ActorsNeRF: Animatable Few-shot Human Rendering with Generalizable NeRFs Supplementary Materials

Jiteng Mu[1], Shen Sang[2], Nuno Vasconcelos[1], Xiaolong Wang[1]
[1]UC San Diego, [2]ByteDance

1. Overview

In the supplementary materials, we provide more details of the submission: We show more quantitative and qualitative results of few-shot generalization (paper Section 4.2.) on various datasets in Section 2; We present details on the design of each module in ActorsNeRF in Section 3; More implementation details are discussed in Section 4.

2. Few-shot Generalization

This section serves to supplement the primary findings of Section 4.2., where more quantitative and qualitative results are presented. Specifically, the setting considers the support set consists of m frames, where $m = \{5, 10, 30, 100, 300\}$, sampled uniformly from 300 consecutive frames of a monocular video of an unseen subject. The objective is to synthesize the actor from novel viewpoints with novel poses.

More Results. In addition to Neural Body and NumanNeRF, in this section, we additionally provide results compared to NeuMan on the ZJU-dataset, as shown in Table 1, Table 2, and Table 3. NeuMan (NM) is designed to jointly model the human subject and scene by training separate NeRFs for foreground and background. To adapt NeuMan to our monocular few-shot setting, we only train the human NeRF and zero out the output of the scene NeRF. Our results demonstrate that ActorsNeRF outperforms NeuMan in all settings. To provide further visual comparison, we show the rendering results for different methods and shot counts in Figure 2, Figure 3, Figure 4, and Figure 5. The results show that ActorsNeRF produces consistently smoother renderings with higher-quality details and less body distortion, while maintaining a valid shape across all few-shot settings.

We present additional qualitative results on the AIST++ dataset to complement Table 2 in the paper. The results are shown in Figure 6, Figure 7, and Figure 8. Our proposed method, ActorsNeRF, is shown to produce high-quality details, including facial features, while maintaining a smooth boundary. In contrast, the baseline method produces noisy renderings with broken body parts. These results demonstrate the superior performance of ActorsNeRF in animating novel actors with few monocular images under challenging poses. The results also suggest that large-scale pretraining allows the model to generalize better to unseen poses and viewpoints, further highlighting the effectiveness of ActorsNeRF in monocular few-shot 3D human body rendering.

Comparison to Category-level NeRFs. Neural Human Performer and MPS-NeRF are two methods that also employ NeRF models trained at the category-level with encoders. However, there are notable differences between these methods and ActorsNeRF. First, both methods require multi-view images for both training and inference, while ActorsNeRF only requires a few monocular images. This makes our setting more challenging, as our network must reason about pose variances in addition to viewpoint differences when aggregating features from monocular images. Moreover, while Neural Human Performer does not support novel pose synthesis for a novel human subject, ActorsNeRF produces an animatable NeRF model for a new observed person. MPS-NeRF implements a canonical space and thus allows for animation of the new person given multi-view images. In comparison, we design a novel two-level canonical space to better fit individual shapes and incorporates localized SMPL local features for better feature aggregation.

3. Network Architectures

Encoder. We employ ResNet-18 as encoder $\mathcal{E}$ to extract a feature set $\mathbf{U}^k = [\mathbf{U}^k_p, \mathbf{U}^k_s]$. $\mathbf{U}^k_p$ is a 256-dim tensor of concatenated features from Res-Conv1, Res-block 1 and Res-block 2. $\mathbf{U}^k_s$ further applied a convolutional layer on the top of $\mathbf{U}^k_p$ and is with 64 channels.

