Supplementary Materials for DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image

1. DAD-3DHeads Dataset

The images in DAD-3DHeads dataset are anonymised without additional metadata. The results of labeling in the form of 3D head model do not contain any private or sensitive information. The data gathered is not being used for identification purposes or in connection with any other personal data.

1.1. Dataset card

As stated in Section 3.2 of the main paper, along with the dataset of images and annotations, we provide additional information per image such as gender, age, illumination conditions, image quality, pose, presense of expression, and occlusions (see Fig. 7 here and Fig.3 in the main paper). We also provide more visual examples from the dataset, and demonstrate the DAD-3DHeads diversity in terms of ethnicity, age, gender, camera pose, expressions (see Fig. 8).

We use multiple sources to construct our dataset, among which WIDER FACE dataset [12], the Adience dataset [4], Compound Facial Expressions of Emotions Dataset [3], WFLW [11], AFW [14], Helen [8], LFPW [2].

1.2. DAD-3DHeads annotations accuracy

We provide more visualizations of the results of DAD-3DHeads annotations compared to GT scans on neutral faces from NoW dataset [9] in Fig. 9. This is in addition to (and in the context of) Figure 4 in the main paper.

1.3. Annotation process

We provide here more details along with visuals to prevent possible misunderstandings around the labeling process. The full video is available via the project webpage.

As was already mentioned in the Sec.3.1, the annotators do not explicitly control or label either the 3DMM parameters or the blendshapes. They also do not have to disambiguate between identity and expression. The annotators only have to ”pin” a 3d head model - a mesh - to the head image. They do so iteratively having the initial generic mesh optimized after every ”pin” being placed (see Fig. 1). The nonlinear optimization is performed, indeed, over the shape and expression parameters, along with the pose.

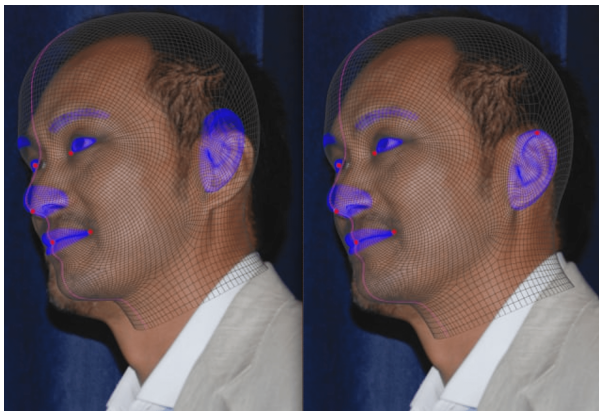

Figure 1. The mesh is being deformed due to the under-the-hood optimization after the ”pin” is placed on the ear.

Figure 2. The rendered texture here has ”holes” due to the occlusions, but can also be ”torn” if the ”pins” are placed poorly.

After each step, the annotators can inspect if the fitted mesh aligns well with the image in several ways: they might view (i) the reprojected set of selected landmarks that correspond to the recognizible features on any face or head (see Fig. 6 as an example), such as contours of the eyes (eyelash line), lips, etc.; (ii) the image rendered onto the mesh as a texture given the fitting - it helps to inspect if there appeared texture ”holes” due to poor labeling or occlusions (see Fig. 2); (iii) the mesh itself is visible in 360^o to inspect

if the skull shape is not deformed (see Fig. 2, right). These measures help to partially overcome limitations introduced by the absence of camera parameters for the input images, thus possible ambiguities caused by the effects of perspective projection.

Another important issue is that our annotations consist only of the 3D vertices and transformation matrices, we do not interpret the resulting mesh w.r.t. identity, expression, or any other ambiguous and ill-defined over a single-image input feature. The concerns arise as there might be cases, e.g., faces with extreme expressions, where it is impossible to perfectly detect the shape of the head with neutral expression based on a single image. Indeed, the 3D scanners would provide an accurate 3D head model that the manual annotation cannot guarantee. However, they operate under controlled capture, i.e., *3D scans have been in-the-lab instead of in-the-wild*. We provide the community with the complementary data. It has a *known* trade-off in accuracy, but the performance is sufficient for many applications that operate in-the-wild.

