

You Only Train Once: Multi-Identity Free-Viewpoint Neural Human Rendering from Monocular Videos

Jaehyeok Kim¹ Dongyoon Wee² Dan Xu¹

¹The Hong Kong University of Science and Technology ²Clova AI, NAVER Corp.

jkimbf@connect.ust.hk, dongyoon.wee@navercorp.com, danxu@cse.ust.hk

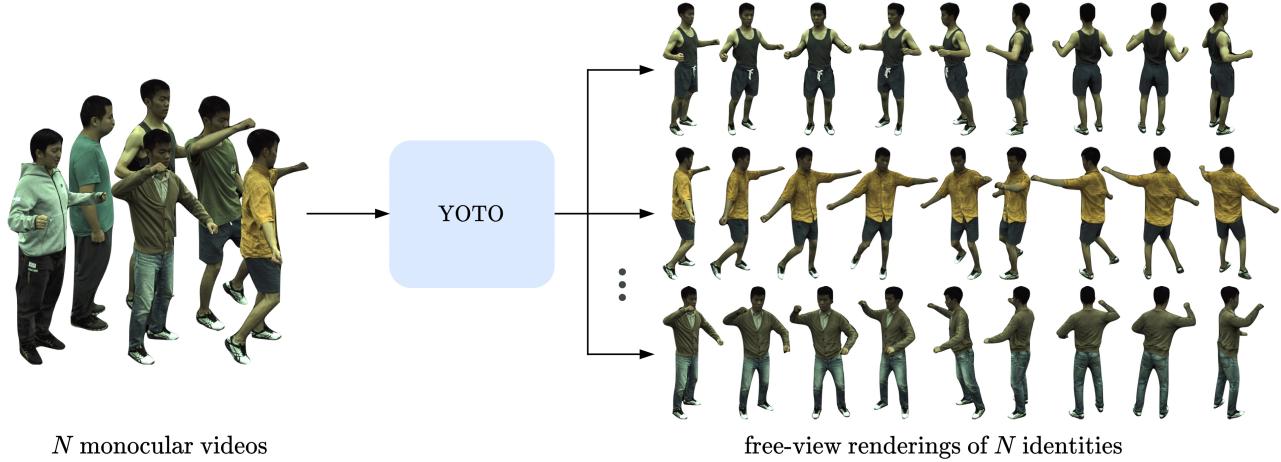


Figure 1: The proposed YOTO is a dynamic human generation framework, which can simultaneously train and model multiple people with distinct appearances and motions, while also allowing coherent rendering of them in the unified framework in high fidelity. The identity animation can be performed by using either seen poses or novel poses.

Abstract

We introduce *You Only Train Once* (YOTO), a dynamic human generation framework, which performs free-viewpoint rendering of different human identities with distinct motions, via only one-time training from monocular videos. Most prior works for the task require individualized optimization for each input video that contains a distinct human identity, leading to a significant amount of time and resources for the deployment, thereby impeding the scalability and the overall application potential of the system. In this paper, we tackle this problem by proposing a set of learnable identity codes to expand the capability of the framework for multi-identity free-viewpoint rendering, and an effective pose-conditioned code query mechanism to finely model the pose-dependent non-rigid motions. YOTO optimizes neural radiance fields (NeRF) by utilizing designed identity codes to condition the model for learning various canonical T-pose appearances in a single shared volumetric representation. Besides, our joint learning of multiple identities within a unified model incidentally enables flexible motion transfer in high-quality photo-realistic render-

ings for all learned appearances. This capability expands its potential use in important applications, including Virtual Reality. We present extensive experimental results on ZJU-MoCap and PeopleSnapshot to clearly demonstrate the effectiveness of our proposed model. YOTO shows state-of-the-art performance on all evaluation metrics while showing significant benefits in training and inference efficiency as well as rendering quality. The code and model will be made publicly available soon.

1. Introduction

Novel view synthesis of a person with dynamic motions from a monocular video is an especially challenging and long-standing problem. Unlike other similar tasks dealing with dynamic scenes, it requires modeling not only complicated motions generated by body joints but also non-rigidities of finer-granularity components such as the body and clothes. Moreover, the monocular setting further complicates the problem as information about every body motion from a single-image view is extremely limited. Therefore, the free-viewpoint rendering of moving people has

mostly been investigated under multi-view settings without taking into account non-rigid fine-grained motions.

Recently, Weng *et al.* [35] successfully address the above-mentioned problem with an architecture comprising structures for learning of a motion field and a NeRF [19]. Although they address the monocular free-viewpoint rendering with state-of-the-art performance, its identity-specific nature largely limits its potential for practical application scenarios. It requires an independent model that is trained from scratch for every human identity from each monocular video. This constraint severely downgrades the efficiency and generalization performance of the method. For instance, it is clearly not scalable if there is a considerable number of human-specific videos to be learned, as the overall model size and the training time would significantly increase, proportional to the number of identities following their pipeline. Besides, it is obviously not flexible to perform any interactions between the different identities (*e.g.*, performing motion transfer among the identities in different videos), due to the fact that there are no correlations constructed during training and inference among the different human models. We believe that each video comprises generic and distinctive information and the generic one could be learned collaboratively from different videos by having an unified NeRF framework, thus benefiting the rendering performance as well.

