Accurate Facial Landmark Detector via Multi-scale Transformer

Yuyang Sha, Weiyu Meng, Xiaobing Zhai, Can Xie, and Kefeng Li(✉)

Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, China
kefengl@mpu.edu.mo

Abstract. Facial landmark detection is an essential prerequisite for many face applications, which has attracted much attention and made remarkable progress in recent years. However, some problems still need to be solved urgently, including improving the accuracy of facial landmark detectors in complex scenes, encoding long-range relationships between keypoints and facial components, and optimizing the robustness of methods in unconstrained environments. To address these problems, we propose a novel facial landmark detector via multi-scale transformer (MTLD), which contains three modules: Multi-scale Transformer, Joint Regression, and Structure Loss. The proposed Multi-scale Transformer focuses on capturing long-range information and cross-scale representations from multi-scale feature maps. The Joint Regression takes advantage of both coordinate and heatmap regression, which could boost the inference speed without sacrificing model accuracy. Furthermore, in order to explore the structural dependency between facial landmarks, we design the Structure Loss to fully utilize the geometric information in face images. We evaluate the proposed method through extensive experiments on four benchmark datasets. The results demonstrate that our method outperforms state-of-the-art approaches both in accuracy and efficiency.

Keywords: Facial landmark detection · Vision transformer · Multi-scale feature · Global information

1 Introduction

Facial landmark detection aims to find some pre-defined locations on human face images, which usually have specific semantic meanings, such as the eyebrow or pupil. It has become one of the most fundamental tasks in computer vision and is used for many real-world applications. Thanks to the development of deep learning and computer vision techniques, facial landmark detection algorithms have achieved significant progress in accuracy and efficiency over the past decades.

Since 2012, methods based on deep neural networks have been the dominant solution for many fields in computer vision. Similarly, facial landmark detectors based on deep learning show significant advantages over traditional methods in

Q. Liu et al. (Eds.): PRCV 2023, LNCS 14429, pp. 278–290, 2024.
https://doi.org/10.1007/978-981-99-8469-5_22

terms of accuracy, generalization, and robustness. Recently, several facial landmark detection algorithms [1–3] with excellent performance have been proposed. For instance, Feng *et al.* [2] proposed the Wing-Loss to increase the contribution of the samples with small and medium size errors to the training of the regression framework. The designed Wing-Loss enables coordinate-based methods to achieve promising performance under wild environments. Xia *et al.* [3] leveraged coordinate regression and Transformer to explore the inherent relationships between facial keypoints and achieve impressive results.

In order to achieve excellent performance, existing mainstream methods attempt to utilize a more complex backbone for learning discriminative representations, such as ResNet [4], HRNet [5], *etc.* Other approaches involve complex data augmentation technologies [6], while some methods [1,3] focus on optimizing the regression schemes with the carefully designed detection head or vision transformer. Although these approaches perform well on public benchmark datasets, they are still hard to apply in unconstrained environments and complex scenes. One issue is that most works take deep convolution networks (CNN) as the backbone to extract features for input samples, which may pay more attention to local information but ignore some meaningful global representations and long-range relationships. Additionally, these frameworks often overlook essential prior knowledge of human face images, such as structural information and geometric relationships of different facial components. That may limit the model's performance, especially on occluded and blurred face samples. Moreover, the commonly used approaches struggle to balance accuracy and inference speed.

