Robust Face Alignment via Inherent Relation Learning and Uncertainty Estimation

Jiahao Xia, Min Xu, Haimin Zhang, Jianguo Zhang, Wenjian Huang, Hu Cao and Shiping Wen

Abstract—Human tends to locate the facial landmarks with heavy occlusion by their relative position to the easily identified landmarks. The clue is defined as the landmark inherent relation while it is ignored by most existing methods. In this paper, we present Dynamic Sparse Local Patch Transformer (DSLPT), a novel face alignment framework for the inherent relation learning and uncertainty estimation. Unlike most existing methods that regress facial landmarks directly from global features, the DSLPT firstly generates a rough representation of each landmark from a local patch cropped from the feature map and then adaptively aggregates them by a case dependent inherent relation. Finally, the DSLPT predicts the coordinate and uncertainty of each landmark by regressing their probability distribution from the output features. Moreover, we introduce a coarse-to-fine framework to incorporate with DSLPT for an improved result. In the framework, the position and size of each patch are determined by the probability distribution of the corresponding landmark predicted in the previous stage. The dynamic patches will ensure a fine-grained landmark representation for inherent relation learning so that a rough prediction result can gradually converge to the target facial landmarks. We integrate the coarse-to-fine model into an end-to-end training pipeline and carry out experiments on the mainstream benchmarks. The results demonstrate that the DSLPT achieves state-of-the-art performance with much less computational complexity. The codes and models are available at https://github.com/Jiahao-UTS/DSLPT.

Index Terms—Face Alignment, Coarse-to-fine Regression, Inherent Relation Learning, Uncertainty Estimation.

1 Introduction

Face alignment aims at predicting a group of pre-defined landmarks from a face image. It is the basis for many computer vision applications, such as facial emotion recognition [1], [2], face recognition [3], face parsing [4] and face reenactment [5]. Despite recent progress, it still suffers from some limitations, such as high computational complexity and low robustness with heavy occlusion, illumination variation and profile view.

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. Although both heatmap regression methods [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] [22], [23] and coordinate regression methods [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37] [38], [39], [40], [41] show an impressive improvement in recent years, none of them take the inherent relation among landmarks into consideration, which results in fragile robustness. In terms of heatmap regression methods, these methods usually fail to capture the relations of landmarks farther away in a global manner since convolutional neural network kernels focus locally. A coherent inherent relation should be learned together with a fine-grained local appearance while coordinate regression methods lose the information by directly projecting the feature map into fully connected (FC) layers. Moreover, the weights of FC layers are frozen during inference, resulting in the model not being able to aggregate feature adaptively for a case dependent relation.

The other issue that limits the performance of face alignment is that most existing face alignment methods are based on an assumption: the variance of the probability distribution for all landmarks is a constant number. 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 [6], [34], [41] 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.

In this paper, we propose a novel framework, Dynamic Sparse Local Patch Transformer (DSLPT) to solve the two aforementioned problems. Unlike existing coordinate regression methods that directly project the feature map into FC layers, DSLPT firstly crops a local patch for each landmark according to an initial mean face calculated from training samples [30] 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 repre-

- *Jiahao Xia, Min Xu, Haimin Zhang and Shiping Wen are with the Faculty of Engineering and IT, University of Technology Sydney, NSW 2007, Australia. E-mail: Jiahao.Xia@student.uts.edu.au, {Min.Xu, Haimin.Zhang, Shiping.Wen}@uts.edu.au*
- *Jianguo Zhang and Wenjian Huang are with the Department of Computer Science and Engineering, Southern University of Science and Technology, Guangdong 518055, China. Email: {zhangjg, huangwj}@sustech.edu.cn*
- *Hu Cao is with Chair of Robotics, AI and Real time system, Technical University of Munich, Munich 85748, Germany. Email: hu.cao@tum.de*
- *Corresponding author: Min Xu and Jianguo Zhang*

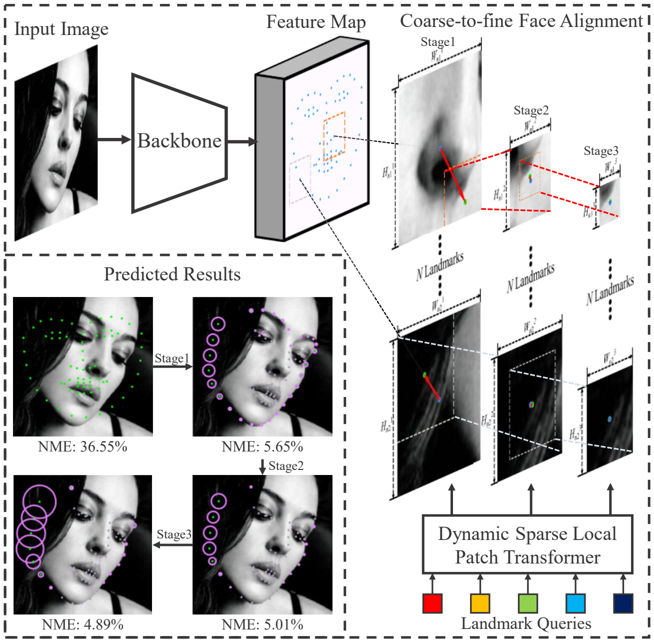


Fig. 1: The proposed coarse-to-fine framework leverages the dynamic patches for robust face alignment. The initial local patches are cropped according to a mean face. Then, the size and position of each patch are adjusted dynamically according to the location and uncertainty predicted in the previous 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.

sentations based on the attention mechanism, which enables DSLPT to learn a case dependent inherent relation. Instead of predicting the numerical coordinates directly, DSLPT predicts the probability distribution of each landmark with log-likelihood estimation. The mean and variance of the distribution can be viewed as the coordinate and uncertainty respectively.

To further improve the performance of DSLPT, we introduce a coarse-to-fine framework to incorporate with DSLPT so that a rough predicted result can converge to the target facial landmarks gradually, as shown in Fig. 1. Instead of using the patch with a fixed size like other patch-based regression methods, the position and size of each patch 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 for more contextual information and a relatively small patch size to the landmarks with low uncertainty for high feature resolution. It enables the model to obtain the advantages of large patch size and small patch size simultaneously.

Moreover, compared to heatmap regression and coordinate regression methods, DSLPT has the following extra unique advantages: 1) The sparse local patches significantly decrease the token number in Transformer [42], making DSLPT more efficient than other methods, especially heatmap regression methods. 2) DSLPT retains the spatial information and aligns the features to the corresponding landmarks by the dynamic local patches. 3) DSLPT does not have a quantization error as heatmap regression methods. Most heatmap regression methods cannot predict the fractional part of landmark coordinates, which results in an inevitable quantization error during transforming the predicted heatmaps into the numerical coordinates.

To verify the effectiveness of DSLPT, we combine DSLPT with a series of backbone and carry out extensive experiments on eight popular benchmarks with different number of pre-defined landmarks. The results demonstrate that DSLPT achieves the state-of-the-art performance on all benchmarks with only $1/5 \sim 1/2$ computational complexity. Moreover, to further verify the transferable capability of DSLPT, we extend DSLPT to human pose estimation. The results and discussions can be found in the **supplementary file**.

A preliminary version of this work appeared as [43]. In this extended journal version, we further address the limitations of the previous work from the following aspects: 1) DSLPT predicts the probability distribution of each landmark rather than a numerical coordinate like the original Sparse Local Patch Transformer (SLPT). The distribution prediction promises a more coherent result and the variance can further serve as the uncertainty of the corresponding landmark. 2) The patch size in DSLPT is determined by the predicted uncertainty while the patch size in the original version is a constant number. The dynamic patch size enables DSLPT to employ more contextual information to locate the landmark with high uncertainty and high feature resolution to locate the landmark with low uncertainty. 3) We implement DSLPT with four well-designed backbones and evaluate them on eight benchmarks. A more comprehensive analysis of their performance and computational complexity in different conditions is given. 4) We extend DSLPT to human pose estimation to verify its transferable capability. The results illustrate it is possible to apply DSLPT in other more challenging tasks.