To target the issues above-discussed, in this paper, we propose YOTO, a more versatile framework for the monocular free-viewpoint rendering of people with distinct motions. We introduce an effective set of learnable identity codes into the framework to enable learning global human-specific representations, which can be utilized in our framework as a perfect switcher to allow multi-subject modeling and renderings, using a single unified model by only *one-time* optimization. Furthermore, in this paper, we present a novel mechanism for querying a separate identity code to learn identity-specific non-rigid motions. This involves utilizing cross-attention between the identity code and the 3D pose of the current frame. By adopting this process, our framework is able to extract discriminative features that are tailored to each identity with a particular pose, resulting in better non-rigid motion estimation. YOTO conditions two different NeRFs, each for non-rigid motions and appearances learned in a unified manner. Unlike Weng *et al.* [35], it incorporates all subjects in interest at the same time for training while not requiring proportionally longer training time. This significantly improves the efficiency of the framework on a number of people and further enhances both qualitative and quantitative performance. Overall, YOTO achieves state-of-the-art performance on free-viewpoint rendering of multiple moving people while showing remarkable enhancements in flexibility and training/inference efficiency compared to [35]. We

present our experimental results on ZJU-MoCap [24] and PeopleSnapshot [1] to demonstrate that YOTO can competently handle hard cases (*e.g.* input videos in the wild) and achieve state-of-the-art performances.

In summary, the contribution of our work is threefold:

- We resolve the issue of subject-specific training by proposing a new framework with learnable identity codes that allows multi-human-identity representation learning.
- The proposed framework, YOTO, better models pose-dependent non-rigid motions by conditioning the non-rigid modeling with a pose-conditioned code queried by cross-attention.
- YOTO not only achieves state-of-the-art performance on all quantitative metrics but also remarkably improves the model efficiency. Incidentally, YOTO can animate all the learned appearances of different identities in high fidelity with any novel poses, thereby enabling high-quality animations for various applications.

2. Related Work

We review closely related works from three perspectives, *i.e.*, deformable neural rendering, neural rendering of humans, and monocular neural human rendering.

Deformable neural rendering. NeRF [19] leverages a multi-layer perceptron (MLP) to learn a static 3D representation of a scene from a dense set of images from diverse viewpoints. Among various research directions, recent works have enhanced NeRF in terms of its efficiency and performance. For instance, several works [5, 27, 38, 6, 28, 20, 2] boost the efficiency of NeRF in training or inference stages to allow more practical usage. Some others [39, 9, 3, 32, 16] extend NeRF to handle the sparse view setting, which adapts NeRF to more realistic training scenarios and expands its practicality. However, it has been observed that they are restricted to static scenes while the majority of the real-world objects are dynamic.

Recent works including [4, 14, 26, 22, 31, 23, 37] have broadened the modeling capabilities of NeRF to dynamic scenes containing movements or deformations. Park *et al.* [22, 23] handle natural deformations on faces by introducing an MLP to estimate per-point deformations and [23] additionally models topological changes by proposing a canonical hyper-space. Pumarola *et al.* [26] and Tretschk *et al.* [31] propose the same idea as [22] to model dynamic objects and non-rigidity, respectively. It should be noted that these motions and deformations are small and simple; otherwise, the aforementioned approaches would not exhibit the anticipated level of performance. Nonetheless, there are deformations in the world spanning from small-scale motions to more complex articulated motions. In contrast, our approach enables learning of human-articulated motions from monocular videos, and performing free-viewpoint rendering of a human performer at any time frame of the videos.

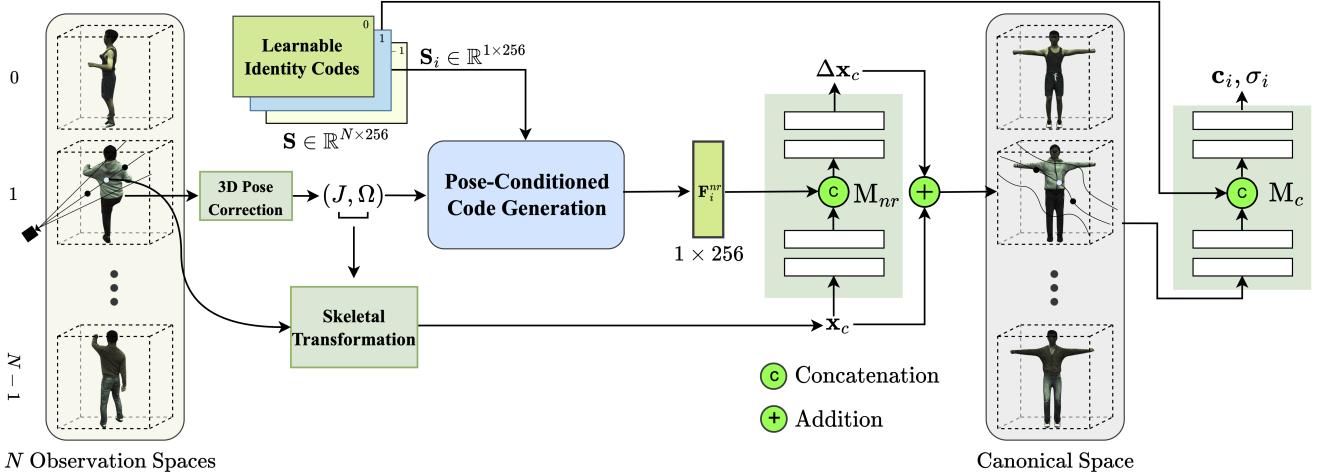


Figure 2: Overview of the proposed framework YOTO for free-view human rendering from monocular videos. Our framework is able to simultaneously train multiple identities while also achieving state-of-the-art performances on rendering quality. This is achieved by the proposed learnable identity codes and the body-pose-conditioned code generation module for subject-specific non-rigid motion estimation and canonical radiance field prediction.