To address the above issues, we present a novel facial landmark detector via Multi-scale Transformer named MTLD. The proposed method mainly consists of three modules: Multi-scale Transformer, Joint Regression, and Structure Loss. In order to optimize the disadvantages of the facial landmark detector based on CNN, we proposed the Multi-scale Transformer for face alignment by making full use of multi-scale feature maps to capture the global representations and explore long-range relationships between different facial keypoints. The Joint Regression can be regarded as coordination regression, which would generate a group of heatmaps with the output multi-scale feature maps of backbone during the training stage, then apply them as an auxiliary heatmap loss to accelerate convergence. Notable, the heatmap loss is only used in the training stage, so it would not affect the model inference speed. The proposed Joint Regression takes full advantage of both heatmap and coordinate regression, which can improve the accuracy of facial landmark detectors without scarifying the inference speed. Prior knowledge of the human face's structural information can improve the accuracy of facial landmark detection models. However, current methods do not make full use of this information. Therefore, we designed a loss function called Structure Loss to constrain the specific information between facial keypoints. This loss function aims to improve the continuity and consistency of predicted localization, especially in occluded and blurred environments. In summary, the primary contributions of this paper are as follows:

- We propose a Multi-scale Transformer for facial landmark detection to enhance model performance by processing multi-scale feature maps, which can capture global information and long-range relationships between different facial keypoints.
- We introduce the Joint Regression, which applies the auxiliary heatmap loss to accelerate convergence and forces the model to learn more discriminate representations. Additionally, we design the Structure Loss to constrain the structural correlations between keypoints, thus significantly improving the model performance under occlusion, blur, large pose, *etc.*
- We conduct extensive experiments to verify the model effectiveness in four benchmark datasets, including 300W, WFLW, COFW, and AFLW. The results demonstrate that our method obtains competitive results and fast inference speed compared to state-of-the-art methods.

2 Related Work

2.1 Facial Landmark Detection

Facial landmark detection is a crucial technique in numerous applications involving face recognition and emotion estimation. Therefore, optimizing the performance of facial landmark detectors can make excellent benefits for these related tasks. Currently, CNN-based facial landmark detectors in this field primarily fall into coordinate and heatmap regression. Coordinate-based methods directly map the input face samples into 2D coordinates, which usually enjoy a faster inference speed. However, the accuracy of coordinate-based methods still needs to be improved. Therefore, Feng *et al.* [2] introduced a new loss function, termed Wing-Loss, which improved the accuracy of the coordinate regression method, especially for face samples under the occlusion and blur situations. Heatmap-based methods utilize CNNs to encode face images into a group of heatmap representations, each indicating the probability of a landmark localization. Now, most high-performance facial landmark detectors are based on heatmap regression. For instance, HRNet [5] maintained multi-resolution representations in parallel and exchanged information between these streams to obtain high-accuracy prediction results.

2.2 Vision Transformer

Transformer is a deep-learning model originally designed for machine translation. Now, transformer-based models have been shown to significantly enhance the performance of many natural language processing tasks. Inspired by the success of sequence-to-sequence tasks, there is growing interest in exploring the use of Transformer models for various computer vision tasks. For example, DETR [7] proposed a novel object detection system by combining CNN and Transformer, which predicts bounding boxes via bipartite matching. ViT [8] directly extracted representations from flatted image patches with a pure transformer encoder for image classification. In this study, we utilize the transformer to take full advantage of multi-scale feature maps, which can help the model to establish the global information and long-range relationships between different facial keypoints, thus improving the model performance.

