H3D-Net: Few-Shot High-Fidelity 3D Head Reconstruction Supplementary Material

Eduard Ramon[1,2] Gil Triginer[1] Janna Escur[1] Albert Pumarola[3] Jaime Garcia[1]
Xavier Giro-i-Nieto[2,3] Francesc Moreno-Noguer[3]
[1]*Crisalix SA* [2]*Universitat Politècnica de Catalunya* [3]*Institut de Robòtica i Informàtica Industrial, CSIC-UPC*
crisalixsa.github.io/h3d-net

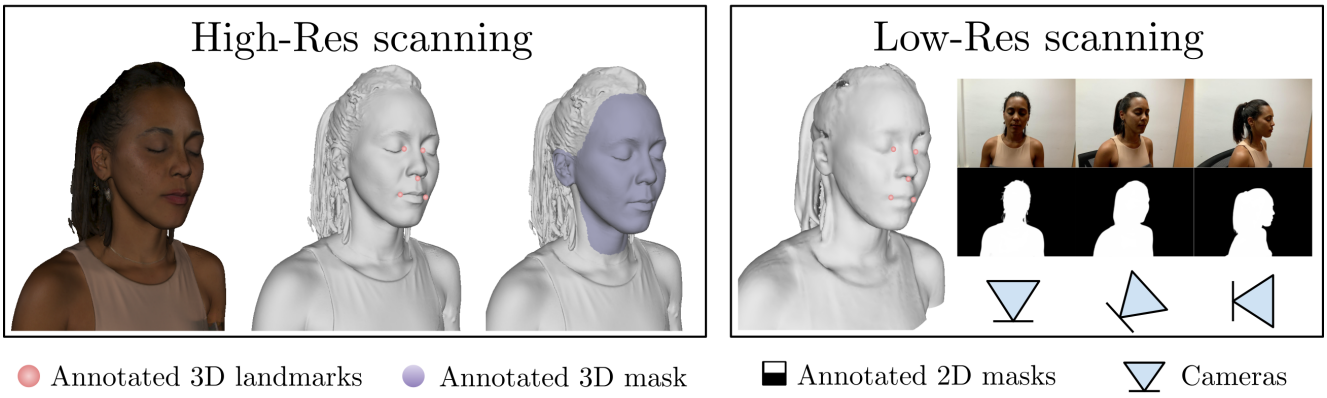

Figure 1. **Acquisition and annotations** for each of the two simultaneous scanning processes per subject. **Left.** Data acquired with the high resolution scanner, from which we obtain the 3D ground truth geometry. **Right.** Data acquired with the low resolution scanner, from which we obtain the images, masks and cameras. The coarse geometry and the annotated 3D landmarks are used to align both scenes.

In this document, we describe the H3DS dataset, which was introduced in the main text, and provide more details about our evaluation procedure. In Section 1, we explain the acquisition setup, and provide a visual overview of H3DS. In Section 2, we explain how the data from H3DS and 3DFAW [5] have been used for evaluation. Finally, in the accompanying video, we provide more examples of 3D head reconstructions obtained with our model H3D-Net, as well as visual comparisons to the baselines IDR [8], MVFNet [7] and DFNRMVS [2].

1. H3DS dataset

H3DS is a new high-resolution dataset for evaluating 3D head reconstruction methods. The dataset contains multiple scenes, each representing the head of a different subject. For each scene, we provide a high resolution 3D scan of the head as a triangular mesh, and a set of RGB images with associated masks and camera parameters. Next, we provide more details about how the dataset has been collected, and show some visual examples.

1.1. Data acquisition setup

During the data acquisition process, each individual is placed on a rotating stand that rotates 360 degrees. We perform two scanning processes simultaneously on each subject, one responsible for acquiring the high resolution 3D ground truth and a second one to obtain multiview images with associated camera parameters. Simultaneity is a necessary condition to ensure that the recorded images faithfully represent the captured ground truth geometry. This data acquisition process, which we detail below, is illustrated in Figure 1.

We use the Artec Eva scanner[1], which has a precision of 0.1 mm and a resolution of 0.2 mm, to obtain a fully-textured, high-resolution (High-Res) 3D mesh of the head, including hair and shoulders, that is established as the 3D ground truth. At the same time, in a parallel scanning process (Low-Res), we obtain a set of images and their associated cameras, as well as a coarse geometry that is only used to align the High-Res and Low-Res reconstructed

[1] https://www.artec3d.com/portable-3d-scanners/artec-eva

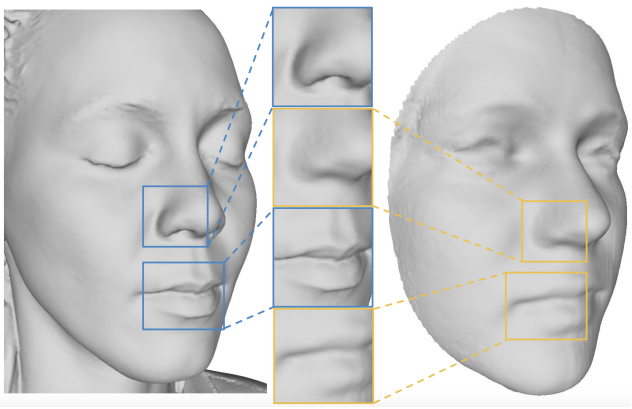

Figure 2. **Comparison** between H3DS and 3DFAW ground truth geometries.

