ActorsNeRF: Animatable Few-shot Human Rendering with Generalizable NeRFs

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

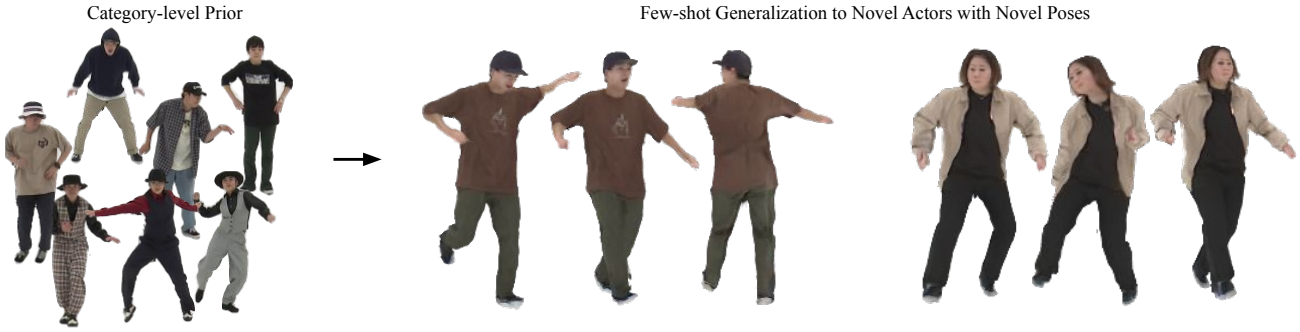

Figure 1: Animatable NeRF from a few images. We present ActorsNeRF, a category-level human actor NeRF model that generalizes to unseen actors in a few-shot setting. With only a few images, e.g., 30 frames, sampled from a monocular video, ActorsNeRF synthesizes high-quality novel views of novel actors in the AIST++ dataset with unseen poses (shown on the right).

Abstract

While NeRF-based human representations have shown impressive novel view synthesis results, most methods still rely on a large number of images / views for training. In this work, we propose a novel animatable NeRF called ActorsNeRF. It is first pre-trained on diverse human subjects, and then adapted with few-shot monocular video frames for a new actor with unseen poses. Building on previous generalizable NeRFs with parameter sharing using a ConvNet encoder, ActorsNeRF further adopts two human priors to capture the large human appearance, shape, and pose variations. Specifically, in the encoded feature space, we will first align different human subjects in a category-level canonical space, and then align the same human from different frames in an instance-level canonical space for rendering. We quantitatively and qualitatively demonstrate that ActorsNeRF significantly outperforms the existing state-of-the-art on few-shot generalization to new people and poses on multiple datasets. Project page: https://jitengmu.github.io/ActorsNeRF/.

1. Introduction

Recent advances in Neural Radiance Fields (NeRF) [42] have enabled significant progress in free-viewpoint rendering of humans performing complex movements. The possibility of achieving photo-realistic rendering is of major interest for various real-world applications in AR or VR. However, to achieve high-quality rendering, existing approaches [33, 23, 32, 29, 42] require a combination of synchronized multi-view videos and an instance-level NeRF network, trained on a specific human video sequence. While results are encouraging, the multi-view requirement is a significant challenge to applications involving videos in the wild. Recently, progress has been made to eliminate this constraint, by enabling human rendering from a monocular video [48, 15]. However, these approaches still require a large number of frames, which covers a person densely from all viewpoints.

In this work, we consider the more practical setting and ask the question: Can an animatable human model be learned from just a few images? We hypothesize that this is possible by introducing a class-level encoder, trained over multiple people, as shown in Figure 1. This hypothesis has been demonstrated by recent works on generalizable NeRFs [53, 2], where an encoder network is trained across multiple scenes or objects within the same category to construct NeRF. By parameter sharing through the encoder, the prior learned across different scenes can be re-used to perform synthesis even with a few views. However, most approaches can only model static scenes. We investigate how generalizable NeRF can be extended to the learning of a

good prior for the much more complex setting of videos of humans performing activities involving many degrees of freedom, large motions, and complex texture patterns.

For this, we introduce **ActorsNeRF**, a category-level human actor NeRF model that is transferable, to unseen humans in novel action poses, in a few-shot setup. This setup requires the animation of a previously unseen human actor into unseen views and poses, from a few frames of monocular video. Such generalization requires more than simple parameter sharing via an encoder network, and can benefit from the incorporation of explicit human priors. Our insight is that, while human actions and appearances are complex, all humans can be *coarsely* aligned in a *category-level canonical space* using a parametric model such as SMPL [24]. *Fine-grained* alignment can then benefit from an *instance-level canonical space* derived from both this prior and the few-shot data available for the target actor.

To implement this insight, we endow ActorsNeRF with a 2-level canonical space. Given a body pose and a rendering viewpoint, a sampled point in 3D space is first transformed into a canonical (T-pose) space by linear blend skinning (LBS) [24], where the skinning weights are generated by a skinning weight network that is shared across various subjects. Since LBS only models a coarse shape (similar to an SMPL mesh), we refer to this T-pose space as the *category-level canonical space.* Direct rendering from the latter fails to capture the shape and texture details that distinguish different people. To overcome this limitation, points in the category-level canonical space are further mapped into an *instance-level canonical space* by a deformation network. A rendering network finally maps the combination of pixel-aligned encoder features and points into corresponding colors and densities.

ActorsNeRF is designed such that the combination of feature encoder and skinning weight network forms a category-level shape and appearance prior, and the deformation network learns the mapping to the instance-level canonical space. To adapt to a novel human actor at test time, only the deformation network (instance-level) and rendering network are fine-tuned with the few-shot monocular images. The image encoder and skinning weight network are frozen.

We quantitatively and qualitatively demonstrate that ActorsNeRF outperforms the existing approaches by a large margin on various few-shot settings for both ZJU-MoCap Dataset [33] and AIST++ Dataset [21]. To the best of our knowledge, we are the first to explore few-shot generalization from few-shot monocular videos in the context of NeRF-based human representations.

