# Supplementary Material: 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

*In this document, we provide more details of network architecture and parameter setting. Further investigation into the selection of patch sizes for different mark-up annotations are provided. Moreover, some qualitative results of GlomFace predicting on extremely occluded faces are visualized to prove structural reasoning of GlomFace.*

## 1. Detailed Network Architecture

GlomFace is functionally divided into a part-whole hierarchical module (PHM) and a whole-part hierarchical module (WHM). The PHM is easy to understand and follow, because it only involves part combining and non-local operation [7]. Therefore, in this section, we mainly show details of WHM architecture. Figure 1 shows the detailed network architecture of the proposed WHM. For the whole representation, we use a recurrent neural network (RNN, 1024-D) to memorize historical information for imposing self-relation. We formulate the recurrent neural network (RNN) for the current iteration step as:

$$\mathbf{y} = \mathbf{RNN}(\mathbf{x} \oplus \mathbf{y}^*), \tag{1}$$

where $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{y}^*$ denote the current self-relation, the whole representation and the previous self-relation, respectively. The RNN is simple and sufficient for memorizing short-term information (four iterations in our implementation). Therefore, in our implementation, replacing the RNN with more complex operations such as LSTM or GRU does not result in more performance gains.

## 2. Ablation study

**Facial hierarchies.** To further investigate the hyper-parameters of the proposed GlomFace, we varied the number of facial levels and iteration steps to report the performance, respectively. All experiment results shown in Table 1 are evaluated on the challenge set of 300W [5]. The table shows that optimal performance is achieved when using five levels and four iterative steps. Two iteration steps can achieve competitive performance when using five levels. We further found that increasing the number of iterations can decrease the average error but overfitting after four iterations. These results demonstrate that GlomFace can significantly benefit from facial hierarchies, and 4 iteration steps are the best. Moreover, the other two investigations are as follows: removing residual connections (3.13→3.16 NME), replacing the PHM with the feature extractor fro [6] (3.13→3.38 NME).

| Setting | $\text{NME}_{ocular}$ |
|---|---|
| GlomFace ($t = 4$,$i = 2$) | 6.39 |
| GlomFace ($t = 4$,$i = 3$) | 5.37 |
| GlomFace ($t = 4$,$i = 4$) | 5.01 |
| GlomFace ($t = 1$,$i = 5$) | 6.73 |
| GlomFace ($t = 2$,$i = 5$) | 5.13 |
| GlomFace ($t = 3$,$i = 5$) | 4.96 |
| GlomFace ($t = 4$,$i = 5$) | 4.87 |
| GlomFace ($t = 5$,$i = 5$) | 4.88 |
| GlomFace ($t = 6$,$i = 5$) | 5.31 |

Table 1. Ablation experiment: NME comparison on the challenge set of 300W. Here,"$t$" and "$i$" denote the iteration step and the number of facial part levels.

**Faces under different levels of occlusion.** COFW and Masked 300W evaluated are two datasets with different levels of occlusion (over 25% and 50%). Moreover, we used five noise block sizes to randomly occlude the full set of 300W.

## 3. Patch Size.

Different annotation schemes of existing datasets have a different number of landmarks, such as COFW68 [3] (68

*Corresponding author.

