This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Occlusion-robust Face Alignment using A Viewpoint-invariant Hierarchical Network Architecture

Congcong Zhu[1], Xintong Wan[1], Shaorong Xie[1], Xiaoqiang Li[1]*, Yinzheng Gu[2]
[1]School of Computer Engineering and Science, Shanghai University
[2]Shanghai HYCloud Network Technology Co. Ltd.
{congcongzhu, wanxintong, srxie, xqli}@shu.edu.cn, guyinzheng@gpushare.com

Abstract

The occlusion problem heavily degrades the localization performance of face alignment. Most current solutions for this problem focus on annotating new occlusion data, introducing boundary estimation, and stacking deeper models to improve the robustness of neural networks. However, the performance degradation of models remains under extreme occlusion (i.e. *average occlusion of over 50%) because of missing a large amount of facial context information. We argue that exploring neural networks to model the facial hierarchies is a more promising method for dealing with extreme occlusion. Surprisingly, in recent studies, little effort has been devoted to representing the facial hierarchies using neural networks. This paper proposes a new network architecture called GlomFace to model the facial hierarchies against various occlusions, which draws inspiration from the viewpoint-invariant hierarchy of facial structure. Specifically, GlomFace is functionally divided into two modules: the part-whole hierarchical module and the whole-part hierarchical module. The former captures the part-whole hierarchical dependencies of facial parts to suppress multi-scale occlusion information, whereas the latter injects structural reasoning into neural networks by building the whole-part hierarchical relations among facial parts. As a result, GlomFace has a clear topological interpretation due to its correspondence to the facial hierarchies. Extensive experimental results indicate that the proposed GlomFace performs comparably to existing state-of-the-art methods, especially in cases of extreme occlusion. Models are available at* https://github.com/zhuccly/GlomFace-Face-Alignment.

1. Introduction

Although great effort [4, 9, 16, 21, 46, 47, 51, 54] has been devoted to face alignment, the localization accuracy remains unsatisfactory under various occlusions. Especially for the current situation, under which people have to wear medical masks due to the COVID-19 pandemic. The following reasons cause the problem. Firstly, some landmarks are inevitably invisible, and partial facial information is not available. Secondly, large-scale occlusion datasets (average occlusion of over 50%) are scarce because annotating landmarks under occlusion is a great challenge. Thirdly, general neural network architectures cannot model the spatial relationship between facial components [35].

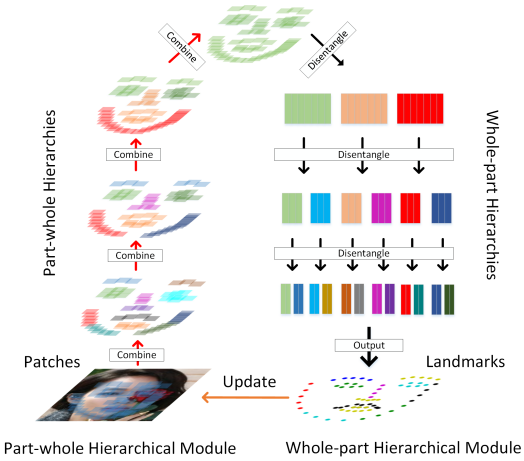

Figure 1. Insight of the proposed GlomFace.

Some studies [30, 53] deal with the occlusion problem by enhancing the coupling of facial context features. However, the excessive coupling may introduce occlusion information over the whole face, resulting in the degradation of localization accuracy on non-occluded areas. Methods such as [22, 23, 40, 41] integrate related tasks (*e.g.* visibility estimation and uncertainty prediction) to improve the occlusion robustness. However, the related tasks cannot directly impose the shape constraint over all landmarks and may even introduce additional annotation and computational costs. Recently, boundary estimation has been in-

*Corresponding author.

tegrated into heatmap-based models [18, 19, 42, 46] and has become a mainstream solution to occlusion, which predicts facial boundaries to provide the shape constraint. Nonetheless, boundary estimation is prone to failure due to the loss of boundary information under extreme occlusion, leading to drift in all landmarks, as shown in the second image of Figure 2. In addition, boundary estimation is computationally complex. Therefore, the performance degradation caused by occlusion remains an unsolved problem.

There is the fact that facial landmarks describe the physiological structure of a human face, which inherently possess viewpoint-invariant hierarchies. The hierarchies are not disturbed by any external environments and thus can be considered to be powerful clues for structural reasoning. There appears to be some works [25, 26, 55] that focus on the facial hierarchies, but they only fine-tune the predicted results for the backbone networks rather than actually model the hierarchies. Geoffrey Hinton indicated that a general neural network can hardly represent viewpoint-invariant hierarchies and proposed GLOM [15] to do so. Unfortunately, GLOM [15] only presents a single idea about representation without describing any working network. Inspired by GLOM, we thought over how a neural network with a fixed architecture can model the viewpoint-invariant hierarchies to handle occlusion. To achieve this, we propose a new neural network architecture called GlomFace, which is functionally divided into two modules: the part-whole hierarchical module (PHM) and the whole-part hierarchical module (WHM).

