Rotationally-Temporally Consistent Novel View Synthesis of Human Performance Video

Youngjoong Kwon[1,2(✉)], Stefano Petrangeli[2], Dahun Kim[3], Haoliang Wang[2], Eunbyung Park[1], Viswanathan Swaminathan[2], and Henry Fuchs[1]

[1] University of North Carolina at Chapel Hill, Chapel Hill, USA
youngjoong@cs.unc.edu
[2] Adobe Research, San Jose, USA
[3] Korea Advanced Institute of Science and Technology, Daejeon, South Korea

Abstract. Novel view *video* synthesis aims to synthesize novel viewpoints videos given input captures of a human performance taken from multiple reference viewpoints and over consecutive time steps. Despite great advances in model-free novel view synthesis, existing methods present three limitations when applied to complex and time-varying human performance. First, these methods (and related datasets) mainly consider simple and symmetric objects. Second, they do not enforce explicit consistency across generated views. Third, they focus on static and non-moving objects. The fine-grained details of a human subject can therefore suffer from inconsistencies when synthesized across different viewpoints or time steps. To tackle these challenges, we introduce a human-specific framework that employs a learned 3D-aware representation. Specifically, we first introduce a novel siamese network that employs a gating layer for better reconstruction of the latent volumetric representation and, consequently, final visual results. Moreover, features from consecutive time steps are shared inside the network to improve temporal consistency. Second, we introduce a novel loss to explicitly enforce consistency across generated views both in *space* and in *time*. Third, we present the Multi-View Human Action (MVHA) dataset, consisting of near 1200 synthetic human performance captured from 54 viewpoints. Experiments on the MVHA, Pose-Varying Human Model and ShapeNet datasets show that our method outperforms the state-of-the-art baselines both in view generation quality and spatio-temporal consistency.

Keywords: Novel view video synthesis · Synthetic human dataset

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58548-8_23) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12349, pp. 387–402, 2020.
https://doi.org/10.1007/978-3-030-58548-8_23

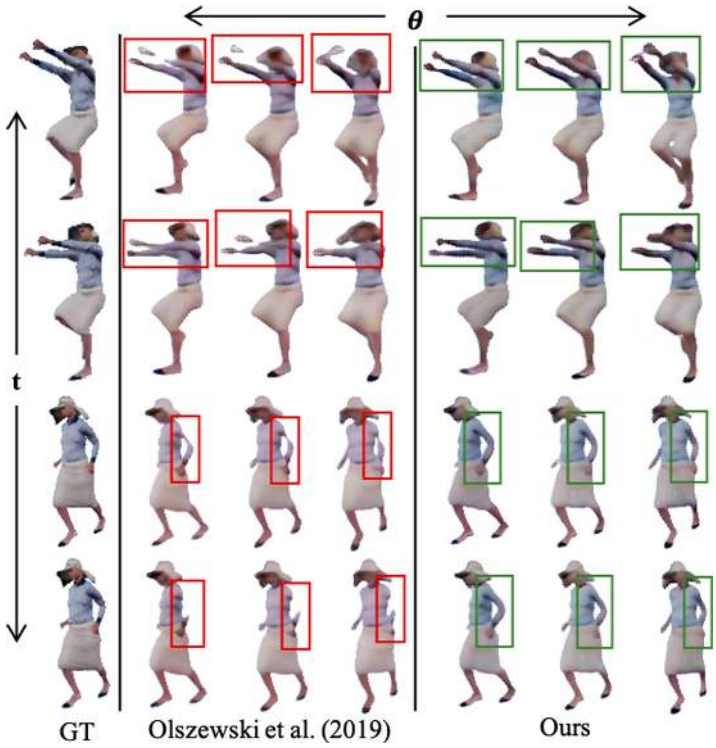

Fig. 1. Our method combines a siamese network architecture and rotational-temporal supervision for higher quality novel view video generation of human performance. Compared to Olszewski et al. [18], ours generate higher quality and more consistent results across views (left-right) and time steps (top-down). Model used for training with an unseen animation sequence, 4 input views.

1 Introduction

Novel View Synthesis (NVS) aims to synthesize new views of an object given different known viewpoints. Recently, a number of learning-based approaches have enabled the view synthesis by direct image or video generation without explicit 3D reconstruction or supervision [18,25]. Applying high-quality, accurate novel view synthesis to human action performance *videos* has a variety of applications in the area of AR/VR, telepresense, volumetric videos, and so on. Existing approaches present three shortcomings when applied to the human novel view video synthesis task. First, they focus on objects with simple shapes and strong symmetry, and perform quite poorly on deformable and asymmetric shapes like the human body. Second, current NVS methods do not enforce explicit consistency between different generated viewpoints, which does not guarantee consistency among generated views (Fig. 1, left). Third, current NVS methods focus on static objects, while human motion cannot be modeled by simple rigid transformations of existing (latent) volumetric representations. Dynamic articulations of

body parts, like limbs and heads, can therefore suffer from significant inconsistencies when synthesizing novel views over time (Fig. 1, left). Moreover, existing NVS datasets are either not designed for the human NVS task [1], are too limited to support learning-based approaches [8], or are not publicly available [6,18]. In this paper, we therefore focus on synthesizing a temporally and spatially consistent novel view *video* of a human performance captured over time from fixed viewpoints. In particular, we present a novel end-to-end trainable video NVS network, combined with effective rotational and temporal consistency supervisions, and a synthetic dataset to further support research in this domain.

