Structure-Coherent Deep Feature Learning for Robust Face Alignment

Chunze Lin*, Beier Zhu*, Quan Wang, Renjie Liao, Chen Qian, Jiwen Lu, *Senior Member, IEEE,* and Jie Zhou, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a structure-coherent deep feature learning method for face alignment. Unlike most existing face alignment methods which overlook the facial structure cues, we explicitly exploit the relation among facial landmarks to make the detector robust to hard cases such as occlusion and large pose. Specifically, we leverage a landmark-graph relational network to enforce the structural relationships among landmarks. We consider the facial landmarks as structural graph nodes and carefully design the neighborhood to passing features among the most related nodes. Our method dynamically adapts the weights of node neighborhood to eliminate distracted information from noisy nodes, such as occluded landmark point. Moreover, different from most previous works which only tend to penalize the landmarks absolute position during the training, we propose a relative location loss to enhance the information of relative location of landmarks. This relative location supervision further regularizes the facial structure. Our approach considers the interactions among facial landmarks and can be easily implemented on top of any convolutional backbone to boost the performance. Extensive experiments on three popular benchmarks, including WFLW, COFW and 300W, demonstrate the effectiveness of the proposed method. In particular, due to explicit structure modeling, our approach is especially robust to challenging cases resulting in impressive low failure rate on COFW and WFLW datasets. The model and code are publicly available at https://github.com/BeierZhu/Sturcture-Coherency-Face-Alignment**

I. Introduction

Face alignment, also known as facial landmark detection is an important topic in computer vision and has attracted much attention over past few years [1], [2], [3], [4], [5], [6], [7]. As a fundamental step for face image analysis, face alignment plays a key role in many face applications such as face recognition [8], expression analysis [9] and face editing [10].

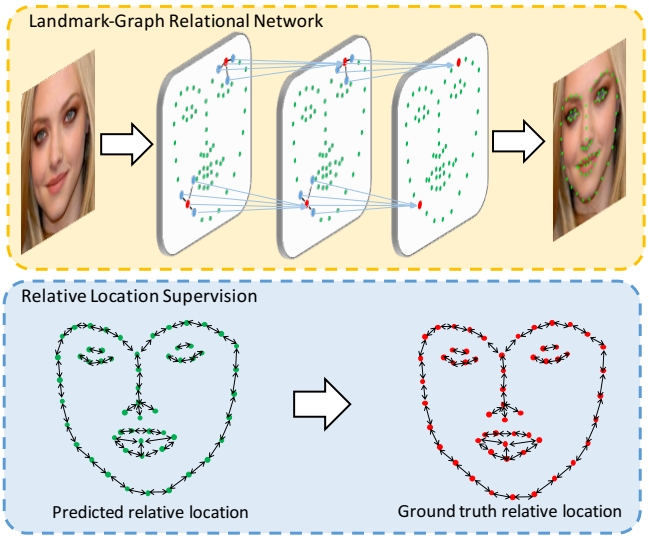


Fig. 1. The proposed structure-coherent deep feature learning method leverages a landmark-graph relational network which provides graph-based inferences among facial landmarks, exploiting facial key points relations to constrain landmarks. Most related landmarks are grouped and convolved together through graph convolutional layers to infer the facial landmarks. In addition, unlike most existing deep learning based methods which focus on penalizing the model to minimize the absolute location of landmarks, we propose a relative location loss to further enhance the facial structure coherency. With this relative location supervision signal, the model is also constrained to minimize the relative location errors of the predicted landmarks.

Although significant progress has been made, face alignment is still a challenging problem due to issues like occlusion, large head pose and complicated expression.

With the success of deep learning in several computer vision tasks such as image classification and object detection, many convolutional neural networks (CNN) based face alignment methods have been proposed. Existing CNN-based face alignment methods can mainly be divided into two categories: heatmap regression based ones [11], [3], [12] and coordinate regression based [13], [2], [1]. Heatmap regression based methods commonly produce higher precise localization for its translation equivariant property [14]. As keeping the high spatial resolution of feature maps and heatmap is essential for high accuracy, heatmap regression based methods commonly utilize stacked hourglass shape networks [15]. However, it leads to computationally heavy models which are impractical for deployment in real-world applications. Coordinate regression based methods are relatively simpler and can be built on lighter convolutional networks. Therefore, in this work,

* means equal contribution.

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, and Grant 61672306, in part by a grant from the Institute for Guo Qiang, Tsinghua University, and in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, in part by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-01-002)

Chunze Lin, Beier Zhu, Quan Wang and Chen Qian are with the Sensetime Research, Beijing, 100084, China. Email: linchunze@sensetime.com; zhubeier@sensetime.com; wangquan@sensetime.com; qianchen@sensetime.com.

Renjie Liao is with the Department of Computer Science, Toronto University, Toronto, Canada. Email: rjliao@cs.toronto.edu

Jiwen Lu and Jie Zhou are with the Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems, and Beijing National Research Center for Information Science and Technology, Beijing, 100084, China. Email: lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn.

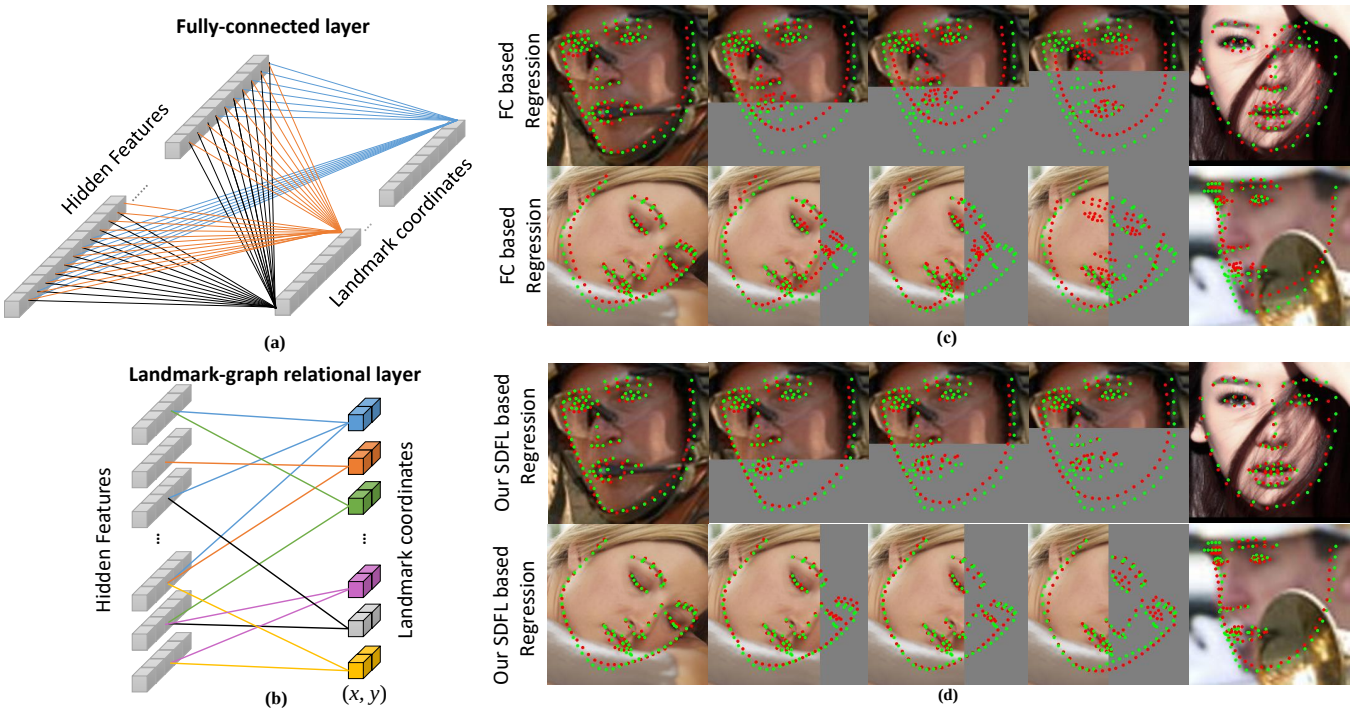


Fig. 2. Comparison between the fully-connected layer and our landmark-graph relational layer. (a) Dense connection in the fully-connected layer where each landmark is correlated with all others. (b) Sparse and relation-aware graph convolutional layer where information propagate only among most related landmarks. (c, d) The performance of fully-connected (FC) based and our structure-coherent deep feature learning (SDFL) based methods under different levels of occlusions. Green and red points correspond to ground-truth and prediction, respectively. The FC based method fails to predict the landmarks due to occlusions, even the visible parts are incorrectly localized due to interference of occluded landmarks. While our SDFL successfully localizes the visible parts and infer the occluded landmarks correctly.

we will focus our attention on coordinate regression methods which result in a better speed/accuracy trade-off.