2. Related Work

Dynamic NeRF. While NeRF [27] was originally proposed for modeling static scenes, recent efforts have successfully extended it to dynamic scenes [49, 22, 9, 30, 31] and deformable objects [46, 34, 17, 36, 51, 5, 8, 12]. One strategy to model dynamic objects and scenes is to align observations from various time steps in a canonical space, decoupling 4D as 3D and time reduces the complexity. For example, Pumarola et al. [34] propose a warping field to map sampled points to template space and then directly render from the canonical space. Our work shares the principle of aligning different observations for efficient modeling. Going beyond an instance canonical space, ActorsNeRF also incorporates the category-level human prior in the modeling of unseen actors with few images.

NeRF-based Human Rendering. Human-specific rendering is a longstanding challenge [16, 26, 52, 4, 7, 39, 11, 25, 1, 38] due to the large modeling space of shape, pose, and appearance. Recently, NeRF-based human representations have shown promise for high-quality view synthesis [33, 32, 50, 29, 19, 23, 42, 48, 55, 15, 10, 20, 45, 40, 41]. For example, Peng et al. [33] propose to attach a set of latent codes to SMPL [24] and render novel views of a performer from sparse multi-view videos. To better animate the human actor, subsequent works [23, 32] introduce a canonical space to align different body poses. These methods require multiview videos, which limits their application. To address this challenge, Weng et al. [48] further decompose shape and pose into skeletal motion and non-rigid shape deformation, and synthesize photorealistic details from just a monocular video. Jiang et al. [15] jointly learns a human NeRF and a scence NeRF from a monocular video. However, these methods are limited to instance-level and does not generalize to novel human actors. To achieve better generalization, category-level NeRFs [19, 55, 10] are introduced but these models require multi-view images for both training and inference. Different from all previous approaches, ActorsNeRF learns a category-level generalizable NeRF that allows for novel pose animation on unseen humans, only requiring few-shot images from a monocular video during inference.

Few-shot NeRF. NeRF generally suffers from suboptimal solutions when trained from a few views. To address this problem, various regularizations [14, 37, 6, 28, 3] have been proposed for the few-shot setting. For example, Jain et al. [14] leveraged the pre-trained CLIP [35] model and enforced semantic consistency in the feature space. Other works proposed to avoid the degenerate solutions by introducing a stronger geometry prior, using either supervised [37, 6] or unsupervised [28, 3] depth information. Instead of designing priors, a data-driven way to achieve few-shot transfer, is to train generalizable NeRFs on a large-scale dataset [44, 2, 53, 43]. Specifically, an encoder is trained to learn the data prior, and the model will be adapted on an unseen example by fine-tuning. We see ActorsNeRF as an integration of the data-driven approach with large-scale data and the regularization manner using

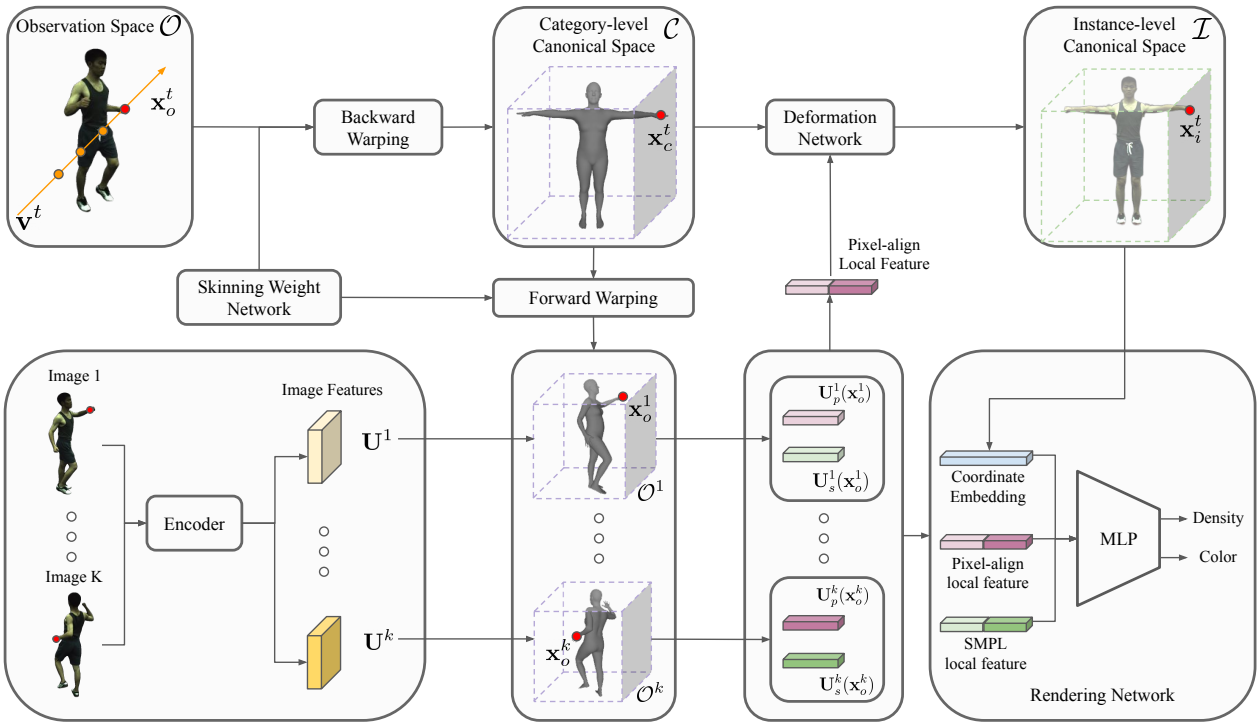

Figure 2: Overview of ActorsNeRF. First, K images are passed through an encoder to extract feature tensors $\mathbf{U}^k$. Given a target pixel location along with a view direction $\mathbf{v}^t$, a point $\mathbf{x}_o^t$ is sampled along the ray. The point is then mapped to the category-level canonical space $\mathbf{x}_c^t$ through a backward warping. Then, $\mathbf{x}_c^t$ is transformed to K other observation spaces through corresponding forward warpings and projected to images to query corresponding local features (pixel-aligned features $\mathbf{U}_p^k$ and SMPL local features $\mathbf{U}_s^k$). Next, a deformation network takes both $\mathbf{x}_c^t$ and its pixel-aligned features to produce a location $\mathbf{x}_i^t$ in the instance-level canonical space. $\mathbf{x}_i^t$ along with its pixel-aligned features are then mapped to color and density for volume rendering.

human-specific priors.

