DSFNet: Dual Space Fusion Network for Occlusion-Robust 3D Dense Face Alignment – Supplementary Material –

Heyuan Li[1], Bo Wang[2], Yu Cheng[1], Mohan Kankanhalli[1], Robby T. Tan[1]
[1] National University of Singapore [2]CtrsVision
liheyuan@u.nus.edu, hawk.rsrch@gmail.com, e0321276@u.nus.edu,
mohan@comp.nus.edu.sg, robby.tan@nus.edu.sg

We arrange the supplemental material in the following sections. In Sec. 1, we provide the details of our method for training, including the details of training data and data augmentation. In Sec. 2, we present the details of the proposed AFLW2000-3D-occlusion dataset. In Sec. 3, we provide more quantitative results on AFLW2000-3D with re-annotated labels provided by LS3D-W [1]. In Sec. 4, we compare our method with more existing methods qualitatively.

1. Training Data

We train our model on the 300W-LP dataset [15], where each 2D image has its corresponding ground truth 3D label. The 3D label is the 3DMM coefficients obtained by fitting 3DMM to 68 manually labeled 2D facial landmarks through the Multi-Features Framework [7]. Because the 3D label is recovered from sparse landmarks, it is not always precise, especially in local facial regions. Additionally, to collect sufficient profile view face images, [15] generates synthetic profile images by 3D rotation around yaw angles, which leads to artifacts in the side face regions, shown in Fig. 1. Since our image space branch relies more on low-level image features, the above two attributes of 300W-LP make this dataset cannot fully reflect the superiority of our method. Therefore, our method has the potential to achieve better performance if more precisely labeled real-world datasets are available in the future.

For data augmentation, we follow previous works [5, 8]. We augment the images by random rotation, translation, scaling, and color channel scaling. Specifically, the rotation ranges from -90 to 90 degree angles, the translation ranges from 10 percent of input size, the scale is from 0.95 to 1.05, color channel scale is from 40 percent of the original color value. We also introduce random synthetic occlusions and some of them have texture from the Describable Textures Dataset [2].

Figure 1. Typical samples from 300W-LP. The left is an image before data augmentation. The right is a corresponding image after data augmentation, which 3D rotates the initial image around the yaw angle. Please note the artifacts in the side face region.

2. AFLW2000-3D-occlusion

We construct AFLW2000-3D-occlusion to evaluate a method's robustness to occlusion. It consists of three subsets: Naturally Occluded Dataset (NOD), Color Synthetically Occluded Dataset (CSOD), and NatOcc Synthetically Occluded Dataset (NSOD).

A. Naturally Occluded Dataset (NOD) This subset contains 127 automatically selected images from AFLW2000-3D. Illustrated in Fig. 2, for every image in AFLW2000-3D, we compute a visible rate v to represent its visibility. We first project the ground-truth face mesh to the image plane and get the projected contour P. Then we use an off-the-shelf face skin segmentation algorithm [14] to get the visible facial region F. The visible rate is computed as the ratio of the area of the intersection of the projected contour and the face skin segmentation to the area of the projected contour:

$$v = \frac{s(P \cap F)}{s(P)}, \tag{1}$$

Profile images always have smaller v, because hairs al-

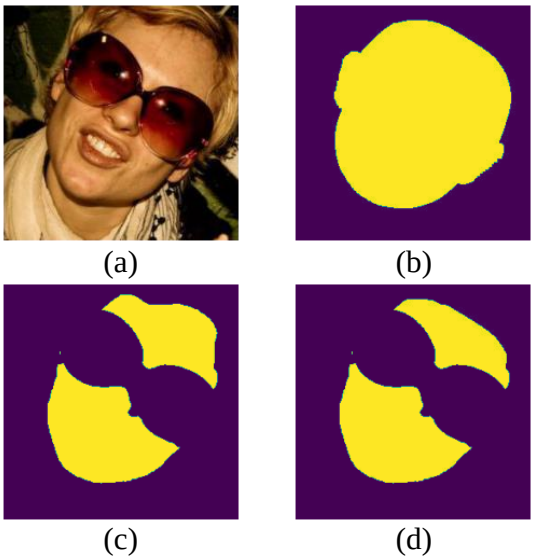

Figure 2. Illustration of computing visible rate. (a) An initial image. (b) P: The corresponding contour of the ground truth 3D face mesh projected to the image plane. (c) F: Face skin region obtained by [14]. (d) $P \cap F$: The intersection of the projected contour and the face skin region.