Contributions. Our model is based on recent NVS methods using latent volumetric representations that can be rendered by a learnable decoder [6,18,25]. It consists of a pair of siamese encoder-decoder networks, each receiving the input RGB video frames of the human performance captured from multiple viewpoints from two consecutive time steps. The temporal features of the inputs are shared between the two networks in order to enhance the novel view reconstruction video quality and temporal consistency. We also present a novel volume gating layer to improve the latent volumetric representation by adaptively attending on the valid volume points only when filling in missing parts during the novel view reconstruction. Moreover, we explicitly enforce rotational and temporal consistency across generated views to provide superior reconstruction performance, and demonstrate the effectiveness of this approach in capturing the complexity of human motion across different viewpoints and time steps. It is worth noting that the proposed rotational consistency supervision is applicable to image-level synthesis and non-human objects as well. Indeed, our approach evaluated on the ShapeNet [1] cars and chairs categories achieves state-of-the-art performance. Finally, we collect and publicly release the synthetic Multi-View Human Action (MVHA) dataset, composed of 30 different 3D human models animated with 40 different Mocap sequences, captured from 54 different viewpoints. Results on our MVHA, ShapeNet [1] and Pose-Varying Human Model (PVHM) [37] datasets confirm both quantitatively and qualitatively the superior performance of the proposed approach compared to start-of-the-art baselines for the NVS task.

The remainder of this paper is structured as follows. Section 2 presents related works, while Sect. 3 presents the network architecture, volume gating convolutions, temporal feature augmentation, and the MVHA dataset. Section 4 reports quantitative and qualitative results, while Sect. 5 concludes the paper.

2 Related Work

In this section, we review prior works in the areas of 2D-based novel view synthesis, 3D-based novel view synthesis, and existing datasets available for the human NVS task.

2D-Based Novel View Synthesis. Tatarchenko *et al.* and Yang *et al.* propose to synthesize novel views by regressing the pixel colors of the target view directly from the input image using a Convolutional Neural Network (CNN) [28,35]. Instead of starting from an empty state, pixel-flow based approaches leverage

pixel-flow to generate high-quality, sharp results [19,27,36]. These works usually use flow prediction to directly sample input pixels to reconstruct the output view. Zhou *et al.* suggest moving pixels from an input to a target view leveraging bilinear sampling kernels [9,36]. The approach by Park *et al.* achieves high-quality results by moving only the pixels that can be seen in the novel view, and by hallucinating the empty parts using a completion network [19]. Park *et al.* also take advantage of the symmetry of objects from ShapeNet [1] by producing a symmetry-aware visibility map [19]. Our human novel view synthesis task cannot fully take advantage of this approach since humans can present highly asymmetric poses. Sun *et al.* further improve these results by aggregating an arbitrary number of input images [27]. Eslami *et al.* use a latent representation that can aggregate multiple input views, which shows good results on synthetic geometric scenes [3]. Unlike the previous NVS works that move or regress pixels, Shysheya *et al.* regress texture coordinates corresponding to a pre-defined texture map [24].

3D-Based Novel View Synthesis. Works embedding implicit spatial consistency in the NVS task using explicit or latent volumetric representations have shown promising reconstruction results. Several recent methods reconstruct an explicit occupancy volume from a single image, and render it using traditional rendering techniques [2,5,10,22,29,32–34]. Methods leveraging signed-distance-field-encoded volumes [20,23], or RGBα-encoded volumes [12] have achieved excellent quality while overcoming the memory limitations of voxel-based representations. Saito *et al.* predict the continuous inside/outside probability of a clothed human, and also infer an RGB value at given 3D positions of the surface geometry, resulting in a successful recovery of intricate details of garments [23]. We do not compare to these methods [20,23] as they require ground truth geometry for the supervision [20,23], and as Lombardi *et al.* does not support generalization to unseen subjects [12]. Rather than generating explicit occupancy volumes, several methods generate latent volumetric representations that can be rendered by a learnable decoder [10,15,16,18,22,25,26]. Sitzmann *et al.* introduce a persistent 3D feature embedding to address the inconsistency between views synthesized by generative networks, which can occur due to a lack of 3D understanding [25,26]. Olszewski *et al.* generate a latent volumetric representation that allows 3D transformations and a combination of different input view images [18]. Moreover, their network does not require any 3D supervision and produces state-of-the-art results for the NVS task on ShapeNet. The main difference between our work and those by Sitzmann *et al.* [25,26] is that we strengthen the spatial consistency by introducing explicit rotational consistency supervision, and by also introducing implicit and explicit temporal consistency to better cope with the human novel view video synthesis task, while Sitzmann *et al.* do not consider temporal aspect.

Datasets for Human Novel View Synthesis. The Human3.6M dataset provides 3.6 million human poses and corresponding images from 4 calibrated cameras [8]. Collecting these datasets requires complex setup with multiple cameras, which is expensive and time-consuming. To address this limitation, synthetic

methods have been proposed. The SURREAL dataset provides a large number of images generated with SMPL body shapes and synthetic textures, together with body masks, optical flow and depth [13,31]. The Pose-Varying Human Model (PVHM) dataset provides RGB images, depth maps, and optical flows of 22 models [37]. The 3DPeople dataset contains 80 3D models of dressed humans performing 70 different motions, captured from 4 different camera viewpoint [21]. Many of the existing learning-based NVS works are built using the ShapeNet dataset [1], which provides 54 reference viewpoints. In contrast, most of the aforementioned human datasets only provide data from relatively few reference viewpoints, preventing existing NVS approaches to be directly applied to the human NVS tasks. The MVHA synthetic dataset we collect contains 30 3D models animated with 40 Mocap sequences each rendered from 54 viewpoints. Compared to the PVHM dataset, we provide much more diverse outfits (*e.g.*, short/long sleeve, pants, and skirts, different types of hats and glasses, etc.), complex motion sequences, and higher resolution images.