Skinning Weight Network. The skinning weight network takes the a global 256-dim embedding as input and transforms it into a 1024-dimension vector. After reshaping the vector into $1 \times 1 \times 1 \times 1024$, 5 subsequential 3D transposed convolution layers are applied to obtain a 3D volume of shape $32 \times 32 \times 32 \times 25$. The skinning weights are defined in the category-level canonical space. The final skinning weights are modeled by the combination of a T-pose prior (3D ellipsoidal Gaussian around each bone) and a residual

		Person 387			Person 393			Person 394		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
5-shot	Neuman	26.93	0.9551	59.65	27.17	0.9544	60.96	27.88	0.9474	59.68
	Ours	**27.26**	**0.9568**	**46.06**	**27.20**	**0.9553**	**46.29**	**28.09**	**0.9577**	**42.48**
10-shot	Neuman	26.90	0.9551	51.14	27.39	0.9547	54.70	28.71	0.9582	48.35
	Ours	**27.15**	**0.9592**	**40.89**	**27.26**	**0.9565**	**42.56**	**28.71**	**0.9613**	**35.93**
30-shot	NeuMan	27.14	0.9580	46.60	27.33	0.9573	48.39	28.73	0.9606	42.55
	Ours	**27.67**	**0.9610**	**36.76**	**27.59**	**0.9577**	**39.51**	**28.97**	**0.9614**	**34.29**
100-shot	NeuMan	27.40	0.9568	46.38	27.47	0.9576	48.02	28.96	0.9614	41.22
	Ours	**27.66**	**0.9614**	**36.39**	**27.57**	**0.9580**	**39.33**	**29.07**	**0.9612**	**34.03**
300-shot	NeuMan	27.61	0.9596	44.64	27.40	0.9577	48.46	28.70	0.9603	42.29
	Ours	**27.61**	**0.9612**	**36.18**	**27.59**	**0.9574**	**39.36**	**28.98**	**0.9611**	**34.17**

Table 1: Few-shot generalization comparison for novel view synthesis of novel actors with unseen poses on the ZJU-MoCap dataset.

		Person 16			Person 17			Person 18			Person 19			Person 20		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
5-shot	NM	24.39	0.9763	29.21	24.77	0.9774	29.19	22.73	0.9745	31.65	24.60	0.9778	29.53	24.01	0.9786	28.17
	ours	**25.22**	**0.9796**	**22.03**	**25.88**	**0.9808**	**22.85**	**24.50**	**0.9811**	**22.38**	**25.24**	**0.9801**	**22.87**	**25.30**	**0.9827**	**21.34**
30-shot	NM	24.54	0.9768	27.44	24.72	0.9775	28.70	23.06	0.9755	30.51	24.32	0.9775	28.75	24.70	0.9801	26.40
	ours	**25.67**	**0.9806**	**20.18**	**26.06**	**0.9826**	**19.45**	**24.82**	**0.9826**	**19.52**	**25.53**	**0.9818**	**19.98**	**25.58**	**0.9840**	**18.49**
300-shot	NM	25.12	0.9793	26.75	25.04	0.9793	28.16	23.76	0.9780	28.47	24.94	0.9795	27.84	24.64	0.9802	26.37
	ours	**25.73**	**0.9812**	**18.93**	**26.14**	**0.9834**	**18.37**	**25.03**	**0.9833**	**18.52**	**25.88**	**0.9827**	**18.58**	25.78	**0.9845**	**17.44**

Table 2: Few-shot generalization comparison for novel view synthesis of novel actors with unseen poses on the AIST++ dataset.

term (generated by the skinning weight network).

Deformation Network. The deformation network takes the concatenated K pixel-aligned local feature, body pose vector, and the sampled point in the category-level canonical space as input and transforms it into a position in the instance-level canonical space. Specifically, the final location in the instance-level canonical space is modeled by a combination of the $\mathbf{x}_c$ and a residual term (output from the deformation network). The network is composed of 6 fully connected layers with ReLU activation functions.

Rendering Network. The rendering network takes as input the point in the instance-level canonical space. As shown in Figure 1, the pixel-aligned local features and the SMPL local features, and output color and density. The pixel-aligned local features and SMPL features are first passed through separate linear layers to form two 256-dim individual embedding. These features are then passed to the following MLP together with the coordinate embedding, to obtain the final predictions. The network consists of multiple fully connected layers with ReLU activations.

