Machine Intelligence Research
www.mi-research.net

21(3), June 2024, 514-525
DOI: 10.1007/s11633-023-1465-1

Structural Dependence Learning Based on Self-attention for Face Alignment

Biying Li[1,2] Zhiwei Liu[1] Wei Zhou[3] Haiyun Guo[1]
Xin Wen[4] Min Huang[2] Jinqiao Wang[1,2]

[1] Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100083, China

[3] Alpha (Beijing) Private Equity, Beijing 100083, China

[4] School of Computer Science, National University of Defense Technology, Changsha 410073, China

Abstract: Self-attention aggregates similar feature information to enhance the features. However, the attention covers nonface areas in face alignment, which may be disturbed in challenging cases, such as occlusions, and fails to predict landmarks. In addition, the learned feature similarity variance is not large enough in the experiment. To this end, we propose structural dependence learning based on self-attention for face alignment (SSFA). It limits the self-attention learning to the facial range and adaptively builds the significant landmark structure dependency. Compared with other state-of-the-art methods, SSFA effectively improves the performance on several standard facial landmark detection benchmarks and adapts more in challenging cases.

Keywords: Computer vision, face alignment, self-attention, facial structure, contextual information.

Citation: B. Li, Z. Liu, W. Zhou, H. Guo, X. Wen, M. Huang, J. Wang. Structural dependence learning based on self-attention for face alignment. *Machine Intelligence Research*, vol.21, no.3, pp.514–525, 2024. http://doi.org/10.1007/s11633-023-1465-1

1 Introduction

Face alignment aims to identify the locations of the facial landmarks, e.g., lips corners, eye corners and nose tips in images or videos. It is a crucial preprocessing step for tasks such as face recognition and facial age estimation and has drawn a surge of interest from both industry and academia.

Self-attention[1] can enhance a feature by aggregating the information from the location where there is a similar feature with it so that contextual information is involved. However, when applying it to the face alignment network, we observe that the distribution of features among all the locations usually has a small variance. The attention of a single feature may cover the whole image space, including the irrelevant areas, as shown in the third row of Fig. 1(b). In addition, when meeting challenging cases such as occlusion, the feature similarity can be easily disturbed by the abnormal appearance changes in the occlusion area. In this situation, it is significant to use a facial structure prior to constraining the landmark structure in the shape distribution of natural faces. To this end, we propose a facial structure prior loss that limits self-attention learning within the scope of facial structure to concentrate on building the dependencies of different landmarks in the training process. As Fig. 1(c) shows, with our proposed facial structural prior, the attention map shows less response on the background, and the network deals with the occlusion case successfully, as the yellow squares mark. Extensive experiments on WFLW, 300W, AFLW and COFW68 demonstrate the effectiveness and robustness of our structural dependence learning based on self-attention for face alignment (SSFA).

In summary, our main contributions are as follows: 1) We explore the problem of using the self-attention mechanism on the face alignment network, which is learning common attention maps that do not focus on the most meaningful areas. 2) We further propose a facial structure prior loss to limit the self-attention learning within the scope of facial structure to concentrate on building the dependencies of different landmarks, leading to a better generalization ability on challenging cases. 3) Our proposed method SSFA effectively improves the face alignment performance on the WFLW, 300W, COFW68 and AFLW datasets.

Research Article
Manuscript received on November 1, 2022; accepted on July 21, 2023; published online on March 20, 2024
Recommended by Associate Editor Jinjun Chen
Colored figures are available in the online version at https://link.springer.com/journal/11633

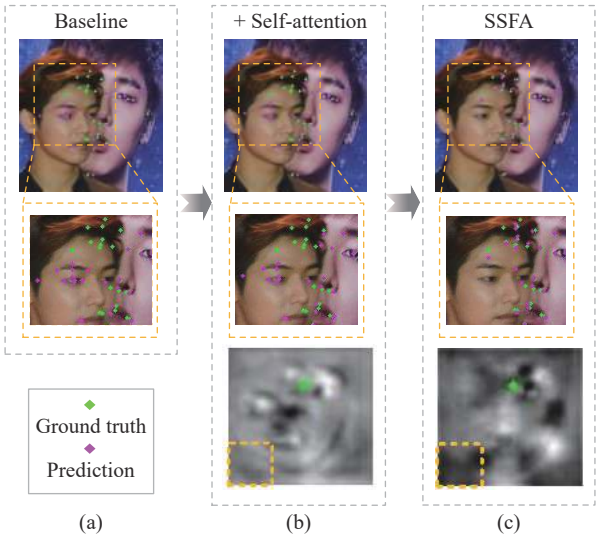

Fig. 1 Landmark prediction results and self-attention heatmap visualization. (a) is the result of the baseline. (b) adds a self-attention module. (c) is the result of SSFA. The top two rows show the landmark prediction results, the green points are ground truth and the magenta points are predictions. The bottom row visualizes the contextual dependencies for the feature at the location of the green point in the facial features. The figures are chosen from WFLW[2] dataset. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

2 Related work

2.1 Face alignment

Recent deep learning based face alignment methods have made great progress. They are mainly categorized into coordinate regression based[3–5] and heatmap regression based[6–8]. The former implicitly learn the statistical characteristics of face shape. However, they ignore the detailed facial structure information, resulting in unsatisfactory detection accuracy.

Heatmap regression-based methods outperform coordinate regression-based methods. They ordinarily use convolutional neural networks (CNNs) to predict one heatmap for one facial landmark and predict the landmark location according to the heatmap response value. Stack hourglass[9] was originally proposed to estimate human pose and then adopted as a backbone by many heatmap regression-based face alignment methods. Liu et al.[7] used a four-stage stack hourglass network and proposed a novel probabilistic model to search for better ground truth while training. HRNet[10] and its improved version HRNet-v2[11] are likewise competitive backbone alternatives in both human pose estimation and face alignment. In this paper, we adopt HRNet-v2 as the backbone to take advantage of its multi-resolution feature extraction and fusion strategy. Specifically, our proposed SSFA does not rely on HRNet-v2. Instead, it is compatible with the backbones of any other heatmap regression-based methods that are end-to-end with CNNs.

2.2 Self-attention

The idea of attention is widely applied in many fields from natural language processing to computer vision, e.g., person re-ID[12] and scene segmentation[13]. It is an important type of attention that can model long-range dependence. The concept of self-attention was first proposed by Vaswani et al.[14] to solve translation tasks. Afterward, Wang et al.[1] combined self-attention with the non-local idea in computer vision, so that distant pixels could contribute to the filtered response according to the pixel feature similarities as well. Later, many methods developed different variants of non-local blocks[12, 15, 16]. Woo et al.[15] used self-attention to improve the performance on the MS-COCO[17] and VOC[18] datasets. Cao et al.[16] reduced the non-local module to learn query-independent pixel-wise relation. LGFA[19] also adopts a self-attention mechanism in face alignment. In contrast, it uses self-attention stage by stage to guide further self-attention, while our SSFA proposes a facial structure prior to guiding self-attention learning directly, which is more concise and effective.