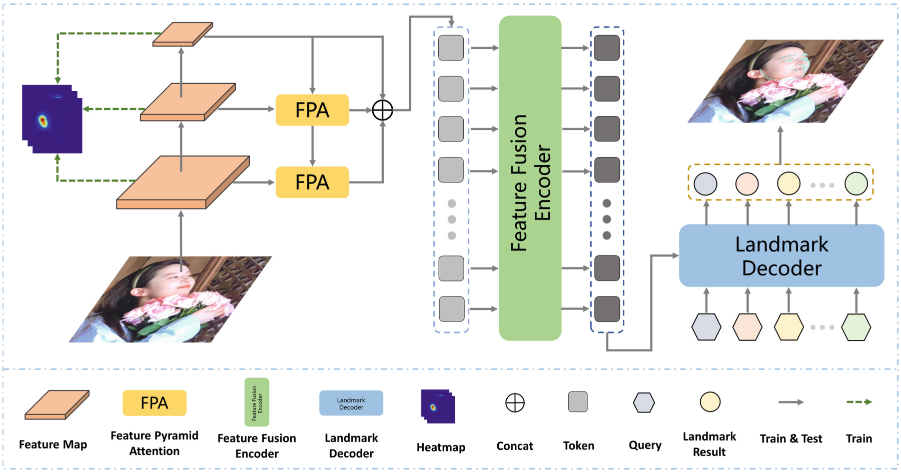

Fig. 1. Overview of the proposed MTLD. Firstly, given a human face image, our framework extracts multi-scale feature maps with the CNN-based backbone. Then, these multi-scale feature maps are processed by two parts simultaneously: one generates heatmaps from the multi-scale feature maps, while the other maps these representations into 2D coordinates. The generated heatmap serves as the auxiliary loss for model training. The proposed multi-scale transformer can fully use cross-level feature maps to extract global information and long-range relationships. Specifically, the heatmap loss is employed in model training and discarded in inference.

3 Method

3.1 Overview

As shown in Fig. 1, we propose a facial landmark detector based on the deep learning module named MTLD. Our method consists of three parts: Multi-scale Transformer, Joint Regression, and Structure Loss. The Multi-scale Transformer can enhance model performance by processing multi-scale feature maps generated from a CNN-based backbone. Joint Regression provides a new scheme for facial landmark detection, which takes advantage of heatmap and coordinate regression. Meanwhile, Structural Loss can constrain the correlation among different facial keypoints, making the model pay more attention to geometric information.

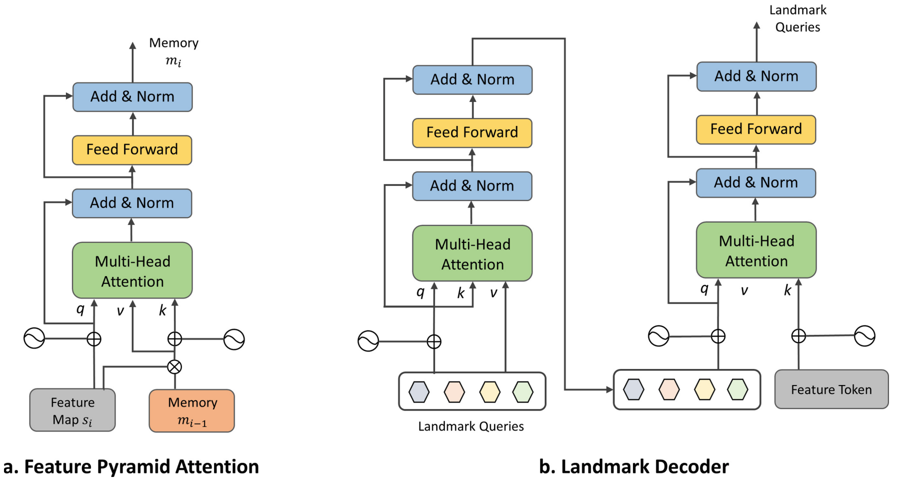

Fig. 2. (a) Detailed structure of the Feature Pyramid Attention. (b) Detailed structure of the Landmark Decoder.

3.2 Multi-scale Transformer

The architecture of our MTLD builds upon the CNN and vision transformer, which is used for exploiting more discriminative representations for facial landmark detection. The introduced network architecture consists of four parts: CNN-based backbone, Feature Pyramid Attention (FPA), Feature Fusion Encoder, and Landmark Decoder. The CNN-based backbone encodes the multi-scale feature maps from input samples, while the FPA obtains cross-scale representations from them. The Feature Fusion Encoder tries to merge and encode the output features from Feature Pyramid Attention to get feature tokens. Then, the Landmark Decoder utilizes landmark queries and feature tokens to predict the coordinates of each facial landmark.