4. Implementation Details

Dataset. We test ActorsNeRF on two benchmark datasets: the ZJU-MoCap dataset and the AIST++ Dataset.

The ZJU-MoCap Dataset dataset contains 10 human subjects recorded from 21/23 multi-view cameras. We use the camera projections, body poses, and segmentations provided by the dataset. Three subjects (person 387, person 393, person 394) were designated as held-out data, while the remaining seven were used for training. 'camera1' is used for learning, whereas other views were solely used for evaluation. Each training video contains 550 frames that cover the person from various angles. For the few-shot experiments, the few-shot support set images are uniformly selected from the a consecutive 300-frame video sequence captured from 'camera1'. For example, for the 5-shot setting, the 5 support set frames were frame0, frame75, frame150, frame225, frame300, respectively. For evaluation, the novel poses are sampled every 10 frames from the rest of the video (frame301-frame550).

The AIST++ Dataset is a collection of dance motion data consisting of 30 human subjects performing various dances, captured from 9 multi-view cameras. We utilized the camera projections, body poses provided by the dataset and use PointRend to acquire foreground masks. We randomly select 30 action sequences, and split the dataset with 25 actors for training and the remaining 5 actors (person 16-20) for evaluation. 'camera1' is used for learning and other views (except 'camera9') are used for evaluation. For each monocular video, starting from frame 200, we used

	Person 387			Person 393			Person 394		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeuMan	24.10	0.9388	78.94	25.33	0.9386	78.99	26.33	0.9422	76.42
Ours	**26.19**	**0.9542**	**50.88**	**26.75**	**0.9546**	**47.02**	**27.84**	**0.9578**	**44.77**

Table 3: Short-video generalization comparison for novel views of novel actors with unseen poses on the ZJU-MoCap dataset.

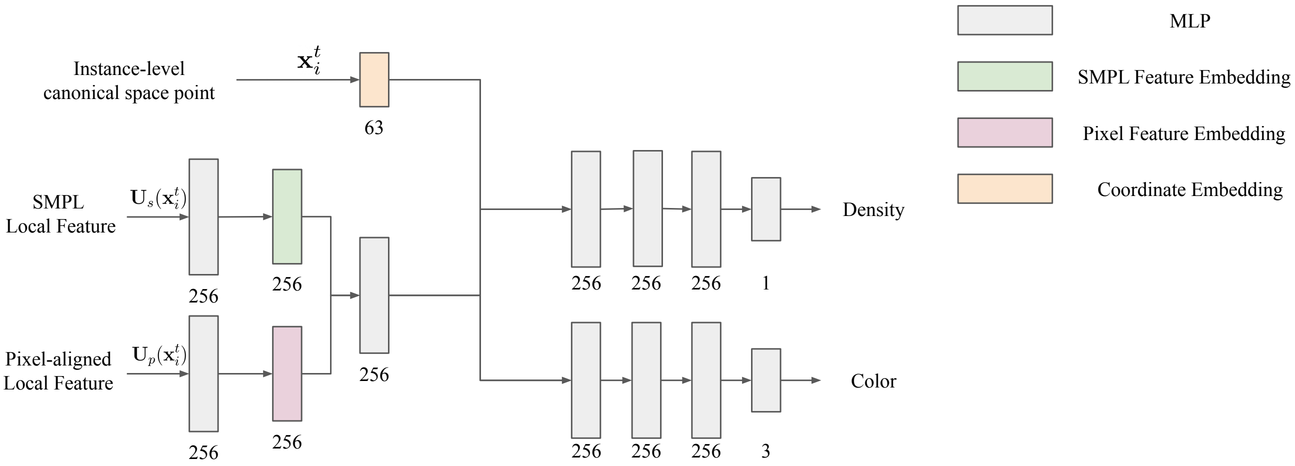

Figure 1: Rendering network architecture.

the following 500 frames sampled every 4 frames to build the dataset. For the few-shot experiments, the few-shot support set images were selected from a consecutive 300-frame video sequence captured from ‘camera1’. Specifically, except for 5 frames that are mannually selected that roughly covered the person from both front and back, others are uniformly sampled images. During the evaluation stage, novel poses were sampled every 10 frames from the rest of the video.