The main contributions of this work can be summarized as:

- A Dynamic Sparse Local Patch Transformer is proposed to explicitly learn a case dependent inherent relation so that the landmark with heavy occlusion can be located robustly according to their relative position to the easily identified landmarks.
- This paper proposes the dynamic patch, a kind of patch whose position and size adjust according to the predicted probability distribution of the corresponding landmark. It ensures more fine-grained landmark representations for better robustness compared to the patches with a fixed size.
- We further introduce a coarse-to-fine framework to incorporate with the DSLPT and integrate it into an end-to-end training pipeline. The framework enables a rough predicted result to converge to the target facial landmarks gradually.
- Extensive experiments and ablation studies are carried out on eight widely used face alignment benchmarks. The results demonstrate that the pro-

posed method achieves competitive performance with much less computational complexity. Besides, we extend DSLPT to human pose estimation and the results show DSLPT has good transferable capability.

2 Related Work

As a fundamental technique for many applications, face alignment has been investigated for decades. In the early stage, face alignment methods commonly rely on generative PCA-based shape models, including Active Appearance Model (AAM) [44], Active Shape Model (ASM) [45], Constrained Local Model (CLM) [46] and their extensions [47], [48]. Because of the fragile robustness and low generalization, they can only be applied in constrained scenarios. To achieve face alignment in the wild, a more robust architecture, Cascade Shape Regression (CSR) model [49], [50], [51] [52], [53] is proposed and it dominates face alignment until convolutional neural network (CNN) is widely applied in face alignment. The existing CNN based methods can be roughly divided into two categories: coordinate regressing methods [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37] [38], [39], [40], [41] and heatmap regression methods [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] [22], [23]. With stronger expressive capability, CNN based methods have boosted the performance of face alignment significantly.

2.1 Coordinate Regression Methods

Coordinate regression methods regress the coordinate of facial landmarks from the input image directly. To improve the robustness of coordinate regression methods, diverse cascaded framework [27], [30], [34], recurrent networks [24], [25] and optimized loss [32], [40] are proposed. Zadeh et al. [6], Liu et al. [34] and Zhu et al. [41] further develop novel patch based regression models to predict a fine-grained result on the local patches. Nevertheless, their patch size is fixed and determined by prior knowledge, which cannot ensure an effective feature for all cases. Besides, patch based methods regress the coordinate of each facial landmark merely from its corresponding patch, ignoring the contextual information of other patches. Despite DSLPT also being a patch based regression method, these drawbacks can be tackled by the proposed dynamic patch and inherent relation learning.

Different from other tasks, the number of training samples in face alignment is very limited, which usually leads to overfitting. To address the problem, Qian et al. [54] and Dong et al. [35] expand the training samples by style transfer; Browatzki et al. [20] and Dong et al. [11] leverage unlabeled samples for training by semi-supervised learning. Zhu et al. [28], [29], Guo et al. [37] and Wu et al. [39] further achieve 3D face alignment by directly regressing the parameters of the 3D morphable model (3DMM) [55]. In recent years, many methods have noticed that face structure is crucial to the performance of face alignment. Lin et al. [40] retain the structure information by an adjacency matrix defined by prior knowledge. Li et al. [38] further improve the performance by a learnable adjacency matrix so that the network can explore a task-specific structure. The structure information is also the cornerstone of inherent relation learning. Therefore, the proposed structure encoding is introduced into DSLPT for retaining face structure information, encoding the landmark distance of a regular face into cosine similarity.

2.2 Heatmap Regression Methods

Heatmap regression methods regress an intermediate heatmap for each landmark by a backbone with a series of downsampling and upsampling layers [7], [8], [9], [10] and 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, Lan et al. [22] adopt an additional decimal heatmap for subpixel estimation; Zhang et al. [18] utilize another network for subpixel offset estimation; Chen et al. [15], Tai et al. [17] and Kumar et al. [19] further predict landmark probability distribution on the heatmap for subpixel coordinate.

Moreover, most heatmap regression methods also ignore the inherent relation between landmarks. Wu et al [31], Wang et al. [13] and Huang et al. [23] set facial boundary heatmap regression as an additional regressing objective for learning the relation between neighboring landmarks. Zou et al. [16] 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) [42] breaks the record of many computation vision tasks, such as image classification [42], [56], [57], object detection [58], [59] and semantic segmentation [60]. Although ViT models also break the record of a very similar task, human pose estimation [61], [62], directly applying ViT in face alignment does not promise an improvement because training ViT requires a large number of training samples. Lan et al. [22] generates decimal heatmaps by ViT. Unfortunately, the ViT based model fails to outperform CNN based model. To address the problem, Zheng et al. [63] pretrain ViT on a very large dataset by contrastive learning and fine-tune the model with annotated samples for heatmap regression. Different from other works, the patch size of DSLPT is dynamic and determined by the predicted uncertainty, which significantly augments the training data. Therefore, DSLPT achieves state-of-the-art performance with very limited training samples.

3 Methodology

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 face alignment and we will describe them in the following sections.

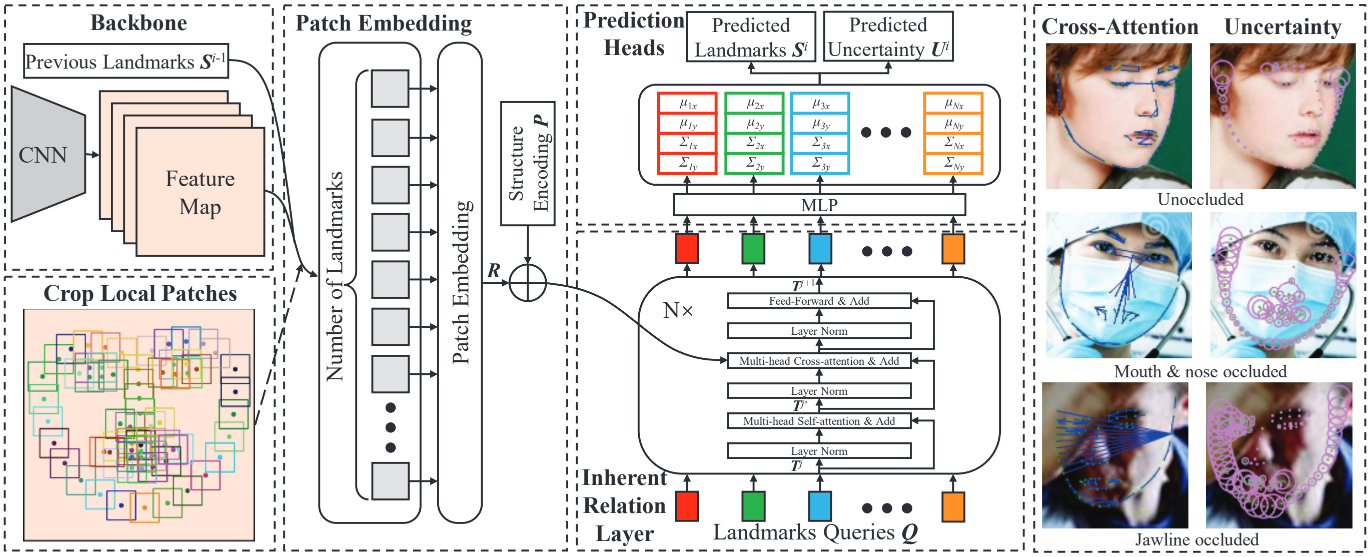
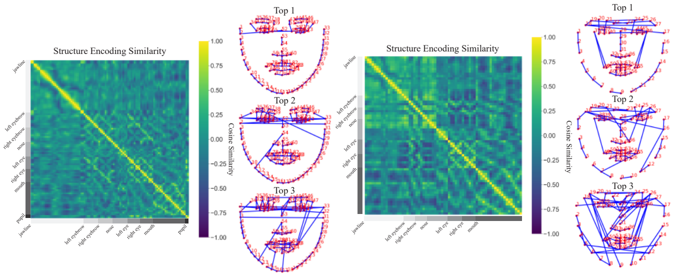

