# Supplemental Materials for TAVA: Template-free Animatable Volumetric Actors

Ruilong Li[1,3], Julian Tanke[2,3], Minh Vo[3], Michael Zollhöfer[3], Jürgen Gall[2], Angjoo Kanazawa[1], Christoph Lassner[3]

1. UC Berkeley 2. University of Bonn 3. Meta Reality Labs Research

## 1 Inverse Skinning Gradients.

As described in Sec. 3.3, for each sample $\mathbf{x}_v$ in the view space, we find its canonical correspondence $\mathbf{x}_c$ through root finding:

$$\text{Find } \mathbf{x}_c^*, \quad \text{s.t. } f(\mathbf{x}_c^*) = LBS(\mathbf{w}(\mathbf{x}_c^*; \Theta_s), \mathbf{P}, \mathbf{x}_c^*) + \Delta_w(\mathbf{x}_c^*, \mathbf{P}; \Theta_\Delta) - \mathbf{x}_v = \mathbf{0} \quad (1)$$

In order to optimize the skinning deformation defined by ($F_{\Theta_s}$, $F_{\Theta_\Delta}$), we need to determine the gradients of the overall loss $\mathcal{L}$ w.r.t the network parameters ($\Theta_s$, $\Theta_\Delta$):

$$\frac{\partial \mathcal{L}}{\partial \Theta_s} = \left[\frac{\partial \mathcal{L}}{\partial \mathbf{x}_c^*}\right]\left[\frac{\partial \mathbf{x}_c^*}{\partial \Theta_s}\right]. \qquad \frac{\partial \mathcal{L}}{\partial \Theta_\Delta} = \left[\frac{\partial \mathcal{L}}{\partial \mathbf{x}_c^*}\right]\left[\frac{\partial \mathbf{x}_c^*}{\partial \Theta_\Delta}\right]. \quad (2)$$

The first term $[\partial \mathcal{L}/\partial \mathbf{x}_c^*]$ can be easily calculated through back-propagation. The second terms $[\partial \mathbf{x}_c^*/\partial \Theta_s]$ and $[\partial \mathbf{x}_c^*/\partial \Theta_\Delta]$ can be calculated analytically via implicit differentiation [2]:

$$\text{Let,} \quad \Gamma(\mathbf{x}_c^*, \mathbf{P}; \Theta_s, \Theta_\Delta) = LBS(\mathbf{w}(\mathbf{x}_c^*; \Theta_s), \mathbf{P}, \mathbf{x}_c^*) + \Delta_w(\mathbf{x}_c^*, \mathbf{P}) \quad (3)$$

$$\text{Then,} \quad \Gamma(\mathbf{x}_c^*, \mathbf{P}; \Theta_s, \Theta_\Delta) - \mathbf{x}_v = \mathbf{0} \quad (4)$$

$$\Leftrightarrow \quad \frac{\partial \Gamma(\mathbf{x}_c^*, \mathbf{P}; \Theta_s, \Theta_\Delta)}{\partial \Theta_s} + \frac{\partial \Gamma(\mathbf{x}_c^*, \mathbf{P}; \Theta_s, \Theta_\Delta)}{\partial \mathbf{x}_c^*} \cdot \frac{\partial \mathbf{x}_c^*}{\partial \Theta_s} = \mathbf{0} \quad (5)$$

$$\Leftrightarrow \quad \frac{\partial \mathbf{x}_c^*}{\partial \Theta_s} = -\left[\frac{\partial \Gamma(\mathbf{x}_c^*, \mathbf{P}; \Theta_s, \Theta_\Delta)}{\partial \mathbf{x}_c^*}\right]^{-1}\left[\frac{\partial \Gamma(\mathbf{x}_c^*, \mathbf{P}; \Theta_s, \Theta_\Delta)}{\partial \Theta_s}\right] \quad (6)$$

$$\Leftrightarrow \quad \frac{\partial \mathbf{x}_c^*}{\partial \Theta_s} = -\left[\frac{\partial \mathbf{x}_v}{\partial \mathbf{x}_c^*}\right]^{-1}\left[\frac{\partial \mathbf{x}_v}{\partial \Theta_s}\right] \quad (7)$$

$$\text{Similarly,} \quad \frac{\partial \mathbf{x}_c^*}{\partial \Theta_\Delta} = -\left[\frac{\partial \mathbf{x}_v}{\partial \mathbf{x}_c^*}\right]^{-1}\left[\frac{\partial \mathbf{x}_v}{\partial \Theta_\Delta}\right] \quad (8)$$

## 2 Dataset Splits and Pose Clustering.

[†]Work done partially while Ruilong and Julian were at Meta Reality Labs Research.