CNN-Based Backbone. In the proposed method, we select the output of the last three stages in the CNN model as multi-scale feature maps. Generally, these feature map contains a large amount of multi-scale semantic and spatial information. Classical CNN models can be directly used in the proposed framework without modification, such as VGG, ResNet, and MobileNet. For instance, we use ResNet-18 as the backbone to illustrate some details. Specifically, given an RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ as input, we can get three feature maps s_3, s_4, and s_5, with the stride of 8, 16, and 32, respectively. Then, these multi-scale feature maps are fed into FPA to get refinement cross-scale representations.

Feature Pyramid Attention (FPA). Encoding multi-scale feature maps can efficiently improve model performance in complex computer vision tasks, such as object detection, instance segmentation, and image generation. Inspired by [9], we introduce a Feature Pyramid Attention formed by the vision transformer to enhance the power of model to utilize multi-level features and capture the long-range relationships between keypoints and facial components. The proposed FPA is shown in Fig. 2-(a). FPA uses the multi-scale feature maps $s_i \in \{s_3, s_4, s_5\}$ generated by CNN-based backbone to output multi-scale feature memory M. In order to make full use of these multi-scale representations, the i^{th} FPA would use the $(i\text{-}1)^{th}$ FPA's output as input.

The input of FPA includes three parties: queries q, keys k, and values v. The q should maintain the relative position by positional embedding p, which is defined as $q_i = s_i + p_i$. We can get the v by Hadamard product with the prior FPA's output m_{i-1}, which can be expressed as $v_i = s_i \circ m_{i-1}$. The k still needs the positional embedding to align semantic meaning as $k_i = v_i + p_i$. Finally, the output of i^{th} FPA can be calculated as:

$$m_i = \mathrm{LN}(t_i + \mathrm{FFN}(t_i)), \tag{1}$$

where, the LN denotes the layer normalization, and FFN is a feed forward network. The t_i is represented by: $t_i = \mathrm{LN}(s_i + \mathrm{MHA}(q_i, k_i, v_i))$, the MHA infers the multi-head attention.

Feature Fusion Encoder. In our method, we employ FPA to encode the global information and long-range relationships from multi-scale feature maps. However, the FPA might pay much attention to high-level representations but ignore some details information contained in these multi-scale feature maps. Therefore, we introduce the Feature Fusion Encoder consisting of multiple transformer encoders to combine and improve these representations from the CNN-based backbone and FPA. The Feature Fusion Encoder can map the input representations into cross-region feature tokens involving rich local and global information.

Landmark Decoder. The proposed MTLD can directly output the coordinates of facial keypoints with Landmark Decoder. The detailed structure is shown in Fig. 2-(b). First, the Landmark Decoder would process the input landmark queries with a self-attention module to make them interact with each other. Then, each landmark query extracts discriminate representations from the input multi-scale feature tokens. Finally, we employ a group of MLPs as the detection head to predict the coordinate results for each facial landmark. In our setting, all the detection heads should output prediction results. Therefore, the first detection head output rough positions, then the subsequent ones can gradually refine the previous results in a coarse-to-fine manner.

3.3 Joint Regression

We propose a novel face alignment scheme named Joint Regression to address the challenging problem of balancing the model accuracy and inference speed. The proposed Joint Regression can be viewed as a combination of heatmap and coordinate regression, which takes full advantage of them. Our framework first employs the CNN-based backbone to encode the input face samples into a group of mulita-scale feature maps. Then, these feature maps would be processed in two parts simultaneously: one generates heatmaps from the multi-scale feature maps, while the other maps the representations into 2D coordinates.

