MIGS: Multi-Identity Gaussian Splatting via Tensor Decomposition - Supplementary -

Aggelina Chatziagapi[1], Grigorios G. Chrysos[2], and Dimitris Samaras[1]

[1] Stony Brook University {aggelina,samaras}@cs.stonybrook.edu
[2] University of Wisconsin-Madison chrysos@wisc.edu

Contents

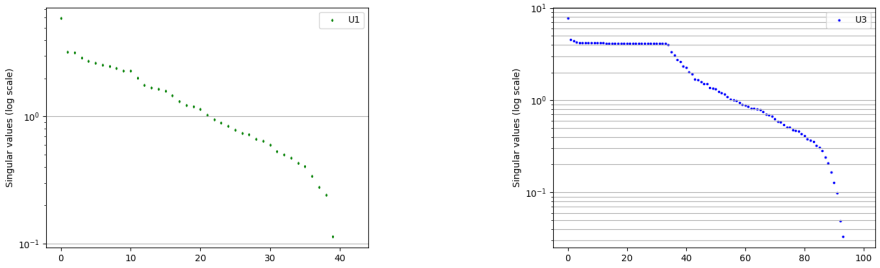
The supplementary document is organized as follows:

- Additional Ablation Study in Sec. A.
- Additional Results in Sec. B.
- Implementation Details in Sec. C.
- Limitations in Sec. D.
- Ethical Considerations in Sec. E.

We strongly encourage the readers to watch our supplementary video on our project page: https://aggelinacha.github.io/MIGS/.

A Additional Ablation Study

Learned Matrices. In Fig. 1, we plot the singular values of our learned matrices $\boldsymbol{U}_1 \in \mathbb{R}^{M \times R}$ and $\boldsymbol{U}_3 \in \mathbb{R}^{N_g \times R}$, where $M = 43$ is the number of parameters per Gaussian, $N_g = 5 \times 10^4$ is the number of the 3D Gaussians, and $R = 100$



(a) Singular values of $\boldsymbol{U}_1$ **(b)** Singular values of $\boldsymbol{U}_3$

Fig. 1: Singular values. The singular values of our learned $\boldsymbol{U}_1 \in \mathbb{R}^{M \times R}$ and $\boldsymbol{U}_3 \in \mathbb{R}^{N_g \times R}$, where $M = 43$ is the number of parameters per Gaussian, $N_g = 5 \times 10^4$ is the number of 3D Gaussians, and $R = 100$ is the rank of our CP tensor decomposition.

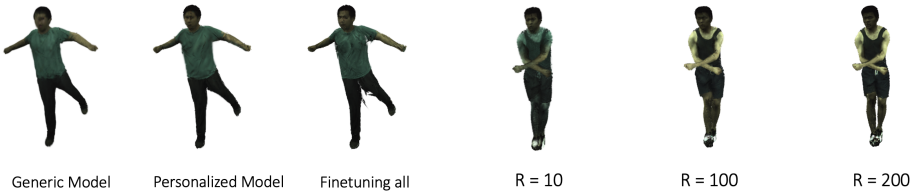
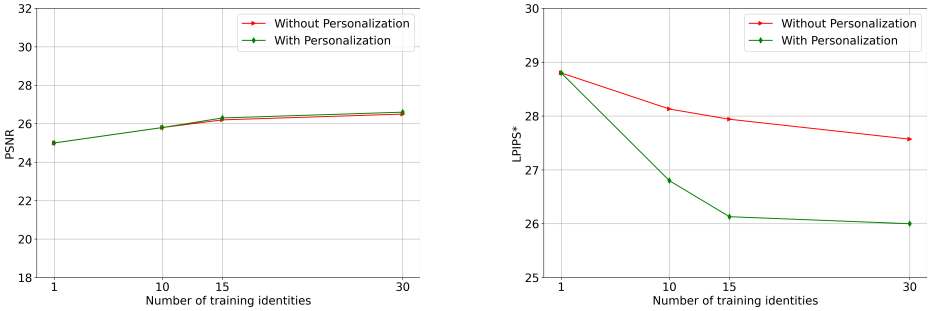


Fig. 2: Ablation study. (a) Ablation study on our proposed personalization procedure. In order to capture individual details, we fine-tune our color MLP of our generic model (left), leading to the personalized model for a particular subject (middle). We do not fine-tune all parameters (see artifacts on the right). **(b)** Ablation study on the rank R of our tensor decomposition. $R = 10$ leads to a mixture of identities (notice different shirt colors), while $R = 100$ is sufficient to capture the training identities.