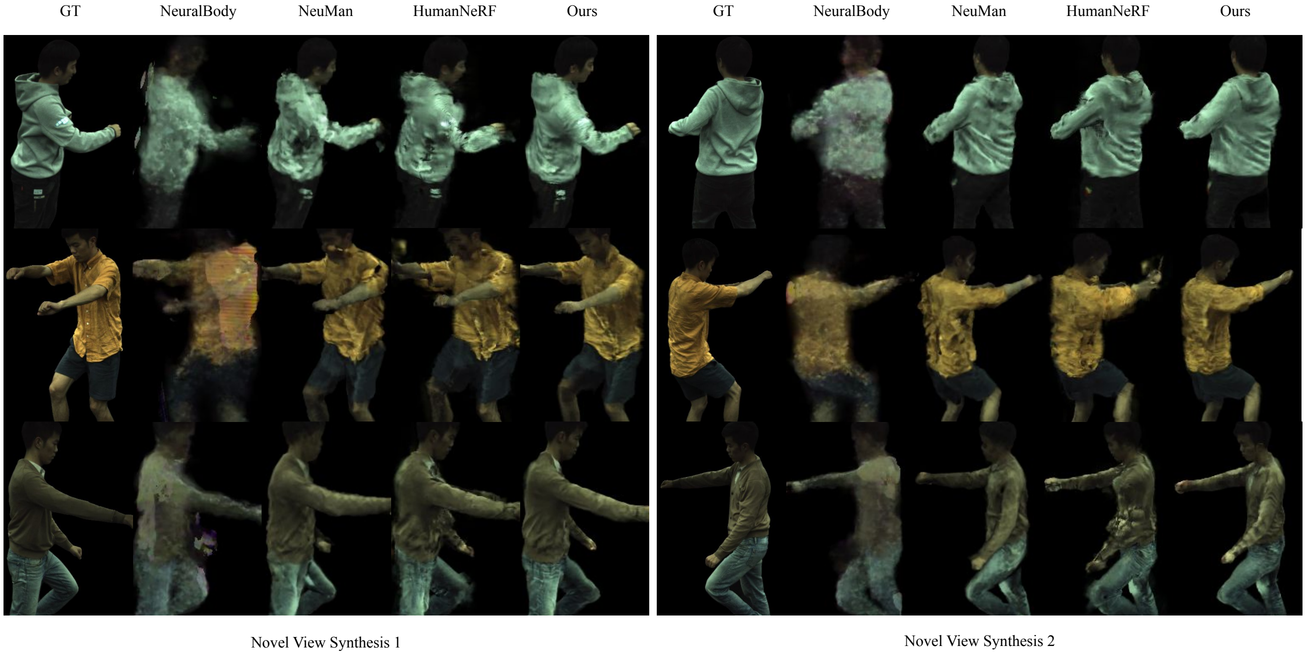

Figure 2: Qualitative comparison for 5-shot novel view synthesis of novel actors with unseen poses on the ZJU-MoCap dataset. Our method achieves high-quality animation with sharp boundary and details.