We employ several convolutional layers at the top of the CNN-based backbone for converting the input multi-scale feature maps into a set of heatmap representations $\mathbf{F_H} \in \mathbb{R}^{H_h \times W_h \times N_h}$, where H_h and W_h represent the height and the width. The N_h denotes the number of facial keypoints. The extracted heatmap representation can be seen as the probability of landmark location. Inspired by heatmap regression, we employ the L_2 loss function to compare the ground-truth heatmap $\mathbf{L_H}$ and predict ones $\mathbf{F_H}$. The $\mathcal{L}_{heat}$ is defined as:

$$\mathcal{L}_{heat} = \|\mathbf{L_H} - \mathbf{F_H}\|_2^2. \tag{2}$$

Notable, the heatmap loss can only be used for auxiliary supervision during the training stage and removed when the model testing and deployment. Therefore, it would not affect the inference speed and model efficiency.

At the same time, the selected multi-scale feature maps generated by the CNN-based backbone are also fed to the designed Multi-scale Transformer, which can map them into 2D coordinates of facial landmarks. We adopt L_1 loss to minimize the error $\mathcal{L}_{coord}$ between the predicted results and 2D ground truths:

$$\mathcal{L}_{coord} = \|\mathbf{L_C} - \mathbf{J_C}\|_1, \tag{3}$$

where, the $\mathbf{L_C}$ and $\mathbf{J_C}$ denote the 2D annotations and predicted results, respectively.

3.4 Structure Loss

The human face contains large amounts of geometric information, which are beneficial to improve model performance, especially face under occlusion, blur, lighting, and extreme pose. In order to make full use of these dependencies, we propose Structure Loss to exploit the structural information among facial landmarks effectively. Specifically, the location of facial landmarks is relatively fixed, such as the pupil locates in the center of the iris. Therefore, the structural information can be used to infer the location of adjacent facial keypoints and prevent some abnormal prediction results. The proposed Structure Loss aims to ensure that the distances between predicted keypoints are the same as those calculated from the ground truth. We formulate the Structure Loss as follows:

$$\mathcal{L}_{struc} = \sum_{i \in N} \sum_{j \in C} \| \|\mathbf{J}_i - \mathbf{J}_j\|_2^2 - \|\mathbf{L}_i - \mathbf{L}_j\|_2^2 \|_1, \tag{4}$$

where $\mathbf{J}$ and $\mathbf{L}$ denote the prediction result and ground-truth labels, respectively. The N indicates the total number of facial landmarks, and i represents the i^{th} landmark. C is a collection, which contains M closet landmarks to the i^{th} one, and j denotes the j^{th} landmarks in C. In our setting, the number of adjacent landmarks $Num.$ is 5. Structure Loss is beneficial to enhance the stability and robustness of facial landmark detection approaches.

3.5 Training Objective

We formulate the goals of MTLD and get the overall training objective, which is computed by:

$$\mathcal{L}_{total} = \lambda_h \mathcal{L}_{heat} + \lambda_c \mathcal{L}_{coord} + \lambda_s \mathcal{L}_{struc}, \tag{5}$$

where λ_h, λ_c, and λ_s denote the balancing parameters used to reweight these above loss functions.

Table 1. Facial landmark detection results about NME (%) on 300W, AFLW, and COFW. Lower is better. **Red** denotes the best, and **blue** indicates the second best.

Method	Backbone	300W			AFLW		COFW
		Full	Comm.	Chal.	Full	Fron.	
LAB [6]	ResNet-18	3.49	2.98	5.19	1.85	1.62	3.92
Wing [2]	ResNet-50	4.04	3.27	7.18	1.47	-	5.07
ODN [10]	ResNet-18	4.17	3.56	6.67	1.63	1.38	-
HRNet [5]	HRNet-W18	3.32	2.87	5.15	1.56	1.46	3.45
AWing [11]	Hourglass	**3.07**	**2.72**	**4.52**	-	-	-
PIPNet [1]	ResNet-101	3.19	2.78	4.89	1.42	-	3.08
SDFL [12]	ResNet-18	3.28	2.88	4.93	-	-	3.63
SLPT [3]	ResNet-34	3.20	2.78	4.93	-	-	4.11
MTLD	ResNet-18	3.28	2.81	4.96	1.42	1.31	3.25
MTLD	ResNet-50	3.20	2.75	4.94	**1.40**	**1.30**	**3.06**
MTLD	ResNet-101	**3.15**	**2.74**	**4.85**	**1.39**	**1.28**	**3.04**