(a) PSNR vs number of training identities

(b) LPIPS* vs number of training identities

Fig. 3: Ablation study. Visual quality (PSNR and LPIPS*) on the "Advanced Test" set, with or without personalization, for different number of training identities and $R = 100$.

is the rank of our CP tensor decomposition (see Sec. 4.2 of the main paper). We use SVD from PyTorch [10] [3] and plot them in logarithmic scale. These matrices correspond to the model trained on the 30 subjects from the AIST++ dataset [5, 13].

Personalization. As mentioned in Sec. 4.4 of the main paper, we can personalize our generic model for a particular subject, in order to capture individual details (*e.g.* face or cloth details). We show an example output of our generic model trained on 30 identities and the corresponding personalized result in Fig. 2a. We notice that this personalization procedure is needed only for models trained with identities more than 20. We only fine-tune the color MLP for 5×10^3 iterations, keeping the rest of the parameters frozen, using a short video of the target subject. We do not fine-tune all parameters, since in this case our network would forget the large variety of human body deformations learned from multiple subjects, leading to artifacts in novel poses (see Fig. 2a right).

[3] https://pytorch.org/docs/stable/generated/torch.svd.html

Fig. 4: Animation of human avatars under novel poses. Qualitative comparison with state-of-the-art approaches, namely HumanNeRF [14], MonoHuman [15], GauHuman [1], and 3DGS-Avatar [12]. The training subjects and the target poses are from the ZJU-MoCap dataset [11]. Our method demonstrates significant robustness.

Ablation Study on the Rank R. In Fig. 2b, we show a qualitative comparison for different values R of our tensor decomposition (see also Sec. 5.3). We observe that $R = 10$ is not enough to capture a larger number of identities. For example, it can lead to a mixture of colors (notice the shirt colors for $R = 10$). On the other hand, $R = 100$ seems sufficient to capture all the training identities.

Ablation Study on the Number of Identities. Figure 3 shows the visual quality for different number of training identities, with and without personalization. As also mentioned in the main paper (see also Fig. 6), increasing the number of identities leads to an increase in robustness under novel poses. Further personalization captures identity-specific details, enhancing the output visual quality.

B Additional Results

Figure 4 demonstrates additional qualitative results for our model trained on subjects from the ZJU-MoCap dataset [11]. Animating them under novel poses leads to artifacts under the arms and legs in all other methods, namely Hu-

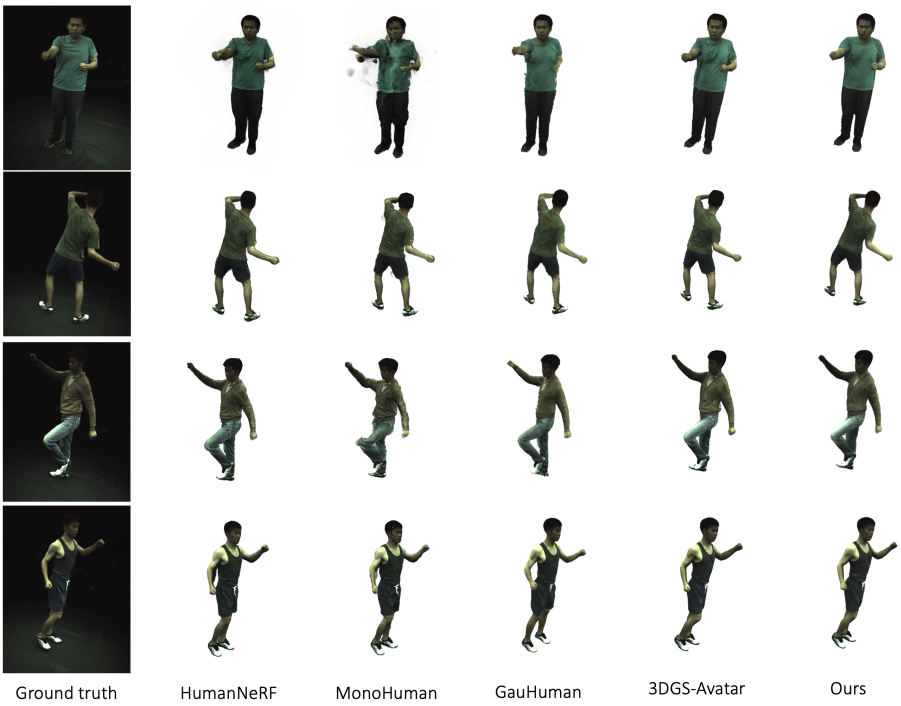


Fig. 5: Novel view synthesis on ZJU-MoCap. Qualitative comparison with state-of-the-art approaches, namely HumanNeRF [14], MonoHuman [15], GauHuman [1], and 3DGS-Avatar [12] on novel view synthesis on the test set of the ZJU-MoCap dataset [11].

manNeRF [14], MonoHuman [15], GauHuman [1], and 3DGS-Avatar [12]. Our proposed method demonstrates significant robustness.

