APSIPA Transactions on Signal and Information Processing, 2023, 12, e28

Original Paper

Boundary-Aware Face Alignment with Enhanced HourglassNet and Transformer

Yingxin Li[1,2], Dongmei Niu[1,2]* and Jingliang Peng[1,2]*

[1] *Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, China*
[2] *School of Information Science and Engineering, University of Jinan, China*

ABSTRACT

In this work, we propose a neural network for boundary-aware face alignment. The proposed network is composed of two stages with the first one estimating boundary heatmaps and the second one predicting landmark positions. We build the first stage by enhancing a baseline HourglassNet. Major enhancements include the addition of a CoordConv layer and addition of shallow and deep feature fusion (SDFusion) blocks. For the second stage, we design a subnet that firstly fuses information of the original image, a latent feature from the first stage and the boundary heatmap generated by the first stage, and secondly uses a Transformer to map the fused feature to the landmark coordinates. As shown by experiments, the proposed algorithm achieves state-of-the-art performance on the benchmark datasets.

Keywords: Face alignment, facial landmark detection, boundary heatmap, HourglassNet.

*Co-corresponding authors: Dongmei Niu, ise_niudm@ujn.edu.cn and Jingliang Peng, jingliap@gmail.com.

Received 26 February 2023; Revised 06 April 2023
ISSN 2048-7703; DOI 10.1561/116.00000115

1 Introduction

Face alignment, also known as facial landmark detection (FLD), is to automatically localize a pre-defined set of semantic feature points in a face image. Accurate face alignment is key to a variety of face applications including face recognition and verification, face reenactment, face morphing and so forth. As such, face alignment has attracted intensive research attention with many successful algorithms especially deep-learning-based ones published in recent years.

The majority of the deep-learning-based FLD algorithms conduct landmark coordinate regression or landmark heatmap regression to localize the landmarks. The former directly predicts the position of each landmark, while the latter estimates each landmark's heatmap and localizes the landmark at the highest-response point. Nevertheless, with these algorithms, the semantic and geometric correlation among the landmarks may not have been fully exploited to guide the regression. By contrast, boundary heatmap was initially proposed and used by [19] for FLD. Each boundary in the heatmap connects landmarks with close semantic and geometric relationships (*e.g.* an eyebrow boundary connects all landmarks residing on an eyebrow) and provides an extra-level of semantic abstraction on the landmarks.

Therefore, we adopt boundary heatmap in our FLD algorithm as well. Similar to [19], we take a two-stage approach: estimating the boundary heatmap at the first stage and using it to assist the landmark coordinate prediction at the second stage. But differently, we make major optimizations to the network structure. Specifically, major algorithmic contributions of this work include:

- **Enhanced HourglassNet for the boundary heatmap regression**. For the boundary heatmap regression in the first stage, we introduce significant enhancements to the baseline HourglassNet including a CoordConv layer and several shallow and deep feature fusion blocks.
- **Newly designed subnet for the landmark coordinate regression**. For the landmark coordinate regression in the second stage, we design a Transformer-based subset. For the input to the Transformer, we design fine modules to fuse information of the original image, a latent feature from the first stage and the boundary heatmap generated by the first stage.

2 Related Works

In recent years, deep learning based methods have achieved great success and become the mainstream for FLD. They can be divided into two categories - coordinate regression methods and heatmap regression methods. Due to the space limit, we review only a few most related ones in the following.

Coordinate regression methods [4, 13, 19–21, 23–25] learn a direct mapping from the input image containing the face to the coordinates of landmarks. ODN [25] obtains clean feature representations at occluded areas and complements semantic features with facial geometric features. GReg+LRefNets [13] combines global regression and local refinement and shares low level features between them. LAB [19] firstly estimates a boundary heatmap which is then used to guide the prediction of landmark coordinates. Wing Loss ([4]) designs a new loss function for landmark prediction. SLPT [20] proposes a sparse local patch transformer to learn landmarks inherent relation for robust face alignment. AnchorFace [21] configures a set of anchor templates for different poses of faces and refines the templates with a network. SRN [23] reduces the semantic ambiguity caused by occlusion with the capture of structural relationships between different facial components. GlomFace [24] solves the problem of occlusion by modelling the facial hierarchies of various occlusions. ResNet [6] is usually used as a backbone for these methods.