3. Method

We introduce ActorsNeRF, a category-level generalizable NeRF model capable of synthesizing unseen humans with novel body poses in a few-shot setup. In order to achieve generalization across different individuals, a category-level NeRF model is first trained on a diverse set of subjects. During the inference phase, we fine-tune the pre-trained category-level NeRF model using only a few images of the target actor, enabling the model to adapt to the specific characteristics of the actor.

Mathematically, the goal is to, given the small set of M frames $I = \{I^m\}$ capturing a human actor with random poses, learn the parameters θ of a network $\mathcal{Q}$ that maps a sampled $3D$ point $\mathbf{x} \in \mathbb{R}^3$ into a color vector $\mathbf{c} = (r, g, b)$ and a density σ, for any randomly sampled rendering viewpoint $\mathbf{v} \in \mathbb{R}^3$ and target pose $\mathbf{S} \in \mathbb{R}^{24\times3}$ (specified by 24 joints in this paper),

$$(\sigma, \mathbf{c}) = \mathcal{Q}(\mathbf{x}, \mathbf{v}, \mathbf{S}, I; \theta). \tag{1}$$

Prior works [48, 15] learns the model by overfitting to the observations. Though it works well when M is sufficiently large to densely cover a person from various viewpoints, in real world, there are many cases where only a few images are provided, e.g., $M = 5$. To address this problem, instead of relying on overfitting to the observed frames, we argue that a generalizable NeRF model that can quickly adapt with a few-shot setup is required. To implement the insight, we propose two key ideas: 1) leveraging a large-scale dataset of monocular videos to learn a category-level prior for the mapping such that the model can quickly adapt to the new person in a few-shot setting, while 2) incorporating human-specific knowledge to align people of diverse shapes and body poses, using a two-level canonical space.

3.1. ActorsNeRF

To learn the mapping of Eq.(1), at a high-level, we assume having access to a training set of multiple monocular videos capturing different people to learn a category-level prior. Then the model is finetuned on M frames to adapt to the new person.

For both category-level pre-training and finetuning, our idea is to first map a sampled point to a category-level

canonical space and then an instance-level canonical space sequentially, where the human body is represented in the canonical pose of Figure 2, and then rendered to color and density conditioned on the encoder features. Mathematically, we define the 3D space associated with a frame of the actor as *observation space* $\mathcal{O} \subset \mathbb{R}^3$, the corresponding category-level canonical space $\mathcal{C} \subset \mathbb{R}^3$, the corresponding instance-level canonical space $\mathcal{I} \subset \mathbb{R}^3$. We then define, a forward warping $\mathcal{T} : \mathcal{C} \rightarrow \mathcal{O}$ and a backward warping $\mathcal{T}^{-1} : \mathcal{O} \rightarrow \mathcal{C}$. In addition, K out of M images are randomly sampled, and an encoder $\mathcal{E}$ is used to obtain the corresponding feature tensor $\mathbf{U}^k$ (for extracting local features).

As shown in Figure 2, to render a target image, a pixel and target rendering viewpoint $\mathbf{v}^t$ define a ray of points $\mathbf{x}_o^t \in \mathcal{O}^t$. Each point is first mapped to a point $\mathbf{x}_c^t$ in the category-level canonical space. The transformation is through a *backward warping* $\mathcal{T}^{-1}$, guided by *skinning weight network* $\mathcal{W}$.

To adapt the canonical, category-level, shape to the shape details of different human actors, a *deformation network* $\mathcal{D}$ is implemented to transform $\mathbf{x}_c^t$ to a location $\mathbf{x}_i^t$ in the *instance-level canonical space* $\mathcal{I}$, where the warping is guided by image local features extracted from a set of feature tensors $\mathbf{U}^k$. To extract the pixel-aligned features, the category-level canonical space point $\mathbf{x}_c^t$ is first mapped by corresponding *forward warpings* $\mathcal{T}$, into observation spaces $\mathcal{O}^k$ $(k \neq t)$ corresponding to image I^k, and then projected to corresponding feature maps.

Finally, using $\mathbf{x}_i^t$ and the image local features, a NeRF-based rendering network $\mathcal{R}$ outputs its color and density. We next detail each component of ActorsNeRF, consisting of image encoder $\mathcal{E}$, skinning weight network $\mathcal{W}$, deformation network $\mathcal{D}$, and rendering network $\mathcal{R}$.

Feature Encoder $\mathcal{E}$. Prior works [53, 2, 19, 55] have shown that encoder features learned at a category-level greatly improve NeRF generalization. Inspired by the observation, we use an encoder $\mathcal{E}$, e.g., ResNet-18 [13], to extract a feature tensor $\mathbf{U}^k = \mathcal{E}(I^k)$ per image I^k. Let Π^k be the camera mapping associated with I^k and $\mathbf{U}^k(.)$ denote a value query operation from feature tensor $\mathbf{U}^k$ via interpolation. Given $\mathbf{U}^k$, pixel-aligned local features $\mathbf{U}_p^k$ and SMPL local features $\mathbf{U}_s^k$ are produced as follows.

Pixel-aligned local features $\mathbf{U}_p^k$ are obtained by extracting local features from $\mathbf{U}^k$ aligned with each pixel. A 3D point $\mathbf{x} \in \mathcal{O}^k$ has pixel-aligned local features $\mathbf{U}_p^k(\mathbf{x}) = \mathbf{U}^k(\Pi^k(\mathbf{x}))$.

SMPL local features $\mathbf{U}_s^k$ are pixel-aligned local features localized by the SMPL model, a parametric mesh model with 6890 vertices. Let $\mathbf{s}_j^k \in \mathbb{R}^3, j = \{1, \cdots, 6890\}$, be the vertices of SMPL fitted to I^k. Each vertex is assigned a local feature $\mathbf{U}_s^k(\mathbf{s}_j^k) = \mathbf{U}^k(\Pi^k(\mathbf{s}_j^k))$.

Skinning Weight Network $\mathcal{W}$. The skinning weight network generates the linear blend skinning weights for different individuals in the category-level canonical space, which enables the transformations between the category-level canonical space and observation spaces. The design of the network follows [48], given B ($B = 24$ in this paper) joints defined on the human body, the linear blend skinning weights defined in the category-level canonical space $\mathcal{C}$ are represented by a 3D volume $\mathbf{W}$. The difference is that, in ActorsNeRF, network parameters are shared across all actors in the training set to capture the category-level shape prior. An operator $\mathbf{W}(\cdot)$ is defined for a value query operation of the feature volume $\mathbf{W}$ via tri-linear interpolation.

