DOI: 10.1111/cgf.14465
EUROGRAPHICS 2022 / R. Chaine and M. H. Kim
(Guest Editors)
Volume 41 (2022), Number 2

MoCo-Flow: Neural Motion Consensus Flow for Dynamic Humans in Stationary Monocular Cameras

Xuelin Chen[1] Weiyu Li[2,1] Daniel Cohen-Or[3] Niloy J. Mitra[4,5] Baoquan Chen[6]

[1]Tencent AI Lab [2]Shandong University [3]Tel Aviv University [4]University College London [5]Adobe Research [6]CFCS, Peking University

Figure 1: *Given only a video (a) captured from a **stationary monocular** camera, our method can synthesize novel views (b) of the dynamic human from arbitrary viewpoints and at any time. An input view is shown in the middle on the right, with imagery synthesized from 8 novel views spreading around the performer. For dynamic results, please see the supplementary video.*

Abstract

*Synthesizing novel views of dynamic humans from stationary monocular cameras is a specialized but desirable setup. This is particularly attractive as it does not require static scenes, controlled environments, or specialized capture hardware. In contrast to techniques that exploit multi-view observations, the problem of modeling a dynamic scene from a single view is significantly more under-constrained and ill-posed. In this paper, we introduce Neural Motion Consensus Flow (*MoCo-Flow*), a representation that models dynamic humans in stationary monocular cameras using a 4D continuous time-variant function. We learn the proposed representation by optimizing for a dynamic scene that minimizes the total rendering error, over all the observed images. At the heart of our work lies a carefully designed optimization scheme, which includes a dedicated initialization step and is constrained by a* motion consensus *regularization on the estimated motion flow. We extensively evaluate MoCo-Flow on several datasets that contain human motions of varying complexity, and compare, both qualitatively and quantitatively, to several baselines and ablated variations of our methods, showing the efficacy and merits of the proposed approach. Pretrained model, code, and data will be released for research purposes upon paper acceptance.*

CCS Concepts

• ***Computing methodologies*** → *Shape modeling; Rendering;*

1. Introduction

We address the challenging problem of synthesizing novel views of dynamic humans in stationary monocular cameras. View syn-

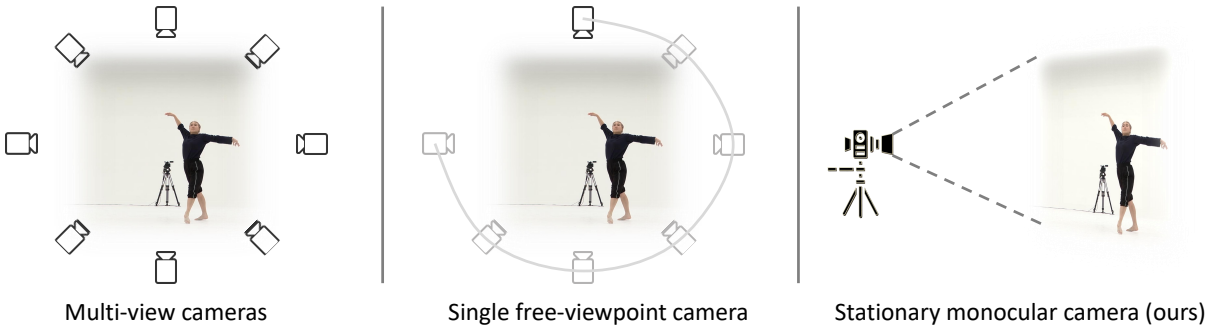

Figure 2: *(Left) Multi-view cameras setup for full observation of a dynamic scene; (middle) single free-viewpoint camera setup that captures the dynamics from varying viewpoints; (right) stationary monocular camera to observe a dynamic scene from only a single fixed viewpoint.*

thesis has been a long-standing problem in both computer vision and computer graphics. Neural Radiance Field (NeRF) [MST*20] has recently revolutionized novel view synthesis of *static* structures by directly optimizing parameters of a continuous 5D scene representation to minimize the error of rendering multiple captured images. Subsequently, there has been a surge of followups extending it to deal with dynamic scenes [PSB*21, LNSW21, PCPMMN21, XHKK20, LSZ*21, GTZN21, PZX*21, TTG*21, DZY*20, PSH*21].

These dynamic NeRFs have shown impressive performance in view synthesis but require different setups to capture the dynamics. The most natural extension is to still use a multi-view camera setting [PZX*21, LSZ*21] to acquire sufficient observation of the dynamic scene from multiple viewpoints. While the multi-view setup significantly constrains modeling of the dynamics, the capture process relies on controlled environments and specialized hardware to synchronize the different acquisitions. Another line of work exploits a *single* free-viewpoint camera to capture dynamic scenes from *varying viewpoints* [PSB*21, LNSW21, PCPMMN21, XHKK20, TTG*21, DZY*20]. While these methods bypass the need for expensive equipment, they still require the capture device to be suitably moved around to allow capturing dynamic scenes, and inevitably rely on Structure-from-Motion (SfM) systems for accurate camera extrinsic parameters to constrain the modeling.

In this paper, we present a dynamic NeRF technique for synthesizing novel views of dynamic humans from stationary monocular cameras. The problem of modeling dynamic scenes from monocular cameras is typically under-constrained, as shown in [PSB*21, LNSW21, PCPMMN21, XHKK20, TTG*21, DZY*20]. Moreover, our setting is even more challenging. Unlike the multiple viewpoints setting, in our stationary monocular camera setting, we only observe the dynamic scene from a single fixed viewpoint. Hence, the extrinsic camera parameters cannot be obtained from SfM to constrain the dynamic scene modeling from multiple viewpoints as in aforementioned works (Figure 2 illustrates different capture setups). On the other hand, a technique for stationary monocular cameras has a significant scope and is applicable to a wide range of everyday captures. Synthesizing novel views from stationary videos would offer creating strong immersive experience for existing videos, and could be embraced in the future by millions of video content producers. Furthermore, stationary videos are easy to capture and require no special assistance, environment or hardware.

In general, to model a dynamic scene for view synthesis, it can be decomposed into a shared canonical static scene for representing the appearance and the geometry of the subjects, and a motion flow that model the dynamics between the canonical space and the observation space at each frame. Both representations can be approximated by Multi-layer Perceptron (MLP) networks and optimized to re-produce the observed frames via differentiable volume rendering. However, a naive approach cannot deal with this overly ambiguous single fix-viewpoint setting, and the network update is prone to an erroneous overfit, as multiple solutions comprised of meaningless canonical representation and motion flows can be combined together to reproduce the observation.

Hence, the key to solving the above formulation is to harness this challenging optimization, throughout the whole optimization process. This results in human bodies and their dynamics that follow faithfully the human perception of the video. To this end, we devise a carefully designed and yet *easy-to-implement* optimization scheme, which disambiguates relatively much worse local minima early at the initialization phase, and imposes a crucial regularization on the update of the motion flow to reach a high degree of *consensus* across the observations, denoted as Motion Consensus Flow (MoCo-Flow in short), consequently preventing the optimization from deviating too much from the initialization and landing on bad local minima. It is worth noting that, in contrast to heuristic regularization terms, the regularization imposed in MoCo-Flow is general and does *not* assume any dynamic characteristics on the moving subjects.

We demonstrate our method on several publicly available datasets, namely AIST [TFHG19], People-Snapshot [AMX*18], and ZJU-MoCap [PZX*21], where we have access to human performance videos filmed by stationary monocular cameras, to show the effectiveness of our method on synthesizing novel views with high visual quality and motion dynamics. We extensively evaluate and compare our method against existing methods, both qualitatively and quantitatively, showing that our method can exhibit state-of-the-art novel view synthesis of dynamic humans from stationary monocular cameras. The comparisons demonstrate that directly applying existing methods that are not dedicated for our stationary monocular setting would simply lead to failure. We also conduct ablation experiments to compare our method against several variants to understand the importance of these key designs.

2. Related Work

Digitizing human bodies has received much attention in both computer graphics and computer vision, with a vast literature of human body performance capture. We first review the works on capturing human performance. Our work is built upon the success of the Neural Radiance Field technique, so we shall also briefly cover recent developments related to NeRF for novel view synthesis.