SSFA aims to address the lack of global facial structure information in heatmap regression-based methods. Jiang et al.[20] noted that the self-attention mechanism may notice some redundant long-range dependencies. We presume that for one landmark, the interdependency within the scope of facial structure matters more. Therefore, we use the self-attention mechanism to model the long-range dependencies and propose a facial structure prior loss to suppress the dependencies of landmarks on irrelevant areas to improve the effectiveness.

3 Method

In this section, we introduce the proposed structural dependence learning based on self-attention for face alignment (SSFA), as Fig. 2 illustrates. SSFA consists of a heatmap regression-based backbone and a facial structure prior guided self-attention module. We introduce the framework structure in Section 3.1. Section 3.2 shows how to inject the self-attention module. Section 3.3 describes the proposed facial structure prior loss.

3.1 Framework structure

As shown in Fig. 2, we take HRNet-v2[11] as the backbone. HRNet-v2 consists of a stem, a CNN-based feature extraction module and a head. The stem extracts features preliminarily and outputs features with $\frac{1}{4}$ resolution of the input. The feature extraction module includes parallel multi-resolution convolutions and cross-resolution fusions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$. Denote the input as $\{X, Y\}$, where X is the input and Y is the ground truth. We

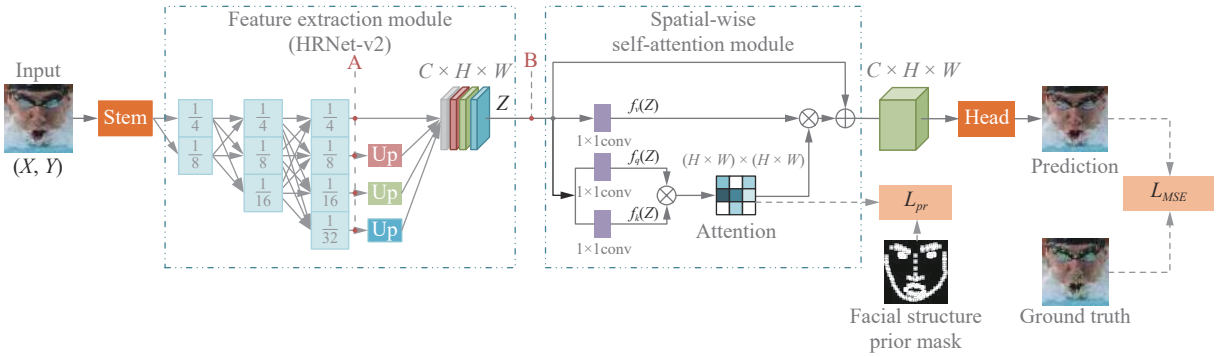

Fig. 2 Framework of the SSFA. We adopt HRNet-v2 as the backbone, which consists of a stem, a feature extraction module and a head. "Up" means upsampling. The self-attention module is added after the feature extraction module. The facial structure prior mask generated according to the ground truth is used to guide the self-attention learning by optimizing the proposed facial structure prior loss L_{pr}. L_{MSE} denotes the MSE loss of the landmark predictions. The figures are chosen from WFLW[2] dataset.

use Z for the multi-resolution features at location B, $Z = [Z_1, Z_2, Z_3, Z_4]$. Z_{1-4} refers to heatmaps at resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ of the input image respectively. The extracted multi-resolution features are sent to the head to predict landmark heatmaps.

To capture the global contextual information from these multi-resolution features explicitly, we insert a self-attention module into the backbone. We compare the performance of several insertion methods and consider their effectiveness and efficiency. Accordingly, adding the self-attention module to location B between the feature extraction module and the head performs best (It is introduced in Section 3.2). In addition, we use a facial structure prior mask to guide the self-attention module to learn.

3.2 Self-attention module

Since convolutional operations only have local receptive fields, the structural relations between distant landmarks are ignored. To overcome this, SSFA adds a self-attention module to introduce contextual information and global dependencies. Furthermore, we consider two types of self-attention named spatial-wise and channel-wise to explore the global contextual information of the landmark heatmaps in different views, and choose the spatial-wise information according to the experiment.

The $C \times H \times W$ feature Z output from the multi-resolution feature extraction module is sent into the spatial-wise self-attention module, as shown in Fig. 2. C is the channel number, H is the height and W is the width. The feature is allocated in three ways. The bottom way further sends the feature to 1×1 convolutional operations separately and obtains two $C' \times H \times W$ features. Then they are reshaped to $C' \times (H \times W)$, indicated as $f_k(Z)$ and $f_q(Z)$, and the $(H \times W) \times (H \times W)$ spatial-wise self-attention matrix $Sa_{sp}(Z)$ is calculated as (1) shows:

$$Sa_{sp}(Z) = f_q^{\mathrm{T}}(Z) \times f_k(Z). \tag{1}$$

C' is a designated channel number smaller than C for saving computational cost. The feature in a middle way is transformed by a 1×1 convolution and then reshaped to $C \times (H \times W)$ indicated as $f_v(Z)$. The calculation of the spatial-wise self-attention $feat_{sp}(Z)$ follows the non-local method[1]:

$$feat_{sp}(Z) = Z + \gamma \times (f_v(Z) \times (Sa_{sp}(Z)))_{C \times H \times W} \tag{2}$$

where the multiplication of $Sa_{sp}(Z)$ and $f_v(Z)$ is resized to $C \times H \times W$. γ is a learnable weight. The summation in (2) is element-wise.

3.3 Facial structure prior loss

Although the self-attention mechanism brings global contextual dependencies, the innate face structure is still not involved. In addition, we find that the correlation response on the background is not sufficiently weak, which could introduce disturbance, as illustrated in Fig. 1.

Considering these shortcomings of the self-attention mechanism, we propose a facial structure prior loss to supervise the learning of the self-attention module. The loss is based on such an intuition: On feature maps, the most valuable locations in terms of facial structure are the landmark locations and their neighboring areas.