Forward and Backward Transformation. Aggregating image features from the same actor under various body poses requires identifying correspondences between the matching body points $\mathbf{x}_o^k$ of different observation spaces $\mathcal{O}^k$. The alignment through the point in category-level canonical space $\mathcal{C}$ enables this, by introduction of forward $\mathcal{T} : \mathcal{C} \rightarrow \mathcal{O}$ and backward $\mathcal{T}^{-1} : \mathcal{O} \rightarrow \mathcal{C}$ warpings.

Given body pose $\mathbf{S}$, a transformation set $\mathbf{T}(\mathbf{S}) = \{\mathbf{T}_1, \cdots, \mathbf{T}_B\}$ is computed for each of the B joints, where each transformation matrix $\mathbf{T}_i \in SE(3)$ is defined with respect to the root joint. Location $\mathbf{x}_c \in \mathcal{C}$ is then mapped into point $\mathbf{x}_o \in \mathcal{O}$ by linear blend skinning (LBS) [24],

$$\mathbf{x}_o = \mathcal{T}(\mathbf{x}_c) = \left(\sum_{b=1}^{B} \mathbf{W}^b(\mathbf{x}_c)\mathbf{T}_b\right)\mathbf{x}_c \tag{2}$$

where $\mathbf{W}^b(\mathbf{x}_c)$ denotes the bth channel of the sampled blending weights at location $\mathbf{x}_c$.

Similarly, the backward mapping $\mathcal{T}^{-1}$ is defined as,

$$\mathbf{x}_c = \mathcal{T}^{-1}(\mathbf{x}_o) = \left(\sum_{b=1}^{B} \mathbf{W}_o^b(\mathbf{x}_o)\mathbf{T}_b^{-1}\right)\mathbf{x}_o \tag{3}$$

where $\mathbf{W}_o^b(\mathbf{x}_o)$ denotes the bth channel of the sampled observation-space blending weights for point $\mathbf{x}_o$, given by[48, 47]

$$\mathbf{W}_o^b(\mathbf{x}_o) = \frac{\mathbf{W}^b(\mathbf{T}_b^{-1}\mathbf{x}_o)}{\sum_{b=1}^{B} \mathbf{W}^b(\mathbf{T}_b^{-1}\mathbf{x}_o)} \tag{4}$$

As shown in Figure 2, given a point $\mathbf{x}_o^t$ in observation space $\mathcal{O}^t$, corresponding points $\mathbf{x}_o^k$ in support set observation spaces $\mathcal{O}^k$ can be established by first backward mapping $\mathbf{x}_o^t$ to the category-level canonical space point $\mathbf{x}_c^t$ and then forward mapping this point to the points $\mathbf{x}_o^k$ using corresponding body poses. These mappings are key to allow the query of pixel-aligned features from the sampled K images capturing different body poses, without requiring a multi-view imaging setup.

Deformation Network $\mathcal{D}$. To compensate for the missing details in the category-level canonical space, a deformation network $\mathcal{D}$, parameterized by $\theta_{\mathcal{D}}$, is implemented to

transform a point $\mathbf{x}_c$ to a point $\mathbf{x}_i$ in a fine-grained instance-level canonical space $\mathcal{I} \subset \mathbb{R}^3$. This deformation is conditioned on the target body pose $\mathbf{S}$ and K pixel-aligned local features $\mathbf{U}_p^k$,

$$\mathbf{x}_i = \mathcal{D}(\mathbf{x}_c, \{\mathbf{U}_p^k\}_{k=1}^K, \mathbf{S}; \theta_{\mathcal{D}}). \tag{5}$$

Rendering Network $\mathcal{R}$. A rendering network parameterized by $\theta_{\mathcal{R}}$ then predicts the color and density for an instance-level 3D location $\mathbf{x}_i$,

$$(\sigma, \mathbf{c}) = \mathcal{R}(\mathbf{x}_i, \{\mathbf{U}_p^k\}_{k=1}^K, \{\mathbf{U}_s^k\}_{k=1}^K; \theta_{\mathcal{R}}), \tag{6}$$

conditioned of the sets of pixel-aligned local features $\{\mathbf{U}_p^k\}_{k=1}^K$ and SMPL local features $\{\mathbf{U}_s^k\}_{k=1}^K$. Note that, since features $\mathbf{U}_s^k$ are only defined on SMPL vertices, features can not be directly queried for sampled points in a continuous space. Therefore, we expand SMPL features to continuous space as follows: K SMPL local features $\mathbf{U}_s^k$ are first concatenated and then passed through a sparse 3D convolution network to generate a 3D volume, such that features at any sampled location (e.g., $\mathbf{x}_i$) can be obtained through tri-linear interpolation. Similar ideas were used by [33, 19]. Different from the prior works where the sparse convolution is implemented in the observation space, the SMPL feature diffusion process is implemented in the canonical space and serves as a category-level prior.

Volume Rendering. As in NeRF [27], the expected rendering color $\mathbf{C}(\mathbf{r})$ along camera ray $\mathbf{r}$ is obtained by aggregating the predicted color $\mathbf{c}$ and density σ generated by (1) using standard volume rendering.

3.2. Category-level Training

During training, given N monocular videos sequences, in each iteration, ActorsNeRF randomly samples K frames from a monocular video and renders another sampled target image I^t in the same video sequence. In practice, we use $K = 3$ to achieve a good balance between computation complexity and feature quality. The image encoder $\mathcal{E}$ is used to extract corresponding features $\mathbf{U} = \{\mathbf{U}^1, \mathbf{U}^2, \cdots, \mathbf{U}^K\}$ for the K images. With these extracted image features, shown in Figure 2, a ray of sampled points for a target rendering viewpoint $\mathbf{v}^t$ is aggregated to producing the corresponding color $\mathbf{C}(\mathbf{r})$. To ensure high-quality rendering, both the mean square error $\mathcal{L}_{mse}$ and the perceptual loss [54] $\mathcal{L}_{LPIPS}$ are used as objective functions. Additionally, another skinning weight regularizer $\mathcal{L}_W$, which is an $\mathcal{L}_1$ loss, is employed to encourage the output skinning weights from the skinning weight network to be close to the prior obtained from the skeleton. All three objective functions are jointly optimized to update all network parameters,

$$\mathcal{L} = \lambda_{mse}\mathcal{L}_{mse} + \lambda_{LPIPS}\mathcal{L}_{LPIPS} + \lambda_W \mathcal{L}_W \tag{7}$$

where λ_{mse}, λ_{LPIPS}, and λ_W are corresponding coefficients to balance different loss functions.

