Sparse Local Patch Transformer for Robust Face Alignment and Landmarks Inherent Relation Learning

Jiahao Xia[1], Weiwei Qu[2], Wenjian Huang[2], Jianguo Zhang[*,2], Xi Wang[3], Min Xu[*,1]
[1]Faculty of Engineering and IT, University of Technology Sydney
[2] Dept. of Comp. Sci. and Eng., Southern University of Science and Technology, [3]CalmCar
Jiahao.Xia@student.uts.edu.au, 11930667@mail.sustech.edu.cn, {huangwj, zhangjg}@sustech.edu.cn, Xi.Wang@calmcar.com, Min.Xu@uts.edu.au

Abstract

Heatmap regression methods have dominated face alignment area in recent years while they ignore the inherent relation between different landmarks. In this paper, we propose a Sparse Local Patch Transformer (SLPT) for learning the inherent relation. The SLPT generates the representation of each single landmark from a local patch and aggregates them by an adaptive inherent relation based on the attention mechanism. The subpixel coordinate of each landmark is predicted independently based on the aggregated feature. Moreover, a coarse-to-fine framework is further introduced to incorporate with the SLPT, which enables the initial landmarks to gradually converge to the target facial landmarks using fine-grained features from dynamically resized local patches. Extensive experiments carried out on three popular benchmarks, including WFLW, 300W and COFW, demonstrate that the proposed method works at the state-of-the-art level with much less computational complexity by learning the inherent relation between facial landmarks. The code is available at the project website[1].

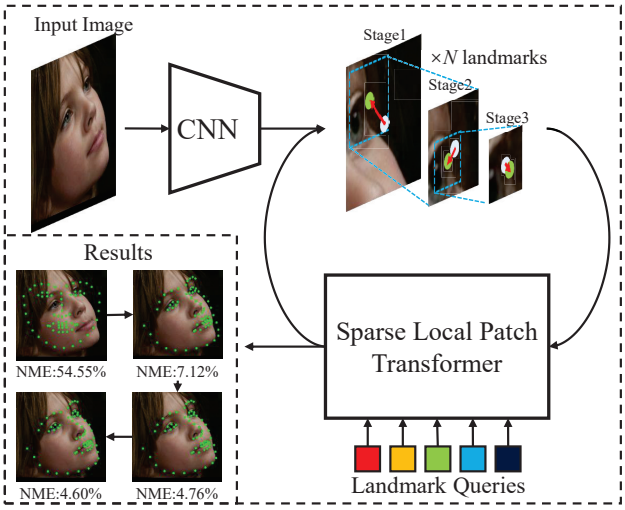


Figure 1. The proposed coarse-to-fine framework leverages the sparse local patches for robust face alignment. The sparse local patches are cropped according to the landmarks in the previous stage and fed into the same SLPT to predict the facial landmarks. Moreover, the patch size narrows down with the increasing of stages to enable the local features to evolve into a pyramidal form.

1. Introduction

Face alignment is aimed at locating a group of predefined facial landmarks from images. Robust face alignment based on deep learning technology has attracted increasing attention in recent years and it is the fundamental algorithm in many face-related applications such as face reenactment [40], face swapping [21] and driver fatigue detection [1]. Despite recent progress, it still remains a challenging problem, especially for images with heavy occlusion, profile view and illumination variation.

The inherent relation between facial landmarks play an important role in face alignment since human face has a regular structure. Although heatmap regression methods achieve impressive performance [7, 18, 33–35] in recent years, they still ignore the inherent relation because convolutional neural network (CNN) kernels focus locally, thus failed to capture the relations of landmarks farther away in a global manner. In particular, they consider the pixel coordinate with highest intensity of the output heatmap as the optimal landmark, which inevitably introduces a quantization error, especially for common downsampled heatmap. Coordinate regression methods [9,10,12,24,36,37,42] have an innate potential to learn the relation since it regresses the coordinates from global feature directly via fully-connected layers (FC). Nevertheless, a coherent relation should be learned together with local appearance while coordinate regression methods lose the local feature by projecting the

*Corresponding Author
[1]https://github.com/Jiahao-UTS/SLPT-master

global feature into FC layers.