We first define the face hierarchies into different levels, and further divide a face into the different facial parts at each level. The part-whole hierarchical module (PHM) hierarchically captures the part-whole spatial dependencies of each facial part, as shown on the left of Figure 1. The shape-indexed patches are fed into the PHM as the lowest-level facial part, and then, this module captures the short-range spatial dependencies within each patch. Subsequently, adjacent patches are combined into neighborhood parts. PHM then enlarges the range of spatial dependencies to the neighborhood level. The above operation is repeated until all patches are combined into a whole. With hierarchical spatial dependencies, PHM can suppress multi-scale occlusion information. Using shape-indexed patches instead of the raw image as input is done for two reasons, 1) providing a clear part-whole hierarchy and 2) the low computational complexity when capturing spatial dependencies in each part. When PHM outputs a whole representation, the whole-part hierarchical module (WHM) starts to build hierarchical relations among facial parts for structural reasoning. To achieve this, WHM hierarchically disentangles the whole representation into low-level part representations, as shown on the right of Figure 1. In each representation disentangling, WHM considers the coupling relations (part-part relationship) between adjacent facial parts at the same level and the constrained relations (whole-part relationship) of a high-level facial part over its internal low-level facial parts. When representation disentangling is completed level by level, part-part and whole-part relations are simultaneously built. With hierarchical relations, WHM can achieve structure reasoning against the shape damage of facial landmarks. Finally, the predicted landmarks are used to update the position of all shape-indexed patches that will be fed into GlomFace again to refine all spatial dependencies and relations. With the viewpoint-invariant hierarchical architecture, the proposed GlomFace can handle various occlusions (*e.g.* self-occlusion and external occlusion) and achieve promising performance even for extreme occlusion cases. Figure 2 shows an example compared with the mainstream Hourglass [32] backbone equipped with boundary estimation. Experiments demonstrated that GlomFace is more robust to occlusion and has a smaller number of FLOPS compared with hourglass-based methods [32, 42, 46].

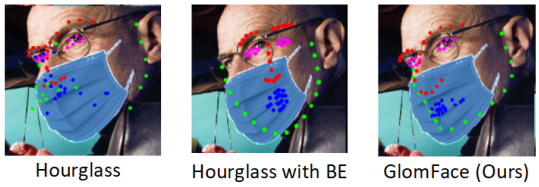

Figure 2. Hourglass *v.s.* GlomFace on a Masked face. "BE" denotes boundary estimation, which imposes shape constraint [42, 46]. We can see that the boundary estimation will make the global shape drift when real facial boundaries are lost. Note that three models are fairly trained on 300W [34], not masked faces.

2. Related Work

Facial alignment aims to localize the key points for a given face image. In [6, 39, 49, 54], facial landmark localization was implemented in a coarse-to-fine manner, which iteratively refines the initial landmarks to the final results. With CNNs applied in this task, studies such as [17, 31, 38] achieved competitive performance by extracting discriminative features from pixels. With the large pose issues taken into consideration, 3D face pose has been introduced to address the large-pose issue, which fits a 3D morphable model (3DMM) to a 2D image [1, 4, 16, 20, 27]. However, methods that apply 3DMM cannot handle occlusion because it is extremely difficult to reconstruct a 3D face under occlusion.

To address the occlusion problem, RCPR [5] reduces exposure to outliers by detecting occlusions explicitly to extract robust shape-indexed features. PCD-CNN [22] used a Dendritic CNN to develop a cascade local prediction model, which ignores the mutual constraints between different facial components. Some methods stack hourglass networks and combine related tasks to handle occlusion faces. LU-

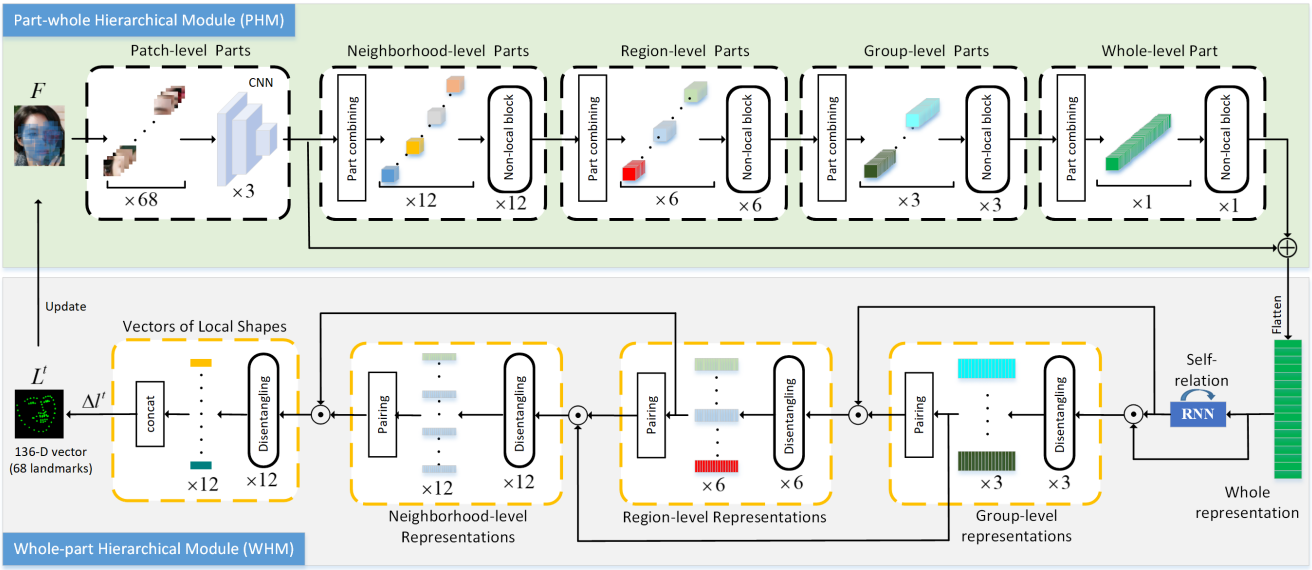

Figure 3. Illustration of GlomFace for 68 landmark prediction. The input is a shape-indexed patch set surrounding the predicted landmarks of the previous iteration. Following the part-whole hierarchies shown in Figure 4, PHM combines patch features into higher-level facial parts based on their indexes and captures the spatial dependencies across patches within each part. It outputs a whole representation with hierarchical dependencies. WHM then hierarchically parses the whole representation into the offset vectors of local landmarks by representation disentangling. This operation builds the hierarchical relations among facial parts following the whole-part hierarchies shown in Figure 4. Here, "$\bigoplus$" and "$\bigodot$" denote residual connection and skip concatenation operations, respectively. In fact, skip concatenation is performed between each i level representation and its internal i-1 level representations. Due to limited space, we only show a skip concatenation in each representation disentangling. More details can be found in the following sections and supplementary materials.