Backbone
Previous Landmarks S^{i-1}
CNN
Feature Map
Crop Local Patches
Patch Embedding
Number of Landmarks
Patch Embedding
Structure Encoding P
R
Prediction Heads
Predicted Landmarks S^i
Predicted Uncertainty U^i
μ_{1x} μ_{1y} Σ_{1x} Σ_{1y}
μ_{2x} μ_{2y} Σ_{2x} Σ_{2y}
μ_{3x} μ_{3y} Σ_{3x} Σ_{3y}
μ_{Nx} μ_{Ny} Σ_{Nx} Σ_{Ny}
MLP
N×
T^{i+1}
Feed-Forward & Add
Layer Norm
Multi-head Cross-attention & Add
Layer Norm
T'
Multi-head Self-attention & Add
Layer Norm
T^i
Inherent Relation Layer
Landmarks Queries Q
Cross-Attention
Uncertainty
Unoccluded
Mouth & nose occluded
Jawline occluded

Fig. 2: A pipeline of the DSLPT. The DSLPT firstly crops the supporting patch for each landmark from the feature map according to the facial landmarks predicted in the previous stage. The patch is then embedded into a vector as the representation $\boldsymbol{R}$ of the corresponding landmark. Subsequently, they are added with the structure encoding $\boldsymbol{P}$ to retain the structure information of face. A fixed number of landmark queries $\boldsymbol{Q}$ are fed into the inherent relation layer, adaptively aggregating the representations based on a learned inherent relation. Finally, the probability distribution of each landmark is predicted independently from the output features. The rightmost images demonstrate the predicted uncertainty and the inherent relation of different cases. Each landmark is connected with the landmark with the highest cross attention weight.



(a) 98 landmarks

(b) 68 landmarks

Fig. 3: **Left**: cosine similarity of the structure encodings learned from 98 landmarks and 68 landmarks datasets. The high cosine similarity between two encodings indicates the corresponding landmarks are close in the regular face shape. **Right**: to better visualize, we connect each landmark to the landmark with the highest, second highest and third highest similarity respectively.

3.1 Dynamic Sparse Local Patch Transformer

As shown in Fig. 2, the DSLPT consists of three complementary components: patch embedding & structure encoding, inherent relation layers and prediction heads.

3.1.1 Patch Embedding & Structure Encoding

Instead of learning a global representation like other ViT models [42], DSLPT aims at learning the inherent relation among landmarks. Therefore, the input to DSLPT should be landmark representations rather than regular patches. To generate the landmark representation, 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 as the structure information in many works [40], [38]. 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 [64]. 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. 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 [40], which means the structure information learned by unsupervised learning in DSLPT is quite close to human prior knowledge.

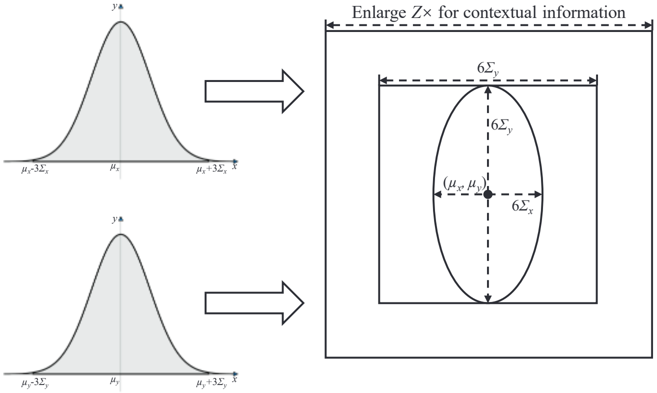


Fig. 4: An intuitive example of how the predicted probability distribution determines the patch size. DSLPT takes $[\mu - 3\Sigma, \mu + 3\Sigma]$ as the confidence interval, which promises that the probability that the landmarks are within the interval is more than 95%. Then, the size of ROI in the patch coordinate system can be written as $\max(6\Sigma_x, 6\Sigma_y)$. The size is finally enlarged $Z\times$ for background information.

3.1.2 *Inherent Relation Layer*

Inspired by the attention mechanism of Transformer [64], we propose the inherent relation layer for learning a case dependent inherent relation. 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:

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

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:

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

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. 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:

$$\boldsymbol{A}'_m = softmax\left(\frac{\left(\boldsymbol{T}_m^{j'} + \boldsymbol{Q}_m\right)\boldsymbol{W}_m^{q'}\left(\left(\boldsymbol{R}_m + \boldsymbol{P}_m\right)\boldsymbol{W}_m^{k'}\right)^T}{\sqrt{d_m}}\right), \tag{3}$$

where $\boldsymbol{T}_m^{j'} \in \mathbb{R}^{N\times d_m}$ 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; $\boldsymbol{W}_m^{q'} \in \mathbb{R}^{N\times d_m}$ and $\boldsymbol{W}_m^{k'} \in \mathbb{R}^{N\times d_m}$ are the weights of FC layers. The output of MCA block can be written as:

$$MCA\left(\boldsymbol{T}^{j'}\right) = \left[\boldsymbol{A}'_1\boldsymbol{T}_1^{j'}\boldsymbol{W}_1^{v'}; ...; \boldsymbol{A}'_M\boldsymbol{T}_M^{j'}\boldsymbol{W}_M^{v'}\right]\boldsymbol{W}'_P, \tag{4}$$

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.

Compared to dividing the feature map in a dense grid manner adopted in other Transformers [42], [58], [59], the dynamic sparse local patches significantly decrease the token number of the MCA blocks, which leads to much lower computational complexity. The computational complexity of the MCA block with the feature in grid manner $\Omega(G)$ and local patch manner $\Omega(S)$ can be calculated as follows:

$$\Omega(S) = 4MNd_m^2 + 2MN^2d_m, \tag{5}$$

$$\Omega(G) = \left(2N + 2\frac{W_FH_F}{P_wP_h}\right)Md_m^2 + 2NM\frac{W_FH_F}{P_wP_h}d_m, \tag{6}$$

where (W_F, H_F) is the size of feature map. $\frac{H_F}{P_h} \times \frac{W_F}{P_w}$ is set to 16×16 as in many works [58], [59] for best performance. For a dataset with 19 landmark annotations, $\Omega(S)$ is only $1/9$ of $\Omega(G)$. Therefore, DSLPT has lower computational complexity than other methods despite the multi-stages for coarse-to-fine locating.

3.1.3 *Prediction Heads*

The heads predict the parameters of a Gaussian or Laplace distribution $(\mu_{x_n}^i, \Sigma_{x_n}^i, \mu_{y_n}^i, \Sigma_{y_n}^i)$ for each landmark, where $\mu_{x_n}^i$ and $\mu_{y_n}^i$ are the distribution mean of n-th landmark in i-th stage on the X axis and Y axis respectively. $\Sigma_{x_n}^i$ and $\Sigma_{y_n}^i$ 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_n^i, H_n^i) is normalized in $[0, 1]$). Therefore, DSLPT does not require positional encoding to retain a global spatial information. The global coordinate (x_n^i, y_n^i) and uncertainty $(U_{x_n}^i, U_{y_n}^i)$ of each landmark can be calculated as follows:

$$\begin{aligned} x_n^i &= x_{\mathrm{lt}_n}^i + W_n^i\mu_{n_x}^i, \\ y_n^i &= y_{\mathrm{lt}_n}^i + H_n^i\mu_{n_y}^i, \end{aligned} \tag{7}$$

$$\begin{aligned} U_{x_n}^i &= W_n^i\Sigma_{x_n}^i, \\ U_{y_n}^i &= H_n^i\Sigma_{y_n}^i, \end{aligned} \tag{8}$$

where $(x_{\mathrm{lt}_n}^i, y_{\mathrm{lt}_n}^i)$ is the global coordinate of the top left point of the n-th patch in the i-th stage, and (W_n^i, H_n^i) is the corresponding patch size.