Human performance capture. Decades of research has been devoted to faithfully capture humans performance. Most of them follow a model-and-render procedure for producing novel view imagery. These methods typically exploit multi-view systems [DHT*00, GLD*19, OERF*16, CCS*15, LFB17, SH07], depth cameras or fusion of depth sensors [IKH*11, NIH*11, ZK14, CCNS12, LVG*13, SFW*14, ZZCL13, DKD*16, SXZ*20, NFS15] to reconstruct geometry, static or motion-factored, by aggregating observation gained from various viewpoints. Various approaches [MBPY*18, WWHY20, LSS*19, PZX*21] have also incorporated emerging neural rendering techniques for compensating the visual loss in the reconstruction.

In the sparse views case, which is more unconstrained and ambiguous, model based methods utilize the prior knowledge of parametric template models to help significantly constrain the solution space for the unobserved body parts. The final geometry is obtained by deforming parametric coarse templates or pre-scanned accurate shapes to fit the captured images [AMX*18, CTMS03, DAST*08, GSDA*09, SGDA*10, PMR11, HSS*09]. Another line of works [LMR*15, KAB20, KBJM18, KPBD19, NSH*19, SHN*19, SSSJ20, ZYW*19] learn human body priors by training networks on a large collection of images data, enabling the inference of complete human models from single images or monocular videos. It remains, however, very difficult for these learning-based methods to produce plausible results for out-of-distribution human samples. In general, while seeking to explicitly model the geometry and texture of the subjects, these current state-of-the-art methods have difficulty in producing realistic view synthesis by rendering from the explicitly reconstructed 3D models.

Neural representations for view synthesis. Neural representation has been one of the key infrastructures to the neural rendering technique that is able to render photo-realistic imagery. Generative Query Network (GQN) [ERB*18, KER*18], the pioneering work in this direction, perceives the underlying 3D scene from a set of input images based on a neural representation and generation network. With an implicit notion of 3D, GQN can synthesize arbitrary views with correct occlusion. Following that, a variety of methods [LSS*19, SZW19, STH*19] emerged that include a more explicit representation of the 3D, exploiting components of the graphics pipeline. We strongly recommend [TFT*20] for a thorough summary of this emerging field. Lately, the Neural Radiance Field (NeRF) [MST*20] technique has revolutionized novel view synthesis of static structures by training an MLP-based radiance and opacity field. Through a differentiable volume rendering technique, NeRF achieved unprecedented success in producing photo-realistic novel view imagery. An explosion of NeRF techniques occurred in the research community since then that improves the NeRF in various aspects of the problem [LGL*20, LMW21, RJY*20, ZRSK20, WPYS21, MBRS*21, WWX*21, LMTL21, LSS*21]. Nevertheless, all these achievements were made on static structures, it remains challenging to extend the static NeRF to deal with dynamic scenes.

Dynamic NeRFs. Lately, there has been a surge of developments related to NeRF extending to deal with dynamic scenes. These dynamic NeRFs have shown impressive performance in view synthesis with different setups to capture the dynamics. The most natural extension is to still use the multi-view camera setting [PZX*21, LSZ*21] to acquire sufficient observation of the dynamic scene. The multi-camera setup helps constrain the modeling of the dynamics significantly. However, these extensions require to have controlled environments and specialized hardware to acquire full observation on the dynamic scene, implying a difficulty in popularizing applications.

Another effort is to relax the need of heavy setups, by exploiting a *single* free-viewpoint camera to capture dynamic scenes from *varying viewpoints* [PSB*21, LNSW21, PCPMMN21, XHKK20, TTG*21, DZY*20]. Although the extrinsic camera parameters obtained from SfM can help constrain the modeling to some extent, the key in this monocular camera setting is to deal with the ambiguity of the geometry, texture, and motion of the subjects due to occlusion. The canonical NeRF and motion flow formulation in this work are similar in spirit with those recently emerging dynamic NeRFs but differ in how we regularize the challenging optimization. Nerfies [PSB*21] proposes an as-rigid-as-possible (ARAP) regularization of the deformation, which assumes elastic deformation behavior of the subjects. In addition, this ARAP regularization has extremely high computational complexity. Analogically, NR-NeRF [TTG*21] proposes regularizers on the estimated deformations which constrain the problem by encouraging small volume preserving deformations. NSFF [LNSW21] leverages external supervision such as rough monocular depth estimation and flow-estimation, which unfortunately are not further jointly optimized during the optimization, to resolve ambiguities. Surprisingly, although D-NeRF [PCPMMN21] also uses the single free-viewpoint setting, the spirally flying camera in their setup covers the entire dynamic scene and the observation turns out to gain sufficient information. Although these methods bypass the need for expensive equipment, they require the capture device to be empowered with certain mobility to allow capturing dynamic scenes. NerFACE [GTZN21] shares the most similar setup to ours, but is highly specialized for human faces. They also do not optimize to obtain face dynamics but purely rely on the high precision of the face tracking method to capture the dynamics.

3. Method

The input to our method is a human performance video $\boldsymbol{I} = \{\boldsymbol{I}_i\}$ captured by a stationary monocular camera, where $i \in \{0, \ldots, m-1\}$ for m observation frames. The performer can perform arbitrary motions in front of the camera so as to show the body sufficiently for a full 360° novel view navigation. Nevertheless, our method can also work well for scenarios where the camera only observes the body partially. In addition, we assume a static background image captured without the human performer. If inapplicable, we simply set the background to white via foreground detection [HGDG17].

To represent dynamic scenes, we decompose the dynamic scene that contains the moving subject into a shared canonical space repre-

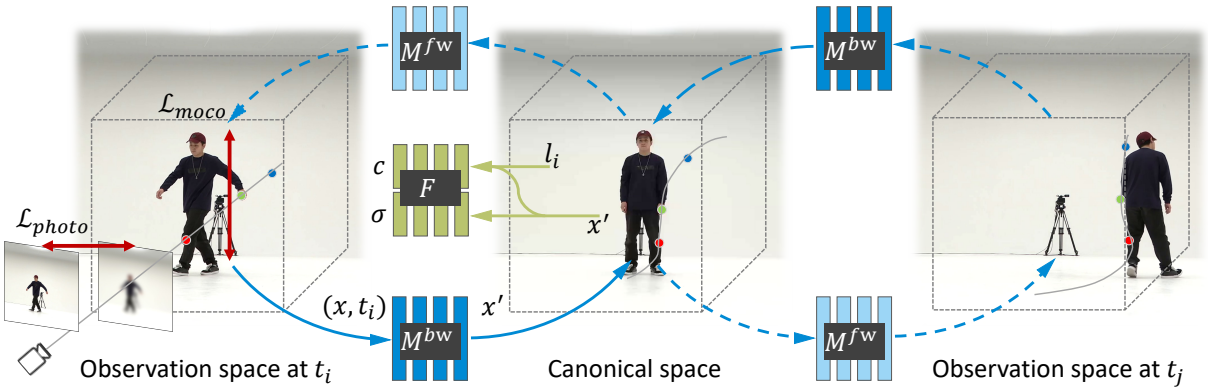

Figure 3: ***MoCo-Flow****. The dynamic scene is represented by a shared canonical NeRF and motion flows. We trace rays in the observation space* t_i *and transform the samples* x *along the ray to 3D samples* x' *in the canonical space via the neural backward motion flow* $\mathrm{M}^{bw} : (\mathrm{x}, \mathrm{t}_i) \rightarrow \mathrm{x}'$. *We evaluate the color and density of* x *at* t_i *through the canonical NeRF with a condition appearance code* $\boldsymbol{l}_i$*:* $\mathrm{F}(\mathrm{x}', \boldsymbol{l}_i) \rightarrow (\mathrm{c}, \sigma)$. *The networks are initialized with rough human mesh estimation and then optimized to minimize the error* $\mathcal{L}_{photo}$ *of rendering captured images. An auxiliary neural forward motion network (*M^{fw}*) is introduced to constrain the optimization with motion consensus regularization* $\mathcal{L}_{moco}$ *(see the loop formed by the blue arrows).*

sented as a neural radiance field (NeRF) and a motion flow that models, for each time step, per-coordinate correspondences between the canonical space and the observation space (Sec. 3). Both representations are *simultaneously* optimized to model a dynamic scene that minimizes the error of reproducing all observation images through differentiable volume rendering. Without the loss of generality, we set the canonical space to be the one at the first frame.

