Rotationally-Consistent Novel View Synthesis for Humans

Youngjoong Kwon
University of North Carolina at Chapel Hill, USA

Stefano Petrangeli
Adobe Research, USA

Dahun Kim
Korea Advanced Institute of Science and Technology, South Korea

Haoliang Wang
Adobe Research, USA

Henry Fuchs
University of North Carolina at Chapel Hill, USA

Viswanathan Swaminathan
Adobe Research, USA

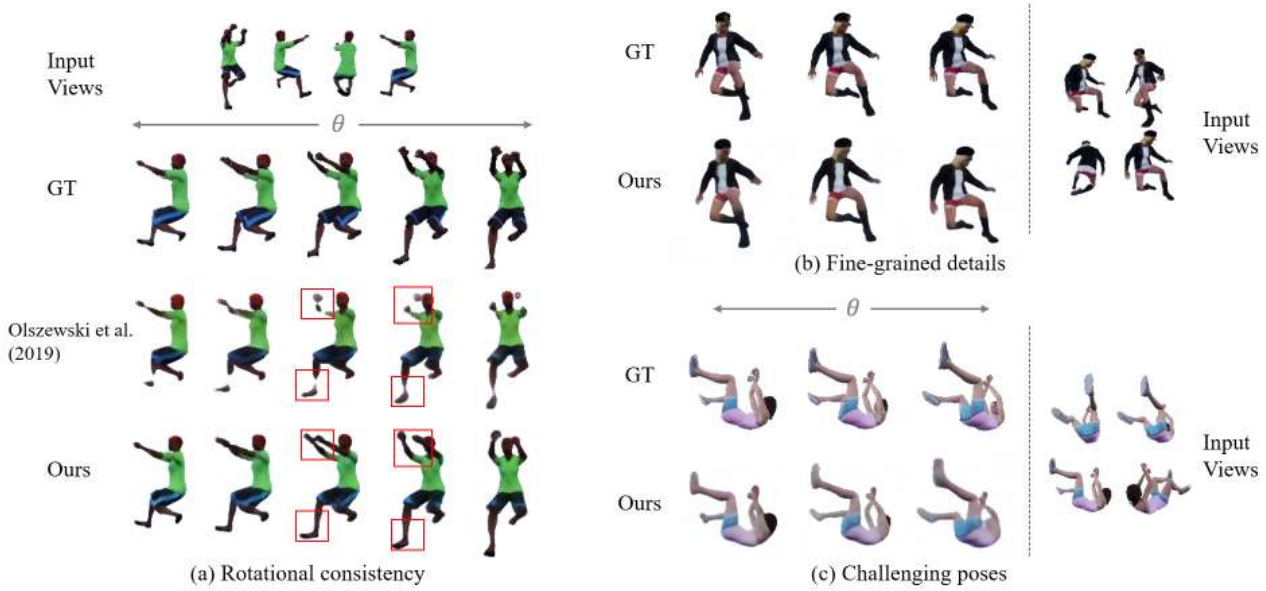

Figure 1: Our proposed method combines novel *multi-view supervision* and *rotational consistency* losses for high-quality human novel view synthesis. Compared to the baseline Olszewski *et al.* [14], our results are rotationally consistent across adjacent views (e.g., arms/legs in (a)). Our method is also able to synthesize fine-grained details (e.g., hat/jacket in (b)) and views with challenging poses as in (c). GT indicates ground truth views. Models are used in training, and tested with unseen poses.

ABSTRACT

Human novel view synthesis aims to synthesize target views of a human subject given input images taken from one or more reference viewpoints. Despite significant advances in model-free novel view synthesis, existing methods present two major limitations when applied to complex shapes like humans. First, these methods mainly focus on simple and symmetric objects, e.g., cars and chairs, limiting their performances to fine-grained and asymmetric shapes. Second, existing methods cannot guarantee visual consistency across different adjacent views of the same object. To solve these problems, we present in this paper a learning framework for the novel view synthesis of human subjects, which explicitly enforces consistency across different generated views of the subject. Specifically, we introduce a novel *multi-view supervision* and an explicit *rotational loss* during the learning process, enabling the model to preserve detailed body parts and to achieve consistency between adjacent synthesized views. To show the superior performance of our approach, we present qualitative and quantitative results on the Multi-View Human Action (MVHA) dataset we collected (consisting of 3D human models animated with different Mocap sequences and captured from 54 different viewpoints), the Pose-Varying Human Model (PVHM) dataset, and ShapeNet. The qualitative and quantitative results demonstrate that our approach outperforms the state-of-the-art baselines in both per-view synthesis quality, and in preserving rotational consistency and complex shapes (e.g. fine-grained details, challenging poses) across multiple adjacent views in a variety of scenarios, for both humans and rigid objects.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413754

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Novel view synthesis.*

KEYWORDS

Human novel view synthesis

ACM Reference Format:
Youngjoong Kwon, Stefano Petrangeli, Dahun Kim, Haoliang Wang, Henry Fuchs, and Viswanathan Swaminathan. 2020. Rotationally-Consistent Novel View Synthesis for Humans. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413754

1 INTRODUCTION

Novel View Synthesis (NVS) aims to synthesize new views of an object given different known viewpoints. Recently, several Deep Neural Network-based NVS approaches have enabled view synthesis by direct image generation without the need for explicit 3D reconstruction or supervision [14, 15, 25, 34]. Despite its rapid development, it is still challenging to apply existing NVS approaches to synthesize complex shapes like the human body. Human novel view synthesis could indeed have a variety of applications in the area of virtual and augmented reality, telepresence, volumetric video reconstruction, virtual try-on systems, and so on.