To address the aforementioned problems, we propose a Sparse Local Patch Transformer (SLPT). Instead of predicting the coordinates from the full feature map like DETR [5], the SLPT firstly generates the representation for each landmark from a local patch. 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. Finally, the subpixel coordinate of each landmark in their corresponding patch is predicted independently by a MLP. Due to the use of sparse local patches, the number of the input token decreases significantly compared to other vision transformer [5, 11].

To further improve the performance, a coarse-to-fine framework is introduced to incorporate with the SLPT, as shown in Fig.1. Similar to cascaded shape regression method [13, 17, 44], 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.

To verify the effectiveness of the SLPT and the proposed framework, we carry out experiments on three popular benchmarks, WFLW [36], 300W [28] and COFW [4]. The results show the proposed method significantly outperforms other state-of-the-art methods in terms of diverse metrics with much lower computational complexity. Moreover, we also visualize the attention map of SLPT and the inner product matrix of landmark queries to demonstrate the SLPT can learn the inherent relation of facial landmarks.

The main contributions of this work can be summarized as:

- We introduce a novel transformer, Sparse Local Patch Transformer, to explore the inherent relation between facial landmarks based on the attention mechanism. The adaptive inherent relation learned by SLPT enables the model to achieve SOTA performance with much less computational complexity.
- We introduce a coarse-to-fine framework to incorporate with the SLPT, which enables the local patch to evolve in a pyramidal form and get closer to the target landmark for the fine-grained feature.
- Extensive experiments are conducted on three popular benchmarks, WFLW, 300W and COFW. The result illustrates the proposed method learns the inherent relation of facial landmarks by the attention mechanism and works at the SOTA level.

2. Related Work

In the early stage of face alignment, the mainstream methods [4, 6, 13, 24, 27, 31, 39, 44] regress facial landmarks directly from the local feature with classical machine learning algorithms like random forest. With the development of CNN, the CNN-based face alignment methods have achieved impressive performance. They can be roughly divided into two categories: heatmap regression method and coordinate regression method.

2.1. Coordinate Regression Method

Coordinate regression methods [12,37,41,42] regress the coordinates of landmarks from feature map directly via FC layers. To further improve the robustness, diverse cascaded networks [17, 30] and recurrent networks [38] are proposed to achieve face alignment with multi stages. Despite coordinate regression methods have an innate potential to learn the inherent relation, it commonly requires a huge number of samples for training. To address the problem, Qian et al. [26] and Dong et al. [9] expand the number of training samples by style transfer; Browatzki et al. [3] and Dong et al. [10] leverage the unlabeled dataset to train the model. In recent years, state-of-the-art works employ the structure information of face as the prior knowledge for better performance. Lin et al. [24] and Li et al. [22] model the interaction between landmarks by a graph convolutional network (GCN). However, the adjacency matrix of GCN is fixed during inference and cannot adjust case by case. Learning an *adaptive* inherent relation is crucial for robust face alignment. Unfortunately, there is no work yet on this topic, and we propose a method to fill this gap.

2.2. Heatmap Regression Method

Heatmap regression methods [7, 25, 29, 34] output an intermediate heatmap for each landmark and consider the pixel with highest intensity as the optimal output. Therefore, it leads to quantization errors since the heatmap is commonly much smaller than the input image. To eliminate the error, Kumar et al. [18] estimate the uncertainty of predicted landmark locations; Lan et al [19] adopt an additional decimal heatmap for subpixel estimation; Huang et al. [15] further regress the coordinate from an anisotropic attention mask generated from heatmaps. Moreover, heatmap regression methods also ignore the relation between landmarks. To construct the relation between neighboring points, Wu et al. [36] and Wang et al. [35] take advantage of facial boundaries as the prior knowledge; Zou et al. [47] cluster landmarks with a graph model to provide structural constraints. However, they still cannot explicitly model an inherent relation between the landmarks with long distance.