VLi [23] jointly predicts landmark locations, associated uncertainties among these predicted locations, and landmark visibility. It models multi-tasks as mixed random variables and estimates them using a four-stage stacked hourglass network (HG). Look at boundary (LAB) [46] imposes global shape constraint over all landmarks by introducing boundary estimation, in which stacked hourglass is used to estimate boundary heatmaps that are computationally complex. PropNet [18], AWing [42] and ADNet [19] follow LAB to stacks massive parameters for estimating boundary heatmaps. Compared to LAB [46], their number of FLOPS are greater. All these methods based on boundary estimation require high computational costs. Although LUVLi [23] proposed a new occlusion dataset MERL-RAV, the dataset cannot provide a complete facial structure because the self-occluded landmarks have no labels. Furthermore, increasing training data does not really provide neural networks the structural reasoning. Therefore, the occlusion problem is still a great challenge.

3. Methods

3.1. Overall Architecture

Figure 3 presents an overview of the proposed GlomFace. Given a face image F and the initial landmarks L^0, we crop N shape-indexed patches surrounding these landmarks in specified index order. These patches are fed into GlomFace, which iteratively refine these landmarks by modeling the viewpoint-invariant hierarchies. Here, the initial landmarks of each iteration step are the prediction of the previous iteration. The initial landmarks are coordinate values of a mean shape from the current training set for the first iteration. Let L^t denote the predicted landmarks at iteration step t, which is refined incrementally based on the results of the previous iteration:

$$L^t = L^{t-1} + \triangle l^t, \tag{1}$$

where $\triangle l^t$ is an offset vector of all landmarks at the t-th iteration, which is predicted using GlomFace:

$$\triangle l^t = \mathbf{GlomFace}(\rho(F|L^{t-1})), \tag{2}$$

where $\rho(\cdot)$ denotes a patch-extracting operation, which crops the shape-indexed patches surrounding L^{t-1}.

After performing all iteration steps, we minimize the following loss function to update GlomFace's parameters:

$$Loss = \sum_{t=1}^{T} \left\| \bar{L} - \left(L^{t-1} + \triangle l^t\right) \right\|_2^2, \tag{3}$$

where T and $\bar{L}$ denote the maximum iteration step and the ground-truth landmarks. Following MDM [38], this $L2$ loss

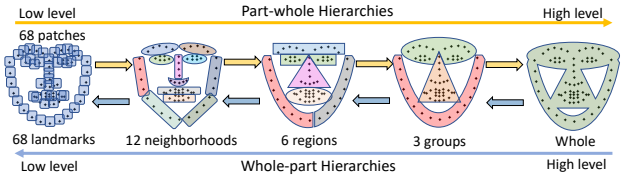

Figure 4. Viewpoint-invariant hierarchies with five levels. Patches and landmarks share a fixed index order. Each block represents a facial part. Part combining and representation disentangling are performed according to the part-whole hierarchies and the whole-part hierarchies, respectively.

function simply computes the point-to-point Euclidean error. It can be replaced to improve further the performance of GlomFace by recent works such as Wing loss [10].

Specifically, GlomFace is functionally divided into two modules: the part-whole hierarchical module (PHM) and the whole-part hierarchical module (WHM). All patches are first sent into the PHM, and then it captures the short-range spatial dependencies within each patch. These patch-level dependencies can suppress the small-scale occlusion information. As the hierarchies get higher, PHM gradually enlarges the range of the spatial dependencies by part combining and thus handles the larger-scale occlusion. The part combining merges the low-level facial parts into the high-level parts by following the part-whole hierarchies until all patches are combined into a whole.

When PHM outputs a whole representation with part-whole hierarchical dependencies, the whole-part hierarchical module (WHM) starts to build the whole-part hierarchical relations by representation disentangling according to the whole-part hierarchies. To achieve this, WHM hierarchically disentangles the representation of high-level facial parts into the representations of its internal lower-level facial parts. In this operation, whole-part hierarchical relations are built between face parts at the same level and between high and low-level face parts. WHM finally disentangles each neighborhood-level representation into the offset of a specified local shape. The predicted offsets update the landmarks and the position of shape-indexed patches.

The part combining and representation disentangling follow the viewpoint-invariant hierarchies with five levels, from low to high, including patch/landmark, neighborhood, region, group and whole, shown in Figure 4.

3.2. Part-whole Hierarchical Module (PHM)

The part-whole hierarchical module (PHM) starts from three convolutional layers, followed by a max-pooling operation behind each layer. In general, the self-attention mechanism is the common way to capture spatial dependencies of a feature map. We find that suitable receptive fields ($7*7$, $5*5$ and $3*3$) in these three convolutional layers can achieve almost the same performance as self-attention due to the small size of each patch ($40 * 40 * 3$). As CNNs reduce the size of patches, the computational complexity of subsequent hierarchical dependencies is significantly reduced.