Neural rendering of humans. As our approach targets the problem of neural human rendering, we discuss related works in this direction. Martin-Brualla *et al.* [18] propose a neural re-rendering approach via U-Net-like architecture for reducing the generated artifacts. By utilizing a few calibration images of the target subject, Pandey *et al.* [21] introduce semi-parametric learning from a single or few input RGBD frames. Similarly, Liu *et al.* [15] propose to use a character model to generate priors for learning time-coherent dynamic textures. Wu *et al.* [36] learn explicit 3D features on point clouds produced from multi-view stereo [29] and use U-Net for free-viewpoint rendering. To enable learning implicit representations from a highly sparse set of input views, structured latent codes are introduced by Peng *et al.* [25] to be applied on a shared deformable mesh (SMPL [17]). There are also several prior works exploring learning animatable avatars [12, 10, 33, 7], which however are based on explicit human parametric models [17], instead of implicit representations. Besides, Jiang *et al.* [11] adopt Instant-NGP [20] to boost the training efficiency. While most of these existing works either use the explicit SMPL as a prior or require multi-view videos, our approach does not rely on the parametric models and utilizes only monocular videos and 3D poses as inputs.

Monocular neural human rendering. Recently, Weng *et al.* [35] propose an approach to conduct free-viewpoint rendering of a human performer within a monocular video. It learns a canonical T-pose representation of the performer by modeling rigid body motions and pose-dependent non-rigid motions. However, it has a severe limitation of subject-specific modeling, which requires a new model to be trained from scratch for *each* input monocular video. In this paper, we introduce a set of novel learnable identity codes and an

effective pose-conditioned code query mechanism, which allows our single model to represent an arbitrary number of human subjects while outperforming the baseline [35] on all evaluation metrics.

3. The Proposed YOTO Approach

3.1. Framework Overview

Given a number of monocular videos each containing a single distinct human subject, the proposed YOTO framework learns discriminative representations of all the moving identities by one-time training, for free-viewpoint rendering, as shown in the framework overview (see Fig. 2). Specifically, we propose a novel idea to enable a simultaneous optimization of a collaborative canonical representation of all the subjects, via introducing a set of *learnable* identity codes. To condition the learning of identity-specific non-rigid motions, we further propose a module to generate pose-conditioned identity codes.

For each identity in a monocular video, the module utilizes a learnable identity code and joint poses corresponding to the target subject as input, and produces a subject-specific pose-conditioned code by a cross-attention mechanism. The generated code is then embedded into an MLP to condition the non-rigid motion estimation. It accepts an input point from a skeletal transformation [35] that maps input deformed joint poses to a canonical T-pose based on subject-specific blend weights with inverse linear-blend skinning. Then, for the transformed canonical points with non-rigid motions applied, we further enable multi-identity rendering by conditioning a canonical MLP with the proposed learnable identity codes, to versatiliely predicts the radiance and density that correspond to each different target subject.

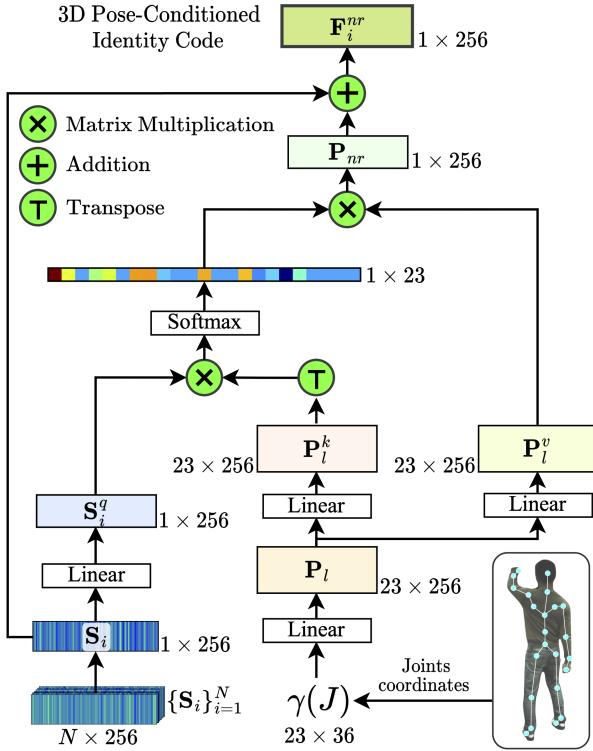


Figure 3: Illustration of the proposed pose-conditioned identity code generation. It accepts a learnable identity code and joint poses of a subject as input and produces an identity-specific code for non-rigid motion estimation.

3.2. Preliminaries: HumanNeRF

In this section, we describe two technical components of HumanNeRF that our proposed framework is based on, *i.e.*, the skeletal motion estimation which learns sets of blend weights for skeletal poses, and the pose correction which corrects estimated error-prone 3D body poses.

Skeletal motion. The skeletal motion learns an inverse formulation of the linear blend skinning, which is an algorithm to render high-order deformation of objects caused by low-order skeletons [8]. It transforms a vertex from the canonical pose to the target pose by weighted-summing a point with a set of skinning weights that describe the degree of influence of each bone and transformation matrices. Weng *et al.* [35] reformulate the linear blend skinning to transform a point \mathbf{x} in the observation space to a point \mathbf{x}_c in the canonical space, which can be written as:

$$\mathbf{x}_c = \sum_{k=1}^K w_k^o (R_k \mathbf{x} + \mathbf{t}_k), \quad w_k^o = \frac{w_k^c (R_k \mathbf{x} + \mathbf{t}_k)}{\sum_{j=1}^K w_j^c (R_j \mathbf{x} + \mathbf{t}_j)}, \quad (1)$$

where w_k^o and w_k^c respectively represent observation and canonical skinning weights of bone k on a point \mathbf{x} in the observation space, and R_k and \mathbf{t}_k are the rotation and translation of the bone k . Given Eq. 1, HumanNeRF optimizes a CNN network to predict the set of canonical skinning

weights $\{w_k\}_{k=1}^K$ for each observation point where K is the total number of joints. YOTO adopts the same formulations for skeletal motion by having distinct motion priors of each identity taken by the CNN network as inputs.