The vision transformer [11] proposed recently enables the model to attend the area with a long distance. Besides, the attention mechanism in transformer can generate

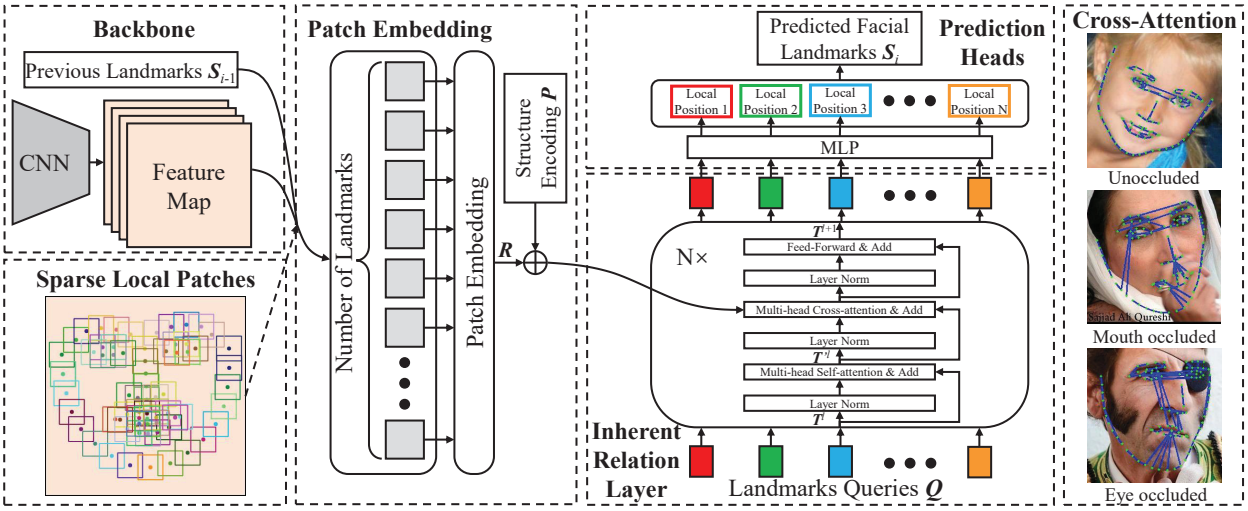


Figure 2. An overview of the SLPT. The SLPT crops local patches from the feature map according to the facial landmarks in the previous stage. Each patch is then embedded into a vector that can be viewed as the representation of the corresponding landmark. Subsequently, they are supplemented with the structure encoding to obtain the relative position in a regular face. A fixed number of landmark queries are then input into the decoder, attending the vectors to learn the inherent relation between landmarks. Finally, the outputs are fed into a shared MLP to estimate the position of each facial landmark independently. The rightmost images demonstrate the adaptive inherent relation of different samples. We connect each point to the point with highest cross-attention weight in the first inherent relation layer.

an adaptive global attention for different tasks, such as object detection [5,46] and human pose estimation [23], and in principle, we envision that it can also learn an adaptive inherent relation for face alignment. In this paper, we demonstrate the capability of SLPT for learning the relation.

3. Method

3.1. Sparse Local Patch Transformer

As shown in Fig.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 [11] 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, 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.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 [32], 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:

$$\boldsymbol{A}_h = softmax\left(\frac{\left(\boldsymbol{T}_h^l + \boldsymbol{Q}_h\right)\boldsymbol{W}_h^q\left(\left(\boldsymbol{T}_h^l + \boldsymbol{Q}_h\right)\boldsymbol{W}_h^k\right)^T}{\sqrt{C_h}}\right), \tag{1}$$

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:

$$MSA\left(\boldsymbol{T}^l\right) = \left[\boldsymbol{A}_1\boldsymbol{T}_1^l\boldsymbol{W}_1^v; ...; \boldsymbol{A}_H\boldsymbol{T}_H^l\boldsymbol{W}_H^v\right]\boldsymbol{W}_P, \tag{2}$$

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.