Figure 8 demonstrates additional qualitative results when the target poses are from the AIST++ dataset [5, 13]. In this case, the poses are more challenging, completely unseen during training (out of the training distribution). Again, MIGS outperforms the other methods, robustly animating the identities under novel poses.

Figure 5 shows qualitative comparisons for novel view synthesis on the test set of the ZJU-MoCap dataset [11]. Corresponding quantitative results are shown in Table 1 of the main paper and Tab. 1. Our method demonstrates comparable performance with 3DGS-Avatar on novel view synthesis, while trained on multiple identities simultaneously.

Figure 6 shows qualitative comparisons with approaches that use multiple views as input. These works address the problem in a different way. They extract features from nearby views and infer a novel view in a feed-forward manner. Most of them cannot animate humans under novel poses, *e.g.* see results of NHP [6] and TransHuman [9] (we use black background to be consistent with

Table 1: Quantitative evaluation on ZJU-MoCap. We compare our method with state-of-the-art approaches (HumanNeRF [14], MonoHuman [15], GauHuman [1], 3DGS-Avatar [12]) on novel view synthesis, using the standard test set of ZJU-MoCap. We report PSNR, SSIM, and LPIPS* = LPIPS $\times 10^3$ on 2 subjects (387, 393). See Table 1 in the main paper for the other 4 subjects (377, 386, 392, 394).

	387			393		
Method	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓
HumanNeRF	28.18	0.9632	35.58	28.31	0.9603	36.72
MonoHuman	27.93	0.9601	41.76	27.64	0.9566	43.17
GauHuman	27.95	0.9608	40.70	27.88	0.9578	43.01
3DGS-Avatar	28.33	0.9642	**34.24**	28.88	0.9635	35.26
Ours	**30.70**	**0.9643**	35.33	**31.57**	**0.9640**	**30.44**

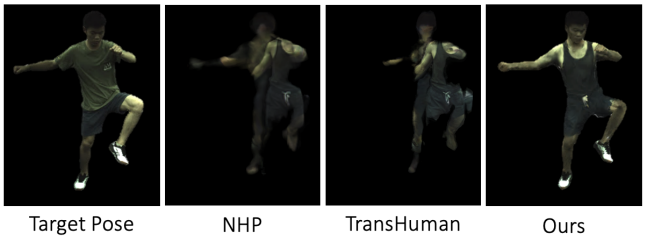


Fig. 6: Comparison with multi-view approaches: Neural Human Performer [4] and TransHuman [9].

their pretrained models - compare with row 2 in Fig. 4). ActorsNeRF [8] does not have code available. GPS-Gaussian [16] does not use body pose (SMPL parameters) as input at all, and thus cannot render novel poses.

C Implementation Details

In this section, we include implementation details of our proposed method (see also Tab. 2). Our implementation is based on PyTorch [10]. We built our architecture upon 3DGS-Avatar [12], but with some important modifications in order to learn multiple identities (see also Sec. 3.2 of the main paper). We use $N_g = 5 \times 10^4$ 3D Gaussians, that we initialize by randomly sampling N_g points on the canonical SMPL mesh surface of the first identity. We initialize the matrices $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$ as described in Sec. 4.3 of the main paper.

The positions $\boldsymbol{\mu}_c$ of the Gaussians in the canonical space are first encoded into a multi-level hash grid and passed through the non-rigid deformation MLP f_d. The hash grid has 16 levels of 2-dimensional features each, hash table size 2^{16}, coarse and fine resolution 16 and 2048 correspondingly [12]. The non-rigid MLP f_d is also conditioned on a latent code $\boldsymbol{z}_p$ that is the output of an hierarchical encoder [7, 12]. The rigid MLP f_r inputs the non-rigidly deformed positions $\boldsymbol{\mu}_d$ and outputs the skinning weights that sum up to 1 through a softmax layer [12]. Similarly with 3DGS-Avatar [12], we normalize the coordinates in the canonical space by proportionally padding the bounding box enclosing the canonical SMPL

mesh of the identity. The color MLP is conditioned on the output $\boldsymbol{z}$ of the non-rigid network, a per-Gaussian feature vector $\boldsymbol{f}$ and the canonicalized viewing direction (see Sec. 3.2). We do not learn any per-frame latent codes, in order to avoid any overfitting to the training frames. We use a larger color MLP compared to 3DGS-Avatar, of 3 layers and 256 hidden units each, in order to learn the diverse colors of multiple identities.

