Supplementary Material for InstantGeoAvatar: Effective Geometry and Appearance Modeling of Animatable Avatars from Monocular Video

Alvaro Budria[1], Adrian Lopez-Rodriguez[2],
Òscar Lorente*[3], and Francesc Moreno-Noguer*[4]

[1] Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
alvaro.francesc.budria@upc.edu
[2] Vody
[3] Floorfy
[4] Amazon

In this **supplementary document**, we provide additional details and results to complement our main submission. In the **supplementary video** we report showcase of our method on in-distribution and out-of-distribution poses. We recommend watching the whole video as it contains several video subsequences and results.

1 Implementation Details

1.1 Network Architecture

In Fig. 1 we show the modeling pipeline of InstantGeoAvatar. The SDF field is implemented as a 16-level hash grid with minimum resolution of 32 and maximum resolution of 512, feature size of 2 per level and codebook size of 2^{19}, followed by an MLP with a single hidden layer of size 64 and an output size of 12 for the feature vector and 1 for the signed distance. The texture field module consists of another MLP with two hidden layers of size 64. All network layers are ReLU-activated, except for the last layer of the SDF field which has no activation, and the last layer of the texture field which is activated with a sigmoid function.

1.2 Training Details

We train our network using the Adam optimizer for 20 epochs with initial learning rate of $1e-2$ and exponential decay, and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. A training cycle takes about 10 minutes on a single A10 NVIDIA GPU.

*Work done while at Institut de Robòtica i Informàtica Industrial (CSIC-UPC).

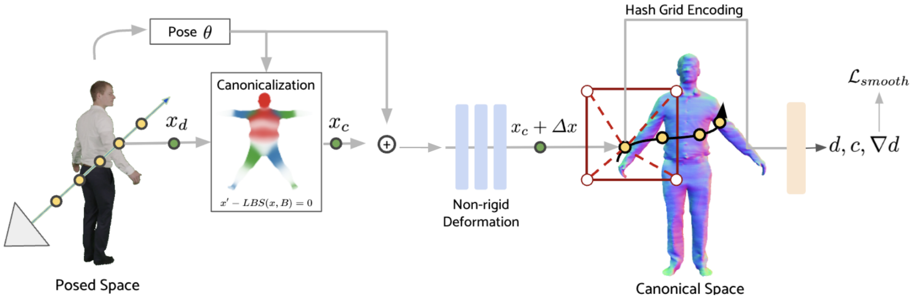

Fig. 1: InstantGeoAvatar pipeline. We densely sample points along camera rays and canonicalize them with an efficient deformation module, after which we non-rigidly deform the points. These points are then used to query into a canonical representation for shape and appearance. During training, we use the surface normals to apply a smoothing regularization term onto the learned signed distance function.

1.3 Baselines

For both Vid2Avatar and InstantAvatar, we follow the authors' original architecture and training schedule, as per their public implementations.

The hybrid positional encoding is reimplemented as described in the original paper [23] by passing Fourier features through a small MLP with output size 3 and concatenating its output to the hash grid features.

The implementation for the architectural and training tweaks from Neuralangelo [56] (the Eikonal loss with finite differences, and the curvature loss) are taken from the github repository `instant-nsr-pl`.

2 More Results

2.1 Quantitative Results on Eikonal Weighting

We report numerical results on the comparison between different weightings of the Eikonal loss and our method (Figure 5 in the main paper).

Table 1: Effect of modifying the weighting for the Eikonal loss weighting.

	PSNR ↑	SSIM ↑	CD ↓	NC ↑
$\lambda_{Eik} = 0.1$	29.24	0.962	0.87	0.82
$\lambda_{Eik} = 0.5$	29.14	**0.966**	0.83	0.84
$\lambda_{Eik} = 1.0$	28.98	0.956	0.83	0.83
$\lambda_{Eik} = 1.5$	29.10	0.959	0.88	0.82
w/ our $\mathcal{L}_{smooth}$	**29.26**	0.964	**0.77**	**0.85**

https://github.com/bennyguo/instant-nsr-pl/tree/main

Although the Eikonal loss weighting does have a significant impact on the resulting geometry, our model obtains better results than the best performing weighting of the Eikonal loss.

2.2 Qualitative Results on Ablation

We show in Figures 2 and 3 a comparison of our method and different SDF training schemes and tricks from prior works, relating to Table 6 and Figure 5 of the main document. We can see that our method achieves better human reconstruction results in terms of the global shape and the surface detail recovery.