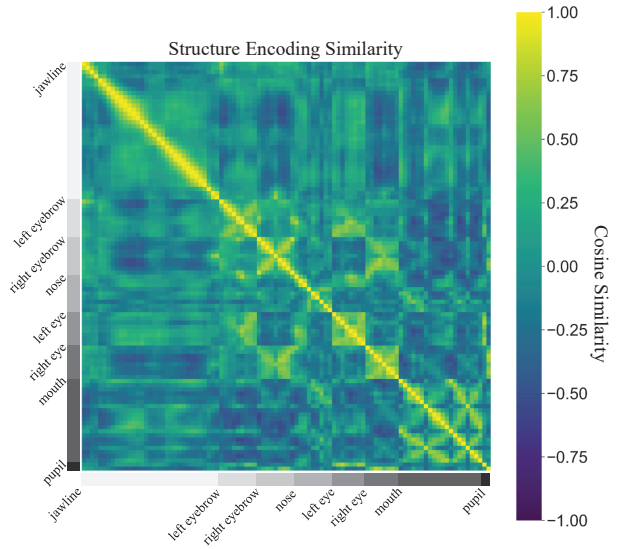


Figure 3. Cosine similarity for structure encodings of SLPT learned from a dataset with 98 landmark annotations. High cosine similarities are observed for the corresponding points which are close in the regular face structure.

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.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:

$$\boldsymbol{A}'_h = softmax\left(\frac{\left(\boldsymbol{T}'^l_h + \boldsymbol{Q}_h\right)\boldsymbol{W}'^q_h\left(\left(\boldsymbol{R}_h + \boldsymbol{P}_h\right)\boldsymbol{W}'^k_h\right)^T}{\sqrt{C_h}}\right). \tag{3}$$

Where $\boldsymbol{W}'^q_h$ and $\boldsymbol{W}'^k_h \in \mathbb{R}^{C_h \times C_h}$ are learnable parameters of two linear layers in the h-th head. $\boldsymbol{T}'^l_h \in \mathbb{R}^{N \times C_h}$ 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:

$$MCA\left(\boldsymbol{T}'^l\right) = \left[\boldsymbol{A}'_1\boldsymbol{T}'^l_1\boldsymbol{W}'^v_1; ...; \boldsymbol{A}'_H\boldsymbol{T}'^l_H\boldsymbol{W}'^v_H\right]\boldsymbol{W}'_P, \tag{4}$$

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

$$\Omega(S) = 4HNC_h^2 + 2HN^2C_h, \tag{5}$$

$$\Omega(F) = \left(2N + 2\frac{W_I H_I}{P_w P_h}\right)HC_h^2 + 2NH\frac{W_I H_I}{P_w P_h}C_h. \tag{6}$$

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

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 [5]), which decreases the computational complexity significantly. For a 29 landmark dataset [4], $\Omega(S)$ 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 $\left(x^i, y^i\right)$ can be calculated by:

$$\begin{aligned} x^i &= x_{lt}^i + w^i t_x^i, \\ y^i &= y_{lt}^i + h^i t_y^i, \end{aligned} \tag{7}$$

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

3.2. 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

Method	NME(%)↓	$FR_{0.1}$(%)↓	$AUC_{0.1}$↑
LAB [36]	5.27	7.56	0.532
SAN [9]	5.22	6.32	0.535
Coord⋆ [34]	4.76	5.04	0.549
DETR† [5]	4.71	5.00	0.552
Heatmap⋆ [34]	4.60	4.64	0.524
AVS+SAN [26]	4.39	4.08	0.591
LUVLi [18]	4.37	3.12	0.557
AWing [35]	4.36	2.84	0.572
SDFL⋆ [24]	4.35	**2.72**	0.576
SDL⋆ [22]	4.21	3.04	0.589
HIH [19]	4.18	2.84	**0.597**
ADNet [15]	**4.14**	**2.72**	**0.602**
SLPT‡	**4.20**	3.04	0.588
SLPT†	**4.14**	**2.76**	0.595

Table 1. Performance comparison of the SLPT and the state-of-the-art methods on WFLW. The normalization factor is inter-ocular and the threshold for FR is set to 0.1. Key: [**Best**, **Second Best**, ⋆=HRNetW18C, †=HRNetW18C-lite, ‡=ResNet34]