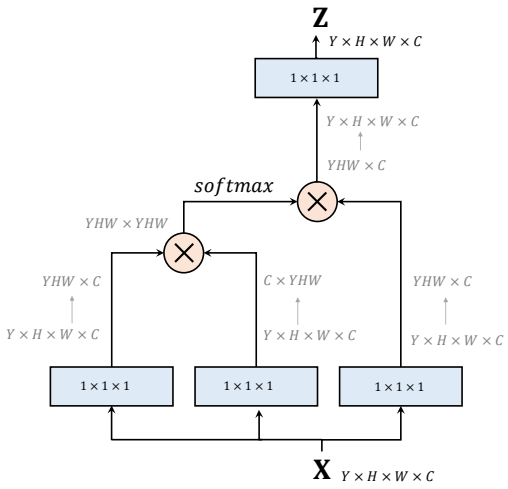

Figure 5. A non-local block. "**X**" is a feature map set with a shape of $Y \times H \times W \times C$, where Y denotes the number of feature maps with a size of "$H \times W \times C$" (height, width and channel). "$\otimes$" denotes the matrix multiplication. The blue boxes denote 1×1×1 convolutions.

After extracting patch-level part features using CNNs, PHM performs a part combining to produce the feature maps of higher-level facial parts by following the part-whole hierarchies shown in Figure 4. The PHM then enlarges the scope of the spatial dependencies to the range of current facial parts. Since each facial part contains more than one patch feature, the common self-attention mechanisms cannot capture the dependencies across patches. Therefore, we use a non-local operation [3] to perform the capturing of spatial dependencies across patches. It computes the response at a position as a weighted sum of the features at all positions. Following [44], we build a non-local block to do so, shown in Figure 5. Different from [44] focusing temporal dependencies across frames, we exploit this block to capture spatial dependencies across patches. Finally, a whole representation is obtained, which represents the part-whole hierarchical dependencies.

3.3. Whole-part Hierarchical Module (WHM)

There is strong psychological evidence that people can parse an object into parts at different levels and model the viewpoint-invariant spatial relationship between a part and a whole as the coordinate transformation [14]. This evidence would appear to explain why people excel at structural reasoning (*e.g.* playing jigsaw puzzles). Inspired by existing studies [14, 15], the whole-part hierarchical module (WHM) learns how to understand the whole-part hierarchi-

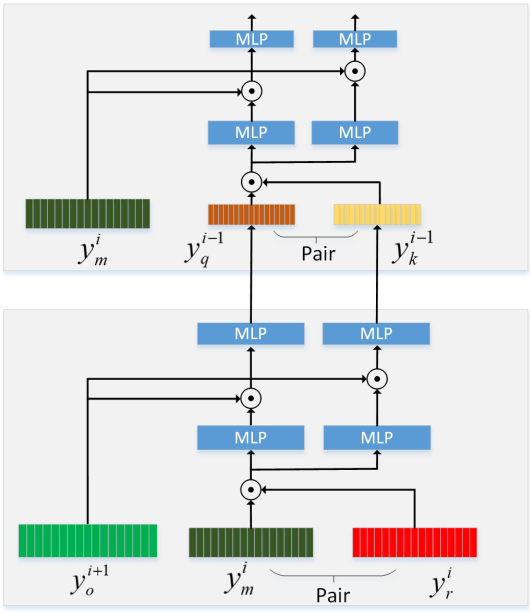

Figure 6. Successive disentangling operations. Here, y_o^{i+1} denotes the representation of a facial part o at the i+1 level, which is disentangled into y_m^i and y_r^i. Subsequently, y_m^i is disentangled into y_q^{i-1} and y_k^{i-1}. For simplicity, we omit operations of disentangling y_k^{i-1}, y_r^i and y_o^{i+1}.

cal relations of facial parts. With the top-down hierarchical architecture, WHM hierarchically parses the representation of each high-level facial part to the representations of its internal low-level parts by representation disentangling, level by level, following the whole-part hierarchies shown in Figure 4. The representation disentangling is similar to picking out the best fit for each part from all pieces when people are playing a puzzle. In this operation, the representations of adjacent parts that are all subordinate to the same higher-level part are paired into a relation pair or triple for building the coupling relationship. Meanwhile, this higher-level part representation imposes the constrained relation over this pair or triple. Taking i level facial part m as an example, let $\mathbf{y}_m^i$ denote its representation. We suppose it has a adjacent part representation $\mathbf{y}_r^i$. Where (m, r) is a i level pair that is subordinate to i+1 level part o with feature representation $\mathbf{y}_o^{i+1}$. If m contains two i-1 level parts, the representation disentangling is shown in Figure 6 and can be expressed as follows:

$$\mathbf{y}_q^{i-1} = \mathbf{MLPs}_\pi\left((\mathbf{y}_m^i, \mathbf{y}_r^i)|\mathbf{y}_o^{i+1}\right), \tag{4}$$

$$\mathbf{y}_k^{i-1} = \mathbf{MLPs}_\varphi\left((\mathbf{y}_m^i, \mathbf{y}_r^i)|\mathbf{y}_o^{i+1}\right), \tag{5}$$

where facial parts q and k are subordinate to part m, $\mathbf{y}_q^{i-1}$ and $\mathbf{y}_k^{i-1}$ denote the representations of parts q and k. $\mathbf{MLPs}_\pi$ and $\mathbf{MLPs}_\varphi$ disentangles $\mathbf{y}_m^i$ into $\mathbf{y}_q^{i-1}$ and $\mathbf{y}_k^{i-1}$, respectively. With building hierarchical relations, WHM injects the structural reasoning into neural networks. Note that we use its historical information for the whole representation to build the self-relation by leveraging a recurrent neural network (RNN). Although the extraction of facial features from local patches can reduce data dimensionality, some facial information may be missed. Memorizing previous information can meaningfully supplement facial information for subsequent iterations. Moreover, in the cascaded regression framework, every iteration should interact rather than be performed independently. Using an RNN to memorize facial information across all iterations allows joint optimization of all iteration components during end-to-end training. This helps our model to achieve smooth offset predictions and stability during training, as shown by MDM [38].

