**	2.59
	TUFA (ViT-S/8)*	**1.52**	**7.36**	**1.27**	**0.86**	1.46	**2.99**	**2.33**
$AUC_{0.1}$↑	3FabRec*	0.484	0.192	0.448	0.496	0.473	0.398	0.434
	HRNet	0.524	0.251	0.510	0.533	0.545	0.459	0.452
	ATF*	0.560	0.301	0.546	0.566	0.581	0.487	0.489
	LUVLi	0.557	0.310	0.549	0.584	0.588	0.505	0.525
	AWing	0.590	0.334	0.572	0.596	0.602	0.528	0.539
	HIH	0.597	0.342	**0.590**	0.606	0.604	0.527	0.549
	ADNet	0.602	0.344	0.523	0.580	0.601	0.530	0.548
	FaRL*	0.602	-	-	-	-	-	-
	AV w. SAN*	0.591	0.311	0.549	0.609	0.581	0.516	**0.551**
	SDFL	0.576	0.315	0.550	0.585	0.583	0.504	0.515
	SDL	0.589	0.315	0.566	0.595	0.604	0.524	0.533
	SLPT	0.595	0.348	0.574	0.601	0.605	0.515	0.535
	SPIGA	0.606	0.353	0.580	0.613	**0.622**	0.533	**0.553**
	DSLPT	**0.607**	0.353	0.586	**0.614**	**0.623**	**0.535**	0.549
	TUFA (ViT-S/8)	0.585	0.321	0.553	0.593	0.591	0.513	0.535
	TUFA (ViT-S/16)*	0.604	**0.359**	0.584	0.613	0.616	0.534	**0.551**
	TUFA (ViT-S/8)*	**0.610**	**0.371**	**0.592**	**0.620**	**0.622**	**0.540**	**0.553**

Table 6.2: Performance comparison against state-of-the-art heatmap regression and coordinate regression methods on WFLW and its subsets. Corresponding parameter size, flops are also reported. Key: [**Best**, **Second Best**, *=trained with multiple datasets]

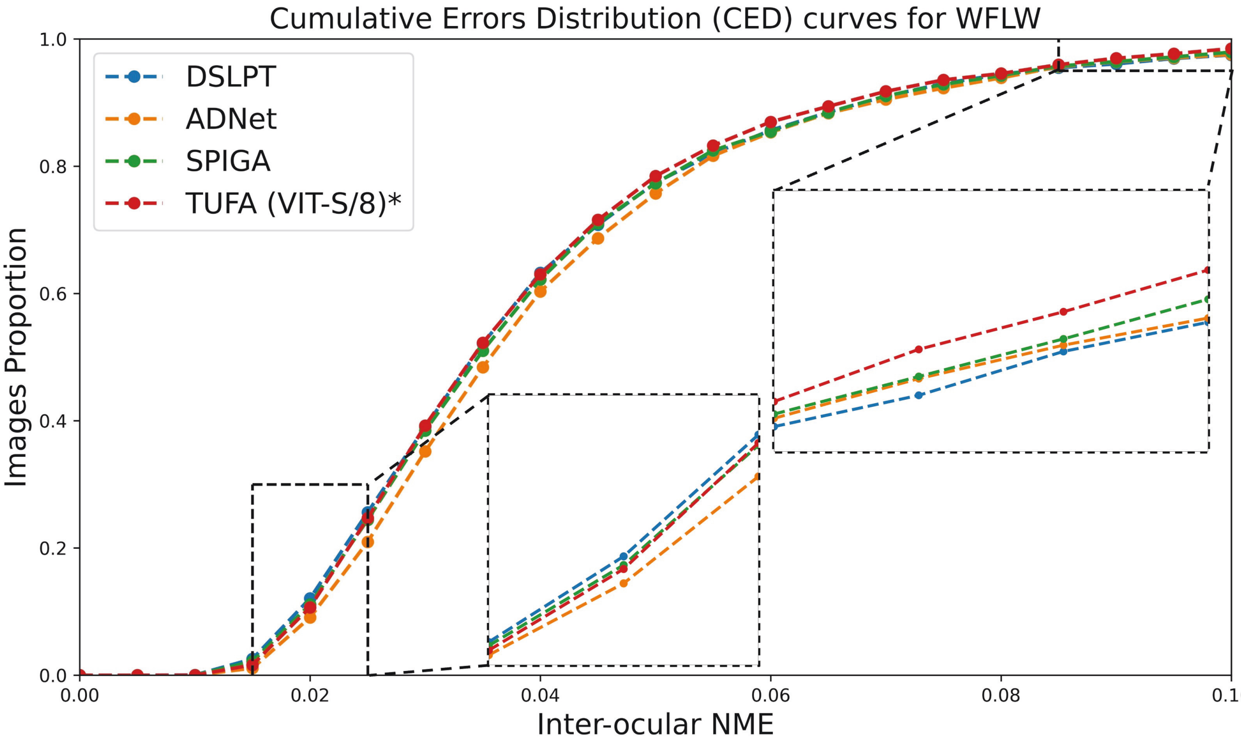


Figure 6.5: The comparisons of CED curves between TUFA and other state-of-the-art methods on WFLW full set.

utilization of extra training samples.

Most existing methods, such as DSLPT, SPIGA, and ADNet, achieve competitive performance based on a multi-stage approach, while TUFA locates facial landmarks using only a single stage. For a comprehensive analysis, we plot the CED curves of these multi-stage methods and TUFA (using multi-dataset learning) in Fig. 6.5. Note that the official implementation of ADNet performs slightly better than the results reported in their original paper. The CED curves of DSLPT and SPIGA are higher than that of TUFA in the range of $[0, 0.025]$. It suggests the proportion of the samples with very small NME predicted by DSLPT and SPIGA is larger than that of TUFA, despite that TUFA (using multi-dataset learning) performs better in NME. This indicates that these multi-stage facial landmark detection methods primarily reduce the NME on easy samples for better numerical results. However, their improvements under challenging conditions are relatively insignificant. Despite the fact that these conditions are long-tailed, they are critical as they ultimately determine the quality of facial landmark detection in real-world scenarios. By employing extra data, TUFA significantly improves the performance on these challenging conditions. As a result, TUFA achieves improvements of 26.9%, 36.9% and

33.3% in $\mathrm{FR}_{0.1}$ on full set, largepose subset and occlusion subset respectively compared to SPIGA.

Considering efficiency, we have also implemented a TUFA with ViT-S/16. Despite having comparable flops to other methods, it yields the second best performance in NME and $\mathrm{FR}_{0.1}$. Moreover, it does not require any post-processing as it regresses the coordinates of landmarks. Therefore, this TUFA runs faster than heatmap regression methods even though they have similar flops. The entire training process of this TUFA can be completed within 5 hours on a single A40 GPU.

300W

As tabulated in Table 6.3, multi-dataset learning also improves the performance of TUFA on common set, challenging set and full set by 8.16%, 9.18% and 8.67% respectively in the metric of NME. Heatmap regression methods, such as ADNet and Awing, commonly demonstrate superior performance on 300W, especially the challenging subset, compared to coordinate regression methods. This is because they provide semantic supervision to the network by encoding the annotated landmarks into heatmaps, which delivers better performance with limited training samples. However, the unique properties of TUFA successfully address this limitation of coordinate regression methods, achieving the best performance in NME on the challenging set, as low as 4.45%, even though TUFA is a single-stage facial landmark detection method. This demonstrates that the multi-dataset learning of TUFA promises a very competitive generalization ability. Compared to FaRL and ATF, which are also trained with multiple face datasets, TUFA outperforms them by 4.22% and 6.94% respectively in NME despite fewer extra faces used in training.