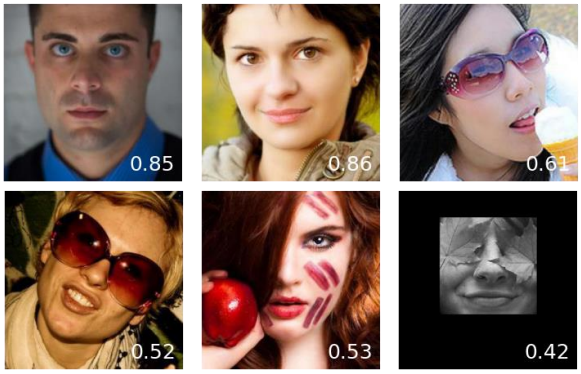

Figure 3. Images from AFLW2000-3D with different visible rates. Images in the second row are selected to the Naturally Occluded Dataset (NOD).

ways cover the ear region which accounts for a large proportion of the projected contours especially viewed from the side. To prevent entangling with the problem of large yaw angles, we filter out samples with yaw angles larger than 60 degrees and only select samples with visible rates less than 0.6. Fig. 3 shows some samples with different visible rates.

B. Color Synthetically Occluded Dataset (CSOD) This subset contains 6000 images augmented from AFLW2000-3D. We follow the occlusion patterns in [11]. Every image in AFLW2000-3D is occluded by three different types of color synthetic occluder: 1. Single-square occlusion, which is a square-shaped occlusion. 2. Double-square occlusion, which consists of two intersected square-shaped occlusions. 3. Column-shaped occlusion, which is a rectangular occlusion whose height is equal to the side length of the input image. Examples of this subset are shown in Fig. 4.

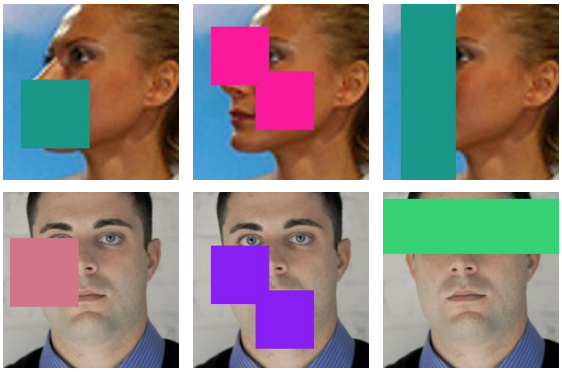

Figure 4. Samples from Color Synthetically Occluded Dataset (CSOD). Left: Single-square occlusion. Middle: Double-square occlusion. Right: Column-shaped occlusion.

Figure 5. Samples from NatOcc Synthetically Occluded Dataset (NSOD).

C. NatOcc Synthetically Occluded Dataset (NSOD) This subset contains 2000 images augmented from AFLW2000-3D. We use Naturalistic Occlusion Generation (NatOcc) technique from [12] to add daily objects on top of images from AFLW2000-3D. Examples of this subset are shown in Fig. 5.

3. Evaluation on Re-annotated AFLW2000-3D

Since AFLW2000-3D is semi-automatically annotated, there are some cases that have inaccurate annotations. As shown in Fig. 6, some of our method's predictions with large NME errors actually are caused by inaccurate annotations. LS3D-W [1] provides more accurate 2D annotations of 68 facial landmarks. We evaluate our method on 2D