The fully-connected layers (FC) are commonly used in such methods to convert convolutional feature maps to facial landmark coordinates [13], [2], [1]. However, the dense connections of fully-connected layers make every landmark correlate to each other. As shown in Fig. 2 (a), in the FC layer, every landmark coordinate is connected to the same hidden features. The error of one landmark leads to error of all other landmarks, especially in hard cases such as occlusion. As shown in Fig. 2 (c), when we progressively occlude human face, the error of face contour leads to the error of other parts of human face. Therefore, the fully-connected architecture is inefficient which incorporates redundant and noisy information from occluded landmarks. This raises a natural question: How to effectively leverage the information among landmarks for better inference?

In this paper, we propose a structure-coherent deep feature learning (SDFL) method for robust face alignment by explicitly exploring the relation among facial landmarks. Since human face has a regular structure, coherence among different facial parts provides important cues for effectively localizing facial landmarks, which helps keep the structure of face and infer occluded landmarks. With the help of deep geometric learning, we treat the features of each landmark as a node, and leverage a graph relational network to propagate features among the neighboring nodes. Illustrations of graph relational framework are depicted on the top of Fig. 1 and Fig. 2 (b). The sparse graph structure endows the model with the capability of using the facial structure coherence appropriately. The sparse graph structure is learnt by data-driven based neighborhood construction and dynamic weight adjustment. Our model can adaptively update the weight of neighborhood to highlight landmarks with high confidence while inhibit the information propagation from noisy landmarks. Fig. 2 (d) shows that reasoning with structure coherence cues bolster our model to correctly localize the key points in challenging real-world situations such as occlusion and large pose. With extreme occlusion situations, our SDFL correctly localizes visible facial key points and infer the occluded ones. To be more specific, our SDFL consists of three parts: node embedding module, dynamic adjacency matrix weighting module and graph relational network. The node embedding module converts the convolutional features into graph node representations and the relation is learnt via dynamic adjacency matrix weighting module, based on which, the graph relational network effectively regresses the coordinate of facial landmarks.

Moreover, we propose a relative location loss function to provide a supervision signal on relative position of landmarks, see bottom of Fig. 1. Unlike most existing methods which only utilize absolute position of landmarks as supervision, our relative location loss acts as a regularizer to penalize infeasible local landmark shape and make the model infer correct facial landmarks.

We evaluate the proposed method on three widely-used face alignment benchmarks including WFLW [1], COFW [16] and

300W [17]. Experimental results demonstrate the effectiveness of our approach, which outperforms existing state-of-the-art regression based methods by a large margin. We conduct extensive ablation studies to show the effectiveness of each proposed module and design. In addition, our SDFL can be easily implemented on top of different convolutional backbones. Comparing to the FC layers, when using our landmark-graph relational network as predictor head, we observe consistent improvement across different backbones including MobilenetV2 [18], EfficientNet [19], ResNet [20], Res2Net [21], VGG16 [22] and HRNet [12].

In summary, the main contributions of this paper are as follows:

1) We propose a sparse dynamic graph relational network to explicitly model the interaction among most related landmarks, making the detector more robust to occlusion, large pose and expression issues.
2) We introduce a relative location loss function to consider the relative position of landmarks. This supervision signal allows the network to further pay attention to the facial structure.
3) We conduct extensive experiments on three challenging face alignment benchmarks and achieve very competitive and state-of-the-art performance. Furthermore, we perform comprehensive ablation study to analyze the contribution of main components and examine the effect of sparse and dense interaction among landmarks.

II. Related Work

In this section, we briefly review four related topics: conventional landmarks regression, deep learning based coordinate and heatmap regression based face alignment, and graph neural network.

A. Conventional Methods

Conventional facial landmark detection models mainly fall into two categories, *i.e.*, fitting models and Constrained Local Models (CLMs). Taylor *et al.* introduced the active appearance model [23][24] to fit the facial images with a small number of coefficients, controlling both the facial appearance and the facial shape. CLMs [25][26] predict the landmarks based on the global facial shape constraints as well as the independent local appearance information around each landmark. Locating facial landmarks with graph structure is related to some previous works [27] [28], which apply deformable part models [29] to face analysis. These methods belong to probabilistic graphical models, which entail hand-crafted potential functions and iterative optimization for inference. However, our method is deep learning based graph network, which generates richer and more expressive feature embeddings and enjoys the faster inference.

B. CNN based Coordinate Regression

Coordinate regression models directly map the face image to the landmark coordinates. Zhang *et al.* [30] improved the robustness of detection through multi-task learning, *i.e.*, learning landmark coordinates and predicting facial attributes at the same time. MDM [13] utilizes the recurrent neural network for end-to-end model training, and locates landmarks from coarse to fine. Feng *et al.* [2] introduced a modified log loss, named Wing loss, to increase the contribution of small and medium errors to the training process. LAB [1] regresses facial landmark coordinates with the help of boundary information to reduce the annotation ambiguities. In spite of the advantage of explicit inference of landmark coordinates without any post-processing, the coordinate regression models generally underperform heatmap regression models in terms of accuracy. Unlike these methods which overlook the relation between landmarks, we propose landmark-graph relational network and relative location loss function to take the interactions among facial key points into consideration.

C. CNN based Heatmap Regression

Heatmap regression models consider facial landmark regression task as a heatmap regression problem. These methods generally leverage fully convolutional networks (FCNs) to transform the input image into heatmap which highlights facial key points. In recent work, stacked hourglass (HG) [15] is widely used to achieve the state-of-the-art performance. Yang *et al.* [11] first normalized faces with a supervised transform and then predicted heatmap using a HG. Dapogny *et al.* [3] proposed an end-to-end deep convolutional cascade architecture to improve the localization accuracy of HGs. Liu *et al.* [31] developed a latent variable optimization strategy to reduce the impact of ambiguous annotations when training a 4-stacked HG. In addition to HG, architecture such as HRNet [12] is also able to yield excellent performance. Chandran *et al.* [32] presented a fully convolutional attention driven regional architecture for predicting landmarks on very high resolution facial images without downsampling. Supervision by Registration and Triangulation (SRT) [33] is an unsupervised approach that utilizes unlabeled multi-view video to improve the accuracy and precision of landmark detectors. Dong *et al.* proposed Teacher Supervises StudentS (TS3) [34] which is an interaction mechanism between one teacher network and two student networks, to explore unlabeled data. The student detection networks tend to generate pseudo labels for unlabeled images, while the teacher network learns to judge the quality of the generated pseudo labels. Despite their higher accuracy, heatmap regression models are much more costly from a computational point of view compared to coordinate regression models. The expensive computation of the headmap based approaches is an obstacle for the deployment of such methods in real-time facial analysis systems. In this work, we focus on efficient coordinate regression-based methods.

D. Graph Neural Networks

Graph Neural Networks (GNNs) are a class of models which try to generalize deep learning to handle graph-structured data. They are first introduced in [35] and become more and more popular recently [36]. There are mainly two types of GNNs: Message Passing based Neural Networks [35], [37], [38] and Graph Convolution based Neural Networks [39], [40], [41], [42]. Many recent works have shown that GNNs are

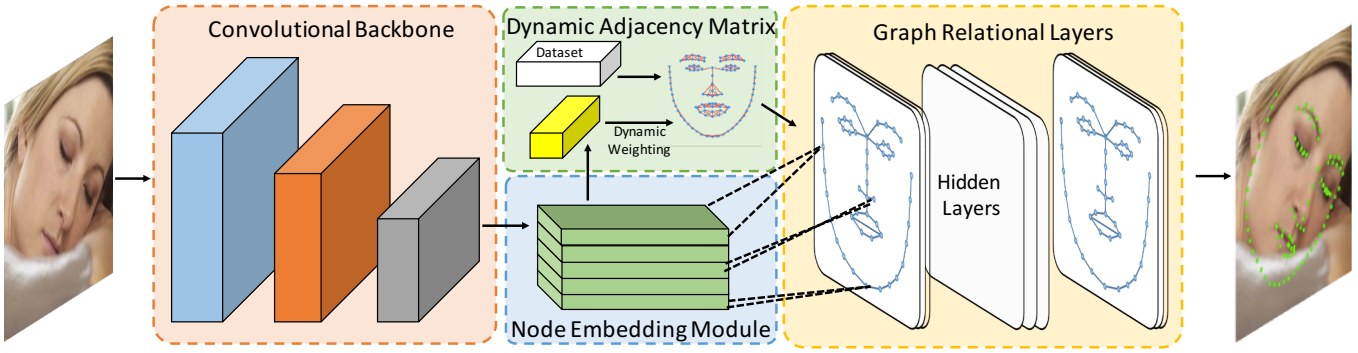


Fig. 3. An overview of the proposed method. The convolutional backbone computes feature maps from the input face image. The node embedding module then maps the convolutional features into graph node representations. Together with the dynamic graph adjacency matrix that learned from the dataset and features extracted from the node embedding module, they are fed into the graph relational layers to infer the facial landmarks. The graph relational layers permit to interact and propagate information among most related facial key points.

very effective in many computer vision tasks, *e.g.*, RGBD semantic segmentation [43], action recognition [44], scene graph generation and reasoning [45], [46], image annotation [47], object detection [48] and 3D shape analysis [49]. Specifically, in this work, we closely follow the so-called graph convolutional network (GCN) [40] which greatly simplifies the graph convolution operator by exploiting approximation to the Chebyshev polynomial based graph spectral filters. It provides a simple yet effective way to integrate local neighboring node feature following the graph topology.