Method	Inter-Ocular NME (%) ↓		
	Common	Challenging	Fullset
SAN [9]	3.34	6.60	3.98
Coord⋆ [34]	3.05	5.39	3.51
LAB [36]	2.98	5.19	3.49
DeCaFA [7]	2.93	5.26	3.39
HIH [19]	2.93	5.00	3.33
Heatmap⋆ [34]	2.87	5.15	3.32
SDFL⋆ [24]	2.88	4.93	3.28
HG-HSLE [47]	2.85	5.03	3.28
LUVLi [18]	2.76	5.16	3.23
AWing [35]	2.72	**4.53**	3.07
SDL⋆ [22]	**2.62**	4.77	**3.04**
ADNet [15]	**2.53**	**4.58**	**2.93**
SLPT‡	2.78	4.93	3.20
SLPT†	2.75	4.90	3.17

Table 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]

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. (see Section 4.5).

3.3. 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 [25, 29], 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:

$$L = \frac{1}{SDN}\sum_{i=1}^{S}\sum_{j=1}^{D}\sum_{k=1}^{N}\frac{\left\|\left(x_{gt}^{k}, y_{gt}^{k}\right) - \left(x^{ijk}, y^{ijk}\right)\right\|_2}{d}, \tag{8}$$

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. Experiment

4.1. Datasets

Experiments are conducted on three popular benchmarks, including WFLW [36], 300W [28] and COFW [4].

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 [45], the training subset of HELEN [20] and LFPW [2]. The test set is further divided into a challenging subset that includes 135 images (IBUG fullset [28]) 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 [14]). Both contains 507 images. We employ the COFW68 set for *cross*-dataset validation.

4.2. Evaluation Metrics

Referring to other related work [18, 24, 35], we evaluate the proposed methods with standard metrics, Normalized Mean Error (NME), Failure Rate (FR) and Area Under Curve (AUC). **NME** is defined as:

$$NME\left(\boldsymbol{S}, \boldsymbol{S}_{gt}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\boldsymbol{p}^{i} - \boldsymbol{p}_{gt}^{i}\right\|_2}{d} \times 100\%, \tag{9}$$

where $\boldsymbol{S}$ and $\boldsymbol{S}_{gt}$ denote the predicted and annotated coordinates of landmarks respectively. $\boldsymbol{p}^{i}$ and $\boldsymbol{p}_{gt}^{i}$ indicate the coordinate of i-th landmark in $\boldsymbol{S}$ and $\boldsymbol{S}_{gt}$. N is the number of

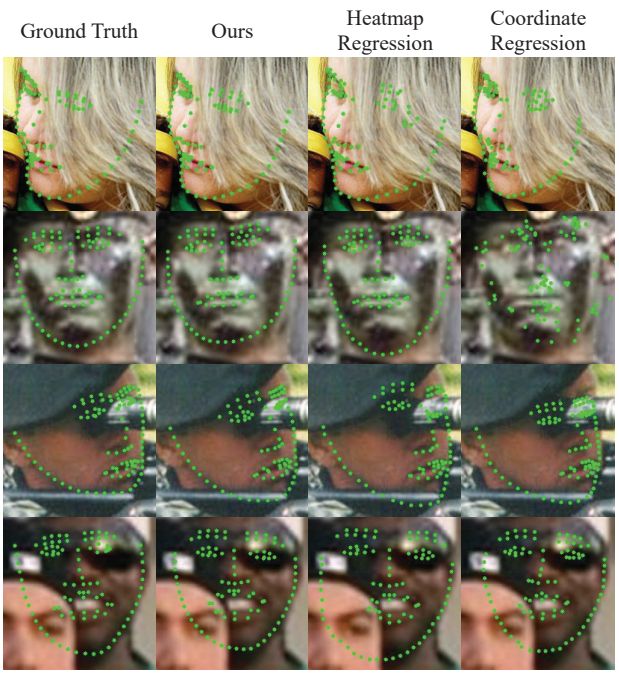


Figure 4. Visualization of the ground truth and face alignment result of SLPT, heatmap regression (HRNetW18C) and coordinate regression (HRNetW18C) method on the faces with blur, heavy occlusion and profile face.