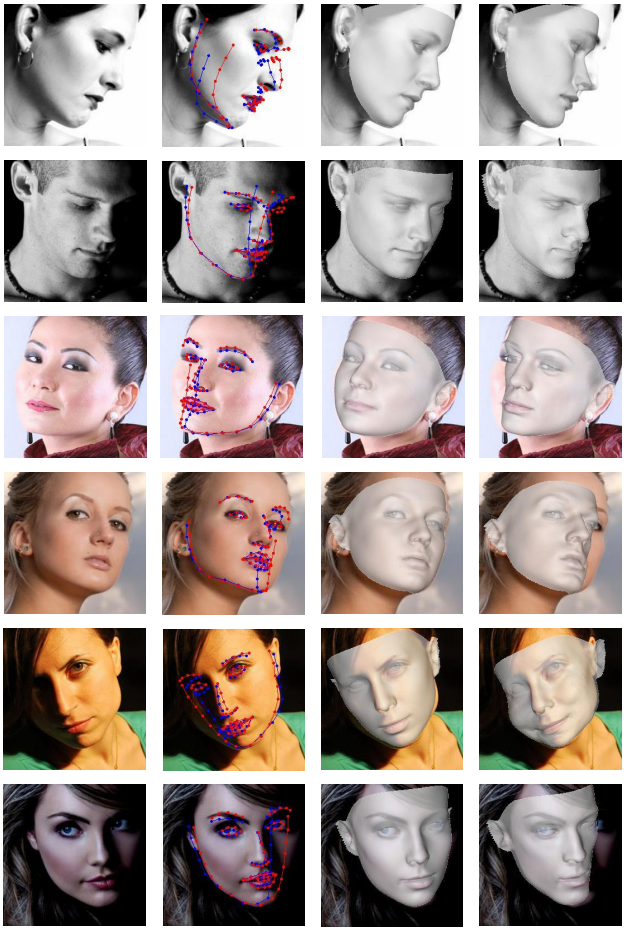

Figure 6. Comparing results from our DSFNet and ground truth on AFLW2000-3D dataset. From left to right are the input image, facial landmarks comparison (blue for ground truth and red for our method), our method's prediction, and ground truth.

Method	2D Sparse Face Alignment			
	0 to 30	30 to 60	60 to 90	Mean
DHM [10]	2.28	3.10	6.95	4.11
3DDFA [15]	2.84	3.52	5.15	3.83
PRNet [5]	2.35	2.78	4.22	3.11
MGCNet [9]	2.72	3.12	3.76	3.20
Deep3D [3]	2.56	3.11	4.45	3.37
3DDFA-V2 [16]	2.84	3.03	4.13	3.33
SADRNet [8]	2.31	2.46	3.41	2.73
SynergyNet [13]	**2.05**	2.49	3.52	2.69
DSFNet-f (ours)	2.11	**2.31**	**3.28**	**2.57**

Table 1. Sparse face alignment (68 landmarks) on AFLW2000-3D Reannotated. The NME (%) for faces with different yaw angles are reported.

sparse alignment using these re-annotated labels. As shown in Tab. 1, our method has a better performance compared to existing methods.

4. Additional Qualitative Results

We compare our method qualitatively with PRN [5], MGCNet [9], DECA [4], 3DDFA-V2 [6], SynergyNet [13] and SADRNet [8] in Fig. 7 and Fig. 8. Images are from AFLW2000-3D and [12].

The qualitative results demonstrate our method's robustness to severe occlusion and large view angles. Even if a large proportion of the face is occluded, our method can filter out non-facial regions to prevent interference from occluders and deduce the whole face geometry from the visible regions. Therefore, our method conducts reasonable predictions in severe occluded cases where other methods' predictions are distorted due to occlusion. Because our method fully utilizes low-level image information which is less variant to geometry transformation, our method shows high robustness to large view angles. We notice MGCNet and SynergyNet have certain robustness to large view angles, and DECA and SADRNet are robust to some kinds of occlusion. Our method is the only one that is robust to both problems.

References

[1] 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. 1, 2

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1

[3] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 3

[4] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021. 3

[5] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 1, 3

[6] 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. 3

[7] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 986–993. IEEE, 2005. 1

[8] Zeyu Ruan, Changqing Zou, Longhai Wu, Gangshan Wu, and Limin Wang. Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruc-