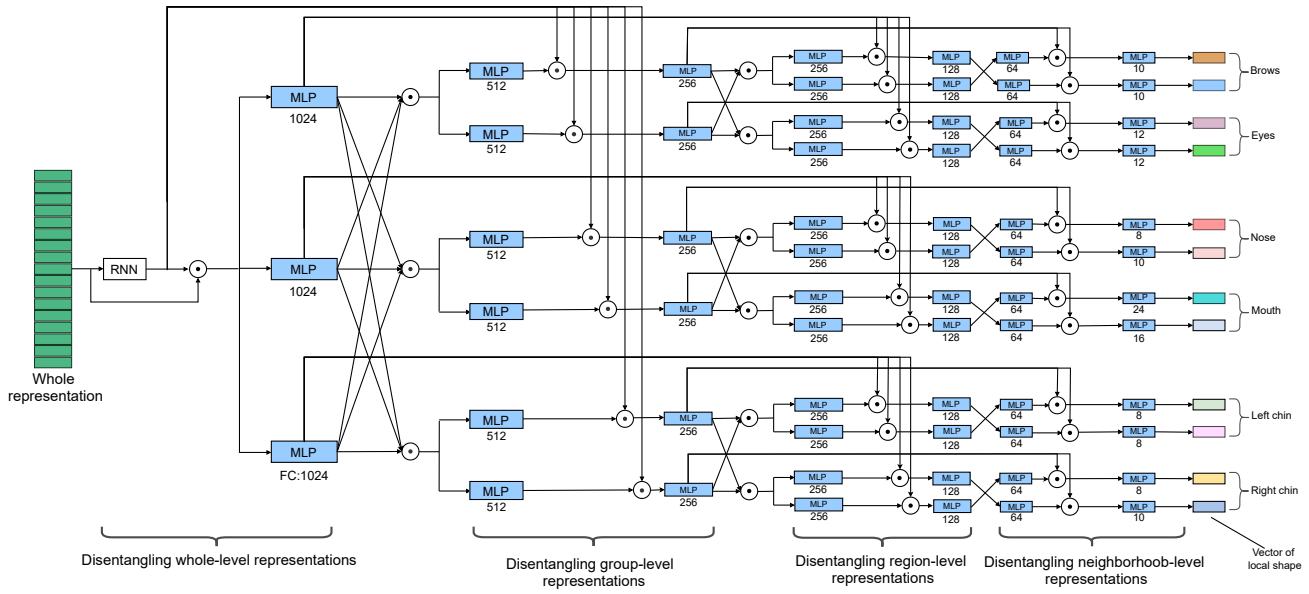


Figure 1. Detailed Architecture of whole-part hierarchical module (WHM). Where each 1-layer MLP (fully connected layer) with RELU nonlinearity is followed by a batch-norm layer. "$\odot$" denotes concatenation operation. Each number refers to the output dimension of the current MLP.

| Occlusion size | 20 × 20 | 40× 40 | 60× 60 | 80× 80 | 100× 100 |
|---|---|---|---|---|---|
| MDM [6] | 3.79 | 4.28 | 5.31 | 6.81 | 8.23 |
| SAAT [9] | 3.31 | 3.92 | 4.71 | 5.69 | 6.12 |
| GlomFace (Ours) | 3.15 | 3.67 | 4.08 | 4.57 | 5.38 |

landmarks), 300W [5] (68 landmarks), WFLW [8] (98 landmarks) and COFW29 [1] (29 landmarks). For 68 and 98 landmarks, we set the patch size to $42 * 42$, and large patch size leads to heavy overlap of face patches and information redundancy. For sparse 29 landmarks, we set the size of patches to $46 * 46$ and remove the neighborhood level due to its sparse hierarchical information. Reducing patch size will result in not cropping enough facial information. In Table 2, we show the ablation experiments on patch size. The results show that when GlomFace predicts sparse landmarks, the patch size should be increased accordingly. In addition, increasing the size does not significantly increase the computational effort due to the reduced number of patches.

## 4. Prediction under Extreme Occlusion

Additional qualitative results of the proposed GlomFace on Masked 300W are illustrated in Figure 2. We can see that the proposed GlomFace can efficiently reason facial structure under extreme occlusion. The sixth column illustrate that GlomFace achieves structural reasoning rather than captures edge information. The last column shows the failure examples. These samples undergo structural incongruities under large poses. More results can be found by running our evaluation code.

| Patch size | COFW68 | COFW29 |
|---|---|---|
| S=36 | 5.16 | 6.14 |
| S=38 | 4.62 | 5.49 |
| S=40 | 4.24 | 4.81 |
| S=42 | 4.21 | 4.75 |
| S=44 | 4.22 | 4.42 |
| S=46 | 4.29 | 4.37 |
| S=48 | 4.65 | 4.37 |

Table 2. Ablation experiment: NME comparison of different patch sizes ($\text{NME}_{ocular}$ for COFW68 and $\text{NME}_{pupil}$ for COFW29 ).

## 5. Discussion

The proposed GlomFace is a new network architecture focusing on occlusion, which injects the power of structural reasoning into the neural network by leveraging hierarchical spatial dependencies and relations. It differs from any existing backbone network for face alignment tasks, its structural reasoning power comes from the architecture rather than from additional prediction tasks (*e.g.* visibility estimation and boundary estimation). We believe that Glom-

Face can serve as a strong baseline, existing incremental works [2, 4, 9] can be integrated to further improve the performance of the proposed method.

## References


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

[2] 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. 3

[3] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015. 1

[4] 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. 3

[5] 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. 1, 2

[6] 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. 1, 2

[7] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 1

[8] 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. 2

[9] 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. 2, 3

Figure 2. Qualitative results on Masked-300W Dataset. The last column shows the failure examples.