2. Experimental results on DAD-3DHeads benchmark

2.1. 3D Landmark localization

We evaluate the state-of-the-art methods such as JVCR [13], FaceSynthetics [10], 3DDFA-v2 [6], and proposed DAD-3DNet on DAD-3DHeads benchmark for the task of 3D Landmark Localization. We report normalized mean error (NME) for the predicted landmarks reprojected onto the image plane (see Tab. 1). We analyse the NME metric on full test dataset as well as across challenging subgroups (atypical poses, compound expressions, heavy occlusions). *DAD-3DNet shows superior performance in all cases.*

When computing NME of the competitor methods, we only use the images where the landmarks have been localized, otherwise the NME is ill-defined. See examples where the landmarks are not localized in Fig. 10, along with other challenging cases.

Performance on the "Expr." subset is better than the overall average (Table 1) for all of the methods. We attribute this to the fact that for heavily occluded faces and large extreme poses, where the landmarks are not clearly visible (and therefore the emotions), the images are labeled as "neutral" by default.

2.2. DAD-3DNet performance across multiple subgroups

We present an extension of Table 4 from the main paper, performing in-depth analysis of the results across multiple subgroups. We define subgroups w.r.t. *pose* (front, side, atypical), *age* (child, young, middle age, senior), *image quality* (high, low), *occlusions* (true, false), *expressions* (neutral, non-neutral), *lightning* (standard, non-standard). The results of the DAD-3DNet model are reported in 2. This analysis shows robustness of the proposed approach across various conditions (distribution shifts) in-the-wild.

Model	**NME↓**			
	Overall	Pose	Expr.	Occl.
3DDFA-V2 [6]	3.580	7.630	3.168	3.195
FaceSynthetics [10]	4.363	15.781	3.159	4.934
JVCR [13]	4.455	12.514	3.843	4.949
DAD-3DNet	**2.302**	**6.049**	**1.748**	**2.036**

Table 1. **3D Landmark Localization on DAD-3DHeads benchmark.** We compute the normalized mean error (NME, the lower the better) on full test dataset as well as on challenging atypical poses (Pose), compound expressions (Expr.) and heavy occlusions (Occl.) subsets. DAD-3DNet performs superior in all cases.

.

Pose	**NME**	***Z5* Accuracy**	**Chamfer Dist.**	**Pose Error**
front	1.496	0.965	3.146	0.089
side	2.257	0.952	3.180	0.143
atypical	6.190	0.916	4.027	0.343

Age	**NME**	***Z5* Accuracy**	**Chamfer Dist.**	**Pose Error**
child	1.662	0.960	3.546	0.103
young	2.421	0.953	3.178	0.150
middle	2.438	0.955	3.393	0.133
senior	1.756	0.958	2.989	0.113

Quality	**NME**	***Z5* Accuracy**	**Chamfer Dist.**	**Pose Error**
high	2.065	0.957	3.194	0.129
low	4.755	0.928	3.643	0.259

Occlusion	**NME**	***Z5* Accuracy**	**Chamfer Dist.**	**Pose Error**
True	4.242	0.9436	3.784	0.1986
False	2.134	0.955	3.182	0.1359

Expression	**NME**	***Z5* Accuracy**	**Chamfer Dist.**	**Pose Error**
Non-neutral	1.156	0.959	3.412	0.116
Neutral	1.644	0.950	3.088	0.164

Lighting	**NME**	***Z5* Accuracy**	**Chamfer Dist.**	**Pose Error**
Standard	2.350	0.954	3.182	0.142
Non-standard	2.235	0.954	3.569	0.142

Table 2. In-depth analysis of the benchmark results reported in Table 4 of the main paper across multiple subgroups such as *camera pose*, *age*, *image quality*, *occlusions*, *expressions*, *lighting*. This analysis shows robustness of the proposed approach across various conditions (distribution shifts) in-the-wild.

2.3. 3D Head Pose Estimation

The comparison of DAD-3DNet with the state-of-the-art 3D Head Pose Estimation method img2pose [1] is provided in Tab. 3. We calculate two metrics on the rotation matrices R_1 (GT) and R_2 (prediction), similar to Table 2 in the main paper: (i) Frobenius norm of the matrix $I - R_1 R_2^T$, and (ii) the angle in axis-angle representation of $R_1 R_2^T$.