There are two key features in our work for addressing this overly ambiguous and under-constrained optimization problem. First, we utilize domain-specific data priors for building up a canonical NeRF and a neural motion flow that serves as good initial values to our optimization (Sec. 3.2.1). This still leaves a significant search space with too many irrelevant local minima. Second, while constraining the solution to remain close to the initial guess, we also introduce a novel motion consensus regularization. This regularization does not limit the dynamic characteristics of the moving subjects, is computationally efficient, and effectively encourages the motion to reach a high degree of consensus among all observation spaces (Sec. 3.2.2).

3.1. Neural Dynamic Scenes

Canonical neural radiance field. The neural radiance field, which is approximated using an MLP network F, is a continuous scene representation that maps a 3D coordinate $\mathrm{x} = (x, y, z)$ and viewing direction d to an emitted color $\mathrm{c} = (r, g, b)$ and volume density σ. The original NeRF was introduced for synthesizing novel views of a static scene from multi-view images, where the color prediction c is additionally conditioned on viewing directions. While only having constant viewing directions over training, conditioning the color on viewing directions does not hold true in our stationary monocular camera setting, this condition is thus removed from our model. Last, similar to [MBRS*21], to modulate the appearance variation across all observation images, we condition the color prediction on an optimizable appearance latent code $\boldsymbol{l}_i$ associated to each image $\boldsymbol{I}_i$: $\mathrm{F} : (\mathrm{x}, \boldsymbol{l}_i) \rightarrow (\mathrm{c}, \sigma)$.

Neural motion consensus flow. To represent moving subjects in the dynamic scene, we introduce another MLP-parameterized network M^{bw} to model the motion between the canonical space and the observation space at each time step. Formally, given a 3D coordinate x at time t_i, M^{bw} is optimized to transform x *back* to a 3D coordinate x' in the canonical space via predicting an SE(3) transformation [PSB*21]: $\mathrm{M}^{bw} : (\mathrm{x}, \mathrm{t}_i) \rightarrow \mathrm{x}'$, where we can evaluate the color and density of x at time t_i with the canonical NeRF. In addition, to enforce the motion consensus over time for penalizing the deviation of the optimization as aforementioned, we introduce an auxiliary motion flow network M^{fw} that inversely transforms a 3D coordinate x' in the canonical space to a 3D coordinate x in each observation space: $\mathrm{M}^{fw} : (\mathrm{x}', \mathrm{t}_i) \rightarrow \mathrm{x}$.

Since directly passing raw coordinates to MLP networks would fail to learn high-frequency functions in low-dimensional problem domains [MST*20], we lift the input 3D coordinates x and the time step t to higher dimension spaces for both the canonical NeRF and the neural motion flow networks. The lifting is performed using the positional encoding function $\gamma(\mathrm{x})$ and $\gamma(\mathrm{t})$ as proposed in [MST*20].

Volume rendering. With the backward neural motion flow, we can simply evaluate the radiance field at each time step as: $\mathrm{F}(\mathrm{M}^{bw}(\mathrm{x}, \mathrm{t}_i), \boldsymbol{l}_i) \rightarrow (\mathrm{c}, \sigma)$, for volume rendering the observation space, thus accounting for the inferred dynamics. The volume rendering equation as in [MST*20] is employed to render images from the radiance field of each observation space. Recall that we assume to have a decoupled static background image (i.e., without the human body). When rendering, the last sample on the ray is assigned with the color of the pixel corresponding to the ray on the background image. This encourages the networks to predict high density values only for 3D coordinates of moving subjects so as to reproduce a clean background though the differentiable volume rendering.

3.2. Optimization

3.2.1. Initialization

We use VIBE [KAB20] to estimate a SMPL [LMR*15] mesh sequence from the video, with which we can sample points on the body in observation spaces and obtain their corresponding samples in the canonical space, and vice versa. However, although VIBE has superior generalizability, it assumes an orthogonal camera projection during training over a large collection of images, which implies it cannot estimate the 3D spatial location while predicting the shape and pose parameters of the SMPL model. Hence, to obtain the location, we render the mask of the mesh using known intrinsic camera parameters and compare against the Mask R-CNN [HGDG17] detected mask. We perform a location search on a 3D grid to minimize the matching loss. Figure 4 shows a resultant mesh sample.

Figure 4: *A VIBE based reconstruction provides only a rough estimate of the geometry, pose, and location.*

We update the weights of both M^{bw} and M^{fw} to fit the extracted observation-canonical point pairs to serve as initialization to the subsequent optimization. Moreover, due to the infeasibility of building explicit correspondences between observation-space non-human samples and canonical-space non-human samples, we simply suppress the density value on these free samples using Binary Cross Entropy (BCE) loss without specifying the target location for them. Overall, the loss for initializing the motion flow networks is defined as:

$$\begin{aligned}\mathcal{L}_{mo}^{fit} := &|\mathrm{M}^{bw}(\mathrm{x}_h,\mathrm{t}_i)-\mathrm{x}'_h|+|\mathrm{M}^{fw}(\mathrm{x}'_h,\mathrm{t}_i)-\mathrm{x}_h| \\ &+\mathrm{BCE}(\mathrm{F}_\sigma(\mathrm{M}^{bw}(\mathrm{x}_f,\mathrm{t}_i),\boldsymbol{l}_i),0),\end{aligned} \tag{1}$$

where x_h denotes observation-space human samples, x'_h the corresponding canonical-space samples, and x_f free samples in observation spaces.

Further, we use the geometry of the canonical mesh to initialize the density branch F_σ of the canonical NeRF. Departing from this initialization, the neural dynamic scene represented by the canonical NeRF and the motion flow networks is subsequently optimized by the MSE photometric loss $\mathcal{L}_{photo}$ between the image rendered from the fixed viewpoint at each time step and the input image.

3.2.2. Regularization

As the initialization derived from the estimated SMPL meshes apparently contains errors regarding the geometry, location, and motion of the human (see Figure 4), the canonical NeRF and the motion flow need to be further optimized and corrected. We use two more strategies to guide the optimization process.

Motion consensus regularization. We apply a global all-to-all motion consensus regularization to the update of the motion flow networks, which enforces bidirectional flows between two observation spaces to achieve high consensus though the canonical space (see the loop formed by the blue solid line and dash lines in Figure 3),

$$\mathcal{L}_{moco}^{global} := |\mathrm{M}^{fw}(\mathrm{M}^{bw}(\mathrm{M}^{fw}(\mathrm{M}^{bw}(\mathrm{x},\mathrm{t}_i),\mathrm{t}_j),\mathrm{t}_j),\mathrm{t}_i)-\mathrm{x}|, \tag{2}$$

where t_i and t_j are random samples of time steps. Additionally, we apply a local motion self-consensus regularization, enforcing the motion flow to be invertible locally at each time step t_i:

$$\mathcal{L}_{moco}^{local} := |\mathrm{M}^{fw}(\mathrm{M}^{bw}(\mathrm{x},\mathrm{t}_i),\mathrm{t}_i)-\mathrm{x}|. \tag{3}$$

The total motion consensus regularization is thus defined as: $\mathcal{L}_{moco} := \mathcal{L}_{moco}^{global} + \mathcal{L}_{moco}^{local}$. Note that we only apply this motion consensus regularization on samplings of moving subjects, while imposing no regularization on the dynamics of free space samplings. This is achieved by simply filtering out free space samplings with a density threshold ε=0.01.

Coarse-to-fine flow regularization. We further employ a coarse-to-fine annealing strategy, as in [PSB*21], to gradually optimize the neural dynamic scene from modeling low-frequency details to high-frequency details [TSM*20]. This regularization is carried out by gradually increasing the frequency bands used in the positional encoding of 3D coordinates. The positional encoding for each of the three coordinate values in x is then changed to: $\gamma(\mathrm{x}) = (\mathrm{x},...,w_{f-1}(\alpha)\sin(2^{\mathrm{f}-1}\pi\mathrm{x}), w_{f-1}(\alpha)\cos(2^{\mathrm{f}-1}\pi\mathrm{x}))$, where f is the maximum number of frequency bands, $w_k(\alpha) = (1-\cos(\pi\mathrm{clamp}(\alpha-k,0,1)))/2$, and $\alpha(n) = \mathrm{f}n/\mathrm{N}$ with n being the current optimization iteration and N the total iterations for the coarse-to-fine optimization stage. Note that this coarse-to-fine regularization is only applied to the positional encoding function of 3D locations $\gamma(\mathrm{x})$, the positional encoding of time steps $\gamma(\mathrm{t})$ uses constant number of frequency bands during the optimization.