COFW

With the very limited number of training samples and an average occlusion of over 23% on the testing faces, COFW presents a significant challenge for all facial landmark detection methods. As shown in Table 6.4, despite the competitive performance on other datasets, existing state-of-the-art methods, such as DSLPT, ADNet and SDFL, often suffer from overfitting on this dataset. Similarly, TUFA, when trained with a single dataset, also tends to overfit the training set, achieving only 3.40% in the metric of NME. Nevertheless, TUFA can successfully address the problem by utilizing knowledge from other face datasets, thereby significantly improving the

Method	Inter-Ocular NME (%) ↓		
	Common	Challenging	Fullset
AV w. SAN [88]*	3.21	6.49	3.86
3FabRec [20]*	3.36	5.74	3.82
LAB [43]	2.98	5.19	3.49
DeCaFA [60]	2.93	5.26	3.39
HIH [69]	2.93	5.00	3.33
HRNet [55]	2.87	5.15	3.32
SDFL [52]	2.88	4.93	3.28
LUVLi [67]	2.76	5.16	3.23
ATF [71]*	2.75	4.86	3.17
AWing [61]	2.72	**4.53**	3.07
SDL [50]	2.62	4.77	3.04
ADNet [70]	**2.53**	4.58	**2.93**
SLPT [133]	2.75	4.90	3.17
FaRL [137]*	2.70	4.64	3.08
SPIGA [54]	2.59	4.66	2.99
DSLPT [136]	**2.57**	4.69	2.98
TUFA (ViT-S/8)	2.82	4.90	3.23
TUFA (ViT-S/16)*	2.68	4.58	3.05
DSLPT (ViT-S/8)*	2.59	**4.45**	**2.95**

Table 6.3: Performance comparison against state-of-the-art methods on 300W. Key: [**Best**, **Second Best**, *=trained with multiple datasets]

Method	Inter-Ocular		Inter-Pupil	
	NME(%)↓	FR$_{0.1}$(%)↓	NME(%)↓	FR$_{0.1}$(%)↓
LAB [43]	3.92	0.39	-	-
Wing [44]	-	-	5.44	3.75
DCFE [38]	-	-	5.27	7.29
SDFL [52]	3.63	**0.00**	-	-
HRNet [55]	3.45	**0.20**	-	-
ATF [71]*	3.32	-	-	-
AWing [61]	-	-	4.94	0.99
ADNet [70]	-	-	4.68	**0.59**
SLPT [133]	3.32	**0.00**	4.79	1.18
DSLPT [136]	3.33	**0.20**	4.79	1.36
TUFA (ViT-S/8)	3.40	**0.20**	4.91	1.18
TUFA (ViT-S/16)*	**3.18**	**0.20**	**4.58**	**0.39**
TUFA (ViT-S/8)*	**3.07**	**0.20**	**4.43**	**0.39**

Table 6.4: Performance comparison against state-of-the-art methods on COFW. Key: [**Best**, **Second Best**, *=trained with multiple datasets]

metric by 9.71% in NME and yielding the best performance. Although ATF also leverages extra datasets to train HRNet, it fails to consider the semantic differences between the landmarks with similar definitions, which inadvertently introduces noise into the training and leads to a less significant improvement.

MERL-RAV

The numerical results on MERL-RAV are shown in Table 6.5. The proportions of unoccluded, externally occluded and self-occluded landmarks in MERL-RAV full set are 76.56%, 10.81% and 12.63% respectively. Because the position annotations of the self-occluded landmarks are not provided, the numerical results on the full set and its subsets exclude the results of these landmarks. Therefore, these numerical results are largely determined by the performance on the easy samples. As mentioned before, multi-stage methods tend to perform better on the easy samples compared to one-stage methods. Therefore, TUFA only demonstrates comparable performance to these multi-stage methods. Nevertheless, when we investigate the externally

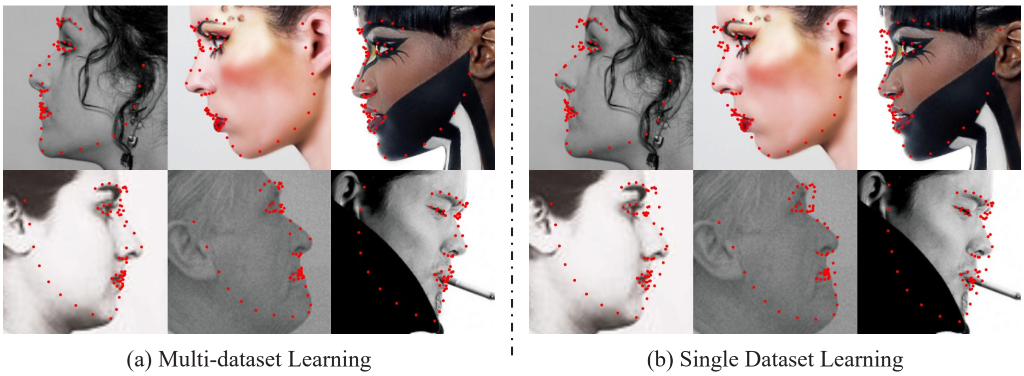


Figure 6.6: (a) Results in profile view predicted by the TUFA with multi-dataset learning. (b) Results in profile view predicted by the TUFA with single dataset learning. The **red** points represent the predicted landmarks.

occluded landmarks separately, we can find that the multi-dataset learning still improves the performance on these challenging landmarks effectively, despite the relatively fewer extra samples provided by WFLW, 300W and COFW compared to the training set of MEAL-RAV. Moreover, the qualitative results shown in Fig. 7 illustrate TUFA performs well on the self-occluded landmarks based on the knowledge of other datasets, even though MERL-RAV does not provide their position labels for training.

6.3.5 Comparisons in Cross-dataset Validation

COFW68

The cross-dataset validation aims at evaluating the generalization ability of facial landmark detection methods. In the case of TUFA with single dataset learning, the model is trained on 300W training set. For TUFA trained with multiple datasets, the model used in four within-dataset validations is directly evaluated on COFW68. The evaluation results are show in Table 6.6. Despite not re-training the TUFA for COFW68, it still outperforms existing state-of-the-art methods by a significant margin. Compared to single dataset learning, the multi-dataset learning approach of TUFA demonstrates much stronger generalization ability, achieving an impressive improvement of 9.98% in NME.

Masked 300W

The setting of cross-dataset validation on Masked-300W is kept consistent with the setting on COFW68, and the evaluation results are displayed in Table 6.7. The testing faces are with an

Method	Full Set		Frontal Subset		Half-Profile Subset		Profile SubSet		Unoccluded		Externally Occluded	
	$\text{NME}_{\text{box}} \downarrow$	$\text{AUC}_{\text{box}}^{0.07} \uparrow$	$\text{NME}_{\text{box}} \downarrow$	$\text{AUC}_{\text{box}}^{0.07} \uparrow$	$\text{NME}_{\text{box}} \downarrow$	$\text{AUC}_{\text{box}}^{0.07} \uparrow$	$\text{NME}_{\text{box}} \downarrow$	$\text{AUC}_{\text{box}}^{0.07} \uparrow$	$\text{NME}_{\text{box}} \downarrow$	$\text{AUC}_{\text{box}}^{0.07} \uparrow$	$\text{NME}_{\text{box}} \downarrow$	$\text{AUC}_{\text{box}}^{0.07} \uparrow$
DU-Net [57]	1.99%	71.80%	1.89%	73.25%	2.50%	64.78%	1.92%	72.79%	-	-	-	-
LUVLi [67]	1.61%	77.08%	1.74%	75.33%	1.79%	74.69%	1.25%	82.10%	1.60%	-	3.53%	-
SLPT [133]	**1.51%**	78.33%	**1.62%**	76.82%	**1.68%**	**76.01%**	1.21%	82.74%	**1.50%**	**79.22%**	3.33%	**56.24%**
SIPGA [54]	**1.51%**	**78.47%**	**1.62%**	**76.96%**	**1.68%**	75.64%	**1.19%**	**83.00%**	-	-	-	-
DSLPT [136]	**1.48%**	**78.87%**	**1.60%**	**77.24%**	**1.64%**	**76.58%**	**1.16%**	**83.46%**	**1.48%**	**79.37%**	**3.26%**	**57.01%**
TUFA (ViT-S/8)	1.55%	77.93%	1.65%	76.42%	1.72%	75.38%	1.23%	82.46%	1.56%	78.29%	3.30%	56.10%
TUFA (ViT-S/16)*	1.61%	77.09%	1.72%	75.52%	1.78%	74.57%	1.28%	81.73%	1.63%	77.11%	3.31%	56.12%
TUFA (ViT-S/8)*	1.54%	78.01%	1.65%	76.48%	1.71%	75.50%	1.22%	82.55%	1.56%	78.17%	**3.21%**	**57.01%**

