3D-aware Facial Landmark Detection via Multiview Consistent Training on Synthetic Data

Libing Zeng [1*], Lele Chen[2], Wentao Bao[3*], Zhong Li[2], Yi Xu[2], Junsong Yuan[4], Nima K. Kalantari[1]
[1]Texas A&M University, [2]OPPO US Research Center, InnoPeak Technology, Inc,
[3]Michigan State University, [4]University at Buffalo

Abstract

In this supplementary document, we begin by elaborating on our proposed novel approach and providing additional analysis. Next, we present the experimental protocol and provide further numerical and visual comparisons across multiple datasets. Finally, we include an additional ablation study. Specifically, We offer further analysis of annotation inconsistency in Sec. 1, pose distribution illustration of datasets in Sec. 2, visual show of projection-prodiction distance in Sec. 3, experiment protocol in Sec. 4, more quantitative evaluations in Sec. 5, comprehensive qualitative evaluations in Sec. 6 on all four datasets against all forementioned approaches, and an extra ablation study in Sec. 7.

1. Annotation Inconsistency

As we explained in our main draft, the landmark annotation inconsistency issue is inevitable, though each annotation may seem reasonable to the given image. DAD-3DHeads [6] incorporates the FLAME fitting method to help annotation. However, it still suffers from this annotation inconsistency problem. To illustrate this, we run the procedure as shown in Fig. 1. Given an image with the annotated landmark, we first obtain a fitted mesh through the FLAME fitting, and then project the mesh to another view based on the corresponding camera parameters. Finally, we can extract landmarks from the projected mesh for the new view. As shown in the zoom-in inset of Fig. 1, the projected landmark does not fit the image. For instance, the points of the mouth area indicate the multiview inconsistency caused by annotation inconsistency. Motivated by this observation, we propose to train facial landmark detectors via multiview consistent synthetic data.

2. Dataset Distribution

In addition to the multiview consistency, another benefit of our synthetic dataset is the removal of pose distribution bias in training data. General datasets, such as DAD-3DHeads [6], are biased in the small range of head pose distribution. As shown in Fig. 2, the histograms of pitch and yaw angles of head pose in DAD-3DHeads [6] indicate approximate normal distributions with means at around zero degrees. In contrast, the pose distributions of our synthetic dataset are much more balanced across the full range, thus, help the model generate better estimations.

3. Projection-Prediction Distance

With the ready of the multiview consistent synthetic dataset of well-balanced pose distribution, we propose to incorporate multiview consistency into landmark detection by minimizing the distance of predicted landmark and projected landmark of the given image. As shown in Fig. 3, predictions and projections are denoted as green points and white points respectively, between which are the distances shown in red lines. DAD-3DNet+(Ours) generates much more consistent results indicated by shorter red lines in Fig. 3. Next, we will provide more numerical and visual results to demonstrate the superiority of our approach.

4. Experimental Protocol

Pre-training: weights of baselines (*e.g.* 3DDFA, DAD-3DNet, FAN, 3DDFA-V2) are provided by the official implementations. For example, 3DDFA is pretrained on 300W-LP, AFLWs, and DAD-3DNet is pretrained on DAD-3DHead train set; ***Fine-tuning:*** we initialize the model with the official weights (*e.g.* 3DDFA, DAD-3DNet), then conduct the 3D consistency training with our proposed DAD-3DHeads-Syn train set for another 100 epochs to obtain 3DDFA+ and DAD-3DNet+. ***Inference:*** our method only fine-tunes the existing networks to learn the 3D consistency,

[*]This work was done when Libing Zeng and Wentao Bao were interns at OPPO US Research Center, InnoPeak Technology, Inc.