This self-relation significantly reduces the data dimensionality of the whole representation without losing important information and thus saves computational costs. Finally, WHM disentangles each neighborhood-level representation to the offset vector of a specified local shape whose landmarks are indexed in the current neighborhood. We concatenate all vectors to the offset of a global shape with N landmarks and use it to update the landmarks and patches for the next iteration.

4. Experiments

4.1. Datasets and Metrics

General datasets. The evaluation datasets for this setting are typically 300W [34] and WFLW [46]:

300W [34]. Following the widely used evaluation setting [38, 46, 51], the training set contains 3148 images. The test sets consist of LFPW and HELEN test sets as the Common set, the IBUG set as the Challenging set, and the union of them as the Full set.

WFLW [46]. This dataset is a new facial dataset based on WIDER Face [48], which was proposed by LAB [46]. It contains $10,000$ images ($7,500$ for training and $2,500$ for testing) with 98 landmarks.

Occlusion datasets. This evaluation setting includes three occlusion datasets:

COFW29 [5]. The Caltech Occluded Face in the Wild (COFW29) dataset [5] consists of 1345 training face images and 507 test face images collected from the Internet, all of which are annotated with 29 landmarks.

COFW68 [12]. To evaluate the cross-data robustness, we train Glomface on 300W dataset (68 landmarks) and test on COFW68 re-annotated with 68 landmarks by [12]. Therefore, this dataset is used only for test and not training.

Masked 300W [51]. This dataset contains masked faces (average occlusion of over 50%), which is synthesized by

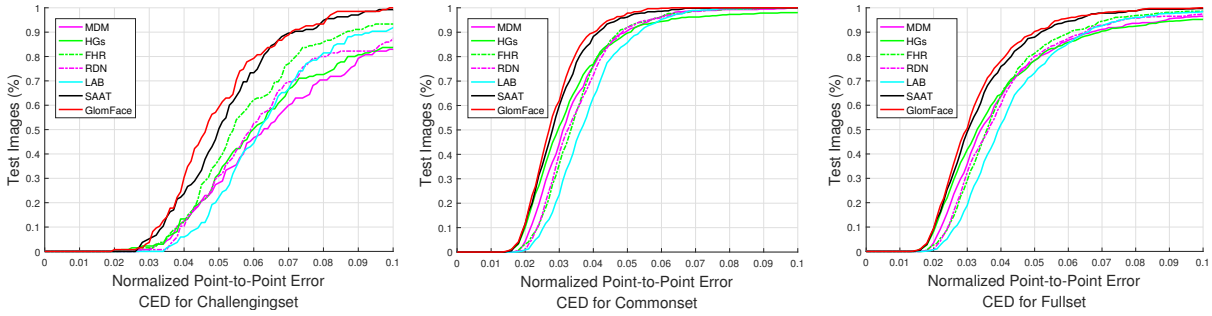

Figure 7. The CED curves of our proposed method compared to state-of-the-art methods on three subsets of 300W.

SAAT [51] based on 300W [34]. Note that this dataset is used only for test and not training.

Evaluation Metrics. We employ the Normalized Error (NME), cumulative error distribution (CED) curves and failure rate (FR) as evaluation metrics. Some state-of-the-art methods [10, 23, 38, 41] use different normalizing terms to compute the NME, we follow their normalizing terms to conduct our experiments for comparison with these methods. Here, NME_{ocular}, NME_{pupil} and NME_{bbx} set normalizing terms as inter-ocular distance, the inter-pupil distance and the geometric mean of the bounding box, respectively.

4.2. Implementation Details

We use the ground-truth bounding boxes to crop the face image into a size of $224 * 224$ and each pact into a size of $40 * 40$. Following the existing methods [10, 38, 42, 46], we augment training data by handcrafted transformation (rotation, flipping, scaling, random blocking,*etc*.). Our model has trained with about $50,000$ steps in an end-to-end manner on four NVIDIA GTX 1080 Ti cards. We set an initial learning rate of 0.0002, a decay factor of 0.97, a batch size of 64 and 4 iteration steps. More implementation details and parameters of network architecture can be found in our code and supplementary materials.

4.3. Evaluation on General Datasets

Results of evaluation on 300W. To fairly compare existing methods, we categorized all methods, based on their prediction manner, into two categories: heatmap-based models and regressor-based models. The former outputs the heatmaps of landmarks always using stacked networks, such as hourglass [32] or Unet-like network. The latter directly predicts the coordinate values of landmarks. Moreover, following Awing [42], we replace CNNs and $L2$ loss, using the CoordConv layer [29] and Wing loss [10], to improve our model termed GlomFace*. We show comparison results of the proposed GlomFace with state-of-the-art methods in Table 1. Experiments show that our GlomFace achieves the best performance compared with all regressor-based models. There is two heatmap-based method (Awing [43] and ADnet [19]) that is significantly stronger than the proposed GlomFace on the challenge set. However, GlomFace surpasses them by a large margin on the occlusion dataset COFW29 (see Table 3). We further observe that almost all approaches to achieve promising performance work incrementally based on the existing backbones [13, 32]. However, this backbone does not achieve structural reasoning to combat occlusion. In the subsequent comparisons on the occlusion data, GlomFace shows a greater advantage. We can see that GlomFace* performs comparably to the best heatmap-based model [43]. This shows that GlomFace is extensible and can be used as a strong baseline.

