2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICME57554.2024.10687648

TOWARDS ACCURATE 3D FACE ALIGNMENT UNDER EXTREME SCENARIOS VIA MULTI-GRANULARITY PERTURBATION RELEARNING

Xinyu Li[1], Xing Wang[2*], Xiaoxiao Yang[1], Suping Wu[1†], Xiangzheng Li[3], Xitie Zhang[1], Zhiyuan Zhou[1], Xiang Zhang[1]*

[1]NingXia University, [2]Beijing China-Power Information Technology Co., LTD., [3]Ningxia Normal University
Email: pswuu@nxu.edu.cn

ABSTRACT

3D face alignment from monocular images in challenging scenarios such as large poses and occlusions presents a huge challenge. To overcome this challenge, we propose a Multi-granularity Perturbation Relearning Network (MPRN), utilizing relearning attention to capture crucial features. Specifically, MPRN employs an attention mechanism to highlight effective features and further conducts relearning for attention to refine its accuracy. However, in extreme scenarios, the loss of key 3D facial information hampers the effective functioning of relearning attention. To this end, we construct multi-granularity perturbation graphs to infer the missing key 3D facial information and correspondingly guide the multiple times learning of attention module using perturbation graphs at various granularities. By doing this, our MPRN could effectively capture crucial 3D facial features in extreme scenarios, thereby achieving precise 3D face alignment. Experiments on the AFLW 2000-3D and AFLW datasets demonstrate the effectiveness of our MPRN.

Index Terms— 3D face alignment, 3D multimedia, computer vision

1. INTRODUCTION

The 3D face alignment task is to automatically locate the 3D key feature points of a face including eyes, nose tip, mouth corner points, eyebrows, and contour points of various parts of the face based on the input face image. It plays an important role in many vision analysis applications such as face recognition, expression recognition, and facial makeup changing. In 3D face alignment, estimating the key feature points of a 3D face from a single image has become a challenging problem due to a variety of complex factors such as lighting, external occlusions, and large facial poses. To address the above problems, [1] proposes a three-dimensional deformable model (3DMM) for 3D faces based on principal component analysis (PCA). 3DMM is a statistical parametric model that includes identity, expression, and texture parameters. It approximates the 3D face as a linear combination of basic identity, expression, and texture. With the development of deep learning, deep neural networks were introduced into 3DMM estimation [2, 3, 4, 5, 6, 7, 8, 9]. 3DDFA[10] employs deep neural network to predict the 3DMM parameters of face image to generate 3D face. Then, it completes the alignment by extracting the key point coordinates of generated 3D face. Unlike 3DMM-based methods, PRNet[11] predicts UV position maps to generate 3D faces, reducing the data processing workload. DFIN[12] designs a feature extraction network containing deep and shallow modules and introduces facial geometry consistency loss to extract different levels of semantic information. DAMDNet[13] proposes a dual-attention mechanism dense network, including a spatial and a channel attention mechanism. DAMDNet could adaptively focus on the spatial and channel distribution of feature maps, thereby enhancing more important feature information. RADAN[14] improves the feature extraction ability of the network especially the discriminative features containing local and global facial information. However, when under the extreme scenarios such as large pose and occlusion, the invisible part of the representation forms missing and the these methods are usually unable to use the representation information to align the face landmarks, as shown in Figure 1. This poses a huge challenge for neural network prediction.

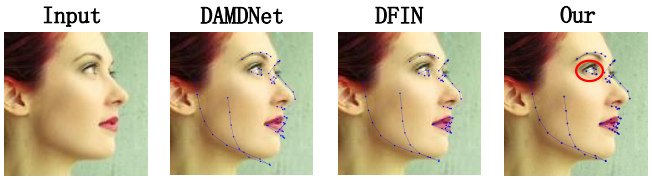

Fig. 1. 3D face alignment comparison in large pose extreme scenarios. Our alignment of eye is more accurate.

To achieve better 3D face alignment, we propose a Multi-granularity Perturbation Relearning Network (MPRN). Specifically, we employ the Convolutional Block Attention Module (CBAM) [15] to discern which features extracted by the backbone network are effective. To refine the accuracy of CBAM, we incorporate relearning by correcting CBAM's

* represents equal contributions and †represents corresponding author. This work was supported by the National Natural Science Foundation of China under Grant 62062056 and 61662059, and in part by the Ningxia Natural Science Foundation of China under Grant 2022AAC03327.

