Supplementary Material for Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression

Aaron S. Jackson[1] Adrian Bulat[1] Vasileios Argyriou[2] Georgios Tzimiropoulos[1]

[1] The University of Nottingham, UK [2] Kingston University, UK

[1]{aaron.jackson, adrian.bulat, yorgos.tzimiropoulos}@nottingham.ac.uk
[2] vasileios.argyriou@kingston.ac.uk

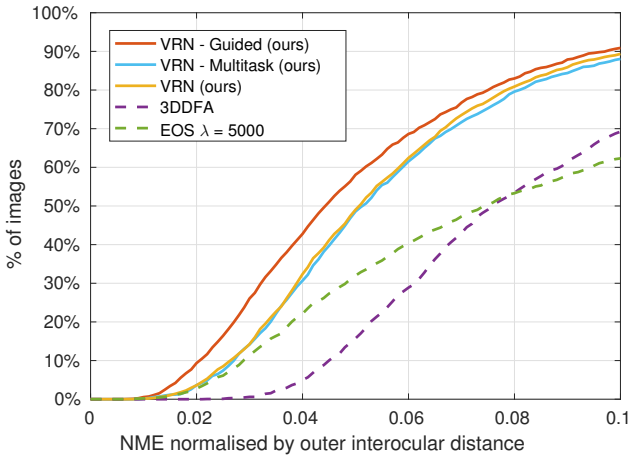

Figure 1: NME-based performance on the in-the-wild AFLW2000-3D dataset, where ICP has been used to remove the rigid transformation. The proposed *Volumetric Regression Networks*, and EOS and 3DDFA are compared.

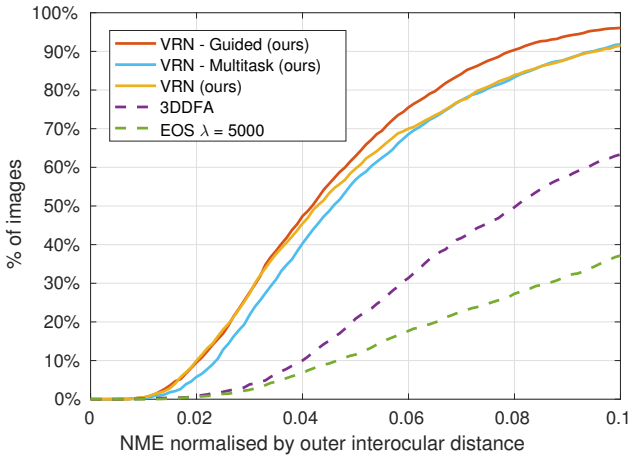

Figure 2: NME-based performance on our large pose and expression renderings of the BU4DFE dataset, where ICP has been used to remove the rigid transformation. The proposed *Volumetric Regression Networks*, and EOS and 3DDFA are compared.

1. Results with ICP Registration

We present results where ICP has been used not only to find the correspondence between the groundtruth and predicted vertices, but also to remove the rigid transformation between them. We find that this offers a marginal improvement to all methods. However, the relative performance remains mostly the same between each method. Results on AFLW2000 [5], BU4DFE [3] and Florence [1] can be seen in Figs. 1, 2 and 3 respectively. Numeric results can be found in Table 1.

Table 1: Reconstruction accuracy on AFLW2000-3D, BU4DFE and Florence in terms of NME where ICP has been used to remove the rigid transformation. Lower is better.

Method	AFLW2000	BU4DFE	Florence
VRN	0.0605	0.0514	0.0470
VRN - Multitask	0.0625	0.0533	0.0439
VRN - Guided	**0.0543**	**0.0471**	**0.0429**
3DDFA [5]	0.1012	0.1144	0.0784
EOS [2]	0.0890	0.1456	0.1200

2. Results on 300VW

To demonstrate that our method can work in unconstrained environments and video, we ran our *VRN - Guided* method on some of the more challenging Category C footage from the 300VW [4] dataset. These videos are challenging usually for at least one of the following reasons: large pose, low quality video, heavy motion blurring and