Existing NVS methods commonly focus on simple and symmetric objects, and perform quite poorly on irregular and asymmetric shapes like a human body, as shown in Figure 1(a). Moreover, visual results are often inconsistent between different generated viewpoints (see Figures 1 and 5). This happens because current NVS works mainly focus on the accurate synthesis of a single target view at a time, without considering spatial and rotational consistency, which in turn can cause body parts (or other fine-grained details) to disappear across adjacent views. Moreover, current methods usually focus on simple and symmetric objects, like those available in the ShapeNet dataset [1], and can therefore fail to reconstruct the complex details of a human subject.

In order to capture the complexity of the human body and overcome the aforementioned limitations, we propose a novel end-to-end learning model that explicitly leverages multi-view constraints for superior reconstruction results. We first introduce the concept of *Multi-View Supervision*, in contrast to classical NVS methods that only compute losses for a single target viewpoint. In our multi-view supervision, the NVS network generates not only the target view, but also additional views adjacent to the target one, during a single forward pass. By generating these additional views, the network has more opportunities to observe the fine-grained details that should not be omitted from the target view, leading to a better reconstruction of the human body shape, even for highly asymmetric poses. Additionally, we propose a new *Rotational Consistency Loss* to explicitly enforce consistency among generated views, an aspect that is often overlooked in previous works. This loss is computed by comparing A', the generated view, with the ground-truth B warped according to the optical flow between ground-truth views A and B. Intuitively, this allows the network to share features between adjacent synthesized views. Consequently, this encourages the network to generate views that are consistent with each other. It is worth noting that our approach is applicable to non-human cases as well. Indeed, our approach evaluated on the ShapeNet cars and chairs categories achieves state-of-the-art performance.

We evaluate the performance on the proposed approach on the Multi-View Human Action (MVHA) dataset we collected [10], composed of diverse 3D human models animated in various poses - each captured from 54 different viewpoints, the Pose-Varying Human Model (PVHM) dataset [35], and ShapeNet [1]. Our method is based on recent NVS works using latent volumetric representations [14, 23], and it consistently outperforms the state-of-the-art baselines on the three aforementioned datasets, both quantitatively and qualitatively.

In summary, our contribution is three-fold. First, we propose the concept of multi-view supervision in the loss formulation of the human novel view synthesis task, which allows the network to learn from multiple views at the same time when generating a specific target view. This allows our approach to generate and retain fine-grained details of the target subject. Second, we introduce the concept of rotational consistency in the loss formulation, to guarantee visual consistency across adjacent generated views. Together with multi-view supervision, this approach results in improved quality of human NVS results. Third, we experiment on our MVHA dataset, PVHM and ShapeNet datasets, and demonstrate that our approach consistently outperforms state-of-the-art baselines on all datasets, both qualitatively and quantitatively.

The rest of this paper is organized as follows. Section 2 reports related works in the area of novel view synthesis. Sections 3 and 4 detail the proposed framework and loss formulation, and the characteristics of our synthetic human dataset, respectively. Comprehensive qualitative and quantitative results are presented in Section 5, while Section 6 concludes the paper.

2 RELATED WORKS

In this section, we review prior works in the area of novel view synthesis. In particular, we review the Convolutional Neural Networks (CNNs)-based methods that have shown promising results on NVS and are therefore widely used in NVS research.

2.1 2D-based novel view synthesis

A large number of works focus on transforming 2D view features to decode a novel view. Tatarchenko *et al.* [26] and Yang *et al.* [33] propose to synthesize novel views by regressing the pixel colors of the target view directly from the input image using CNNs. Several other previous works leverage pixel-flow [15, 25, 34, 35] to generate high-quality, sharp results. These works would usually use flow prediction to directly sample input pixels to construct the output view. Specifically, instead of starting from an empty state, Zhou *et al.* [34] suggest to move pixels from an input to a target view leveraging bilinear sampling kernels [8]. The approach proposed by Park *et al.* [15] achieves high-quality synthesis result by moving only the pixels that can be seen in the novel view, and by hallucinating the empty parts using a completion network. It takes advantage of the symmetry of objects from ShapeNet by producing a symmetry-aware visibility map, facilitating the synthesis with large viewpoint changes. Sun *et al.* [25] further improve the results by aggregating an arbitrary number of input images. Zhu *et al.* [35] presents a human NVS dataset (Pose-Varying Human Model) and method leveraging the depth and optical flow information, but it shows artifacts in the region that is not visible in the source views.