3.3. Implementation details

In our implementation, both the canonical NeRF and the motion consensus flow networks are approximated by an 8-layer MLP network with hidden width 256, ReLU activation, and a skip connection at the 4th layer. We use hierarchical volume sampling strategy [MST*20], sampling 64 coarse locations and 128 fine locations along the rays. We set the number of frequency bands f used in $\gamma(t)$ to 16, and the maximum number of frequency bands f used in $\gamma(\mathrm{x})$ to 8. We use 8 dimensions for the appearance latent codes, which are randomly initialized prior to the optimization.

Adaptive bounding volume. Evaluating 3D coordinates over the whole 3D space of the dynamic scene requires the samples along the rays to be sufficiently dense for synthesizing high-quality imagery, consequently leading to a time-consuming optimization process. Since the moving subjects typically occupy a small portion of the large space, inspired by [LGL*20, LSS*21], in each observation space, we set an adaptive bounding volume for bounding the ray marching. Concretely, at each observation space, we compute an axis-aligned bounding box (AABB) from the estimated mesh, and empirically enlarge this AABB by an offset of 0.2 meters along XY dimension and 0.4 meters along Z dimension to ensure it is sufficient to entirely cover the underlying human body. Then, we only evaluate samples along the rays that are intersecting with the AABB, and bound the sampling to be within the AABB.

Initialization and optimization. To initialize the density branch F_σ

Figure 5: *Qualitative results on People-snapshot dataset. Odd rows show the input observation across time from the stationary monocular view; even rows show the associated imagery synthesized from a novel view across time. Note that the logos on clothing (see results in the middle of the 6th row) are recovered in the novel view, although unseen from the input view. Please use digital zoom.*

of the canonical NeRF, we render a set of multi-view images of the canonical SMPL mesh, and follow the optimization as in [MST*20] to obtain the weights of F_σ. Next, we initialize the motion flow networks M^{bw} and M^{fw} and the color branch F_c of the canonical NeRF to overfit with the following loss, while freezing the density branch F_σ: $\mathcal{L}^{init} := \mathcal{L}_{photo} + \lambda\mathcal{L}_{moco} + \mu\mathcal{L}^{fit}_{mo}$, where we use $\lambda = 0.2$, and $\mu = 10$. Subsequently, we unfreeze F_σ and jointly optimize all networks with the loss: $\mathcal{L}^{joint} := \mathcal{L}_{photo} + \lambda\mathcal{L}_{moco}$. Typically, the entire optimization takes around 3 days on 8 V100 GPUs. Rendering a novel view image at resolution 512 x 512 takes roughly 30 seconds. Since it is ambiguous to model the background from stationary monocular cameras, we simply set a background image with solid colors (a virtual background image is also possible) for rendering novel views. See also the supplementary.

4. Experiments

In this section, we describe the datasets and metrics for evaluating our method, and show both qualitative and quantitative results of our method. We also present quantitative and qualitative comparisons against several baseline methods, along with experiments conducted for evaluating different aspects of our method.

Figure 6: *Qualitative results on AIST. From top to bottom: male pop, female pop, and female ballet dancer. Each two rows shows the input observation across time (the row showing gray background) and associated imagery from a novel view. The first column visualizes the camera setup and the initial body orientation in the first frame.*

Datasets. We evaluate on three data sources that contain human performances of varying complexity: *(A) People-snapshot* [AMX*18], which captures human performers that rotate while holding an A-pose in front of a stationary camera. This dataset contains motions of rather low complexity, and does not offer ground truth images from other viewpoints. So we further evaluate on more challenging datasets, where ground truth novel view images are available as well. *(B) AIST* [TFHG19,LYRK21], which is a shared database containing dance videos. The videos are captured using multiple cameras (9 at most) surrounding a dancer to simultaneously shoot from various directions. We use only the monocular video captured from the lower front by default (i.e., the camera with ID C09 as described in the database) for modeling, and use videos filmed from the rest positions as ground truth for evaluation. *(C) ZJU-MoCap* [PZX*21], which is another human performance dataset created for evaluating dynamic human reconstruction from multi-view videos. The humans perform arbitrary and complex motions, including twirling, arm swings, punching, kicking and so forth, in a multi-camera system that has 21 synchronized cameras. Again, we use only the monocular video captured at the camera with ID 01 by default, and use the remaining cameras for evaluation. We refer readers to the supplementary for more details of data processing.

Evaluation measures. Since it is ambiguous to infer the surrounding environment in novel views given a stationary monocular camera only, we simply mask out the background of ground truth images and crop an image patch (512 x 512) around the human center for computing the metrics.

There are several commonly standard metrics: peak signal-to-noise ratio (PSNR), and perceptual similarity through LPIPS [ZIE*18]. However, we found that these metrics are very unsuitable for our task due to severe misalignment between the ground truth images and the synthesized views (which we shall discuss in the limitations), exhibiting an irrational trend. This is also pointed out in [PSB*21]. Here we first conduct experiments to investigate the influence of the misalignment on the PSNR and LPIPS metrics.

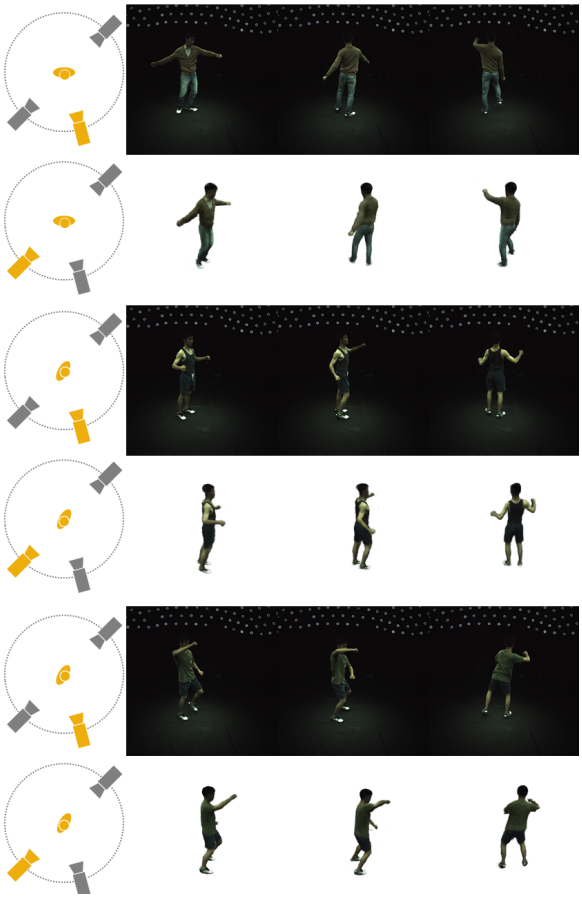

Figure 7: *Qualitative results on ZJU-MoCap. From top to bottom: Swing2, Warmup, and Swing1. Each two rows shows the input observation across time (the row showing garyish background) and associated imagery from a novel view. The first column visualizes the camera setup and the initial body orientation in the first frame.*

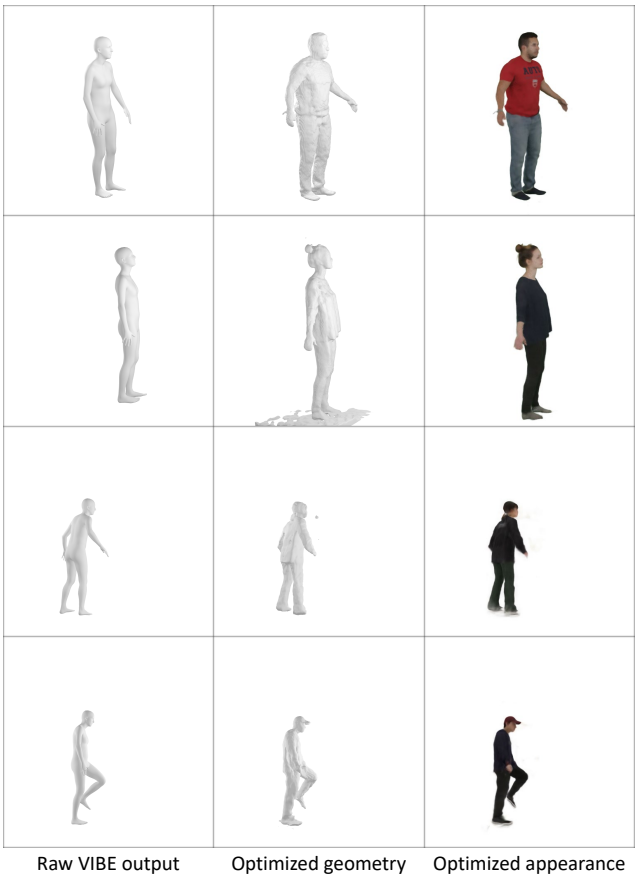

Figure 8: *Side-by-side visual comparisons to raw VIBE outputs, showing the improvements on both the geometry and the appearance over the optimization.*

Concretely, we generate a set of images via translating the ground truth image by offsets (in pixel) along random directions or rotating around the human center by degrees along random directions. Table 1 shows the PSNR and LPIPS scores degrade dramatically while the visual content remains correct but misalignment increases.

