CAT-NeRF: Constancy-Aware Tx^{2}Former for Dynamic Body Modeling
Supplementary Material

In this supplementary material, we present some further background knowledge, experimental details, ablation studies, and visualization results that do not fit into the paper due to space limitations. We first introduce the background of NeRF, followed by the detailed network architecture description and the ablation study for the size of the latent feature and the number of views we used in the experiment. Finally, we present more visualization results on H36M and ZJU-MoCap datasets with novel views and poses.

Review of Neural Radiance Field. NeRF is a continuous volumetric field representing the density and color of the object or scene. For the color and density of a point for the viewing direction d, NeRF finds the ray r that goes through the field and predicts the density $\sigma(x')$ and color value $c(x')$ for the 3-D point x' on this ray following $\sigma(x'), c(x') = \mathbf{F}(x', d, l)$ using an MLP network $\mathbf{F}(\cdot)$. l is a latent representation for the appearance of the object following [23]. The final predicted color $\tilde{C}(r)$ projected on the 2-D plain for the ray r is accumulated with a random set of quadrature points $\{h_p\}_{p=1}^{m} \in [h_n, h_f]$ following [24]

$$\tilde{C}(r) = \sum_{p=1}^{m} exp(-\sum_{q=1}^{p-1} \sigma_q \delta_q)(1 - exp(-\delta_p \sigma_p))c(p) \quad (9)$$

where h_n and h_f represent the near and far bound of the field for the sampled points. δ_p is the distance between two quadrature points h_p and h_{p+1}. Due to the performance of under-fitting on high-frequency patterns, NeRF includes a position encoding $\gamma(x)$ for projecting the point x, normalized to $[-1, 1]$, to the high dimension space following

$$\gamma(x) = (sin(2^0 \pi x), cos(2^0 \pi x), ..., sin(2^{L-1} \pi x), cos(2^{L-1} \pi x)) \quad (10)$$

We follow [24] to use the three normalized coordinates of the point x' with L as 10 and three of the Cartesian viewing direction unit vector for d with L as 4.

Network Architectures. In our experiment, we have five trainable modules in the proposed CAT-NeRF: neural radiance field, neural blend weight field, frame-unique field decoder, constant field decoder, and Tx^{2}Former $G(\cdot)$.

For the neural radiance field, we follow [32] to build an 11-layer MLP for decoding color and density. The dimensionality is set to 256 for the first 10 layers and 128 for the final layer. The positional encoding $\gamma(x)$ is sent to the first MLP layer and concatenated with the output of the 4^{th} layer as the input for 5^{th} MLP layer. The position encoding for the viewing direction $\gamma(d)$ and appearance code l is sent to the 9^{th} and 10^{th} concatenated with the previous layer's output. Density σ and color c are predicted from the 8^{th} and the last MLP layers, respectively.

For the frame-unique and constant field decoders, we use two identical 2-layer MLP networks but do not share weights. The dimensionality for all layers is set to 64. For the neural blend weight field, we use an 8-layer MLP following [32] where every layer is 128-d. We use the concatenation of the outputs from the constant and frame-unique field encoders along with $\gamma(x)$ as input for the first and fifth MLP layers. The final projection of Δw is produced with the exponential output from the last layer. We use ReLU as the activation function following every MLP layer for all our networks except the final output layer. We set the dimensionality of features in Tx^{2}Former $G(\cdot)$ as 128 for both two layers and include one layer of Transformer encoder [46] for T_1 and T_2 in our experiment.

Size of Latent Features. In our model, we introduce using the 128-dimension features for both frame-unique feature and constant feature representation. We show five different variations in the dimensionality for features representation in Table 6. Note that when we only have one 128-D feature vector for the frame-unique feature, the method defaults to Animatable NeRF. Compared with using the unique feature representation for every individual frame, we see that the constant feature alone boosts the performance for dynamic human body rendering and outperforms the original Animatable NeRF. When using the frame-unique feature along with the constant representation, the network shows the best PSNR value, while results for using 64-dim or 128-dim features do not change greatly.

Number of Training Viewpoints. In addition to the main experiments where we use three training camera viewpoints for H36M, we also assess the performance and robustness of using fewer camera viewpoints comparing CAT-NeRF with Animatable NeRF. We train our network on one (viewpoint 0), two (viewpoint 0 and 1), and three (viewpoint 0,1 and 2) camera viewpoints, respectively, and report the PSNR results on the reconstruction of viewpoint 3 for inference.

We show the results in Table 7. With the maximum available viewpoints, both methods show the best performance for dynamic body shape rendering. When we reduce the available number of available viewpoints in the train-

(a) AN (b) Ours (c) Groundtruth (a) AN (b) Ours (c) Groundtruth

Novel View Novel Pose

Figure 6. Visualizations for novel pose and novel view on H36m datasets comparing (a) Animatable NeRF and (b) ours. Images in the column of AN are the outputs from Animatable NeRF.

Feat. Dim	ψ^c	0	64	128	64	128
	ψ^u	128	64	128	128	0
PSNR		22.05	24.16	24.52	24.35	23.14

Table 6. Ablation results for different feature dimensions used by constant feature ψ^c and frame-unique features ψ^u of the network.