Methods	Challenging	Common	Full	Backbone
Heatmap-based				
HGs [32](2016)	7.23	3.72	4.41	Designed
FAN [4](2017)	5.52	3.08	3.56	HGs [32]
SAN [7] (2018)	6.60	3.34	3.98	CMP [45]
LAB [46](2018)	5.19	2.98	3.49	HGs [32]
FHR [37] (2018)	6.28	3.02	3.66	HGs [32]
AWing [42] (2019)	4.52	2.72	3.07	HGs [32]
AS+SAN [33] (2019)	6.49	3.21	3.86	CMP [45]
LUVLi [23] (2020)	5.16	2.76	3.23	HGs [32]
3FabRec [2] (2020)	5.74	3.36	3.82	ResNet [13]
SRT [8] (2020)	5.61	2.80	3.39	HGs [32]
SDL [25] (2020)	4.77	2.62	3.04	HRnet [36]
HIH [24] (2021)	5.00	2.93	3.33	HGs [32]
HGs+SAAT [51] (2021)	5.03	2.82	3.25	HGs [32]
ADnet [19] (2021)	4.58	2.53	2.93	HGs [32]
Regressor-based				
MDM [38](2016)	7.56	4.36	4.99	Designed
TSR [31] (2017)	7.56	4.36	4.99	Designed
RDN [28] (2018)	7.04	3.31	4.23	MDM [38]
Wing [10] (2018)	5.23	2.93	3.38	Designed
ODN [53] (2019)	6.67	3.56	4.17	ResNet [13]
SDFL [26] (2021)	4.93	2.88	3.28	ResNet [13]
GlomFace (Ours)	4.87	2.79	3.20	Designed
GlomFce*(Ours)	4.79	2.72	3.13	Designed

Table 1. NME_{ocular} comparison to state-of-the-art methods on 300W. "Designed'' means that the method is designed with a new backbone network. [Key: Top-1, Top-2]

To further evaluate the performance of GlomFace on

Datasets	Full		Pose		Occlusion	
Metrics	NME	FR	NME	FR	NME	FR
LAB [46]	5.27	7.56	10.24	28.83	6.79	13.72
SRT [8]	5.13	7.07	-	-	-	-
3FabRec [2]	5.62	8.28	10.23	34.35	6.92	15.08
SAAT [51]	5.11	5.63	-	-	-	-
LUVLi [23]	4.37	3.12	-	-	-	-
Awing [42]	4.36	2.84	7.38	13.50	5.19	5.98
SDL [25]	4.21	3.04	7.36	15.95	4.98	5.29
GlomFace (Ours)	4.81	3.77	8.17	17.48	5.14	6.73
PropNet* [18]	4.05	2.96	6.92	12.58	4.58	5.16
ADNet* [19]	3.98	2.00	6.56	9.20	4.36	4.48

Table 2. Comparison to state-of-the-art methods on WFLW. Note that PropNet* and ADNet* employ focal wing loss [10] by using the attribute labels provided by WFLW [46]. [Key: Top-1, Top-2]

300W, Figure 7 shows the CED curves compared with state-of-the-art open-source methods whose mean error values are presented in Table 1. As shown in Figure 7, GlomFace significantly surpasses the other methods.

Results of evaluation on WFLW. We followed LAB [46] and used the inter-ocular normalization to normalize errors. We did not compare with some state-of-the-art methods such as PropNet [18] and ADNet [19] because they employ focal wing loss [10] by using the attribute labels provided by WFLW [46]. Moreover, we chosen the pose subset and the occlusion subset for further comparison, since both subsets contain self-occlusion and external occlusion. NME and failure rate (a threshold value of 0.1) are reported in Table 2. Although Awing [42] and SDL [25] outperform ours on the full and large pose sets, GlomFace is competitive on the occlusion set.

4.4. Evaluation on Occlusion Data

In this evaluation, COFW68 and Masked 300W were used to conduct cross-data experiments due to having no training data.

Results of evaluation on COFW29. For COFW29 dataset, we used the inter-pupil distance to normalize the point-to-point error by following [5]. Table 3 indicate that the proposed method outperforms all state-of-the-art methods by a large margin in term of NME. It even surpasses Awing [42] that is the best on 300W. This also proves that GlomFace is more robust on occluded faces than most state-of-the-art methods.

Results of evaluation on COFW68. In Table 4, we report the comparison results with existing state-of-the-art methods on COFW68 [12]. The results indicate that the proposed GlomFace outperforms all state-of-the-art methods by a large margin. Compared with the state-of-the-art SAAT [51], our method achieves decreases in NME of 8.68%. The failure rate of our model is only 0.79% which is far less all existing methods. This lowest failure rate clearly shows that GlomFace achieves excellent performance under occlusion. This cross-dataset evaluation indicates that GlomFace has outstanding robustness against occlusion and generalization in occluded environments.

Method	**NME**$_{pupil}$	**FR**$_{8\%}$	**FR**$_{10\%}$
PCD-CNN [22](2018)	5.77	-	3.73
Wing [10] (2018)	5.44	-	3.75
3DDE [41] (2019)	5.11	6.50	-
AWing [42] (2019)	4.94	5.52	0.99
MNN [40] (2020)	5.04	-	-
ADNet [19] (2021)	4.68	-	0.59
GlomFac (Ours)	4.37	4.53	1.56

Table 3. Comparison of averaged errors and failure rates on COFW29 dataset. [Key: Top-1, Top-2].

Method	**NME**$_{ocular}$	**FR**$_{10\%}$
TCDCN [50] (2016)	8.05	6.31
MDM [38](2016)	6.12	5.13
FAN [4](2017)	5.85	3.94
LAB [46] (2018)	4.62	2.17
ODN [53] (2019)	5.87	2.84
SDL [25] (2020)	4.22	0.39
SRN [52] (2021)	4.67	1.97
ODN [53] (2019)	5.87	2.84
SAAT [51](2021)	4.61	1.58
GlomFac (Ours)	4.21	0.79
	NME$_{bbx}$	
LUVLi [23]	2.75	-
GlomFac (Ours)	2.09	0.39

Table 4. Comparison of averaged errors and failure rates on COFW68 dataset. GlomFace achieves the best performance on this occlusion dataset. [Key: Top-1, Top-2].