3 Proposed Method

3.1 Problem Definition

A human performance captured from view k, out of K available views, consists of T consecutive RGB frames $I^k_{1:T} := \{I^k_1, I^k_2, ..., I^k_T\}$. Given a set of input $I^k_{1:T}$ with $k = 1, \ldots, K$, our goal is to directly synthesize a novel view video $\hat{I}^q_{1:T}$ so that: 1) given time step t, views generated across different query viewpoints q are consistent among each other and 2) given query q, temporally consecutive frames are consistent among each other. We simplify the problem of optimizing $p(\hat{I}^q_{1:T}|I^k_{1:T})$ by factorizing the conditional distribution to a product form:

$$p(\hat{I}^q_{1:T}|I^k_{1:T}) = \prod_{t=1}^{T} p(\hat{I}^q_t|I^k_t, I^k_{t-1}). \tag{1}$$

In our experiments, we sample two consecutive frames for each network feed-forward. During training, we augment the supervision signal using the symmetry between time steps t and $t-1$ by learning $p(\hat{I}^q_t, \hat{I}^q_{t-1}|I^k_t, I^k_{t-1})$. The volume used in our paper is centered on the target object and its axis is aligned with the camera coordinate. Perspective effects caused by pinhole camera projection and camera intrinsic parameters are approximately learned by the encoder and decoder networks, rather than handled explicitly.

3.2 Network Architecture

Inspired by 3D structure-aware view synthesis pipelines [10,17,18,22,25], input pixels I^k_τ (with $\tau = t, t-1$) are transformed from 2D-to-3D and then 3D-to-2D throughout the layers to be mapped onto the target view pixels I^q_τ. An overview of the proposed architecture design is given in Fig. 2. Our model is a two-tower

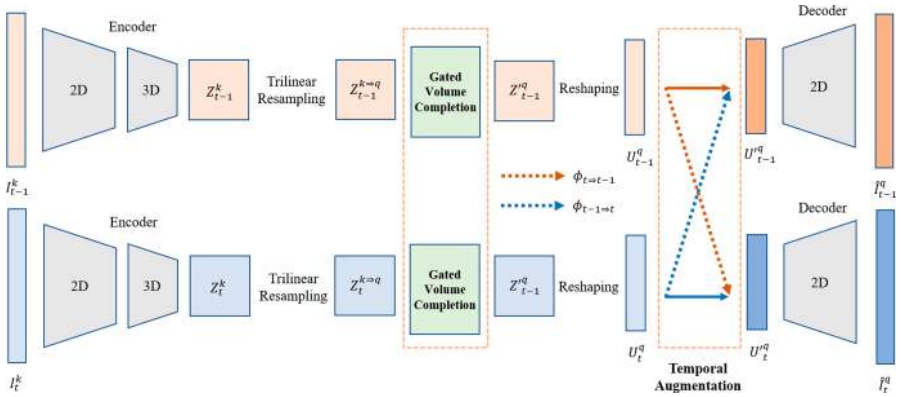

Fig. 2. Our network is a two-tower siamese network consisting of four blocks: encoder, volume completion, temporal augmentation, and decoder.

siamese network consisting of four blocks: encoder, volume completion, temporal augmentation, and decoder. For each tower, the encoder E takes I_τ^k as input, and generates a volumetric representation Z_τ^k of I_τ^k through a series of 2D-conv, reshape and 3D-conv layers. The 3D representation is then transformed to a target view $Z_\tau^{k\to q}$ via trilinear resampling. Multiple input views can be used by adding the representation over all k, $Z_\tau^q = \sum_k Z_\tau^{k\to q}$. These 3D features are fed into the proposed volume completion and are reshaped along the depth axis to become 2D features U_τ^q. U_τ^q then become $U_\tau'^q$ after going through the proposed temporal augmentation. The decoder D generates the output image I_τ^q starting from $U_\tau'^q$. The siamese network is coupled with two input-output pairs from consecutive time steps as $I_t^q, I_{t-1}^q = f(I_t^k, I_{t-1}^k)$, and the two towers are connected at the temporal fusion module.

Gated Volume Completion. The volumetric representation is initialized with the visible volume points generated from the source views I_τ^k (with $\tau = t, t-1$). The remaining *unseen* volume points should be hallucinated so as to generate plausible target views when rendered by the decoder. Our volume completion module consists of a series of 3D convolutions to inpaint the missing volume points. In vanilla 3D convolutional layers, all feature points are treated as the same valid ones, which is appropriate for tasks with complete inputs such as 3D object detection. In the presence of *empty* voxels in our completion problem, however, it is ambiguous whether current locations belong to the foreground voxels that should be hallucinated, or to the background that must remain unchanged. Vanilla 3D convolutions apply the same filters on all seen and unseen foreground, background and mixed voxels/features, leading to visual artifacts such as color discrepancy, blurriness and omission of shape details when synthesized to a target view by the decoder (Fig. 1).