scenes. For the latter, we use a Structure Sensor [2] mounted on an iPad Pro. Note that while the two scanners are based on structured light, their different wavelengths allow us to use them at the same time without risking interference. Finally, both scenes are aligned using 5 manually annotated 3D landmarks on the coarse and fine geometries, and applying the Iterative Closest Point (ICP) algorithm [1]. Furthermore, we semi-automatically annotate the foreground/background masks on each of the images, as well as a facial mask on the 3D ground truth for evaluation purposes.

We scan a total of 10 subjects, 5 female and 5 male, with a number of views per scene ranging from 65 to 75. The whole process takes approximately 1 hour per subject, which includes the raw data acquisition and the 2D and 3D annotations. We evaluate H3D-Net using the 10 scanned subjects from the H3DS dataset available at submission time, a number of test samples comparable to other recent SOTA works [8, 4, 3]. Yet, we plan increasing the number of H3DS scenes, and make them publicly available.

1.2. Dataset analysis

In Figure 2 we provide a qualitative comparison of our acquired ground truth geometries with those provided in the 3DFAW [5] dataset. Recall that the scans from 3DFAW were estimated via multi-view stereo optimization, which is less precise than using structured light. We compare two ground truth geometries from the H3DS and 3DFAW datasets. Despite the geometries necessarily corresponding to different subjects, since both datasets have been acquired on disjoint groups of people, the higher definition of the H3DS geometry can be appreciated.

Finally, in Figure 3 we provide a visualization of some H3DS scenes. As it can be observed, the captured geometry contains very fine details, such as hair, beards and upper body clothes. The data covers full 360 degrees around each subject. We also provide carefully annotated 2D masks for each image of each scene.

[2] https://structure.io

2. Evaluation details

We have benchmarked H3D-Net and the baselines on the 3DFAW and H3DS datasets. In the following, we provide more details on the evaluation on each dataset.

3DFAW. We have used 10 scenes selected from the 3DFAW training set, since it is the only of the three available splits that provides 3D ground truth. Note that we do not use the data for any pre-training, but only for evaluation purposes. The scenes are split into 5 male and 5 female subjects, with identifiers `subject-313`, `subject-315`, `subject-316`, `subject-318`, `subject-320`, `subject-321`, `subject-323`, `subject-325`, `subject-338` and `subject-345`. For each scene, we decompress the video `iPhone_{ID}.MOV` using ffmpeg with a sampling rate of 5 Hz. Then, we manually select three views at approximately -30, 0, and 30 degrees, and semi-automatically annotate the ground truth masks for each image. Finally, we run a 3DMM-based monocular 3D reconstruction algorithm to obtain a coarse mesh and a camera pose for each of the views [6]. The coarse geometry is used to align the predicted cameras with the prior geometry pre-learnt in H3D-Net. The aligned predicted cameras are then used as ground truth by H3D-Net and IDR.

H3DS. For the evaluation using H3DS, we use different subsets of all the acquired images, ground truth masks and ground truth cameras. The subset configurations are defined by their yaw angles as follow: $\mathcal{V}_3 = \{0, \pm 45\}$, $\mathcal{V}_4 = \{\pm 45, \pm 90\}$ and $\mathcal{V}_N = \{\frac{360}{N} i\}_{i=1}^{N}$ for $N = 8, 16, 32$. Given that for the smallest subsets $\mathcal{V}_3$ and $\mathcal{V}_4$ the back of the head is not visible, for alignment purposes we annotate in the 3D ground truth the set of vertices belonging to the facial region as shown in Figure 1-left. In order to evaluate full-head 3D reconstructions from H3D-Net or IDR, we align the reconstruction and the ground truth geometries using ICP only on the vertices defined by this region.

References

[1] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *CVPR*, 2007. 2

[2] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *CVPR*, 2020. 1

[3] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *arXiv preprint arXiv:2003.08934*, 2020. 2

Figure 3. **H3DS** dataset examples. On the left, a subset of the the views per scene, which range full 360 degrees. On the right, the textured high definition ground truth geometries.

[4] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 2

[5] Rohith Krishnan Pillai, László Attila Jeni, Huiyuan Yang, Zheng Zhang, Lijun Yin, and Jeffrey F Cohn. The 2nd 3d face alignment in the wild challenge (3dfaw-video): Dense reconstruction from video. In *ICCV Workshops*, 2019. 1, 2

[6] Eduard Ramon, Janna Escur, and Xavier Giro-i Nieto. Multi-view 3d face reconstruction in the wild using siamese networks. In *ICCV Workshops*, 2019. 2

[7] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, 2019. 1

[8] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *NeurIPS*, 33, 2020. 1, 2