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 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. Implementation Details

Each input image is cropped and resized to 256×256. We train the proposed framework with Adam [8], 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 ($\pm 5\%$), rotation ($\pm 30°$), translation ($\pm 10px$). We implement our method with two different backbone: a light HRNetW18C [34] (the modularized block number in each stage is set to 1) and Resnet34 [16]. For the HRNetW18C-lite, the resolution of feature map is 64×64, and for the Resnet34, we extract representations from the output feature maps of stages C2 through C5. (see Appendix A.1).

Method	Inter-Ocular NME(%)↓	Inter-Ocular FR(%)↓	Inter-Pupil NME(%)↓	Inter-Pupil FR(%)↓
DAC-CSR [13]	6.03	4.73	-	-
LAB [36]	3.92	0.39	-	-
Coord* [34]	3.73	0.39	-	-
SDFL* [24]	3.63	**0.00**	-	-
Heatmap* [34]	3.45	**0.20**	-	-
Human [4]	-	-	5.60	-
TCDCN [42]	-	-	8.05	-
Wing [12]	-	-	5.44	3.75
DCFE [31]	-	-	5.27	7.29
AWing [35]	-	-	4.94	**0.99**
ADNet [15]	-	-	**4.68**	**0.59**
SLPT‡	**3.36**	0.59	4.85	1.18
SLPT†	**3.32**	**0.00**	**4.79**	1.18

Table 3. NME and $FR_{0.1}$ comparisons under Inter-Ocular normalization and Inter-Pupil normalization on *within*-dataset validation. The threshold for failure rate (FR) is set to 0.1. Key: [**Best**, **Second Best**, *=HRNetW18C, †=HRNetW18C-lite, ‡=ResNet34]

Method	Inter-Pupil NME(%)↓	$FR_{0.1}$(%)↓
TCDCN [42]	7.66	16.17
CFSS [44]	6.28	9.07
ODN [43]	5.30	-
AVS+SAN [26]	4.43	2.82
LAB [36]	4.62	2.17
SDL* [22]	4.22	**0.39**
SDFL* [24]	4.18	**0.00**
SLPT‡	**4.11**	0.59
SLPT†	**4.10**	0.59

Table 4. Inter-ocular NME and $FR_{0.1}$ comparisons on 300W-COFW68 *cross*-dataset evaluation. Key: [**Best**, **Second Best**, *=HRNetW18C, †=HRNetW18C-lite, ‡=ResNet34]

4.4. Comparison with State-of-the-Art Method

WFLW: as tabulated in Table 1 (more detailed results on the subset of WFLW are in Appendix A.2), SLPT demonstrates impressive performance. With the increasing of inherent layers, the performance of SLPT can be further improved and outperforms the ADNet (see Appendix A.5). Referring to DETR, we also implement a Transformer based method that employs the full feature map for face alignment. The number of the input tokens is 16×16. 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 (see Appendix A.3). Moreover, the SLPT also outperforms the coordinate regreesion and heatmap regression methods significantly. Some qualitative results are shown in Fig. 4. It is evident that our method could localize the landmarks accurately, in partic-

Model	Intermediate Stage											
	1st stage			2rd stage			3rd stage			4th stage		
	NME	FR	AUC	NME	FR	AUC	NME	FR	AUC	NME	FR	AUC
Model† with 1 stage	4.79%	5.08%	0.583	-	-	-	-	-	-	-	-	-
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 5. Performance comparison of the SLPT with different number of coarse-to-fine stages on WFLW. The normalization factor for NME is inter-ocular and the threshold for FR and AUC is set to 0.1. Key: [**Best**, †=HRNetW18C-lite]

Method	MSA	MCA	NME	FR	AUC
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 6. NME(↓), $FR_{0.1}$(↓) and $AUC_{0.1}$(↑) with/without Encoder and Decoder. Key: [**Best**, †=HRNetW18C-lite]

Method	NME	FR	AUC
w/o structure encoding†	4.16%	2.84%	0.593
w structure encoding†	**4.14%**	**2.76%**	**0.595**

Table 7. NME(↓), $FR_{0.1}$(↓) and $AUC_{0.1}$(↑) with/without structure encoding. Key: [**Best**, , †=HRNetW18C-lite]

ular for face images with blur ($2nd$ row in Fig.4), profile view ($1st$ row in Fig.4) and heavy occlusion ($3rd$ and $4th$ row in Fig.4).