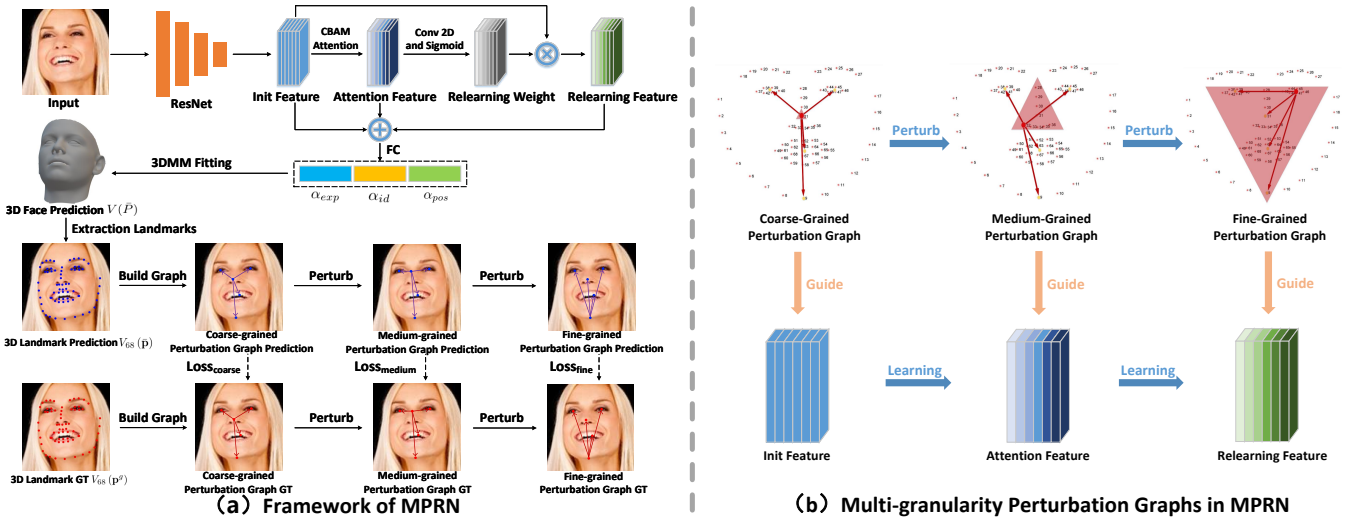

Fig. 2. (a) is the framework of the proposed MPRN, while (b) is multi-granularity perturbation graphs in MPRN.

outputs. However, in extreme scenarios, the loss of critical 3D facial information makes it difficult for both CBAM and the relearning process to capture effective features. Therefore, we need to infer comprehensive 3D facial structural information to guide them. To this end, we build a deterministic face structure graph with the nose tip as the center. As shown in Figure 2 (b), the center of the picture and the center of the cheek outline are the tip of the nose (No. 31) and the chin (No. 9) respectively. The centers of the left and right eyes and the mouth are the 3D coordinate averages of their respective parts. We employ such a graph for the predicted landmarks and GT landmarks, making the predicted landmarks more accurate by taking the length and direction of the corresponding edges in the two graphs consistently. However, the graph is uniquely determined and the attention module will only focus on specific edges of the graph, resulting in only a part of the face structure information could be mined. To mine more complete information, we propose an idea of multi-granularity perturbation and elaborate design a perturbation method which changes the anchor point of the graph from the tip of the nose to a random center point of the eye, nose, mouth, and cheek area. Thus, the face structure graph is not uniquely determined and is constantly updated in the training process, enabling it to represent a more complete 3D facial structure information. In order to adapt to this random change process, multi-granularity perturbation graphs motivate the neural network to spread its attention from specific edges of the graph to random edges, and then the attention module will mine more face geometric structure information, so as to make more accurate inferences about the invisible points of the face with large poses. We utilize three granularity levels of perturbation graphs to guide attention through three learning stages, enhancing the effective extraction of 3D facial features, thus achieving more accurate 3D face alignment. Our core contributions to this work are summarized as follows:

- Our MPRN builds multiple 3D face graphs at multi-granularity perturbation, which represent 3D face structure information from global to local respectively. Multi-granularity perturbation graphs could gradually infer 3D face information of the occluded area, thereby solving the significant challenge brought by extreme scenarios such as occlusion.
- In MPRN, multi-granularity perturbation graphs guide the learning of the attention module in a corresponding manner, making it more explicit. This process plays a crucial role in enhancing the effective extraction of 3D face features.
- Extensive experimental results on the AFLW 2000-3D and AFLW datasets prove that the 3D face alignment performance and generalization of our MPRN is superior.