Here we take the single-sample-training process as an example. We denote the ground truth of the sample as Y. Y is the set of N facial landmarks, $Y = [a^{(j)}, b^{(j)}]_N$. $a^{(j)}$ and $b^{(j)}$ represent the xy coordinate of the j-th landmark, $j \in [0, N)$. For each coordinate $[a^{(j)}, b^{(j)}]$, we generate an $H \times W$ submask $mask_j(a, b)$ with a Gaussian distribution:

$$mask_j(a,b) = \begin{cases} \mathrm{e}^{-\frac{(a-a^{(j)})^2+(b-b^{(j)})^2}{2\sigma^2}}, \\ \qquad \text{if } |a-a^{(j)}|, |b-b^{(j)}| \le R \\ 0, \qquad \text{if } |a-a^{(j)}|, |b-b^{(j)}| > R \end{cases} \tag{3}$$

where R is the radius of a Gaussian distribution. The

mean of the Gaussian distribution is the landmark ground truth, and the standard deviation σ is a hyperparameter. The facial structure prior $mask(a,b)$ is the element-wise summation of all the submasks:

$$mask(a,b) = \sum_j mask_j(a,b). \tag{4}$$

We define the landmark neighborhood as

$$NEI = \{[m,n] | mask(m,n) > 0\}. \tag{5}$$

The non-neighborhood is defined as $\overline{NEI} = \{[p,q] | mask(p,q) = 0\}$. An L1 loss L_{pr} is used in our method to restrain the accumulated self-attention value, which evaluates the similarity of the landmark neighborhood features and the non-neighborhood features:

$$L_{pr}(NEI, Sa_{sp}(Z)) = \frac{1}{|NEI|} \sum_{m,n} \sum_{p,q} mask(m,n)(Sa'_{sp}(Z)[m,n,p,q]) \tag{6}$$

where $[m,n] \in NEI$, $[p,q] \in \overline{NEI}$. The $(H \times W) \times (H \times W)$ spatial-wise self-attention $Sa_{sp}(Z)$ is reshaped to $H \times W \times H \times W$ as $Sa'_{sp}(Z)$, and $[m,n,p,q]$ is the index to $Sa'_{sp}(Z)$. The mask value $mask(m,n)$ of the corresponding landmark neighborhood is adopted as the weight. The closer to the landmarks and their neighbours, the larger the weight is.

The visual interpretation of (6) is shown in Fig. 3. First, the $H \times W$ facial structure prior mask is generated centered on facial landmarks, as Fig. 3 (a) shows. The mask is processed with binarization to obtain the landmark neighborhood NEI in (5) (here we use the 2D map to visualize the landmark neighborhood as the white area). Then the $H \times W$ NEI map is reshaped to $(H \times W) \times 1$ as the indication for the following steps: for any $[m,n]$ in NEI, the corresponding index is $((m-1) \times W + n)$.

The calculation of L_{pr} is shown in Fig. 3(b). The $(H \times W) \times (H \times W)$ spatial-wise self-attention matrix Sa_{sp} is reshaped to $H \times W \times H \times W$ Sa'_{sp}. We select a landmark neighborhood feature at $[m,n]$ for illustration. Its corresponding self-attention features are at the $((m-1) \times W + n)$-th row in Sa_{sp}, and are reshaped to an $H \times W$ feature map $Sa'_{sp}(m,n,:,:)$.

Then we obtain the response on the non-neighborhood area of the $Sa'_{sp}(m,n,:,:)$ according to $\overline{NEI}$. To constrain the response on the non-neighborhood area and avoid irrelevant information, we calculate the L1 summation of the non-neighborhood response, weighted with the corresponding $mask(m,n)$. The closer $[m,n]$ is to a landmark location, the larger $mask(m,n)$ is. All the non-neighborhood self-attention values of neighborhood features are weighted and added together to finally obtain the facial structure prior loss L_{pr}.

Meanwhile, following the heatmap-based methods, we adopt MSE loss for the facial landmark detection results Y_{pred}, as shown in (7) and (8).

$$Y_{pred} = F_{head}(feat_{sp}) \tag{7}$$

$$L_{MSE}(Y, Y_{pred}) = \frac{1}{N} \sum_j \left\| Y^{(j)} - Y^{(j)}_{pred}) \right\|^2_{L_2} \tag{8}$$

where F_{head} denotes the head block of HRNet-v2. In addition, we use a weighting parameter β as (9) indicates, to balance the two losses:

$$Loss = L_{MSE}(Y, Y_{pred}) + \beta \times L_{pr}(NEI, Sa_{sp}(Z)). \tag{9}$$

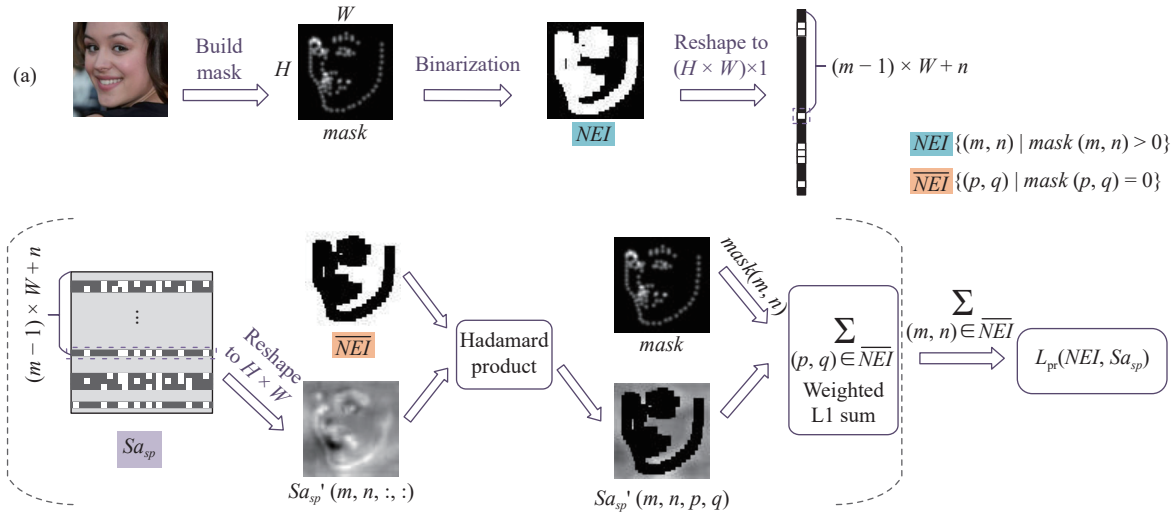

Fig. 3 The visual interpretation of (6). (a) The generalization process of the facial structure mask. (b) The calculation process of the facial structure prior loss L_{pr}. The figures are chosen from WFLW[2] dataset.

4 Experiment

4.1 Datasets, evaluation metrics and implementation details

4.1.1 Datasets

WFLW[2] is built on the WIDER Face[21]. It has a 7 500 image training set and a 2 500 image testing set. There are 98 manually annotated landmarks for every face. It also contains several testing subsets that belong to different topics: large pose, illumination, occlusion, make-up, blur and expression.

300W is made up of several datasets: HELEN[22], AFW[23], IBUG, LFPW[24] and XM2VTS. The training set contains 3 148 images, including the full set of AFW and the training subsets of HELEN and LFPW. The evaluations are on three protocols: full test set, challenging subset and common subset.

AFLW[25] contains 20 000 images for training, and 4 386 images for testing, providing 19 landmarks for each face.

COFW68[26] is the re-annotated for 68 landmarks in the COFW testing dataset. It is only used for tests. In our experiment, we train SSFA on the 300W dataset (68 landmarks) and test it on COFW68 to evaluate the cross-data performance.

