One-shot Implicit Animatable Avatars with Model-based Priors
Supplemental Material

Yangyi Huang[1,4*] Hongwei Yi[2*] Weiyang Liu[2,3] Haofan Wang[4]
Boxi Wu[5] Wenxiao Wang[5] Binbin Lin[5,6†] Debing Zhang[4] Deng Cai[1]
[1] State Key Lab of CAD & CG, Zhejiang University
[2] Max Planck Institute for Intelligent Systems, Tübingen [3] University of Cambridge
[4] Xiaohongshu Inc. [5] School of Software Technology, Zhejiang University [6] Fullong Inc.
huangyangyi@zju.edu.cn hongwei.yi@tuebingen.mpg.de

1. Implementation Details

In this section, we provide important implementation details for our experiments. We also publicly release our experiment code, results, and model checkpoints at https://huangyangyi.github.io/ELICIT for research purposes.

1.1. Optimization

In this section, we provide details about the two-stage optimization process of ELICIT. For loss weights settings in Eq. (7), we set $\lambda_{\text{CLIP}} = 0.1$, $\lambda_{\text{sil}} = 0.01$ are the loss weights. We do not use text prompts in our experiments unless specified, for a fair comparison with baseline methods. The initialization stage takes $T_{\text{init}} = 15,000$ iterations of optimization, while the one-shot training stage takes $T_{\text{train}} = 20,000$ iterations. The entire training process for each subject takes approximately 5 hours on 4 NVIDIA Tesla V100 GPUs. We follow the hyper-parameter settings of the HumanNeRF[15] code for the optimizer, learning rate, and ray sampling configurations. Specifically, we only train $T_{\text{train}} = 5,000$ for quantitative comparison on novel view synthesis in Tab. 2.

1.2. Details of hybrid sampling strategy

In this section, we provide a detailed description of our hybrid sampling strategy, which combines body-part-aware sampling and rotation-aware sampling in one-shot training.

For each iteration, we randomly decide whether to sample a novel view from $\{(\theta_i, \mathbf{e}_j)\}_{i=1,j=1}^{L,M}$ or the input view $V_s = (\theta_s, \mathbf{e}_s)$ with a probability of $p_{\text{novel}} = 0.5$. If $V_{\text{train}} = V_s$, we follow HumanNeRF to sample a pair of patches for reconstruction. Otherwise, we randomly select a body part k (including the whole body) with weighted probability$\{p_{\text{part}}^k\}_{k=1}^K$, and sample a training patch V_{train}^k which is decided by the bounding box of SMPL rendered body-part segmentation $S_{\text{SMPL}}^k(V_{\text{train}})$.

After sampling the training patch, we sample the reference patch from the V_s or other views of the same pose $\{(\theta_{\text{train}}, \mathbf{e}_j)\}_{j=1,j\neq i}^M$. The camera views of the current pose are divided into front views, rear views, left views, and right views according to the body rotation angle. We assume that the input image is close to the front view of the character. If a rear view of specific body parts (e.g. head, upper body, or whole body) is sampled as the training view, we randomly sampled nearest views from left views and right views as V_{ref}. Then we render body-part patch V_{ref}^k by our NeRF model as reference. Otherwise, the reference patch will be constructed by the resized patch V_s^k cropped from the input image. We set the size of patches in training to 224×224 for all experiments, the same as the input resolution of the CLIP ViT/L-14 model we use for semantic prior.

1.3. Detailed configuration of evaluation

In this section, we provide the detailed configuration of our quantitative comparison on ZJU-MoCap dataset and Human 3.6M dataset.

1.3.1 Data splitting

For per-subject optimization methods Animatable NeRF[11] (Ani-NeRF) and NeuralBody[13] (NB), we use all subjects of ZJU-MoCap data-set (313, 315, 377, 386, 387, 390, 392, 393, 394) and the "Posing" sequences of Human 3.6M dataset (S1, S5, S6, S7, S8, S9, S11). We provide information on the single input frame of each subject to evaluate novel pose synthesis, and the 10 frames of each subject we sampled to evaluate novel view synthesis in our experiment code.