3 Limitations and Societal Impact

InstantGeoAvatar inherits the weaknesses of methods trained solely on 2D images. In particular, we require all regions of the human subject to be visible for obtaining an accurate result. As all other methods learning a canonical representation, InstantGeoAvatar struggles with garments whose topology significantly departs from that of the body, such as skirts. Moreover, our method is depends on accurate body pose and camera parameters to properly deform points to canonical space. An interesting line of future work is to leverage data-driven, purely image-based methods to mitigate the above mentioned limitations.

Virtual avatars can revolutionize fields like gaming, virtual reality, and online communication. However, creating personalized avatars raises privacy concerns. These avatars can also be misused for identity theft or creating fake identities. Such issues must be addressed by companies before deploying these techniques in market products.

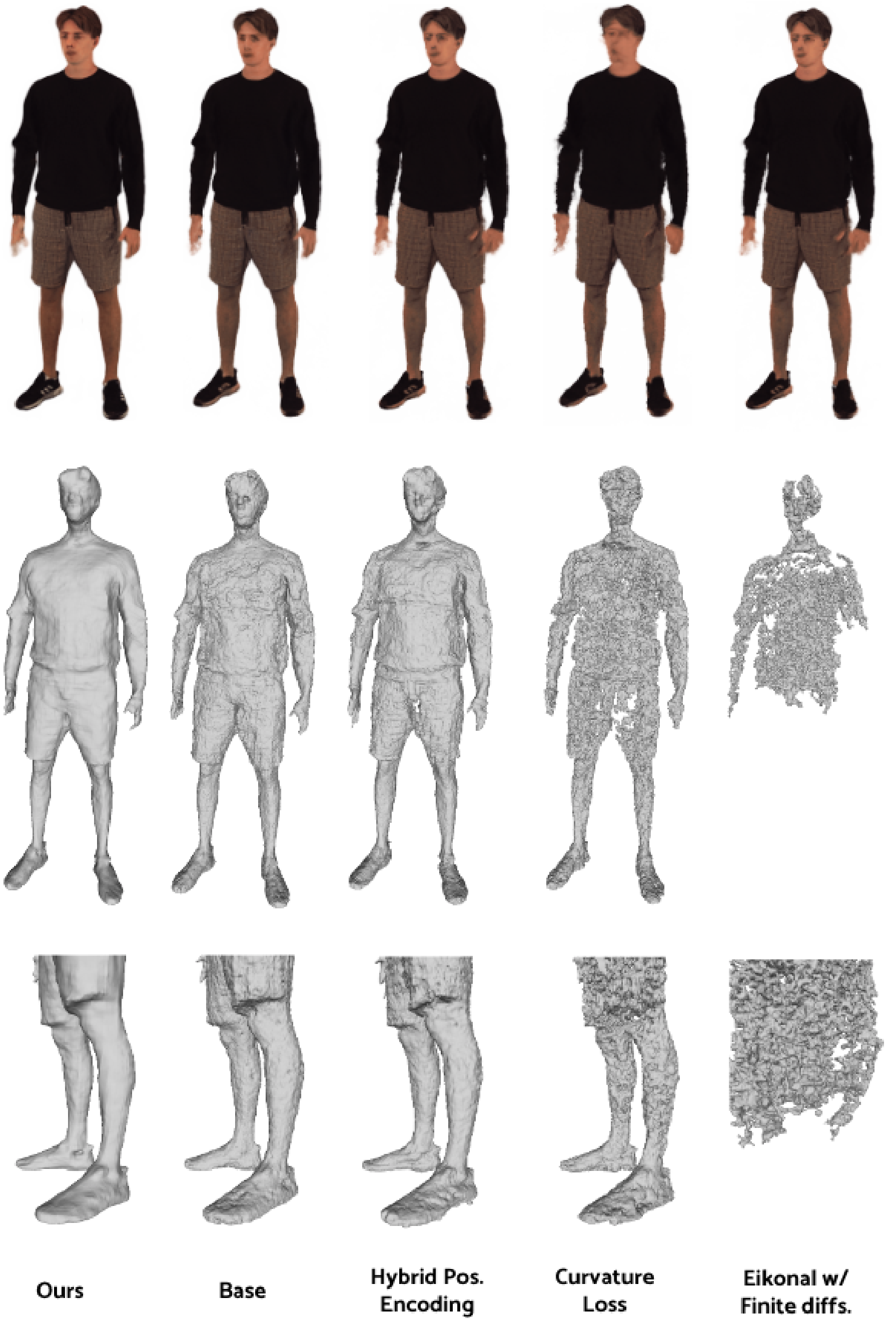

Fig. 2: Qualitative comparison of SDF regularization schemes. We show the full body and a close up of the legs.