Algorithm 1 Training pipeline of the 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 $(\mu^g_{x_n})$, the number of stage $N_S$
while the training epoch is less than a specific number do
    Forward $B$ for feature map by $\boldsymbol{F} = B(\boldsymbol{I})$;
    Initialize the local patch size $(W^1_n, H^1_n) \leftarrow (\frac{W}{4}, \frac{H}{4})$
    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
```

3.2 Distribution Estimation

3.2.1 Maximum Log-likelihood Estimation

The L1 loss and L2 loss, which are widely used in face alignment [32], [40], 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:

$$P_\Theta(z|\boldsymbol{I}) = \frac{1}{\Sigma\sqrt{2\pi}}\exp(-\frac{1}{2}\left(\frac{z-\mu}{\Sigma}\right)^2), \tag{9}$$

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:

$$\mathcal{L} = -\log P_\Theta(z|\boldsymbol{I})|_{z=\mu^g} \propto \log\Sigma + \frac{(\mu^g-\mu)^2}{2\Sigma^2}, \tag{10}$$

As mentioned by [19], 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 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.

3.2.2 Dynamic Local Patches

Previous patch based regression methods [6], [34] 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. 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 $\max(6W^i_n\Sigma_x, 6H^i_n\Sigma_y)$. 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: $W^{i+1}_n = H^{i+1}_n = \max(6ZW^i_n\Sigma_x, 6ZH^i_n\Sigma_y)$

Moreover, the patch size of n-th landmark in $(i+1)$-th stage should also be limited in $\left[L_{\text{down}}W^i_n, L_{\text{up}}W^i_n\right]$ and $\left[L_{\text{down}}H^i_n, L_{\text{up}}H^i_n\right]$. 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.

3.3 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 1** 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.

3.4 Auxiliary Inherent Relation Loss

Similar to other face alignment methods [8], [9], 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:

$$\mathcal{L}_t = \sum_{i=1}^{N_S}\sum_{j=1}^{N_I}\sum_{n=1}^{N} \frac{\mathcal{L}\left(\mu^{g_i}_{n_x}, \mu^i_{jn_x} \Sigma^i_{jn_x}\right) + \mathcal{L}\left(\mu^{g_i}_{n_y}, \mu^i_{jn_y}, \Sigma^i_{jn_y}\right)}{2}, \tag{11}$$

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:

$$\begin{aligned} \mu^{g_i}_{n_x} &= \frac{x^g_n - x^i_{\mathrm{lt}_n}}{W^i_n}, \\ \mu^{g_i}_{n_y} &= \frac{y^g_n - y^i_{\mathrm{lt}_n}}{H^i_n}. \end{aligned} \tag{12}$$

4 Experiments

In this section, we evaluate the proposed face alignment method on eight popular benchmarks and carry out extensive experiments to verify the effectiveness. Specifically, we first introduce the eight popular face alignment 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.

4.1 Benchmarks

- **WFLW** [31]: the WFLW is a very challenging face alignment 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** [65]: 300W includes 3,148 faces for training and 689 faces for testing. The faces in the training set come from the fullset of AFW [66] and the training subset of HELEN [67] and LFPW [68]. 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 [65]). Moreover, 300W also annotates additional 600 face images with 68 landmarks to form the 300W-private subset.
- **COFW** [49]: 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 [69]).
- **Menpo** [70] [71]: 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** [72]: AFLW-19 consists of 24,386 faces from AFLW [73], 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** [19]: MERL-RAV re-annotates 19,314 faces from AFLW [73] 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** [21]: Masked 300W synthesizes 689 masked faces from the test set of 300W [65]. The average occluded area in Masked 300W is over 50% of the face area.
- **300W-LP & AFLW2000-3D** [28]: 300W-LP synthesizes 122,450 samples from 300W [65] via face profiling. Each sample is annotated with 68 3D landmarks. AFLW2000-3D selects 2,000 faces from AFLW [73] and each face is also labeled with 68 3D landmarks.

4.2 Evaluation Metrics

Referring to related work [11], [23], [31], 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:

$$NME = \frac{1}{N}\sum_{n=1}^{N} \frac{||\,(x^g_n, y^g_n) - (x_n, y_n)\,||}{d_{\mathrm{norm}}}, \tag{13}$$

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 ($\sqrt{H_{\mathrm{box}} \times W_{\mathrm{box}}}$) or the diagonal of annotated bounding box ($\sqrt{H^2_{\mathrm{box}} + W^2_{\mathrm{box}}}$) 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_α 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_α.

4.3 Implementation Details

The proposed face alignment framework is implemented in Pytorch [74], and we employ four different networks as the backbone: ResNet34 [75], ResNet50 [75], HRNetW18C [7], HRNetW18C-lite (the modularized block number in each stage is set to 1). Each backbone is pre-trained on the ImageNet dataset [76] as the related works [7], [38].

TABLE 1: Performance comparisons with heatmap regression and coordinate regression methods on WFLW full set and its subsets. The normalization factor of NME is inter-ocular distance. Key: [**Best**, **Second Best**, *=initialized from scratch]

Method	Backbone	type	ImageNet pretraining	Flops ↓	Params ↓	Inter-Ocular NME (%) ↓						
						Full	Pose	Exp.	Ill.	Mu.	Occ.	Blur
HRNet [7]	HRNetW18C	heatmap	Y	4.75G	9.66M	4.60	7.94	4.85	4.55	4.29	5.44	5.42
LUVLi [19]	8 DU-Net	heatmap	N	-	-	4.37	7.56	4.77	4.30	4.33	5.29	4.94
AWing [13]	4 Hourglass	heatmap	N	26.8G	24.15M	4.21	7.21	4.46	4.23	4.02	4.99	4.82
HIH [22]	2 Hourglass	heatmap	N	10.38G	14.47M	4.18	7.20	**4.19**	4.45	3.97	5.00	4.81
ADNet [23]	4 Hourglass	heatmap	N	17.04G	13.37M	4.14	6.96	4.38	4.09	4.05	5.06	4.79
SDFL [40]	ResNet34	coordinate	N	-	-	4.55	-	-	-	-	-	-
AV w. SAN [54]	ResNet152	coordinate	Y	33.87G	35.02M	4.39	8.42	4.68	4.24	4.37	5.60	4.86
DETR (R50) [58]	ResNet50	coordinate	Y	10.62G	35.25M	4.32	7.64	4.67	4.24	4.19	5.12	4.90
SDL [38]	HRNetW18C	coordinate	Y	-	-	4.21	7.36	4.49	4.12	4.05	4.98	4.82
SLPT [43]	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 2: Performance comparisons with state-of-the-art methods in $\text{FR}_{0.1}$ and $\text{AUC}_{0.1}$ on WFLW. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C=HRNetW18C, W18C-l=HRNetW18C-lite, *=initialized from scratch]

Metric	Method	Full	Pose	Exp.	Ill.	Mu.	Occ.	Blur
$\text{FR}_{0.1}$(%)↓	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
$\text{AUC}_{0.1}$↑	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

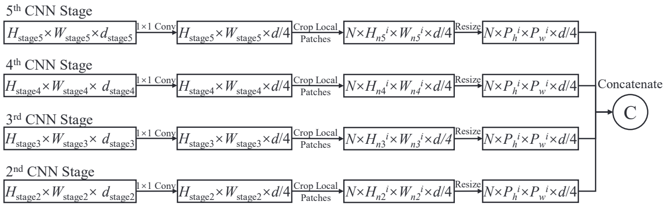

5th CNN Stage
4th CNN Stage
3rd CNN Stage
2nd CNN Stage
$H_{stage5}\times W_{stage5}\times d_{stage5}$
$H_{stage4}\times W_{stage4}\times d_{stage4}$
$H_{stage3}\times W_{stage3}\times d_{stage3}$
$H_{stage2}\times W_{stage2}\times d_{stage2}$
1×1 Conv
$H_{stage5}\times W_{stage5}\times d/4$
$H_{stage4}\times W_{stage4}\times d/4$
$H_{stage3}\times W_{stage3}\times d/4$
$H_{stage2}\times W_{stage2}\times d/4$
Crop Local Patches
$N\times H_{n5}^{i}\times W_{n5}^{i}\times d/4$