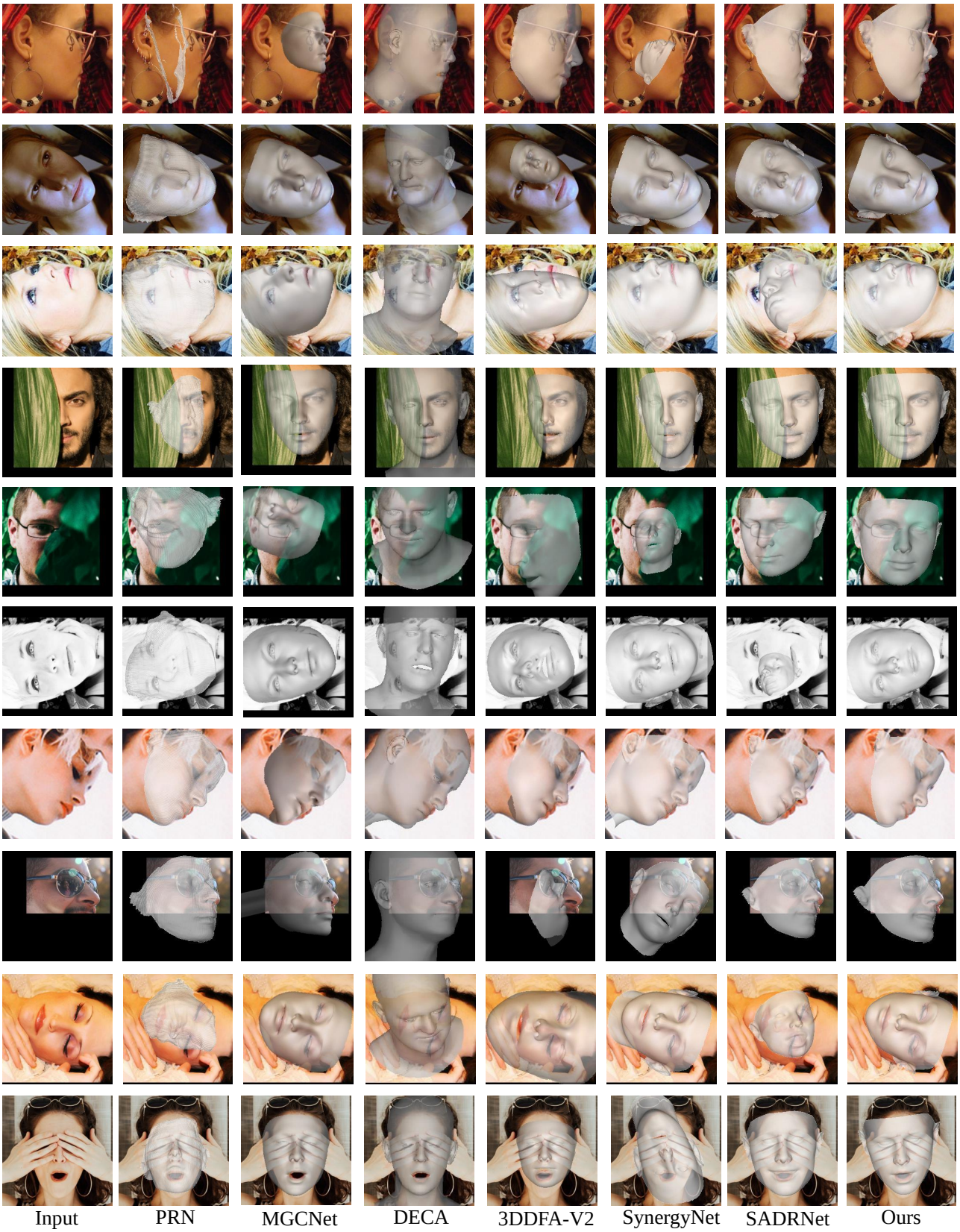

Figure 7. Qualitative Comparison on AFLW2000-3D and [12].

Figure 8. (Continued) Qualitative Comparison on AFLW2000-3D and [12].

tion. *IEEE Transactions on Image Processing*, 30:5793–5806, 2021. 1, 3

[9] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. *arXiv preprint arXiv:2007.12494*, 2020. 3

[10] Bin Sun, Ming Shao, Siyu Xia, and Yun Fu. Deep evolutionary 3d diffusion heat maps for large-pose face alignment. In *BMVC*, page 256, 2018. 3

[11] Hitika Tiwari, Vinod K Kurmi, KS Venkatesh, and Yong-Sheng Chen. Occlusion resistant network for 3d face reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 813–822, 2022. 2

[12] Kenny T. R. Voo, Liming Jiang, and Chen Change Loy. Delving into high-quality synthetic face occlusion segmentation datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2, 3, 4, 5

[13] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *2021 International Conference on 3D Vision (3DV)*, 2021. 3

[14] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 1, 2

[15] 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*, pages 146–155, 2016. 1, 3

[16] Xiangyu Zhu, Fan Yang, Di Huang, Chang Yu, Hao Wang, Jianzhu Guo, Zhen Lei, and Stan Z Li. Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In *ECCV*, pages 343–358. Springer, 2020. 3