3.3. Few-shot Optimization

To transfer the knowledge for ActorsNerf learned at a category level to a novel human actor with M frames provided, we propose to fine-tune the model to match the observations. During fine-tuning, we select K out of M frames for the encoder to extract features, such that all M images are synthesized with these K image features. Note that M here can be much smaller compared to the category-level pretraining stage. In addition, different from the pretraining stage, where different K frames are used for each iteration, the K frames are fixed in the few-shot optimization stage for a stable performance in the few-shot setting. As the combination of the feature encoder and skinning weight network forms a category-level shape and appearance prior, only the deformation network and rendering network are fine-tuned, and the encoder and skinning weight network are frozen. Additionaly, only the mean square error and perceptual loss are used in the few-shot optimization stage. After fine-tuning, ActorsNeRF is capable of rendering the novel actor with novel viewpoints and poses using the K frames.

4. Experiments

We test ActorsNeRF on multiple benchmark datasets, e.g., ZJU-MoCap dataset [33] and AIST++ dataset [21], and ActorsNeRF significantly outperforms multiple representative state-of-the-art baselines.

4.1. Datasets and Baselines

Dataset. We test ActorsNeRF on two datasets: ZJU-MoCap [33] dataset and AIST++ dataset [21]. 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. Follow [19], We leave 3 (387, 393, 394) subjects as held-out data and use the remaining 7 for training. The AIST++ dataset is a dancing motion dataset capturing 30 human subjects performing various dances from 9 multi-view cameras. We randomly select one action sequence for each subject and then split the dataset with 25 actors for training and the other 5 actors (16-20) for testing. For both datasets, 'camera1' is used for training and other views are only used for evaluation. More details are discussed in the supplementary materials.

Baseline and Metric. We compare our method with the most representative state-of-the-art view synthesis methods. HumanNeRF [48] (HN) aligns various poses in a canonical space and achieves state-of-the-art rendering performance from a monocular video. NeuralBody [33] is a representative method for rendering from observation space. Neural Human Performer [19] and MPS-NeRF [10] require multi-view images for both training and inference so they are

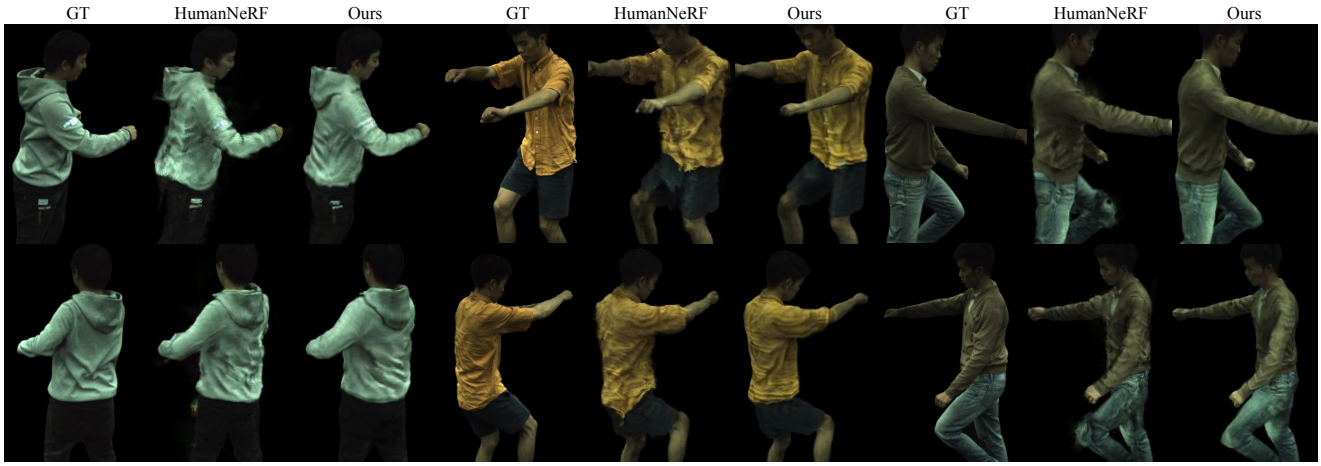

Figure 3: Qualitative comparison for few-shot novel view synthesis of novel actors with unseen poses on the ZJU-MoCap dataset. Ground-truth (GT), HumanNeRF, and our results are shown from left to right. Top and bottom rows are rendered from different viewpoints. Our method renders high-quality images with sharp boundaries and details.

		Person 387			Person 393			Person 394		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
5-shot	NeuralBody	26.19	0.9404	85.91	27.54	0.9476	84.15	27.52	0.9453	82.20
	HumanNeRF	26.28	0.9465	65.64	26.84	0.9466	63.53	27.53	0.9451	64.15
	Ours	**27.26**	**0.9568**	**46.06**	**27.20**	**0.9553**	**46.29**	**28.09**	**0.9577**	**42.48**
10-shot	NeuralBody	**27.18**	0.9494	76.32	27.17	0.9469	78.73	28.14	0.9492	77.23
	HumanNeRF	26.66	0.9501	59.38	26.96	0.9516	56.29	28.51	0.9515	52.62
	Ours	27.15	**0.9592**	**40.89**	**27.26**	**0.9565**	**42.56**	**28.71**	**0.9613**	**35.93**
30-shot	NeuralBody	26.32	0.9457	65.43	27.08	0.9482	72.85	28.10	0.9530	69.55
	HumanNeRF	27.25	0.9555	47.59	27.37	0.9558	46.04	28.38	0.9559	43.76
	Ours	**27.67**	**0.9610**	**36.76**	**27.59**	**0.9577**	**39.51**	**28.97**	**0.9614**	**34.29**
100-shot	NeuralBody	26.91	0.9513	60.60	27.13	0.9515	67.11	27.85	0.9544	59.42
	HumanNeRF	27.13	0.9589	40.00	27.20	0.9550	42.91	28.25	0.9568	40.45
	Ours	**27.66**	**0.9614**	**36.39**	**27.57**	**0.9580**	**39.33**	**29.07**	**0.9612**	**34.03**
300-shot	NeuralBody	26.95	0.9518	59.89	27.22	0.9535	64.60	27.69	0.9557	56.17
	HumanNeRF	27.30	0.9584	40.72	27.24	0.9557	43.59	28.46	0.9578	39.44
	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.

not directly comparable to our monocular setting. Following [48], we use three metrics for quantitative evaluation: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and perceptual quality (LPIPS) [54] (reported by $\times 10^3$). More comparisons and details are included in the supplementary materials.