III. Structure-coherent Deep Feature Learning

As the relative spatial relationship of facial landmarks is stable, it is desirable to capture and exploit such important cues for accurate localization. For that, we propose a structure-coherent deep feature learning (SDFL) method to enforce the detected landmarks be correct and coherent. Specifically, we leverage a landmark-graph relational network by considering each landmark as a node and exploring their relation. As illustrated in Fig. 3, our SDFL is mainly composed of four key parts: a convolutional backbone, a node embedding module, a dynamic adjacency matrix weighting module and graph relational layers. The convolutional backbone computes and extracts feature maps from an input image. The node embedding module maps the convolutional features into graph node features, and then a sparse graph structure is learnt by dynamic adjacency matrix weighting module. Finally, the node features and the sparse graph structure are fed into the graph relational layers to output the coordinates of facial landmarks. We propose a relative location loss and a soft wing loss to supervise the learning. The former one tends to regularize the local structure of facial landmarks while the latter one makes the model focus on medium errors of absolute location.

A. Node Embedding

Given the input image, a convolutional backbone first extracts the convolutional features maps from this image. However, a graph convolution network cannot directly take these feature maps as input. In order to train our model in an end-to-end manner, we design the map-to-node module to seamlessly map convolutional feature maps to graph node representations. Specifically, the input convolutional feature maps $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ (where C, H and W denote number of channels, height and width of the feature maps) are first transformed to the hidden feature maps by non-linear function $\boldsymbol{H} = \phi(\boldsymbol{F}) \in \mathbb{R}^{Nn\times H\times W}$, where $n \in \mathbb{Z}^+$ is the expansion coefficient and N is the number of landmarks. In this paper, we consider two convolution-BN-ReLU blocks with $n = 4$ as the non-linear function $\phi(\cdot)$. $\boldsymbol{H}$ is then reshaped to $\boldsymbol{H}^0 \in \mathbb{R}^{N\times nHW}$ to represent the input node features.

B. Sparse Graph Construction

Landmark-graph relational network propagates information among nodes based on a adjacency matrix which determines the edge between nodes. In the face alignment task, due to the lack of pre-defined adjacency matrix for facial landmarks, we build it through a two-step process: we first determine the neighborhood statically for each dataset, then dynamically adjust the weights of the adjacency matrix during inference.

Neighborhood Construction: Prior to training or inference, we build the neighborhood in a data-driven way, *i.e.*, treating each landmark as a node and mining the correlation among landmarks within the dataset. Specifically, we assemble the landmark coordinates of the training set into a rank-three data tensor $\boldsymbol{T} \in \mathbb{R}^{M\times N\times 2}$ where M is the number of images, and the last dimension represents the (x, y) coordinates. We then slice the tensor $\boldsymbol{T}$ along the last dimension to generate $\boldsymbol{T}_x$ and $\boldsymbol{T}_y$. Based on $\boldsymbol{T}_x \in \mathbb{R}^{M\times N}$ and $\boldsymbol{T}_y \in \mathbb{R}^{M\times N}$, we calculate Pearson's correlation coefficient in x and y direction respectively to form correlation matrices $\boldsymbol{C}_x \in \mathbb{R}^{N\times N}$ and $\boldsymbol{C}_y \in \mathbb{R}^{N\times N}$. Then, the correlation between nodes is defined as

$$\boldsymbol{C} = \frac{1}{2}(\text{abs}(\boldsymbol{C}_x) + \text{abs}(\boldsymbol{C}_y)), \tag{1}$$

where $\text{abs}(\cdot)$ returns element-wise absolute value of matrix. Considering the computation cost and noisy edges, we only retain the top $k + 1$ largest value of each row of $\boldsymbol{C}$ to form a sparse binary matrix $\boldsymbol{M} \in \mathbb{R}^{N\times N}$. In other words, most

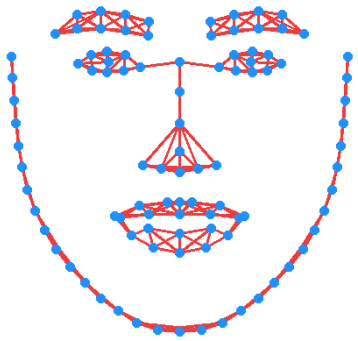


Fig. 4. Example of 4-neighborhood adjacency matrix of our landmark-graph relational network, computed from the WFLW dataset with 98 landmarks.

k relevant landmarks are picked as the neighborhood of each landmark. The binary matrix with self-loops can be written as:

$$\boldsymbol{M}_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{C}_{ij} \in \text{Top}^{k+1}_{t=1,\ldots,N}(\boldsymbol{C}_{it}) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

An example of 4-neighborhood adjacency matrix computed with 98 landmarks is shown in Fig. 4. We can see that according to this data-driven strategy, each landmark-node is correlated with its closest 4 neighborhood points.

Dynamic Adjacency Matrix Weighting: The neighborhood is constructed based on the geometric structure of facial landmarks, while learning relationship among landmarks for each face aims to take the facial appearance factors like occlusion and head pose into consideration. Given the binary matrix $\boldsymbol{M}$ which determines the node neighborhood, we seek to adaptively adjust its weights, so that we can reduce the influence of noisy nodes and augment the information from high confident key points. Formally, given the features $\boldsymbol{H}^0$ extracted from the map-to-node modules, we utilize the global average pooling layer followed by two fully-connected layers to map $\boldsymbol{H}^0$ to a vector $\boldsymbol{a}$ whose size is equal to the non zeros in $\boldsymbol{M}$. Finally, we replace the non zeros value in $\boldsymbol{M}$ with $\boldsymbol{a}$ to form the dynamic adjacency matrix $\boldsymbol{A}$. We employ the binary matrix $\boldsymbol{M}$ to hold the neighborhood and only learn their weights because the facial shape pattern is stable. By fixing the sparse connection, we greatly reduce the training parameters which makes the learning process easier.

C. Graph Relational Layers

Now we have the embed node feature and the dynamic adjacency matrix, we can infer the landmark coordinate via the graph relational reasoning. Specifically, the input embed node feature $\boldsymbol{H}^0$ is first fed to a graph convolution, followed by several graph residual blocks. The last graph convolution (without batch normalization and ReLU operations) maps the hidden node features to landmark coordinates $\boldsymbol{P} \in \mathbb{R}^{N\times 2}$. An overview of the graph relational layers architecture is shown in Fig. 5 for more clarity.

Unlike standard convolutions that operate on local Euclidean structures, *e.g.*, an image grid, the goal of graph convolution is to learn a function $f(\cdot,\cdot)$ on a graph $\mathcal{G}$, which takes node feature $\boldsymbol{H}^l \in \mathbb{R}^{N\times d_l}$ and the corresponding adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ as input, and outputs the node features as $\boldsymbol{H}^{l+1} \in \mathbb{R}^{N\times d_{l+1}}$. Here N, l, d_l and d_{l+1} denote the number of nodes, index of graph layer, the dimension of input node features and the dimension of output node feature, respectively. Every graph convolutional layer can be written as a non-linear function by,

$$\boldsymbol{H}^{l+1} = f(\boldsymbol{H}^l, \boldsymbol{A}) \tag{3}$$

With the specific graph convolutional operators employed by [40], the layer can be represented as,

$$\boldsymbol{H}^{l+1} = \psi(\sigma(\boldsymbol{A})\boldsymbol{H}^l\boldsymbol{W}^l) \tag{4}$$

where $\boldsymbol{W}^l \in \mathbb{R}^{d_l\times d_{l+1}}$ is a transformation matrix to be learned, $\sigma(\cdot)$ denotes a normalization operation, and $\psi(\cdot)$ denotes BN-ReLU operation. Following the strategy in [50], we adopt a row-wise softmax operation as σ. Softmax operation makes the weights of each node like probabilities over its neighboring nodes, which stabilizes the training process.