Method	$\|\|I - R_1 R_2^T\|\|_F$	Angle error (degrees)
Img2Pose [1]	0.226	9.122
DAD-3DNet	**0.138**	**5.360**

Table 3. **3D Head Pose estimation on DAD-3DHeads benchmark**. DAD-3DNet outperforms state-of-the-art img2pose [1]. The measure of $R_1 R_2^T$ deviation from identity matrix lies in the $(0, 2\sqrt{2})$ range [7].

2.4. Failure cases

DAD-3DNet still sometimes fails on some cases of severe occlusions, extremely low quality, or very atypical poses (like faces upside-down) (see Fig. 3). Our current model was trained with no augmentations, this creates a suitable venue for future explorations.

Figure 3. Failure cases of DAD-3DNet on DAD-3DHeads benchmark, 3D Landmark Localization.

3. Miscellaneous

***Z_n* Metrics.** We measure performance of DAD-3DNet via Z_n for different values of n, see the results in Fig. 4. The accuracy does not change dramatically with n, so for DAD-3DHeads benchmark we use $n = 5$ as a trade-off between computational complexity and robustness.

"Head" and "face" mesh vertices. In the paper, we refer to two subsets of the FLAME mesh vertices: "head" and "face". The former ones are used in the loss terms calculation (see Shape+Expression loss and Reprojection loss in Sec. 4.2) and in the Z_n accuracy measurements for DAD-3DHeads benchmark (see Sec. 5.1), and the latter ones - in the Chamfer distance measurements for DAD-3DHeads benchmark (see Sec. 5.1). We provide visual examples of what these subsets represent on meshes with various head shape and face expressions in Fig. 5.

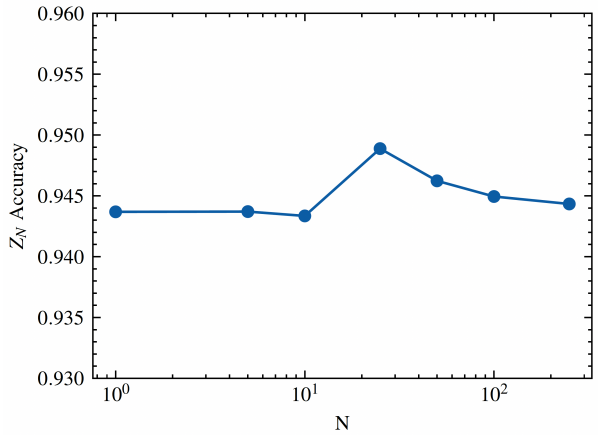

Figure 4. Z_N accuracy of DAD-3DNet for different values of n. x-axis is in log-scale.

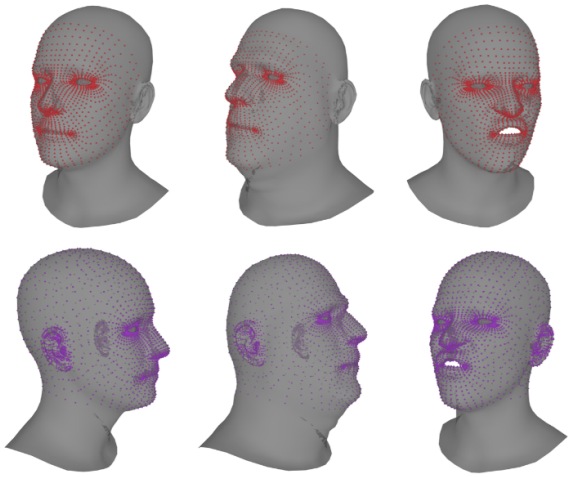

Figure 5. **"Face" and "head" subsets of FLAME mesh vertices.** Upper row: "face", capture frontal part of the head without ears. Lower row: "head", capture the head without neck. The eyeballs are excluded in both.

Various number of landmarks. As DAD-3DHeads dataset is **dense**, it allows for training different models, localizing many more than the usual 68 landmarks [5]. This flexibility saves the human annotator efforts, because the data should not be relabeled every time a different setup is needed. Moreoever, DAD-3DNet training pipeline allows for inference on any subset of head vertices, given its 3DMM prediction branch, as they can be subsampled after the entire mesh is predicted (see examples of different landmark subsets in Fig. 6).

Figure 6. DAD-3DHeads dataset allows for flexibly choosing the desired landmark subset for predicting as many dense landmarks as needed. Left to right: 68 landmarks [5], 191 landmarks, 445 landmarks.