Training Details. We implement patch-based rendering with 128 samples per ray. The category-level model is trained with $200K$ iterations and the fine-tuning takes 50k iterations. During category-level training, $\lambda_{mse} = 0.2$, $\lambda_{LPIPS} = 1$, $\lambda_W = 0.01$. The deformation network is not optimized until 10000 iterations. The Adam optimizer is implemented with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is set to 5×10^{-5} for the skinning weight network and the deformation network, and 5×10^{-4} for other modules. Positional embedding is applied to sampled points before input to each module.

4.2. Generalization

In this section, we analyze how ActorsNeRF generalizes in a few-shot setup. We demonstrate the generalization of the learned category-level prior in two settings: few-shot generalization in Section 4.2.1 and short-video generalization in Section 4.2.2. The few-shot generalization setting samples images where the human actor is mostly observed (e.g., both front and back). In contrast, in the short-video

Figure 4: Qualitative comparison for few-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. In contrast, HumanNeRF outputs blurry images with broken body parts.

		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	HN	24.37	0.9752	29.59	24.86	0.9762	29.39	22.77	0.9738	33.02	24.51	0.9759	28.68	24.55	0.9791	27.63
	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**
10-shot	HN	24.22	0.9737	31.20	24.84	0.9758	28.81	23.96	0.9763	28.15	24.32	0.9758	28.86	24.94	0.9803	25.17
	ours	**25.25**	**0.9794**	**22.22**	**25.87**	**0.9812**	**22.45**	**24.71**	**0.9822**	**21.26**	**25.33**	**0.9809**	**21.55**	**25.53**	**0.9833**	**20.71**
30-shot	HN	25.08	0.9773	25.46	25.62	0.9793	24.06	24.53	0.9800	23.35	25.34	0.9793	23.77	25.41	0.9822	21.76
	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**
100-shot	HN	25.31	0.9783	21.87	25.81	0.9801	21.74	24.59	0.9811	20.84	25.69	0.9809	20.57	25.58	0.9837	18.29
	ours	**25.78**	**0.9812**	**19.05**	**26.14**	**0.9833**	**18.47**	**25.02**	**0.9834**	**18.64**	**25.72**	**0.9827**	**18.84**	**25.86**	**0.9846**	**17.69**
300-shot	HN	25.65	0.9795	21.54	26.12	0.9817	21.00	24.96	0.9824	19.98	25.79	0.9816	20.54	**26.00**	0.9840	17.92
	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.

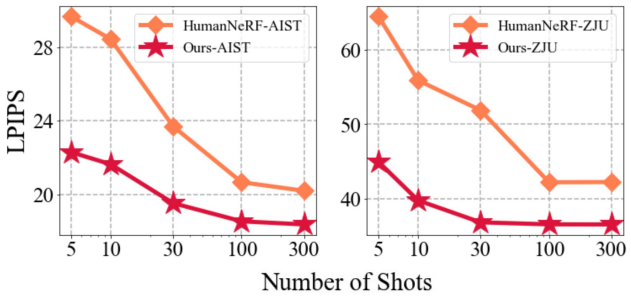

Figure 5: Comparison for few-shot novel view synthesis of novel actors with unseen poses on ZJU-MoCap dataset (right) and AIST++ dataset (left). LPIPS scores are obtained by averaging across all test subjects in a dataset.

generalization setting, the input frames are selected such that a subject is not densely covered. This experiment is designed in a way that the model needs to 'imagine' the missing portions.

4.2.1 Few-shot Generalization

In this section, we vary the number of input images to test the novel view synthesis of novel actors with unseen poses. 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.

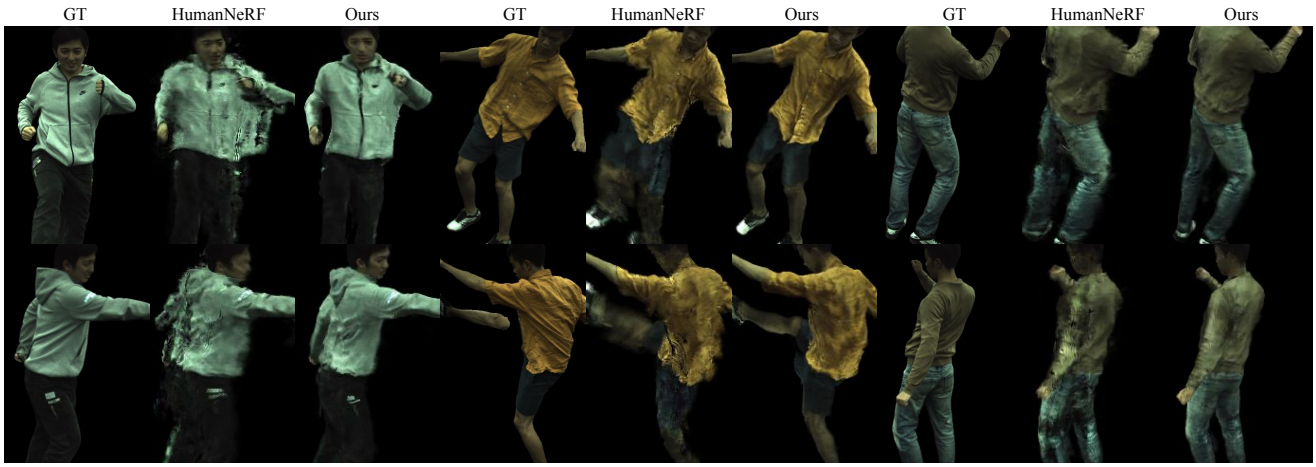

Figure 6: Qualitative comparison for short-video novel view synthesis of novel actors with unseen poses on the ZJU-MoCap dataset. Each row presents a test actor with different poses. Ground-truth (GT), HumanNeRF, and our results are shown from left to right. ActorsNeRF generates sharp boundaries and maintains a better overall shape, indicating that the category level prior can be leveraged to smoothly synthesizing unobserved portions.