As described in Sec 4.1, to avoid similar poses appearing in both the training and the validation set, we split the dataset by clustering the frames based on pose similarity. Fig. 1 shows an example of the pose similarity matrix on ZJU-Mocap subject 313. It clearly shows that this actor moves with a repetitive motion pattern. Thus the previous way [5,6] of splitting the dataset into two chunks with consecutive frames will cover similar poses in both sets, which is not suitable for evaluating the pose generalization ability. This motivated us to introduce our new data split protocol based on pose clustering.

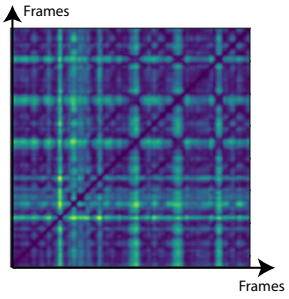



**Fig. 1.** *Pose Similarity Matrix on ZJU Subject 313.*

Our pose clustering process is as follows: We first disable the global (root) transformation for all poses. Then, the difference of two poses is measured by the Euclidian distance of their corresponding mesh vertices (SMPL mesh for ZJU-Mocap). As we only use pose clustering to construct the dataset splits, the mesh information is considered accessible here. Next the K-Medoids algorithm is adopted to cluster the poses into $K = 10$ clusters (Examples shown in Fig. 2). Finally we calculate the average difference between the K medoids to find the most different one, which we regard as the out-of-distribution poses to form the OOD validation set.

## 3 Implementation Details.

TAVA employs four MLPs ($F_{\Theta_r}$, $F_{\Theta_\Delta}$, $F_{\Theta_r}$, $F_{\Theta_a}$) in our method. Both, $F_{\Theta_r}$ and $F_{\Theta_\Delta}$, consist of 4 layers with 128 hidden units, with 4-degree positional encoding [3] on the input coordinates. $F_{\Theta_r}$ is an 8-layer MLP with 256 hidden units and uses 10-degree integrated positional encoding, similar to Mip-NeRF [1]. $F_{\Theta_a}$ is a single-layer MLP with 128 hidden units that connects to the 8-th layer of $F_{\Theta_r}$. We follow the hyper-parameters in Mip-NeRF [1] for the volume rendering, where 64 samples are drawn for each ray at both coarse and fine level.

## 4 Baseline Implementation Details.

As described in Sec. 4.2, for the template-based baselines Animatable-NeRF [5] and NeuralBody [6], we use their official implementations. For the template-free baselines NARF [4] and A-NeRF [7], we re-implemented them in our code base for fair comparison. We also carefully adapt their official implementations to the ZJU-Mocap dataset to verify our re-implementation. As shown in Tab. 1, our re-implementation achieves better performance than the official implementations on ZJU-Mocap subject 313.