4.1.2 Evaluation metrics

Referring to other related work, we evaluate our algorithm using the common standard normalized mean error (NME), failure rate (FR), and area under the curve (AUC) for quantitative analysis. NME is defined as the average point-to-point Euclidean distance between the ground truth Y and the predicted landmarks Y_{pred}, which is normalized by the reference distance d:

$$NME = \frac{1}{N}\sum_{j}\frac{\left|Y^{(j)} - {Y_{pred}}^{(j)}\right|}{d} \tag{10}$$

where $Y^{(j)}$ is the j-th ground truth landmark coordinate in Y, and $Y_{pred}^{(j)}$ is the j-th predicted landmark coordinate in Y_{pred}. N is the number of landmarks of each face on the image. For the normalization factor d, we provide inter-ocular distance (the distance of a person′s outer eye corners) for datasets WFLW, 300W, and COFW68, and image size for AFLW.

FR indicates the percentage of the test images of whose NME is higher than a given threshold. AUC is the area under the CED curve, which indicates the fraction of test images whose NME is less than or equal to the thresholds from zero to a given value.

4.2 Implementation details

The inputs are cropped and resized to 256×256. The preprocessing operations contain a ± 10 degree random rotation, 0.75–1.25 times scaling, and random flipping. The RGB value is normalized. We adopt Adam as the optimizer. The learning rate starts from 2×10^{-3} and drops to 2×10^{-4} and 2×10^{-5} at epochs 40 and 55, respectively. The model is trained for 80 epochs with a batch size of 16. The training is implemented with Python 3.7, PyTorch 1.5.0, and two 1080Ti GPUs. In the test stage, we adopt the flipping strategy to evaluate the average prediction results of the input image and its flipped image.

4.3 Ablation study on WFLW

4.3.1 Ablation study on self-attention module

First, to explore the best insertion fashion of the self-attention module, and introduce the contextual information, we try two locations (A and B in Fig. 2) to add the spatial-wise self-attention module. As Fig. 2 shows, at the end of the feature extraction module (location B), the outputs of four convolutional branches are upsampled to the $\frac{1}{4}$ input resolution and concatenated. Three insertion methods are evaluated: Adding the self-attention module at location A, at location B, and at A and B simultaneously. The first method sets four self-attention modules separately to the four branches at A, upsamples the features enhanced by self-attention, and concatenates them before sending them to the head block. The upsampling of the features leads to the degeneration of the effect of the self-attention mechanism. In contrast, the feature scale at B is the same as the feature in the landmark prediction stage, so the self-attention module can explore the landmark dependencies more accurately. As the NME results on WFLW in Table 1 show, the performance is the best when the spatial-wise self-attention module is added at B. Therefore, we keep the self-attention module at B in the later experiments.

Table 1 The WFLW NME test results of the self-attention module added at location A, location B, locations A and B , as shown in Fig. 2. Bold is the best.

	Baseline	A	B	A&B
Test(NME)	4.17	4.13	**4.09**	4.11

Second, we explore different self-attention modules such as spatial-wise and channel-wise[13], and evaluate the effectiveness of self-attention in different views. We try several combinations of modules, including a single module and parallel/series combinations of the channel-wise and spatial-wise self-attention modules. For the parallel case, the feature at position B is modified by spatial-wise and channel-wise self-attention modules separately and then added together. For the serial case, the feature at position B is sent to the two kinds of self-attention modules sequentially. As Table 2 shows, on WFLW, the single spatial-wise self-attention module performs best. The additional channel-wise module causes degeneration, which as we consider has great relations with the property of channel-wise operation. Channel-wise self-atten-

Table 2 The WFLW NME test results of different sets of the spatial-wise (sp_att) and channel-wise(ch_att) self-attention at location B, as shown in Fig. 2. Bold is the best.

	Test	Largepose	Expression	Illumination	Makeup	Occlusion	Blur
Baseline(pretrained)	4.17	7.33	4.61	4.14	4.05	5.03	4.80
ch_att	4.26	7.34	4.54	4.23	4.28	5.18	4.80
ch_att+sp_att(series)	4.24	7.29	4.49	4.17	4.31	5.16	4.76
sp_att+ch_att (parallel)	4.10	7.08	**4.39**	4.07	3.99	4.95	4.67
sp_att+ch_att (series)	4.12	7.12	4.47	4.06	3.97	4.92	4.71
sp_att	**4.09**	**7.06**	4.43	**4.04**	**3.96**	**4.90**	**4.65**

tion compresses the spatial information like global pooling, as CBAM[15] manipulations, which abandon the beneficial detailed spatial information. However, one of the significant characteristics of the human face is the special confirmed spatial structure. Therefore, we choose spatial-wise self-attention to explore the landmark interdependency.

4.3.2 Ablation study on the facial structure prior mask

Although self-attention introduces the interdependency of facial landmarks, there is a moderate response in the irrelevant regions due to a lack of facial structure guidance. It disturbs the landmark localization, as Fig. 1 shows. To guide the self-attention to concentrate more on the facial structure, we employ a facial structure prior mask. Given a facial image with landmark ground truth, we obtain the facial structure prior mask by building Gaussian distributions centered around each landmark, as (3) and (4) show. The facial structure prior loss is the L1 summation of the self-attention map response in the irrelevant regions of all the landmarks and their neighbor points as (6) indicates. In addition, we adopt a weighting parameter β to balance the losses. As Table 3 shows, we explore the value of β at 0.1, 0.01 and 0.001, the Gaussian distribution standard deviation σ at 1, 2 and 3, and the range R for Gaussian distribution at 3 and 5. The

Table 3 The NME test results of training with masks of different standard deviations σ, ranges R, and weights β on WFLW. Bold is the best.

	β=0.001		β=0.01		β=0.1	
	R=3	R=5	R=3	R=5	R=3	R=5
σ=1	4.07	4.07	**4.05**	4.06	4.12	4.14
σ=2	4.09	4.10	4.08	4.09	4.11	4.11
σ=3	4.09	4.12	4.10	4.09	4.17	4.15

NME performance is the best when $\sigma = 1$, $R = 3$ and $\beta = 0.01$. The visualization of masks with different σ and R is as shown in Fig. 4.