Heatmap regression methods [2, 8, 9, 15–17, 22] estimate the heatmap of each landmark and predict the landmark to be at the point with the highest response or proximity value. AWing [17] improves Wing loss so that it is not only microscopic near zero, but also more friendly to small errors. HIH [9] uses two types of heatmaps in collaboration to address the negative impact of quantization. HRNetV2 [16] predicts heatmaps by combining the outputs of HRNet. SAAT [22] generates adversarial images by conditional GAN for training a network with two hourglass modules. LUVLi [8] proposes a method to predict landmark visibility and algorithm confidence for each landmark. MMDN [15] enhances the robustness of detection results by exploring the high-order feature correlations. SAAT [22] improves facial landmark detection as a defence against sample-adaptive black-box attacks. ACHR [2] proposes a network architecture that does not require downsampling and is specifically designed to predict landmarks on very high resolution facial images. Hourglass networks [10] or UNet [11] are often used as backbones in these methods. PIPNet [7] makes a joint prediction of landmark heatmaps and offset values. In general, heatmap regression methods produce more accurate results at the cost of more computing and memory resources.

Transformer [14] has been recently used for vision tasks for its unique global feature attention and superior dynamic feature extraction. Examples include ViT [3] for image classification and DETR [1] for target detection.

3 Proposed Approach

Our proposed neural network consists of two stages, boundary heatmap estimation and landmark coordinate prediction. The structural diagram is shown

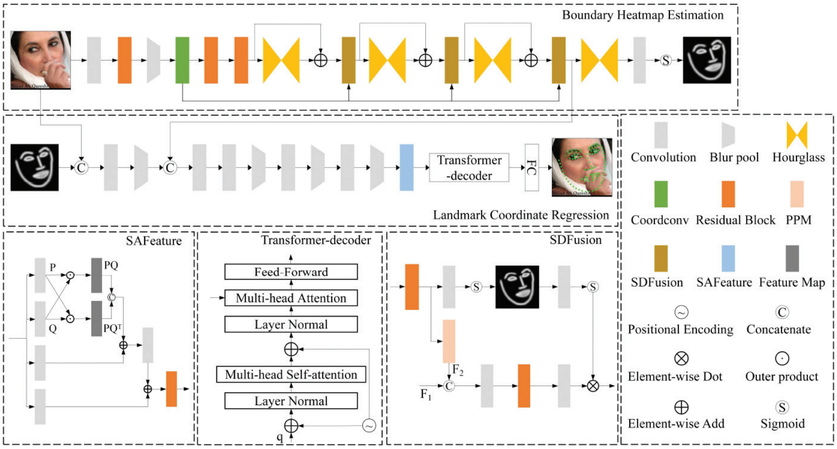


Figure 1: Architectural overview of our proposed boundary-aware face alignment model. The first row of the structural diagram shows the boundary heatmap estimation subnet while the second row shows the landmark coordinate prediction subnet. The remaining part shows fine structures of some modules in the neural network. More explanation of the model is provided in Section 3.

in Figure 1 where the first (resp. second) row corresponds to the first (resp. second) stage.

Specifically, let $I^{W\times H\times C}$ be an input image, where W, H, C are the width, height and channels of image I. Then facial landmark detection problem of our model can be defined as such function $\Phi : I \to (B, P)$. That means, from the input image I, we predict a boundary heatmap sequence $B = \{b_1, b_2, \ldots, b_K\}$ and a landmark matrix $P = X^{N\times 2}$, where b, K and N denote the predicted heatmap, the number of hourglass module from the first stage and the number of facial landmarks from the second stage, respectively. In addition, the first stage will transport the input image I, the predicted boundary heatmap b_K and the input of the last hourglass module to the second stage network.

3.1 Boundary Heatmap Estimation

Attention mechanisms [14, 18] and coarse-to-fine frameworks [5] are effective techniques applied in many computer vision tasks. Inspired by attention mechanisms and coarse-to-fine frameworks, the Boundary Heatmap Estimation subnet aims to focus on boundary region information for more accurate prediction of boundary heatmaps, and to appropriately use the predicted heatmaps as an explicit guidance to enhance the feature map. As the Hourglass network has shown superior performance in LAB, we also use Hourglass modules to construct our coarse-to-fine subnet but make significant improvements to the baseline Hourglass network. The improvements are motivated by our pursuit for robustness against translation of faces and shifting of features and full utilization of features from multiple stages and multiple scales.