Pose correction. As pointed out by Weng *et al.* [35], the aforementioned input 3D body pose (J, Ω) (J and Ω denote the joint locations and orientations, respectively) tends to be error-prone as it is from an off-the-shelf pose estimator. Therefore, it introduces a pose correction MLP that takes the joint orientations $\Omega = \{\Omega_k\}_{k=1}^K$ and predicts their offsets for pose refinement. Over multiple training iterations, the joint angles of the 3D body pose undergo continuous refinements, thus resulting in state-of-the-art performance.

3.3. Learnable Identity Codes

We now introduce the proposed learnable identity codes that are essential and effective for enabling discriminative multi-identity rendering using only one-time training. We propose to impose the learnable identity codes to condition both the non-rigid MLP M_{nr} for fine-grained motion estimation and the canonical MLP M_c for identity-specific radiance field prediction. To achieve this goal, a set of N learnable identity codes $\{\mathbf{S}_i \in \mathbb{R}^{256}\}_{i=1}^N$ are defined corresponding to N identities in the training video data, and each identity code \mathbf{S}_i is learned globally based on gradient descent to represent each identity. \mathbf{S}_i is jointly optimized with the objectives of the framework to obtain subject-specific representations. For each monocular video with a specific identity, we use its corresponding identity code as input, and thus the learned identity codes are discriminative. Then, we directly use the identity codes to condition the canonical MLP to predict the subject-specific radiance field, while for the non-rigid motion estimation, we further introduce an effective mechanism to generate pose-conditioned identity code to better facilitate the subject-specific fine-grained motion estimation.

3.4. 3D Body Pose Conditioned Identity Code

Each learned identity code can represent a subject globally from all the monocular videos. However, different video frames of the same identity may present distinct body motions. To further enhance the N learnable identity codes $\{\mathbf{S}_i\}_{i=1}^N$ for the learning of non-rigid motions, we propose to employ the 3D body pose as guidance for generating a pose-condition identity code. A cross-attention mechanism is designed to perform interactions between the learnable identity codes and the 23 joint positions (*i.e.* $J = \{j_i\}_{i=1}^{23}$). A detailed overview of the mechanism is illustrated in Fig. 3. To learn implicit representations of the 3D body joints, we conduct positional encoding [19] for the input joint points. We first project each joint position into a higher dimension by using a sinusoidal positional encoding function, $\gamma(j_i) =$

$(\sin(2^0\pi j_i), \cos(2^0\pi j_i), \dots, \sin(2^{L-1}\pi j_i), \cos(2^{L-1}\pi j_i))$, where L is the number of frequency bands; $\gamma(\cdot)$ is independently applied to each joint. After this procedure, we generate a 36-dimension representation for each input joint. We then project the encoded points (*i.e.*, $\gamma(J) \in \mathbb{R}^{23 \times 36}$) to an implicit pose code \mathbf{P}_l by feeding it to a single linear layer with parameters \mathbf{W}_p as below:

$$\mathbf{P}_l = \mathbf{W}_p \cdot \gamma(j_i). \quad (2)$$

Then, we generate one query signal \mathbf{S}_i^q from the corresponding identity code \mathbf{S}_i of the target identity via a projection matrix \mathbf{W}_Q , and key and value signals from the pose code \mathbf{P}_l via two other projection matrices \mathbf{W}_K and \mathbf{W}_V as:

$$\mathbf{S}_i^q = \mathbf{W}_Q \cdot \mathbf{S}_i, \quad \mathbf{P}_l^k = \mathbf{W}_K \cdot \mathbf{P}_l, \quad \mathbf{P}_l^v = \mathbf{W}_V \cdot \mathbf{P}_l. \quad (3)$$

Finally, we generate the pose-conditioned identity code \mathbf{F}_i^{nr} for the non-rigid motion estimation as:

$$\mathbf{P}_{nr} = \text{softmax}(\mathbf{S}_i^q \cdot (\mathbf{P}_l^k)^\top) \cdot \mathbf{P}_l^v, \quad \mathbf{F}_i^{nr} = \mathbf{P}_{nr} + \mathbf{S}_i. \quad (4)$$

We further employ this code for conditioning the learning of the non-rigid MLP (*i.e.* \mathbf{M}_{nr}). This step is intended to encourage this MLP to estimate pose-coherent non-rigid motions. The details are discussed in Section 3.5.