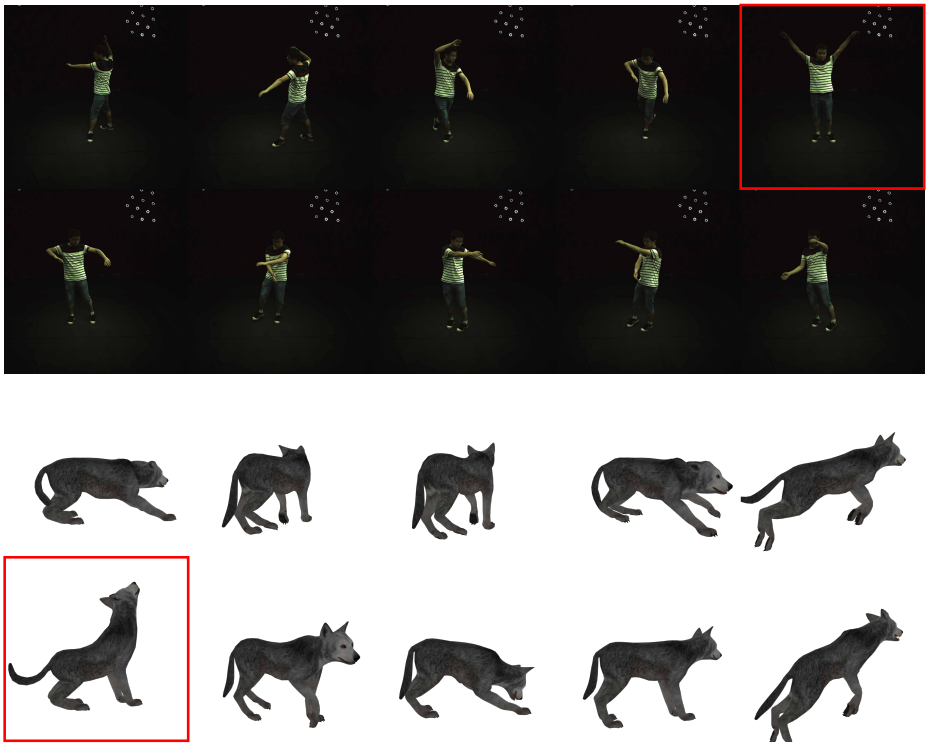


**Fig. 2.** *Pose Clustering*. Here we show the K-Medoids clustering results on ZJU-Mocap subject 315 and the Wolf subject. The one cluster marked as red is automatically identified as the most different one thus is selected as the OOD validation set.

| | Novel-view | | Novel-pose (ind) | | Novel-pose (ood) | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| A-NeRF(official) | 28.36 | 0.945 | 28.80 | 0.947 | 27.69 | 0.928 |
| A-NeRF(re-impl.) | **32.14** | **0.976** | **33.39** | **0.980** | **28.53** | **0.953** |
| NARF(official) | 30.65 | 0.962 | 32.22 | 0.969 | **28.10** | 0.944 |
| NARF(re-impl.) | **33.17** | **0.979** | **34.67** | **0.983** | 28.05 | **0.951** |

**Table 1.** *Re-implementation verification on ZJU-Mocap Subject 313.*

## 5 Visualizations for Skinning Weights.

Fig. 3 shows results of our learned skinning weights and canonical geometry for animal subjects and ZJU-Mocap data. The surfaces are extracted with marching cube algorithm with threshold 5.0 on the density field.

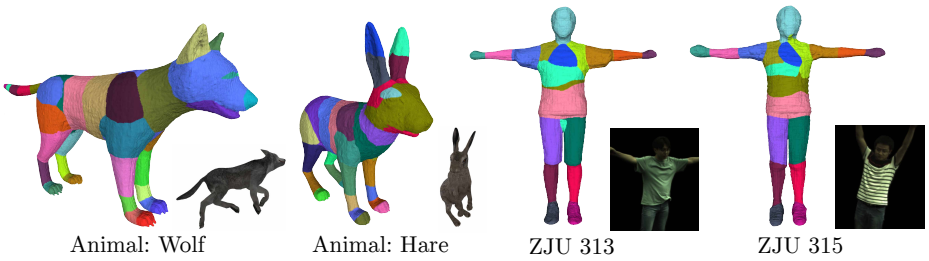


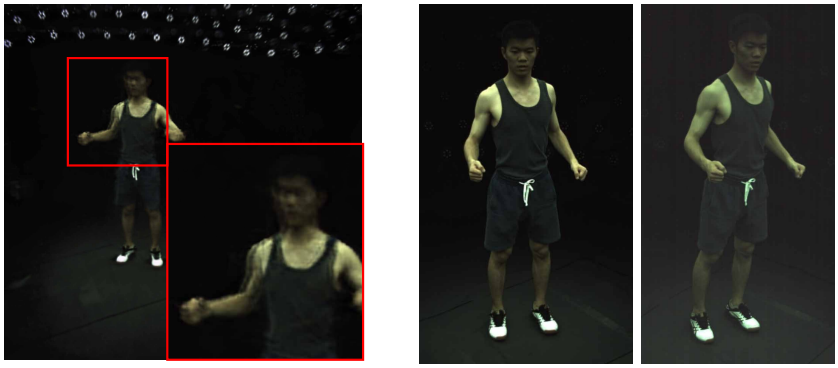

**Fig. 3.** *Learned skinning weights and canonical geometry.* Color denotes the top-1 bone from skinning weights.

## 6 Challenges in ZJU-Mocap Dataset.

The ZJU-Mocap dataset has become an increasingly popular dataset to study human performance capture, reconstruction, and neural rendeirng [5,6,8]. Yet we notice that there are a few issues in this dataset that are neither addressed nor mentioned in the previous works, including *imperfect camera calibrations* and *various camera exposures* (as shown in Fig. 4). **We also get acknowledged from the authors of ZJU-Mocap Dataset on those issues.** We discovered these issues after the submission so they are not considered in our designs. Yet they greatly affect both our performance and the baselines'. We believe it is worth to point them out so that they can be considered in the future research.