Table 6.5: NME_{box} and $\text{AUC}_{\text{box}}^{0.07}$ on the MERL-RAV full set and its subsets. The NME_{box} and $\text{AUC}_{\text{box}}^{0.07}$ of unoccluded and externally occluded landmarks are also reported. Key: [**Best**, **Second Best**, *=trained with multiple datasets]

Method	NME(%)↓	$\text{FR}_{0.1}$(%)↓
CFSS [12]	6.28	9.07
ODN [48]	5.30	-
AV w. SAN [88]*	4.43	2.82
LAB [43]	4.62	2.17
GlomFace [19]	4.21	0.79
SDL [50]	4.22	0.39
SDFL [52]	4.18	**0.00**
SLPT [133]	4.10	0.59
DSLPT [136]	4.03	**0.20**
SPIGA [54]	3.93	-
TUFA (ViT-S/8)	4.11	**0.20**
TUFA (ViT-S/16)*	**3.72**	**0.00**
TUFA (ViT-S/8)*	**3.70**	**0.00**

Table 6.6: Performance comparison of *cross*-dataset validation on COFW68. Key: [**Best**, **Second Best**, *=trained with multiple datasets]

Method	Inter-Ocular NME (%) ↓		
	Common	Challenging	Fullset
CFSS [12]	11.73	19.98	13.35
Hourglass [56]	8.17	13.52	9.22
MDM [36]	7.66	11.67	8.44
FAN [58]	7.36	10.81	8.02
LAB [43]	6.07	9.59	6.76
SAAT [68]†	5.42	11.36	6.58
GlomFace [19]†	5.29	8.81	5.98
DSLPT [136]	6.01	10.19	6.83
DSLPT [136]†	**4.78**	**8.10**	**5.42**
TUFA (ViT-S/8)	7.12	9.96	7.67
TUFA (ViT-S/16)*	5.21	8.25	5.80
TUFA (ViT-S/8)*	**4.98**	**8.08**	**5.58**

Table 6.7: Performance comparison of *cross*-dataset validation on Masked 300W. Key: [**Best**, **Second Best**, *=trained with multiple datasets, †=data augmentation adjustment]

average occlusion of over 50%, leading to a significant domain gap between the training and testing faces. To minimize this gap, existing state-of-the-art methods have optimized the data augmentation, randomly masking each training sample with several blocks of varying sizes. Although this approach can improve the quantitative results in cases with heavy occlusion, it also results in performance degradation under other conditions. Without any modification in data augmentation techniques, the TUFA learned from multiple datasets still achieves an impressive result, as low as 5.58% in NME. This outcome also demonstrates the remarkable generalization ability and robustness of TUFA.

300W-private

We carry out cross-dataset validation on 300W-private using the same settings as those for COFW and Masked 300W. Similarly, the improvement brought to TUFA by multi-dataset learning is also significant, as demonstrated in Table 6.8. Even though most samples from 300W-private are under common conditions, which are easily handled by other multi-stage methods, TUFA still yields the best performance across all metrics on the full set. The $\mathrm{FR}_{0.08}$

Method	Indoor subset			Outdoor subset			Full set		
	NME↓	AUC$_{0.08}$ ↑	FR$_{0.08}$ ↓	NME↓	AUC$_{0.08}$ ↑	FR$_{0.08}$ ↓	NME↓	AUC$_{0.08}$ ↑	FR$_{0.08}$ ↓
DAN [42]	-	-	-	-	-	-	4.30%	47.00%	2.67%
SHN [139]	4.10%	-	-	4.00%	-	-	4.05%	-	-
DCFE [38]	3.96%	52.28%	2.33%	3.81%	52.56%	1.33%	3.88%	52.42%	1.83%
SPIGA [54]	**3.43%**	**57.35%**	1.00%	**3.43%**	**57.17%**	**0.33%**	**3.43%**	**57.27%**	0.67%
DSLPT [136]	3.47%	56.60%	**0.33%**	3.47%	56.68%	**0.00%**	3.47%	56.64%	**0.17%**
TUFA (ViT-S/8)	3.86%	51.98%	1.33%	3.86%	51.90%	1.00%	3.86%	51.94%	1.17%
TUFA (ViT-S/16)*	3.51%	56.09%	**0.33%**	3.52%	55.96%	0.67%	3.52%	56.03%	**0.50%**
TUFA (ViT-S/8)*	**3.40%**	**57.48%**	**0.00%**	**3.42%**	**57.31%**	**0.33%**	**3.41%**	**57.40%**	**0.17%**

Table 6.8: Performance comparison of *cross*-dataset validation on 300W-private. Key: [**Best**, **Second Best**, *=trained with multiple datasets]

of TUFA is as low as 0.17%, which means only one sample failed to be aligned by TUFA in the entire dataset. This demonstrates that TUFA is highly robust compared to all existing methods.

6.3.6 Few-shot Facial Landmark Detection

We pretrain TUFA (ViT-S/8) on COFW, 300W and MERL-RAV (**29 & 68 landmarks**) and implement few-shot facial landmark detection on WFLW (**98 landmarks**) as described in Section 6.2.4. Note that there are large differences between the semantic definitions of 29 & 68 landmarks and 98 landmarks. We randomly select a certain number of training samples in each experiment. With varying training set sizes, we repeat each experiment five times, and the mean NME with each training set size is reported in Table 6.9. With only *10* training samples, TUFA still demonstrates superior performance compared to other methods, outperforming 3FabRec and He et al. trained with *7500* annotated faces. When we fine-tune TUFA on WFLW with 100% of the training set, the NME on the testing set drops to as low as 3.99%. This performance is comparable to the NME of the TUFA when trained with all four datasets together. This result indicates that TUFA can effectively leverage the knowledge learned from pretraining. This is achieved because TUFA successfully unifies the learning target of different facial landmark detection datasets, rather than treating landmark regression as an independent target. Because of this unique property, TUFA can be transferred to a group of newly defined landmarks efficiently, even with a very limited number of annotated samples. 3FabRec, Autolink

Table 6.9: Performance comparison in inter-ocular NME (%)↓ with reduced training set on WFLW full set. Key: [**Best**]

Method	Training set size						
	1	10	20	50	5%	20%	100%
AV w. SAN [88]	-	-	-	-	-	6.00	4.39
Xiao et al. [140]	43.0	21.9	19.3	17.6	10.6	7.08	5.62
Autolink [26]	14.9	13.5	13.3	11.2	7.68	7.31	6.35
3FabRec [20]	15.8	9.66	-	8.39	7.68	6.51	5.62
He et al. [21]	12.4	9.19	8.62	7.90	6.22	5.61	5.38
TUFA	**7.62**	**5.20**	**4.90**	**4.77**	**4.43**	**4.28**	**3.99**

and He et al. set image reconstruction as a pre-text task to encourage model to perform better with very limited training images. However, He et al. also found that this pre-text task may also lead to performance degradation when the number of training samples is sufficient. The main reason is that the learning targets of image reconstruction and facial landmark detection are not consistent.

6.3.7 Zero-shot Facial Landmark Detection

CelebA-aligned

To evaluate the semantic consistency of the landmarks predicted by TUFA in a zero-shot manner, we follow the same evaluation protocol as other unsupervised landmark detection methods [25], [24], [26] used on CelebA-aligned [135]. We randomly generate a scratch shape with a certain number of points N_{pre}. Then, the TUFA (ViT-S/8), trained with the four datasets, directly predicts the corresponding landmarks on both MAFL training and testing subset. The predicted coordinates on MAFL training subset are used to calculate a matrix that projects the coordinates to the positions of the labeled landmarks. We use this matrix to transfer the predicted landmark coordinates on MAFL testing subset to the coordinates of the labeled landmarks. Finally, we quantitatively measure the performance with inter-ocular NME.