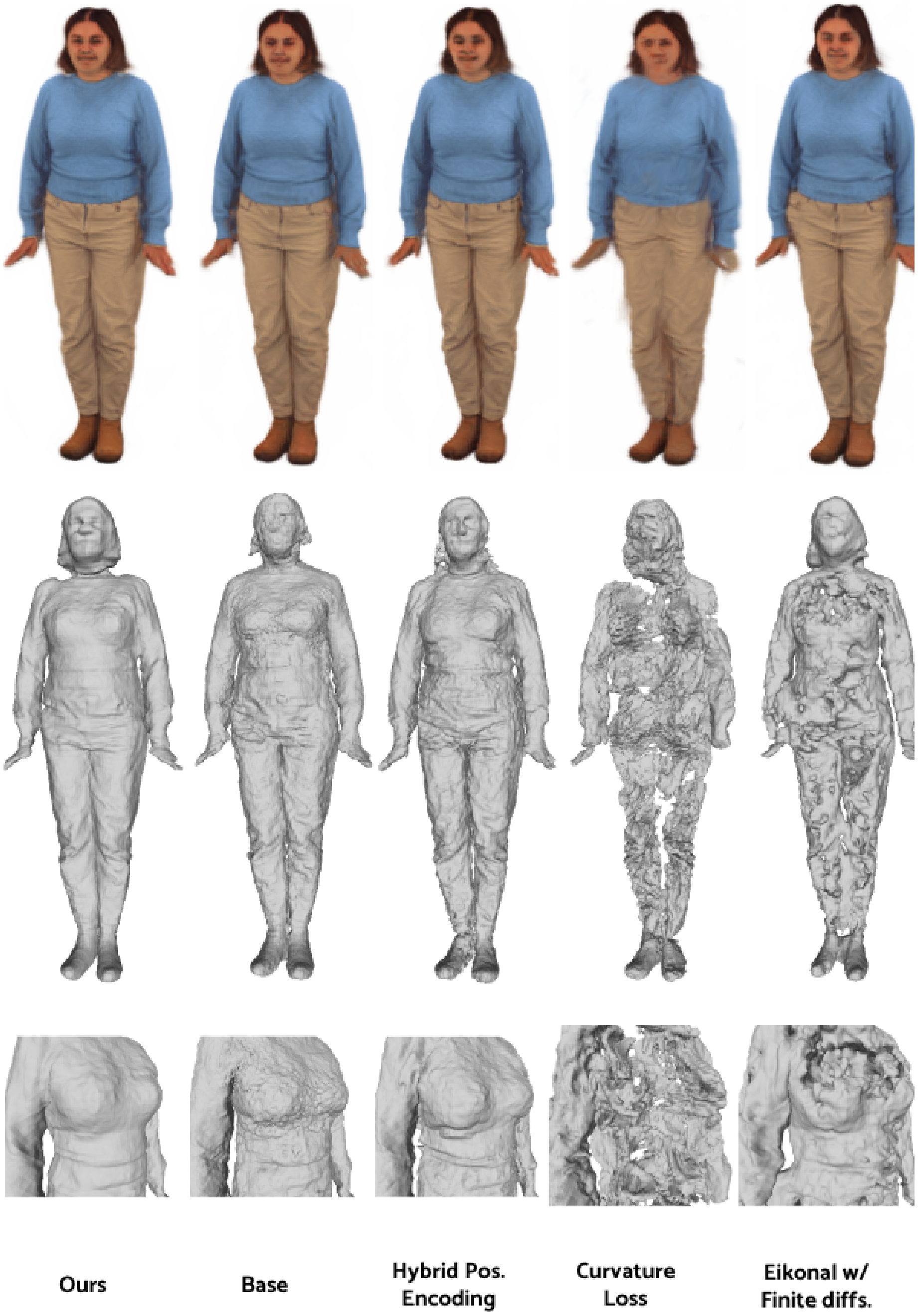

Fig. 3: Qualitative comparison of SDF regularization schemes. We show the full body and a close up of the chest region.