## 7 Ablation Studies.

Thanks to our model design, we can train a full model with the non-linear deformation $\Delta_v$ and the ambient occlusion AO enabled, then strip them out at inference time. Fig. 5 shows a qualitative result to visually demonstrate the impacts on the full model. Notice that without AO, the shading effects are removed during rendering, which produces overall brighter images than the ground-truth. This is an expected effect, but prohibits quantitative evaluation. Furthermore, we ablate these two design decisions during training. For the non-linear deformation, our ablation is to simply disable it during training to see if the LBS is enough to model the deformation. For the AO, our ablation is to compare it with predicting a pose-dependent color by conditioning pose to the



Imperfect Camera Calibrations Various Camera Exposures

**Fig. 4.** *Challenges in ZJU-Mocap Dataset.* **Left**: We here train a standard NeRF [9] on a single frame with all the views. Imperfect camera calibrations cause the ghost effects on some views. **Right**: We here compare two groundtruth images side by side for the same subject with different cameras.

color branch of $F_{\Theta_r}$, and disabling the AO branch $F_{\Theta_a}$. As shown in Tab. 2, both design decisions contribute to the final model performance. Lastly, we show the two different strategies to deal with root finding failures described in the paper in Tab. 2. Using the interpolation strategy results in slightly better performance.

| | Novel-view | | Novel-pose (ind) | | Novel-pose (ood) | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| w/o non-linear $\Delta_v$ | 31.86 | 0.974 | 32.19 | 0.975 | 30.56 | 0.965 |
| w/o AO (pose-dep color) | 32.94 | 0.980 | 33.20 | 0.980 | 30.57 | 0.968 |
| w/o r.f. interplation | 33.02 | 0.980 | 33.31 | 0.981 | **30.72** | **0.969** |
| Ours | **33.11** | **0.981** | **33.35** | **0.981** | 30.69 | **0.969** |

**Table 2.** *Model ablations on the ZJU Mocap subjects.*

## 8 Per-subject Breakdown Comparisons.

We also report a per-subject breakdown of the quantitative metrics against all baseline methods in Tab. 3 and Tab. 4.

| | Novel-view | | Novel-pose (ind) | | Novel-pose (ood) | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| ***Subject 313*** | | | | | | |
| Animatable-NeRF [5] | 29.15 | 0.967 | 28.67 | 0.968 | 28.00 | 0.952 |
| NeuralBody [6] | 33.81 | 0.982 | 34.20 | 0.984 | 28.33 | 0.958 |
| Pose-NeRF | 31.94 | 0.974 | 33.22 | 0.979 | 27.03 | 0.941 |
| A-NeRF [7] | 32.14 | 0.976 | 33.39 | 0.980 | 28.53 | 0.953 |
| NARF [4] | 33.17 | 0.979 | 34.67 | 0.983 | 28.05 | 0.951 |
| Ours | 33.14 | 0.979 | 34.51 | 0.984 | 29.28 | 0.957 |
| ***Subject 315*** | | | | | | |
| Animatable-NeRF [5] | 27.62 | 0.962 | 26.50 | 0.956 | 25.52 | 0.949 |
| NeuralBody [6] | 31.41 | 0.982 | 30.35 | 0.984 | 25.87 | 0.957 |
| Pose-NeRF | 28.96 | 0.970 | 28.81 | 0.969 | 24.30 | 0.930 |
| A-NeRF [7] | 29.67 | 0.974 | 29.46 | 0.973 | 26.82 | 0.959 |
| NARF [4] | 30.18 | 0.977 | 29.96 | 0.976 | 27.05 | 0.960 |
| Ours | 30.84 | 0.980 | 30.61 | 0.979 | 26.55 | 0.960 |
| ***Subject 377*** | | | | | | |
| Animatable-NeRF [5] | 32.17 | 0.979 | 30.20 | 0.974 | 28.95 | 0.969 |
| NeuralBody [6] | 33.86 | 0.985 | 32.96 | 0.983 | 31.55 | 0.978 |
| Pose-NeRF | 32.34 | 0.978 | 32.14 | 0.978 | 29.58 | 0.970 |
| A-NeRF [7] | 32.62 | 0.980 | 32.53 | 0.980 | 31.77 | 0.978 |
| NARF [4] | 32.87 | 0.982 | 32.83 | 0.981 | 31.83 | 0.978 |
| Ours | 33.08 | 0.982 | 33.05 | 0.982 | 32.26 | 0.980 |
| ***Subject 386*** | | | | | | |
| Animatable-NeRF [5] | 34.07 | 0.975 | 32.00 | 0.967 | 32.23 | 0.974 |
| NeuralBody [6] | 36.55 | 0.985 | 36.19 | 0.984 | 35.57 | 0.983 |
| Pose-NeRF | 34.30 | 0.977 | 34.19 | 0.977 | 32.81 | 0.973 |
| A-NeRF [7] | 35.37 | 0.981 | 35.23 | 0.980 | 34.50 | 0.979 |
| NARF [4] | 35.53 | 0.981 | 35.39 | 0.981 | 35.46 | 0.982 |
| Ours | 35.38 | 0.981 | 35.25 | 0.980 | 34.68 | 0.980 |