We conduct this experiment 10 times with 10 randomly generated scratch shapes. The mean and variance of the NMEs are reported in Table 6.10. Although TUFA is not trained with the

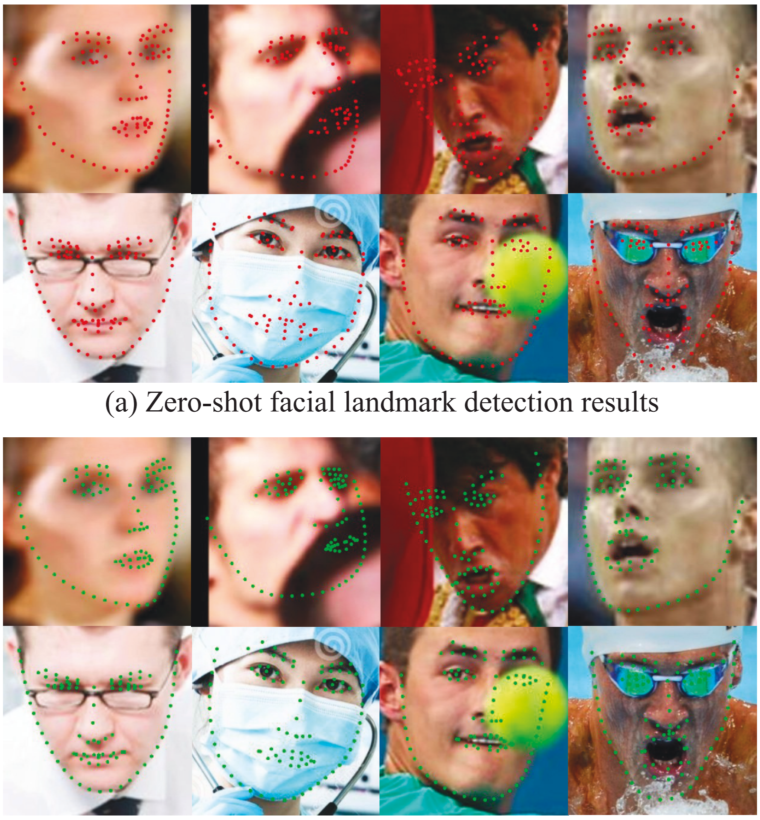
Table 6.10: Performance of TUFA in inter-ocular NME on CelebA under the setting of zero-shot facial landmark detection. The evaluation results of state-of-the-art self-supervised and unsupervised methods also reported. Key: [**Best**]

Method	type	NME ↓ (N_{pre}=10)
Thewlis et al. [22]	unsupervised	7.95%
Zhang et al. [23]	unsupervised	3.46%
Lorenz et al. [24]	unsupervised	3.24%
IMM [25]	unsupervised	3.19%
AutoLink [26]	unsupervised	3.92±0.69%
Mallis et al. [27]	self-supervised	3.83%
TUFA	zero-shot	**2.65±0.33%**

160,000 in-domain faces from CelebA training set, unlike other self-supervised or unsupervised methods, it still yields the best performance, significantly outperforming Mallis et al. and IMM, by 30.81% and 16.93% respectively, in terms of NME. This result demonstrates the landmarks predicted by TUFA in the zero-shot manner display better semantic consistency in various cases, as compared to other methods. Moreover, the semantic definitions of the landmarks predicted by TUFA can be assigned by humans via a manually generated shape, whereas those in self-supervised or unsupervised methods are learned randomly. Consequently, TUFA has a much broader application range.

WFLW

To better evaluate the performance of zero-shot facial landmark detection in practical application, we carry out an additional experiment. We first train TUFA (ViT-S/8) on MERL-RAV (**68 landmarks**). Then, we align the mean shape on WFLW (**98 landmarks**) to the mean shape on MERL-RAV using affine transformation. Finally, we evaluate TUFA on the WFLW full set directly without re-training. Since TUFA is the first framework for zero-shot facial landmark detection, we can only compare TUFA in the zero-shot setting with some fully-supervised methods to quantitatively demonstrate its performance. The evaluation results of zero-shot facial landmark detection with TUFA, and some fully supervised methods released before 2018, are listed in Table 6.11. Despite the large disparity in landmark semantics and numbers between WFLW and MERL-RAV, TUFA in the zero-shot setting still outperforms SDM, CFSS



(a) Zero-shot facial landmark detection results

(b) Ground truth

Figure 6.7: Zero-shot facial landmark detection results predicted by TUFA on WFLW. The red points and green points represent predicted landmarks and ground truth respectively.

and DVLN by 42.66%, 34.95% and 2.96% respectively in the metric of NME. The qualitative results shown in Fig. 6.7 demonstrate the semantics of the landmarks predicted by the zero-shot facial landmark detection remain highly consistent, regardless of heavy occlusion or profile view. Therefore, TUFA can be easily applied to a group of newly defined landmarks without any annotated training samples and re-training processes, significantly broadening the application range of existing facial landmark detection methods. Moreover, we can also conclude that TUFA learns a more universal knowledge for facial landmark detection, which explicitly represents the regular structure of human face.

6.3.8 Ablation Studies

Influence of the number of training datasets

The performance of TUFA, when trained with different numbers of datasets, is tabulated in Table 6.12. The quantitative results on each dataset are significantly improved as the num-

Method	Full	Pose	Exp.	Ill.	Mu.	Occ.	Blur
Fully-supervised Facial Landmark Detection							
SDM [9]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
CFSS [12]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
DVLN [141]	6.08	11.54	6.78	**5.73**	5.98	7.33	6.88
Zero-shot Facial Landmark Detection							
TUFA	**5.90**	**10.03**	**6.16**	5.78	**5.50**	**6.68**	**6.69**

Table 6.11: Performance of TUFA in inter-ocular NME (%)↓ on WFLW under the setting of zero-shot facial landmark detection. The evaluation results of some fully supervised methods before 2018 are also reported. Key: [**Best**]

Method	Training Datasets				NME (%)↓			
	300W	COFW	WFLW	MERL	300W	COFW	WFLW	MERL
TUFA1 (ViT-S/8)	✓	-	-	-	3.23	-	-	-
TUFA2 (ViT-S/8)	✓	✓	-	-	3.14	3.27	-	-
TUFA3 (ViT-S/8)	✓	✓	✓	-	3.08	3.18	4.17	-
TUFA4 (ViT-S/8)	✓	✓	✓	✓	**2.95**	**3.07**	**3.93**	**1.54**

Table 6.12: The influence of different numbers of datasets used in training. The inter-ocular NME on 300W, COFW and WFLW, and the NME_{box} on MERL-RAV are reported. Key: [**Best**]

ber of training datasets increases, despite the large difference between their annotation forms. This demonstrates TUFA effectively unifies the learning target of different datasets and maximizes the utilization of their knowledge. Therefore, even though COFW only provides 1,345 labeled faces, it still improves NME from 3.23% to 3.14% on 300W. This improvement is quite competitive in facial landmark task. MERL-RAV provides the largest number of annotated faces, leading to the most significant improvement on 300W, COFW and WFLW. The unique property of TUFA also enables its performance to be further improved with the release of more high-quality facial landmark detection datasets in the future.

Method	Semantic Alignment Embeddings	NME (%)↓						
		300W	COFW	WFLW	MERL	COFW68	300W-M	300W-P
TUFA1 (ViT-S/8)	w/o	3.42	3.08	3.97	1.57	3.71	5.87	4.08
TUFA2 (ViT-S/8)	w	**2.95**	**3.07**	**3.93**	**1.54**	**3.70**	**5.58**	**3.40**

Table 6.13: The influence of semantic alignment embeddings. The inter-ocular NME on 300W, COFW, WFLW, COFW68, Masked 300W and 300W-private, and the NME_{box} on MERL-RAV are reported. Key: [**Best**, 300W-M=masked 300W, 300W-P=300W-private]

Influence of the semantic alignment embeddings

To further investigate the influence of semantic alignment embeddings, we also implement a version of TUFA that does not incorporate these embeddings. Instead, we align the mean shapes from different datasets using affine transformation and train TUFA using these aligned mean shapes, without adding the semantic alignment embeddings. The comparison results between the TUFA with/without these embeddings are shown in Table 6.13. The aligned mean shape based on affine transformation cannot explicitly represent the semantic differences between landmarks. Consequently, it inevitably introduces a quantitative error into the predicted results. Because 300W, 300W-private and Masked 300W are labeled using semi-supervised methodology [134], their annotations exhibit much smaller variance than manual annotations. Therefore, the quantitative errors result in significant performance degradation on these three datasets. With an average occlusion of over 23%, the manually annotated landmarks on COFW and COFW68 exhibit much larger variance than other datasets. Contrary to the results on 300W, 300W-private and Masked 300W, the evaluation results on these datasets are less sensitive to the quantitative errors. Nevertheless, the absence of semantic alignment still leads to slight performance degradation. Overall, the subtle semantic differences between the landmarks with similar definitions cannot be ignored. The semantic alignment embeddings, while only slightly increasing the number of parameters in the training process, effectively eliminate the quantitative error caused by these semantic differences.

