ANIM: Accurate Neural Implicit Model for Human Reconstruction from a single RGB-D image

Supplementary Material

Marco Pesavento$^{1,3}$  Yuanlu Xu$^3$  Nikolaos Sarafianos$^3$  Robert Maier$^3$  Ziyang Wang$^3$
Chun-Han Yao$^2$  Marco Volino$^1$  Edmond Boyer$^3$  Adrian Hilton$^1$  Tony Tung$^3$

1University of Surrey, CVSSP, UK  2UC Merced  3Meta Reality Labs

7. Overview

In this supplementary material we provide:
1. Additional details about the implementation of ANIM
2. Details about the architectural design of VFE (Voxel Feature Extractor)
3. Further details on ANIM-Real
4. Limitations of the proposed approach
5. Additional results obtained by applying ANIM on real noisy data captured with Azure Kinect
6. Qualitative results for the ablation studies presented in the main paper
7. Additional qualitative results, to further demonstrate that ANIM and the new technical contributions we propose clearly outperform prior works on reconstruction quality.

8. Implementation details

In our proposed network architecture, the normals and the RGB images are concatenated and processed by the two hourglass architectures with four stacks each: the HR-FE outputs an embedding of resolution $256 \times 256 \times 256$ while the resolution of the features obtained from LR-FE is $256 \times 128 \times 128$. The former are bi-linearly interpolated with the ground-truth points projected on the input image to align the point and the feature in the 2D space. The latter are given as input to the VFE along with a voxel created from the input depth map. 3D points from the depth map are obtained by transforming 2D image coordinates to 3D world coordinates using the camera parameters, prior to normalization. The voxel is created from these 3D points. The LR features are aligned with the voxel, which is created with as many voxels as the number of channels of the LR feature (256). The VFE is a novel SparseConvNet U-Net style architecture, based on SparseConvNet [19] that has shown to be efficient for the task of 3D object detection when the input is sparse. Following [44], for any point in 3D space, we tri-linearly interpolate the latent codes from multi-scale code volumes with the ground truth point. The VFE and the HR-FE features are concatenated and finally classified by the MLP with a number of neurons equal to (369, 512, 256, 128, 1). The same features extracted from the VFE and the HR-FE are then interpolated with the point cloud for the depth-supervision. We implement our proposed framework using PyTorch and run training and testing with NVIDIA Tesla V100 GPUs. We train the neural networks with Adam optimizer and a learning rate $lr = 1e^{-4}$ and $\delta = 1.25$. Inference time for one image, without code optimization, is in the order of the second.

For the comparisons in Sec. 5.3 IF-Net, PaMIR, ICON, SuRS, OcPlans, and (6) PIFu and IF-Net variants are re-trained with the same dataset and configuration as ANIM. PIFuhD, ECON, PHORHUM and NormalGAN are not re-trained due to unavailability of training code. We used their checkpoints for evaluation. All methods are tested on the same datasets (RenderPeople [1], THuman2.0 [60]).

Ethical concerns. ANIM was trained on public datasets that do not reveal the identity of subjects. ANIM aims at faithfully capturing full-body humans without alteration and body distortion, avoiding potential misuse or misrepresentation.