2. PROPOSED METHOD

2.1. Overview

Figure 2(a) shows our MPRN, it uses ResNet to extract the initial feature map of the input image and then obtains the attention feature map via CBAM. Next, it calculates the relearning weight map through 2D convolution and the sigmoid function, and multiplies it with the initial map to get the relearning feature map. Then, it sums up these three maps and uses a FC layer to obtain 3DMM coefficients for fitting the 3D face of the input image. Finally, it extracts 68 key points based on 3DMM and constructs a multi-granularity perturbation graph sequence accordingly. These graphs guide the

three feature extraction process, resulting in more accurate 3D face alignment, as shown in Figure 2(b).

2.2. Face Feature Extraction

After extracting the features of the input image using ResNet, we integrate the CBAM attention module to compute the interdependencies between channel layers or spatial pixels, thereby discerning between effective and ineffective features. In order to refine CBAM's modeling of these interdependencies, we adopt a relearning approach to recalibrate CBAM. This empowers the network to enhance its representational capacity through iterative learning, facilitating the explicit identification of effective features. Formally, the process could be described as the following formulas:

$$F_{init} = ResNet(X) \tag{1}$$

$$F_{att} = CBAM(F_{init}) \tag{2}$$

$$W_{rel} = Sigmoid(Conv_{3\times3}(ReLU(Conv_{3\times3}(F_{att})))) \tag{3}$$

$$F_{rel} = W_{rel} \odot F_{init} \tag{4}$$

where X, F_{init}, F_{att}, $Conv_{3\times3}$, W_{rel}, $\odot$, and F_{rel} represent the input image, initial feature map, attention feature map, 3×3 convolution, relearning weight map, element-wise product, and relearning feature map, respectively. All $F_{init,att,rel}$ and $W_{rel} \in R^{C\times H\times W}$. Through the process, the attention feature map is obtained by the learning of CBAM module, then it is transformed into a weight map by the relearning of two-layers 2D convolution with sigmoid function. The weight map recalibrates CBAM and generates a relearning feature map by performing an element-wise product with the initial feature map.

Then, we sum up these three feature maps and obtain the 3DMM coefficients through utilizing the fully connected (FC) layer. Formally, this process could be described as the following formulas:

$$\alpha_{id,exp,pos} = FC(F_{init} + F_{att} + F_{rel}) \tag{5}$$

where $\alpha_{id,exp,pos}$ represents the 3DMM coefficients, which has 62 dimensions. We select the first 40 dimensions as the id parameters α_{id}, the following 10 dimensions as the expression parameters α_{exp}, and the last 12 dimensions as the pose parameters α_{pos}. These three parameters are the most important features in 3D face alignment and everything we do is aimed at extracting these three parameters more accurately.

2.3. Multi-granularity Perturbation Graph

3DMM model is fitted to an accurate 3D face based on these coefficients α_{pos}, α_{id}, and α_{exp}. Then, we construct the multi-granularity perturbation graphs with 68 landmarks, where the landmarks could be obtained by the fixed index of the fitted 3D face. Formally, it could be expressed in turn as follows:

$$V = (\alpha_{id} * A_{id} + \alpha_{\text{exp}} * A_{\text{exp}}) * \alpha_{\text{pos}} \tag{6}$$

$$V_{68} = V[ind, :, :] \tag{7}$$

where α_{id}, α_{exp}, α_{pos}, V, A_{id}, and A_{exp} represent the calculation of the identity parameters, expression parameters, pose parameters, fitted 3D face, identity base of 3DMM, and expression base of 3DMM, respectively. While V_{68} and ind are the predicted 3D landmarks and the index of the 68 face landmarks in the fitted 3D face, respectively.

With the predicted 3D landmarks V_{68}, we could proceed to construct multi-granularity perturbation graphs, as shown in Figure 2(b). We first set up the base nodes of the graph, which consist of landmarks with fixed ordinal numbers. For cheek contour, eyes, and mouth, we use No.9 point, No.37-No.42 average point (same for the right eye), and No.49-No.68 average point as base nodes respectively. Next, we select the central node of the graph, which is connected to the basic node to obtain the edge, and the edge in turn forms the graph together with the node.