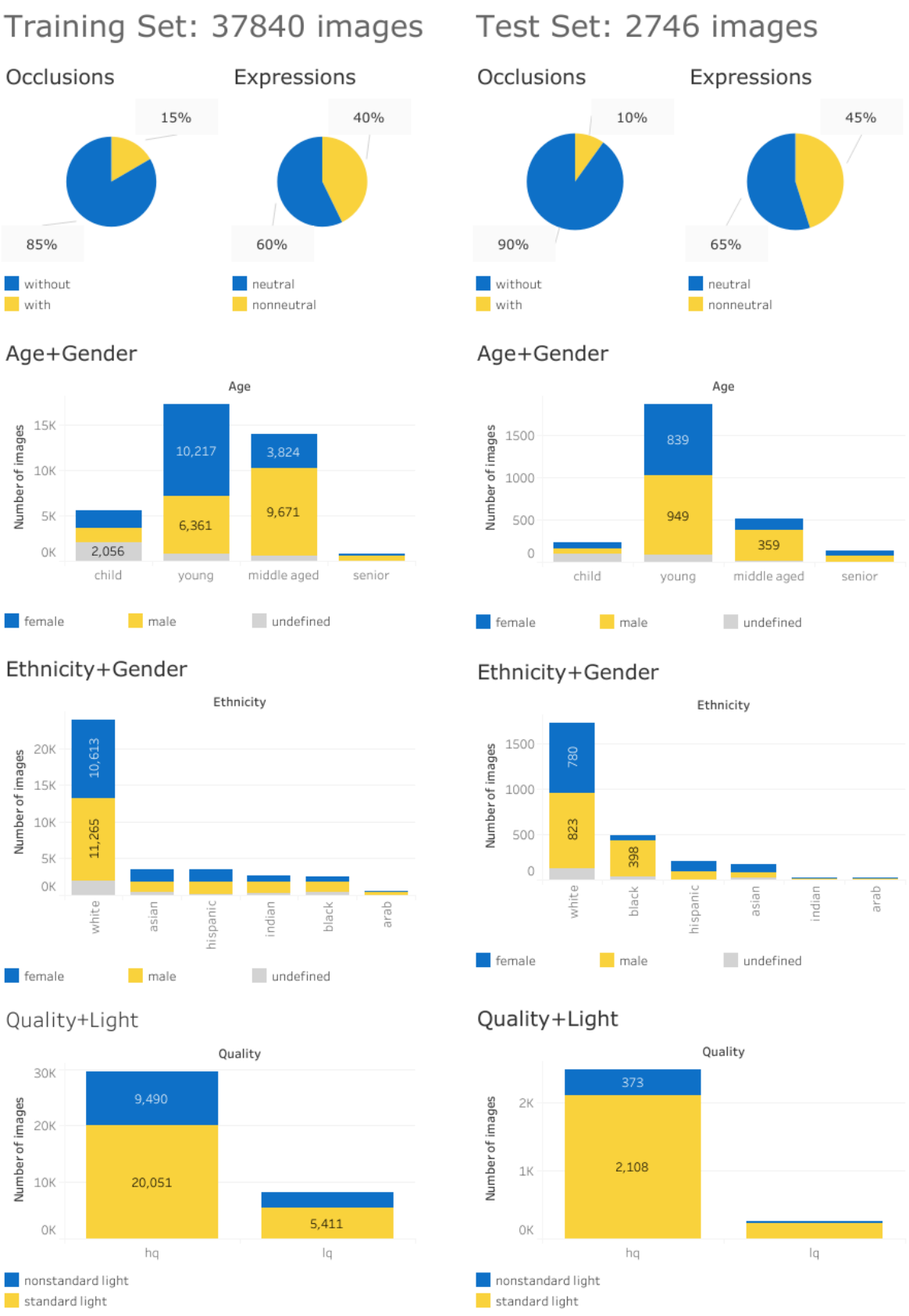

Figure 7. Attribute labels (gender, age, illumination, and image quality) and the distribution across ethnic groups in DAD-3DHeads.

Figure 8. **DAD-3DHeads dataset, more visual examples.** The source images cover large variation in poses [2, 4, 9, 11, 14, 16, 18-20], expressions [3, 6, 7, 8, 15], occlusions [1-3, 8, 19], non-standard illumination conditions [6, 13, 14], low image quality [2, 4, 13].