As shown in Figure 1, the layers before the first hourglass module is similar to those in the baseline model, but we replace the pooling layer by blur pool and add a CoordConv layer afterwards. Every two adjacent hourglass modules are connected by an SDFusion module, which generates an attention map to enhance the features obtained by fusing the outputs F_1 and F_2 of the shallow CoordConv layer and the pyramid pooling module (PPM) module, respectively. The fusion is performed using two 1×1 convolutions and a residual module. Each SDFusion module predicts the boundary heatmap, which is not only used for intermediate supervision but also for generating the attention map. In contrast to the baseline model, which uses a simple 1×1 convolution to extract features from the predicted boundary heatmap and then fuses them with other features for the next module, the SDFusion module can focus the network on boundary features and provide features with more boundary information for the input of the next module.

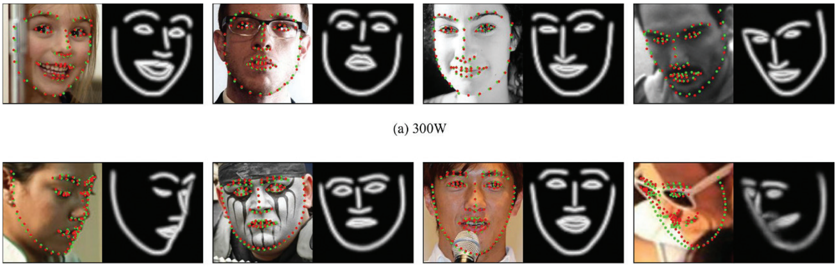
In summary, we propose to enhance the baseline HourglassNet in several aspects, as shown in Figure 1. First, we use blur pooling in place of max pooling in an early stage to introduce shift-invariance. Second, we introduce a CoordConv layer that processes translation better than a plain convolution layer. Third, we design and add SDFusion blocks to effectively fuse shallow and deep features. Specifically, each SDFusion block fuses the feature from the shallower CoordConv layer and the feature from a deeper hourglass block. It is worthwhile to point out that, in order to capture multi-scale context information, PPM is used in SDFusion.

During training, the network is supervised by ground truth heatmaps. As for how to produce ground truth heatmaps, we use the method provided by LAB [19]. In addition, we use 13 boundaries. In order to better display the results, the boundary heatmap results we show in Figure 2 are processed, that is, we draw all the boundaries on one map.

3.2 Landmark Coordinate Prediction

As Transformer has achieved great success in solving a variety of vision tasks, we use it as the core of our landmark coordinate regressor, as shown by the second row of the structural diagram in Figure 1.

As can be seen in Figure 1, prior to the Transformer, we design network blocks to effectively fuse information of the original image I, a latent feature F from the first stage and the boundary heatmap b_K generated by the first stage. To ensure that the original image I and the boundary heatmap b_K have the same resolution, downsampling and upsampling are respectively applied to them. In particular, continuous channel adjustment, resolution reduction



(a) 300W

(b) WFLW

Figure 2: Results on some test images selected from the 300W and WFLW datasets, including an inaccurate result placed at the bottom right. For the result of each test image, the red and green points within the left image represent the ground-truth and predicted landmark locations, respectively, while the right image is the boundary heatmap estimated by the proposed method.

and feature extraction are performed while the three inputs are fused with concatenation, convolution and blur pooling. Then, a self-attention-based feature re-extraction (SAFeature) module is designed and used.

As shown in Figure 1, the SAFeature module contains four branches. The top two branches generate the two feature maps P and Q, and use matrix outer product to obtain two new feature maps PQ and PQ^T that are then concatenated. The concatenated feature map is then fused with the outputs of the other two branches in turn, one by addition and convolution and the other by addition and residuals. Finally, after dimension adjustment, the fused feature gets through the Transformer decoder to obtain the final result.

As shown in Figure 1, there are several highlights in our design of the landmark coordinate prediction subnet. Firstly, blur pooling is used multiple times in the pipeline to achieve shift-invariance. Secondly, a SAFeature block is designed and used. The self-attention mechanism used in the SAFeature block helps capture long-range dependencies in the image. Thirdly, the Transformer also depends heavily on the self-attention mechanism that helps promote the accuracy of prediction.

3.3 Loss Function

The total loss includes two terms, L_{lm} and L_{bh}, corresponding to the landmark coordinate loss and the boundary heatmap loss, respectively.