Inspired by the success of ResNet [20], we adopt the graph residual block architecture. Each graph block consists of two graph convolutional layers and can be formulated based on Eq. (3) as

$$\begin{aligned} \boldsymbol{H}^{l+1} &= f(\boldsymbol{H}^l, \boldsymbol{A}) \\ \boldsymbol{H}^{l+2} &= f(\boldsymbol{H}^{l+1}, \boldsymbol{A}) + \boldsymbol{H}^l \end{aligned} \tag{5}$$

D. Loss Functions

We propose two loss functions to train our model: the relation location loss and the soft wing loss.

Relative location Loss: Conventionally, given predicted landmark points $\hat{\boldsymbol{P}}$ and ground-truth landmark points $\boldsymbol{P}$, the objective is to minimize the error of the absolute location of landmark points, *i.e.*, $||\boldsymbol{P} - \hat{\boldsymbol{P}}||$. However, by simply minimizing the error of absolute position, this loss function ignores the relative location between landmark points. Such relative position cues are crucial for preserving the facial structure and allow the model to predict coherent facial landmarks. In order to provide such information to our model during the training, we propose a relative location aware loss function which is based on the Laplacian prior from geometric modeling or 3D meshing [51]. Specifically, given a landmark point $\boldsymbol{p}_i$ and its neighborhood set $\mathcal{N}_i$, where $\mathcal{N}_i = \{j : j \neq i,\ \boldsymbol{M}_{ij} = 1\}$, the Laplacian of $\boldsymbol{p}_i$ can be written as:

$$\boldsymbol{\delta_i} = \sum_{j\in\mathcal{N}_i} \omega_{ij}(\boldsymbol{p_i} - \boldsymbol{p_j}) = \boldsymbol{p_i} - \sum_{j\in\mathcal{N}_i} \omega_{ij}\boldsymbol{p_j}, \tag{6}$$

where $\sum_{j\in\mathcal{N}_i} \omega_{ij} = 1$ denotes the weight between the nodes. Our relative location loss minimizes the difference between the predicted $\hat{\boldsymbol{\delta}}_i$ and the ground truth $\boldsymbol{\delta}_i$, which acts as a regularization term to penalize infeasible local landmark shape. As far as our knowledge, we are the first to apply such loss in 2D facial landmark detection.

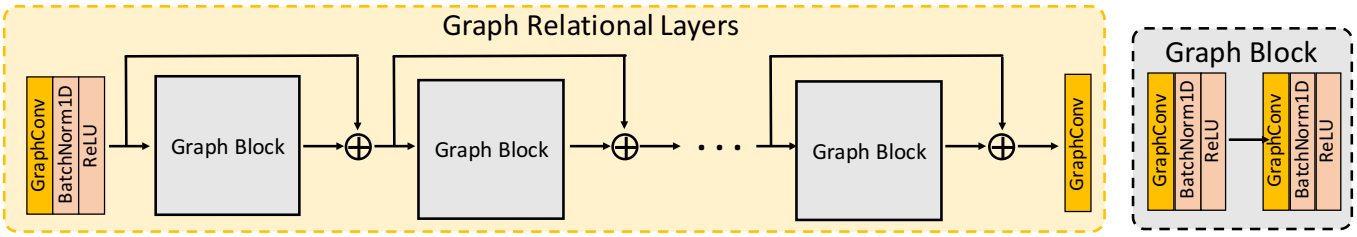


Fig. 5. Illustration of the architecture of our graph relational layers. It consists of a graph convolution-BN-ReLU and several residual graph blocks and a graph convolution to output the final landmark coordinate. Each graph block is composed of two graph convolution-BN-ReLU. $\oplus$ denotes the element-wise addition operation and GraphConv represents the graph convolution.

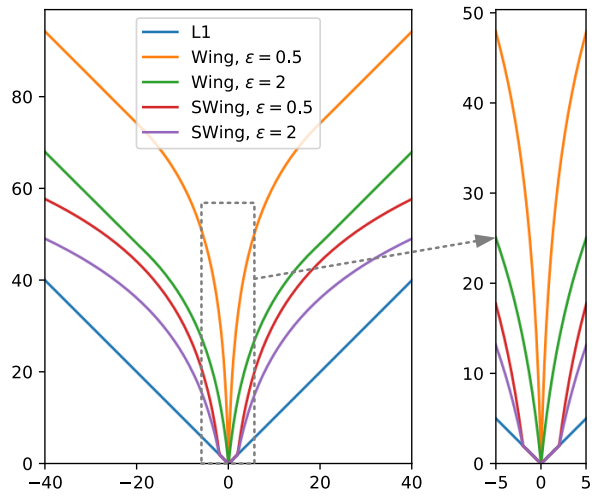


Fig. 6. Illustration of L1, Wing and Soft Wing loss functions. ω_1 is set 2. ω and ω_2 are set to 20. Unlike Wing loss, our loss is linear for small errors.

To facilitate the Laplacian computation for all landmarks $\boldsymbol{P} \in \mathbb{R}^{N\times 2}$, we construct a $N \times N$ Laplacian Matrix $\boldsymbol{L}$ as follows:

$$\boldsymbol{L}_{i,j} = \begin{cases} -\omega_{ij} & \text{if } j \in \mathcal{N}_i \\ 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and the Laplacians $\Delta = [\boldsymbol{\delta_1}, \boldsymbol{\delta_2}, \cdots, \boldsymbol{\delta_N}]^T$ can be easily computed via a matrix multiplication:

$$\Delta = \boldsymbol{LP}. \tag{8}$$

In this work, we utilize the uniform Laplacian where the neighbors are equally weighted, *i.e.*, $\omega_{ij} = \frac{1}{|\mathcal{N}_i|}, \forall j \in \mathcal{N}_i$, with $|\mathcal{N}_i|$ the number of neighborhood. The uniform Laplacian is quite simple for implementation and has the nice property that its weights are independent of the landmark positions.

Soft Wing Loss: While the relative location loss permits to enhance the structure among landmarks, the absolute location loss plays an important role for accurate face alignment. The L2 and L1 losses are mainly utilized as absolute location loss functions in most existing deep learning based methods. However, as pointed in [2] the L2 loss tends to pay more attention to large errors and overlooks relatively small errors. This property prevents the face alignment model to precisely localize facial landmarks. Therefore, we present Soft Wing loss to make our model more focus on the errors of medium range:

$$\text{SoftWing}(x) = \begin{cases} |x| & \text{if } |x| < \omega_1 \\ \omega_2 \ln(1 + \frac{|x|}{\epsilon}) + B & \text{otherwise} \end{cases} \tag{9}$$

which is linear for small values, and take the curve of $\ln(\cdot)$ for medium and large values. Similar to Wing loss [2], we use the non-negative ω_1 to switch between linear and non-linear part, and ϵ to limit the curvature of the non-linear part. B is set to $\omega_1 - \omega_2 \ln(1 + \omega_1/\epsilon)$ to make function continuous at ω_1. The soft wing loss is plotted in Fig. 6.

The overall loss function for N landmarks is:

$$\mathcal{L} = \sum_{i=1}^{N} l(\boldsymbol{p}_i - \hat{\boldsymbol{p}}_i) + \lambda l(\hat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i), \tag{10}$$

where the first term corresponds to the soft wing loss and the second one is the relative location loss. $l(\cdot)$ corresponds to the soft wing function and λ is the parameter to regularize the importance of loss functions. We set $\lambda = 1$ by cross-validation.

E. Discussion

Comparison with Fully-connected layer: The fully-connected layer and our graph relational layers embed the feature of landmarks in two different ways. As illustrated in Fig. 2(a) The CNN backbone and the hidden fully-connected layer map the input facial image to the hidden vector, which embed the feature of landmarks globally. Thus, the errors of some parts of the prediction effects the other parts, as they share the same hidden feature. As we can observe in Fig. 2(b), for the FC-based method, the errors of occluded part interfere the prediction of other visible parts, resulting in wrong localization. Meanwhile, our graph relational layers embeds the node feature for each landmark, and propagates node feature according to their relationship and the dynamic edge weights. If some parts of predictions fail because of the occlusion, large pose or other hard condition, the node feature of other parts degrade gracefully because of the sparse connection among the node features and the dynamic adjustment of the relationship. As shown in Fig. 2(d), our SDFL based method is more robust to hard cases and produces correct localization even in extreme occlusion. Besides, fully-connected layers are prone to overfit because of the large number of trainable parameters, while

the graph convolution layer requires much fewer trainable parameters. Taking $N = 98$ (number of landmarks annotated in WFLW dataset) and the hyperparameters used in [2] as an example, the hidden and last fully-connected layer contain $(2\times2\times512)\times1024 = 2,097,152$ and $1024\times196 = 200,704$ parameters, respectively, while our hidden and last graph convolutional layer only need to learn $128 \times 128 = 16,384$ and $128 \times 2 = 256$ parameters, respectively. Our graph relational layers require thus approximatively 138 times fewer parameters than fully-connected layers.