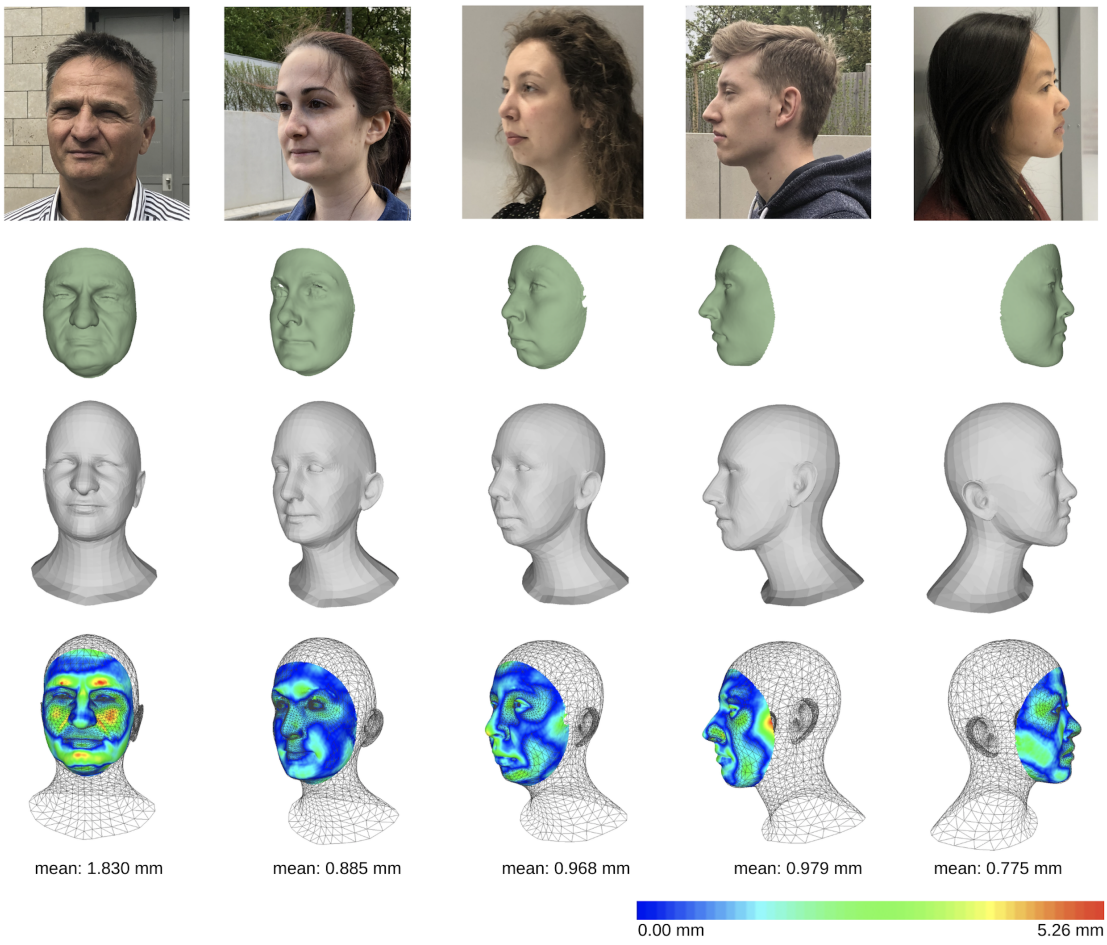

Figure 9. **DAD-3DHeads accuracy** on selected samples from the NoW dataset. **First row:** input image; **second row:** GT scan; **third row:** the result of our annotation; **fourth row:** alignment of the mesh (wireframe) and the GT scan (with color-coded errors overlayed).