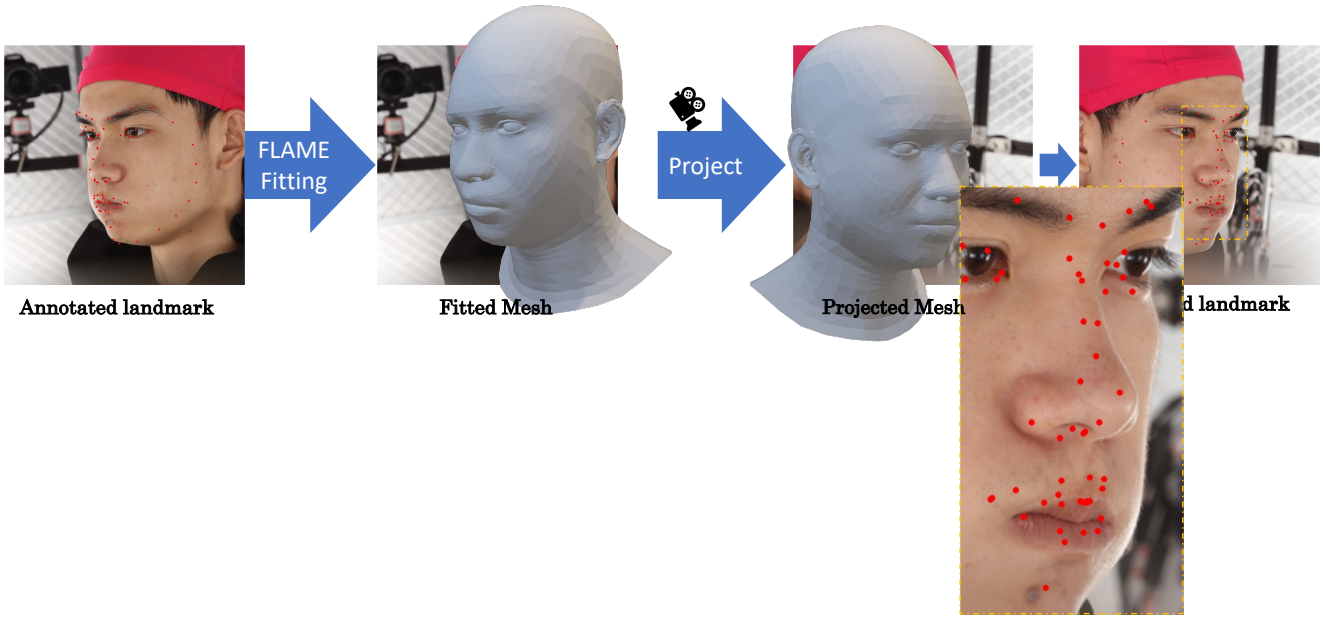


Figure 1. Multiview inconsistency caused by landmark annotation inconsistency.

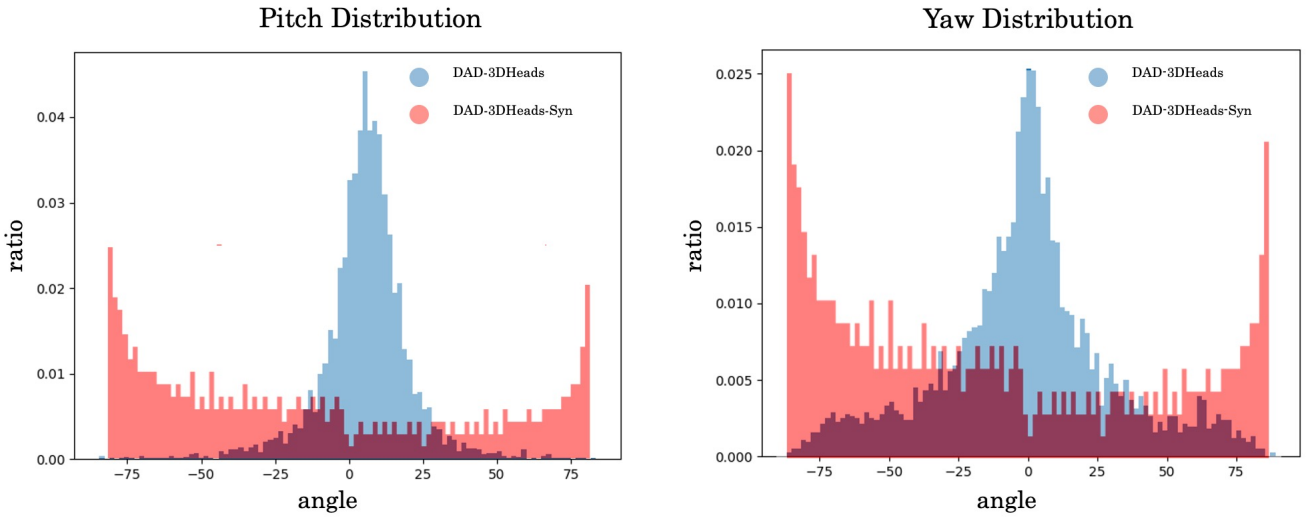


Figure 2. Pose distributions of datasets. Our DAD-3DHeads-Syn is much more balanced in pose distributions.