To address this problem, we propose the volume gating convolutions to improve the latent volumetric representation computed by the network. The

volume gating convolution learns a dynamic feature gating mechanism for each spatial location, *e.g.*, foreground or background voxels, seen or unseen voxels. Together, the gating operation can properly handle the uncertainty of voxel occupancy. Such explicit decoupling of the foreground object is also necessary to deal with real images with arbitrary background, which should remain unchanged in any possible target views. Specifically, we consider the formulation where the input 3D features are firstly used to compute gating values $g^{k\to q} \in \mathbb{R}^{H\times W\times D}$ (with H, W, and D being the width, height, and depth of the latent volumetric representation). By construction, however, the employed 3D volume is memory inefficient, *i.e.*, $O(n^3)$ w.r.t. the resolution, which is another limitation in creating the voxel-wise gating values [25]. Thus, we propose that full 3D gating may be conveniently approximated by decomposing it into 2D spatial gatings, *i.e.*, $O(n^2)$, along three canonical directions, *i.e.*, $g^{HW_{k\to q}} \in \mathbb{R}^{H\times W\times 1}$, $g^{HD_{k\to q}} \in \mathbb{R}^{H\times 1\times D}$, and $g^{WD_{k\to q}} \in \mathbb{R}^{1\times W\times D}$. It is worth noting that the gates depend only on input and query viewpoints k and q, and not on the timestep τ.

To obtain the 2D gating values, we first average-pool the input volume features $Z_\tau^{k\to q}$ along one spatial dimension (we will omit the superscript $k \to q$ for the remaining of this paragraph for ease of notation). For example, $Z_\tau^{HW} \in \mathbb{R}^{C\times H\times W\times 1}$ is obtained by average-pooling along the depth axis. We then apply average-pooling and max-pooling operations along the channel axis and concatenate them to generate an efficient feature descriptor $F_\tau^{HW} \in \mathbb{R}^{H\times W\times 2}$. The 2D gating is generated by applying a 5×5 convolution on the concatenated feature as $g^{HW} = \sigma(w(F_\tau^{HW})) \in \mathbb{R}^{H\times W\times 1}$. g^{HD} and g^{WD} are computed similarly and each of them achieves canonical-view volume carving. The full 3D gating is a geometrical mean of the three 2D gating values with shape broadcasting as $g = (g^{HW} \odot g^{HD} \odot g^{WD})^{1/3} \in \mathbb{R}^{H\times W\times D}$, which encodes *where* in HWD volume space to emphasize or suppress. The final output is a multiplication of the learned feature and gating value $Z_\tau'^{k\to q} = Z_\tau^{k\to q} \odot g^{k\to q}$, where g is copied along the channel axis. Our proposed volume gating is memory-efficient, easy to implement and performs significantly better at correcting color discrepancy and missing shape details in the generated views (see Sect. 4).

Temporal Feature Augmentation. Up to the completion of the 3D features, our siamese network treats human performance at each time step separately, producing $Z_{t-1}'^q$ and $Z_t'^q$ from each tower independently. These 3D features are then reshaped along the depth dimension to become frontal-view 2D features U_{t-1}^q and U_t^q with respect to the query viewpoint. At this point, we propose a temporal feature augmentation module to leverage complementary information between $t-1$ and t. First, this approach allows us to improve the per-frame quality synthesis, since consecutive time steps might reveal more visible pixels of the same model and, therefore, leads to better occlusion handling. Second, temporal coherency is greatly improved as the generation of the final output is conditioned on both consecutive time steps.

Given the projected 2D features U_{t-1}^q and U_t^q, our temporal augmentation module learns the flow warping to align U_{t-1}^q onto U_t^q. The flow submodule receives an initial optical flow $\phi_{t\to t-1}^{init}$ computed by FlowNet2 [7], and refine it

for more accurate deep feature flow. The warped feature $\hat{U}^q_{t-1}$ is then multiplied with a non-occlusion mask $M_{t\to t-1}$ calculated from the warping error as $e^{-\alpha\|I_t - W_{t\to t-1}(I_{t-1})\|_2^2}$, and then element-wise summed with the current features as $U'^q_t = U^q_t + M_{t\to t-1}\cdot\hat{U}^q_{t-1}$. The above operations are designed to be performed in a temporally bi-directional manner, *i.e.*, from $t-1$ to t and vice versa. Once U'^q_{t-1} and U'^q_t are obtained, they are fed into the decoders to generate the two consecutive frames of the query viewpoint $\hat{I}^q_{t-1}$ and $\hat{I}^q_t$.

3.3 Rotational and Temporal Supervision

Our goal is to generate novel view videos of a human performance that is consistent across query viewpoints and time steps, which cannot be guaranteed by simply providing multiple input viewpoints at each time step. To this end, we design a loss function that can explicitly enforce both rotational and temporal consistency in the generated views. It is worth noting that, during training, our network generates additional target views in addition to the query viewpoint q that are used to enforce rotational consistency. At testing time instead, only the query viewpoint is generated.

Query Loss. We first calculate the reconstruction loss on the individual query view (*e.g.*, between I^q_τ and $\hat{I}^q_\tau$) as follows:

$$\mathcal{L}_{query} = \lambda_R\mathcal{L}_R + \lambda_P\mathcal{L}_P + \lambda_S\mathcal{L}_S + \lambda_A\mathcal{L}_A \tag{2}$$

where $\mathcal{L}_R$ denotes the L_1 reconstruction loss, $\mathcal{L}_P$ is the L_2 loss in the feature space of the VGG-19 network, $\mathcal{L}_S$ is the SSIM loss, $\mathcal{L}_A$ is the adversarial loss calculated using the discriminator architecture from Tulyakov *et al.* [30].