Metric	masking ration						
	0%	25%	50%	75%	80%	85%	90%
NME(%)↓	4.265	4.256	4.233	**4.229**	**4.229**	4.241	4.682
$\mathrm{FR}_{0.1}$(%)↓	2.800	2.360	2.360	2.440	**2.320**	2.440	3.920
$\mathrm{AUC}_{0.1}$↑	0.582	0.582	0.584	**0.585**	0.584	0.582	0.546

Table 6.14: The influence of masking ratio on WFLW. The inter-ocular NME is reported. Key: [**Best**]

Influence of the masking ratio

To further explore the influence of masking ratio, we train TUFA with different masking ratios on WFLW. The results are reported in Table 6.14. As the masking ratio increases, the performance of TUFA is improved. The primary reason is that randomly masking a certain ration of landmarks encourages TUFA to learn the mapping using the relationship between distant landmarks. This relationship ensures a more coherent mapping between the 2D plane and target faces, while the landmarks that are close together often introduce bias into the learned mapping. The results show TUFA works well within the range of $[50\%, 85\%]$. If the masking ratio exceeds 85%, the number of remaining landmarks, which serves as anchors during training, is insufficient to construct a mapping from a 2D plane to a highly complex face surface. As a result, the performance of TUFA degrades significantly.

Cross attention visualization

We visualize the cross attention maps of a scratch shape predicted by TUFA in Fig. 6.8. The attention maps indicate the mean weights from the structure prompts to image patches in MCA blocks, and the definition of the scratch is shown in the second row of Fig. 6.4. Even though the landmarks of the scratch shape are unseen during training, they can still guide TUFA to focus on the corresponding parts of faces regardless of various poses and conditions. It reveals the reliability of the zero-shot facial landmark detection achieved by TUFA. Moreover, if the corresponding parts on faces are occluded or lack significant features (row 3, landmark 3 and 6), TUFA can also look at nearby facial components and utilize their relative positions for landmark locating. That is why TUFA demonstrates much stronger robustness than other methods in very challenging cases.

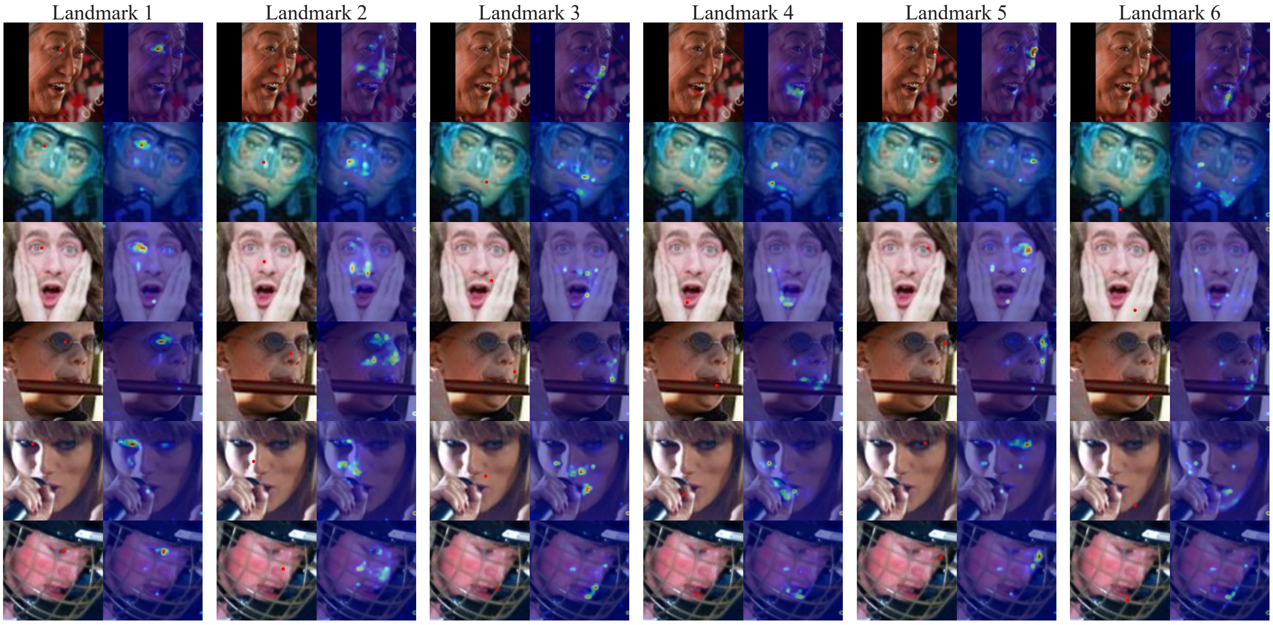


Figure 6.8: The zero-shot landmark predicted results of TUFA (column 1, 3, 5, 7, 9 and 11) and the corresponding mean cross attention maps (column 2, 4, 6, 8, 10 and 12). The **red** points represent the predicted landmarks.

6.4 Conclusion

In this chapter, we introduce a task-agnostic, unified face alignment (facial landmark detection) framework, named TUFA, that notably enhances the performance of both many-shot and few-shot facial landmark detection, while also implementing zero-shot facial landmark detection for the first time. Compared to existing facial landmark detection methods, TUFA represents a significant breakthrough in three critical aspects: 1) TUFA successfully models a mapping between an interpretable plane and target faces, and further bridges the gap between the seen and unseen facial landmarks using the proposed face structure prompts. This makes TUFA the first framework capable of tackling the challenge of zero-shot facial landmark detection. 2) TUFA successfully unifies the learning targets of various facial landmark detection datasets and mitigates the noise introduced by different annotation methods through the utilization of the proposed semantic alignment embeddings. Thus, we provide a pre-trained model with very robust generalization ability based on the multi-dataset learning. 3) the unified learning target also enables the learned knowledge to be easily transferred to a new set of defined landmarks. Consequently, TUFA significantly boosts the performance of few-shot facial landmark detection. Overall, we believe the unique properties of TUFA can propel facial landmark detection to a new stage and provide a competitive baseline for future works.

Chapter 7

Conclusions and Future Work

In this chapter, we conclude the work of this thesis and introduce the potential future work.

7.1 Summary of Outcomes

Facial landmark detection has been investigated more than 20 years, yet it still suffers from fragile robustness and reliability in real scenarios. This thesis focuses on improving the robustness and reliability of existing facial landmark detection methods under three very challenging conditions: mobile devices with limited computational capacity, heavy occlusion, and unseen facial landmarks. To boost their performance, we explore four aspects: utilizing boundary information in low-level features and multi-task learning to improve efficiency, learning case-dependent inherent relationships between landmarks, dynamically adjusting the receptive field according to the predicted uncertainty, and transforming landmark regression into an agnostic and unified learning task.

To validate the effectiveness of boundary information contained in low-level features and multi-task learning to efficient facial landmark detection, we propose an Alignment & Tracking & Pose Network (ATPN) and carry out extensive experiments on several benchmarks. The experimental results show that this boundary information can significantly boost the robustness and reliability of facial landmark detection without increasing too much parameters and computational complexity. It illustrates that this method is quite suitable for the lightweight model. Moreover, we also find the features learned from facial landmark detection can improve the performance for other face related tasks, such as face tracking and head pose estimation.

Therefore, we also design a new training strategy for multi-task learning to make full use of the features learned from facial landmark detection. With the limited computational complexity and number of parameters, ATPN achieves superior performance compared to other lightweight models.