4 Experiments

4.1 Implementation Details and Datasets

In the training phase, all input images need to be cropped by bounding boxes, then resized to the size of 256 × 256. The data augmentations are applied for model training, including random rotation, occlusion, scaling, horizontal flipping, and blurring. We adopt pre-trained ResNet-18 as the default CNN-based backbone. In order to get more accurate results, we also conduct experiments based on ResNet-50 and ResNet-101. The total epochs are 150, and the mini-batch size is 64. We choose Adam as the optimizer with an initial learning rate is 3.0×10^{-4}, and then decay by 10 at 70^{th} and 120^{th} separately. The implementation of our method is based on PyTorch with one NVIDIA Tesla A100 GPU.

Table 2. Facial landmark detection results about NME (%) on WFLW test and 6 subsets: pose, expression (expr.), illumination (illu.), make-up (mu.), occlusion (occu.) and blur. For the NME and FR, lower is better.

Method	backbone	Test	Pose	Expr.	Illu.	Mu.	Occl.	Blur
LAB [6]	ResNet-18	5.27	10.24	5.51	5.23	5.15	6.79	6.32
Wing [2]	ResNet-50	5.11	8.75	5.36	4.93	5.41	6.37	5.81
HRNet [5]	HRNet-W18	4.60	7.94	4.85	4.55	4.29	7.33	6.88
Awing [11]	Hourglass	4.36	7.38	4.58	4.32	4.27	**5.19**	4.96
PIPNet [1]	ResNet-101	4.31	7.51	**4.44**	4.19	**4.02**	5.36	5.02
SDFL [12]	ResNet-18	4.35	7.42	4.63	4.29	4.22	**5.19**	5.08
SLPT [3]	ResNet-34	**4.20**	**7.18**	4.52	**4.07**	4.17	**5.01**	**4.85**
MTLD	ResNet-18	4.47	7.80	4.54	4.40	4.31	5.52	5.23
MTLD	ResNet-50	4.39	7.70	4.41	4.22	4.15	5.43	5.12
MTLD	ResNet-101	**4.25**	**7.29**	**4.37**	**4.10**	**4.03**	5.31	**4.91**

In order to evaluate the performance of our proposed method, we conduct extensive experiments on four benchmark datasets: 300W [13], COFW [14], AFLW [15], and WFLW [6]. Most of the experiment setting about datasets follow [5]. We adopt the normalized mean error (NME) to evaluate the performance of our approach on the benchmark dataset. Specifically, the inter-ocular distance is used as the normalization distance for 300W, COFW, and WFLW, while using the face bounding box as the normalization distance in AFLW.

4.2 Main Results

We compare our proposed method with several state-of-the-art approaches on four benchmark datasets in terms of NME. To further explore the effectiveness of the backbone, we conduct experiments with different CNN modules, including ResNet-18, ResNet-50, and ResNet-101. Some visualization results of our proposed MTLD on 300W and WFLW are shown in Fig. 3.

300W. We compare the proposed MTLD with other state-of-the-art methods on 300W and its subsets. Table 1 shows that MTLD with ResNet-18 obtains comparable results with existing approaches. We can find that our method with ResNet-101 achieves the second best detection accuracy on 300W-Full, Common, and Challenging sets, which is only slightly behind AWing [11].

AFLW. The AFLW dataset is a challenging benchmark for evaluating facial landmark detectors. We compare the proposed MTLD with the existing methods, and the results are shown in Table 1. Obviously, our framework with ResNet-18 gets 1.42% NME, which is comparable with SOTA methods. Furthermore, MTLD with ResNet-50 or ResNet-101 outperforms all the existing methods on AFLW-Full and AFLW-Frontal datasets.