$N\times H_{n4}^{i}\times W_{n4}^{i}\times d/4$
$N\times H_{n3}^{i}\times W_{n3}^{i}\times d/4$
$N\times H_{n2}^{i}\times W_{n2}^{i}\times d/4$
Resize
$N\times P_h^{i}\times P_w^{i}\times d/4$
Concatenate
C

Fig. 5: Constructing multi-level feature maps for DSLPT.

For ResNet34 and ResNet50, we employ multi-level feature maps for face alignment, as shown in Fig. 5. 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^1_{nk}, W^1_{nk}) is $\left(\frac{H_{\text{stage}k}}{4}, \frac{W_{\text{stage}k}}{4}\right)$. The patch size of following stages (H^i_{nk}, W^i_{nk}) is calculated using **Algorithm 1**. For HRNetW18C and HRNet18C-lite, we only utilize a single level feature map following the heatmap regression method [7].

We employ AdamW [77] 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.

4.4 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 1 and Table 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 [58]) with 6 encoders and decoders to estimate the probability distribution of each landmark. The token number of the DETR is **16 × 16**. Compared to predicting landmark coordinates from dense patches as DETR, the inherent

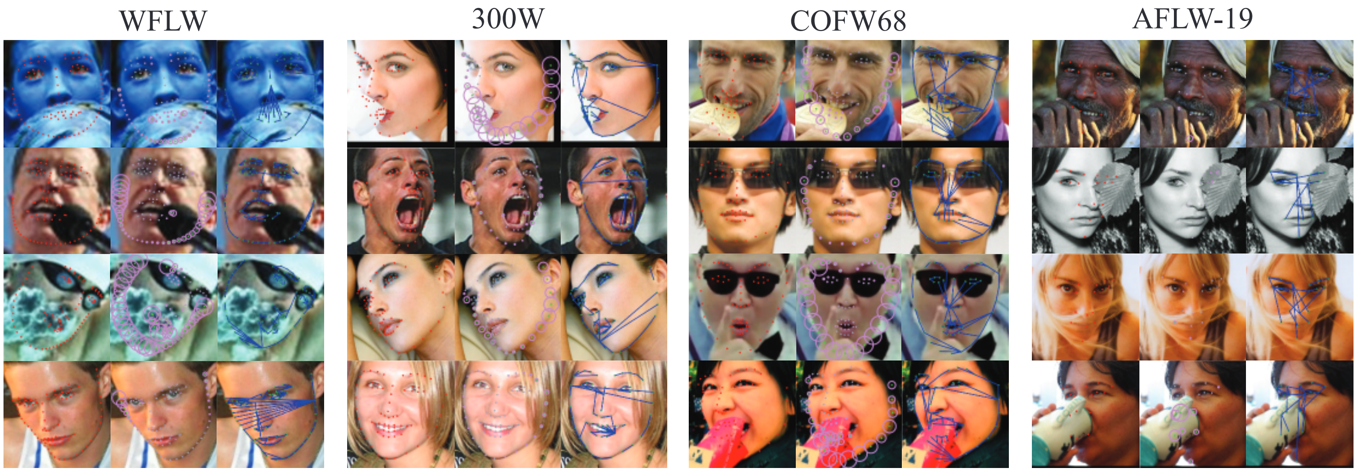


Fig. 6: Visualized results on WFLW, 300W, COFW68, AFLW-19 testset. The **red** point and **green** point indicate the ground truth and the predicted landmark respectively. The uncertainty of each landmark is shown by **pink** circle. Each landmark is connected with the landmark with highest cross attention weigh by **blue** line.

TABLE 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]

Method	Inter-Ocular NME (%) ↓		
	Common	Challenging	Fullset
MHHN [14]	3.18	6.01	3.74
LAB [31]	2.98	5.19	3.49
DeCaFA [12]	2.93	5.26	3.39
HIH [22]	2.93	5.00	3.33
HRNet [7]	2.87	5.15	3.32
SDFL (W18C) [40]	2.88	4.93	3.28
HG-HSLE [16]	2.85	5.03	3.28
DETR (R50) [58]	2.86	4.96	3.27
LUVLi [19]	2.76	5.16	3.23
SHN-GCN [18]	2.73	4.64	3.10
AWing (4HG) [13]	2.72	**4.53**	3.07
SDL [38]	2.62	4.77	3.04
ADNet (R50) [23]	-	-	3.11
ADNet (4HG) [23]	**2.53**	**4.58**	**2.93**
SLPT (W18C-l) [43]	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 4: Performance comparisons under Inter-Ocular normalization and Inter-Pupil normalization on *within*-dataset validation. The threshold for FR is set to 0.1. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C]

Method	Inter-Ocular		Inter-Pupil	
	NME(%)↓	FR(%)↓	NME(%)↓	FR(%)↓
MHHN [14]	4.95	1.78	-	-
LAB [31]	3.92	0.39	-	-
SDFL (W18C) [40]	3.63	**0.00**	-	-
HRNet [7]	3.45	**0.20**	-	-
TCDCN [33]	-	-	8.05	-
SHN-GCN [18]	-	-	5.67	2.36
DETR (R50) [58]	3.79	0.59	5.46	2.37
Wing [32]	-	-	5.44	3.75
DCFE [26]	-	-	5.27	7.29
AWing [13]	-	-	4.94	0.99
ADNet [23]	-	-	**4.68**	**0.59**
SLPT (W18C-l) [43]	**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

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. 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.

300W: as shown in Table 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 results are shown in Table 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

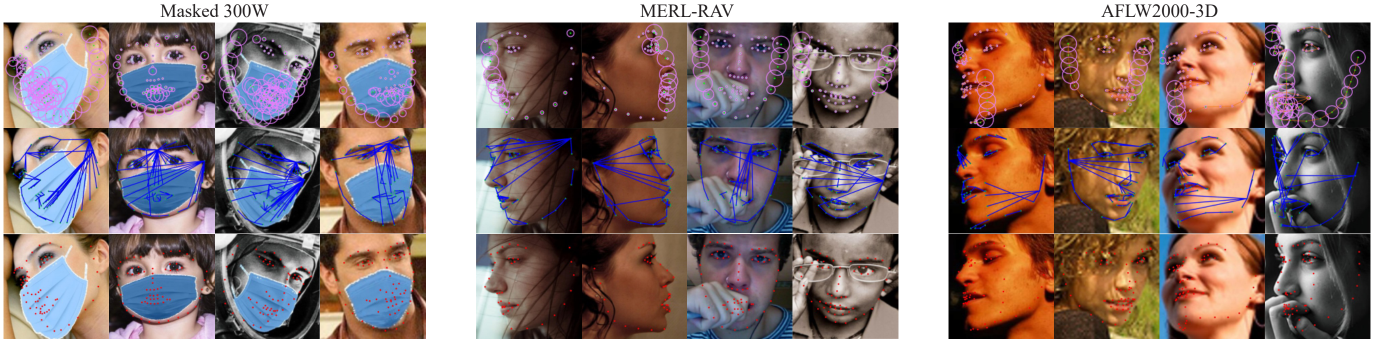


Fig. 7: Visualized results of the externallly occluded and self-occluded cases on Masked 300W, MERL-RAV and AFLW2000-3D. The **red** point and **green** point indicate the ground truth and the predicted landmark respectively. The uncertainty of each landmark is shown by pink circle. Each landmark is connected with the landmark with highest cross attention weigh by **blue** line. The **Orange** lines represent the 3D facial boundaries.

TABLE 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]

Method	$\text{NME}_{\text{box}}(\%)\downarrow$			$\text{AUC}^{0.07}_{\text{box}}(\%)\uparrow$		
	Menpo	300W-p	COFW68	Menpo	300W-p	COFW68
SAN† [35]	2.95	2.86	3.50	61.9	59.7	51.9
2D-FAN† [10]	2.16	2.32	2.95	69.0	66.5	57.5
KDN [15]	2.26	2.49	-	68.2	67.3	-
Softlabel† [15]	2.27	2.32	2.92	67.4	66.6	57.9
KDN† [15]	2.01	2.21	2.73	71.1	68.3	60.1
LUVLI [19]	2.18	2.24	2.75	70.1	68.3	60.8
LUVLI† [19]	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 6: NME_{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]

Method	Unocculded		Externally Occluded	
	$\text{NME}_{\text{box}}\downarrow$	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$	$\text{NME}_{\text{box}}\downarrow$	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$
Softlabel [15]†	2.30	5.99	5.01	7.32
KDN [15]†	2.34	1.63	4.03	11.62
LUVLi [19]†	2.15	9.37	**4.00**	32.49
SLPT (W18C-l) [43]	**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 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	Inter-Ocular NME(%)↓	$\text{FR}_{0.1}$(%)↓
TCDCN [33]	7.66	16.17
CFSS [52]	6.28	9.07
ODN [36]	5.30	-
AV w. SAN [54]	4.43	2.82
LAB [31]	4.62	2.17
SDL [38]	4.22	0.39
SDFL (W18C) [40]	4.18	**0.00**
DETR (R50) [58]	4.15	0.59
SLPT (W18C-l) [43]	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**

learn regressing landmarks from different scales of patches. Besides, a deeper backbone commonly leads to more severe overfitting on this condition. Compared to coordinate 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 [19]. 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. 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 illustrate 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 [28] 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

TABLE 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]

Method	$\mathrm{NME}_{\mathrm{diag}} \downarrow$		$\mathrm{NME}_{\mathrm{box}} \downarrow$	$\mathrm{AUC}^{0.07}_{\mathrm{box}} \uparrow$
	Full	Frontal	Full	Full
DeepReg [27]	2.12%	-	-	-
RND [34]	2.06%	-	-	-
SAN [35]	1.91%	1.85%	-	-
Wing [32]	-	-	3.56%	0.535
KDN [15]	-	-	2.80%	0.603
ODN [36]	1.63%	1.38%	-	-
HRNet [7]	1.57%	1.46%	-	-
LUVLi [19]	1.39%	1.19%	2.28%	0.680
MHHN [14]	1.38%	1.19%	-	-
SHN-GCN [18]	-	-	2.15%	-
LAB [31]	1.25%	1.14%	-	-
DETR (R50) [58]	**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 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]

Method	Inter-Ocular NME (%) ↓		
	Common	Challenging	Fullset
CFSS [52]	11.73	19.98	13.35
Hourglass [8]	8.17	13.52	9.22
MDM [24]	7.66	11.67	8.44
FAN [10]	7.36	10.81	8.02
LAB [31]	6.07	9.59	6.76
SAAT [21]	5.42	11.36	6.58
GlomFace [41]	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 10: Performance comparisons on MERL-RAV. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, W18C-l=HRNetW18C-lite, W18C=HRNetW18C, ⋆=initialized from scratch]

Metrics (%)	Method	Full	Frontal	Half-Profile	Profile
$\mathrm{NME}_{\mathrm{box}} \downarrow$	DU-Net [9]	1.99	1.89	2.50	1.92
	LUVLi [19]	1.61	1.74	1.79	1.25
	SLPT [43]	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**
$\mathrm{AUC}^{0.07}_{\mathrm{box}} \uparrow$	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 11: $\mathrm{NME}_{\mathrm{box}}$ and $|\mathbf{\Sigma}|^{\frac{1}{2}}$ 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]

Method	Self-occluded		Unoccluded		Externally Occluded	