Using the sparse local patch and attention mechanism, we successfully enable a model, the Sparse Local Patch Transformer (SLPT), to learn the case-dependent inherent relationships, which are crucial for the performance of facial landmark detection. With this clue, the model can act as human, using the relative positions of those occluded landmarks to the easily identified landmarks for facial landmark detection. As a result, our method achieves robust facial landmark detection, especially for the faces with heavy occlusion and profile view, and signifcantly outperforms the state-of-the-art methods with much less computational complexity. Ablation studies verify the effectiveness of the proposed method, and the visualized results show that the learned case-dependent inherent relation is quite coherent and close to human perception.

By using the negative log-likilyhood function to constrain the learning, the model learns to predict the uncertainty of each landmark in an unsupervised manner. We further incorporate the predicted uncertainty with SLPT to develop a novel model, dynamic sparse local patch transformer (DSLPT), which can dynamically adjust the receptive field. It successfully addresses the limitation of existing patch based methods: setting the variance of the probability distribution to a constant number for all landmarks. As a result, the robustness and reliability for heavily occluded landmarks are further enhanced, achieving impressive performance on existing benchmarks. The visualized results demonstrate that DSLPT can apply a larger patch to the landmark with high uncertainty for more contextual information and a smaller patch to the landmark with low uncertainty for higher resolution features.

We successfully improve the robustness and reliability of facial landmark detection for landmarks unseen during training by introducing a task-agnostic, unified face alignment (TUFA) framework. It notably enhances the performance of both many-shot and few-shot face alignment, while also implementing zero-shot facial landmark detection for the first time. Compared to existing facial landmark detection methods, TUFA represents a significant breakthrough in three critical aspects: 1) TUFA successfully models a mapping between an interpretable plane and target faces, and further bridges the gap between the seen and unseen facial landmarks using the proposed face structure prompts. This makes TUFA the first framework capable of tackling the challenge of zero-shot face alignment. 2) TUFA successfully unifies the learning targets of various facial landmark datasets and mitigates the noise introduced by different annotation methods through the utilization of the proposed semantic alignment embeddings. Thus, we provide a pre-trained model with very robust generalization ability based on the multi-dataset learning. 3) the unified learning target also enables the learned knowledge to be easily transferred to a new set of defined landmarks. Consequently, TUFA significantly boosts the performance of few-shot face alignment. Overall, we believe the unique properties of TUFA can propel facial landmark detection to a new stage and provide a competitive baseline for future work.

7.2 Future Work

Despite the significant improvement on robustness and reliability of facial landmark detection, there remain many possible directions. We will discuss some important possible future works below:

1. Compared to 2D facial landmarks, 3D facial landmarks contain more abundant information, such as depth and orientation, which significantly widen the application range of many downstream tasks. Although 2D facial landmark can be well localized under all conditions, the performance of 3D facial landmark detection still needs to be improved. Therefore, developing a robust method for 3D facial landmark detection is an important potential research direction.

2. Compared to 2D facial landmarks, labeling 3D facial landmarks is much more expensive and time-consuming. Additionally, considering the privacy issues, it is very hard to establish a large-scale 3D facial landmark dataset. Therefore, an intuitive method is to establish a synthetic 3D facial landmark dataset using rendering engine. Therefore, developing a domain adaptation method is crucial to bridge the gap between these synthetic images and real images.

3. Unsupervised learning is a crucial method for eliminating the reliance on annotation. Although unsupervised 2D facial landmark detection has been investigated for more than 10 years, there is no existing work on unsupervised 3D facial landmark detection. Devel-

oping a robust unsupervised 3D facial landmark detection method can maximize the use of raw images and provide a strong pretrained model.

4. Compared to other tasks, the number of training samples for facial landmark detection is very insufficient. Using semi-supervised learning to train a large-scale model with a very large dataset is also crucial for the performance of downstream tasks. Unfortunately, there is no related work on this topic yet.

Bibliography

[1] B. Liang, Y. Pan, Z. Guo, *et al.*, "Expressive talking head generation with granular audio-visual control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3377–3386. DOI: 10.1109/CVPR52688.2022.00338.

[2] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5073–5082. DOI: 10.1109/CVPR42600.2020.00512.

[3] N. Otberdout, M. Daoudi, A. Kacem, L. Ballihi, and S. Berretti, "Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 848–863, 2022. DOI: 10.1109/TPAMI.2020.3002500.

[4] A. B. Tanfous, H. Drira, and B. B. Amor, "Sparse coding of shape trajectories for facial expression and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2594–2607, 2020. DOI: 10.1109/TPAMI.2019.2932979.

[5] X. Xu, Q. Meng, Y. Qin, *et al.*, "Searching for alignment in face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 3065–3073.

[6] B. Bakker, B. Zabłocki, A. Baker, *et al.*, "A multi-stage, multi-feature machine learning approach to detect driver sleepiness in naturalistic road driving conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4791–4800, 2022. DOI: 10.1109/TITS.2021.3090272.

[7] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2887–2894. DOI: 10.1109/CVPR.2012.6248015.

[8] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, Dec. 2013, pp. 1513–1520. DOI: 10.1109/ICCV.2013.191.

[9] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 532–539. DOI: 10.1109/CVPR.2013.75.

[10] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874. DOI: 10.1109/CVPR.2014.241.

[11] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859–1866. DOI: 10.1109/CVPR.2014.240.

[12] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4998–5006. DOI: 10.1109/CVPR.2015.7299134.

[13] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1862–1874, 2015. DOI: 10.1109/TPAMI.2014.2382106.

[14] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3659–3667. DOI: 10.1109/CVPR.2015.7298989.

[15] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu, "Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp. 3681–3690. DOI: 10.1109/CVPR.2017.392.

[16] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1233–1245, 2016. DOI: 10.1109/TIP.2016.2518867.

[17] A. Zadeh, T. Baltrušaitis, and L.-P. Morency, "Convolutional experts constrained local model for facial landmark detection," in *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition Workshops, 2017, pp. 2051–2059. DOI: 10.1109/CVPRW.2017.256.

[18] H. Liu, J. Lu, M. Guo, S. Wu, and J. Zhou, "Learning reasoning-decision networks for robust face alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 679–693, 2020. DOI: 10.1109/TPAMI.2018.2885298.

[19] C. Zhu, X. Wan, S. Xie, X. Li, and Y. Gu, "Occlusion-robust face alignment using a viewpoint-invariant hierarchical network architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 102–11 111. DOI: 10.1109/CVPR52688.2022.01083.

[20] B. Browatzki and C. Wallraven, "3fabrec: Fast few-shot face alignment by reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp. 6109–6119. DOI: 10.1109/CVPR42600.2020.00615.

[21] X. He, G. Bharaj, D. Ferman, H. Rhodin, and P. Garrido, "Few-shot geometry-aware keypoint localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2023, pp. 21 337–21 348.

[22] J. Thewlis, H. Bilen, and A. Vedaldi, "Unsupervised learning of object landmarks by factorized spatial embeddings," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3229–3238. DOI: 10.1109/ICCV.2017.348.

[23] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, "Unsupervised discovery of object landmarks as structural representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2694–2703. DOI: 10.1109/CVPR.2018.00285.

[24] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer, "Unsupervised part-based disentangling of object shape and appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 947–10 956. DOI: 10.1109/CVPR.2019.01121.

[25] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, "Unsupervised learning of object landmarks through conditional image generation," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 4020–4031.

[26] X. He, B. Wandt, and H. Rhodin, "Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 36 123–36 141.

[27] D. Mallis, E. Sanchez, M. Bell, and G. Tzimiropoulos, “From keypoints to object landmarks via self-training correspondence: A novel approach to unsupervised landmark discovery,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8390–8404, 2023. DOI: 10.1109/TPAMI.2023.3234212.

[28] F. Pourpanah, M. Abdar, Y. Luo, *et al.*, “A review of generalized zero-shot learning methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4051–4070, 2023. DOI: 10.1109/TPAMI.2022.3191696.

[29] P. Huang, J. Han, D. Cheng, and D. Zhang, “Robust region feature synthesizer for zero-shot object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7612–7621. DOI: 10.1109/CVPR52688.2022.00747.

