Towards Fast, Accurate and Stable 3D Dense Face Alignment -Supplementary Material-

Jianzhu Guo[1,2⋆][0000−0002−8493−3689], Xiangyu Zhu[1,2⋆][0000−0002−4636−9677], Yang Yang[1,2][0000−0003−0559−5464], Fan Yang[3][0000−0003−4348−3148], Zhen Lei[1,2†][0000−0002−0791−189X], and Stan Z. Li[4][0000−0002−2961−8096]

[1] CBSR&NLPR, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] College of Software, Beihang University
[4] School of Engineering, Westlake University
{jianzhu.guo,xiangyu.zhu,yang.yang,zlei,szli}@nlpr.ia.ac.cn,
fanyang@buaa.edu.cn

A. Checkerboard Artifacts

The checkerboard artifacts of dense vertices regression [2,4] are shown in Fig. 1.

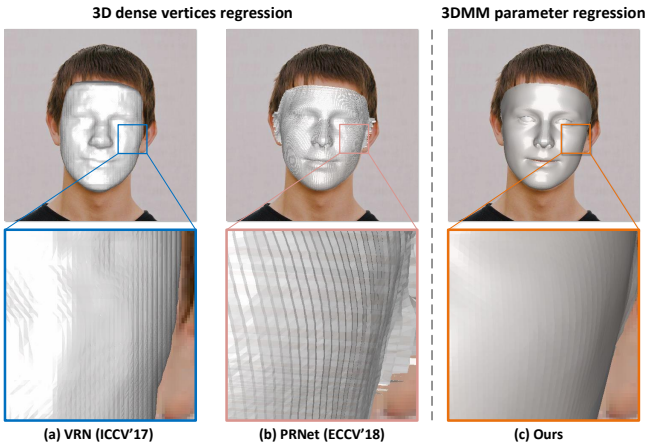

Fig. 1: A result of PRNet [2], VRN [4] and our method. The upper row is the dense mesh overlapped with the original image, the bottom row is the local details enlarged (better view in the electronic version). Local details show that the output mesh of PRNet is jagged and has checkerboard artifacts, VRN also has slight checkerboard artifacts, and our result is the smoothest.

⋆ Equal contribution.
† Corresponding author.

B. Impact of Dimension Reduction

The NME error heatmap caused by different size of shape and expression dimensions is shown in Fig. 2.

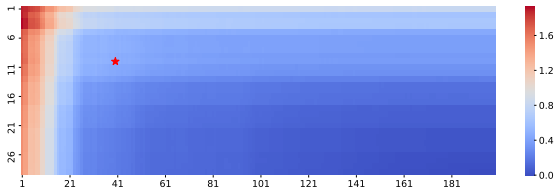

Fig. 2: The 29×199 heatmap of NME (%) with different dimensions of shape and expression parameter (x-axis is shape, y-axis is expression). When the dimensions are set to [40, 10] (shown as the red star marker), the NME increase is about 0.4%, which is acceptable.

C. Implementation Details

Our experiments are based on PyTorch [1]. During training, all faces are cropped and resized to 120×120, then normalized by subtracting 127.5 and being divided by 128. We use SGD with a batch size B of 128 to optimize the network, with the weight decay of 0.0005 and momentum of 0.9. For our model *MobileNet (M+R+S)*, k is 100 for the meta-joint optimization, and for the short-video-synthesis, each still image is synthesized with $n = 8$ frames and the perturbation settings are: $\Delta s \in [0.95, 1.05]$, $\Delta\theta \in [-3°, 3°]$, $\Delta t1, \Delta t2 \in [-5, 5]$ pixels, $\Delta\phi, \Delta\gamma \in [-5°, 5°]$.

D. Generalization and Scaling-up Ability

We compare the performance and speed with different architectures and scaling-up options in Table 1 and Fig. 3. Note that the proposed methods are all applied on them. The results in Table 1 and Fig. 3 reveal the generalization and scaling-up ability of our proposed methods: (i) when equipped with a more powerful backbone like ResNet-22, our methods perform better, which demonstrates the generalization ability across architectures; (ii) with different multipliers and input size, our methods show the great scaling-up ability. Users can choose the proper scaling-up option according to their need. Besides, MobileNet-V3 [3] performs better than MobileNet and MobileNet-V2 [5], and MobileNet-V3 ×0.5 gives similar performance to PRNet with only 27.4M MACs, indicating that it is 225x faster than PRNet (6190M MACs) theoretically.

Table 1: Comparisons of performance and speed on AFLW2000-3D, AFLW and Menpo-3D with different channel numbers and backbones. We ignore the reconstruction time (1ms in CPU) of 3D dense vertices in this table.