**Table 3.** *Per-Subject Comparisons on the ZJU Mocap Dataset.*

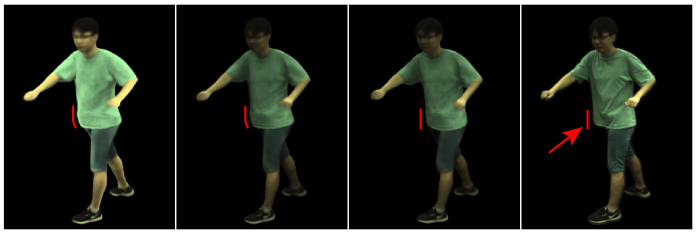



**Fig. 5.** *Ablation on the ambient occlusion (AO) and non-linear deformation ($\Delta_v$) terms.* Due to our designs, we can train our full model with both enabled, then disable them during inference to ablate their effects.

| | Novel-view | | Novel-pose (ind) | | | Novel-pose (ood) | | |
|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | P2P ↓ | PSNR ↑ | SSIM ↑ | P2P ↓ |
| ***Subject Hare*** | | | | | | | | |
| Pose-NeRF | 23.97 | 0.949 | 22.28 | 0.942 | 197.01 | 15.35 | 0.925 | 100.72 |
| A-NeRF [7] | 31.33 | 0.974 | 31.28 | 0.974 | 35.44 | 23.00 | 0.960 | 26.53 |
| NARF [4] | 36.45 | 0.986 | 36.56 | 0.986 | 10.90 | 29.40 | 0.979 | 5.64 |
| Ours | 37.35 | 0.990 | 37.57 | 0.990 | 5.04 | 35.24 | 0.989 | 3.91 |
| ***Subject Wolf*** | | | | | | | | |
| Pose-NeRF | 22.83 | 0.946 | 21.57 | 0.941 | 197.21 | 17.90 | 0.925 | 76.97 |
| A-NeRF [7] | 31.20 | 0.979 | 31.16 | 0.979 | 27.60 | 28.31 | 0.974 | 11.55 |
| NARF [4] | 36.64 | 0.989 | 36.74 | 0.990 | 7.65 | 32.43 | 0.985 | 11.27 |
| Ours | 37.26 | 0.992 | 37.33 | 0.992 | 3.57 | 36.30 | 0.992 | 2.85 |

**Table 4.** *Per-Subject Comparisons on the animal subjects Hare and Wolf.*

## References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: International Conference on Computer Vision (2021)
2. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11594–11604 (2021)
3. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
4. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5762–5772 (2021)
5. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: International Conference on Computer Vision (2021)
6. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021)
7. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. Advances in Neural Information Processing Systems **34** (2021)
8. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. CVPR (2022)
9. Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Conference on Computer Vision and Pattern Recognition (2022)