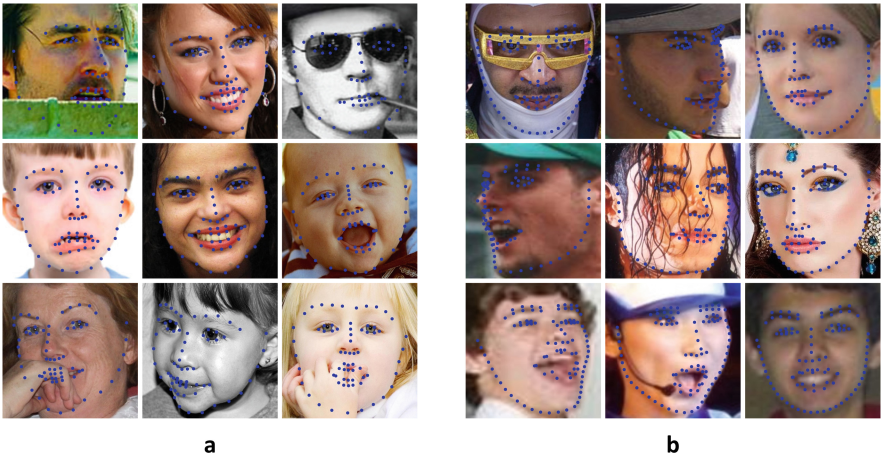

Fig. 3. Visualization results of our method. (a) Results on 300W. (b) Results on WFLW.

COFW. In Table 1, we report the comparison results with existing SOTA methods on the COFW dataset. The results indicate that MTLD with ResNet-18 obtains 3.25% NME, which is slightly higher than Wing-Loss [2] and HRNet [5]. Furthermore, the MTLD with ResNet-101 gets 3.04% NME and outperforms the previous methods by a significant margin.

WFLW. The WFLW dataset is more challenging than 300W and COFW, which provides many face images with various senses. We conduct experiments on WFLW and six subsets. Table 2 demonstrates NME results about the SOTA methods and MTLD with different backbones. We observe that MTLD with ResNet-18 achieves comparable performance to the SOTA methods equipped with more complex models, such as Hourglass and HRNet. Furthermore, MTLD with ResNet-101 achieves SOTA performance on one subset and second best results on four subsets.

4.3 Ablation Study

In this section, we conduct several experiments to verify the effectiveness of the proposed module. Then, we evaluate the model size, computational cost, and inference speed of MTLD. Besides, we also design experiments to explore the appropriate number of adjacent landmarks in Structure Loss.

Effectiveness of Proposed Modules. In this section, we investigate the effectiveness of these modules and conduct experiments on the 300W-Full dataset with NME. For easy comparison, we make the coordinate regression framework with ResNet-18 as the baseline model. Then, add different modules proposed in this paper and analyze their impact on the results. The results are shown in Table 3. We observe that each proposed module is beneficial to improve model

Table 3. The NME (%) of different modules on 300W-Full dataset, including: Baseline (Base.), Multi-scale Transformer (MST.), Heatmap Loss (Heat.), Structure Loss (Sturc.). Lower is better.

Base.	MST.	Heat.	Struc.	NME
✓				4.02
✓	✓			3.39
✓	✓	✓		3.34
✓	✓		✓	3.30
✓	✓	✓	✓	**3.25**

accuracy. The designed MTLD equipped with Multi-scale Transformer, Heatmap Loss, and Structure Loss can achieve 3.25% NME in the 300W-Full dataset. Specifically, the Multi-scale Transformer can significantly boost the model performance.

Table 4. The comparison of different approaches in backbone, model size (Param) computational cost (GFLOPs), and inference speed (fps) on CPU and GPU.