Backbone	AFLW2000-3D	AFLW	Menpo-3D	Params	MACs	Inference Time (CPU)
PRNet [2]	3.62	4.77	1.90 / 0.54	13.4M	6190M	175ms
PRNet ×0.25	4.77	6.54	-	0.84M	434M	48.7ms
PRNet ×0.125	5.24	7.06	-	0.21M	134M	38.4ms
ResNet-22	**3.49**	**4.32**	**1.67** / **0.45**	18.45M	2663M	67.5ms
MobileNet	**3.51**	**4.43**	**1.71** / **0.48**	3.27M	183.5M	6.2ms
MobileNet ×0.75	3.62	4.49	1.74 / 0.50	1.86M	105.9M	4.2ms
MobileNet-V3 ×0.5	3.61	4.48	1.80 / 0.51	1.65M	27.4M	3.4ms

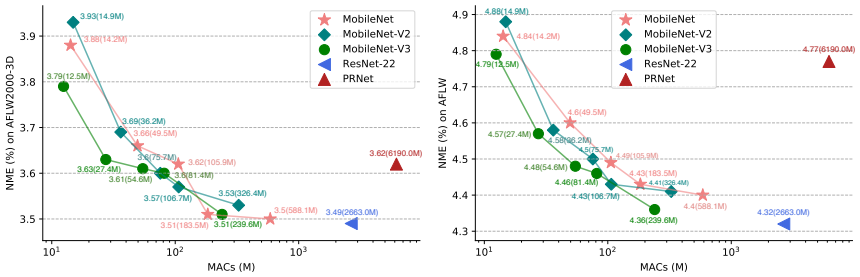

Fig. 3: The trade-off between the computation complexity MACs and NME (%) on AFLW2000-3D and AFLW. MobileNet, MobileNet-V2 and MobileNet-V3 (large mode) use multipliers 0.25, 0.5, 0.75 and 1 with input size 120 or 128 and the multiplier 1 with input size 224. ResNet uses 120. PRNet is shown here for comparison. Lower NME (%) is better.

E. Qualitative Results

We present more qualitative results (Fig. 4) for comparisons with VRN [4] and PRNet [2] on AFLW2000-3D and AFLW. The supplementary video presents 3D sparse and dense face alignment results.

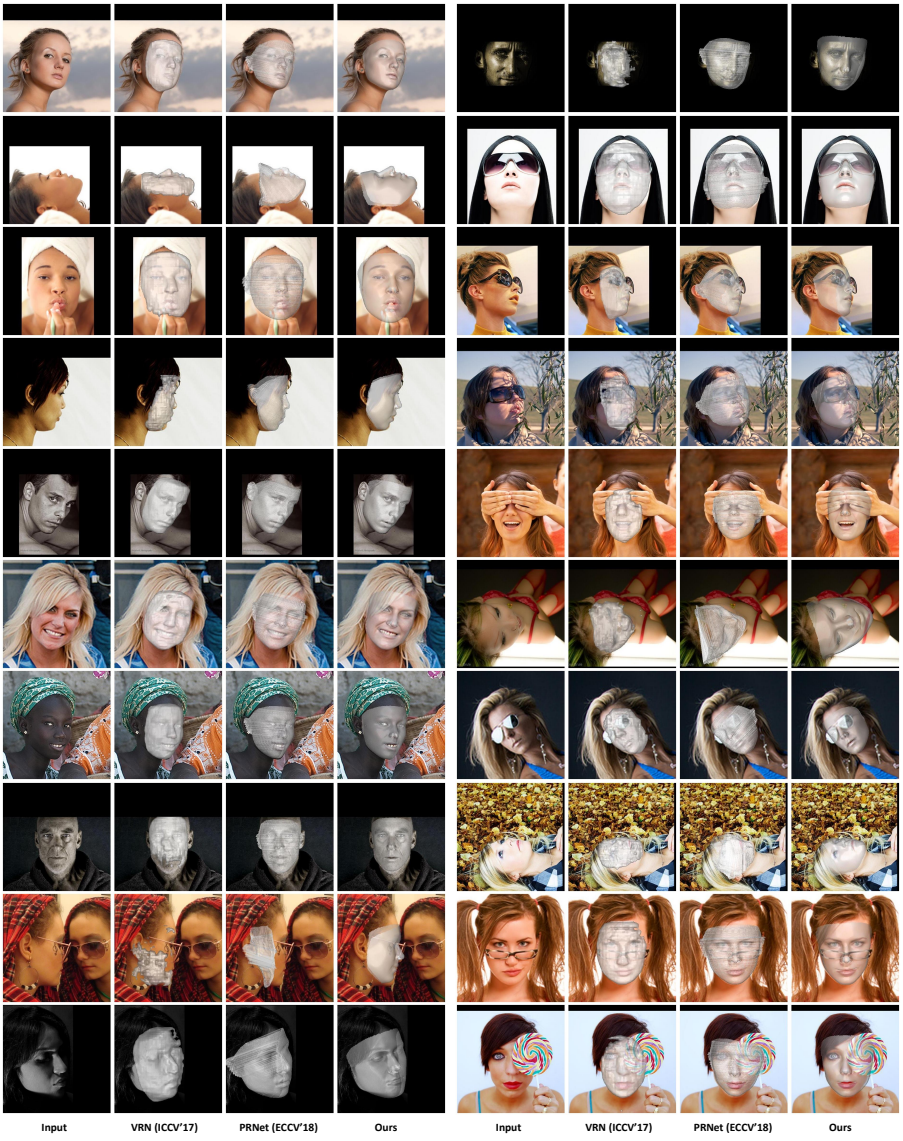

Fig. 4: Qualitative results on AFLW2000-3D and AFLW. Our results are from the *MobileNet (M+R+S)* model, which runs at over 50fps on a single CPU core. Please zoom in to see local details. (better view in the electronic version)

References

1. Adam, P., Sam, G., Soumith, C., et al.: Automatic differentiation in PyTorch. In: NIPS Autodiff Workshop (2017) 2
2. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: ECCV (2018) 1, 3
3. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. arXiv preprint arXiv:1905.02244 (2019) 2
4. Jackson, A., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In: ICCV (2017) 1, 3
5. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR (2018) 2