4.4 Comparison with the state of the art

4.4.1 Results on WFLW

We compare SSFA with previous methods on WFLW in Table 4. The baseline here is the recurrent result of HRNet-v2, and the NME of SSFA is 0.12 better than the baseline on the test set. Both DeCaFA[27] and AC-DC[30] are trained with multiple datasets while SSFA uses only one. AWING[28] is dedicated to the loss function design. DAG[6] focuses on the different global and local features and uses adaptive graph learning. LGFA[19] adopts two

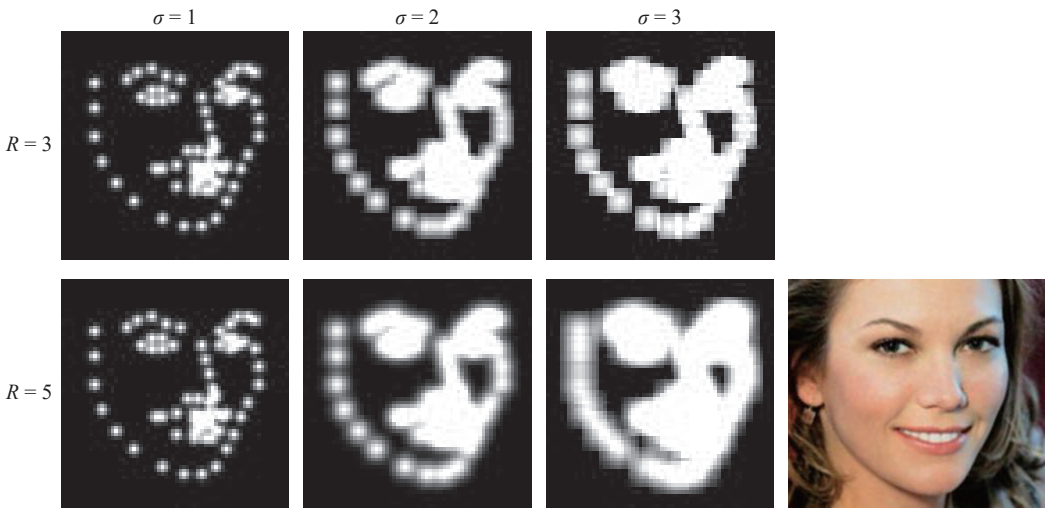

Fig. 4 Visualization of a sample and its facial structure prior masks with different standard deviations σ and ranges R . The figures are chosen from WFLW[2] dataset.

Table 4 Face alignment results on WFLW

Metric	Method	Test	Pose	Express.	Illum.	Makeup	Occlusion	Blur
Mean error (%)	DeCaFA[27]	4.62	8.11	4.65	4.41	4.63	5.74	5.38
	AWING[28]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	LUVLi[29]	4.37	–	–	–	–	–	–
	AC-DC[30]	4.49	7.76	4.45	4.35	4.25	5.57	5.21
	DAG[6]	4.21	7.36	4.49	4.12	4.05	4.98	4.82
	LGFA[19]	4.28	7.63	4.33	4.16	4.27	5.33	4.95
	SLPT[31]	4.14	–	–	–	–	–	–
	GlomFace[32]	4.81	8.17	–	–	–	5.14	–
	RHT[33]	**3.96**	6.77	4.38	4.02	4.03	4.77	**4.58**
	DSLPT[34]	4.01	6.87	**4.29**	3.99	**3.86**	4.79	4.66
	CHS[35]	4.04	**6.76**	4.33	**3.98**	3.87	**4.71**	4.64
	Baseline(ours)	4.17	7.33	4.61	4.14	4.05	5.03	4.80
	Baseline+sp_att(ours)	4.09	7.06	4.43	4.04	3.96	4.90	4.65
	SSFA(ours)	4.05	6.96	4.37	4.04	3.92	4.84	4.61
FR @0.1 %	DeCaFA[27]	4.84	21.40	3.73	3.22	6.15	9.26	6.61
	AWING[28]	2.84	13.50	2.23	2.58	2.91	5.98	3.75
	LUVLi[29]	3.12	–	–	–	–	–	–
	AC-DC[30]	4.29	17.30	2.69	2.45	4.66	9.20	5.82
	DAG[6]	3.04	15.95	2.86	2.72	1.45	5.29	4.01
	LGFA[19]	3.44	16.26	2.23	2.58	2.91	7.47	4.40
	SLPT[31]	2.72	–	–	–	–	–	–
	GlomFace[32]	3.77	17.48	–	–	–	6.73	–
	DSLPT[34]	2.52	13.19	2.23	2.44	**0.97**	4.89	3.49
	CHS[35]	**1.80**	**9.51**	**1.59**	**1.72**	1.46	**3.13**	**2.46**
	SSFA(ours)	2.60	12.47	1.81	2.40	2.19	4.83	3.57
AUC @0.1 %	DeCaFA[27]	56.30	29.20	54.60	57.90	57.50	48.50	49.40
	AWING[28]	57.20	31.20	51.50	57.70	57.10	50.20	51.20
	LUVLi[29]	57.70	–	–	–	–	–	–
	AC-DC[30]	57.50	31.50	56.60	58.70	58.30	49.50	51.10
	DAG[6]	58.90	31.50	56.60	59.50	60.30	52.30	53.30
	LGFA[19]	58.70	31.70	58.10	59.90	58.80	50.40	52.90
	SLPT[31]	59.50	–	–	–	–	–	–
	DSLPT[34]	**60.70**	35.30	58.60	61.40	**62.30**	53.50	**54.90**
	CHS[35]	60.15	**35.52**	57.92	60.80	61.55	**54.03**	54.62
	SSFA(ours)	59.60	35.04	**59.21**	**61.70**	60.60	53.92	54.86

stacked hourglass modules, uses self-attention stage by stage to explore contextual information, and guides further self-attention. Both LGFA and our SSFA use self-attention. However, we propose the facial structural prior to guide the learning of self-attention directly, which is concise and effective. SLPT[31] and its improved version DSLPT[34] both utilize a transformer to arrange local patches and obtain the inherent relation, where the transformer framework is time-consuming compared to SSFA. GlomFace[32] concentrates on the occlusion data using hierarchical facial information, while our SSFA gets 0.3 better on the occlusion subset of WFLW. RHT[33] introduces the face with ground truth landmark heatmaps as the reference face to learn the facial structure commonality. This method requires landmark information at first, which has requirements for the application scenarios. CHS[35] achieves better results than SSFA and other state-of-the-art methods. However, the model complexity of CHS is much higher than that of SSFA, as stated in Section 4.5.

4.4.2 Results on 300W, AFLW and COFW68

We also carry out experiments on the 300W and AFLW datasets separately, as shown in Tables 5 and 6. We also evaluate the performance on the COFW68 of the model trained on the 300W as the cross-dataset testing to further check out the effectiveness, as shown in Table 7.