Rotational Consistency Loss. To improve consistency, we let the network generate two additional target views $l \in L_a$ immediately adjacent to the query view. These additional views help the network to synthesize complex shapes, by providing additional information during the training process of the target view q on the actual shape of the human subject, and are generated from the same volumetric representation Z^k_τ. Similarly as for the query viewpoint, we compute the reconstruction loss $\mathcal{L}_{rot1}$ for each additional adjacent view (*e.g.*, between I^l_τ and $\hat{I}^l_\tau$) in the same way as in Eq. 2 (excluding term $\mathcal{L}_P$ in this case).

Next, to minimize inconsistencies between synthesized views, we consider the warping error between the query view and the adjacent views. We compute $W_{l\to q}$, the function warping an image according to the backward flow between the ground-truth images I^q_τ and I^l_τ computed by FlowNet2 [7]. Next, we compute $M_{l\to q}$, the binary occlusion mask between the I^l_τ and $W_{l\to q}(I^q_\tau)$ as $M_{l\to q} = e^{-\alpha\|I^l_\tau - W_{l\to q}(I^q_\tau)\|_2^2}$ [11]. We use a bi-linear sampling layer to warp images and set $\alpha = 50$ (with pixel range between $[0, 1]$) [9]. We apply this occlusion mask to both $\hat{I}^l_\tau$ (the additional generated view) and the warped generated query view $W_{l\to q}(\hat{I}^q_\tau)$, and compute the warping loss as follows:

$$\mathcal{L}_{rot2} = \left\| M_{l\to q} * (\hat{I}^{l}_{\tau} - W_{l\to q}(\hat{I}^{q}_{\tau})) \right\|_1 \tag{3}$$

The final rotational loss is given by:

$$\mathcal{L}_{rot} = \frac{1}{|L_a|} \sum_{l\in L_a} \mathcal{L}_{rot1} + \lambda_R \mathcal{L}_{rot2} \tag{4}$$

Temporal Consistency Loss. The temporal loss is calculated as the sum of the bi-directional warping errors (*e.g.*, from $t-1 \to t$ and $t \to t-1$) between the generated query images $\hat{I}^{q}_{t-1}$ and $\hat{I}^{q}_{t}$. First, we calculate $W_{t\to t-1}$, the function warping the ground-truth image I^{q}_{t-1} towards I^{q}_{t}, computed by FlowNet2 [7]. Next, we calculate the binary occlusion mask $M_{t\to t-1}$, and apply it to the generated view at time t, $\hat{I}^{q}_{t}$, and the warped generated view at time $t-1$, $W_{t\to t-1}(\hat{I}^{q}_{t-1})$. We repeat the same calculation for the warping error $t \to t-1$ and compute the final temporal loss as follows:

$$\mathcal{L}_{temp} = \left\| M_{t\to t-1} * (\hat{I}^{q}_{t} - W_{t\to t-1}(\hat{I}^{q}_{t-1})) \right\|_1 + \left\| M_{t-1\to t} * (\hat{I}^{q}_{t-1} - W_{t-1\to t}(\hat{I}^{q}_{t})) \right\|_1 \tag{5}$$

Overall Loss. The overall training loss $\mathcal{L}_{tot}$ is given by:

$$\mathcal{L}_{Total} = \frac{1}{|\{t, t-1\}|} \Big(\sum_{\tau\in\{t,t-1\}} (\lambda_{query}\mathcal{L}_{query} + \lambda_{rot}\mathcal{L}_{rot}) + \lambda_{temp}\mathcal{L}_{temp} \Big), \tag{6}$$

where $t-1$ and t are the two consecutive time steps. The weights λ_{query}, λ_{rot}, λ_{temp} are set to 2, 1, 1, respectively.

3.4 Multi-View Human Action (MVHA) Dataset

In order to support the development of learning-based NVS solutions that are applicable to human performance, we introduce the Multi-View Human Action (MVHA) dataset. Compared with previous similar datasets that only provide 4 captured viewpoints [8], ours provide 54 different viewpoints for each unique model (18 azimuths and 3 elevations). Moreover, our dataset is not composed of static captures, but of synthetic human subjects moving in extremely diverse modality. The detailed description and samples of the MVHA dataset can be found in the supplementary material.

Body and Clothing Models. We generate fully textured meshes for 30 human characters using Adobe Fuse [4]. The distribution of the subjects' physical characteristics covers a broad spectrum of body shapes, skin tones, outfits and hair geometry. Each subject is dressed in a different outfit including a variety of garments, combining tight and loose clothes.

Table 1. Quantitative results (with 4 input views) on the MVHA dataset for the ablation study (1a) and comparison with TBN approach [18] under different testing scenarios (1b). Arrows indicate the direction of improvement for the metric under consideration.