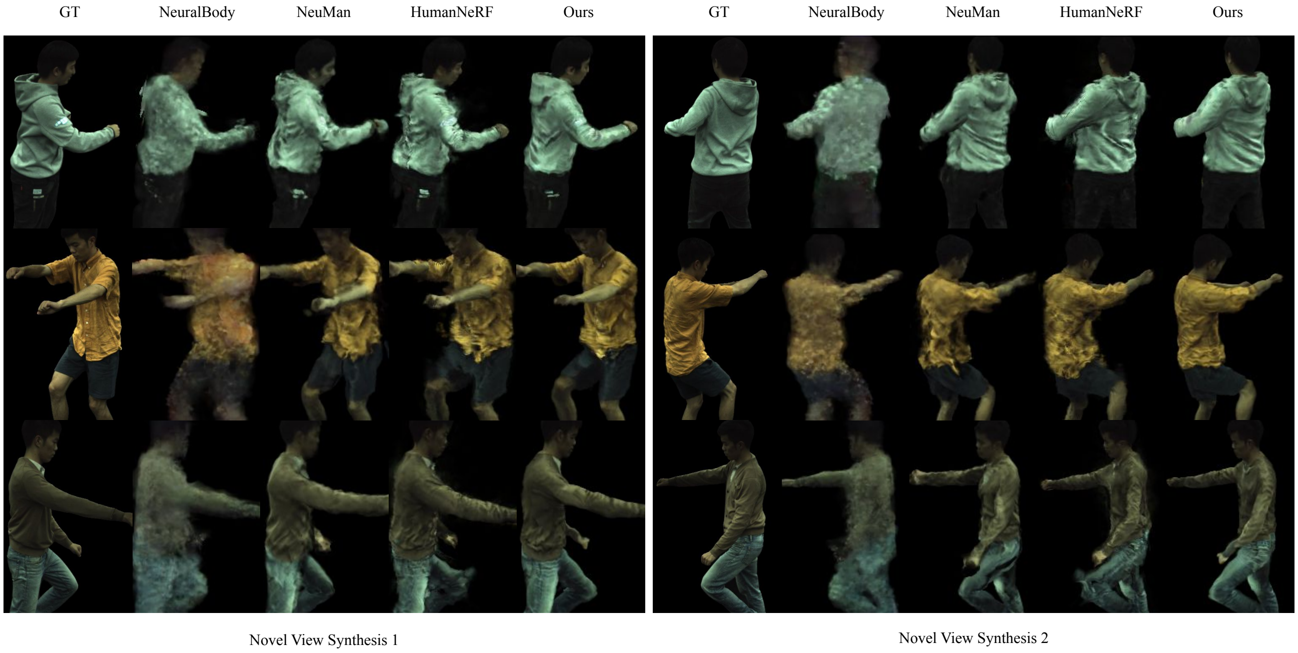

Figure 3: Qualitative comparison for 10-shot novel view synthesis of novel actors with unseen poses on the ZJU-MoCap dataset. Our method achieves high-quality animation with sharp boundary and details.