Table 5 Face alignment results (NME) on 300W

Methods	Common	Challenging	Full
DAN[36]	3.19	5.24	3.59
DSRN[37]	4.12	9.68	5.21
SAN[38]	3.34	6.60	3.98
LAB(w/B)[2]	2.98	5.19	3.49
DeCaFA[27]	2.93	5.26	3.39
HSLE[39]	3.21	5.69	3.70
DAG[6]	2.62	4.77	**3.04**
3FabRec[40]	3.36	5.74	3.82
LUVLi[29]	2.76	5.16	3.23
LGSA[19]	2.92	5.16	3.36
SLPT[19]	2.75	4.90	3.17
RHT[33]	2.34	**4.37**	**2.74**
DSLPT[34]	2.57	4.69	2.98
CHS[35]	**2.52**	4.48	2.91
SSFA(ours)	2.78	4.90	3.21

Table 6 Face alignment results (NME) on AFLW

Methods	Pretrained	NME
LAB(w/B)[2]	–	1.85
Wing[8]	Y	1.47
AWing[28]	N	1.53
LUVLi[29]	N	1.39
PIPNet-18[41]	Y	1.48
DTLD-s[42]	N	1.39
RHT[33]	N	1.87
DSLPT[34]	Y	1.36
CHS[35]	N	**0.96**
SSFA(ours)	N	1.40

Among the state-of-the-art methods, SSFA achieves competitive results on these datasets. DAG[6] adopts a graph convolutional neural network for better performance, however, it also introduces more parameters. LUVLi[29] performs competitively in 300W and AFLW, however, SSFA performs much better on WFLW. Considering the complexity of these datasets, SSFA is more adaptive to difficult data. SLPT[31] and DSLPT[34] both take the face image as a set of landmark local patches and put them into the transformer, while DTLD[42] uses cascaded decoders to refine the regression of the landmarks. All three methods mention the effect of self-attention. However, SSFA not only takes advantage of self-attention to design a framework to capture the global contextual information, but also proposes the facial structure prior mask to guide the self-attention learning and further solves its problem of the disturbance from irrelevant areas.

Table 7 Cross-dataset evaluation results (NME) on COFW68

Method	NME_{ocular}	FR@0.1(%)
LAB(w/B)[2]	4.62	2.17
SLD[6]	4.22	0.39
GlomFace[32]	4.21	0.79
SDFL[43]	4.18	**0**
SLPT[19]	4.11	0.59
CHS[35]	**3.78**	–
SSFA(ours)	4.03	0.40

4.5 Model complexity

The comparison of the model complexity with state-of-the-art is shown in Table 8. SSFA is based on the HRNet-v2 framework, with the additional self-attention mechanism and the facial structure prior mask. Compared to HRNet-v2, the additional design in SSFA only slightly increases the FLOPs and parameters. Furthermore, the complexity of SSFA is lower than that of the compared SLPT, CHS and DSLP methods. It is worth mentioning that although the face alignment performance of CHS[35] is better than that of other methods, it requires a relatively high computational load.

Table 8 Model complexity

Method	#Params(M)	FLOPs(G)
LAB(w/B)[2]	9.66	18.85
AWING[28]	24.15	26.8
HRNet-v2[11]	9.66	4.75
SLPT[19]	13.19	6.12
CHS[35]	154.04	41.69
DSLPT[34]	19.35	7.83
SSFA(ours)	9.67	4.99

4.6 Visualization

In this section, we visualize some landmark predictions in the challenging cases and their corresponding self-attention heatmaps on WFLW, as Fig. 5 shows. For each person, the pictures in the top row are the results of adding only a self-attention module to the baseline, and the pictures in the bottom row show the results of SSFA.

The first man's face is half illuminated and half dark,

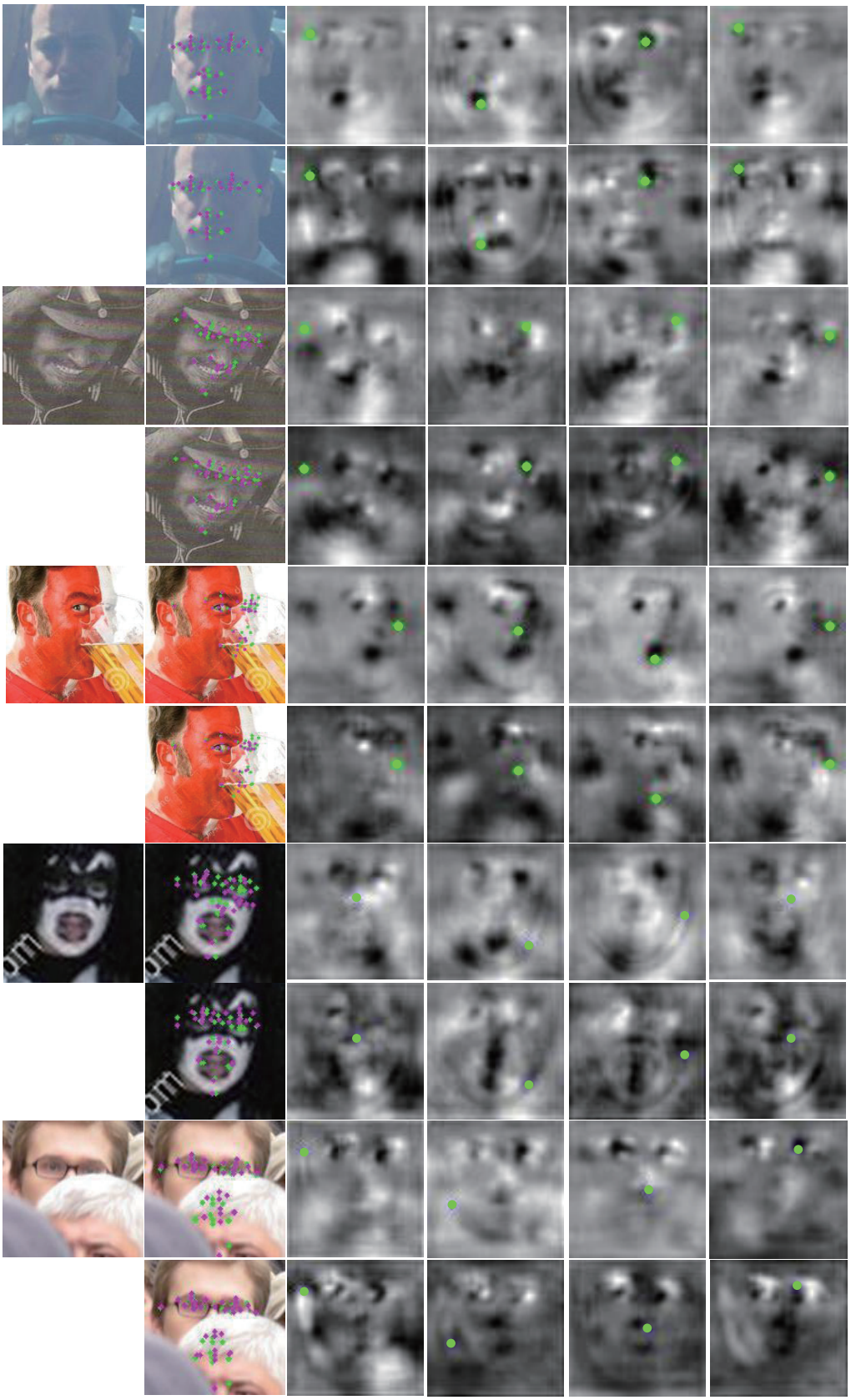

Fig. 5 Visualization of samples and their heatmaps on WFLW[2]. Green points are the ground truth and magenta points are predictions. The green point on a heatmap represents the position of the landmark whose self-attention heatmap is calculated. For each person, the pictures in the top row show the performance of training with the self-attention module only, while the pictures below refer to the results of SSFA. Better viewed in color. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

which disturbs the self-attention, and results in the predicted deviation to the lighter half. SSFA constrains the landmark prediction by the facial landmark interdependency so that the disturbance of illumination is overcome. Similarly, the colors of red and white on the third man's face and the black and white color of the fourth person also influence the self-attention mechanism. Both the second person's and the fifth person's images show the occlusion case. The former's hat covers the eyes and eyebrows and the latter's nose and mouth are occluded by the foreground. Compared to the prediction on the top, SSFA achieves more structural and robust predictions.