Method	$L_1 \downarrow$	SSIM↑	RL↓	TL↓
Base	.0825	.7364	.0433	.0280
Base+RC	.0736	.7541	.0390	.0163
Base+TC	.0743	.7506	.0414	.0149
Base+RC+TC	**.0722**	**.7596**	**.0370**	**.0146**

(a) Ablation study (seen model/unseen pose)

Testing Scenario	Method	$L1 \downarrow$	SSIM↑	RL↓	TL↓
1. Seen model	TBN [18]	.1068	.7072	.0541	.0303
+ Unseen pose	Ours	**.0722**	**.7596**	**.0370**	**.0131**
2. Unseen model	TBN [18]	.0982	.7109	.0504	0.024
+ Seen pose	Ours	**.0806**	**.7438**	**.0458**	**.0146**
3. Unseen model	TBN [18]	.1130	.6867	.0564	.0282
+ Unseen pose	Ours	**.0863**	**.7279**	**.0480**	**.0150**

(b) Results and Generalization Capabilities

Mocap Sequences. We gather 40 realistic motion sequences from Adobe Mixamo [14]. These sequences include human movements with different complexity, from relatively static body motions (*e.g.*, standing) to very complex and dynamic motion patterns (*e.g.*, break-dance or punching).

Camera, Lights and Background. A 3D rendering software is used to apply different Mocap animation sequences to the 30 3D models. The illumination is composed of an ambient light plus a directional light source. We use a projective camera with 512×512 pixel resolution. The distance to the subject is fixed to ensure the whole body is in view at all times. Every sequence is rendered from 54 camera views, 18 azimuths at 20-degree intervals and 3 elevations at 10-degree intervals. For every rendered view and time step, we provide the final RGB image and associated binary segmentation mask. Custom background can be added by taking advantage of the released segmentation masks.

4 Experiments

In this section, we present the results in terms of view generation quality, and rotational and temporal consistency. We present quantitative and qualitative results on our MVHA dataset, as well as the PVHM and ShapeNet datasets [1,37]. To evaluate the view generation quality, we use the L_1 and SSIM scores between the generated and ground truth views. The consistencies between adjacent generated views and between consecutive time steps are evaluated using the Rotational Loss (RL), Eq. 4, and Temporal Loss (TL), Eq. 5, respectively. Unless otherwise stated, all results are reported for models not used during training. The details on the dataset splits, training process, additional results and video results are available in the supplementary material.

4.1 Results on the MVHA Dataset

In this section, we present the performance of the proposed approach for the MVHA dataset. Note that we used 18 azimuths with a fixed elevation throughout the experiments. We investigate the ablative impact of the proposed rotational

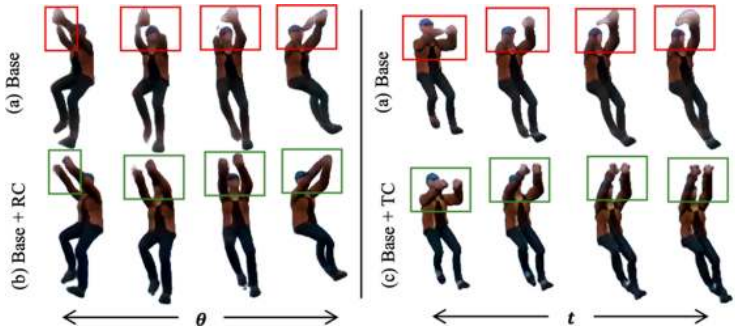

Fig. 3. Ablation study (left: fixed time step, varying query viewpoint; right: fixed query viewpoint, varying time step). Rotational and temporal supervision remarkably improve quality and consistency of the novel view synthesis, both across query viewpoints (left) and time steps (right). Model seen during training, unseen pose, 4 input views.

and temporal supervisions, the generalization ability of our method to unseen models and poses, and provide a visual analysis of the learned volume gating.

Ablation Study. The ablation study investigates the impact of the proposed Rotational Consistency (RC) and Temporal Consistency (TC) on the novel view synthesis quality, both quantitatively (Table 1a) and qualitatively (Fig. 3). Both methods improve the baseline performance and significantly reduce the rotational (Table 1a, rows 1–2) and temporal (Table 1a, rows 1–3) warping errors in the final results. Visual results confirm the gains brought by the proposed supervisions (Fig. 3). In addition, we find that rotational and temporal supervisions have a complementary contribution to the per-frame generation quality, each improving the baseline. With both RC and TC together, we achieve the best performances overall, across all metrics (Table 1a, row 4).

Visualization of Learned Volume Gating. As presented in Sect. 3.2, we allow the network to learn the volumetric mask g automatically. Memory-intensive 3D gating operations are decomposed into three 2D gatings along the depth, height and width axes of the latent volumetric representation. We visualize these learned gating values g_{HW}, g_{WD} and g_{HD} in Fig. 5. We observe the gating masks have different values at each spatial location, especially based on whether the current location is on the foreground or not. More interestingly, the three gating masks *attend to* the foreground shape captured at each corresponding canonical view. This also implies that our learned volumetric representation is indeed aware of 3D structure, and consequently, the learned volume gating layer can achieve soft volume carving.

4.2 Results on PVHM and ShapeNet Datasets

We compare our view generation quality with diverse well-known NVS methods [6,18,19,28,36,37]. We first compare the performance of these methods

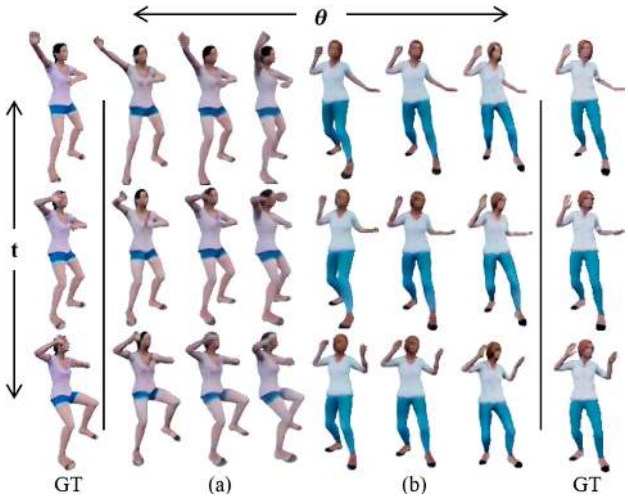

Fig. 4. Our method generates high-quality visual results in case of models unseen during training, both for seen poses (a) and unseen poses (b) (4 input views).