For Neural Human Performer[7] (NHP), since it requires pre-training on subjects from the same dataset, we only evaluated NHP with 3 testing subjects from each dataset:

ZJU Mocap (313, 315, 387), Human 3.6M (S8, S9, S11), and use remaining subjects for pre-training.

1.3.2 Baseline settings

Neural Human Performer[7]. We modify NHP to take only one input view from the first camera of ZJUMoCAP or the third camera of H36M and train the model with novel view ground truth from all other available cameras. We keep other hyperparameters the same as original paper and trained each model with 1000 epochs.

NeuralBody[11]. We train NB models for each input frame by optimizing the model only on the single input image. We set the number of optimization iterations to 50K, which is enough for NB to converge on the input image (total loss $<$ 0.0001). We keep other hyper-parameter the same as original paper.

Animatable NeRF[11]. We choose Ani-NeRF with pose-dependent fields (PDF), which presents the best results in the original paper. We also train Ani-NeRF models until convergence, similar to the setting of NB.

2. Additional Results

2.1. Comparison with MonoNHR

To compare our method with MonoNHR[2], which reports state-of-the-art results on human-specific novel view synthesis from a single monocular input, we present qualitative results of MonoNHR and ELICIT on the ZJU-MoCAP dataset. As full results from MonoNHR are not available, we use the novel view synthesis results from its official qualitative video and compare them with the same input view on ELICIT.

As shown in Figure R.1, while MonoNHR can estimate approximate clothed body geometry, it produces blurry contents on the novel views, whereas ELICIT generates more realistic details on human faces, bodies, and clothing.

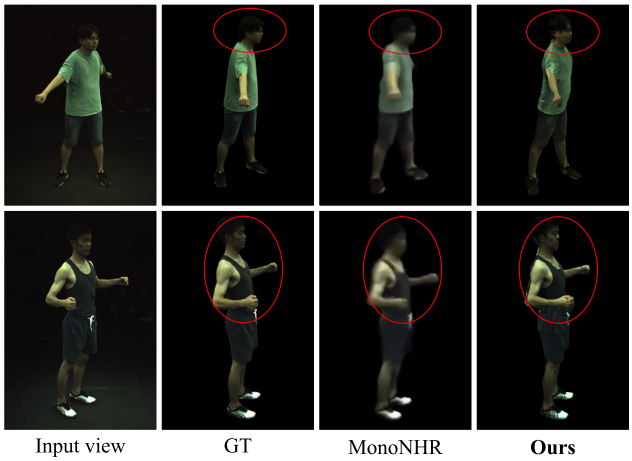

Figure R.1: **Qualitative comparison with MonoNHR[2].** While MonoNHR produces blurry faces, ELICIT generates realistic facial details, demonstrating the superior performance of our method.

2.2. Ablation Study

2.2.1 Differerent pretrained visual models

As discussed in Section 4.4, we also compare the performance of different pre-trained visual models, including an DINO [1] ViT used by SinNeRF [16], an ImageNet pre-trained ViT/L-14 [4, 3], an unsupervised pre-trained ViT/L-14 by MAE [5], also a lighter version of CLIP ViT/B-32. As shown in Figure R.2, CLIP ViT/L-14 shows best performance in capturing 3D-aware human body structure and generating vivid visual details, and the two CLIP pre-trained models have a better performance on head structure than Image pre-trained models. This comparison suggests that the rich pre-training data of the CLIP model, as well as the larger model capacity of CLIP ViT/L-14 compared to CLIP ViT/B-32, are key factors contributing to the effectiveness of our semantic loss.

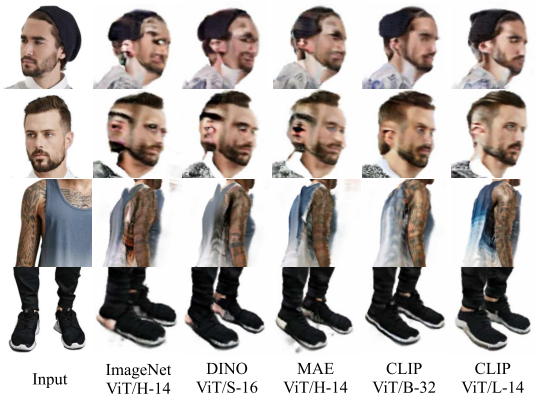

Figure R.2: Qualitative results for the ablation studies of vision models used for the semantic loss, selected from DeepFashion[8] dataset. The CLIP ViT/L-14 model we use produce best detailed geometry and textures.