Comparison with Wing Loss: Wing loss [2] has constant gradient when error is large, and large gradient for small or medium range errors, which is defined as:

$$\text{Wing}(x) = \begin{cases} \omega \ln(1 + \frac{|x|}{\epsilon}) & \text{if } |x| < \omega \\ |x| - C & \text{otherwise} \end{cases} \tag{11}$$

where x is error and C is $\omega - \omega \ln(1 + \omega/\epsilon)$ to smoothly link two piece-wise functions. According to our experiment, the performance of Wing loss is not consistently better than L1 loss, especially when we train the neural networks on difficult dataset with heavy occlusion and blur, such as WFLW. As mentioned in [31], this may be caused by inconsistent annotations due to various reasons, e.g., unclear or inaccurate definition of some landmarks, poor quality of some facial images. Imposing a large gradient magnitude around very small error to force the model exactly fit the ground truth landmarks makes the training process unstable. The visualization of L1, Wing and our Soft Wing loss is shown in Fig. 6. Note that we discard the linear part of Wing loss, since our proposed loss can adaptively adjust the magnitude of gradient between medium ($\omega_1 < |x| < \omega_2$) and large errors ($|x| > \omega_2$). The magnitude of gradient of the non-linear part is $\frac{\omega_2}{|x|+\epsilon} \approx \frac{\omega_2}{|x|}$ (ϵ is commonly set to small value). Our proposed loss is insensitive to outliers where the gradient varies between $[\frac{\omega_2}{C}, 1]$ (C is the image size). Note that ω_2 should not set to small value because it will cause gradient vanishing problem.

IV. Experiments and Analysis

In this section, we first introduce the face alignment datasets and evaluation metrics that we utilized for experiments. We then provide the implementation details, show the effectiveness of proposed method on top of different convolutional backbones, and then compare the proposed method with the state-of-the-art approaches. Finally, we present the comprehensive ablation study which shows the contribution of different key designs of our framework.

A. Datasets and Evaluation Protocols

We conducted our experiments on three widely-adopted challenging datasets: WFLW [1], COFW [16] and 300W [17]. Here we provide a brief description of these datasets and present the evaluation metrics.

Wider Facial Landmarks in-the-Wild (WFLW) dataset is among the most challenging face alignment benchmark which includes various hard cases such as heavy occlusion, blur and large pose. The whole WFLW dataset consists of 10,000 facial images which are splitted into 7,500 training images and 2,500 testing images. Each image in this dataset is manually annotated with 98 facial landmarks. The testing set is further divided into several subsets such as large pose, expression, illumination, make-up, occlusion and blur, which permits to evaluate the performance of methods facing different issues.

Caltech Occluded Faces in the Wild (COFW) dataset is collected to present faces with large variations in shape and occlusions in real-world conditions. COFW is an extension of the Labeled Facial Parts in the Wild (LFPW) dataset [52], by complementing additional training and test examples with heavy occlusions. Various types of occlusions are introduced and result in a 23% occlusion on facial parts on average. The dataset includes 1,345 training and 507 test images, manually annotated with 29 landmarks. We also use the re-annotated test set [53] with 68 landmarks annotation for cross-dataset validation.

300 Faces In-the-Wild Challenge (300W) dataset contains face images with moderate variations in pose, expression and illumination. The training set (3,148 images) includes the fullset of AFW and the training subsets of LFPW and HELEN. The full testing set is divided into common subset (554 images) and challenging subset (135 images). Namely, the common subset consists of face image in relatively simple scenarios, while the challenging subset is composed of relatively difficult samples. The face images in this dataset are semi-automatically annotated with 68 facial landmarks.

Evaluation Metric: We followed the standard evaluation protocol and evaluated the proposed method with normalized mean error (NME), failure rate (FR) and area under curve (AUC).

The NME for each image is defined as:

$$NME(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{N}\sum_{i=1}^{N}\frac{||\boldsymbol{p}_i - \hat{\boldsymbol{p}}_i||_2}{d} \tag{12}$$

where $\boldsymbol{p}_i$ and $\hat{\boldsymbol{p}}_i$ are the i-th ground truth and predicted landmark coordinates respectively, N denotes the number of landmarks in an image, the inter-ocular distance is utilized as the normalization factor d.

Failure Rate, namely evaluates the number failed prediction. Given an image, if the NME is larger than a threshold, then it is considered as a failed estimation. Following the protocol in [1], the failure rate for a maximum error of 10% is reported.

Area Under Curve is calculated based on the cumulative error distribution curve which presents the NME to the proportion of total test samples. Larger AUC means that larger portion of the test dataset is correctly predicted.

B. Implementation details

We present here some experimental settings and training strategies.

Data augmentation: All training and testing images are center-cropped and resized to 256×256 before feeding to our model. In order to improve the generalization capability of our model, we augmented the training data with random rotation ($\pm40°$), translation (±30 px), flipping (50 %), rescaling (±10 %) and random occlusion (20 % of image size). To

TABLE I
EVALUATION ON THE WFLW DATASET USING FULLY-CONNECTED (FC) LAYERS AND OUR LANDMARK-GRAPH RELATIONAL LAYERS (L-GRL) AS PREDICTOR HEAD ON TOP OF VARIOUS BACKBONES. CONSISTENT IMPROVEMENT IS OBSERVED WITH OUR SDFL METHOD AND SOFT WING LOSS (SWING) COMPARED TO FC AND THE L1 LOSS.

Backbone	Head	NME(%) L1	NME(%) SWing
MobileNetV2	FC	4.99	4.86
	L-GRL	4.77 (↓0.22)	4.70 (↓0.16)
EfficientNet-B0	FC	5.00	4.80
	L-GRL	4.79 (↓0.21)	4.61 (↓0.19)
EfficientNet-B1	FC	4.92	4.81
	L-GRL	4.79 (↓0.13)	4.60 (↓0.21)
VGG16	FC	5.57	4.83
	L-GRL	5.42 (↓0.15)	4.71 (↓0.12)
ResNet18	FC	5.02	4.95
	L-GRL	4.76 (↓0.26)	4.65 (↓0.30)
ResNet34	FC	4.88	4.83
	L-GRL	4.67 (↓0.21)	4.55 (↓0.28)
Res2Net50_26w_4s	FC	4.89	4.85
	L-GRL	4.60 (↓0.29)	4.55 (↓0.30)
HRNetW18C	FC	4.56	4.54
	L-GRL	4.41 (↓0.15)	4.35 (↓0.16)

mitigate the issue of pose variations, we adopt the Pose-based Data Balancing (PDB) [2] strategy with 9 bins.

Model architecture: We utilized different convolutional backbones including MobileNetV2 [18], EfficientNet [19], VGG16 [22], ResNet [20], Res2Net [21] and HRNet [12] as our backbone for the experiments. For HRNet, we downsample feature maps of different resolutions into 8×8. For the architecture of our graph relational layers, we deployed 4 graph residual blocks with hidden feature dimension set to 128. We set the number of neighborhood k to 3 for adjacency matrix.

Training: During the training, we employed vanilla SGD for optimization with a batch size of 64 for 500 epochs. We set the weight decay and the momentum to 0.0005 and 0.9 respectively. The initial learning rate is 0.01 which is dropped by 5 every 100 epochs. The parameters of the Soft Wing loss are set to $\omega_1 = 2$, $\omega_2 = 20$ and $\epsilon = 0.5$. Our models are trained from scratch using Pytorch.

C. SDFL with Different Backbones

Our structure-coherent learning method can be easily implemented with different convolutional backbones. Indeed, the landmark-graph relational layers (L-GRL) in our SDFL can be easily constructed on top of different convolutional backbones as a predictor head. To assess the effectiveness of our SDFL method, we compare it with the commonly used predictor head, *i.e.*, fully-connected layers across several popular convolutional backbones such as MobilenetV2 [18], EfficientNet [19], ResNet [20], VGG16 [22], Res2Net [21] and HRNet [12]. For the fully-connected layers based predictor head, we utilized a hidden FC layer with 256 units followed by a ReLU activation, and a FC layer to output a vector of $2N$ real numbers for the 2D coordinates of N landmarks. We employed Soft wing loss and L1 loss as supervision signals to train both the FC based and SDFL based models. The results on the WLFW dataset are listed in Tab. I. We observe noticeable improvements when our graph relational layers are utilized as the predictor head, regardless of the convolutional backbones and the loss functions. The improvement is particularly significant when the convolutional backbone is small, *e.g.*, ResNet18 with 0.26% and 0.30% gains using L1 and Soft wing losses, respectively. This means that the proposed SDFL method is more robust than FC layers in case where the features extracted from the input image have limited presentation capability. These results demonstrate superiority of our SDFL compared to the coordinate regression based methods which commonly employed FC layers. Furthermore, the graph relational layers in our SDFL entail fewer learnable parameters because of the parameter sharing scheme in graph convolution operations. While in the fully-connected layers, due to their dense connection property, there are a large number of parameters to be optimized. Note that, to make the comparison more convincing, we conducted a series of experiments with various settings of FC layers (number of layers and units) and selected the best result listed in Tab. I.