	$\mathrm{NME}_{\mathrm{box}} \downarrow$	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$	$\mathrm{NME}_{\mathrm{box}} \downarrow$	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$	$\mathrm{NME}_{\mathrm{box}} \downarrow$	$\lvert\mathbf{\Sigma}\rvert^{\frac{1}{2}}$
LUVLI [19]	-	-	1.60%	9.28	3.53%	34.41
SLPT [43]	-	-	**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

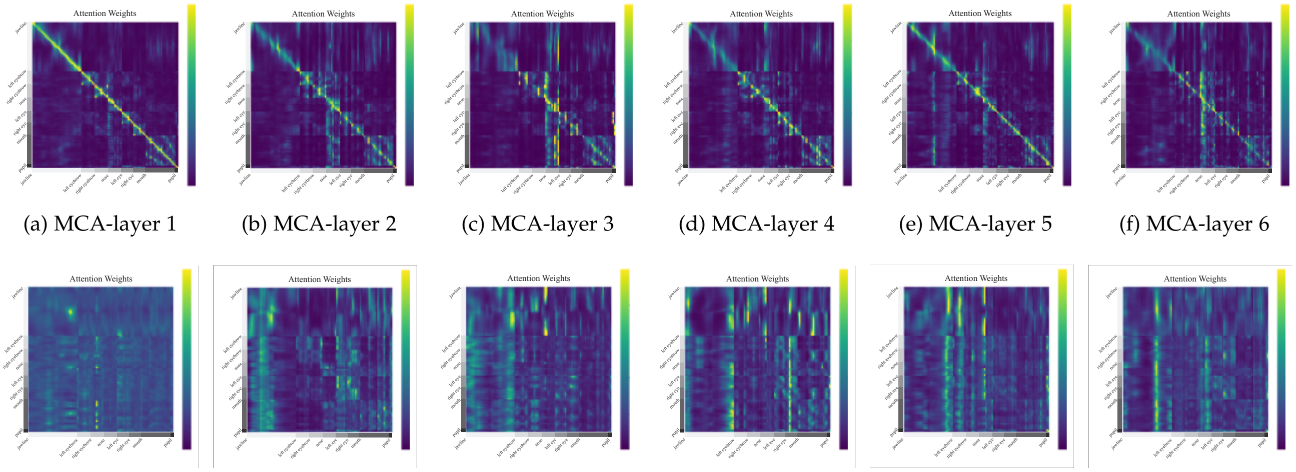
pretraining, we can observe an improvement of 4.54%, 4.48% and 4.81% in $\mathrm{NME}_{\mathrm{box}}$ on Menpo, 300W-private and COFW68 respectively.

To further explore the influence of pretraining and the dynamic patches, we tabulate the $\mathrm{NME}_{\mathrm{box}}$ and square root of the determinant of uncertainty (SQDU) on the externally occluded and unoccluded landmarks of COFW68 in Table 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 unoccluded landmarks predicted by DSLPT is $1/4 \sim 1/3$ of the SQDU on occluded landmarks, as shown in Table 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 $\mathrm{NME}_{\mathrm{box}}$ 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 $\mathrm{NME}_{\mathrm{box}}$ 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 7 to compare DSLPT to other state-of-the-art methods.

AFLW-19: we report the $\mathrm{NME}_{\mathrm{box}}$ (set d_{norm} to the geometric mean of bounding box size) and $\mathrm{NME}_{\mathrm{diag}}$ (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 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.

Masked 300W: following [41], we use the training set of 300W [65] 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 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



(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

Fig. 8: The statistical attention interactions of MCA and MSA in the final stage on the WFLW test set. Each row indicates the attention weight of corresponding landmark to other landmarks.

TABLE 12: Performance comparisons with the state-of-the-art methods on AFLW2000-3D. Key: [**Best**, **Second Best**, R34=ResNet34, R50=ResNet50, *=initialized from scratch]

Method	NME$_{\text{box}}$ (%)↓			
	$[0°, 30°]$	$[30°, 60°]$	$[60°, 90°]$	Mean
SDM [50]	3.67	4.94	9.67	6.12
3DDFA [28]	3.78	4.54	7.93	5.42
3DDFA+SDM [28]	3.43	4.24	7.17	4.94
3D-FAN [10]	3.15	3.53	4.60	3.76
3DDFA-TPAMI [29]	2.84	3.57	4.96	3.79
3DDFAV2 (MR) [37]	2.75	3.49	4.53	3.59
3DDFAV2 (MRS) [37]	2.63	3.42	4.48	3.51
SynergyNet [39]	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**

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. 7.

MERL-RAV: as shown in Table 10, DSLPT improves the metric by 8.07% and 25.63% in NME$_{\text{box}}$ over LUVLi and DU-Net respectively. Some cases with external occlusion and self-occlusion are demonstrated in Fig.7. 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$_{\text{box}}$ and SQDU on three types of landmarks of MERL in Table 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 [19]. As a result, the adaptive receptive field brings an improvement of 2.38% and 8.38% in NME$_{\text{box}}$ on the externally occluded landmarks compared to SLPT and LUVLi.

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 [28], we use the 300W-LP samples synthesized from the training set of LFPW, HELEN and the whole AFW for training. As shown in Table 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.7, 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 [28] 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.

4.5 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.

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 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, 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,

TABLE 13: The influence of coarse-to-fine framework when using different number of stages. Key: [**Best**, W18C-l=HRNetW18C-lite]

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

Method	MSA	MCA	NME ↓	FR ↓	AUC ↑
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 15: Influence of different kinds of encodings. Key: [**Best**, W18C-l=HRNetW18C-lite]

Method	encoding	NME↓	FR↓	AUC↑
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 16: Computational complexity, parameters and performance of the DSLPT with different inherent relation layer number on WFLW testset. Key: [**Best**, W18C=HRNetW18C]

Method	Layer number	Flops	Params	NME↓	FR↓	AUC↑
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**

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.

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 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. 8. 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 [64]. Their performance on WFLW testset is reported in Table 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 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 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 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** [43]. 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 [19], both the Laplace and Gaussian negative log-likelihood function demonstrate comparable performance in DSLPT. The main reason is that

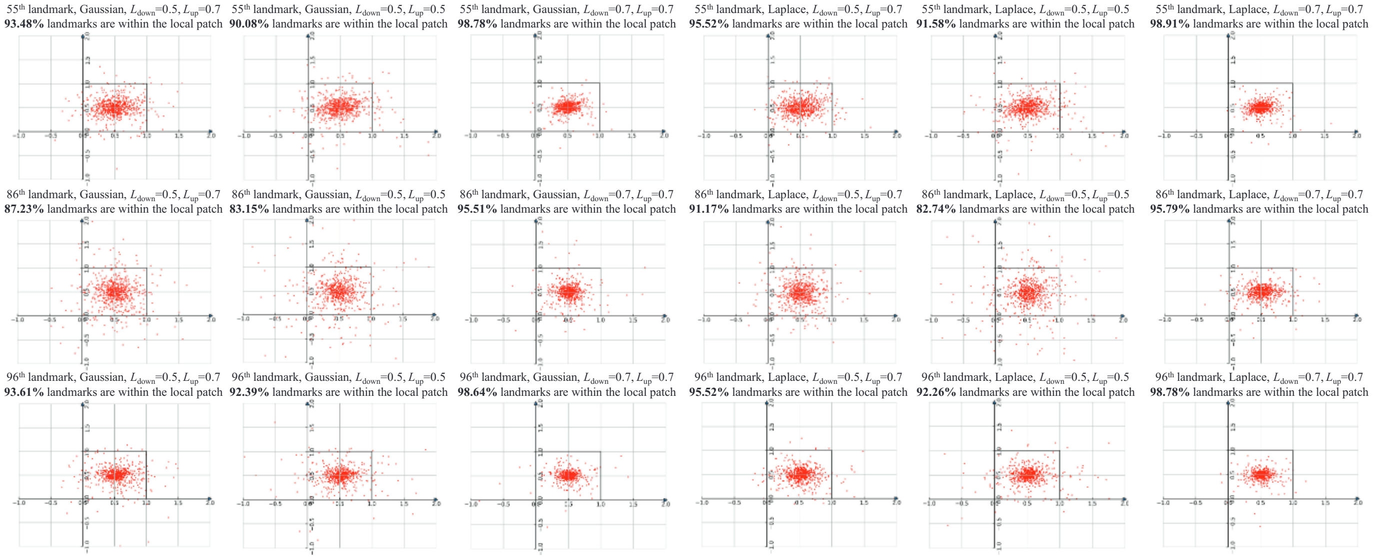


Fig. 9: The distribution of 55^{th}, 86^{th} and 96^{th} landmark in patch coordinate system on the occlusion subset of WFLW. The percentage of the landmarks that are within the local patch is also reported in captions.

TABLE 17: Performance of the DSLPT with different loss function on WFLW testset. Each model is with HRNetW18C-lite as the backbone. If L_{down} and L_{up} are a same value, the patch sizes of all landmarks are a constant number. Key: [**Best**, Gaussian=Gaussian negative log-likelihood function, Laplace=Laplace negative log-likelihood function]

Loss function	L_{down}	L_{up}	NME↓	FR↓	AUC↑
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

[19] 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.9. 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 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.

TABLE 18: Performance of the DSLPT with different L_{down} and L_{up} on WFLW testset. Key: [**Best**]

	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

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 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-

TABLE 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 $\sqrt{W_{\text{box}} \times H_{\text{box}}}$.

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 [58] (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	-	-	-	-	-	-

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

5 Conclusion

In this paper, we propose the Dynamic Sparse Local Patch Transformer to address two main issues in face alignment: 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 three problems in face alignment and outperforms other methods with much less computational complexity.

References