300W: the comparison result is shown in Table 2. Compared to the coordinate and heatmap regression methods (HRNetW18C [34]), 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.

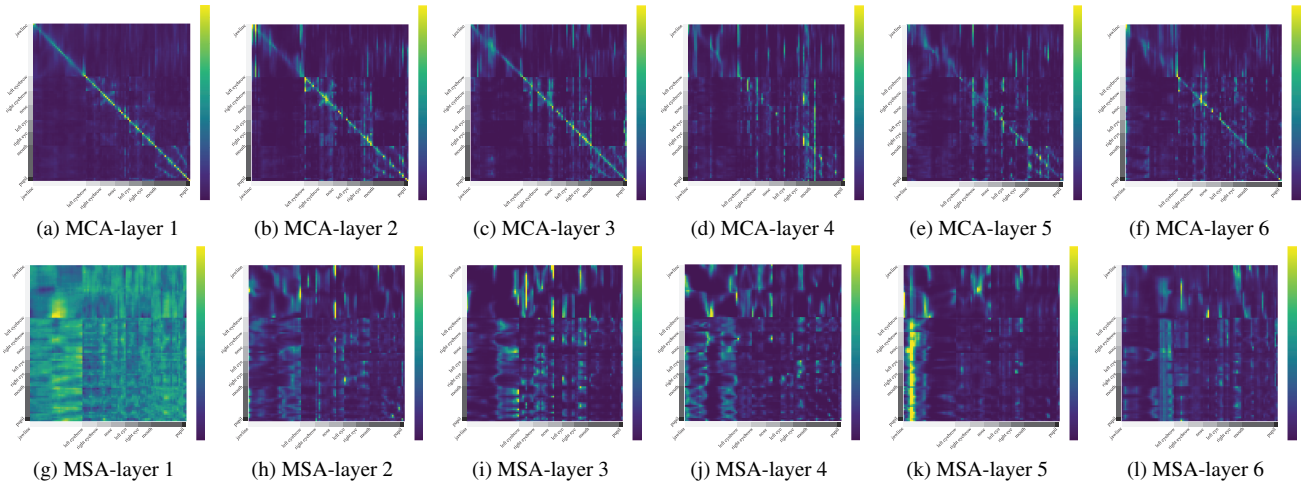
COFW: We conduct two experiments on COFW for comparsion, the *within*-dataset validation and *cross*-dataset validation. For the *within*-dataset validation, the model is trained with 1,345 images and validated with 507 images on COFW. The inter-ocular and inter-pupil NME of SLPT and the state-of-the-art methods are reported in Table 3 respectively. In this experiment, the number of training sample is quite small, which leads to the significant degradation of 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.

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. 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 face alignment, especially for the condition of heavy occlusion.

4.5. Ablation Study

Evaluation on different 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 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 (eq.1) and *representation-query* inter relation (eq.3) 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 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-



(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. 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.

Method	FLOPs(G)	Params(M)
HRNet* [34]	4.75	9.66
LAB [36]	18.85	12.29
AVS + SAN [26]	33.87	35.02
AWing [35]	26.8	24.15
DETR† (98 landmarks) [5]	4.26	11.00
DETR† (68 landmarks) [5]	4.06	11.00
DETR† (29 landmarks) [5]	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 8. Computational complexity and parameters of SLPT and SOTA methods. Key: [*=HRNetW18C, †=HRNetW18C-lite]]

attention and cross-attention can be combined to improve the performance of model further.

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 7.

Evaluation on computational complexity: the computational complexity and parameters of SLPT and other SOTA methods are shown in Table 8. 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.

Besides, the influence of patch size and inherent layer number are shown in the Appendix A.4 and A.5.

4.6. Visualization

We calculate the mean attention weight of each MCA and MSA block on the WFLW test set, as shown in Fig.5. 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.

5. Conclusion

In this paper, we find out that the inherent relation between landmarks is significant to the performance of face alignment while it is ignored by the most state-of-the-art methods. To address the problem, we propose a sparse local patch transformer for learning a *query-query* and a *representation-query* relation. 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 face alignment, 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.