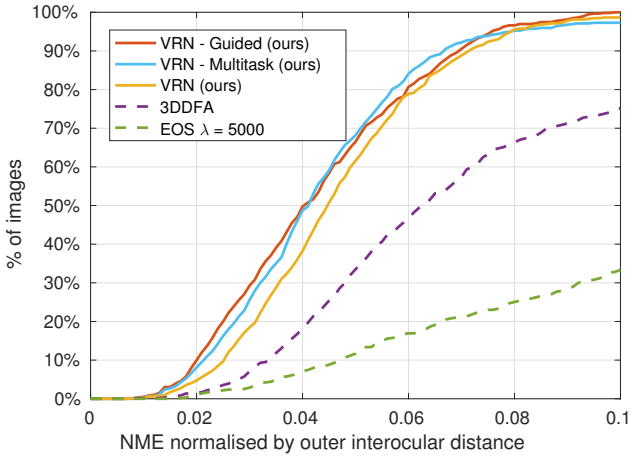

Figure 3: NME-based performance on our large pose renderings of the Florence dataset, where ICP has been used to remove the rigidla transformation. The proposed *Volumetric Regression Networks*, and EOS and 3DDFA are compared.

occlusion. We produce these results on a frame-by-frame basis, *each frame is regressed individually without tracking*. Videos will be made available on our project website and can also be found in the supplementary material.

3. Additional qualitative results

This section provides additional visual results and comparisons. Failure cases are shown in Fig. 4. These are mostly unusual poses which can not be found in the training set, or are not covered by the augmentation as described Section 3.4 of our paper. In Fig. 5 we show some visual comparison between *VRN* and *VRN - Guided*. These differences are quite minor. Finally, in Fig. 6 we show some typical examples from our renderings of BU-4DFE [3] and Florence [1], taken from their respective testing sets.

References

[1] A. D. "Bagdanov, I. Masi, and A. Del Bimbo. The florence 2d/3d hybrid face datset. In *Proc. of ACM Multimedia Int.l Workshop on Multimedia access to 3D Human Objects (MA3HO11)*. ACM, ACM Press, December 2011.

[2] P. Huber, G. Hu, R. Tena, P. Mortazavian, W. P. Koppen, W. Christmas, M. Rätsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework.

[3] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[4] S. Zafeiriou, G. Tzimiropoulos, and M. Pantic. The 300 videos in the wild (300-vw) facial landmark tracking in-the-wild challenge. In *ICCV Workshop*, 2015.

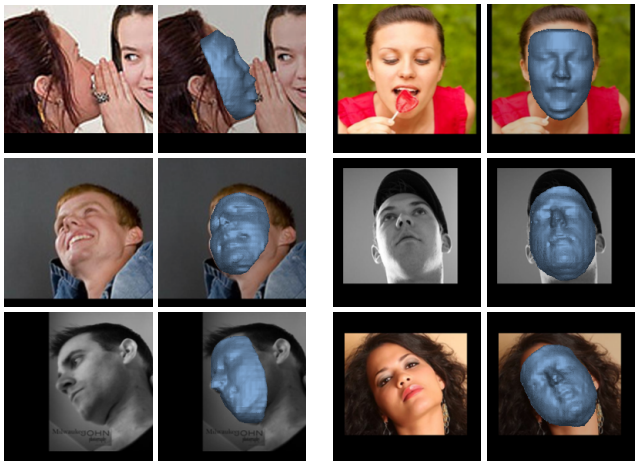

Figure 4: Some failure cases on AFLW2000-2D from our *VRN - Guided* network. In general, these images are difficult poses not seen during training.

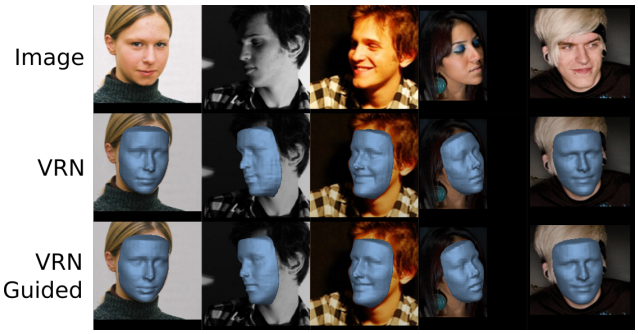

Figure 5: A visual comparison between *VRN* and *VRN - Guided*. The main difference is that the projection of the volume has a better fit around the shape of the face.

Figure 6: Examples of rendered images from (a) BU4DFE (containing large poses and expressions), and (b) Florence (containing large poses) datasets.

[5] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. 2016.