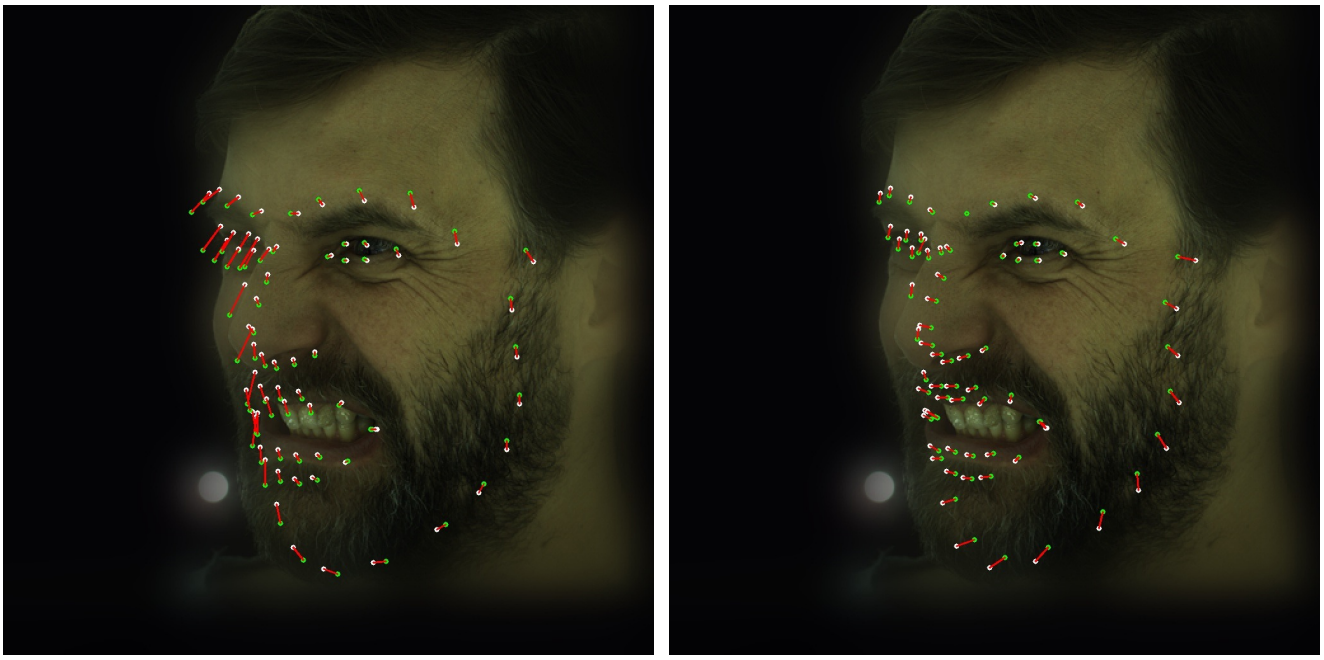
thus the inference time and memory complexity are same as the original networks.

5. More Quantitative Evaluations

In this section, we will first provide the landmark detection comparison on DAD-3DHeads-Syn covering all algorithms shown previously in the main paper. We further present pose estimation comparisons on both DAD-3DHeads-Syn and MultiFace [7] across the same existing methods.

5.1. Landmark Detection Results

In the main paper, we have shown the quantitative results of landmark detection on DAD-3DHeads [6], FaceScape [8], MultiFace [7]. Here, we provide an extra test on our DAD-3DHeads-Syn dataset. As shown in Tab. 1, the algorithms incorporated with our plug-in module, 3DDFA+(Ours) and DAD-3DNet+(Ours), generate much better numerical results than their base models. Both of them make important improvements in the NME of landmark accuracy.