Acknowledgment

This work was sponsored by the program of China Scholarships Council (No. 202006130004).

References

[1] Bram Bakker, Bartosz Zabłocki, Angela Baker, Vanessa Riethmeister, Bernd Marx, Girish Iyer, Anna Anund, and Christer Ahlström. A multi-stage, multi-feature machine learning approach to detect driver sleepiness in naturalistic road driving conditions. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–10, 2021.

[2] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, pages 545–552, 2011.

[3] Björn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *CVPR*, pages 6109–6119, 2020.

[4] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, pages 1513–1520, 2013.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[6] David Cristinacce and Tim Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 3, page 929–938, 2006.

[7] Arnaud Dapogny, Matthieu Cord, and Kevin Bailly. Decafa: Deep convolutional cascade for face alignment in the wild. In *ICCV*, pages 6892–6900, 2019.

[8] Kingma Diederik and Ba Jimmy. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[9] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, pages 379–388, 2018.

[10] Xuanyi Dong, Yi Yang, Shih-En Wei, Xinshuo Weng, Yaser Sheikh, and Shoou-I Yu. Supervision by registration and triangulation for landmark detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3681–3694, 2021.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[12] Zhenhua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiaojun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, pages 2235–2245, 2018.

[13] Zhenhua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiaojun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *CVPR*, pages 3681–3690, 2017.

[14] Golnaz Ghiasi and Charless C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1899–1906, 2014.

[15] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *2021 ICCV*, pages 3060–3070, 2021.

[16] He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[17] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPRW*, pages 2034–2043, 2017.

[18] Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, pages 8233–8243, 2020.

[19] Xing Lan, Qinghao Hu, and Jian Cheng. Revisting quantization error in face alignment. In *2021 ICCVW*, pages 1521–1530, 2021.

[20] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. Interactive facial feature localization. In *ECCV*, pages 679–692, 2012.

[21] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *CVPR*, pages 5073–5082, 2020.

[22] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, and Shun Miao. Structured landmark detection via topology-adapting deep graph learning. In *ECCV 2020*, pages 266–283, Cham, 2020. Springer International Publishing.

[23] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *ICCV*, 2021.

[24] Chunze Lin, Beier Zhu, Quan Wang, Renjie Liao, Chen Qian, Jiwen Lu, and Jie Zhou. Structure-coherent deep feature learning for robust face alignment. *IEEE Transactions on Image Processing*, 30:5313–5326, 2021.

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016.

[26] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *ICCV*, pages 10152–10162, 2019.

[27] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, pages 1685–1692, 2014.

[28] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, pages 397–403, 2013.

[29] Zhiqiang Tang, Xi Peng, Kang Li, and Dimitris N. Metaxas. Towards efficient u-nets: A coupled and quantized approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2038–2050, 2020.

[30] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou.

Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, pages 4177–4187, 2016.

[31] Roberto Valle, José M. Buenaposada, Antonio Valdés, and Luis Baumela. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *ECCV*, pages 609–624, Cham, 2018.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, page 6000–6010, Red Hook, NY, USA, 2017.

[33] Jun Wan, Zhihui Lai, Jun Liu, Jie Zhou, and Can Gao. Robust face alignment by multi-order high-precision hourglass network. *IEEE Transactions on Image Processing*, 30:121–133, 2021.

[34] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021.

[35] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, pages 6970–6980, 2019.

[36] Wenyan Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138, 2018.

[37] Wenyan Wu and Shuo Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *CVPRW*, pages 2096–2105, 2017.

[38] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, pages 57–72, Cham, 2016.

[39] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[40] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *CVPR*, pages 5325–5334, 2020.

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[42] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108, Cham, 2014.

[43] Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *CVPR*, pages 3481–3491, 2019.

[44] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, pages 4998–5006, 2015.

[45] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012.

[46] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2020.

[47] Xu Zou, Sheng Zhong, Luxin Yan, Xiangyun Zhao, Jiahuan Zhou, and Ying Wu. Learning robust facial landmark detection via hierarchical structured ensemble. In *2019 ICCV*, pages 141–150, 2019.