So, to better evaluate our novel view synthesis results, we propose to calculate the *plausibility* of the synthesized images as human imagery, which is evaluated as the human detection accuracy in percentage produced by Mask-RCNN. In addition, we further propose to measure the *pose accuracy* that is evaluated as the commonly used object keypoints similarity (OKS) in the human pose detection field. Concretely, we use the AlphaPose [FXTL17, LWZ*18, XLW*18] method to detect the human keypoints both in the ground truth image and the synthesized novel view image, then canonicalize these two detected human poses by aligning their mean centers, and finally compute the OKS as the pose accuracy for the novel view image.

MoCo-Flow novel view synthesis. We present qualitative results of MoCo-Flow on synthesizing novel views of dynamic humans given a stationary monocular video only. Figure 5 and Figure 7 show the visual results on the aforementioned datasets, wherein we render the imagery from *unseen* viewpoints in the dynamic scene during the performance. We can see, in addition to reconstructing highly plausible motions and human geometries with clothing in novel views, our method can also recover fine details such as patterns on clothing, the thin string-band on the wrist, and the hat brim in the novel view imagery.

Table 1: *PSNR and LPIPS scores degrade dramatically while the misalignment increases but the visual content remains correct.*

translation (px)	10	20	30	40	50
PSNR	20.46	17.44	16.38	15.51	14.84
LPIPS	0.04	0.08	0.10	0.12	0.13

rotation (deg.)	5	10	15	20	25	30
PSNR	21.07	18.24	16.97	16.25	15.74	15.34
LPIPS	0.03	0.06	0.08	0.10	0.11	0.12

Table 2: *Quantitative comparisons on AIST dataset. D-NeRF outputs blank imagery at novel views.*

	PSNR					LPIPS					Plausibility ↑					OKS ↑				
	D-NeRF	NSFF	NB	NerFACE	MoCo-Flow	D-NeRF	NSFF	NB	NerFACE	MoCo-Flow	D-NeRF	NSFF	NB	NerFACE	MoCo-Flow	D-NeRF	NSFF	NB	NerFACE	MoCo-Flow
male pop	15.575	13.514	**18.195**	17.008	16.102	0.133	0.561	**0.094**	0.104	0.109	0.000	0.060	0.430	0.640	**0.941**	0.000	0.092	0.288	0.431	**0.670**
female pop	16.578	15.131	**17.694**	16.425	14.531	0.117	0.511	**0.086**	0.099	0.151	0.000	0.020	0.320	0.563	**0.943**	0.000	0.079	0.337	0.383	**0.495**
female ballet	16.912	**17.890**	17.880	17.215	17.265	0.123	0.104	0.104	0.118	**0.100**	0.000	0.030	0.360	0.605	**0.920**	0.000	0.018	0.134	0.113	**0.503**

In addition, we show more qualitative evaluation of the learned motion flow. In Figure 8, we present side-by-side visual comparisons of MoCo-Flow results against the raw VIBE outputs, showing the improvements of the optimized geometry and appearance by the motion flow over the initialization. In Figure 9, we visualize the dense correspondences derived from the learned motion flow between the canonical and the observation space.

Comparisons. We present both qualitative and quantitative comparisons against several latest works, namely D-NeRF [PCP-MMN21], NSFF [LNSW21], Neural Body (NB) [PZX*21] , and NerFACE [GTZN21], on the AIST dataset, where ground truth novel views are available for evaluation. We obtained the results of the former three using authors' code. As for NerFACE, which is highly specialized for faces and does not optimize the dynamics as aforementioned, we implement it based on our framework to work on SMPLs. More details can be found in the appendix.

Although these methods can still overfit the training view, directly applying existing methods that are not dedicated for our stationary monocular setting would simply lead to failure in synthesizing novel views. D-NeRF even outputs blank imagery. The quantitative comparisons are presented in Table 2. We struggle to obtain higher PSNR and LPIPS due to the *mismatch* between the physical factors of the reconstructed 3D and the ground truth, as PSNR and LPIPS are not ideal metrics for evaluation on our task, while the baselines easily get similar scores with meaningless results (e.g., the D-NeRF even gets high scores with white images). But, our method dominates over the plausibility and pose accuracy score, outperforming the baselines by significant margins, as baselines completely failed to synthesize human imagery from novel views. We present the visual comparisons in Figure 11, where our results have high plausibility and accurate poses, and outperform baselines in consistence with the quantitative comparisons.

Last, we also show side-by-side comparisons with a model-based method - Video Avatar [AMX*18], which works only on A-posed performers and reconstructs only the canonical textured mesh. The visual comparison results are presented in Figure 10.

Ablation study. We first conduct experiments, where several variants are obtained by removing one component from MoCo-Flow, to individually evaluate the contribution of each component. Although most variants are still able to overfit the input observation, they fail to produce high-quality novel view images. More specifically, without the total (*w/o* $\mathcal{L}_{moco}$), the global (*w/o* $\mathcal{L}^{global}_{moco}$), or the local motion consensus regularization (*w/o* $\mathcal{L}^{local}_{moco}$), the body are significantly more distorted; without the dedicated initialization step (*w/o init.*), the optimization completely failed, producing meaningless blank novel view imagery; with the absence of the adaptive bounding volume (*w/o ada. vol.*), we observe plenty of noisy floaters in novel views; without the coarse-to-fine regularization (*w/o c2f*), the optimization is very unstable and leads to noisy imagery; with the appearance branch conditioning on the ray direction (*w/ ray dir.*), the results exhibit abnormal colors under novel views. The quantitative and qualitative results are presented in Table 3 and Figure 13, respectively. Moreover, to offer a more intuitive understanding from another perspective, we conducted additional ablation experiments, wherein variants are created by progressively adding modules to a base model that purely consists of a canonical NeRF and motion networks. The

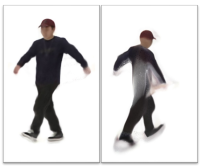

Figure 12: *Effect of partial observation during capture.*

Figure 9: *Each pair visualizes the dense correspondences derived from the backward motion flow between the canonical body (left) and the body reconstructed (right) at an observation frame.*

Figure 10: *Qualitative comparison with Video Avatar. Each pair shows the visual result of ours (left) and Video Avatar (right).*