Methods	**Challenging**	**Common**	**Full**
CFSS [54]	19.98	11.73	13.35
SBR [37]	13.28	8.72	9.6
MDM [38]	11.67	7.66	8.44
HGs [32]	13.52	8.17	9.22
MDM [38]	11.67	7.66	8.44
FHR [37]	11.28	7.02	7.85
FAN [4]	10.81	7.36	8.02
LAB [46]	9.59	6.07	6.76
SRN [52]	9.28	5.78	6.46
SAAT [51]	11.36	5.42	6.58
GlomFace (Ours)	8.81	5.29	5.98

Table 5. NME comparison on Masked 300W. Note that Masked 300W is only used for cross-dataset evaluation, not training. [Key: Top-1, Top-2].

Results of evaluation on Masked 300W. Table 5 shows the comparison results with existing methods on Masked 300W (average occlusion of over 50%). Following [51], this Masked 300W is used only for cross-dataset evaluation in the test phase and is not used during the training

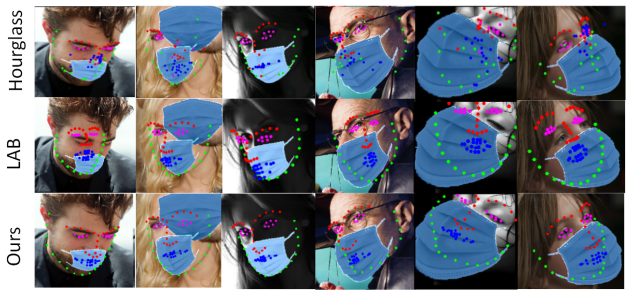

Figure 8. Qualitative results of the proposed GlomFace compared with the Hourglass [32] backbone and LAB [46] with boundary estimation on Masked 300W (average occlusion of over 50%). All methods are trained only on the general 300W [34].

stage. Compared with the state-of-the-art SAAT [51], our method achieves decreases in the error of 22.44%, 6.08% and 11.55% on the Masked 300W Challenging, Common, and Full sets, respectively. These results clearly show that our method achieves the best performance on extreme occluded faces.

We further investigated the occlusion robustness of the proposed GlomFace under extreme occlusion. Figure 8 shows qualitative results of the proposed GlomFace compared with the mainstream Hourglass [32] backbones and boundary estimation model LAB [46] and on severely occluded faces [51] (average occlusion of over 50%). All models are trained only on 300W [34], not any additional training data. These results indicate that GlomFace has powerful structural reasoning capabilities to efficiently combat extreme occlusions.

Analysis. Most state-of-the-art methods improve the occlusion robustness by integrating additional prediction tasks into existing backbone networks. LUVLI [23] introduces visibility and uncertainty estimation into stacked hourglass [32]. LAB [46], AWing [42], ADNet [19] and PropNet [18] exploit boundary estimation to impose the shape constraint to stacked hourglass [32]. Their computational complexity is much greater than that of our GlomFace due to the boundary estimation. SDFL [26] and SDLSDL [25] integrates graph convolution layer into existing backbones to handle the occlusion problem. These SOTAs are all based on pre-existing backbone networks such as HGs, HRnet and ResNet, and thus are incremental works. Unlike these methods mentioned above, GlomFace injects the structure reasoning into the hierarchical architecture, it features a new backbone network with potential for further improvement.

4.5. Self Evaluations

Note that more ablation studies were shown in the supplementary material duo to the page limit.

Computational efficiency. We compared GlomFace with three state-of-the-arts [18,42,46] in the term of computational cost (FLOPS) for predicting 68 landmarks. These three methods are all hourglass-based and integrate boundary estimation by stacking sub-networks. As shown in Table 6, GlomFace has a smaller number of FLOPS than these methods. The results demonstrate that our architecture is more efficient compared to boundary estimation.

Methods	FLOPS	Sub-nets
LAB [46]	18.85G	✓
AWing [42]	26.79G	✓
PropNet [18]	42.83G	✓
GlomFace ($t=4, i=5$)	13.48G	×
GlomFace ($t=1, i=5$)	3.37G	×

Table 6. Comparison of computational complexity. Here, "t" and "i" denote the iteration step and the number of facial part levels.

Limitations. Although GlomFace achieves a promising performance, it has two limitations. The first is that the proposed GlomFace cannot be ported to other prediction tasks, since its network architecture is designed exactly according to the facial hierarchies. Second, GlomFace relies on the indexes of dense landmarks to define facial hierarchies, and it will lose the advantage of the facial hierarchies when predicting sparse landmarks (*e.g.* 5 landmarks).

5. Conclusion

In this paper, we have proposed a new network architecture, GlomFace, which significantly improves the occlusion robustness of face alignment. Experimental results demonstrated that GlomFace achieves competitive performance compared with state-of-the-art methods, especially on occluded faces. The advantages of GlomFace are summarized as follows: (1) structural reasoning; (2) better occlusion robustness than existing methods; (3) smaller number of FLOPS than hourglass-based models. GlomFace's shortcomings mainly regard poor performance in terms of portability and sparse landmark prediction. Future work focuses on more lightweight GlomFace and extending spatial dependencies into the temporal domain to handle dynamic occlusion in facial landmark tracking.

Acknowledgements

This work is supported in part by Shanghai science and technology committee under grant No.21511100600. We appreciate the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No.19DZ2252600) for providing the computing resources and technical support.