3.5. Pose-Conditioned IDs for Non-rigid Motions

Learning the non-rigid body motions from the input monocular video is critical for generating natural and high-fidelity rendering results. Therefore, we make use of an MLP to learn the point-specific non-rigid movements by predicting the corresponding offsets $\Delta\mathbf{x}_c$ to the canonical point \mathbf{x}_c . We first apply the aforementioned positional encoding $\gamma(\cdot)$ to enable the learning of high-frequency details. We concatenate the relative joint coordinates J to $\gamma(\mathbf{x}_c)$ as an input to \mathbf{M}_{nr} so that it can learn the pose-dependent non-rigid motions. Moreover, we additionally concatenate the pose-conditioned identity code \mathbf{F}_i^{nr} to the intermediate logits of \mathbf{M}_{nr} , thus allowing \mathbf{M}_{nr} to render subject and pose-dependent non-rigid deformations as well. We intend the code to play the role of assisting \mathbf{M}_{nr} to learn the non-rigid motions, and thus we concatenate it with the representation of point \mathbf{x}_c in the middle, which can be formulated as:

$$\Delta\mathbf{x}_c = \mathbf{M}_{nr}(J \oplus \gamma(\mathbf{x}_c); \mathbf{F}_i^{nr}), \quad \mathbf{x}_c = \mathbf{x}_c + \Delta\mathbf{x}_c. \quad (5)$$

We simply add the estimated motion offsets from the pose condition to the input 3D canonical point \mathbf{x}_c . As the result, the 3D canonical points reflect both the non-rigid motions and the subject-dependent 3D shape deformations.

3.6. ID-Conditioned Canonical Representations

To supervise the model training and render novel images, we now regress the radiance for each canonical point using another MLP. We embed the learnable identity code \mathbf{S}_i of the i -th identity in the training video, into the middle layer of the MLP via concatenating with the input point representation, so that the same canonical MLP \mathbf{M}_c can also learn

different independent subject-specific radiance fields. Benefiting from the identity codes as conditions, our YOTO framework optimizes only 1 canonical space while representing N different subjects. The color \mathbf{c}_i and density σ_i of each point can then be predicted by:

$$\mathbf{c}_i, \sigma_i = \mathbf{M}_c(\gamma(\mathbf{x}_c); \mathbf{S}_i). \quad (6)$$

Volume rendering. Following Mildenhall *et al.* [19], we adopt stratified sampling and volume rendering to compute the estimated color of each ray. We sample M different points for each ray \mathbf{r} and integrate the colors and densities of them for subject i as follows:

$$C_i(\mathbf{r}) = \sum_{m=1}^M T_m (1 - \exp(-\sigma_{i,m} \delta_{i,m})) \mathbf{c}_{i,m}, \quad (7)$$

where $\delta_{i,m}$ is the adjacent distance from m -th to $m+1$ -th sample and $T_m = \exp(-\sum_{n=1}^{m-1} \sigma_{i,n} \delta_{i,n})$.

3.7. Optimization

We perform one-time optimization of the model by training it with combined image frames of all N subjects $\{I_1^i, I_2^i, \dots, I_{F_i}^i\}_{i=1}^N$, where F_i is the number of training frame for subject i . For each training iteration, YOTO randomly selects one frame and samples rays regardless of the subject identities. For fair comparisons to the baseline, we follow the setup of Weng *et al.* [35] as described below. Our framework also samples rays in a patch \hat{P}_{F_i} from an image $I_{F_i}^i$ to utilize the LPIPS [40] loss term. The LPIPS loss term measures the perceptual distance between two image patches, and thus we also adopt it for more perceptually good renderings. Our framework takes the features of patches extracted from the pre-trained VGGNet [30] and computes the LPIPS loss term $\mathcal{L}_{\text{LPIPS}}$. Moreover, we also compute an $L2$ loss term (*i.e.* \mathcal{L}_2) of the rendered RGB for each ray. We combine the two loss terms with a coefficient λ and write the overall optimization loss \mathcal{L}_o for the whole framework as:

$$\begin{aligned} \mathcal{L}_o &= \mathcal{L}_{\text{LPIPS}} + \lambda \mathcal{L}_2, \quad \mathcal{L}_2 = \sum (C_i(\mathbf{r}) - C_i^{GT}(\mathbf{r}))^2, \\ \mathcal{L}_{\text{LPIPS}} &= \text{LPIPS}(VGG(\hat{P}_i), VGG(\hat{P}_i^{GT})), \end{aligned} \quad (8)$$

where the symbol GT indicates the ground truth.

4. Experiments

We conduct extensive experiments on a publicly available benchmark dataset (*i.e.* ZJU-MoCap [24]) to verify the effectiveness of the proposed approach for human rendering under free-viewpoints with monocular videos. We also illustrate the qualitative performance of YOTO in handling a larger number of individuals and in-the-wild settings with a single-time training on PeopleSnapshot [1]. Our findings indicate that YOTO can learn a considerable number of identities even from the in-the-wild videos, thus highlighting its effectiveness in handling such challenges.

	Subject 377			Subject 386			Subject 387		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
HumanNeRF [35]	30.39	0.9624	25.27	33.18	0.9629	30.29	28.11	0.9515	36.98
YOTO (Ours)	30.57	0.9698	21.88	33.43	0.9655	26.11	28.39	0.9534	34.55
	Subject 392			Subject 393			Subject 394		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
HumanNeRF [35]	31.03	0.9580	33.99	28.29	0.9476	39.22	30.31	0.9507	34.64
YOTO (Ours)	31.21	0.9598	31.06	28.70	0.9504	35.63	30.80	0.9535	32.11

Table 1: Quantitative results on ZJU-MoCap dataset where LPIPS* = LPIPS $\times 10^3$ following Weng *et al.* [35]. As there is no publicly available evaluation protocol from HumanNeRF (*e.g.* their used testing frames), we directly use their released checkpoints that achieved the performance mentioned in [35] to evaluate on our evaluation protocol, in which we choose to evaluate on all existing testing frames instead of sampled frames, in order to have thorough and strict evaluations. YOTO outperforms the baseline on all metrics: PSNR, LPIPS*, and SSIM. It should be noted that our model one-time trains all the identities, while HumanNeRF optimizes each identity separately.