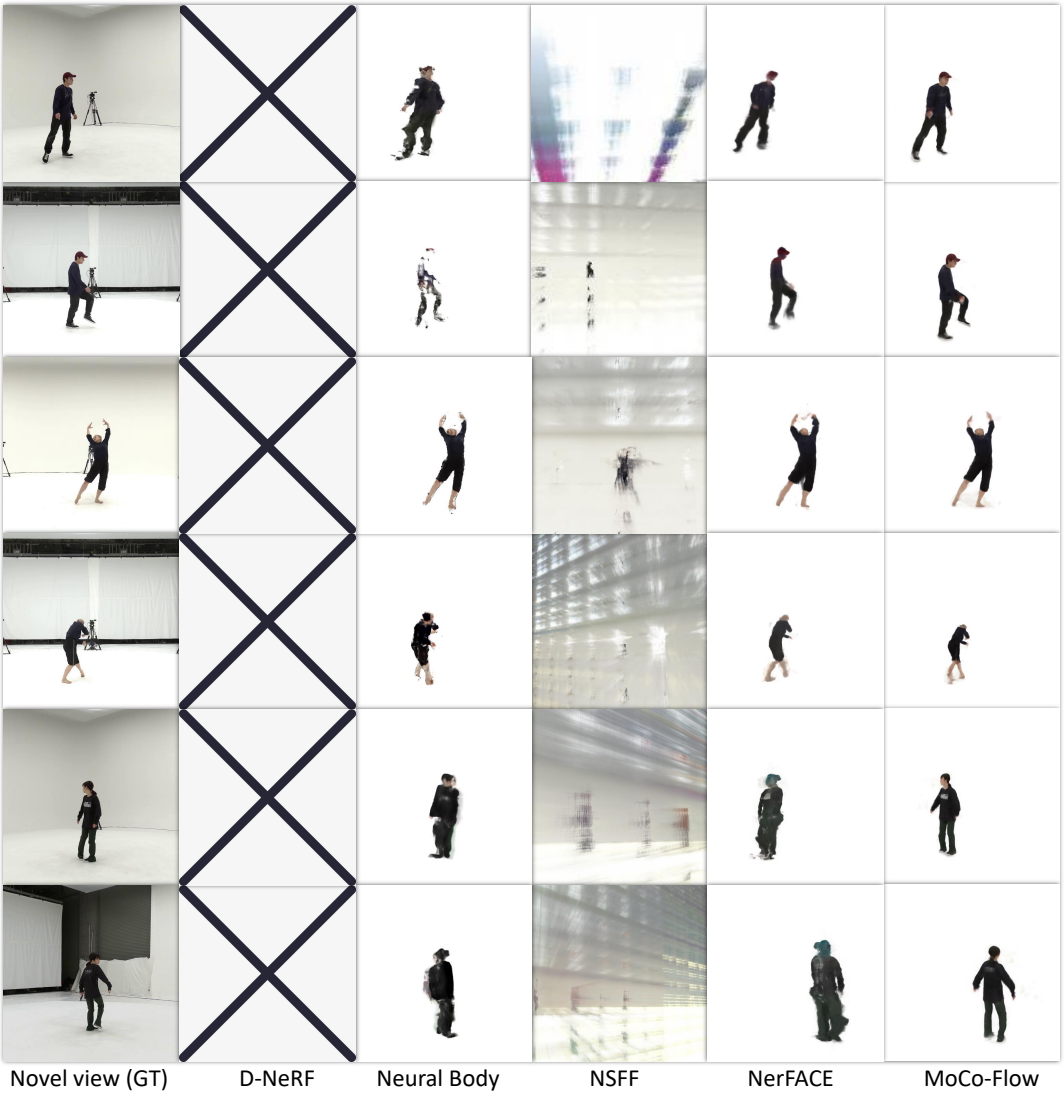

Figure 11: *Side-by-side visual comparisons. D-NeRF failed on the task and outputs blank imagery. Our method has the best visual quality and plausibility.*

quantitative results are presented in Table 4. Note the significant gain achieved when introducing our initialization with SMPL, motion consensus regularization, and adaptive volume.

We also investigate the performance of our method under the partial capture situation, wherein the camera does not fully capture the body. To this end, we simply cut out the forepart from the video of the male pop dancer and stop at where the dancer is about to take the spinning movement and show his back. We observed that our method failed to infer and complete the imagery of the missing regions on the back, producing erroneous colors due to the lack of observation on the back (see the right of the inset Figure 12). Nevertheless, our method is still able to produce correct imagery from the views near the front view of the dancer (see the left of the inset Figure 12).

5. Conclusion

We have presented a dynamic NeRF technique for synthesizing novel views of dynamic humans from stationary monocular cameras. Without the observation from various viewpoints to constrain the dynamics modeling, the problem is overly unconstrained and ambiguous. We address this problem with a carefully designed optimization scheme, which disambiguates bad local minima early at the initialization phase, and imposes a crucial regularization on the motion flow update to reach a high degree of *consensus* across the observations. This regularization has been demonstrated to be effective on preventing the optimization from deviating too much from the initialization and from landing on bad local minima.

Limitations and future work. There are, nevertheless, limitations to our method in its current form. As with any monocular method,

Table 3: *Quantitative results of removing each component from MoCo-Flow.*

	w/o $\mathcal{L}_{moco}$	w/o $\mathcal{L}_{moco}^{global}$	w/o $\mathcal{L}_{moco}^{local}$	w/o init.	w/o ada. vol.	w/o c2f	w/ ray dir.	MoCo-Flow
Plausibility ↑	0.659	0.865	0.851	0.000	0.826	0.726	0.940	**0.941**
OKS ↑	0.609	0.641	0.634	0.000	0.579	0.512	0.663	**0.670**

Figure 13: *Qualitative results of removing each component from MoCo-Flow.*

physical scale remains inherently ambiguous, thus, it is not guaranteed that our method can model the human body with correct physical scale (e.g., we observed that the modeled human has longer legs when the estimated mesh in VIBE output leans forward). Generally speaking, our method inherits erroneous estimation of the human body (VIBE's output in our case), and cannot compensate large errors in pose, location, and geometry of the human body. Moreover, the motion of self-occluded parts may not be correctly modeled, due to vanishing gradients from the photometric reconstruction loss. Nevertheless, with the rapid advances in learning human priors, we believe that the above issues will be alleviated by stronger priors. It would also be an interesting direction to explore how the proposed method can be extended to general objects if given proper priors. Lastly, as we represent the dynamics using per-coordinate dense motion flows, evaluating and optimizing the networks are computational intensive and time-consuming. In the future, we would like to explore the possibility of leveraging a hybrid model that brings together the explicit parametric model and neural-based implicit model, possibly using local implicits, to greatly reduce the optimization space and computation time. This would also be helpful for extending the proposed method to much longer video clips.

Table 4: *Quantitative results of progressively adding modules to a base model consisting of a canonical NeRF and motion networks.*

	Plausibility	OKS
Base	0.000	0.000
Base + c2f	0.000	0.000
Base + c2f + init.	0.618	0.486
Base + c2f + init. + moco	0.826	0.579
Base + c2f + init. + moco + ada. vol.	0.941	0.670

Acknowledgements

We thank the anonymous reviewers for their insightful comments and feedback. This work is supported in part by grants from the Joint NSFC-ISF Research Grant (62161146002), and gifts from Adobe Research.

References

[AMX*18] ALLDIECK T., MAGNOR M., XU W., THEOBALT C., PONS-MOLL G.: Video based reconstruction of 3d people models. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2018), pp. 8387–8397. 2, 3, 7, 9

[CCNS12] CUI Y., CHANG W., NÖLL T., STRICKER D.: Kinectavatar: fully automatic body capture using a single kinect. In *Asian Conference on Computer Vision* (2012), Springer, pp. 133–147. 3

[CCS*15] COLLET A., CHUANG M., SWEENEY P., GILLETT D., EVSEEV D., CALABRESE D., HOPPE H., KIRK A., SULLIVAN S.: High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG) 34*, 4 (2015), 1–13. 3

[CTMS03] CARRANZA J., THEOBALT C., MAGNOR M. A., SEIDEL H.-P.: Free-viewpoint video of human actors. *ACM Transactions on Graphics (TOG) 22*, 3 (2003), 569–577. 3

[DAST*08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. 1–10. 3

[DHT*00] DEBEVEC P., HAWKINS T., TCHOU C., DUIKER H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (2000), pp. 145–156. 3

[DKD*16] DOU M., KHAMIS S., DEGTYAREV Y., DAVIDSON P., FANELLO S. R., KOWDLE A., ESCOLANO S. O., RHEMANN C., KIM D., TAYLOR J., ET AL.: Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG) 35*, 4 (2016), 1–13. 3

[DZY*20] DU Y., ZHANG Y., YU H.-X., TENENBAUM J. B., WU J.: Neural radiance flow for 4d view synthesis and video processing. *arXiv preprint arXiv:2012.09790* (2020). 2, 3

[ERB*18] ESLAMI S. A., REZENDE D. J., BESSE F., VIOLA F., MORCOS A. S., GARNELO M., RUDERMAN A., RUSU A. A., DANIHELKA I., GREGOR K., ET AL.: Neural scene representation and rendering. *Science 360*, 6394 (2018), 1204–1210. 3