[30] T. Cootes, C. Taylor, D. Cooper, and J. Graham, “Active shape models-their training and application,” *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995, ISSN: 1077-3142. DOI: https://doi.org/10.1006/cviu.1995.1004.

[31] T. Cootes, G. Edwards, and C. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001. DOI: 10.1109/34.927467.

[32] X. Liu, “Discriminative face alignment,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1941–1954, 2009. DOI: 10.1109/TPAMI.2008.238.

[33] D. Cristinacce and T. Cootes, “Feature detection and tracking with constrained local models,” in *British Machine Vision Conference*, 2006, pp. 95.1–95.10. DOI: 10.5244/C.20.95.

[34] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

[35] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004, ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000029664.99615.94.

[36] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, “Mnemonic descent method: A recurrent process applied for end-to-end face alignment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4177–4187. DOI: 10.1109/CVPR.2016.453.

[37] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Proceedings of the European Conference on Computer Vision*, Cham, 2016, pp. 57–72.

[38] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, "A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment," in *Proceedings of the European Conference on Computer Vision*, Cham, 2018, pp. 609–624.

[39] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3691–3700. DOI: 10.1109/CVPR.2017.393.

[40] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 146–155. DOI: 10.1109/CVPR.2016.23.

[41] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IIEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, 2019. DOI: 10.1109/TPAMI.2017.2778152.

[42] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 2034–2043. DOI: 10.1109/CVPRW.2017.254.

[43] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2129–2138. DOI: 10.1109/CVPR.2018.00227.

[44] Z. Feng, J. Kittler, M. Awais, P. Huber, and X. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2235–2245. DOI: 10.1109/CVPR.2018.00238.

[45] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 3067–3074, 2018. DOI: 10.1109/TPAMI.2017.2787130.

[46] H. Liu, J. Lu, M. Guo, S. Wu, and J. Zhou, "Learning reasoning-decision networks for robust face alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 679–693, 2020. DOI: `10.1109/TPAMI.2018.2885298`.

[47] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 379–388. DOI: `10.1109/CVPR.2018.00047`.

[48] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3481–3491. DOI: `10.1109/CVPR.2019.00360`.

[49] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 152–168.

[50] W. Li, Y. Lu, K. Zheng, *et al.*, "Structured landmark detection via topology-adapting deep graph learning," in *Proceedings of the European Conference on Computer Vision*, Cham: Springer International Publishing, 2020, pp. 266–283, ISBN: 978-3-030-58545-7.

[51] C.-Y. Wu, Q. Xu, and U. Neumann, "Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry," in *Proceedings of International Conference on 3D Vision*, 2021, pp. 453–463. DOI: `10.1109/3DV53792.2021.00055`.

[52] C. Lin, B. Zhu, Q. Wang, *et al.*, "Structure-coherent deep feature learning for robust face alignment," *IEEE Transactions on Image Processing*, vol. 30, pp. 5313–5326, 2021. DOI: `10.1109/TIP.2021.3082319`.

[53] X. Shao, J. Xing, J. Lyu, X. Zhou, Y. Shi, and S. Maybank, "Robust face alignment via deep progressive reinitialization and adaptive error-driven learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5488–5502, 2022. DOI: `10.1109/TPAMI.2021.3073593`.

[54] A. Prados-Torreblanca, J. M. Buenaposada, and L. Baumela, "Shape preserving facial landmarks with graph attention networks," in *British Machine Vision Conference*, 2022.

[55] J. Wang, K. Sun, T. Cheng, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021. DOI: `10.1109/TPAMI.2020.2983686`.

[56] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 483–499, ISBN: 978-3-319-46484-8.

[57] Z. Tang, X. Peng, K. Li, and D. N. Metaxas, "Towards efficient u-nets: A coupled and quantized approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2038–2050, 2020. DOI: 10.1109/TPAMI.2019.2907634.

[58] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1021–1030. DOI: 10.1109/ICCV.2017.116.

[59] X. Dong, Y. Yang, S.-E. Wei, X. Weng, Y. Sheikh, and S.-I. Yu, "Supervision by registration and triangulation for landmark detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3681–3694, 2021. DOI: 10.1109/TPAMI.2020.2983935.

[60] A. Dapogny, M. Cord, and K. Bailly, "Decafa: Deep convolutional cascade for face alignment in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6892–6900. DOI: 10.1109/ICCV.2019.00699.

[61] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6970–6980. DOI: 10.1109/ICCV.2019.00707.

[62] J. Wan, Z. Lai, J. Liu, J. Zhou, and C. Gao, "Robust face alignment by multi-order high-precision hourglass network," *IEEE Transactions on Image Processing*, vol. 30, pp. 121–133, 2021. DOI: 10.1109/TIP.2020.3032029.

[63] L. Chen, H. Su, and Q. Ji, "Face alignment with kernel density deep neural network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6991–7001. DOI: 10.1109/ICCV.2019.00709.

[64] X. Zou, S. Zhong, L. Yan, X. Zhao, J. Zhou, and Y. Wu, "Learning robust facial landmark detection via hierarchical structured ensemble," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 141–150. DOI: 10.1109/ICCV.2019.00023.

[65] Y. Tai, Y. Liang, X. Liu, *et al.*, "Towards highly accurate and stable face alignment for high-resolution videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8893–8900.

[66] J. Zhang, H. Hu, and S. Feng, "Robust facial landmark detection via heatmap-offset regression," *IEEE Transactions on Image Processing*, vol. 29, pp. 5050–5064, 2020. DOI: 10.1109/TIP.2020.2976765.

[67] A. Kumar, T. K. Marks, W. Mou, *et al.*, "Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8233–8243. DOI: 10.1109/CVPR42600.2020.00826.

[68] C. Zhu, X. Li, J. Li, and S. Dai, "Improving robustness of facial landmark detection by defending against adversarial attacks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 11 731–11 740. DOI: 10.1109/ICCV48922.2021.01154.

[69] X. Lan, Q. Hu, and J. Cheng, "Revisting quantization error in face alignment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2021, pp. 1521–1530. DOI: 10.1109/ICCVW54120.2021.00177.

[70] Y. Huang, H. Yang, C. Li, J. Kim, and F. Wei, "Adnet: Leveraging error-bias towards normal direction in face alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 3060–3070. DOI: 10.1109/ICCV48922.2021.00307.

[71] X. Lan, Q. Hu, and J. Cheng, "Atf: An alternating training framework for weakly supervised face alignment," *IEEE Transactions on Multimedia*, vol. 25, pp. 1798–1809, 2023. DOI: 10.1109/TMM.2022.3164798.

[72] Z. Zhou, H. Li, H. Liu, N. Wang, G. Yu, and R. Ji, "Star loss: Reducing semantic ambiguity in facial landmark detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 475–15 484. DOI: 10.1109/CVPR52729.2023.01485.

[73] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceeding in International Conference on Learning Representations*, 2021.

[74] W. Wang, E. Xie, X. Li, *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2021, pp. 568–578.

[75] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2021, pp. 10 012–10 022.

[76] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision*, Springer, 2020, pp. 213–229.

[77] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable {detr}: Deformable transformers for end-to-end object detection," in *Proceeding in International Conference on Learning Representations*, 2021.

[78] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 12 077–12 090.

[79] Y. Li, S. Zhang, Z. Wang, *et al.*, "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 11 293–11 302. DOI: `10.1109/ICCV48922.2021.01112`.

[80] Y. YUAN, R. Fu, L. Huang, *et al.*, "Hrformer: High-resolution vision transformer for dense predict," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, 2021, pp. 7281–7293.

[81] C. Yu, B. Xiao, C. Gao, *et al.*, "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 10 435–10 445. DOI: `10.1109/CVPR46437.2021.01030`.

[82] Q. Li, Z. Zhang, F. Xiao, F. Zhang, and B. Bhanu, "Dite-hrnet: Dynamic lightweight high-resolution network for human pose estimation," in *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*, Jul. 2022, pp. 1095–1101. DOI: `10.24963/ijcai.2022/153`.

[83] T. Wen, Z. Ding, Y. Yao, Y. Wang, and X. Qian, "Picassonet: Searching adaptive architecture for efficient facial landmark localization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10 516–10 527, 2023. DOI: `10.1109/TNNLS.2022.3167743`.