D. Comparison with the State-of-the-Art Methods

In this part, we compare the proposed method with best performing approaches on three face alignment benchmarks.

WFLW: We evaluate our approach on the WFLW dataset and compare with state-of-the-art methods in terms of normalized mean error (NME), failure rate (FR) and area under curve (AUC). To better understand the effectiveness of the proposed method, we analyze the performance on six subsets with specific issue, *e.g.*, large pose, exaggerated expression, illumination, make-up, occlusion and blur [1]. The overall results are tabulated in Table II. The proposed method achieves 4.35% NME, 2.72% Failure Rate at 10% and 0.5759 AUC, which outperforms most state-of-the-art approaches. Our method fails on only 2.72% of all images, which demonstrates the robustness of our model. AWing [57] is the most competitive approach which shows very great results on WFLW. Since AWing is a heatmap based method, it can produce much more precise landmark localization than our coordinate regression based method. However, our method still performs slightly better than AWing in terms of NME and by large margin in terms of FR. Since the WFLW consists numerous hard cases, this comparison points out that our model performs much better than AWing on the difficult scenarios, which highlight the robustness of the proposed approach. Some qualitative results are depicted in Fig. 7 (a), where our model successfully localizes landmarks in hard cases, such as occlusion, make up and large pose.

COFW: We compare the proposed method with existing face alignment approaches on the COFW dataset and tabulate the results in Table III. When using the COFW training set for training and evaluating with 29 landmarks, our method achieves state-of-the-art performance with 3.63% normalized mean error and 0% failure rate at 10%. These impressive results outperform the state-of-the-art methods by a large margin, which demonstrate the superiority of the proposed SDFL. Note that the 0% failure rate means that our model correctly localizes the facial landmark on all testing images.

TABLE II
EVALUATION OF OUR METHOD AND STATE-OF-THE-ART APPROACHES ON FULLSET AND SIX TYPICAL SUBSETS OF WFLW. THE RESULTS IN TERMS OF NORMALIZED MEAN ERROR, NME (%), FAILURE RATE AT 10%, FR (%) AND AUC ARE REPORTED.

Metric	Method	Fullset	Pose	Expression	Illumination	Make-up	Occlusion	Blur
NME	DVLN17[4]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
	3FabRec20 [54]	5.62	10.23	6.09	5.55	5.68	6.92	6.38
	LAB18 [1]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	SRT20 [33]	5.13	-	-	-	-	-	-
	Wing18 [2]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	AGCFN19 [55]	4.90	8.78	5.00	4.93	4.85	6.26	5.73
	LAB18 [1] + AVS19 [56]	4.76	8.21	5.14	4.51	5.00	5.76	5.43
	DeCaFA19 [3]	4.62	8.11	4.65	4.41	4.63	5.74	5.38
	HRNet19 [12]	4.60	7.94	4.85	4.55	4.29	5.44	5.42
	AWing19 [57]	4.36	**7.38**	**4.58**	4.32	4.27	5.19	**4.96**
	SDFL (Ours)	**4.35**	7.42	4.63	**4.29**	**4.22**	**5.19**	5.08
FR	DVLN17[4]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	3FabRec20 [54]	8.28	34.35	8.28	6.73	10.19	15.08	9.44
	LAB18 [1]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	SRT20 [33]	7.07	-	-	-	-	-	-
	Wing18 [2]	6.00	22.72	4.78	4.30	7.77	12.50	7.76
	AGCFN19 [55]	5.92	24.23	5.41	4.72	5.82	11.00	8.79
	LAB18 [1] + AVS19 [56]	5.24	20.86	4.78	3.72	6.31	9.51	7.24
	DeCaFA19 [3]	4.84	21.40	3.73	3.22	6.15	9.26	6.61
	AWing19 [57]	2.84	13.50	2.23	2.58	2.91	5.98	3.75
	SDFL (Ours)	**2.72**	**12.88**	**1.59**	**2.58**	**2.43**	**5.71**	**3.62**
AUC	DVLN17[4]	0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	3FabRec20 [54]	0.4840	0.1920	0.4480	0.4960	0.4730	0.3980	0.4340
	LAB18 [1]	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
	SRT20 [33]	0.5464	-	-	-	-	-	-
	Wing18 [2]	0.5504	0.3100	0.4959	0.5408	0.5582	0.4885	0.4918
	AGCFN19 [55]	0.5452	0.2826	0.5267	0.5511	0.5547	0.4621	0.4823
	LAB18 [1] + AVS19 [56]	0.5460	0.2764	0.5098	0.5660	0.5349	0.4700	0.4923
	DeCaFA19 [3]	0.563	0.292	0.546	0.579	0.575	0.485	0.494
	AWing19 [57]	0.5719	0.3120	0.5149	0.5777	0.5715	0.5022	0.5120
	SDFL (Ours)	**0.5759**	**0.3152**	**0.5501**	**0.5847**	**0.5831**	**0.5035**	**0.5147**

To further verify the generalization capability of our method, we conduct a cross-dataset evaluation using 300W for training and evaluate on the COFW-68 dataset annotated with 68 landmarks [53]. Our method outperforms the existing best approaches by a large margin, with 4.18% NME and 0% FR. Since the COFW dataset is mainly composed of occluded faces, this impressive performance indicates the robustness of our structure-coherent framework to handle heavy occlusions. Some qualitative results are plotted in Fig. 7 (c), where the proposed method effectively infers occluded facial parts without letting the occlusion interfere the visible parts.

300W: We compare our approach against the existing best performing methods on the 300W dataset. The results are reported in Table IV. Our model achieves a NME of 2.88% on the common set, 4.9% on the challenging set and 3.28% on the full set. Our method outperforms most existing approaches and achieves very competitive results with the state-of-the-art methods. Since the 300W dataset is composed of facial image with relatively simple scenarios, the localization precision is crucial for the final performance. While the robustness to hard cases such as occlusion counts less in the final results. As LUVLi [70] and AWing [57] are heatmap regression based face alignment methods, they naturally perform better than our coordinate regression approach due to their higher localization precision. Some qualitative results are illustrated in Fig. 7 (b), where our model obtains accurate landmarks localization in large poses and expressions.

Efficiency Analysis: We analyzed the computational complexity with floating point operations (FLOPs). Suppose the number of landmarks is N, input feature size is I and output feature size is O. A graph convolutional layer (without activation) can be viewed as product of three matrices, *i.e.* normalized adjacency matrix $A \in \mathbb{R}^{N\times N}$, input node features $H \in \mathbb{R}^{N\times I}$ and transformation matrix $W \in \mathbb{R}^{I\times O}$. A is a $N(k+1)$-sparse matrix ($k \ll N$). So AHW costs $2N(k+1)I + 2NIO = 2N(k+1+O)I$ FLOPs. While the FLOPs for FC layers under the same input/output condition is $2NINO$. The ratio of the a graph convolutional layer and a

TABLE III
EVALUATION ON THE COFW DATASET IN TERMS OF NME (%) AND FAILURE RATE (%) AT 10%.

Method	Trained on COFW		Trained on 300W	
	NME	FR	NME	FR
TCDCN14[30]	-	-	7.66	16.17
SAPM15[58]	-	-	6.64	5.72
CFSS15[59]	-	-	6.28	9.07
HPM14[53]	7.50	13.00	6.72	6.71
CCR15[60]	7.03	10.9	-	-
DRDA16[61]	6.46	6.00	-	-
RAR16[62]	6.03	4.14	-	-
SFPD17[63]	6.40	-	-	-
DAC-CSR17[64]	6.03	4.73	-	-
Wing18[2]	5.44	3.37	-	-
ODN19[65]	5.30	-	-	-
LAB18[1]	3.92	0.39	4.62	2.17
SAN18[66] + AVS19[56]	-	-	4.43	2.82
AWing19[57]	4.94	0.99	-	-
SDFL (Ours)	**3.63**	**0**	**4.18**	**0**

a) Qualitative results on the WFLW testset

b) Qualitative results on the 300W testset

c) Qualitative results on the COFW testset

Fig. 7. Visualization of results on some testing image from (a) the WFLW dataset with 98 landmarks, (b) the 300W dataset with 68 landmarks and (c) the COFW dataset with 29 landmarks. From these results, we can see that our model outputs accurate landmarks localization in difficult cases such as make-up, occlusion, large pose and expression. Better viewed in color.