DAD-3DNet DAD-3DNet+

Figure 3. Comparisons between DAD-3DNet and DAD-3DNet+(Ours) on projection-prediction distance. Predictions and projections are denoted as green points and white points respectively, between which are the distances shown in red lines.

Table 1. Facial landmark detection result (NME) on DAD-3DHeads-Syn. Lower values mean better results.

Method	DAD-3DHeads-Syn
FAN [2]	2.826
Dlib [5]	3.023
3DDFA-V2 [4]	2.840
3DDFA [3]	3.174
3DDFA+	2.981
DAD-3DNet [6]	2.373
DAD-3DNet+	2.211

5.2. Pose Estimation Results

Table 2 shows the pose estimation on DAD-3DHeads-Syn, and MultiFace [7], DAD-3DNet+(Ours) achieves 24.4%* and 19.0%* on DAD-3DHeads-Syn, and MultiFace [7] respectively. Also, the plug-in module improves significantly on almost all of metrics across pitch, roll, and yaw angles, except yaw estimation on DAD-3DHeads-Syn.

6. More Qualitative Evaluations

In this section, we provide even more extensive visual comparisons against aforementioned approaches, FAN [1], Dlib [5], 3DDFA [3], 3DDFA-V2 [4], and DAD-3DNet [6], covering all the four datasets, DAD-3DHeads-Syn, DAD-3DHeads [6], FaceScape [8], and MultiFace [7]. Specifically, we provide five pairs of comparisons, 3DDFA+(Ours) vs. 3DDFA [3], DAD-3DNet+(Ours) vs. Dlib [5], DAD-3DNet+(Ours) vs. FAN [1], DAD-3DNet+(Ours) vs. 3DDFA-V2 [4], and DAD-3DNet+(Ours) vs. DAD-3DNet [6]. As shown in Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16, Fig. 17, Fig. 18, Fig. 19, Fig. 20, Fig. 21, Fig. 22, and Fig. 23, methods with our plug-in module always show higher quality of landmark detection.

7. Ablation Study on Growing Number of the Training Data.

We analyze the impact of the extra training data (DAD-3DHead-Syn train set) used in the fine-tune stage. Specifically, we continue to train the 3DDFA on extra data for another 100 epochs to obtain **3DDFA**$_C$, similarly, we obtain **DAD-3DNet**$_C$. We test the models on FaceScape

*Head pose error drops from 7.412 to 5.958.
*Head pose error drops 14.962 to 12.578.

Table 2. Head pose estimation results (head pose error) on DAD-3DHeads-Syn, and MultiFace [7]. Lower values mean better results.

	DAD-3DHeads-Syn				MultiFace [7]			
	Pitch	Roll	Yaw	Overall	Pitch	Roll	Yaw	Overall
FAN [1]	21.938	13.093	17.002	17.344	16.840	5.913	21.074	14.609
Dlib [5]	14.525	11.472	8.272	11.430	23.506	4.303	11.093	12.966
3DDFA-V2 [4]	24.428	9.133	19.791	17.784	20.607	8.751	17.418	15.592
3DDFA [3]	24.418	9.364	19.750	17.834	29.059	12.077	17.382	19.506
3DDFA+	22.841	9.008	18.321	16.723	28.086	10.260	16.292	18.213
DAD-3DNet [6]	8.440	11.822	2.183	7.412	23.477	7.285	14.123	14.962
DAD-3DNet+	6.348	8.914	2.613	5.958	21.019	5.808	11.906	12.578

(lab-controlled), and DAD-3DHeads test set (in-the-wild). According to table below, although 3DDFA$_C$ and DAD-3DNet$_C$ are fine-tuned on extra data, their performance improvements are negligible. In contrast, after 3D-consistency training with extra data, 3DDFA+ and DAD-3DNet+ yield the best results comparing against their baselines.

Table 3. Ablation Study on Growing Number of the Training Data

	FaceScape		DAD-3DHeads	
	NME ↓	Pose ↓	NME ↓	Pose ↓
3DDFA	7.988	19.752	4.082	8.956
3DDFA$_C$	7.976	19.634	4.035	8.893
3DDFA+	**7.425**	**18.826**	**3.784**	**8.226**
DAD-3DNet	6.681	14.624	2.599	7.382
DAD-3DNet$_C$	6.662	14.557	2.584	7.186
DAD-3DNet+	**6.050**	**11.863**	**2.503**	**6.500**

References

[1] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3706–3714, 2017.