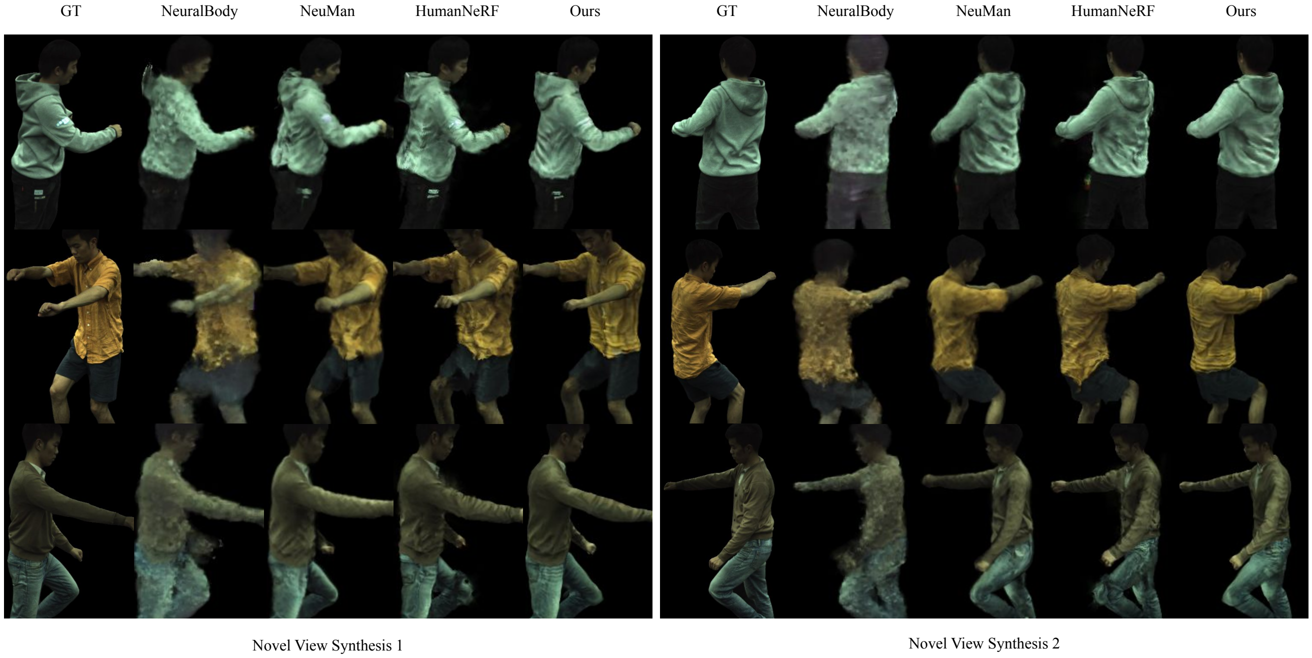

Figure 4: Qualitative comparison for 30-shot novel view synthesis of novel actors with unseen poses on the ZJU-MoCap dataset. Our method achieves high-quality animation with sharp boundary and details.