We use L_2 loss to define each term. Specifically, the loss is defined by

$$\begin{aligned}\text{Loss} &= L_{lm} + \beta L_{bh} \\ &= \frac{1}{N_{lm}} \sum_{i=1}^{N_{lm}} ||p_i - \hat{p_i}||^2 + \frac{\beta}{N_{bh}} \sum_{i=1}^{N_{bh}} \omega_i ||H_i - \hat{H}_i||^2 \end{aligned} \tag{1}$$

where, N_{lm} denotes the number of facial landmarks, p_i and $\hat{p_i}$ denote the predicted coordinates and the ground truth, respectively, N_{bh} denotes the number of predicted boundary heatmaps (equal to the number of hourglass blocks), H_i and $\hat{H}_i$ are the predicted boundary heatmap and the ground truth, respectively, ω_i is the weight, and β is a hyperparameter to regulate the two types of losses and set to 0.0001 by default.

4 Experiments

4.1 Implementation Details

Each input image is cropped and resized to 256×256 and each boundary heatmap has a size of 64×64. Data enhancement is performed on the training data by random translation(±10%), rotation(±30°), horizontal flipping(50%), illumination(±20%), blurring(10%) and occlusion. For the training, we use an Adam optimizer with the initial learning rate set to 1×10^{-4} and β_1 and β_2 set to 0.5 and 0.9, respectively. On one GPU (NVIDIA 3090 24GB), the network is trained for 150 epochs, and the learning rate is reduced to 1/10 of the previous value for twice at the 90th and the 120th epochs. The batch size is 16 and, in the loss function, the weights $\omega_{i=1,2,3,4}$ are 0.25, 0.5, 0.75 and 1.0, respectively.

4.2 Metics and Datasets

Evaluation Metrics. We applied the commonly used evaluation metrics, Normalized Mean Error (NME), Failure Rate (FR) and Area under the Curve (AUC), to compare the proposed method with some state-of-the-art methods. **NME** is defined as:

$$\mathrm{NME}(\mathrm{P}, \hat{\mathrm{P}}) = \frac{1}{N} \sum_{i=1}^{N_{lm}} \frac{||p_i - \hat{p_i}||^2}{d} \times 100 \tag{2}$$

where, P and $\hat{P}$ denote the predicted and annotated coordinates of landmarks, respectively, p_i and $\hat{p_i}$ indicate the coordinate of the i-th landmark in P and $\hat{P}$, respectively, N is the number of the facial landmarks, and d is the reference distance to normalize the error. Here, we use the distance between outer eye corners (inter-ocular) as the reference distance. **FR** represents the percentage of the failed images whose NMEs are higher than a certain threshold in the test set. **AUC** can be calculated based on Cumulative Error Distribution (CED) curve. A larger AUC means that more images are well estimated.

General Datasets. Experiments were conducted on two general datasets, 300W [12] and WFLW [19].

Table 1: Comparing with state-of-the-art methods on 300W. Key: [Best, Second Best].

	Method	Common	Challenging	Full
NME(%) ↓	ACHR	2.83	7.04	4.23
	SRN	3.08	5.86	3.64
	SRN+HG	3.03	5.38	3.49
	LAB	2.98	5.19	3.49
	LUVLi	2.76	5.16	3.23
	HRNetV2	2.87	5.15	3.32
	AnchorFace	3.12	6.19	3.72
	HIH$_C$	2.95	5.04	3.36
	HIH$_T$	2.93	5.00	3.33
	SAAT	2.87	5.03	3.29
	PIPNet	2.78	4.89	3.19
	SLPT	2.75	4.90	3.17
	GlomFace*	2.72	4.79	3.13
	AWing	2.72	4.52	3.07
	Ours	2.71	4.70	3.10

300W dataset contains 3148 images for training and 689 images for testing. Following the widely used evaluation setting, the test sets usually consist of the common set (554 images), the challenging set (135 images) and the full set (the total 689 images). Each image in 300W is annotated with 68 facial landmarks.

WFLW dataset contains 7500 images for training and 2500 images for testing with 98 landmarks and rich attribute labels. It also has six different test subsets with attribute labels, such as occlusion, make-up and illumination.

4.3 Comparison of Experimental Results

Figure 2 shows some results on some test images selected from the 300W and WFLW datasets and Figure 3 also provides more actual results of generated boundary heatmaps from six subtests of the WFLW test. For the result of each test image, the ground-truth and predicted landmark locations are marked with red and green colors in the left image, respectively; while the right image is the boundary heatmap estimated by the proposed method. From these examples, we see that our proposed model adapts well to various challenging situations. The predictions are close to the ground truth for most cases shown here except for the last one at the bottom right. Since the test image at the bottom right contains complex information like large pose and severe occlusion,