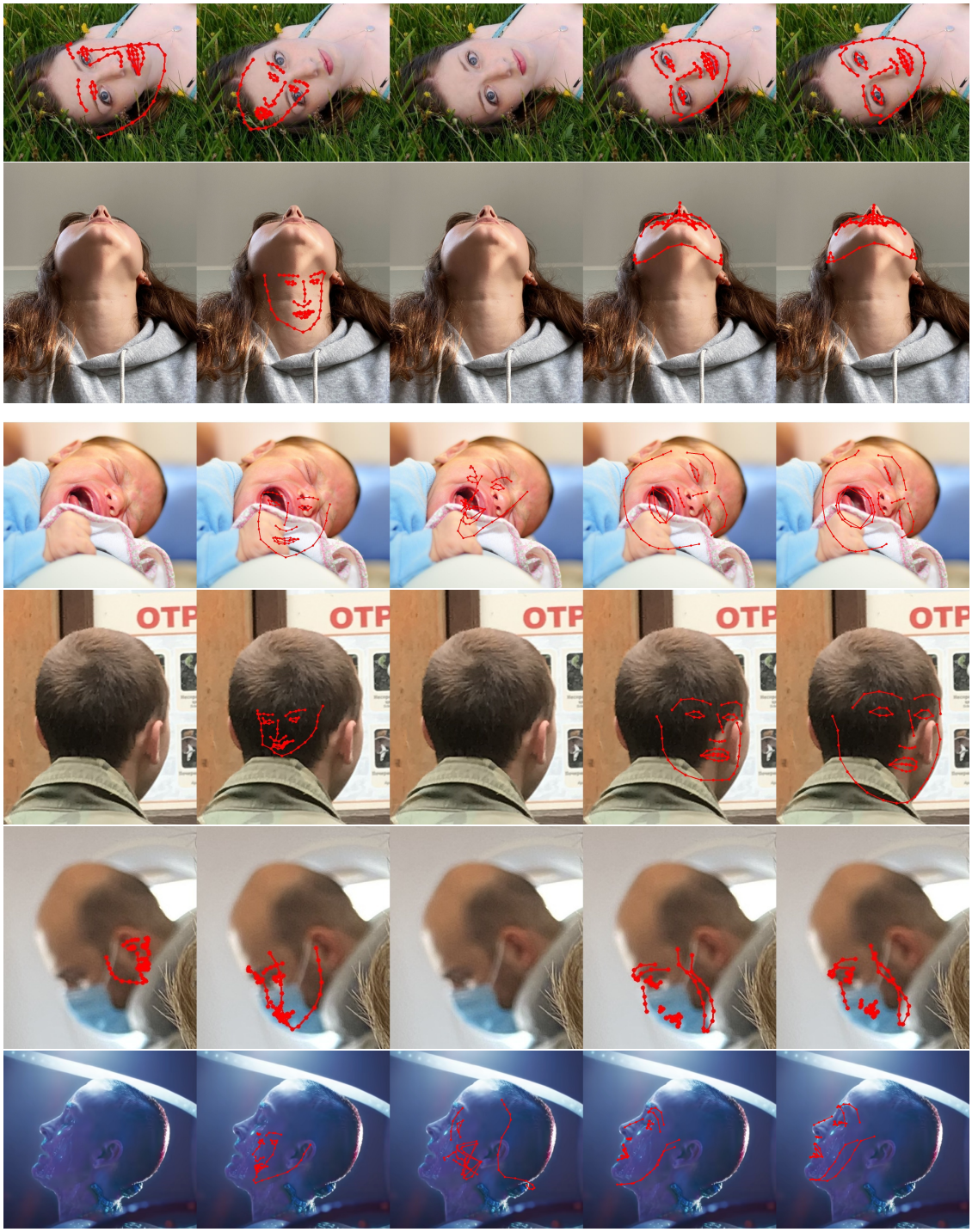

Figure 10. Qualitative comparison of DAD-3DNet and state-of-the-art methods on challenging cases from DAD-3DHeads benchmark. **Left to right:** 3DDFA-v2 [6], FaceSynthetics [10], JVCR [13], DAD-3DNet (ours), ground truth.