[2] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[3] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 3ddfa. https://github.com/cleardusk/3DDFA, 2018.

[4] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020.

[5] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[6] Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Jiři Matas, and Viktoriia Sharmanska. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 20942–20952, 2022.

[7] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. In *arXiv*, 2022.

[8] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 601–610, 2020.

Figure 4. Comparisons between 3DDFA [3] and 3DDFA+(Ours) on DAD-3DHeads-Syn.

Figure 5. Comparisons between Dlib [5] and DAD-3DNet+(Ours) on DAD-3DHeads-Syn.

Figure 6. Comparisons between FAN [1] and DAD-3DNet+(Ours) on DAD-3DHeads-Syn.

Figure 7. Comparisons between 3DDFA-V2 [4] and DAD-3DNet+(Ours) on DAD-3DHeads-Syn.

Figure 8. Comparisons between DAD-3DNet [6] and DAD-3DNet+(Ours) on DAD-3DHeads-Syn.

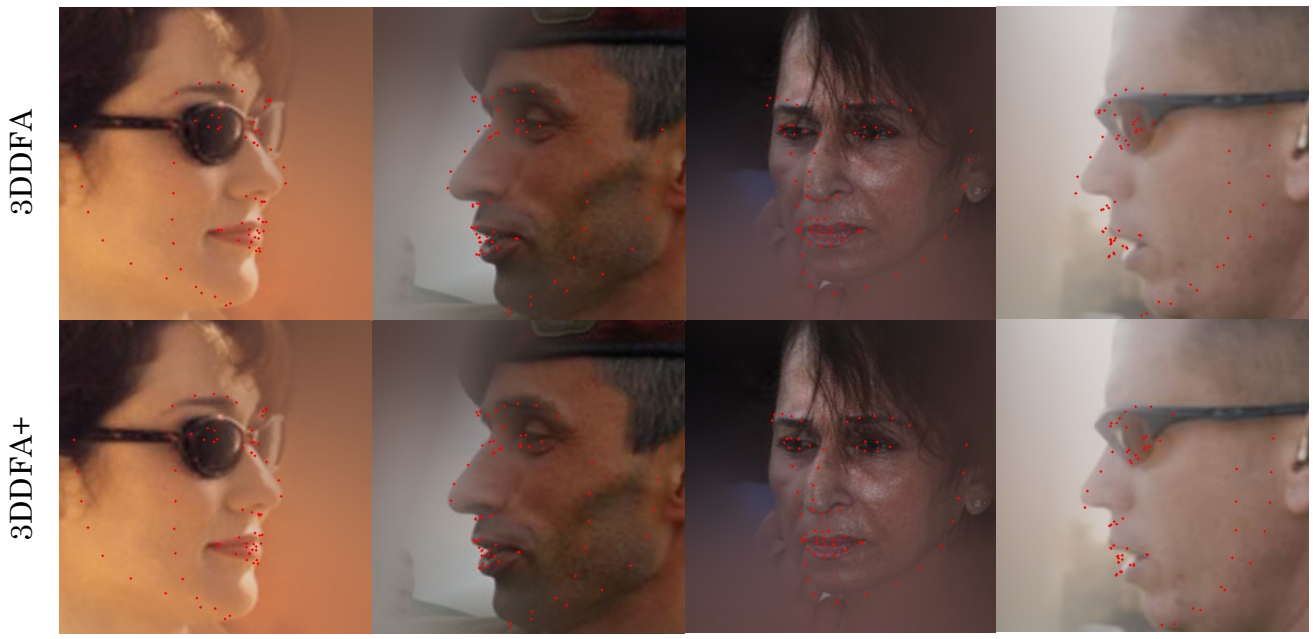


Figure 9. Comparisons between 3DDFA [3] and 3DDFA+(Ours) on DAD-3DHeads [6].

Figure 10. Comparisons between Dlib [5] and DAD-3DNet+(Ours) on DAD-3DHeads [6].