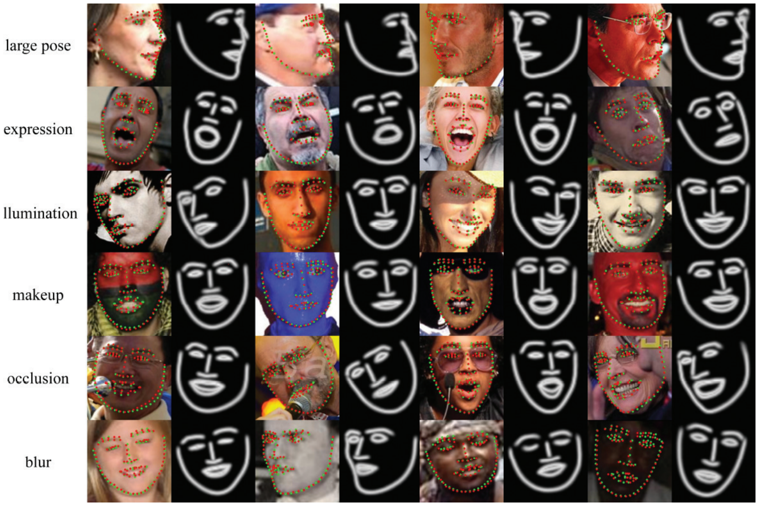


Figure 3: Image samples from six subsets of WFLW test imposed with landmarks and generated boundary heatmaps. Each row comes from different subset.

the boundary heatmap estimated by our method is not that correct, which further affects the accuracy of the predicted landmark locations.

Evaluation on 300W. We compared the proposed method with several state-of-the-art methods on the three test sets, i.e., the common, challenging and full sets. The results are shown in Table 1. The best and second best results are highlighted in red and blue, respectively. Our method performed the best on the common set and relatively well on the full set. Due to the limited number of images for training, our method performed slightly worse than the Awing method [17] on the challenging set.

Evaluation on WFLW. We compared different methods on the test set and several subsets including large pose, expression, illumination, make-up, occlusion and blur. The comparison results are shown in Table 2. Though the NME value of our method is slightly higher than the best value achieved by SLPT [20] on the whole test set, our method performed much better than SLPT under the FR metric. Concerning the results on the subsets, our method performed relatively well and outperformed the other methods on the two subsets, the occlusion and blur subsets. The results demonstrate that our method is effective for face alignment.

4.4 Ablation Study

Evaluation on SDFusion and SAFeature Modules. In order to evaluate the importance of the proposed SDFusion and SAFeature modules to the

Table 2: Comparing with state-of-the-art methods on WFLW. Key: [Best, Second Best].

	Testset			Subset(NME(%))					
Method	NME(%) ↓	FR(%) ↓	AUC ↑	Large pose	Expression	Illumination	Make-up	Occlusion	Blur
LAB	5.27	7.56	0.5323	10.24	5.51	5.23	5.15	6.79	6.32
Wing	5.11	6.00	0.5504	8.75	5.36	4.93	5.41	6.37	5.81
MMDN	4.87	-	-	8.15	4.99	4.61	4.72	6.17	5.72
SRN	4.86	-	-	-	-	-	-	-	-
GlomFace	4.81	3.77	-	8.17	-	-	-	-	4.88
LUVLi	4.37	3.12	0.5777	-	-	-	-	-	-
AWing	4.36	2.84	0.5719	7.38	4.58	4.32	4.27	5.19	4.96
AnchorFace	4.32	2.96	0.5769	-	-	-	-	-	-
PIPNet	4.31	-	-	7.51	4.44	4.19	4.02	5.36	5.02
HIH$_C$	4.18	2.96	0.597	7.20	4.19	4.45	3.97	5.00	4.81
SLPT	4.14	2.76	0.595	6.96	4.45	4.05	4.00	5.06	4.79
Ours	4.16	2.32	0.5927	7.20	4.46	4.07	4.10	4.87	4.66

Table 3: Ablation experiments for SDFusion and SAFeature. Key: [Best].

	WFLW	300W
w/o SDFusion	4.34	3.25
w/o SAFeature	4.31	3.39
w/ neither	4.42	3.45
w/ both	4.16	3.10

Table 4: Ablation experiments for enhanced HourglassNet. Key: [Best].

	WFLW	300W
w/o enhancements	0.955	0.972
w/ enhancements	0.962	0.976

landmark detection results, we conduct experiments on WFLW and 300W and implement four different models with/without SDFusion or SAFeature. The evaluation metric we use is NME, and the results are summarized in Table 3. With one or both of SDFusion and SAFeature removed, the NMEs increase significantly, which indicates the effectiveness of the SDFusion and the SAFeature modules.