2.2.2 Hybrid sampling strategy

To thoroughly evaluate the effectiveness of our proposed hybrid sampling strategy, we conducted a detailed ablation study on both body-part-aware sampling and rotation-aware sampling. As shown in Figure R.3, our results indicate that body-part-aware sampling improves ELICIT's ability to synthesize realistic details on crucial body parts with fine-grained supervision. Additionally, rotation-aware sampling successfully avoids artifacts of mirrored appearance by using neighboring views as a reference to recover heavily occluded body regions.

2.2.3 Comparing CLIP loss with perceptual losses

In our main paper, we compared our CLIP-based semantic loss with various embedding losses that capture high-level

Figure R.3: **Ablation study of hybrid sampling strategy.** Comparison of training with different sampling strategies: without (a) body-part-aware sampling, without (b) rotation-aware sampling, and full hybrid sampling strategy. The absence of either sampling strategy leads to artifacts, such as mirrored appearance or missing details on important body parts.

semantics. However, since the CLIP loss can also capture low-level visual attributes such as color and texture, we further evaluated its effectiveness by comparing it with two commonly-used perceptual losses: LPIPS[17] and VGG-based perceptual loss[6], in generative and reconstruction tasks. As depicted in Figure R.4, LPIPS loss and VGG loss only capture a subset of low-level visual features and cannot synthesize 3D-aware appearance with high-fidelity details in occluded areas, unlike CLIP-loss.

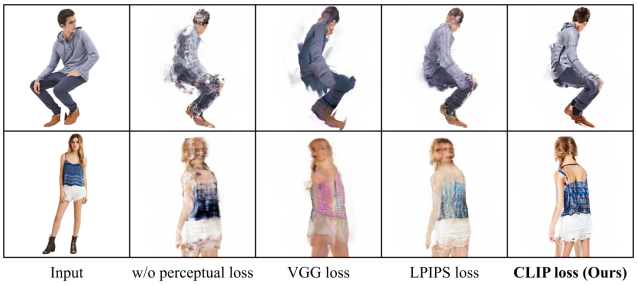

Figure R.4: **Comparison of CLIP loss to other perceptual losses.** LPIPS and VGG-based perceptual losses only capture a subset of low-level visual features, leading to limited performance in synthesizing occluded clothed body appearance compared to CLIP loss.

2.3. Extensions

ELICIT proposes a simple and effective pipeline for creating animatable avatars with implicit representation and model-based prior. The pipeline is also extensible for future improvements with different implicit human representations, semantic priors, geometric priors, and input settings. In this section, we introduce several extensions of ELICIT that can inspire future work.

2.3.1 Alternative human representations

ELICIT can be trained using various implicit human representations. For example, as shown in Figure R.5, we replaced the HumanNeRF model used in ELICIT with an SDF-based model from Animatable NeRF[12, 11]. This alternative representation performed better in surface geometry, while HumanNeRF produced blurry floating artifacts near the body that decreased the rendering quality. Such explorations with different implicit human representations can lead to further improvements in the quality of the synthesized avatars.

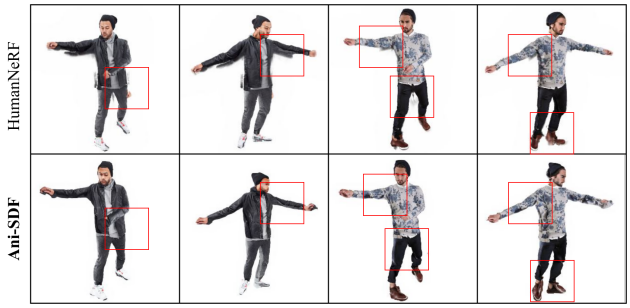

Figure R.5: **Improved human representation.** The SDF-based model from Ani-NeRF[12] reduces floating artifacts (marked with red rectangles), which are commonly present in our HumanNeRF-based model, leading to better surface geometry.