We observed that our network is very sensitive to the learning rate of the different components and the initialization, similarly with other 3DGS methods [2]. Our initialization is described in Sec. 4.3. We experimentally chose the learning rates depicted in Tab. 2. As shown, we use different learning rates for different rows of $\boldsymbol{U}_1$ that roughly correspond to scaling, rotation, features for color, and opacity (see our tensor construction in Sec. 4.1). These learning rates are similar with the learning rates used for scaling, rotation, color, and opacity by other 3DGS implementations [2, 12]. However, in our case, we get the corresponding values after multiplying the matrices $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$ using Eq. (8). We use Adam optimizer [3]. The rest of the Adam hyper-parameters are set at their default values ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$).

We train our multi-identity network for 5×10^4 iterations for 15 subjects, and 10^5 iterations for 30 subjects, that need about 2 and 4 hours correspondingly on a single NVIDIA Quadro RTX 6000 GPU. Following prior work [12, 14], we freeze everything in the first 10^3 iterations and train only the rigid MLP. In this way, we better initialize the rigid mapping from the canonical to the observation space based on estimated SMPL parameters, and avoid any noisy gradients in the beginning. After the first 10^3 iterations, we enable optimization to all the components, except for the non-rigid network that is frozen until the first 3×10^3 iterations. In constrast to 3DGS-Avatar [12], we do not add any learnable module for pose correction for the estimated SMPL parameters, in order to avoid overfitting to the training body poses.

We use the same loss function as 3DGS-Avatar [12]:

$$\begin{aligned}\mathcal{L} = \lambda_{l1}\mathcal{L}_{l1} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{skin}\mathcal{L}_{skin} + \\ \lambda_{isopos}\mathcal{L}_{isopos} + \lambda_{isocov}\mathcal{L}_{isocov} \ ,\end{aligned} \tag{1}$$

where $\mathcal{L}_{l1}$ is the L1 photometric loss and $\mathcal{L}_{perc}$ is the perceptual (LPIPS) loss with VGG as backbone. The mask loss $\mathcal{L}_{mask}$ corresponds to the L1 loss between the ground truth foreground mask and the predicted mask by accumulating the predicted opacities [12]. The skinning loss provides a regularization on the non-rigid MLP [12]. The as-isometric-as possible losses $\mathcal{L}_{isopos}$ and $\mathcal{L}_{isocov}$ restrict the 3D Gaussian positions to preserve a similar distance after deformation to the observed space, with also similar covariance matrices [12]. We set $\lambda_{l1} = 1$, $\lambda_{perc} = 0.01$, $\lambda_{mask} = 0.1$, $\lambda_{isopos} = 1$, $\lambda_{isocov} = 100$, and $\lambda_{skin} = 10$ for the first 10^3 iterations that is then decreased to $\lambda_{skin} = 0.1$ [12].

In order to avoid any non-differentiable gradient updates, we do not apply any densification or pruning of the 3D Gaussians during training, in contrast to the original 3DGS implementation [2]. Instead, we keep N_g Gaussians throughout our optimization, which are moved and deformed according to our network. In

Number of 3D Gaussians N_g	5×10^4
Rank of CP tensor decomposition R	100
Dimension of feature vector $\boldsymbol{f}$	32
Dimension of non-rigid output vector $\boldsymbol{z}$	32
Non-rigid deformation MLP f_d: Linear layers	3
Non-rigid deformation MLP f_d: Hidden units	128
Rigid transformation MLP f_r: Linear layers	3
Rigid transformation MLP f_r: Hidden units	128
Color MLP f_c: Linear layers	3
Color MLP f_c: Hidden units	256
Activation	ReLU
Optimizer	Adam
Learning rate for $\boldsymbol{U}_{1_{3:6,:}}$ (scaling)	5×10^{-3}
Learning rate for $\boldsymbol{U}_{1_{6:10,:}}$ (rotation)	10^{-3}
Learning rate for $\boldsymbol{U}_{1_{10:42,:}}$ (feature)	2.5×10^{-3}
Learning rate for $\boldsymbol{U}_{1_{42:43,:}}$ (opacity)	5×10^{-2}
Initial learning rate for $\boldsymbol{U}_{1_{:3,:}}$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$	1.6×10^{-4}
Final learning rate or $\boldsymbol{U}_{1_{:3,:}}$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$	1.6×10^{-6}
Initial learning rate for rigid MLP	10^{-4}
Initial learning rate for non-rigid and color MLP	10^{-3}
Final learning rate for MLPs	10^{-6}
Learning rate schedule	exponential decay
Max iterations	10^5

Table 2: Hyper-parameters of our architecture.