We construct the graph in the form of a multi-granularity perturbation, which are three granularities of coarse, medium, and fine, as shown in Figure 2(b). The coarse-grained graph takes the tip of the nose (No.31) as the only center point, which ensures that the most important edges representing the most basic geometric structure information of the face are fixed. Furthermore, to diffuse the receptive field of the attention module and focus on the comprehensive face geometric structure information, we impose multi-granularity perturbations on coarse-grained graph. Specifically, we perturb the center point of the coarse-grained graph to obtain a medium-grained graph. The graph takes a random landmark in the nose region (No.28-No.36) as the center point, making the edges connecting the center point and the base point random so that the graph represents higher-order geometric structure information. Similarly, fine-grained graphs could be obtained by applying perturbations to medium-grained graphs. The center point is spread from the nose area (No.28-No.36) to the average point of eyes (No.37-No.42, No.43-No.48), mouth area (No.49-No.68), and the cheek contour center point (No.9 point). The graph randomly selects the center point among these points so that the edge connecting the center point and the base point can pass through the whole 3D face, making the graph represents the complete geometric structure information. We construct multi-granularity perturbation graphs for predicted 3D face and ground truth, respectively.

As shown in Figure 2(b), we employ multi-granularity perturbation graphs to guide the iterative extraction and learning of facial feature maps. Specifically, the coarse-grained perturbation graph guides the initial feature map because both represent the fundamental aspects of 3D face information. Subsequently, due to the broader representation scope of the

medium-grained perturbation graph compared to the coarse-grained one and the presence of more effective features in the attention feature map compared to the initial feature map, they encompass higher-order 3D face information, allowing the medium-grained perturbation graph to effectively guide the attention feature map. Similarly, the fine-grained perturbation graph, along with the relearning feature map, represents more comprehensive 3D face information, making fine-grained perturbation graph suitable for guiding the relearning feature map. In this way, ResNet avoids overlooking the most fundamental and essential three-dimensional facial features. Furthermore, CBAM could more clearly distinguish between effective and ineffective features, while its calibration becomes more targeted.

3. LOSS FUNCTION

We exploit the difference in the length and direction of the corresponding edges of the predicted graph and the ground truth graph. Specifically, under the same granularity, for two edges obtained by connecting the same center point to the same base point on the two graphs, their lengths are used to calculate the mean square error, and their directions are used to calculate the cosine similarity. The weighted summation of the two obtains the loss at this granularity and the weighted summation of the losses at multiple granularity could obtain the total loss. Formally, the above loss function could be expressed in turn as follows:

$$V_{center} \in \begin{cases} \{31\} & \text{, Coarse-grained} \\ \{[28,36]\} & \text{, Medium-grained} \\ \{ave([37,42]), ave([43,48]), ave([49,68]), 9\} & \text{, Fine-grained} \end{cases} \tag{8}$$

Where$\{\cdot\}$ and $[\cdot]$ represent the set of integers and closed intervals, respectively. In addition, ave represents the average operation and the numbers such as 31 are the X, Y, and Z coordinates of face landmarks index. Therefore, V_{center} could be different while the granularity of the perturbation graph is different.

$$\text{Loss}_{length} = \sum_{i=1}^{k} \left\| \left(V_{base^i_{pre}} - V_{center_{pre}} \right) - \left(V_{base^i_{gt}} - V_{center_{gt}} \right) \right\|^2 \tag{9}$$

V_{base} represents the X, Y, and Z coordinates of the base points of the two graphs, respectively. Correspondingly, V_{center} represents the coordinates of the center point. k is the number of base points. In this way, the length difference of the corresponding edges of the predicted graph and the GT graph is calculated.