Model	Train Time Hours	# of Param. Million	Model Size MB
HumanNeRF [35]	147	386.4	4428
YOTO (Ours)	31	65.3	747

Table 2: Efficiency comparison between our proposed framework and HumanNeRF [35] in terms of training time, the total number of parameters, and the model size.

4.1. Datasets

To thoroughly evaluate the performance of YOTO and to fairly compare against the baseline, we use ZJU-MoCap [24] dataset for quantitative evaluation. As ZJU-MoCap dataset has 23 different camera views for each subject, we use camera #1 for the training and the others for the evaluation. We train a single copy of YOTO on all 6 different subjects (*i.e.*, 377, 386, 387, 392, 393, 394) for the evaluation. Therefore, all the qualitative and quantitative results on ZJU-MoCap in the following sections are rendered by one and the same YOTO, whereas 6 different HumanNeRF models are trained for 6 subjects separately. For additional qualitative evaluation, we adopt PeopleSnapshot [1] and use 22 different videos taken in various environments.

4.2. Training Details

We train our model with mostly the same configurations as HumanNeRF [35] did so that we can clearly observe the benefits gained from our contributions. We use Adam optimizer [13] with betas (0.9, 0.999), and learning rates of 5×10^{-4} for the canonical MLP M_{nr} and the subject codes S , and 5×10^{-5} for the others. The number of patches per iteration is 6 with the dimension of 32×32 where each ray has 128 samples. We train all models including the baseline for 400K iterations with 1 Nvidia A100.

4.3. State-of-the-art Comparison

Since HumanNeRF is the state-of-the-art method for free-viewpoint rendering with monocular video, we mainly compare our performance against it. For quantitative evaluation, Weng *et al.* [35] did not release their exact evaluation protocol. For instance, the exact frame IDs of the test set used for their evaluation are not available. Thus, we conduct the quantitative comparison by evaluating the different models using all the test frames, instead of sampling a subset from them. We believe this evaluation protocol is stricter to verify the performance of a model. To ensure fair comparisons, we also directly utilize the pre-trained checkpoints released by Weng *et al.* [35] on all the following evaluations, as the authors confirmed that the best performances are from the released checkpoints.

Quantitative comparison. We compute PSNR, SSIM [34], and LPIPS* and use these metrics to quantitatively evaluate our framework. As mentioned earlier, we use the released pre-trained checkpoints of HumanNeRF for the comparison and it is denoted as HumanNeRF*, and evaluate the models on all the available novel-view frames instead of evaluating on sampled images. As demonstrated in Table 1, YOTO achieves state-of-the-art performances in terms of all the metrics on ZJU-MoCap dataset. It can be observed that our framework shows considerable improvements on LPIPS* across all subjects, while PSNR and SSIM metrics are also clearly improved. It implies that YOTO renders more perceptually reasonable and coherent novel-view images compared to HumanNeRF [35]. It should be noted that our YOTO framework jointly learns all the different identities via one-time training, while the results of HumanNeRF are evaluated on models separately trained on the different identities. These results can effectively demonstrate the performance advantages of our model.