	Person 387			Person 393			Person 394		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeuralBody	25.71	0.9414	79.53	26.66	0.9422	82.15	27.60	0.9468	76.65
HumanNeRF	24.84	0.9352	81.82	25.76	0.9408	70.81	27.03	0.9460	65.82
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 view synthesis of novel actors with unseen poses on the ZJU-MoCap dataset.

As shown in Table 1, on the ZJU-MoCap dataset, ActorsNeRF outperforms all baseline methods by a large margin across all shots, especially on the LPIPS metric. Figure 3 presents the 30-shot rendering with unseen poses of test actors in the ZJU-MoCap dataset from 2 different viewpoints. The rendering produced by ActorsNeRF is smoother and preserves high-quality details, while maintaining a valid shape with less body distortion. In comparison, HumanNeRF produces unsmooth, blurry textures with distorted boundaries in the rendered images.

To further demonstrate the effectiveness of the proposed method, we test our algorithm on the more challenging AIST++ dataset [21], which contains diverse subjects with complex dancing poses and loose clothes. We employ PointRend [18] to obtain foreground masks automatically. This introduces additional challenges as the mask boundaries tend to be noisy. The quantitative results are shown in Table 2, where our method achieves a much better performance. Figure 4 visualizes how, on this dataset, HumanNeRF frequently fails to produce a valid shape (distorted bodies and broken body parts), whereas ActorsNeRF achieves a much better synthesis of the overall shapes and textures. This again demonstrates the effectiveness of the learned category-level prior.

We additionally plot the LPIPS metric, averaged on all test subjects, for different shots in Figure 5. Our insight is that ActorsNeRF gains more from the learned category-level prior when inputting fewer images. However, noting that even when 300 frames are provided, ActorsNeRF still yields significant margin over the baseline. The results suggests that the category-level prior of ActorsNeRF improves the rendering quality over a large few-shot spectrum. More results and visualizations for all different shots are shown in the supplementary materials.

4.2.2 Short-video Generalization

In this section, we study whether our algorithm is capable of imagining missing portions of the body given only a handful of images of a new person, with the goal of rendering novel poses and viewpoints. Specifically, we consider a scenario where only a portion of the person is observed in all frames, which is a natural task for humans who can imagine the missing parts of the body. We examine whether our algorithm can perform similarly under these conditions, by sampling $m = 10$ frames from 100 consecutive frames.

Table 3 shows that, on the ZJU-MoCap dataset, ActorsNeRF again outperforms HumanNeRF (HN) approach significantly. Figure 6 provides a clear visual comparison

	Person 387			Person 393			Person 394		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours full model	27.15	0.9592	40.89	27.26	0.9565	42.56	29.11	0.9613	35.93
w/out instance-level canonical space	27.33	0.9587	41.12	27.32	0.9562	44.13	28.73	0.9591	36.61
w/out pixel-align local features	27.06	0.9585	40.99	27.42	0.9569	42.77	28.96	0.9602	36.31
w/out SMPL local features	27.32	0.9592	41.35	27.32	0.9567	43.34	28.89	0.9608	37.13

Table 4: Ablation on components of ActorsNeRF. The full model achieves the overall best performance.

between different methods. HumanNeRF produces some details for the observed body portions, but fails to maintain a valid shape for the less unobserved body parts. In comparison, ActorsNeRF still generates sharp boundaries and keeps the overall shape, suggesting that the ActorsNeRF is capable of leveraging the category level prior to smoothly synthesize unobserved portions of the body.

4.3. Ablation

In this section, we investigate the efficacy of the two-level canonical space and local encoder features of ActorsNeRF in the context of the 10-shot case of the few-shot generalization setting on the ZJU-MoCap dataset as described in Section 4.2.1. Specifically, we perform ablation experiments by excluding a component in both category-level training and instance-level fine-tuning. Our quantitative results are presented in Table 4, and the qualitative results are provided in Figure 7.

As discussed in Section 3.1, the deformation network is responsible for refining category-level shape into a fine-grained instance-level shape. We ablate the deformation network by directly rendering from the category-level canonical space, i.e., the rendering network directly takes as input the point in the category-level canonical space as input. The figure demonstrates that the absence of instance-level canonical space in the model leads to distorted facial details in the rendered images. This finding highlights the importance of the proposed two-level canonical space design for producing photorealistic details.

ActorsNeRF queries pixel-aligned local features from multiple observed images using category-level points that are transformed via corresponding forward mappings. When rendering without these pixel-aligned local features, the output images appear more distorted, with less smooth boundaries. Similarly, omitting the SMPL surface aggregated features results in worse rendering quality, including distorted facial details and noisy boundaries. These observations emphasize the importance of the SMPL aggregated local features as a strong category-level prior for smooth rendering. Quantitatively, we observe that removing either pixel-aligned local features or smpl local features results in lower LPIPS scores, indicating that all components are essential for achieving the best overall performance.

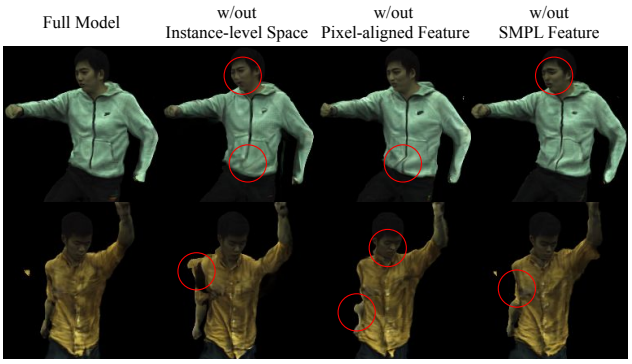

Figure 7: Ablation on components of ActorsNeRF. ActorsNeRF full model, ActorsNeRF without the instance-level canonical space, ActorsNeRF without the pixel-aligned local features, ActorsNeRF without the SMPL local features are shown from left to right. The full model achieves the overall best performance whereas other models show various distortions.