We also visualize the self-attention heatmaps of some landmarks in Fig. 5. The green point on a heatmap marks the location of the landmark feature of which self-attention is calculated. The brighter pixel shows a higher similarity to the marked landmark feature. According to the visualization, heatmaps learned by SSFA have weaker responses on the nonface areas such as the background, the nonface skin area, and the person's clothes. That is with the guidance of facial structure dependencies, the self-attention module can be constrained to learn within the scope of the facial structure better.

5 Conclusions

In this paper, we propose a structural dependence learning based on self-attention for face alignment (SSFA) method. Considering the shortcomings of heatmap regression-based methods, we adopt self-attention to capture global contextual information. Furthermore, we propose a facial structure prior loss to guide self-attention, which helps to focus on areas within the scope of facial structure. The evaluation results on several popular benchmarks show that SSFA can effectively improve the performance of face alignment and deal with challenging situations such as illumination and occlusion.

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2021YFE0205700), the National Natural Science Foundation of China (Nos. 62076235, 62276260 and 62002356), sponsored by the Zhejiang Lab (No. 2021KH0AB07) and the Ministry of Education Industry-University Cooperative Education Program (Wei Qiao Venture Group, No. E1425201).

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] X. L. Wang, R. Girshick, A. Gupta, K. M. He. Non-local neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 7794–7803, 2018. DOI: 10.1109/CVPR.2018.00813.

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

[3] Y. Sun, X. G. Wang, X. O. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, pp. 3476–3483, 2013. DOI: 10.1109/CVPR.2013.446.

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

[5] Z. P. Zhang, P. Luo, C. C. Loy, X. O. Tang. Facial landmark detection by deep multi-task learning. In *Proceedings of the 13th European Conference on Computer Vision*, Zürich, Switzerland, pp. 94–108, 2014. DOI: 10.1007/978-3-319-10599-4_7.

[6] W. J. Li, Y. H. Lu, K. Zheng, H. F. Liao, C. Lin, J. B. Luo, C. T. Cheng, J. Xiao, L. Lu, C. F. Kuo, S. Miao. Structured landmark detection via topology-adapting deep graph learning. In *Proceedings of the 16th European Conference on Computer Vision*, Glasgow, UK, pp. 266–283, 2020. DOI: 10.1007/978-3-030-58545-7_16.

[7] Z. W. Liu, X. Y. Zhu, G. S. Hu, H. Y. Guo, M. Tang, Z. Lei, N. M. Robertson, J. Q. Wang. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 3462–3471, 2019. DOI: 10.1109/CVPR.2019.00358.

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

[9] A. Newell, K. Y. Yang, Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, pp. 483–499, 2016. DOI: 10.1007/978-3-319-46484-8_29.

[10] K. Sun, B. Xiao, D. Liu, J. D. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 5686–5696, 2019. DOI: 10.1109/CVPR.2019.00584.

[11] J. D. Wang, K. Sun, T. H. Cheng, B. R. Jiang, C. R. Deng, Y. Zhao, D. Liu, Y. D. Mu, M. K. Tan, X. G. Wang, W. Y. Liu, B. Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021. DOI: 10.1109/TPAMI.2020.2983686.

[12] Z. Z. Zhang, C. L. Lan, W. J. Zeng, X. Jin, Z. B. Chen. Relation-aware global attention for person re-identification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 3183–3192, 2020. DOI: 10.1109/CVPR42600.2020.00325.

[13] J. Fu, J. Liu, H. J. Tian, Y. Li, Y. J. Bao, Z. W. Fang, H. Q. Lu. Dual attention network for scene segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 3141–3149, 2019. DOI: 10.1109/CVPR.2019.00326.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. Attention is all you

need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 6000–6010, 2017.

[15] S. Woo, J. Park, J. Y. Lee, I. S. Kweon. CBAM: Convolutional block attention module. In *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 3–19, 2018. DOI: 10.1007/978-3-030-01234-2_1.

[16] Y. Cao, J. R. Xu, S. Lin, F. Y. Wei, H. Hu. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, Republic of Korea, pp. 1971–1980, 2019. DOI: 10.1109/ICCVW.2019.00246.

[17] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár. Microsoft COCO: Common objects in context, [Online], Available: https://arxiv.org/abs/1405.0312.

[18] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010. DOI: 10.1007/s11263-009-0275-4.

[19] P. C. Gao, K. Lu, J. Xue, L. Shao, J. Y. Lyu. A coarse-to-fine facial landmark detection method based on self-attention mechanism. *IEEE Transactions on Multimedia*, vol. 23, pp. 926–938, 2021. DOI: 10.1109/TMM.2020.2991507.

[20] Z. H. Jiang, W. H. Yu, D. Q. Zhou, Y. P. Chen, J. S. Feng, S. C. Yan. ConvBERT: Improving BERT with span-based dynamic convolution. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[21] S. Yang, P. Luo, C. C. Loy, X. O. Tang. WIDER FACE: A face detection benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 5525–5533, 2016. DOI: 10.1109/CVPR.2016.596.

[22] V. Le, J. Brandt, Z. Lin, L. Bourdev, T. S. Huang. Interactive facial feature localization. In *Proceedings of the 12th European Conference on Computer Vision*, Florence, Italy, pp. 679–692, 2012. DOI: 10.1007/978-3-642-33712-3_49.

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

[24] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013. DOI: 10.1109/TPAMI.2013.23.

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

[26] G. Ghiasi, C. C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces, [Online], Available: https://arxiv.org/abs/1506.08347.

[27] A. Dapogny, M. Cord, K. Bailly. DeCaFa: Deep convolutional cascade for face alignment in the wild. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, pp. 6892–6900, 2019. DOI: 10.1109/ICCV.2019.00699.

[28] X. Y. Wang, L. F. Bo, L. Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, pp. 6970–6980, 2019. DOI: 10.1109/ICCV.2019.00707.

[29] A. Kumar, T. K. Marks, W. X. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. M. Liu, C. Feng. LUVLi face alignment: Estimating landmarks′ location, uncertainty, and visibility likelihood. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 8233–8243, 2020. DOI: 10.1109/CVPR42600.2020.00826.