9. VFE Architecture

We report in Tab. 4 the detailed architecture of VFE, which consists of a SparseConvNet U-net that we designed for ANIM. The SparseConvNet implements spatially sparse convolutional networks [19]. The VFE architecture is implemented using sub-manifold sparse convolution operations. The table gives the sizes of the different layers and of the receptive fields. We experimented with various variants and report the ones that returned the best results in our experiments.
Table 4. VFE SparseConvNet U-net Architecture.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Layer Description</th>
<th>Output Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 16 \text{ features}, \text{stride } 1) \times 2$</td>
<td>$D \times H \times W \times 16$</td>
</tr>
<tr>
<td>4</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 32 \text{ features}, \text{stride } 2)$</td>
<td>$1/2D \times 1/2H \times 1/2W \times 32$</td>
</tr>
<tr>
<td>5-6</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 32 \text{ features}, \text{stride } 1) \times 2$</td>
<td>$1/2D \times 1/2H \times 1/2W \times 32$</td>
</tr>
<tr>
<td>7</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 64 \text{ features}, \text{stride } 2)$</td>
<td>$1/4D \times 1/4H \times 1/4W \times 64$</td>
</tr>
<tr>
<td>8-10</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 64 \text{ features}, \text{stride } 1) \times 3$</td>
<td>$1/4D \times 1/4H \times 1/4W \times 64$</td>
</tr>
<tr>
<td>11</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 128 \text{ features}, \text{stride } 2)$</td>
<td>$1/8D \times 1/8H \times 1/8W \times 128$</td>
</tr>
<tr>
<td>12-15</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 128 \text{ features}, \text{stride } 1) \times 4$</td>
<td>$1/8D \times 1/8H \times 1/8W \times 128$</td>
</tr>
<tr>
<td>16</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 64 \text{ features}, \text{stride } 1)$</td>
<td>$1/4D \times 1/4H \times 1/4W \times 64$</td>
</tr>
<tr>
<td></td>
<td>concat output 16/10</td>
<td>$1/4D \times 1/4H \times 1/4W \times 64$</td>
</tr>
<tr>
<td>17</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 32 \text{ features}, \text{stride } 1)$</td>
<td>$1/4D \times 1/4H \times 1/4W \times 64$</td>
</tr>
<tr>
<td>18-20</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 32 \text{ features}, \text{stride } 1) \times 3$</td>
<td>$1/4D \times 1/4H \times 1/4W \times 64$</td>
</tr>
<tr>
<td>21</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 32 \text{ features}, \text{stride } 1)$</td>
<td>$1/2D \times 1/2H \times 1/2W \times 32$</td>
</tr>
<tr>
<td></td>
<td>concat output 21/6</td>
<td>$1/2D \times 1/2H \times 1/2W \times 32$</td>
</tr>
<tr>
<td>22</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 32 \text{ features}, \text{stride } 1)$</td>
<td>$1/2D \times 1/2H \times 1/2W \times 32$</td>
</tr>
<tr>
<td>23-24</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 32 \text{ features}, \text{stride } 1) \times 2$</td>
<td>$1/2D \times 1/2H \times 1/2W \times 32$</td>
</tr>
<tr>
<td>25</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 16 \text{ features}, \text{stride } 1)$</td>
<td>$D \times H \times W \times 16$</td>
</tr>
<tr>
<td></td>
<td>concat output 25/3</td>
<td>$D \times H \times W \times 32$</td>
</tr>
<tr>
<td>26</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 16 \text{ features}, \text{stride } 1)$</td>
<td>$D \times H \times W \times 16$</td>
</tr>
<tr>
<td>27-28</td>
<td>$(3 \times 3 \times 3 \text{ conv}, 16 \text{ features}, \text{stride } 1) \times 2$</td>
<td>$D \times H \times W \times 16$</td>
</tr>
</tbody>
</table>

10. ANIM-Real dataset details

As explained in Sec. 4 of the main paper, the performance of neural implicit models significantly deteriorates when tested with raw data from consumer-grade sensors due to the severe input noise. To address this problem, we curated a new dataset (ANIM-Real) consisting of RGB-D noisy data captured with Azure Kinect and high-quality 3D ground-truth meshes reconstructed using a high-resolution camera system that employs active stereo and multi-view cameras [24]. We fine-tune ANIM with this dataset to reconstruct accurate and high-quality 3D human shapes from real-world data, mitigating the impact of the sensor noise. This section provides further details on the system used for data capture and presents examples of data of ANIM-Real. The capture system comprises two subsystems, with 1 Azure Kinect camera and 32 multi-view stereo cameras from [24]. To acquire the data, we calibrate the two systems in order to align the 3D ground-truth meshes with the RGB-D data. The collected dataset consists of 31 subjects, with 16 women and 15 men captured, each subject performing a set of scripted animations (e.g., standing, walking, turning, jogging, stretching, putting on/taking off clothes). Some examples of data are shown in Fig. 10.

Datasets that integrate high-resolution 3D ground-truth shapes with raw RGB-D data are currently unavailable. The introduction of ANIM-Real is a valuable contribution to the research community in the context of neural implicit 3D human reconstruction. This dataset helps to mitigate domain gaps, providing researchers with a resource that facilitates the development of effective techniques in this domain.

11. Limitations

Failure cases can arise from challenging scenes that include arbitrary objects or complex motions (e.g. taking of clothes) as shown in Fig. 8a.
The accuracy of ANIM applied to real-world data is slightly lower than the one achieved with synthetic data since ANIM is still influenced by the noise of the input raw data, which can affect the reconstruction as shown in Fig. 8b where the ankle of the model is not reconstructed. The model could be further fine-tuned to learn specific sensor noise and mitigate domain gaps.

12. Additional results on real-world data

We test ANIM on real-world data obtained with an Azure Kinect after fine-tuning with additional 16k frames, consisting of around 800 frames in average from a single view of 21 subjects. Fig. 10 shows examples of ANIM reconstruction from real single RGB-D images captured with a Kinect Azure. Our approach can retrieve high-quality details on the final mesh even if the input normals and depth are noisy. ANIM can eliminate the noise of the consumer-grade sensor, significantly improving the reconstruction with accurate and high-quality 3D human shapes. We present further qualitative comparisons among different methods using real-world data. Given the inherent challenges associated with the real-world dataset, we show results from one of the most competitive methods, PIFu+VFE+SN, both before and after finetuning. As illustrated in Figure 9, finetuning PIFu+VFE+SN on ANIM-Real yields qualitative improvements, yet not on par with ANIM.