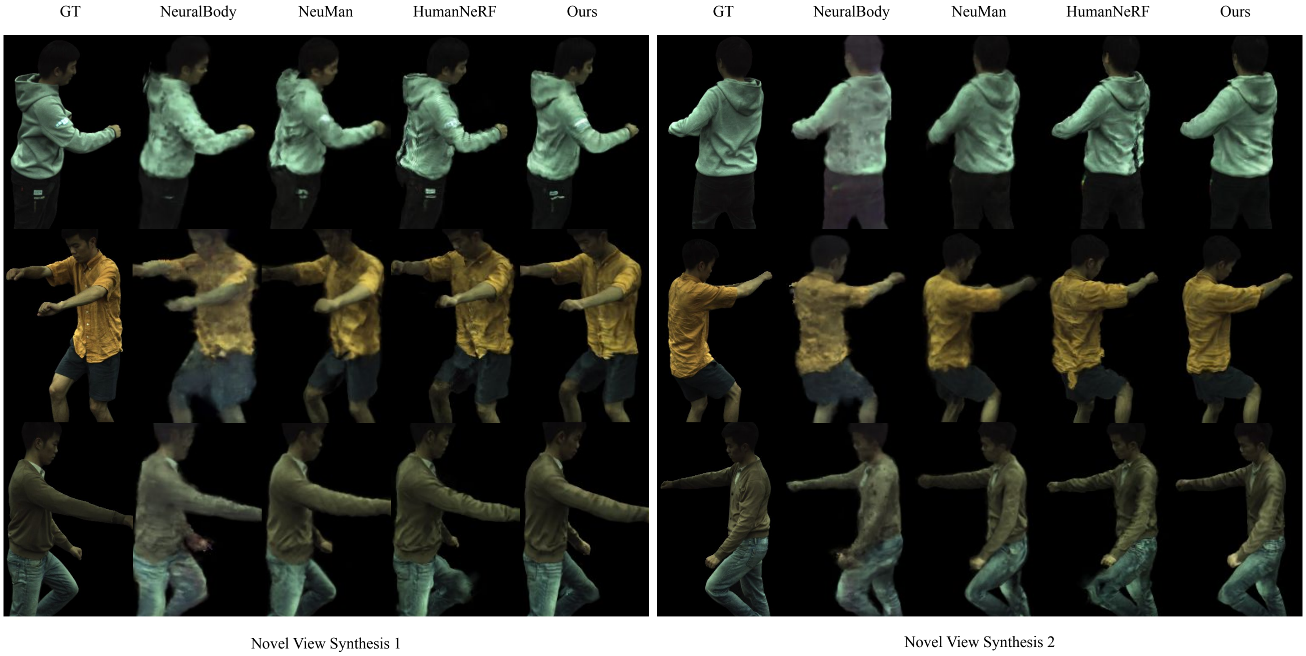

Figure 5: Qualitative comparison for 100-shot novel view synthesis of novel actors with unseen poses on the ZJU-MoCap dataset. Our method achieves high-quality animation with sharp boundary and details.

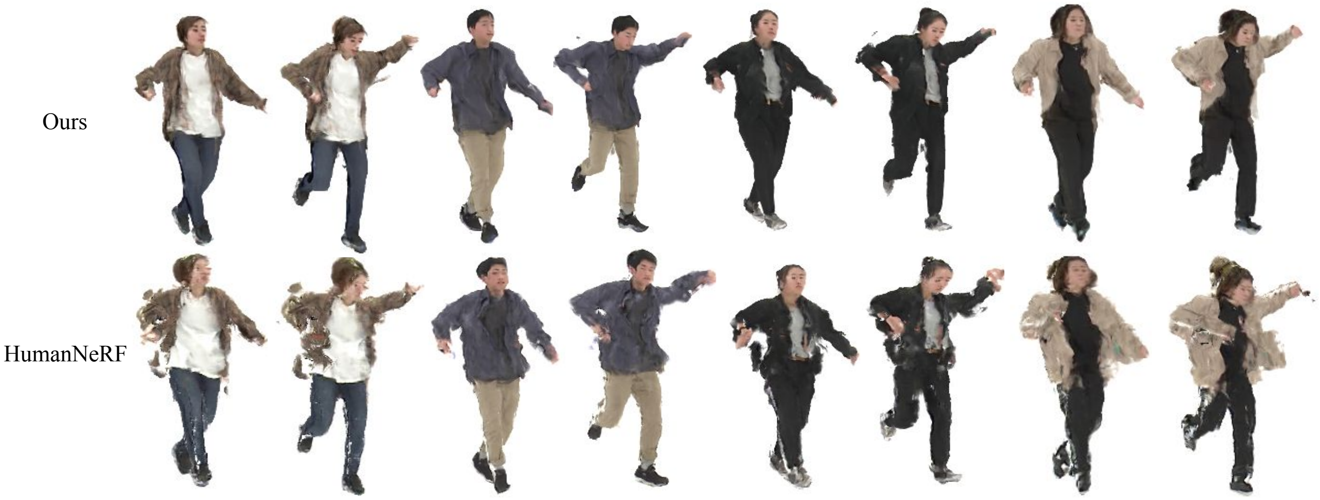

Figure 6: Qualitative comparison for 10-shot novel view synthesis of novel actors with unseen poses on the AIST++ dataset. Our method achieves high-quality animation with sharp boundary and details.