[84] P. Micaelli, A. Vahdat, H. Yin, J. Kautz, and P. Molchanov, "Recurrence without recurrence: Stable video landmark detection with deep equilibrium models," in *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2023, pp. 22 814–22 825.

[85] G. Shapira, N. Levy, I. Goldin, and R. J. Jevnisek, “Knowing when to quit: Selective cascaded regression with patch attention for real-time face alignment,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2372–2380.

[86] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. DOI: 10.1109/LSP.2016.2603342.

[87] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[88] S. Qian, K. Sun, W. Wu, C. Qian, and J. Jia, “Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 152–10 162. DOI: 10.1109/ICCV.2019.01025.

[89] Y. Wang, L. Zhang, Y. Yao, and Y. Fu, “How to trust unlabeled data? instance credibility inference for few-shot learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6240–6253, 2022. DOI: 10.1109/TPAMI.2021.3086140.

[90] G. Zhang, Z. Luo, K. Cui, S. Lu, and E. P. Xing, “Meta-detr: Image-level few-shot detection with inter-class correlation exploitation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 832–12 843, 2023. DOI: 10.1109/TPAMI.2022.3195735.

[91] G. Cheng, C. Lang, and J. Han, “Holistic prototype activation for few-shot segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4650–4666, 2023. DOI: 10.1109/TPAMI.2022.3193587.

[92] J. Thewlis, H. Bilen, and A. Vedaldi, “Unsupervised learning of object landmarks by factorized spatial embeddings,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3229–3238. DOI: 10.1109/ICCV.2017.348.

[93] Y. Jin, W. Sun, J. Hosang, E. Trulls, and K. M. Yi, “Tusk: Task-agnostic unsupervised keypoints,” in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, 2022, pp. 29 538–29 551.

[94] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, “Self-supervised learning of interpretable keypoints from unlabelled videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8784–8794. DOI: 10.1109/CVPR42600.2020.00881.

[95] Y. Kim, S. Nam, I. Cho, and S. J. Kim, “Unsupervised keypoint learning for guiding class-conditional video prediction,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[96] M. Minderer, C. Sun, R. Villegas, F. Cole, K. P. Murphy, and H. Lee, “Unsupervised learning of object structure and dynamics from videos,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[97] S. Tourani, A. Alwheibi, A. Mahmood, and M. H. Khan, “Pose-guided self-training with two-stage clustering for unsupervised landmark discovery,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 041–23 051. DOI: 10.1109/CVPR52733.2024.02174.

[98] P. Gleize, W. Wang, and M. Feiszli, “Silk: Simple learned keypoints,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 22 442–22 451. DOI: 10.1109/ICCV51070.2023.02056.

[99] A. Gupta, S. Narayan, S. Khan, F. S. Khan, L. Shao, and J. van de Weijer, “Generative multi-label zero-shot learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 611–14 624, 2023. DOI: 10.1109/TPAMI.2023.3295772.

[100] P. Huang, J. Han, D. Cheng, and D. Zhang, “Robust region feature synthesizer for zero-shot object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7612–7621. DOI: 10.1109/CVPR52688.2022.00747.

[101] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, “Zegclip: Towards adapting clip for zero-shot semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 175–11 185. DOI: 10.1109/CVPR52729.2023.01075.

[102] C.-C. Lin, K. Lin, L. Wang, Z. Liu, and L. Li, “Crossmodal representation learning for zero-shot action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 946–19 956. DOI: 10.1109/CVPR52688.2022.01935.

[103] H. Zhang, L. Xu, S. Lai, *et al.*, "Open-vocabulary animal keypoint detection with semantic-feature matching," *International Journal of Computer Vision*, pp. 1–18, 2024. DOI: 10.1007/s11263-024-02126-3.

[104] A. Howard, M. Sandler, B. Chen, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324. DOI: 10.1109/ICCV.2019.00140.

[105] X. Huang, W. Deng, H. Shen, X. Zhang, and J. Ye, "Propagationnet: Propagate points to curve to learn structure information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7263–7272. DOI: 10.1109/CVPR42600.2020.00729.

[106] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 936–944. DOI: 10.1109/CVPR.2017.106.

[107] A. Bulat and G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 3726–3734. DOI: 10.1109/ICCV.2017.400.

[108] R. Liu, J. Lehman, P. Molino, *et al.*, "An intriguing failing of convolutional neural networks and the coordconv solution," in *Advances in Neural Information Processing Systems*, 2018, pp. 9628–9639.

[109] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403. DOI: 10.1109/ICCVW.2013.59.

[110] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5525–5533. DOI: 10.1109/CVPR.2016.596.

[111] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016, ISSN: 0262-8856.

[112] K. Diederik and B. Jimmy, "Adam: A method for stochastic optimization," in *Proceeding in International Conference on Learning Representations*, 2015.

[113] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[114] H. Liu, J. Lu, J. Feng, and J. Zhou, “Two-stream transformer networks for video-based face alignment,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2546–2554, 2018. DOI: 10.1109/TPAMI.2017.2734779.

[115] P. Chandran, D. Bradley, M. Gross, and T. Beeler, “Attention-driven cropping for very high resolution facial landmark detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5860–5869. DOI: 10.1109/CVPR42600.2020.00590.

[116] K. Sun, W. Wu, T. Liu, *et al.*, “Fab: A robust facial landmark detection framework for motion-blurred videos,” in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5461–5470. DOI: 10.1109/ICCV.2019.00556.

[117] X. Zou, P. Xiao, J. Wang, L. Yan, S. Zhong, and Y. Wu, “Towards unconstrained facial landmark detection robust to diverse cropping manners,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 2070–2075, 2021. DOI: 10.1109/TCSVT.2020.3006236.

[118] R. Ranjan, V. M. Patel, and R. Chellappa, “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*, vol. 41, no. 1, pp. 121–135, 2019. DOI: 10.1109/TPAMI.2017.2781233.

[119] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 59–66. DOI: 10.1109/FG.2018.00019.

[120] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, “Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1087–1096. DOI: 10.1109/CVPR.2019.00118.

[121] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[122] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2879–2886. DOI: 10.1109/CVPR.2012.6248014.

[123] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 679–692, ISBN: 978-3-642-33712-3.

[124] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2011, pp. 545–552. DOI: 10.1109/CVPR.2011.5995602.

[125] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1899–1906. DOI: 10.1109/CVPR.2014.306.

[126] J. Deng, A. Roussos, G. Chrysos, *et al.*, "The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking," *International journal of computer vision*, vol. 127, pp. 599–624, 2019. DOI: 10.1007/s11263-018-1134-y.

[127] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 2116–2125. DOI: 10.1109/CVPRW.2017.263.

[128] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3409–3417. DOI: 10.1109/CVPR.2016.371.

[129] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2011, pp. 2144–2151. DOI: 10.1109/ICCVW.2011.6130513.

[130] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, 2019.

[131] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[132] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceeding in International Conference on Learning Representations*, 2019.

[133] J. Xia, W. Qu, W. Huang, J. Zhang, X. Wang, and M. Xu, "Sparse local patch transformer for robust face alignment and landmarks inherent relation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 4052–4061.

[134] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 896–903. DOI: 10.1109/CVPRW.2013.132.

[135] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738. DOI: 10.1109/ICCV.2015.425.

[136] J. Xia, M. Xu, H. Zhang, *et al.*, "Robust face alignment via inherent relation learning and uncertainty estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 358–10 375, 2023. DOI: 10.1109/TPAMI.2023.3260926.

[137] Y. Zheng, H. Yang, T. Zhang, *et al.*, "General facial representation learning in a visual-linguistic manner," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp. 18 697–18 709.

[138] M. Caron, H. Touvron, I. Misra, *et al.*, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 9630–9640. DOI: 10.1109/ICCV48922.2021.00951.

[139] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 2025–2033. DOI: 10.1109/CVPRW.2017.253.

[140] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 472–487.

[141] W. Wu and S. Yang, “Leveraging intra and inter-dataset variations for robust face alignment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 2096–2105. DOI: `10.1109/CVPRW.2017.261`.