13. Ablation Studies

We illustrate qualitative comparisons for the ablation studies presented in Sec. 5.2 of the main paper. The labels used in the figures are consistent with the ablation study conducted in Tab. 1 and Tab. 2 in the main paper. Fig. 11 illustrates the role that each module of ANIM plays in representing high-quality details in the final reconstruction, with the highest-quality shapes obtained when all the modules are exploited. More specifically, fewer details are represented in the face and hands of the model when spatial-aware sampling is not applied. The importance of normals and HR feature can also be noticed by the reduced amount of details in the final reconstruction. Less accurate shapes are then obtained if LR feature is not used. The introduction of depth supervision further increases the accuracy and the details in the reconstructed shapes. Fig 12 demonstrates the effectiveness of the architecture of ANIM. Each key component was tested one-by-one and it is proved that the complete model outperforms the others with more accurate and highly-detailed 3D shapes.

14. Additional qualitative Results

Additional qualitative comparisons for approaches that reconstruct the 3D shape from an input different than RGB-D are presented in Fig. 13 while Fig. 14 shows additional results obtained by reconstructing 3D shapes from RGB-D data. ANIM consistently generates high-fidelity reconstructions, with cloth wrinkles and high-quality faces and hands in accordance with the input RGB images thanks to our depth-supervision strategy. Depth ambiguity issues are also solved by leveraging the depth channel of the input data. Moreover, it is shown how the contributions we propose, such as using the VFE and the multi-resolution features of HR-FE and LR-FE, can be used to improve other approaches, but only our complete ANIM model design returns the best results. Fig. 15 shows results of reconstructing 3D shapes from input different than RGB-D for other related methods that are not shown in the paper. Fig. 16 show the the side-view reconstruction of the results showed in Fig. 7 and Fig. 14.
Figure 10. More reconstruction results by ANIM using a consumer-grade RGB-D camera (Azure Kinect) as an input. ANIM is capable of handling various human body and cloth typologies ranging from a skirt to a bath robe and is agnostic to diverse human poses.
Figure 11. We conducted an ablation study on the components of ANIM that influence the reconstruction quality. We show reconstructions of 2 subjects (one from THuman2.0 [60] and the other from RenderPeople [1]) captured by a single-view RGBD image (i.e., partial view), from frontal and 45-deg side views. Our full ANIM model provides high-quality reconstructions with facial expressions, hands, and cloth wrinkles with fine-level details, without shape distortion along the camera view. Please zoom in the figure to better see details.
Figure 12. We conducted an ablation study where components of ANIM were removed one-by-one to prove the superiority of the proposed architecture. We show reconstructions of 2 subjects (one from THuman2.0 [60] and the other from RenderPeople [1]) captured by a single-view RGBD image (i.e. partial view), from frontal and 45-deg side views, with colored normals. Our full ANIM model provides more accurate results. Please zoom in the figure to better see details.
Figure 13. Additional comparisons with approaches that use single RGB image or partial point clouds as input. Data from RenderPeople [1]. ANIM reconstructs full-body models with high accuracy, with cloth wrinkles, face and hand details, and without depth ambiguity (i.e. distortion along camera view).

Figure 14. Additional comparisons with methods that use a single RGB-D image as input. Our core contributions can leverage state-of-the-art models, but only our complete ANIM model design returns the best results. We show reconstruction from the front view. Data from THuman2.0 [60].
Figure 15. Qualitative comparisons with approaches not illustrated in the main paper that use a single RGB image or partial point clouds as input. Data from RenderPeople [1].
<table>
<thead>
<tr>
<th>Input</th>
<th>Normal</th>
<th>OcPlane</th>
<th>PIFu+D</th>
<th>PIFu+D+SN</th>
<th>PIFu+VFE+HR+SN</th>
<th>IFNet</th>
<th>PIFu+VFE+SN</th>
<th>IFNet+HR+SN</th>
<th>ANIM</th>
</tr>
</thead>
</table>

Figure 16. Side-views of the 3D shapes reconstructed from an input RGB-D data showed in Fig. 7 and Fig. 14.
References


[25] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In IEEE Conference on Computer Vision and Pattern Recognition, 2020. 2


[38] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *IEEE International Conference on Computer Vision*, 2021. 2


[40] Aihua Mao, Hong Zhang, Yuxin Liu, Yinglong Zheng, Guqing Li, and Guoqiang Han. Easy and fast construction of a 3d avatar with an rgbd-d sensor. *Sensors*, 17(5), 2017. 2


[57] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. Icon: Implicit clothed humans obtained from normals. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 6, 7