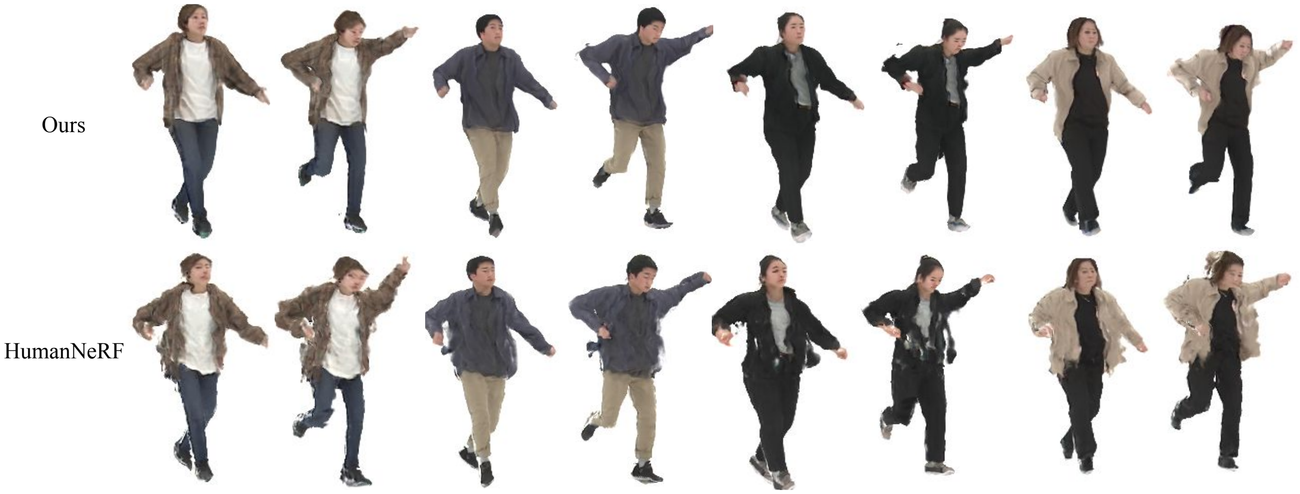

Figure 7: Qualitative comparison for 30-shot novel view synthesis of novel actors with unseen poses on the AIST++ dataset. Our method achieves high-quality animation with sharp boundary and details.

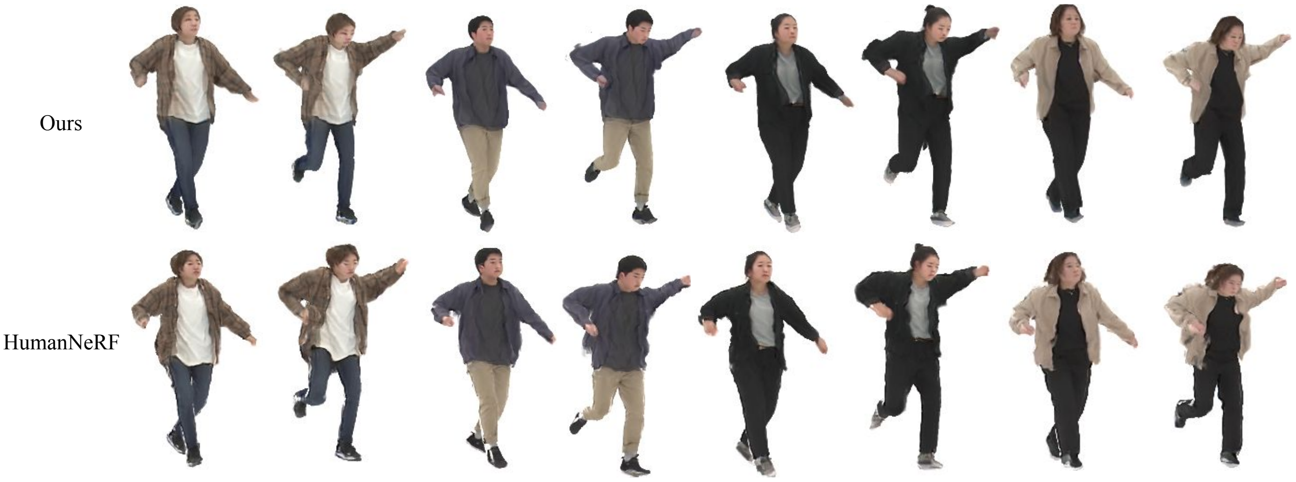

Figure 8: Qualitative comparison for 100-shot novel view synthesis of novel actors with unseen poses on the AIST++ dataset. Our method achieves high-quality animation with sharp boundary and details.