Set the vector $V_{base_{pre}} - V_{center_{pre}}$ = a and $V_{base_{gt}} - V_{center_{gt}}$ = b, then the cosine similarity of the corresponding edges of the two graphs is:

$$\cos\theta = \frac{ab}{\|a\|\|b\|} \tag{10}$$

$$\text{Loss}_{direction} = \sum_{i=1}^{k} (1 - \cos\Theta^i) \tag{11}$$

where $\cos\Theta^i$ represents the cosine similarity of the two No.i corresponding edges between the reconstructed graph and the ground truth graph. K is the number of base points. Thus, the direction difference of the corresponding edges of the predicted graph and the GT graph is obtained.

$$\text{Loss}_{coarse} = \alpha Loss_{length} + \beta Loss_{angle} \tag{12}$$

$Loss_{coarse}$ represents the coarse-grained perturbation graph loss, while α and β are weight coefficients. In the same way, we could obtain the medium-grained and fine-grained perturbation graph loss. Then, the weighted addition of the three granularity perturbation graph loss functions to obtain the multi-granularity perturbation graph (mpg) Loss function.

$$\text{Loss}_{mpg} = \lambda Loss_{coarse} + \mu Loss_{medium} + \xi Loss_{fine} \tag{13}$$

where $Loss_{mpg}$, $Loss_{medium}$, and $Loss_{fine}$ represent the multi-granularity perturbation graph, medium-grained, and fine-grained perturbation graph loss function, respectively. λ, μ, and ξ are weight coefficients.

4. EXPERIMENTS

4.1. Datasets

We train our MPRN on 300W-LP[10] and evaluate MPRN on AFLW[20] and AFLW 2000-3D[10].

4.2. Implementation Details

Our MPRN is trained on GTX 2080Ti and is implemented on the PyTorch deep learning framework. Through the training, the loss function weights α, β, λ, μ, and ξ are 1, 1000, 10, 10, and 1.

4.3. Quantitative Experiment

We present the normalized mean error (NME) results for each method in Table 1. NME calculates the average of the normalized errors between the label landmarks and the predicted landmarks, reflecting the accuracy of the proposed method for 3D face alignment. In Table 1, we present a comparison between our approach and the majority of state-of-the-art methods, including yaw angles of [0°,30°], [30°,60°], [60°,90°], as well as average values for [0°,90°]. Our method not only outperforms the SOTA approaches across various angle angles on two datasets but also achieves the best performance in the most critical average value metrics. This signifies the overall superior 3D face alignment performance of our method.

4.4. Ablation Experiment

Our approach combines relearning attention, coarse-grained, medium-grained, and fine-grained perturbation graphs for 3D face alignment. To assess the effectiveness of each module,

Table 1. The NME of 3D face alignment results on AFLW and AFLW 2000-3D.

	AFLW DataSet(21 pts)				AFLW 2000-3D DataSet(68 pts)			
Method / Absolute Yaw Angle	[0,30)	[30,60)	[60,90]	Mean	[0,30)	[30,60)	[60,90]	Mean
3DDFA (CVPR 16) [10]	5.000	5.060	6.740	5.600	3.780	4.540	7.930	5.420
Yu17 (ICCV 17) [16]	5.940	6.480	7.960	-	3.620	6.060	9.560	-
DAMDNet (ICCVW 19) [13]	4.359	5.209	6.028	5.199	2.907	3.830	4.953	3.897
MFIRRN (ICASSP 21) [17]	4.321	5.051	5.958	5.110	2.841	3.572	4.561	3.658
MARN (ICPR 21) [17]	4.306	4.965	5.775	5.015	2.989	3.670	4.613	3.757
GSRN (MMM 21) [18]	4.253	5.144	5.816	5.073	2.842	3.789	4.804	3.812
Wang22 (ICIP 22) [19]	4.212	4.935	5.787	4.978	2.906	3.572	4.677	3.774
RADAN (T-BIOM 23) [14]	4.129	4.888	5.495	4.837	2.792	3.583	**4.495**	3.623
Our	**4.097**	**4.776**	**5.420**	**4.765**	**2.736**	**3.536**	4.500	**3.591**

Table 2. Ablation results on AFLW 2000-3D and AFLW.

Method / Mean NME Error / Dataset	***AFLW 2000-3D***	***AFLW***
Our w/o RA,Coarse,Medium,Fine-graph	3.692	4.924
Our w/o Coarse,Medium,Fine-graph	3.623	4.837
Our w/o Medium,Fine-graph	3.627	4.788
Our w/o Fine-graph	3.626	4.777
Our	**3.591**	**4.765**

we conducted a comprehensive ablation study, and the results are presented in Table 2, where RA includes relearning attention with a generic data augmentation technique. The first row in Table 2 represents our method's performance without any of these modules, and its performance is relatively poor. In the second row, the inclusion of relearning attention leads to a performance improvement. Rows three to five progressively introduce coarse, medium, and fine-grained perturbation graphs, resulting in corresponding and substantial enhancements in performance relative to the second row. This demonstrates the effective guidance provided by coarse, medium, and fine-grained perturbation graphs for the initial, attention, and relearning feature map, respectively.