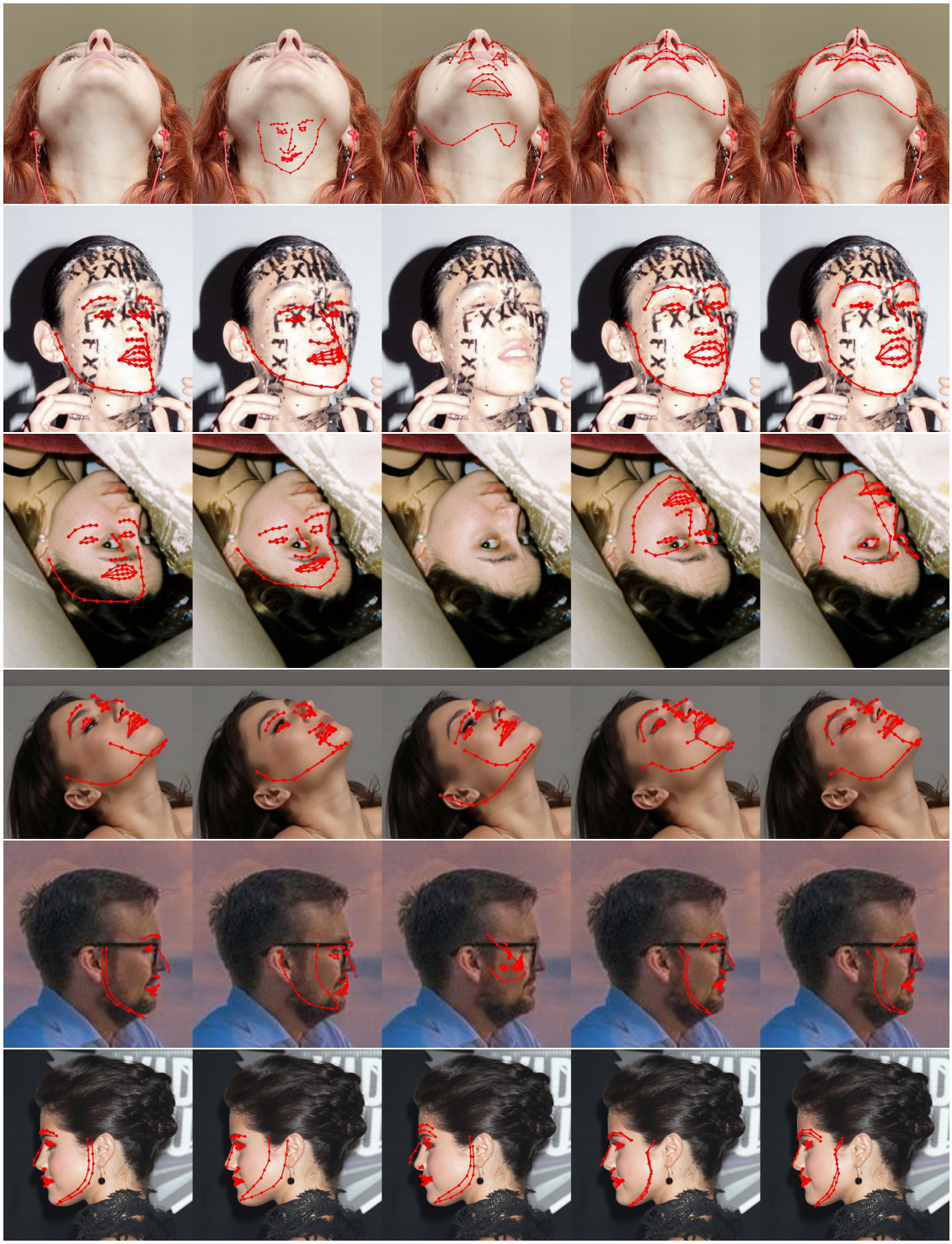

Figure 10. Qualitative comparison of DAD-3DNet against state-of-the-art methods on challenging cases from DAD-3DHeads benchmark (cont.) **Left to right:** 3DDFA-v2 [6], FaceSynthetics [10], JVCR [13], DAD-3DNet (ours), ground truth.

Figure 10. Qualitative comparison of DAD-3DNet against state-of-the-art methods on challenging cases from DAD-3DHeads benchmark (cont.) **Left to right:** 3DDFA-v2 [6], FaceSynthetics [10], JVCR [13], DAD-3DNet (ours), ground truth.

References

[1] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7617–7627, 2021. 2, 3

[2] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2930–2940, 2013. 1

[3] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. 1

[4] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. 1

[5] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 3, 4

[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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 8, 9, 10

[7] Du Q Huynh. Metrics for 3d rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164, 2009. 3

[8] Vuong Le, Jonathan Brandt, Zhe Lin, and Lubomir Bourdev. Interactive facial feature localization. 10 2012. 1

[9] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 1

[10] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3681–3691, October 2021. 2, 8, 9, 10

[11] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[12] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[13] Hongwen Zhang, Qi Li, and Zhenan Sun. Joint voxel and coordinate regression for accurate 3d facial landmark localization. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2202–2208. IEEE, 2018. 2, 8, 9, 10

[14] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012. 1