[30] A. Dapogny, K. Bailly, M. Cord. Deep entwined learning head pose and face alignment inside an attentional cascade with doubly-conditional fusion. In *Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition*, Buenos Aires, Argentina, pp. 192–198, 2020. DOI: 10.1109/FG47880.2020.00038.

[31] J. H. Xia, W. W. Qu, W. J. Huang, J. G. Zhang, X. Wang, M. Xu. Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 4042–4051, 2022. DOI: 10.1109/CVPR52688.2022.00402.

[32] C. C. Zhu, X. T. Wan, S. R. Xie, X. Q. Li, Y. Z. Gu. Occlusion-robust face alignment using a viewpoint-invariant hierarchical network architecture. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 11102–11111, 2022. DOI: 10.1109/CVPR52688.2022.01083.

[33] J. Wan, J. Liu, J. Zhou, Z. H. Lai, L. L. Shen, H. Sun, P. Xiong, W. W. Min. Precise facial landmark detection by reference heatmap transformer. *IEEE Transactions on Image Processing*, vol. 32, pp. 1966–1977, 2023. DOI: 10.1109/TIP.2023.3261749.

[34] J. H. Xia, M. Xu, H. M. Zhang, J. G. Zhang, W. J. Huang, H. Cao, S. P. Wen. Robust face alignment via inherent relation learning and uncertainty estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10358–10375, 2023. DOI: 10.1109/TPAMI.2023.3260926.

[35] J. So, Y. Han. Heatmap-guided selective feature attention for robust cascaded face alignment. *Sensors*, vol. 23, no. 10, Article number 4731, 2023. DOI: 10.3390/s23104731.

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

[37] X. Miao, X. T. Zhen, X. L. Liu, C. Deng, V.s Athitsos, H. Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 5040–5049, 2018. DOI: 10.1109/CVPR.2018.00529.

[38] X. Y. Dong, Y. Yan, W. L. Ouyang, Y. Yang. Style aggregated network for facial landmark detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 379–388, 2018. DOI: 10.1109/CVPR.2018.00047.

[39] X. Zou, S. Zhong, L. X. Yan, X. Y. Zhao, J. H. Zhou, Y. Wu. Learning robust facial landmark detection via hierarchical structured ensemble. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, pp. 141–150, 2019. DOI: 10.1109/ICCV.2019.00023.

[40] B. Browatzki, C. Wallraven. 3FaBrec: Fast few-shot face alignment by reconstruction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recog-*

nition, Seattle, USA, pp. 6109–6119, 2020. DOI: 10.1109/CVPR42600.2020.00615.

[41] H. B. Jin, S. C. Liao, L. Shao. Pixel-in-pixel Net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, vol. 129, no. 12, pp. 3174–3194, 2021. DOI: 10.1007/s11263-021-01521-4.

[42] H. Li, Z. D. Guo, S. M. Rhee, S. Han, J. J. Han. Towards accurate facial landmark detection via cascaded transformers. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 4166–4175, 2022. DOI: 10.1109/CVPR52688.2022.00414.

[43] C. Z. Lin, B. Zhu, Q. Wang, R. J. Liao, C. Qian, J. W. Lu, J. Zhou. Structure-coherent deep feature learning for robust face alignment. *IEEE Transactions on Image Processing*, vol. 30, pp. 5313–5326, 2021. DOI: 10.1109/TIP.2021.3082319.

Biying Li received the B. Eng. degree in automation from Xi'an Jiaotong University, China in 2018. She is currently a Ph.D. degree candidate in the Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, China.

Her research interests include 3D face and human understanding, image and video processing, and pattern recognition.

E-mail: libiying2018@ia.ac.cn

ORCID iD: 0000-0001-5125-7832

Zhiwei Liu received the B. Sc. degree in software engineering from Sichuan University, China in 2015, and the Ph.D. degree in pattern recognition and intelligent system from the Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, China in 2020. Currently, he is an assistant professor. He has published several papers on CVPR, AAAI, ACMMM, ECCV, TMM, and TOMM. He is participating in several national projects, including the National Natural Science Foundation of China.

His research interests include 3D face and human understanding, virtual human generation and control, and human-centric AI-generated content.

E-mail: zhiwei.liu@nlpr.ia.ac.cn

Wei Zhou received the B. Eng. degree in software engineering from the Beijing Institute of Technology, China in 2007, the M. Sc. degree in software engineering in Peking University, China in 2010, and is currently a Ph.D. degree candidate in Tsinghua University, China. He serves as Chief Investment Officer of Wuhan Artificial Intelligence Research Institute.

His research interest is intelligent decisions driven by multimodal heterogeneous data.

E-mail: zhouwei@wair.ac.cn

Haiyun Guo received the B. Sc. degree in electronic information science and technology from Wuhan University, China in 2013, and the Ph.D. degree in pattern recognition and intelligent systems from the University of Chinese Academy of Sciences, China in 2018. Currently, she is an associate research fellow at the Institute of Automation, Chinese Academy of Sciences, China.

Her research interests include image and video analysis, multimodal understanding, large-scale model training, and general model design.

E-mail: haiyun.guo@nlpr.ia.ac.cn

Xin Wen received the B. Eng. degree in communication engineering from Chongqing University of Posts and Telecommunications, China in 2016, and the M. Sc. degree in computer technology from the University of Chinese Academy of Sciences, China in 2021. She is currently a Ph.D. degree candidate at the National University of Defense Technology, China.

Her research interests include image processing, pattern recognition and 3D reconstruction.

E-mail: wenxin21@nudt.edu.cn

Min Huang received the B. Sc. and Ph.D. degrees in computer sciences and technology from Wuhan University, China in 2002 and 2007. From 2017 to 2018, she was a visiting scholar with the School of Informatics at the University of Edinburgh, UK. She is currently an associate professor at the School of Artificial Intelligence, University of Chinese Academy of Sciences, China.

Her research interests include machine learning, knowledge engineering and pattern recognition.

E-mail: huangm@ucas.ac.cn

Jinqiao Wang received the B. Eng. degree in mechanical and electronic engineering from Hebei University of Technology, China in 2001, and the M. Sc. degree in mechanical and electronic engineering from Tianjin University, China in 2004. He received the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, China in 2008. He is currently a professor at the Chinese Academy of Sciences, China.

His research interests include pattern recognition and machine learning, image and video processing, mobile multi-media, and intelligent video surveillance.

E-mail: jqwang@nlpr.ia.ac.cn (Corresponding author)

ORCID iD: 0000-0002-9118-2780

Machine Intelligence Research

Citation: B. Li, Z. Liu, W. Zhou, H. Guo, X. Wen, M. Huang, J. Wang. Structural dependence learning based on self–attention for face alignment. *Machine Inteligence Research*, vol.21, no.3, pp.514–525, 2024. https://doi.org/10.1007/s11633–023–1465–1

Articles may interest you