Table 2. Quantitative results on the PVHM dataset [37], for our method and several baseline methods.

Method	MSE↓	SSIM↑
Tatarchenko *et al.* [28]	96.83	.9488
Zhou *et al.* [36]	131.6	.9527
Park *et al.* [19]	85.35	.9519
Zhu *et al.* [37]	72.86	.9670
Huang *et al.* [6]	89.44	.9301
Olszewski *et al.* [18]	70.34	.9695
Ours	**61.68**	**.9807**

on the PVHM dataset [37], to show that our approach can produce superior results on a different dataset with similar characteristics as the one we collected (Table 2). We use the Mean Squared Error (MSE) and Structural Similarity Index (SSIM) to quantitatively compare the different approaches, as done by Zhu *et al.* [37]. Our method produces the best results overall, for both metrics (Table 2). We also show a qualitative comparison with some of the most recent human-specific methods [6,18,37] in Fig. 6. Zhu *et al.* [37] show artifacts on the face region, due to the failure of transferring pixels from visible parts. Huang *et al.* [6] cannot successfully recover the whole body shape. Olszewski *et al.* [18] show incomplete reconstruction of the arm and noticeable color alterations. Our method is able to reconstruct fine-grained details, including limbs and wrinkles, with higher fidelity with respect to the original colors. Both qualitative

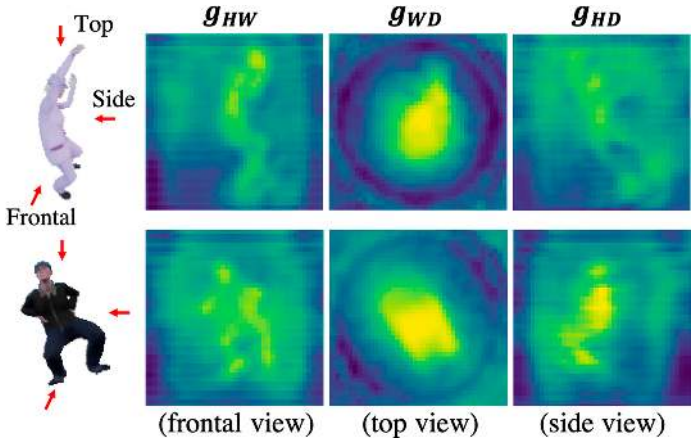

Fig. 5. Example visualization of the gate volumes, for three views and two models of the MVHA dataset.

Fig. 6. Qualitative results on the PVHM dataset, for our approach and different baselines [6,18,37]. Our approach results in overall superior performance.

Table 3. Quantitative results on the ShapeNet dataset [1], for our method and several baseline methods (4 input views), on both car and chair categories. We additionally report the rotational loss for our approach and Olszewski *et al.* [18].

Views	Methods	Car			Chair		
		$L_1 \downarrow$	SSIM↑	RL↓	$L_1 \downarrow$	SSIM↑	RL↓
4	Tatarchenko *et al.* [28]	.112	.890	–	.192	.900	–
	Zhou *et al.* [36]	.081	.924	–	.165	.891	–
	Sun *et al.* [27]	.062	.946	–	.111	.925	–
	Olszewski *et al.* [18]	.059	.946	.076	.107	.939	.073
	Ours	**.051**	**.960**	**.059**	**.087**	**.958**	**.061**

and quantitative results confirm that our approach results in better novel view synthesis reconstruction for human subjects (Fig. 4).

We finally demonstrate that the proposed method can generalize as well for static, inanimate objects such as 'cars' and 'chairs'. In this case, we compare our method to four baselines from recent literature [18,27,28,36]. Table 3 and

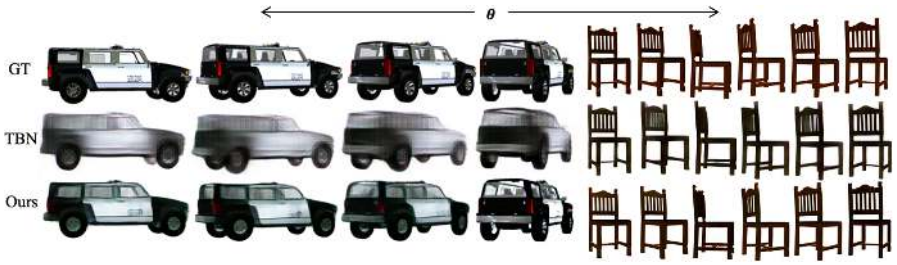

Fig. 7. Qualitative results on the ShapeNet dataset [1], for an example car and chair model (4 input views). Our approach outperforms TBN [18] both in terms of generation quality and consistency across generated views.