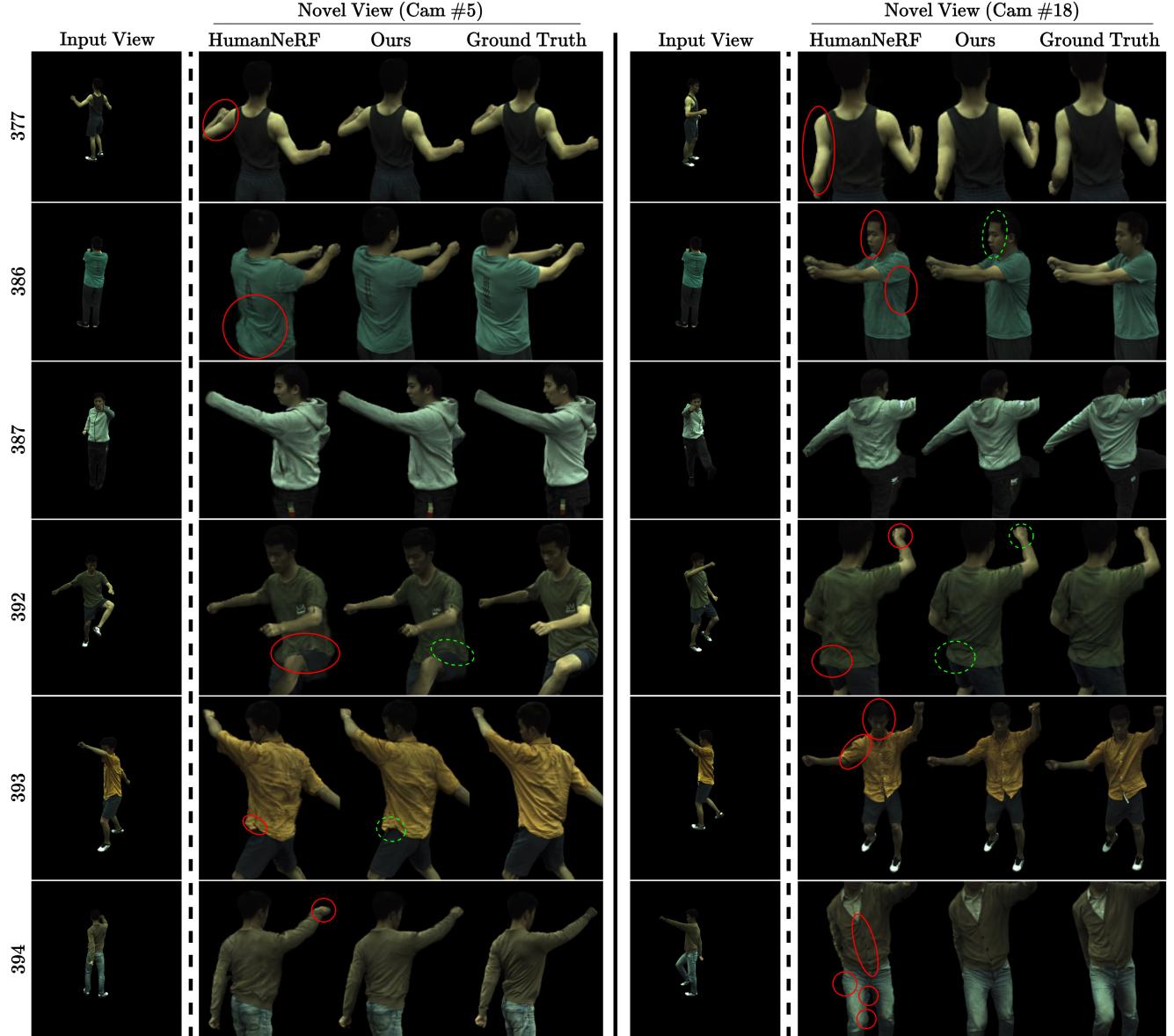


Figure 4: Qualitative comparison of free-viewpoint synthesis against HumanNeRF [35]. We indicate failing cases with red solid circles and successful cases with green dotted circles.

Qualitative comparison. Fig. 4 illustrates images rendered from novel views by both HumanNeRF and our proposed framework. The free-view rendering results for HumanNeRF are generated by their released checkpoints. As shown in Fig. 4, HumanNeRF suffers from various artifacts caused by its motion field, non-rigid and canonical MLPs. The result of HumanNeRF on subject 377 shows that the unseen parts of arms are rendered as black, whereas our framework can coherently render the skin color. Moreover, the results also show that HumanNeRF suffers from incorrect pose transformation. We can observe that HumanNeRF fails to transform the head pose of subject 386, thus showing both eyes as indicated with a red circle. For subject

392, the baseline fails to model the non-rigid motions of the t-shirt’s bottom hem caused by the raised left leg, while our framework can successfully render them benefiting from the proposed pose-conditioned identity codes for non-rigid motion estimation. All other examples in Fig. 4 can further verify that YOTO is better at modeling coherent and pose-dependent non-rigid motions.

In addition, we present qualitative results on PeopleSnapshot [1] in Fig. 5, to demonstrate the capability of our YOTO in joint handling a greater number of identities by one-time training. As can be observed in Fig. 5, mostly distinct appearances, especially the garments, are successfully learned by a single copy of YOTO. By controlling the

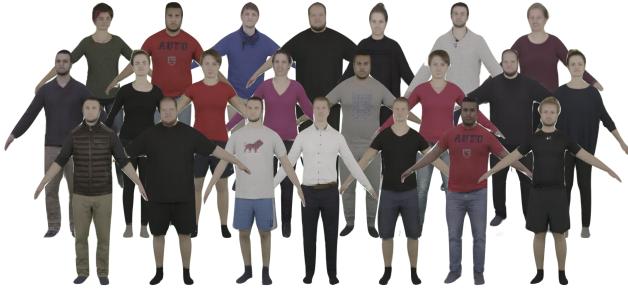


Figure 5: Illustration of the capability of YOTO on PeopleSnapshot [1] for learning with a larger number of input monocular videos. We show the rendering results of 22 different videos taken under various environments. Only one YOTO is used to learn all the representations.

	PSNR↑	SSIM↑	LPIPS*↓
YOTO (Full Model)	30.51	0.9588	30.20
w/o pose-condition	30.41	0.9583	30.39
w/o ID codes & pose-condition	28.72	0.9501	39.35

Table 3: Quantitative model analysis on ZJU-MoCap for our novel contributions, *i.e.* the learnable identity codes and pose-conditioned code generation mechanism.

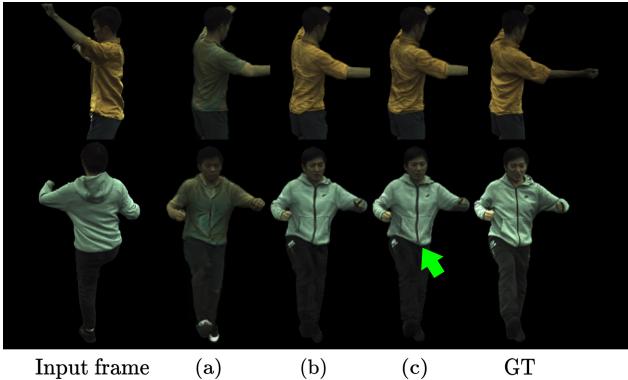


Figure 6: Qualitative ablation on ZJU-MoCap. Without the identity codes and pose-conditioned non-rigid codes (a), the model fails to learn distinct appearances. The identity codes (b) resolve the issue of (a). The pose-conditioned non-rigid codes (c) allow the model to learn more coherent and high-fidelity non-rigid motions that correspond to the 3D pose.

capacity of MLPs of YOTO and the learning rate configurations based on the desired number of identities, we believe that the rendering quality can be further boosted.