References

[1] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *TPAMI*, 25(9):1063–1074, 2003. 2

[2] Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *CVPR*, June 2020. 6, 7

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*, volume 2, pages 60–65. IEEE, 2005. 4

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, pages 1021–1030, 2017. 1, 2, 6, 7

[5] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollar. Robust face landmark estimation under occlusion. In *ICCV*, December 2013. 2, 5, 7

[6] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *IJCV*, 107(2):177–190, 2014. 2

[7] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, pages 379–388, 2018. 6

[8] X. Dong, Y. Yang, S. Wei, X. Weng, Y. Sheikh, and S. Yu. Supervision by registration and triangulation for landmark detection. *TPAMI*, pages 1–1, 2020. 6, 7

[9] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *CVPR*, pages 360–368, 2018. 1

[10] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, pages 2235–2245, 2018. 4, 6, 7

[11] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *CVPR*, pages 2385–2392, 2014.

[12] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015. 5, 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[14] Geoffrey Hinton. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, 3(3):231–250, 1979. 4

[15] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *arXiv preprint arXiv:2102.12627*, 2021. 2, 4

[16] Shi HL et al. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016. 1, 2

[17] Zhibin Hong, Xue Mei, Danil Prokhorov, and Dacheng Tao. Tracking via robust multi-task multi-view joint sparse representation. In *ICCV*, pages 649–656, 2013. 2

[18] Xiehe Huang, Weihong Deng, Haifeng Shen, Xiubao Zhang, and Jieping Ye. Propagationnet: Propagate points to curve to learn structure information. In *CVPR*, pages 7265–7274, 2020. 2, 3, 7, 8

[19] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *ICCV*, pages 3080–3090, 2021. 2, 3, 6, 7, 8

[20] Amin Jourabloo and Xiaoming Liu. Pose-invariant face alignment via cnn-based dense 3d model fitting. *IJCV*, 124(2):187–203, 2017. 2

[21] Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Pose-invariant face alignment with a single cnn. In *ICCV*, Oct 2017. 1

[22] Amit Kumar and Rama Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. *CVPR*, 2018. 1, 2, 7

[23] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, pages 8236–8246, 2020. 1, 3, 6, 7, 8

[24] Xing Lan, Qinghao Hu, and Jian Cheng. Revisting quantization error in face alignment. In *ICCV Workshop*, pages 1521–1530, 2021. 6

[25] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, et al. Structured landmark detection via topology-adapting deep graph learning. In *ECCV*, pages 266–283. Springer, 2020. 2, 6, 7, 8

[26] Chunze Lin, Beier Zhu, Quan Wang, Renjie Liao, Chen Qian, Jiwen Lu, and Jie Zhou. Structure-coherent deep feature learning for robust face alignment. *TIP*, 2021. 2, 6, 8

[27] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *ECCV*, pages 545–560. Springer, 2016. 2

[28] Hao Liu, Jiwen Lu, Minghao Guo, Suping Wu, and Jie Zhou. Learning reasoning-decision networks for robust face alignment. *TPAMI*, 2018. 6

[29] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *NeurIPS*, 31, 2018. 6

[30] Zhiwei Liu, Xiangyu Zhu, Guosheng Hu, Haiyun Guo, Ming Tang, Zhen Lei, Neil M Robertson, and Jinqiao Wang. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *CVPR*, pages 3467–3476, 2019. 1

[31] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *CVPR*, pages 3317–3326, 2017. 2, 6

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 2, 6, 7, 8

[33] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial land-

mark detector with semi-supervised style translation. In *CVPR*, pages 10153–10163, 2019. 6

[34] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, pages 397–403, 2013. 2, 5, 6, 8

[35] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NeurIPS*, pages 4967–4976, 2017. 1

[36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 6

[37] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. In *AAAI*, volume 33, pages 8893–8900, 2019. 6, 7

[38] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, pages 4177–4187, 2016. 2, 3, 5, 6, 7

[39] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, pages 3659–3667, 2015. 2

[40] Roberto Valle, Jose Miguel Buenaposada, and Luis Baumela. Multi-task head pose estimation in-the-wild. *TPAMI*, 2020. 1, 7

[41] Roberto Valle, José M. Buenaposada, Antonio Valdés, and Luis Baumela. Face alignment using a 3d deeply-initialized ensemble of regression trees. *CVIU*, 189:102846, 2019. 1, 6, 7

[42] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, pages 6971–6981, 2019. 2, 3, 6, 7, 8

[43] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, pages 6971–6981, 2019. 6

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 4

[45] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016. 6

[46] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138, 2018. 1, 2, 3, 5, 6, 7, 8

[47] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 1

[48] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, pages 5525–5533, 2016. 5

[49] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, pages 1–16, 2014. 2

[50] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, 38(5):918–930, 2016. 7

[51] Congcong Zhu, Xiaoqiang Li, Jide Li, and Songmin Dai. Improving robustness of facial landmark detection by defending against adversarial attacks. In *ICCV*, pages 11751–11760, October 2021. 1, 5, 6, 7, 8

[52] Congcong Zhu, Xiaoqiang Li, Jide Li, Songmin Dai, and Weiqin Tong. Reasoning structural relation for occlusion-robust facial landmark localization. *Pattern Recognition*, 122:108325, 2022. 7

[53] Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *CVPR*, pages 3486–3496, 2019. 1, 6, 7

[54] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, pages 4998–5006, 2015. 1, 2, 7

[55] Xu Zou, Sheng Zhong, Luxin Yan, Xiangyun Zhao, Jiahuan Zhou, and Ying Wu. Learning robust facial landmark detection via hierarchical structured ensemble. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 141–150, 2019. 2