Figure 11. Comparisons between FAN [1] and DAD-3DNet+(Ours) on DAD-3DHeads [6].

Figure 12. Comparisons between 3DDFA-V2 [4] and DAD-3DNet+(Ours) on DAD-3DHeads [6].

Figure 13. Comparisons between DAD-3DNet [6] and DAD-3DNet+(Ours) on DAD-3DHeads [6].

Figure 14. Comparisons between 3DDFA [3] and 3DDFA+(Ours) on FaceScape [8].

Figure 15. Comparisons between Dlib [5] and DAD-3DNet+(Ours) on FaceScape [8].

Figure 16. Comparisons between FAN [1] and DAD-3DNet+(Ours) on FaceScape [8].

Figure 17. Comparisons between 3DDFA-V2 [4] and DAD-3DNet+(Ours) on FaceScape [8].

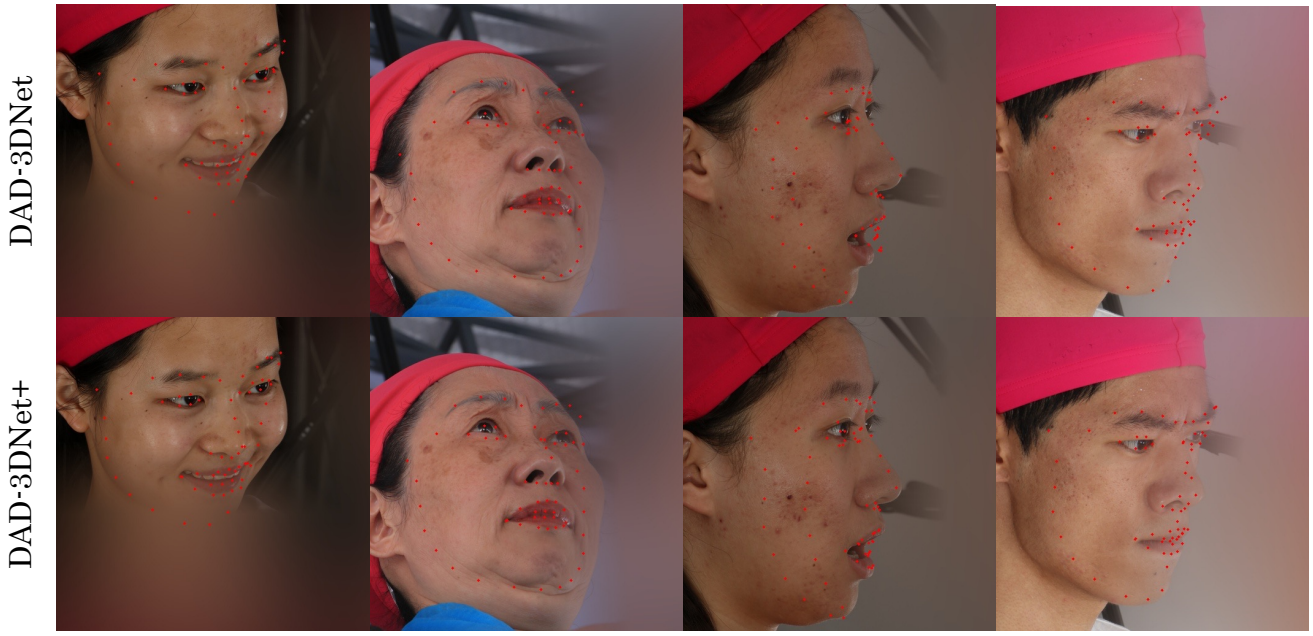


Figure 18. Comparisons between DAD-3DNet [6] and DAD-3DNet+(Ours) on FaceScape [8].

Figure 19. Comparisons between 3DDFA [3] and 3DDFA+(Ours) on MultiFace [7].

Figure 20. Comparisons between Dlib [5] and DAD-3DNet+(Ours) on MultiFace [7].

Figure 21. Comparisons between FAN [1] and DAD-3DNet+(Ours) on MultiFace [7].

Figure 22. Comparisons between 3DDFA-V2 [4] and DAD-3DNet+(Ours) on MultiFace [7].

Figure 23. Comparisons between DAD-3DNet [6] and DAD-3DNet+(Ours) on MultiFace [7].