Efficiency comparison. As shown in Table 2, although we train our model with 6 different subjects jointly, there is no significant increase in the training time and the model size. YOTO requires approximately 5.5 seconds on average for every 20 iterations of training, while the baseline takes about 4.4 seconds. With the same computer resource available, YOTO boosts the total training time by $\times 4.7$ since HumanNeRF requires sequential training for all 6 subjects.

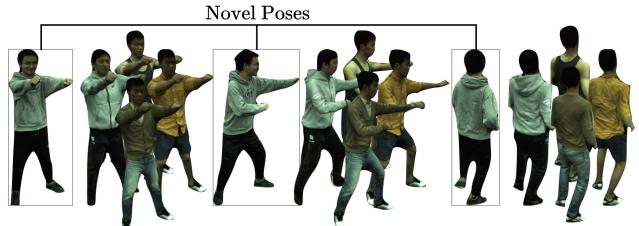


Figure 7: Illustration of motion transfer on ZJU-MoCap.

Moreover, as also stated in Table 2, the model size is only increased by 9MB which is 1.22% of the original model size whereas it increases linearly with the number of subjects in the case of HumanNeRF.

Novel Motion Transfer. YOTO has an incidental advantage in that it can animate the learned identities by simply replacing the input pose with a novel one as illustrated in Fig. 7. This improves the usability and applicability of YOTO since it would not need to train each model for the animation of each identity.

5. Model Analysis

To study the effectiveness of the proposed different components of YOTO, we consider different variants as shown in Table 3: (i) ‘YOTO (Full Model)’ indicates the proposed full version of YOTO framework; (ii) ‘w/o pose-condition’ denotes that we disable the pose-conditioned identity codes, while only using the learnable identity codes; (iii) ‘w/o ID codes & pose-condition’ means that we disable the learnable identity codes and the pose-conditioned codes. As shown in Table 3, the introduction of learnable identity codes significantly improves the quantitative performance, especially in terms of PSNR and LPIPS*. The pose-conditioned code generation module further allows YOTO to achieve state-of-the-art performance. Fig. 6 also clearly demonstrates remarkable qualitative improvements made by both of our proposed contributions. For example, as indicated with a green arrow in Fig. 6 the non-rigid motions of the hem of the jacket caused by the raised right leg is only coherently reproduced with the pose-conditioned codes.

6. Conclusion

In this paper, we presented a novel approach YOTO for simultaneous training of multi-identities from monocular videos for free-viewpoint rendering with higher fidelity and better efficiency. By optimizing a single model with the proposed learnable identity codes, the model becomes able to handle all subjects in the input monocular videos, producing even higher quality compared to the models trained separately. We further propose a novel pose-conditioned identity code to enhance motion coherency in modeling. YOTO has fully demonstrated its effectiveness and established new state-of-the-art performance on the problem.

References

- [1] Thiendo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018. [2](#), [5](#), [6](#), [7](#), [8](#)
- [2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial radiance fields. In *ECCV*, 2022. [2](#)
- [3] 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*, 2021. [2](#)
- [4] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, 2021. [2](#)
- [5] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. FastNeRF: High-fidelity neural rendering at 200fps. *ICCV*, 2021. [2](#)
- [6] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. *ICCV*, 2021. [2](#)
- [7] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debng Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. *arXiv preprint arXiv:2212.02469*, 2022. [3](#)
- [8] Alec Jacobson, Zhigang Deng, Ladislav Kavan, and J. P. Lewis. Skinning: Real-time shape deformation. In *ACM SIGGRAPH 2014 Courses*, 2014. [4](#)
- [9] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021. [2](#)
- [10] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. SelfRecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. [3](#)
- [11] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. *arXiv*, 2022. [3](#)
- [12] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. In *ECCV*, 2022. [3](#)
- [13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. [6](#)
- [14] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. [2](#)
- [15] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, 2020. [3](#)
- [16] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Theobalt Christian, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *CVPR*, 2022. [2](#)
- [17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM SIGGRAPH Asia*, 2015. [3](#)
- [18] Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, Adarsh Kowdle, Christoph Rhemann, Dan B Goldman, Cem Keskin, Steve Seitz, Shahram Izadi, and Sean Fanello. LookinGood: Enhancing performance capture with real-time neural re-rendering. *ACM Transactions on Graphics (TOG)*, 2018. [3](#)
- [19] 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. In *ECCV*, 2020. [2](#), [4](#), [5](#)
- [20] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 2022. [2](#), [3](#)
- [21] Rohit Pandey, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, Shahram Izadi, and Sean Fanello. Volumetric capture of humans with a single rgbd camera via semi-parametric learning. In *CVPR*, 2019. [3](#)
- [22] 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*, 2021. [2](#)
- [23] 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. *SIGGRAPH Asia*, 2021. [2](#)
- [24] 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*, 2021. [2](#), [5](#), [6](#)
- [25] 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*, 2021. [3](#)
- [26] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2020. [2](#)
- [27] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. KiloNeRF: Speeding up neural radiance fields with thousands of tiny mlps. In *ICCV*, 2021. [2](#)
- [28] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. [2](#)
- [29] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. [3](#)
- [30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. [5](#)
- [31] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view

- synthesis of a dynamic scene from monocular video. In *ICCV*, 2021. 2
- [32] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2
- [33] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. ARAH: Animatable volume rendering of articulated human sdbs. In *European Conference on Computer Vision*, 2022. 3
- [34] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6
- [35] 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*, 2022. 2, 3, 4, 5, 6, 7
- [36] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *CVPR*, 2020. 3
- [37] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *CVPR*, 2021. 2
- [38] Alex Yu, Rui long Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021. 2
- [39] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2
- [40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5