4.5. Evaluation of Generalization Performance

Table 3. Generalization performance evaluation results on AFLW 2000-3D and AFLW.

Method / Mean NME Error / dataset	***AFLW 2000-3D***	***AFLW***
DAMDNet	3.897	5.199
DAMDNet + Coarse-graph	3.812	5.073
DAMDNet + Coarse,Medium-graph	3.748	4.997
DAMDNet + Coarse,Medium,Fine-graph	**3.718**	**4.978**

Our multi-granularity perturbation graphs not only guide the CBAM attention network but could also serve as effective guidance for other attention networks, such as DAMDNet, RADAN, and GSRN. Table 3 displays the quantitative results as we progressively introduced coarse-grained, medium-grained, and fine-grained perturbation graphs into

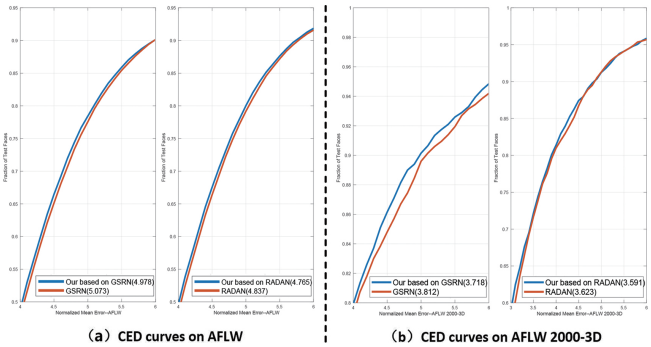

Fig. 3. The cumulative error distribution (CED) curves on AFLW (a) and AFLW 2000-3D (b).

DAMDNet. The results demonstrate that DAMDNet's performance improved significantly in a stepwise manner. Figure 3(a) and Figure 3(b) display the key sections of the CED curve for GSRN and RADAN with the addition of our multi-granularity perturbation graphs on the AFLW and AFLW 2000-3D datasets, respectively. The CED curve calculates the sample ratio (Y-axis) corresponding to each NME (Normalized Mean Error) value (X-axis). The higher a method's curve is positioned in Figure 3, the lower NME and better its performance is. After adding our multi-granularity perturbation graphs to GSRN and RADAN, their curves are almost always positioned above the original methods' curves. It highlights the effectiveness of our multi-granularity perturbation graphs in guiding other attention networks, thereby showcasing our generalization capabilities.

4.6. Qualitative Experiment

In Figure 4, we not only showcase the qualitative results using our approach described in this paper at (a), (b), and (c) but also demonstrate the qualitative results at (d), (e), and (f) when integrating our method into DAMDNet, GSRN, and RADAN, respectively. From Figure 4 (b), (d), (e), and (f),