FC layer with respect to the FLOPs is $\frac{k+1+O}{NO}$, which shows that the computational complexity of a graph convolutional layer is much smaller than FC layer if the input and output size are the same.

In our experiments, we use 4 graph blocks and 1 graph layer. Denote F as the hidden feature size. N, I, F, O, k are set to $98, 256, 128, 2, 3$. The overall FLOPs for our L-GRN is $2N(k+1+F)I + 14N(k+1+F)F + 2N(k+1+O)F = 29.95$M FLOPs by omitting the addition operations of the skip connections which are negligible. We compare the FLOPs with some existing methods and tabulate the results in Table IV-D. The comparison shows that our model achieve better face alignment performance with smaller FLOPs. We also record the runtime of our ResNet34 + L-GRN model on a 1080Ti GPU which takes 23ms to process a 256×256 image.

E. Ablation Study

To better understand our model, we performed ablation experiments on the WFLW dataset using ResNet18 as convolutional backbone. In particular, we analyzed the effect of the number of neighbors k when constructing our adjacency matrix in the landmark-graph relational layers, and then demonstrated the importance of the dynamic edge weighting. We also examined the contribution of the proposed relative location loss and soft wing loss to the performance. The whole results are tabulated in Table VI.

Number of neighbors: We performed the experiments with different values of k from $k = 1$ to $k = 97$ and performed

TABLE IV
EVALUATION ON THE 300W COMMON SUBSET, CHALLENGING SUBSET AND FULLSET IN TERMS OF NME(%).

Method	Common	Challenging	Full
PCD-CNN$_{18}$ [67]	3.67	7.62	4.44
Chandran *et al.*$_{20}$ [32]	2.83	7.04	4.23
CPM+SBR$_{18}$ [68]	3.28	7.58	4.10
SAN$_{18}$ [66]	3.34	6.60	3.98
3FabRec$_{20}$ [54]	3.36	5.74	3.82
LAB$_{18}$ [1]	2.98	5.19	3.49
TS$^3{}_{19}$ [34]	2.91	5.91	3.49
DU-Net$_{19}$ [69]	2.97	5.53	3.47
DeCaFA$_{19}$ [3]	2.93	5.26	3.39
SRT $_{20}$ [33]	2.80	5.61	3.39
HRNet$_{19}$ [12]	2.87	5.15	3.32
LUVLi$_{20}$ [70]	2.76	5.16	3.23
AWing$_{19}$ [57]	**2.72**	**4.52**	**3.07**
SDFL(Ours)	2.88	4.93	3.28

TABLE V
EFFICIENCY COMPARISON IN TERMS OF FLOPS AND NME ON WFLW FULLSET.

Model	FLOPS (G)	NME (%)
LAB [1]	28.583	5.27
Wing [2]	5.396	5.11
Ours	5.165	4.55

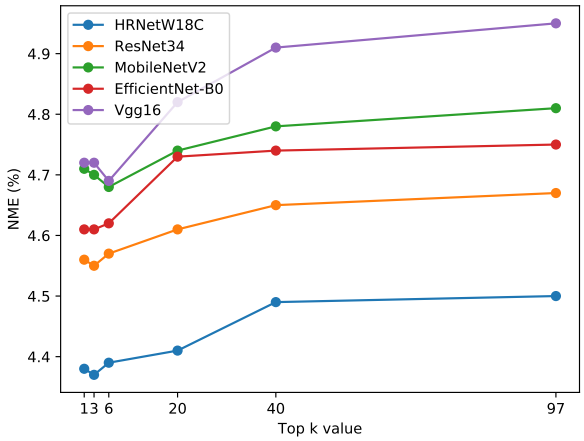


Fig. 8. NME(%) comparison with different number of neighbors k across various convolutional backbones (HRNetW18C, ResNet34, MobileNetV2, EfficientNet-B0 and VGG16).

this design experiment with several convolutional backbones. The results are depicted in Fig. 8. Our model achieves great performances on WFLW dataset with small number of neighbors, *i.e.* $k = 3$ or $k = 6$ for different backbones. According to the experimental results, the performance degrades if the adjacency matrix is too sparse or too dense. When k is too small, each graph node cannot obtain sufficient information from its correlated neighborhood. While when k is too large, the adjacency matrix becomes dense which leads to over-smoothing of the node features. Note that, when the number of neighbors is large enough where landmark-nodes are densely connected, the graph relational layers can be seen as a series of fully-connected layers. The performance degradation with dense neighborhood confirms that dense connection tends to incorporate redundant and noisy information from occluded landmarks. This observation is common to all convolutional backbones used in the experiments, which demonstrate that using an appropriate number of neighborhood is crucial for a great performance.

Neighborhood Construction: In our experiments, we utilized quite simple strategy to describe relationship among landmarks, *i.e.*, separately considering x- and y-axis landmark information. We also jointly considered x- and y-axis landmark information to construct the neighborhood. Using the adjacency matrix built upon this joint strategy, we observe a performance variation of 0.03%. Since the separate and joint consideration of x- and y-axis information strategies lead to negligible localization results difference, we prefer the separate strategy which is more direct and simple for implementation.

Dynamic adjacency matrix weighting: We analyzed the contribution of the dynamic neighborhood learning, which attributes different weights to the edges of graph with respect to the features extracted from input image. For that, we replaced the dynamic adjacency matrix with a binary adjacency matrix where all edges have a same weight set to 1. We observed a degradation of 0.2%, which points out the importance of dynamic neighborhood learning.

Relative location loss: We then investigated the contribution of our relative location loss. When we disable the relative location supervision, we observe a performance drop of 0.1% to 0.2% depending on the designs choices, e.g., using L1 or soft wing loss as absolute supervision signal and fully-connected layers or graph relational layers as predictor head.

Soft wing loss: We compared the effect of different absolute location loss functions such as L1, Wing and our Soft Wing loss. As tabulated in Table VI and VII, our Soft Wing loss consistently outperforms Wing loss and L1 loss. The performance of Wing loss degrades when ϵ decreases, while our loss benefits from imposing larger gradients on medium range errors. The performance of Wing loss is even worse than L1 loss when ϵ is very small. These results demonstrate the superiority of the proposed soft wing loss.

Qualitative analysis of features nodes: We examined how the nodes are processed within our graph relational layers. Different from the spatial features in CNNs which tend to activate spatial regions that correspond to the most salient parts of the input image, the features of graph nodes do not have such correspondence. Thus, visualization of the nodes features are quite meaningless. Instead, we analyzed the evolution of similarity of the nodes within the graph relational layers. For that, we computed the cosine similarity among nodes features and display the visualization in Fig. 9. We can see that the node embedding features present limited similarity, but through the graph relational layers, the features of connected nodes become much more similar. This visualization shows that the graph propagation permits to assemble the features of connected components.

TABLE VI
ABLATION EXPERIMENTS EVALUATED ON THE WFLW TEST SET USING RESNET18 AS THE CONVOLUTIONAL BACKBONE. ANALYSIS SHOW THE EFFECTS OF VARIOUS COMPONENTS AND DESIGN CHOICES ON THE FACE ALIGNMENT PERFORMANCE IN TERMS OF NME.

Design	Choice								
FC layers	✓	✓	✓	✓					
Graph relational layers					✓	✓	✓	✓	✓
L1 loss	✓		✓		✓		✓		
Soft wing loss		✓		✓		✓		✓	✓
Relative location loss			✓	✓			✓	✓	✓
Dynamic adjacency matrix					✓	✓	✓	✓	
NME	5.23	5.06	5.02	4.95	4.87	4.75	4.76	4.65	4.85

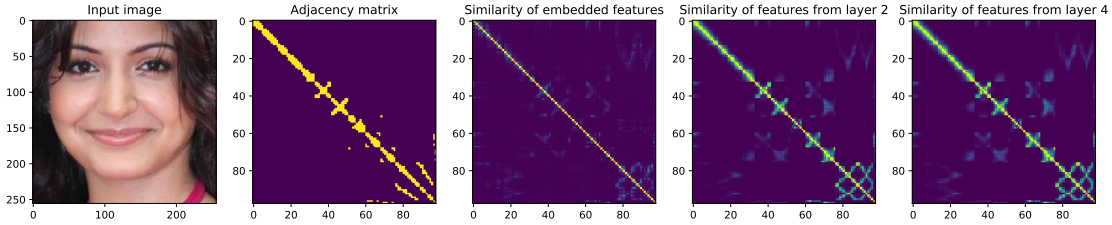


Fig. 9. Visualization of the input image, adjacency matrix and the similarity among features of nodes. We show the evolution of the similarity through the graph relational layers by plotting the input embedded features, features from the 2nd layer and then from the 4th layer of the graph relational layers.