2.3.2 Editing 3D avatars with textual guidance

As we discussed in our main paper, we can improve the performance of semantic prior by incorporating user text prompts through text-based CLIP guidance and image-based CLIP guidance. In addition, as shown in Figure R.6, ELICIT can generate different text-conditioned appearances using different text prompts, such as manipulating the occluded texture of clothing. These results demonstrate the potential for using ELICIT's pipeline for digital human editing tasks with further improvements.

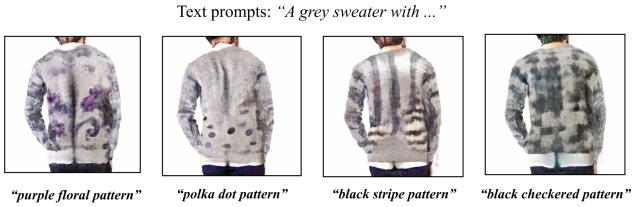

Figure R.6: **Generating text-conditioned appearance.** By using different prompts, we can generate various texture patterns in the occluded area of clothing. While the quality of synthesis is limited, it demonstrates the potential of ELICIT for editing 3D avatars.

2.3.3 Utilizing multiple images

ELICIT can be enhanced by utilizing multiple input images to better recover full-body appearances. It's worth noting that ELICIT can utilize images of different poses without requiring well-aligned pose annotations, by taking one image for reconstruction and using the others as a reference in the CLIP loss. As shown in Figure R.7, we demonstrate the effectiveness of this approach by incorporating an extra back-side image, resulting in better full-body appearance.

Figure R.7: **Utilizing multiple images.** ELICIT can utilize images of different poses as an extra reference to better recover full-body appearance.

3. Limitations

The human body geometry prior utilized by ELICIT requires well-aligned SMPL annotation of body shape and postures. When body parts such as hands and legs are heavily misaligned, artifacts may occur due to the model being initialized incorrectly or failing to sample reference patches for body-part refinement. Furthermore, modeling hand geometry and complex clothing geometry precisely remains a challenge for our method.

Additionally, the computational cost of 5 hours on 4 Tesla V100s per avatar may be prohibitively expensive for certain applications. Future work could focus on developing more efficient human-specific NeRFs that require lower GPU memory, as well as improving the training pipeline to reduce the number of necessary training iterations.

4. Future Work

We plan to further explore model-based priors that can potentially improve ELICIT. For semantic prior, we will investigate the use of image diffusion models [14, 9] which have been applied to text-to-3D tasks, as they are promising options for enhancing the appearance details of ELICIT. For geometric prior, we aim to use a more expressive human-body prior with SMPL-X [10] to improve detailed geometry, such as hand shapes. Regarding implicit representation, we are exploring options and improvements with higher efficiency, better surface geometry, and better rendering quality. Additionally, we are working to enhance the versatility of our one-shot training framework to accept different types of inputs (e.g., multiple images, short videos, and images with a text description).

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 2

[2] Hongsuk Choi, Gyeongsik Moon, Matthieu Armando, Vincent Leroy, Kyoung Mu Lee, and Grégory Rogez. Mononhr: Monocular neural human renderer. *International Conference on 3D Vision (3DV)*, pages 242–251, 2022. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2

[6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 3

[7] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Conference on Neural Information Processing Systems (NeurIPS)*, 34:24741–24752, 2021. 1, 2

[8] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[9] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 4

[10] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 4

[11] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. In *International Conference on Computer Vision (ICCV)*, pages 14314–14323, 2021. 1, 2, 3

[12] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. *arXiv preprint arXiv:2203.08133*, 2022. 3

[13] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit Neural Representations With Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9054–9063, 2021. 1

[14] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 4

[15] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 1

[16] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. *arXiv preprint arXiv:2204.00928*, 2022. 2

[17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 3