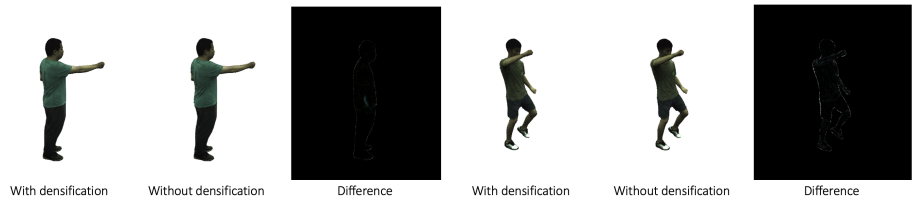


Fig. 7: With or without adaptive densification scheme for the 3D Gaussians [2].

this way, the learnable matrices $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$ are directly optimized with gradient descent, without any non-differentiable updates, and include all parameters for all the training identities. In our preliminary experiments, we observed that we achieve similar results for a single identity with or without densification and pruning, using $N_g = 5 \times 10^4$ (see Fig. 7). However, including an adaptive density control of the Gaussians can be explored as future work.

For $N_i = 30$ identities, MIGS learns only $(M + N_i + N_g)R = (43 + 30 + 5 \times 10^4) \times 100 \approx 5 \times 10^6$ parameters, compared to $MN_iN_g \approx 6.5 \times 10^7$ that would be required by single-identity 3DGS representations, leading to a decrease *by at least one order of magnitude* in the total number of learnable parameters.

D Limitations

We observe that in some cases, our multi-identity network may fail to capture fine-grained details, such as high-frequency texture in clothes or facial details. This tends to happen more when the number of identities increases beyond 20, since the network leverages information from multiple identities for learning, thus smoothing the result. In our work, we tackle this smoothing with the personalization procedure (see Sec. 4.4 and Fig. 2a). In the future, we plan to further enhance high-frequency details by using higher-resolution data and exploring other tensor structures. In addition, in the technical component, we have no theoretical proof that the CP decomposition is the optimal way to factorize the parameters, but we notice empirically that this suffices in our particular case.

E Ethical Considerations

We would like to note the potential misuse of video synthesis methods. With the advances in neural rendering, recent methods can generate photo-realistic human avatars. Our research focuses on human body animation and we presented as main application the animation of human avatars under challenging dance sequences. In contrast to deep fakes, we believe this application cannot be used to spread misinformation or for other harmful purposes. However, we would like to emphasize that there is still a risk of using such methods to generate misleading content. Thus, research on fake content detection and forensics is crucial. We intend to share our source code to help improving such research.

References

1. Hu, S., Liu, Z.: Gauhuman: Articulated gaussian splatting from monocular human videos. arXiv preprint arXiv:2312.02973 (2023)
2. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. Advances in Neural Information Processing Systems **34**, 24741–24752 (2021)
5. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Learn to dance with aist++: Music conditioned 3d dance generation (2021)
6. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6508 (2021)
7. Mihajlovic, M., Zhang, Y., Black, M.J., Tang, S.: LEAP: Learning articulated occupancy of people. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2021)
8. Mu, J., Sang, S., Vasconcelos, N., Wang, X.: ActorsNeRF: animatable few-shot human rendering with generalizable nerfs pp. 18391–18401 (2023)

9. Pan, X., Yang, Z., Ma, J., Zhou, C., Yang, Y.: Transhuman: A transformer-based human representation for generalizable neural human rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3544–3555 (October 2023)
10. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
11. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)
12. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. arXiv preprint arXiv:2312.09228 (2023)
13. Tsuchida, S., Fukayama, S., Hamasaki, M., Goto, M.: Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019. Delft, Netherlands (Nov 2019)
14. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 16210–16220 (2022)
15. Yu, Z., Cheng, W., Liu, X., Wu, W., Lin, K.Y.: Monohuman: Animatable human neural field from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16943–16953 (2023)
16. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. arXiv (2023)

Fig. 8: Animation of human avatars under novel poses. Qualitative comparison with state-of-the-art approaches, namely HumanNeRF [14], MonoHuman [15], and 3DGS-Avatar [12]. The subjects are from ZJU-MoCap [11] and AIST++ [5,13] datasets. The target poses (column 1) are unseen during training, from unseen camera views and advanced dance videos. Our method robustly animates all the identities under challenging novel poses, outperforming the other methods.