[1] A. B. Tanfous, H. Drira, and B. B. Amor, "Sparse coding of shape trajectories for facial expression and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2594–2607, 2020.

[2] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 848–863, 2022.

[3] F. Liu, Q. Zhao, X. Liu, and D. Zeng, "Joint face alignment and 3d face reconstruction with application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 664–678, 2020.

[4] Y. Liu, H. Shi, H. Shen, Y. Si, X. Wang, and T. Mei, "A new dataset and boundary-attention semantic segmentation for face parsing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 11 637–11 644.

[5] J. Zhang, X. Zeng, M. Wang, Y. Pan, L. Liu, Y. Liu, Y. Ding, and C. Fan, "Freenet: Multi-identity face reenactment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020.

[6] A. Zadeh, T. Baltrušaitis, and L.-P. Morency, "Convolutional experts constrained local model for facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 2051–2059.

[7] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2021.

[8] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

[9] Z. Tang, X. Peng, K. Li, and D. N. Metaxas, "Towards efficient u-nets: A coupled and quantized approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2038–2050, 2020.

[10] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.

[11] X. Dong, Y. Yang, S.-E. Wei, X. Weng, Y. Sheikh, and S.-I. Yu, "Supervision by registration and triangulation for landmark detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3681–3694, 2021.

[12] A. Dapogny, M. Cord, and K. Bailly, "Decafa: Deep convolutional cascade for face alignment in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6892–6900.

[13] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6970–6980.

[14] J. Wan, Z. Lai, J. Liu, J. Zhou, and C. Gao, "Robust face alignment by multi-order high-precision hourglass network," *IEEE Trans. Image Process.*, vol. 30, pp. 121–133, 2021.

[15] L. Chen, H. Su, and Q. Ji, "Face alignment with kernel density deep neural network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6991–7001.

[16] X. Zou, S. Zhong, L. Yan, X. Zhao, J. Zhou, and Y. Wu, "Learning robust facial landmark detection via hierarchical structured ensemble," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 141–150.

[17] Y. Tai, Y. Liang, X. Liu, L. Duan, J. Li, C. Wang, F. Huang, and Y. Chen, "Towards highly accurate and stable face alignment for high-resolution videos," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 8893–8900.

[18] J. Zhang, H. Hu, and S. Feng, "Robust facial landmark detection via heatmap-offset regression," *IEEE Trans. Image Process.*, vol. 29, pp. 5050–5064, 2020.

[19] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng, "Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8233–8243.

[20] B. Browatzki and C. Wallraven, "3fabrec: Fast few-shot face alignment by reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6109–6119.

[21] C. Zhu, X. Li, J. Li, and S. Dai, "Improving robustness of facial landmark detection by defending against adversarial attacks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 11 731–11 740.

[22] X. Lan, Q. Hu, and J. Cheng, "Revisting quantization error in face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2021, pp. 1521–1530.

[23] Y. Huang, H. Yang, C. Li, J. Kim, and F. Wei, "Adnet: Leveraging error-bias towards normal direction in face alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3060–3070.

[24] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4177–4187.

[25] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Proc. Eur. Conf. Comput. Vis.*, Cham, 2016, pp. 57–72.

[26] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, "A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment," in *Proc. Eur. Conf. Comput. Vis.*, Cham, 2018, pp. 609–624.

[27] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3691–3700.

[28] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 146–155.

[29] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 78–92, 2019.

[30] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 2034–2043.

[31] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2129–2138.

[32] Z. Feng, J. Kittler, M. Awais, P. Huber, and X. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2235–2245.

[33] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3067–3074, 2018.

[34] H. Liu, J. Lu, M. Guo, S. Wu, and J. Zhou, "Learning reasoning-decision networks for robust face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 679–693, 2020.

[35] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 379–388.

[36] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3481–3491.

[37] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 152–168.

[38] W. Li, Y. Lu, K. Zheng, H. Liao, C. Lin, J. Luo, C.-T. Cheng, J. Xiao, L. Lu, C.-F. Kuo, and S. Miao, "Structured landmark detection via topology-adapting deep graph learning," in *Proc. Eur. Conf. Comput. Vis.* Cham: Springer International Publishing, 2020, pp. 266–283.

[39] C.-Y. Wu, Q. Xu, and U. Neumann, "Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry," in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 453–463.

[40] C. Lin, B. Zhu, Q. Wang, R. Liao, C. Qian, J. Lu, and J. Zhou, "Structure-coherent deep feature learning for robust face alignment," *IEEE Trans. Image Process.*, vol. 30, pp. 5313–5326, 2021.

[41] C. Zhu, X. Wan, S. Xie, X. Li, and Y. Gu, "Occlusion-robust face alignment using a viewpoint-invariant hierarchical network architecture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11 102–11 111.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[43] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4052–4061.

[44] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, 2001.

[45] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[46] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Proc. British Mach. Vis. Conf.*, 2006, pp. 95.1–95.10, doi:10.5244/C.20.95.

[47] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 937–944.

[48] X. Liu, "Discriminative face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1941–1954, 2009.

[49] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1513–1520.

[50] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.

[51] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3681–3690.

[52] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4998–5006.

[53] Z. Feng, J. Kittler, W. Christmas, P. Huber, and X. Wu, "Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3681–3690.

[54] S. Qian, K. Sun, W. Wu, C. Qian, and J. Jia, "Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10 152–10 162.

[55] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, 2003.

[56] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE Int. Conf. Comput. Vis.*, October 2021, pp. 568–578.

[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, October 2021, pp. 10 012–10 022.

[58] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 213–229.