[FXTL17] FANG H.-S., XIE S., TAI Y.-W., LU C.: RMPE: Regional multi-person pose estimation. In *International Conference on Computer Vision (ICCV)* (2017). 8

[GLD*19] GUO K., LINCOLN P., DAVIDSON P., BUSCH J., YU X., WHALEN M., HARVEY G., ORTS-ESCOLANO S., PANDEY R., DOURGARIAN J., ET AL.: The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG) 38*, 6 (2019), 1–19. 3

[GSDA*09] GALL J., STOLL C., DE AGUIAR E., THEOBALT C., ROSENHAHN B., SEIDEL H.-P.: Motion capture using joint skeleton tracking and surface estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2009), IEEE, pp. 1746–1753. 3

[GTZN21] GAFNI G., THIES J., ZOLLHÖFER M., NIESSNER M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 2, 3, 9

[HGDG17] HE K., GKIOXARI G., DOLLÁR P., GIRSHICK R.: Mask r-cnn. In *International Conference on Computer Vision (ICCV)* (2017), pp. 2961–2969. 3, 5

[HSS*09] HASLER N., STOLL C., SUNKEL M., ROSENHAHN B., SEIDEL H.-P.: A statistical model of human pose and body shape. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 337–346. 3

[IKH*11] IZADI S., KIM D., HILLIGES O., MOLYNEAUX D., NEWCOMBE R., KOHLI P., SHOTTON J., HODGES S., FREEMAN D., DAVISON A., ET AL.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (2011), pp. 559–568. 3

[KAB20] KOCABAS M., ATHANASIOU N., BLACK M. J.: Vibe: Video inference for human body pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020). 3, 5

[KBJM18] KANAZAWA A., BLACK M. J., JACOBS D. W., MALIK J.: End-to-end recovery of human shape and pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 7122–7131. 3

[KER*18] KUMAR A., ESLAMI S. A., REZENDE D., GARNELO M., VIOLA F., LOCKHART E., SHANAHAN M.: Consistent jumpy predictions for videos and scenes. 3

[KPBD19] KOLOTOUROS N., PAVLAKOS G., BLACK M. J., DANIILIDIS K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)* (2019). 3

[LFB17] LEROY V., FRANCO J.-S., BOYER E.: Multi-view dynamic shape refinement using local temporal integration. In *International Conference on Computer Vision (ICCV)* (2017), pp. 3094–3103. 3

[LGL*20] LIU L., GU J., LIN K. Z., CHUA T.-S., THEOBALT C.: Neural sparse voxel fields. *Advances in Neural Information Processing Systems (NeurIPS)* (2020). 3, 5

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (SIGGRAPH Asia) 34*, 6 (Oct. 2015), 248:1–248:16. 3, 5

[LMTL21] LIN C.-H., MA W.-C., TORRALBA A., LUCEY S.: Barf: Bundle-adjusting neural radiance fields. *arXiv preprint arXiv:2104.06405* (2021). 3

[LMW21] LINDELL D. B., MARTEL J. N., WETZSTEIN G.: Autoint: Automatic integration for fast neural volume rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 3

[LNSW21] LI Z., NIKLAUS S., SNAVELY N., WANG O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 2, 3, 9

[LSS*19] LOMBARDI S., SIMON T., SARAGIH J., SCHWARTZ G., LEHRMANN A., SHEIKH Y.: Neural volumes: Learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG) 38*, 4 (July 2019), 65:1–65:14. 3

[LSS*21] LOMBARDI S., SIMON T., SCHWARTZ G., ZOLLHOEFER M., SHEIKH Y., SARAGIH J.: Mixture of volumetric primitives for efficient neural rendering. *arXiv preprint arXiv:2103.01954* (2021). 3, 5

[LSZ*21] LI T., SLAVCHEVA M., ZOLLHOEFER M., GREEN S., LASSNER C., KIM C., SCHMIDT T., LOVEGROVE S., GOESELE M., LV Z.: Neural 3d video synthesis. *arXiv preprint arXiv:2103.02597* (2021). 2, 3

[LVG*13] LI H., VOUGA E., GUDYM A., LUO L., BARRON J. T., GUSEV G.: 3d self-portraits. *ACM Transactions on Graphics (TOG) 32*, 6 (2013), 1–9. 3

[LWZ*18] LI J., WANG C., ZHU H., MAO Y., FANG H.-S., LU C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324* (2018). 8

[LYRK21] LI R., YANG S., ROSS D. A., KANAZAWA A.: Learn to dance with aist++: Music conditioned 3d dance generation. In *International Conference on Computer Vision (ICCV)* (2021). 7

[MBPY*18] MARTIN-BRUALLA R., PANDEY R., YANG S., PIDLYPENSKYI P., TAYLOR J., VALENTIN J., KHAMIS S., DAVIDSON P., TKACH A., LINCOLN P., ET AL.: Lookingood: Enhancing performance capture with real-time neural re-rendering. *ACM Transactions on Graphics (TOG)* (2018). 3

[MBRS*21] MARTIN-BRUALLA R., RADWAN N., SAJJADI M. S. M., BARRON J. T., DOSOVITSKIY A., DUCKWORTH D.: NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR* (2021). 3, 4

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)* (2020). 2, 3, 4, 5, 6

[NFS15] NEWCOMBE R. A., FOX D., SEITZ S. M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 343–352. 3

[NIH*11] NEWCOMBE R. A., IZADI S., HILLIGES O., MOLYNEAUX D., KIM D., DAVISON A. J., KOHI P., SHOTTON J., HODGES S., FITZGIBBON A.: Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality* (2011), IEEE, pp. 127–136. 3

[NSH*19] NATSUME R., SAITO S., HUANG Z., CHEN W., MA C., LI H., MORISHIMA S.: Siclope: Silhouette-based clothed people. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4480–4490. 3

[OERF*16] ORTS-ESCOLANO S., RHEMANN C., FANELLO S., CHANG W., KOWDLE A., DEGTYAREV Y., KIM D., DAVIDSON P. L., KHAMIS S., DOU M., ET AL.: Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (2016), pp. 741–754. 3

[PCPMMN21] PUMAROLA A., CORONA E., PONS-MOLL G., MORENO-NOGUER F.: D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 2, 3, 9

[PMR11] PONS-MOLL G., ROSENHAHN B.: Model-based pose estimation. In *Visual Analysis of Humans*. Springer, 2011, pp. 139–170. 3

[PSB*21] PARK K., SINHA U., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., SEITZ S. M., BRUALLA R.-M.: Deformable neural radiance fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 2, 3, 4, 5, 7

[PSH*21] PARK K., SINHA U., HEDMAN P., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., MARTIN-BRUALLA R., SEITZ S. M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics (TOG) 40*, 6 (dec 2021). 2

[PZX*21] PENG S., ZHANG Y., XU Y., WANG Q., SHUAI Q., BAO H., ZHOU X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 2, 3, 7, 9

[RJY*20] REBAIN D., JIANG W., YAZDANI S., LI K., YI K. M., TAGLIASACCHI A.: Derf: Decomposed radiance fields. *arXiv preprint arXiv:2011.12490* (2020). 3

[SFW*14] SHAPIRO A., FENG A., WANG R., LI H., BOLAS M., MEDIONI G., SUMA E.: Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds 25*, 3-4 (2014), 201–211. 3

[SGDA*10] STOLL C., GALL J., DE AGUIAR E., THRUN S., THEOBALT C.: Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics (TOG) 29*, 6 (2010), 1–10. 3

[SH07] STARCK J., HILTON A.: Surface capture for performance-based animation. *IEEE Computer Graphics and Applications 27*, 3 (2007), 21–31. 3

[SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)* (2019), pp. 2304–2314. 3

[SSSJ20] SAITO S., SIMON T., SARAGIH J., JOO H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 84–93. 3

[STH*19] SITZMANN V., THIES J., HEIDE F., NIESSNER M., WETZSTEIN G., ZOLLHOFER M.: Deepvoxels: Learning persistent 3d feature embeddings. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 2437–2446. 3

[SXZ*20] SU Z., XU L., ZHENG Z., YU T., LIU Y., ET AL.: Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *European Conference on Computer Vision (ECCV)* (2020), Springer. 3