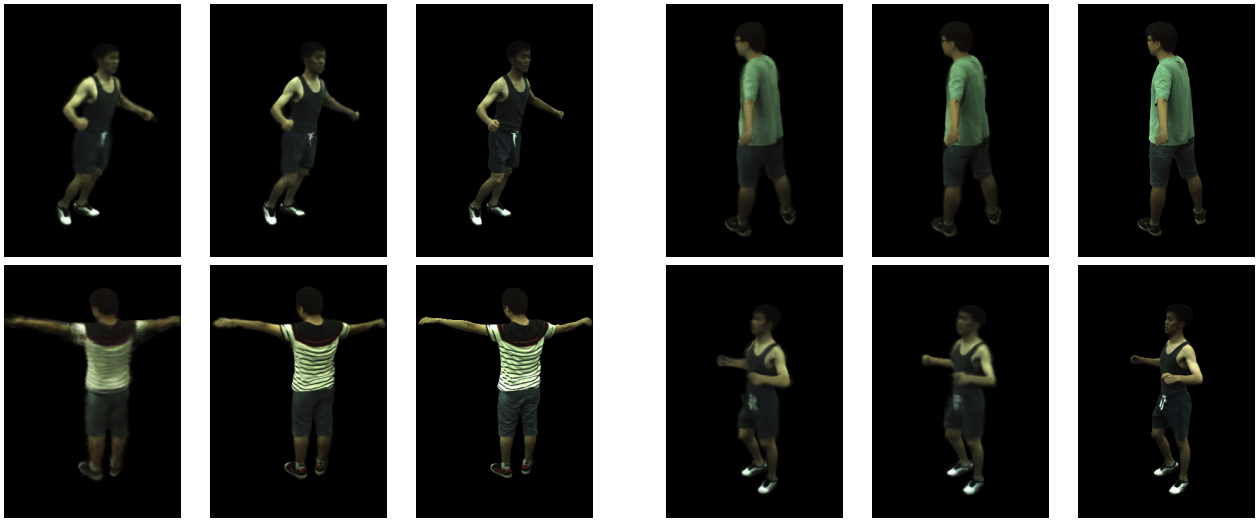
# of Viewpoints	1	2	3
Animatable NeRF	20.78	21.75	22.05
Ours	21.97	24.49	24.52

Table 7. PSNR results for different numbers of training viewpoints compared with Animatable NeRF.

ing set from 3 to 2, our proposed CAT-NeRF shows better consistency than Animatable NeRF, with only a 0.03 drop for PSNR, while Animatable NeRF drops 0.3. When the number of cameras is reduced to one, both methods fall significantly, while CAT-NeRF still achieves a PSNR value of 21.97, similar to the Animatable NeRF three-viewpoint result of 22.05. Mining the temporal constancy assists the model in extracting and extending the frame-level knowledge to video-level knowledge, helping the model achieve better performance even with only one training viewpoint.

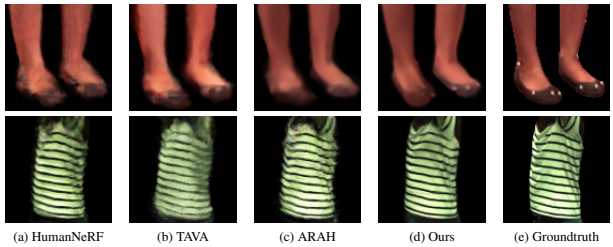
Visualization Results on H36M and ZJU-MoCap. In addition to the visualization results we show in our submission, we present more visualizations for both datasets on both novel view and novel pose settings compared with Animatable NeRF [32]. We show the results for H36M in Figure 6 and ZJU-MoCap in Figure 7.

With CAT-NeRF, we see better-detailed reconstructions than Animatable NeRF on both datasets. For example, for both novel view and novel pose settings in the last row of Figure 6, Animatable NeRF introduces wrinkles that should not appear at the front of the person, while our method creates a more precise image. We can also observe similar differences in the bottom left example in Figure 7, where Animatable NeRF fails to reconstruct the stripes on the shirt. In contrast, CAT-NeRF generates an image with a more precise boundary for the patterns and wrinkles on the clothes. In addition to the framewise results in the figures, we also



(a) AN (b) Ours (c) Groundtruth (a) AN (b) Ours (c) Groundtruth

Figure 7. Visualizations for novel pose and novel view on ZJU-MoCap dataset comparing (a) Animatable NeRF and (b) ours. Images in the column of AN are the outputs from Animatable NeRF.



(a) HumanNeRF (b) TAVA (c) ARAH (d) Ours (e) Groundtruth

Figure 8. Visualization resutls for novel poses and novel views settings for examples in H36m and ZJU-MoCap dataset.

attach a video for each dataset in the supplementary folder.

For the quality of predicted images of two datasets, we observe that, compared with H36M, visualization results on ZJU-MoCap are generally better. Since the number of samples used for training in H36M is smaller and pose differences between the frames are significant, H36M is a much harder dataset to render from a novel viewpoint or for novel poses compared with ZJU-MoCap. In addition, ZJU-MoCap has 4 camera viewpoints available during training, while H36M only has 3. In our submission, we present the visualization results from H36M to compare on a more challenging dataset for dynamic human body rendering.

Comparison with the Latest State-of-the-Art Methods: We compared our method with some of the latest state-of-the-art methods such as HumanNeRF [49], TAVA [18], and ARAH [47] as shown in Figure 8. However, as most of these methods do not have results on H36m, the more complex dataset with fewer cameras and more action variations, and the experimental settings were not exactly the same, we rerun the experiments on the two public datasets used in our paper. We present qualitative comparisons on examples from H36m and ZJU-Mocap, where we zoom in on the corresponding body parts. Our approach achieves more precise boundaries of patterns such as stripes and shoes, with fewer artifacts such as wrinkles. Additionally, the key modules in our method are portable and can be applied to these latest state-of-the-art methods for further improvements.