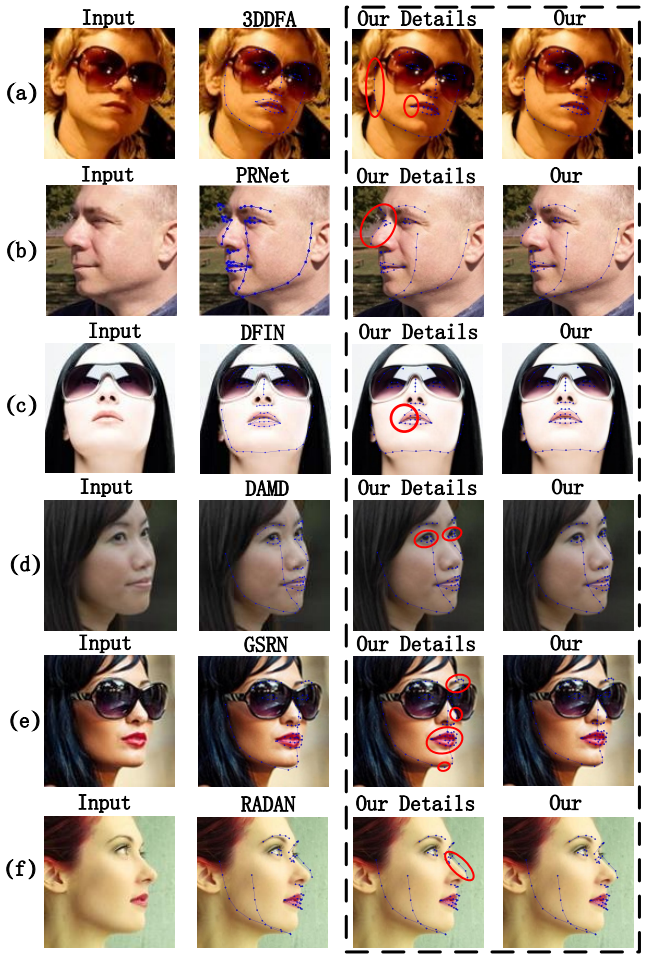

Fig. 4. Qualitative comparison results on AFLW 2000-3D.

it's evident that due to self-occlusion, partial 3D facial structure information is lost, leading existing methods to misalign based on incomplete data. In contrast, our alignment is more accurate as we infer complete 3D facial structure information through multi-granularity perturbation graphs, proving its stronger anti-interference ability under challenging scenarios.

5. CONCLUSION

To overcome the challenges brought by extreme scenarios, we propose MPRN for 3D face alignment. Through applying relearning attention guided by multi-granularity perturbation graphs, MPRN could infer the missing face information in the occluded areas, achieving accurate 3D face alignment under extreme scenarios.

6. REFERENCES

[1] Volker Blanz and Thomas Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[2] Congcong Zhu, Suping Wu, Zhenhua Yu, Xing Wang, and Hao Liu, "Multi-agent deep collaboration learning for face alignment under different perspectives," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1785–1789.

[3] Congcong Zhu, Xiaoqiang Li, Jide Li, Guangtai Ding, and Weiqin Tong, "Spatial-temporal knowledge integration: Robust self-supervised facial landmark tracking," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4135–4143.

[4] Feng Liu, Dan Zeng, Jing Li, and Qijun Zhao, "Cascaded regressor based 3d face reconstruction from a single arbitrary view image," *arXiv preprint arXiv:1509.06161*, vol. 2, 2015.

[5] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu, "Joint face alignment and 3d face reconstruction," in *European Conference on Computer Vision*. Springer, 2016, pp. 545–560.

[6] Xiang Yu, Junzhou Huang, Shaoting Zhang, and Dimitris N Metaxas, "Face landmark fitting via optimized part mixtures and cascaded deformable model," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2212–2226, 2015.

[7] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1513–1520.

[8] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, "Face alignment by explicit shape regression," *International journal of computer vision*, vol. 107, no. 2, pp. 177–190, 2014.

[9] Lei Li, Fuqiang Liu, Junyuan Wang, Yanni Wang, Yifan Chen, and Xinyu Hu, "Exploiting global and instance-level perceived feature relationship matrices for 3d face reconstruction and dense alignment," *Engineering Applications of Artificial Intelligence*, vol. 131, pp. 107862, 2024.

[10] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[11] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 534–551.

[12] Jia Deng, Xiaofei Li, Xing Wang, and Xiangzheng Li, "Deformable feature interaction network and graph structure reasoning for 3d dense alignment and face reconstruction," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[13] Lei Jiang, Xiao-Jun Wu, and Josef Kittler, "Dual attention mobdensenet (damdnet) for robust 3d face alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[14] Zhiyuan Zhou, Lei Li, Suping Wu, Xinyu Li, Kehua Ma, and Xitie Zhang, "Replay attention and data augmentation network for 3d face and object reconstruction," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023.

[15] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[16] Ronald Yu, Shunsuke Saito, Haoxiang Li, Duygu Ceylan, and Hao Li, "Learning dense facial correspondences in unconstrained images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4723–4732.

[17] Lei Li, Xiangzheng Li, Kangbo Wu, Kui Lin, and Suping Wu, "Multi-granularity feature interaction and relation reasoning for 3d dense alignment and face reconstruction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4265–4269.

[18] Xing Wang, Xinyu Li, and Suping Wu, "Graph structure reasoning network for face alignment and reconstruction," in *International Conference on Multimedia Modeling*. Springer, 2021, pp. 493–505.

[19] Jing Wang, Shikun Zhang, Fengyi Song, Ge Song, and Ming Yang, "Exploring occlusion-sensitive deep network for single-view 3d face reconstruction," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1821–1825.

[20] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 2144–2151.