Method	Backbone	Param	GFLOPs	CPU	GPU
LAB [6]	ResNet-18	24.1M	26.7G	2.1	16.7
Wing [2]	ResNet-50	91.0M	5.5G	8.0	30.0
HRNet [5]	HRNet-W18	**9.7M**	4.8G	4.4	11.7
PIPNet [1]	ResNet-18	12.0M	**2.4G**	35.7	200
MTLD	ResNet-18	12.6M	2.7G	**45.8**	**213.5**
MTLD	ResNet-50	27.3M	5.8G	13.5	112.2
MTLD	ResNet-101	46.0M	10.7G	7.6	66.5

Model Size and Speed Analysis. To further evaluate the model's effectiveness, we compare the model size (Params), computational cost (FLOPs), and inference speed (FPS) of our MTLD with SOTA methods. Specifically, the input samples are resized to 256×256, and models are implemented with PyTorch. To compare the inference speed, we evaluate these frameworks on CPU (Intel i7-9700@3.00GHz) and GPU (Nvidia Tesla A100), respectively. Results are shown in Table 4, which indicates that our proposed method with ResNet-18 gets 45.8 FPS and 213.5 FPS on CPU and GPU, respectively. Compared with existing methods, MTLD obtains comparable performance while maintaining a fast inference speed.

Number of Adjacent Points. To explore the appropriate collection number of adjacent points in Structure Loss, we conduct experiments on the 300W-Full dataset in terms of NME. The results are shown in Table 5. It can be observed that when the **Num.** is set to **5**, our method can deliver the best performance.

Table 5. The NME (%) results of our method with different number of adjacent points on 300W-Full dataset. Lower is better.

Num	0	1	3	**5**	8	10	15	20	30
NME	3.29	3.28	3.26	**3.25**	3.27	3.27	3.30	3.36	3.45

5 Conclusion

In this paper, we propose a facial landmark detector named MTLD, which includes three modules: Multi-scale Transformer, Joint Regression, and Structure Loss. Specifically, the carefully designed Multi-scale Transformer enables the model to capture the global dependencies between keypoints and facial components from multi-scale feature maps. The Joint Regression takes advantage of heatmap and coordinate regression, which can achieve superior accuracy results and faster inference speed compared with existing methods. In order to make full use of geometric information contained in the human face, the proposed Structure Loss can force the model to pay more attention to the correlation between landmarks. Additionally, we validate the efficiency and effectiveness of different modules in this paper. Extensive experiments on several benchmark datasets show that MTLD can outperform previous works and offer a better trade-off between accuracy and efficiency.

Acknowledgements. This research was supported by the Macao Polytechnic University (RP/FCSD-02/2022).

References

1. Jin, H., Liao, S., Shao, L.: Pixel-in-pixel net: towards efficient facial landmark detection in the wild. IJCV **129**(12), 3174–3194 (2021)
2. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: CVPR, pp. 2235–2245 (2018)
3. Xia, J., Qu, W., Huang, W., Zhang, J., Wang, X., Xu, M.: Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In: CVPR, pp. 4052–4061 (2022)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
5. Wang, J., et al.: Deep high-resolution representation learning for visual recognition. TPAMI **43**(10), 3349–3364 (2020)

6. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: a boundary-aware face alignment algorithm. In: CVPR, pp. 2129–2138 (2018)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020)
9. Chen, C.-F., Fan, Q., Panda, R.: Crossvit: cross-attention multi-scale vision transformer for image classification. In: CVPR, pp. 357–366 (2021)
10. Zhu, M., Shi, D., Zheng, M., Sadiq, M.: Robust facial landmark detection via occlusion-adaptive deep networks. In: CVPR, pp. 3486–3496 (2019)
11. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: CVPR, pp. 6971–6981 (2019)
12. Lin, C., et al.: Structure-coherent deep feature learning for robust face alignment. TIP **30**, 5313–5326 (2021)
13. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: ICCV Workshops, pp. 397–403 (2013)
14. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: ICCV, pp. 1513–1520 (2013)
15. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: ICCV Workshops, pp. 2144–2151. IEEE (2011)