Evaluation on Enhanced HourglassNet. In order to evaluate the importance of the proposed enhancements to HourglassNet for boundary heatmap estimation, we experiment with the baseline HourglassNet without enhancements and the enhanced HourglassNet on WFLW and 300W. In this experiment, we adopt the Structural Similarity (SSIM) to measure the quality of a predicted boundary heatmap. The closer the SSIM value is to 1, the more identical the predicted boundary heatmap is to the ground truth. The experimental results are summarized in Table 4. On both datasets, we observe higher SSIM values with the enhanced HourglassNet, which indicate the effectiveness of the proposed enhancements to the HourglassNet.

5 Conclusion

In this work, we have proposed a neural network for boundary-aware face alignment. It is boundary-aware in that it first estimates a boundary heatmap and then uses the estimation to guide the prediction of landmark positions. Correspondingly, the proposed neural network is composed of two stages, boundary heatmap estimation and landmark coordinate prediction. For the first stage, a baseline HourglassNet is enhanced by blur pooling, CoordConv, and shallow and deep feature fusion. For the second stage, we design a

Transformer-based subnet that firstly fuses information of the original image, a latent feature from the first stage and the boundary heatmap generated by the first stage, and secondly uses a Transformer to map the fused feature to the landmark coordinates. L_2 losses of both boundary heatmaps and landmarks coordinates are considered for the optimization. Experiments show that the proposed algorithm achieves state-of-the-art performance on the general datasets of 300W and WFLW.

In the future, we plan to further promote the accuracy of boundary heatmap prediction, especially for complex cases (*e.g.*, large pose, severe occlusion), as it plays an essential role in guiding the landmark coordinate regression.

Acknowledgements

This work was supported by Shandong Provincial Natural Science Foundation, China (No. ZR2022MF294), the National Natural Science Foundation of China (No. 62102163), and the Development Program Project of Youth Innovation Team of Institutions of Higher Learning in Shandong Province.

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*, Springer, 2020, 213–29.

[2] P. Chandran, D. Bradley, M. Gross, and T. Beeler, “Attention-driven cropping for very high resolution facial landmark detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 5861–70.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[4] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, “Wing loss for robust facial landmark localisation with convolutional neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 2235–45.

[5] P. Gao, K. Lu, J. Xue, L. Shao, and J. Lyu, “A coarse-to-fine facial landmark detection method based on self-attention mechanism,” *IEEE Transactions on Multimedia*, 23, 2020, 926–38.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–8.

[7] H. Jin, S. Liao, and L. Shao, "Pixel-in-pixel net: Towards efficient facial landmark detection in the wild," *International Journal of Computer Vision*, 129(12), 2021, 3174–94.

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 8236–46.

[9] X. Lan, Q. Hu, and J. Cheng, "Revisting quantization error in face alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 1521–30.

[10] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, Springer, 2016, 483–99.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, 234–41.

[12] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, 397–403.

[13] J. Su, Z. Wang, C. Liao, and H. Ling, "Efficient and accurate face alignment by global regression and cascaded local refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, 0–0.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 30, 2017.

[15] J. Wan, Z. Lai, J. Li, J. Zhou, and C. Gao, "Robust facial landmark detection by multiorder multiconstraint deep networks," *IEEE Transactions on Neural Networks and Learning Systems*, 33(5), 2021, 2181–94.

[16] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 43(10), 2020, 3349–64.

[17] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, 6971–81.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, 3–19.

[19] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 2129–38.

[20] J. Xia, W. Qu, W. Huang, J. Zhang, X. Wang, and M. Xu, "Sparse Local Patch Transformer for Robust Face Alignment and Landmarks Inherent Relation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 4052–61.

[21] Z. Xu, B. Li, Y. Yuan, and M. Geng, "AnchorFace: An anchor-based facial landmark detector across large poses," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 4, 2021, 3092–100.

[22] C. Zhu, X. Li, J. Li, and S. Dai, "Improving robustness of facial landmark detection by defending against adversarial attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 11751–60.

[23] C. Zhu, X. Li, J. Li, S. Dai, and W. Tong, "Reasoning structural relation for occlusion-robust facial landmark localization," *Pattern Recognition*, 122, 2022, 108325.

[24] C. Zhu, X. Wan, S. Xie, X. Li, and Y. Gu, "Occlusion-Robust Face Alignment Using a Viewpoint-Invariant Hierarchical Network Architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 11112–21.

[25] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 3486–96.