Fig. 7 report quantitative and qualitative results, respectively. Together with the L_1 and SSIM metric, we also report the rotational loss for ours and Olszewski et al. [18]. For all metrics, for both the car and chair categories, we outperform all baselines by a consistent margin (Table 3). Figure 7 visually confirms that the quality of the generated views is highly improved with our approach, and that we can keep a high degree of consistency across adjacent generated views.

5 Conclusion

In this paper, we propose a novel siamese network architecture employing volume gating convolutions and temporal feature augmentation to tackle the problem of novel view *video* synthesis of human performance. We also introduce explicit rotational and temporal supervision to guarantee high-quality reconstructions and consistency across the generated views. To support future research in this domain, we collect the Multi-View Human Action (MVHA) dataset, composed of near 1200 synthetic, animated human performance captured from 54 viewpoints. Quantitative and qualitative results on our MVHA, PVHM, and ShapeNet datasets confirm the gains brought by the proposed approach compared to state-of-the-art baselines.

Acknowledgments. Youngjoong Kwon was supported partly by Adobe Research and partly by the National Science Foundation grant 1816148. This work was done while Youngjoong Kwon and Dahun Kim were doing an internship at Adobe Research.

References

1. Chang, A.X., et al.: ShapeNet: an information-rich 3D model repository. arXiv preprint arXiv:1512.03012 (2015)
2. Choy, C.B., Xu, D., Gwak, J.Y., Chen, K., Savarese, S.: 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 628–644. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_38

3. Eslami, S.A., et al.: Neural scene representation and rendering. Science **360**(6394), 1204–1210 (2018)
4. Fuse, A.: https://www.adobe.com/products/fuse.html
5. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 484–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_29
6. Huang, Z., et al.: Deep volumetric video from very sparse multi-view performance capture. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11220, pp. 351–369. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_21
7. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2462–2470 (2017)
8. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
9. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
10. Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. In: Advances in Neural Information Processing Systems, pp. 365–376 (2017)
11. Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Yumer, E., Yang, M.-H.: Learning blind video temporal consistency. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 179–195. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_11
12. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: learning dynamic renderable volumes from images. ACM Trans. Graph. **38**(4) (2019). https://doi.org/10.1145/3306346.3323020
13. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. (TOG) **34**(6), 248 (2015)
14. Mixamo, A.: https://www.mixamo.com
15. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: HoloGAN: unsupervised learning of 3D representations from natural images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7588–7597 (2019)
16. Nguyen-Phuoc, T., Richardt, C., Mai, L., Yang, Y.L., Mitra, N.: BlockGAN: learning 3D object-aware scene representations from unlabelled images. arXiv preprint arXiv:2002.08988 (2020)
17. Nguyen-Phuoc, T.H., Li, C., Balaban, S., Yang, Y.: RenderNet: a deep convolutional network for differentiable rendering from 3D shapes. In: Advances in Neural Information Processing Systems, pp. 7891–7901 (2018)
18. Olszewski, K., Tulyakov, S., Woodford, O., Li, H., Luo, L.: Transformable bottleneck networks. In: The IEEE International Conference on Computer Vision (ICCV), October 2019
19. Park, E., Yang, J., Yumer, E., Ceylan, D., Berg, A.C.: Transformation-grounded image generation network for novel 3D view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3500–3509 (2017)
20. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: learning continuous signed distance functions for shape representation. arXiv preprint arXiv:1901.05103 (2019)

21. Pumarola, A., Sanchez, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F.: 3DPeople: modeling the geometry of dressed humans. arXiv preprint arXiv:1904.04571 (2019)
22. Rezende, D.J., Eslami, S.A., Mohamed, S., Battaglia, P., Jaderberg, M., Heess, N.: Unsupervised learning of 3D structure from images. In: Advances in Neural Information Processing Systems, pp. 4996–5004 (2016)
23. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: PIFu: pixel-aligned implicit function for high-resolution clothed human digitization. arXiv preprint arXiv:1905.05172 (2019)
24. Shysheya, A., et al.: Textured neural avatars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2397 (2019)
25. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: DeepVoxels: learning persistent 3D feature embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2437–2446 (2019)
26. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: continuous 3D-structure-aware neural scene representations. In: Advances in Neural Information Processing Systems, pp. 1121–1132 (2019)
27. Sun, S.-H., Huh, M., Liao, Y.-H., Zhang, N., Lim, J.J.: Multi-view to novel view: synthesizing novel views with self-learned confidence. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 162–178. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_10
28. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Single-view to multi-view: reconstructing unseen views with a convolutional network. arXiv preprint arXiv:1511.06702 6 (2015)
29. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2626–2634 (2017)
30. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: MoCoGAN: decomposing motion and content for video generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1526–1535 (2018)
31. Varol, G., et al.: Learning from synthetic humans. In: CVPR (2017)
32. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., Tenenbaum, J.: MarrNet: 3D shape reconstruction via 2.5 D sketches. In: Advances in Neural Information Processing Systems, pp. 540–550 (2017)
33. Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Advances in Neural Information Processing Systems, pp. 82–90 (2016)
34. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: learning single-view 3D object reconstruction without 3D supervision. In: Advances in Neural Information Processing Systems, pp. 1696–1704 (2016)
35. Yang, J., Reed, S.E., Yang, M.H., Lee, H.: Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In: Advances in Neural Information Processing Systems, pp. 1099–1107 (2015)
36. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View synthesis by appearance flow. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 286–301. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_18
37. Zhu, H., Su, H., Wang, P., Cao, X., Yang, R.: View extrapolation of human body from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4450–4459 (2018)