5. Conclusion

We introduce ActorsNeRF, a generalizable NeRF-based human representation, that is trained from monocular videos and adapted to novel human subjects with a few monocular images. ActorsNeRF encodes the category-level prior through parameter sharing on multiple human subjects, and is implemented with a 2-level canonical space to capture the large human appearance, shape, and pose variations. ActorsNeRF is tested on multiple benchmark datasets, e.g., ZJU-MoCap dataset and AIST++ dataset. Compared to existing state-of-the-art methods, ActorsNeRF demonstrates superior novel view synthesis performance for novel human actors with unseen poses across multiple few-shot settings.

Acknowledgements. This work was supported, in part, by grants from NSF IIS-2041009 and gifts from Amazon. Prof. Wang's lab was supported, in part, by Amazon Research Award, Sony Research Award, Adobe Data Science Research Award, and gifts from Qualcomm.

References

[1] Dan Casas, Marco Volino, John P. Collomosse, and Adrian Hilton. 4d video textures for interactive character appearance. *Comput. Graph. Forum*, 33(2):371–380, 2014. 2

[2] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, pages 14104–14113, 2021. 1, 2, 4

[3] Di Chen, Yu Liu, Lianghua Huang, Bin Wang, and Pan Pan. Geoaug: Data augmentation for few-shot nerf with geometry constraints. In *ECCV*, pages 322–337, 2022. 2

[4] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam G. Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4):69:1–69:13, 2015. 2

[5] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. LISA: learning implicit shape and appearance of hands. In *CVPR*, pages 20501–20511, 2022. 2

[6] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, pages 12872–12881, 2022. 2

[7] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip L. Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts-Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. Fusion4d: real-time performance capture of challenging scenes. *ACM Trans. Graph.*, 35(4):114:1–114:13, 2016. 2

[8] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. *arXiv preprint arXiv:2205.15285*, 2022. 2

[9] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, pages 5692–5701, 2021. 2

[10] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *arXiv preprint arXiv:2203.16875*, 2022. 2, 5

[11] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor S. Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *CVPR*, pages 12135–12144, 2019. 2

[12] Xiang Guo, Guanying Chen, Yuchao Dai, Xiaoqing Ye, Jiadai Sun, Xiao Tan, and Errui Ding. Neural deformable voxel grid for fast optimization of dynamic view synthesis. *arXiv preprint arXiv:2206.07698*, 2022. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[14] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, pages 5865–5874, 2021. 2

[15] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 1, 2, 3

[16] Takeo Kanade, Peter Rander, and P. J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multim.*, 4(1):34–47, 1997. 2

[17] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzcinski, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In *CVPR*, pages 18602–18611, 2022. 2

[18] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9796–9805, 2020. 8

[19] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *NeurIPS*, pages 24741–24752, 2021. 2, 4, 5

[20] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. TAVA: template-free animatable volumetric actors. *arXiv preprint arXiv:2206.08929*, 2022. 2

[21] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3d dance generation with AIST++. In *ICCV*, pages 13381–13392, 2021. 2, 5, 8

[22] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, pages 6498–6508, 2021. 2

[23] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.*, 40(6):219:1–219:16, 2021. 1, 2

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 2, 4

[25] Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien P. C. Valentin, Sameh Khamis, Philip L. Davidson, Anastasia Tkach, Peter Lincoln, Adarsh Kowdle, Christoph Rhemann, Dan B. Goldman, Cem Keskin, Steven M. Seitz, Shahram Izadi, and Sean Ryan Fanello. *LookinGood*: enhancing performance capture with real-time neural re-rendering. *ACM Trans. Graph.*, 37(6):255, 2018. 2

[26] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. Image-based visual hulls. In Judith R. Brown and Kurt Akeley, editors, *SIGGRAPH*, pages 369–374. ACM, 2000. 2

[27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 5

[28] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, pages 5470–5480, 2022. 2

[29] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, pages 5742–5752, 2021. 1, 2

[30] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, pages 5845–5854, 2021. 2

[31] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6):238:1–238:12, 2021. 2

[32] 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 *ICCV*, pages 14294–14303, 2021. 1, 2

[33] 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 *CVPR*, pages 9054–9063, 2021. 1, 2, 5

[34] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, pages 10318–10327, 2021. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2

[36] Amit Raj, Michael Zollhöfer, Tomas Simon, Jason M. Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. PVA: pixel-aligned volumetric avatars. pages 11733–11742, 2021. 2

[37] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*, pages 12882–12891, 2022. 2

[38] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 2

[39] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. *arXiv preprint arXiv:2101.04104*, 2021. 2

[40] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yanpei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *CVPR*, pages 15851–15861, 2022. 2

[41] Shih-Yang Su, Timur M. Bagautdinov, and Helge Rhodin. DANBO: disentangled articulated neural body representations via graph neural networks. *arXiv preprint arXiv:2205.01666*, 2022. 2

[42] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, pages 12278–12291, 2021. 1, 2

[43] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *CVPR*, pages 2846–2855, 2021. 2

[44] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, pages 4690–4699, 2021. 2

[45] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. ARAH: animatable volume rendering of articulated human sdfs. *arXiv preprint arXiv:2210.10036*, 2022. 2

[46] Fangyin Wei, Rohan Chabra, Lingni Ma, Christoph Lassner, Michael Zollhöfer, Szymon Rusinkiewicz, Chris Sweeney, Richard A. Newcombe, and Mira Slavcheva. Self-supervised neural articulated shape and appearance models. In *CVPR*, pages 15795–15805, 2022. 2

[47] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv preprint arXiv:2012.12884*, 2020. 4

[48] 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 *CVPR*, pages 16189–16199, 2022. 1, 2, 3, 4, 5, 6

[49] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *CVPR*, pages 9421–9431, 2021. 2

[50] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *NeurIPS*, pages 14955–14966, 2021. 2

[51] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, pages 2853–2863, 2022. 2

[52] Genzhi Ye, Yebin Liu, Yue Deng, Nils Hasler, Xiangyang Ji, Qionghai Dai, and Christian Theobalt. Free-viewpoint video of human actors using multiple handheld kinects. *IEEE Trans. Cybern.*, 43(5):1370–1382, 2013. 2

[53] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 1, 2, 4

[54] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5, 6

[55] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *CVPR*, pages 7743–7753, 2022. 2, 4