[59] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable {detr}: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021.

[60] 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 *Proc. Advances Neural Inf. Process. Syst.*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, 2021, pp. 12 077–12 090.

[61] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou, "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 11 293–11 302.

[62] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. CHEN, and J. Wang, "HRFormer: High-resolution vision transformer for dense predict," in *Proc. Advances Neural Inf. Process. Syst.*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[63] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, "General facial representation learning in a visual-linguistic manner," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2022, pp. 18 697–18 709.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2017, p. 6000–6010.

[65] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 397–403.

[66] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879–2886.

[67] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 679–692.

[68] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 545–552.

[69] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1899–1906.

[70] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou, "The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking," *Int. J. Comput. Vis.*, vol. 127, pp. 599–624, 2019.

[71] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 2116–2125.

[72] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3409–3417.

[73] 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 *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2144–2151.

[74] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Advances Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, 2019.

[75] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[76] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[77] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.

Jiahao Xia received the B.Eng. degree in automotive engineering from the School of Automobile Engineering, Wuhan University of Technology, Wuhan, China, in 2017, and the M.Eng. degree in mechanical engineering from the College of Mechanical and Vehicle Engineering, Hunan University, Changsha, China, in 2020. He is currently working toward the Ph.D. degree in electrical and data engineering with the Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW, Australia. His current research interests include vision transformer, unsupervised learning, and graph neural network.

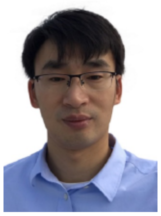
Min Xu (Member, IEEE) received the B.E. degree from the University of Science and Technology of China, Hefei, China, the M.S. degree from the National University of Singapore, Singapore, and the Ph.D. degree from the University of Newcastle, Callaghan, NSW, Australia. She is an Associate Professor with the School of Electrical and Data Engineering (SEDE), Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS), and also the Leader of Visual and Aural Intelligence Laboratory with the Global Big Data Technologies Center (GBDTC), UTS. She has published 170+ research papers in prestigious international journals and conference proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (T-PAMI), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (T-NNLS), IEEE TRANSACTIONS ON MULTIMEDIA (T-MM), IEEE TRANSACTIONS ON MOBILE COMPUTING (T-MC), PR, ICLR, ICML, CVPR, ICCV, ACM MM, AAAI. Her research interests include multimedia, computer vision, and machine learning. Dr. Xu is an Editorial Board Member for Elsevier Journal of Neurocomputing and served as a program chair/area chair for many major conferences.

Haimin Zhang is a Postdoctoral Research Fellow in the School of Electrical and Data Engineering, University of Technology Sydney. He received the Bachelor's degree from Zhejiang Sci-Tech University, Hangzhou, China, the Master's degree from Nankai University, Tianjin, China, and the Ph.D. degree from University of Technology Sydney, Ultimo, NSW, Australia. His research interests include pattern recognition, machine learning, computer vision and multimedia analytics. He serves as a reviewer for many highly rated international conferences and journals, such as IEEE Transactions on Multimedia and Pattern Recognition.

Jianguo Zhang is currently a Professor in Department of Computer Science and Engineering, Southern University of Science and Technology. Previously, he was a Reader in Computing, School of Science and Engineering, University of Dundee, UK. He received a PhD in National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 2002. His research interests include object recognition, medical image analysis, machine learning and computer vision. He is a senior member of the IEEE and serves as an Associate Editor of IEEE Trans on Multimedia.

Wenjian Huang received his Ph.D. degree in mechanics from Peking University, China, in 2020. He is currently working as a Postdoc researcher in the Department of Computer Science and Engineering at the Southern University of Science and Technology, China. His research interests include computer vision, biomedical image analysis, statistical learning, and deep learning.

Hu Cao received his M.Eng. degree in vehicle engineering from HuNan University, China, in 2019. He is currently working toward a Ph.D. degree in Computer Science as a member of the Informatics-6, Chair of Robotics, Artificial Intelligence, and Real-time Systems at Technische Universität München, München, Germany. His research interests include computer vision, neuromorphic engineering, robotics, and deep learning.



Shiping Wen received the M.Eng. degree in Control Science and Engineering, from School of Automation, Wuhan University of Technology, Wuhan, China, in 2010, and received the Ph.D degree in Control Science and Engineering, from School of Automation, Huazhong University of Science and Technology, Wuhan, China, in 2013. He is currently a Professor with the Australian Artificial Intelligence Institute (AAII) at University of Technology Sydney. His research interests include neural network, deep learning, computer vision, and their applications in medical informatics et al. He is a Fellow of Institute of Physics (IOP) and a Fellow of British Computer Society (BCS). He was listed as a Highly Cited Researcher by Clarivate Analytics in 2018 and 2020, respectively. He received the 2017 Young Investigator Award of Asian Pacific Neural Network Association and 2015 Chinese Association of Artificial Intelligence Outstanding PhD Dissertation Award. He currently serves as Associate Editor for Knowledge-Based Systems, Engineering Applications of AI, Neural Processing Letters, et al., and Leading Guest Editor for IEEE Transactions on Network Science and Engineering, Sustainable Cities and Society, et al.