TABLE VII
COMPARISON OF DIFFERENT LOSS FUNCTIONS. ANALYSIS SHOWS THE EFFECTIVENESS OF SOFT WING LOSS IN TERMS OF THE NME (%).

epsilon	0.1	0.2	0.5	1	1.5	2
L1			4.76			
Wing	5.38	4.94	4.76	4.74	4.75	4.73
SoftWing	4.68	4.66	**4.65**	4.71	4.71	4.72

V. CONCLUSION

In this paper, we propose a structure-coherent deep feature learning method for face alignment. We present a landmark-graph relational network which consists of a convolutional backbone, a node embedding module, a dynamic adjacency matrix weighting module and graph relational layers, to explore the relation among landmarks. By appropriately considering the interaction among facial key points, our model achieves correct facial landmarks localization under hard cases. In addition, we introduce a relative location loss function to further enhance the facial structure coherency and a soft wing loss as an improved version of wing loss which permits our model to focus on medium rate error during the training, resulting in a better convergence. Experimental results on three challenging face alignment benchmarks demonstrate the effectiveness of the proposed method.

Although our method performs well on hard scenarios such as occlusion, the localization precision is a shortage of the proposed method which is a coordinate regression based framework. We think that incorporating the relationship among landmarks into the heatmap based face alignment pipeline, merging the high precision and the structure coherence, would be an interesting future work.

REFERENCES

[1] 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.

[2] Z.-H. 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.*, 2017, pp. 2235–2245.

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

[4] W. Wu and S. Yang, "Leveraging intra and inter-dataset variations for robust face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, July 2017.

[5] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1233–1245, 2016.

[6] H. Liu, J. Lu, J. Feng, and J. Zhou, "Learning deep sharable and structural detectors for face alignment," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1666–1678, 2017.

[7] Q. Liu, J. Deng, J. Yang, G. Liu, and D. Tao, "Adaptive cascade regression model for robust face alignment," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 797–807, 2016.

[8] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 787–796.

[9] Y. Zhang, R. Zhao, W. Dong, B.-G. Hu, and Q. Ji, "Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7034–7043.

[10] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2387–2395.

[11] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2017, pp. 2025–2033.

[12] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *CoRR*, vol. abs/1904.04514, 2019.

[13] 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.*, June 2016.

[14] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International Conference on Machine Learning*, 2016, pp. 2990–2999.

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

[16] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, December 2013.

[17] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: database and results," *Image and Vision Computing*, vol. 47, pp. 3 – 18, 2016, 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[19] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *International Conference on Machine Learning*, 2019.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[21] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[24] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 300–305.

[25] D. Cristinacce and T. Cootes, "Feature detection and tracking with constrained local models," *Pattern Recognit.*, vol. 41, pp. 929–938, 01 2006.

[26] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, pp. 200–215, 01 2011.

[27] 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. 2385–2392.

[28] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2729–2736.

[29] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2009.

[30] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[31] Z. Liu, X. Zhu, G. Hu, H. Guo, M. Tang, Z. Lei, N. M. Robertson, and J. Wang, "Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[32] P. Chandran, D. Bradley, M. Gross, and T. Beeler, "Attention-driven cropping for very high resolution facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5861–5870.

[33] 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.*, 2020.

[34] X. Dong and Y. Yang, "Teacher supervises students how to learn from partially labeled images for facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 783–792.

[35] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE NN*, vol. 20, no. 1, pp. 61–80, Jan 2009.

[36] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," *IEEE SPM*, vol. 34, no. 4, pp. 18–42, July 2017.

[37] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," *arXiv preprint arXiv:1511.05493*, 2015.

[38] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*, 2017, pp. 1263–1272.

[39] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," *arXiv preprint arXiv:1312.6203*, 2013.

[40] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017, pp. 1–10.

[41] R. Liao, Z. Zhao, R. Urtasun, and R. S. Zemel, "Lanczosnet: Multi-scale deep graph convolutional networks," *arXiv preprint arXiv:1901.01484*, 2019.

[42] L. Liu, T. Zhou, G. Long, J. Jiang, and C. Zhang, "Learning to propagate for graph meta-learning," *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[43] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5199–5208.

[44] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020.

[45] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 670–685.

[46] J. Shi, H. Zhang, and J. Li, "Explainable and explicit visual reasoning over scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[47] J. Zhang, Q. Wu, J. Zhang, C. Shen, and J. Lu, "Mind your neighbours: Image annotation with metadata neighbourhood graph co-attention networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[48] H. Xu, C. Jiang, X. Liang, and Z. Li, "Spatial-aware graph relation network for large-scale object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[49] N. Verma, E. Boyer, and J. Verbeek, "FeaStNet: Feature-Steered Graph Convolutions for 3D Shape Analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2598–2606.

[50] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3425–3435.

[51] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rossl, and H.-P. Seidel, "Laplacian surface editing," in *Eurographics/ACM SIGGRAPH symposium on Geometry processing*, 2004, pp. 175–184.

[52] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, 2013.

[53] 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.*, June 2014.

[54] B. Browatzki and C. Wallraven, "3fabrec: Fast few-shot face alignment by reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6110–6120.

[55] X. Liu, H. Wang, J. Zhou, and L. Tao, "Attention-guided coarse-to-fine network for 2d face alignment in the wild," *IEEE Access*, vol. 7, pp. 97 196–97 207, 2019.

[56] 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.

[57] 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. 6971–6981.

[58] G. Ghiasi and C. Fowlkes, "Using segmentation to predict the absence of occluded parts," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 22.1–22.12.

[59] 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.

[60] Z. Feng, G. Hu, J. Kittler, W. Christmas, and X. Wu, "Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3425–3440, Nov 2015.

[61] J. Zhang, M. Kan, S. Shan, and X. Chen, "Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2016.

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

[63] Y. Wu, C. Gou, and Q. Ji, "Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, July 2017.

[64] 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.*, July 2017.

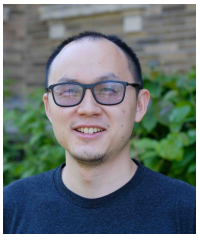
[65] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019.

[66] 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.
[67] A. Kumar and R. Chellappa, "Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 430–439.
[68] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 360–368.
[69] 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.*, 2019.
[70] 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. 8236–8246.

Chunze Lin received the B.S. degree in engineering from Ecole Centrale de Nantes, France and the M.Eng degree in control science and engineering from the department of Automation, Tsinghua University, China. He is currently a research scientist at Sensetime. His research interests include computer vision, pattern recognition and deep learning.

Beier Zhu received the B.S and the M.Eng degree in electrical engineering from Tsinghua University, China. He is currently a Ph.D. candidate with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include computer vision and causal inference.

Quan Wang received the B.S. degree in electronic engineering from Tsinghua University, China. He is currently a research scientist at Sensetime. His research interests include computer vision and 3D reconstruction.



Renjie Liao is a visiting faculty researcher at Google Brain. He graduated with a Ph.D. from the machine learning group at the University of Toronto. During his Ph.D., he also worked as a research scientist at Uber Advanced Technologies Group and was affiliated with the Vector Institute. He obtained his M.Phil. and B.Eng. from the Chinese University of Hong Kong and Beihang University respectively.

Chen QIAN currently the Executive Research Director of SenseTime, is responsible for leading the team in AI content generation, and end-edge computing research in 2D and 3D scenarios. The technology is widely used in the Top 4 mobile companies in China, APPs both home and abroad in augmented reality, video sharing and live streaming, vehicle OEMs and smart industry. He has published dozens of papers on top conferences and journals, e.g. CVPR, ICCV, ECCV and PAMI with more than 3000 citations. He has also led the team to achieve the 1st place in the competition of Face Identification and Face Verification in Megaface Challenge.

Jiwen Lu (M'11-SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and intelligent robotics, where he has authored/co-authored over 270 scientific papers in these areas. He serves the Co-Editor-of-Chief of the Pattern Recognition Letters, an Associate Editor of the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Biometrics, Behavior, and Identity Science, and the Pattern Recognition journal. He also serves as the General Co-Chair of IEEE ICME'2022, and the Program Co-Chair of IEEE FG'2023, IEEE VCIP'2022, IEEE AVSS'2021 and IEEE ICME'2020. He is an IAPR Fellow.

Jie Zhou (M'01-SM'04) received the BS and MS degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and CVPR. He is an associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence and two other journals. He received the National Outstanding Youth Foundation of China Award. He is an IAPR Fellow.