[SZW19] SITZMANN V., ZOLLHÖFER M., WETZSTEIN G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NeurIPS)* (2019). 3

[TFHG19] TSUCHIDA S., FUKAYAMA S., HAMASAKI M., GOTO M.: Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019* (Delft, Netherlands, Nov. 2019), pp. 501–510. 2, 7

[TFT*20] TEWARI A., FRIED O., THIES J., SITZMANN V., LOMBARDI S., SUNKAVALLI K., MARTIN-BRUALLA R., SIMON T., SARAGIH J., NIESSNER M., ET AL.: State of the art on neural rendering. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 701–727. 3

[TSM*20] TANCIK M., SRINIVASAN P. P., MILDENHALL B., FRIDOVICH-KEIL S., RAGHAVAN N., SINGHAL U., RAMAMOORTHI R., BARRON J. T., NG R.: Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems (NeurIPS)* (2020). 5

[TTG*21] TRETSCHK E., TEWARI A., GOLYANIK V., ZOLLHÖFER M., LASSNER C., THEOBALT C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)* (2021), IEEE. 2, 3

[WPYS21] WIZADWONGSA S., PHONGTHAWEE P., YENPHRAPHAI J., SUWAJANAKORN S.: Nex: Real-time view synthesis with neural basis expansion. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 3

[WWHY20] WU M., WANG Y., HU Q., YU J.: Multi-view neural human rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 1682–1691. 3

[WWX*21] WANG Z., WU S., XIE W., CHEN M., PRISACARIU V. A.: Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064* (2021). 3

[XHKK20] XIAN W., HUANG J.-B., KOPF J., KIM C.: Space-time neural irradiance fields for free-viewpoint video. *arXiv preprint arXiv:2011.12950* (2020). 2, 3

[XLW*18] XIU Y., LI J., WANG H., FANG Y., LU C.: Pose Flow: Efficient online pose tracking. In *British Machine Vision Conference (BMVC)* (2018). 8

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 586–595. 7

[ZK14] ZHOU Q.-Y., KOLTUN V.: Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 1–10. 3

[ZRSK20] ZHANG K., RIEGLER G., SNAVELY N., KOLTUN V.: Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492* (2020). 3

[ZYW*19] ZHENG Z., YU T., WEI Y., DAI Q., LIU Y.: Deephuman: 3d human reconstruction from a single image. In *International Conference on Computer Vision (ICCV)* (2019), pp. 7739–7749. 3

[ZZCL13] ZENG M., ZHENG J., CHENG X., LIU X.: Templateless quasi-rigid shape modeling with implicit loop-closure. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2013), pp. 145–152. 3

Appendix A: Effect of the initialization errors

As discussed in the limitations, the misalignment occurs in our results as our method inherits erroneous estimation of the human body, and cannot compensate large errors in pose, location, and geometry of the human body. Moreover, we found that the VIBE tends to estimate human poses that are parallel to the image projection plane, which implies that the pose estimation is less accurate when the camera is not placed horizontally, thus the derived initialization leads to worse local minima (see Figure 14). So, we also show the

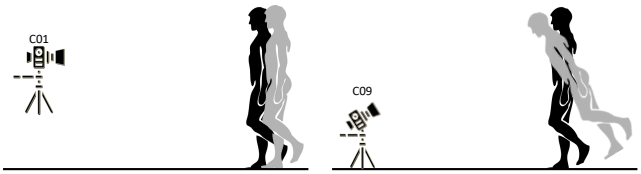

Figure 14: *Left: with the camera C01 in AIST, which is horizontally placed, the pose of the VIBE estimated mesh (in light gray color) is more accurate. But note that there is still a considerable amount of translation error; right: with the camera C09, which has a upward view, the pose of the VIBE estimated mesh has more rotation errors to the ground truth (in aterrimus).*

results of our algorithm trained on AIST camera C01 (AIST-C01 in short) data to demonstrate the performance of our method with different camera viewpoint and initialization (see Figure 15).

Appendix B: Additional implementation details

Optimization details We optimize the networks using the Adam optimizer with a learning rate linearly decayed by a factor of 0.9999 until the maximum number of iterations is reached. We sample 384 rays on a randomly selected image and sample 192 points (64 at the coarse level and 128 at the fine level) along each ray for each iteration of the optimization. The initialization takes N_1 iterations to converge, followed by N_2 iterations for the coarse-to-fine joint optimization (i.e., linearly anneal α from 0 to 8 over the iterations). We further keep α at 8 for N_3 iterations to fine-tune for more high-frequency details. More specifically, we set $(N_1, N_2, N_3) = (200K, 1500K, 1000K)$, $(N_1, N_2, N_3) = (500K, 1500K, 1000K)$, and $(N_1, N_2, N_3) = (800K, 2000K, 1500K)$ for People-snapshot, AIST, and ZJU-MoCap, respectively.

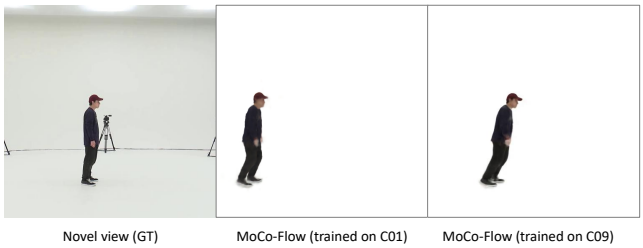

Figure 15: *Left: the ground truth view from the side; middle: training on AIST-C01 data produces imagery with more accurate human poses but larger translation errors; right: training on AIST-C09 data leads to imagery wherein the human body leans more forward.*

Test As discussed in the limitations, our method inherits the erroneous estimation of the human body from the VIBE output. Typically, we found that, on AIST dataset, VIBE often estimates human bodies with large pose and location errors. So, in order to synthesize more visually pleasing results on AIST camera C09 data (AIST-C09, in short), we further introduce a post-processing to manually adjust the orientation of the reconstructed dynamic scenes to be roughly upright. Note that the misalignment to the GT would still exist even with this rough re-orientation post-processing.

Baseline details The results of D-NeRF, Neural Body, and NSFF are obtained with their released code. As for NerFACE, since it is highly specialized for faces and takes as input the face parameters, we implement it in our framework to work on SMPLs. More concretely, instead of rigidly transforming the whole observation-space volume with the estimated face parameter, we convert only the points that are near to the SMPL estimate within a distance threshold (0.2m in our implementation) using the transformation matrix of its nearest vertex on SMPL. Then, as described in their paper, the transformed point, a learnable code, and the corresponding SMPL parameters are fed into the neural radiance field for training. NSFF only supports reconstruction in NDC space, it is non-trivial to adapt it to work on non-NDC space, which is also mentioned in their official code repository. Although the novel view test results of NSFF are obtained in an approximate way where the learned NDC cubic space is scaled to fit the target physical volume, we highlight that it is the inability to model the target volume from a single stationary view that accounts for the improper content that is revealed to be unmeaning at novel views.

Last, it is rather easy to configure the cameras, for all methods, in our setting. All cameras are stationary and hence are set to be aligned with a world coordinate system. Near and far planes are set correctly in baselines following their description and instructions, so samplings are concentrated in a proper volume.

Appendix C: Datasets

We present more details of data processing: *(A) People-snapshot*: each video in this dataset lasts around 10 seconds. We downsample each original video at 24 frames per-second (FPS) to obtain a video at 12 FPS, resulting in around 115 frames in total for each video input. Since a static background image is not available for each video captured in this dataset, we simply mask out the background to be white using the provided foreground mask. *(B) AIST*: we clip out several video clips from the original videos, each lasts around 6-10 seconds. We then downsample each original clip at 60 FPS to obtain a video at 12 FPS, resulting in around 80-120 frames in total for each video input. To obtain the static background image for each input video, we set the color at each location of the background image to be the median value of this location across the whole video clip. To remove the considerable amount of shadows contained in this dataset, we run Mask R-CNN detection to obtain the human segmentation of each frame, which is then composited with the static background image to obtain the final image. *(C) ZJU-MoCap*: each video in this dataset lasts around 10 seconds. We downsample each original video at 24 frames per-second (FPS) to obtain a video at 12 FPS, resulting in around 150 frames in total for each video input.

The static background image is also obtained via the background image extraction as is done in AIST.