Eslami *et al.* [3] develop a latent representation that can aggregate multiple input views, which shows good results on synthetic geometric scenes. Unlike the previous NVS works that move or regress pixels, Shysheya *et al.* [22] regress texture coordinates corresponding to a pre-defined texture map. Our human novel view synthesis task cannot fully take advantage of these approaches since humans can present highly asymmetric poses that are hardly modeled by the 2D-based synthesis methods that lack 3D understanding.

2.2 3D-structure aware novel view synthesis

There have been promising results from works embedding implicit spatial consistency in the NVS task using volumetric representations. Several recent methods reconstruct an explicit occupancy volume from a single image, and render it using traditional rendering techniques[2, 5, 9, 19, 27, 30–32]. Methods leveraging signed-distance-field-encoded volume [16, 21] or RGBα-encoded volume [12], which can be rendered by ray-marching algorithms, have achieved excellent quality while overcoming the memory limitations of voxel-based representations. However, it cannot support generalization to unseen models and poses. Saito *et al.* [21] predict the continuous inside/outside probability of a clothed human, and also infers an RGB value at given 3D positions of the surface geometry, resulting in a successful recovery of intricate details of garments. Instead, it requires ground-truth 3D supervision. Rather than generating explicit occupancy volumes, several methods [9, 14, 19, 23] generate latent volumetric representations that can be rendered by a learnable decoder. Sitzmann *et al.* introduce latent 3D [23] and implicit feature embeddings [24] to address the inconsistency between views synthesized by generative networks, which can occur due to a lack of 3D understanding. The DeepVoxel method [23], however, requires scene-specific optimizations. Olszewski *et al.* [14] generate a latent voxel representation that allows transformations including scaling, rotation, and combination of different input view images. Moreover, their so-called Transformable Bottleneck Network (TBN) does not require any 3D supervision and is therefore easy to train. This approach has produced state-of-the-art results for the NVS task on the ShapeNet dataset. However, it presents only implicit 3D consistency at an intermediate feature level.

Our approach is based on latent volumetric representation approaches, which we further improve by introducing spatial consistency across generated view using *explicit* rotational consistency constraint. Our proposed method better copes with the human novel view synthesis task, and is able to generalize to new models and poses that are unseen during training.

3 HUMAN NOVEL VIEW SYNTHESIS

In this section, we first introduce the baseline network architecture used in our approach [14], and the extensions we propose for the human novel view synthesis task. We then detail the novel loss function designed to, a) learn complex asymmetric shapes via multi-view supervision, and b) enforce rotational consistency across generated views via rotational loss.

3.1 Network architecture

Our proposed approach is built on top of the *Transformable Bottleneck Network* (TBN) architecture proposed by Olszewski *et al.* [14]. A TBN consists of three blocks: an encoder, a bottleneck resampling layer, and a decoder. An overview of the original TBN architecture is given in Figure 2(a). The encoder network E takes an RGB image I_k (taken from input view k) as input, and generates a volumetric representation V_k of I_k through a series of 2D convolutions, reshaping and 3D convolutions. The bottleneck resampling layer S transforms the input volumetric representation V_k into the target view V_t via trilinear interpolation $T_{k \Rightarrow t}$ (where k and t are the input and target views, respectively). This resampling operation is purely computational, i.e., it is not learned but instead provided as an input to the network. This approach allows the network to learn complex transformations between input and target views. Multiple input views can be used by averaging the intermediate volumetric representations before decoding. The decoder D, whose architecture mirrors that of the encoder E, takes the aggregated volumetric representation and generates the output image I'_l. TBN requires only the RGB image of the input and target views, without any form of 3D supervision, which makes it easy to train.

In light of the above, we extend the TBN architecture by incorporating the proposed novel loss functions for the human NVS task. The extended TBN architecture used in this work is presented in Figure 2(b). In addition to the target view t, our network generates a set of additional views, l, starting from the same intermediate volumetric representation. These additional views are combined with the proposed loss system to guarantee both high-quality reconstruction of the target view and visual consistency across adjacent views of the same target, as described in Sections 3.2 and 3.3.

3.2 Multi-view supervision

The goal of our network is to reconstruct the complex and asymmetric shape of a human subject. We therefore propose the concept of *multi-view supervision*, where the network is trained by generating not only the desired target view, but also a set of additional views. These additional views help the network to synthesize complex shapes, by providing additional information during the training process of the target view t on the actual global shape of the human subject. Multi-view supervision is performed by generating both views immediately *adjacent* to the target one and views which are *farther away* with respect to the target. We denote these two sets of additional views as L_a and L_f. As an example, L_a can indicate the two views immediately adjacent to the target one, and L_f can indicate two views that are a certain degree apart from the target.

To obtain the additional views, we resample the volumetric representation V_t to V_l, with $l \in \{L_a, L_f\}$, and decode it into image I'_l, which is compared to ground-truth image I_l. For the multi-view supervision, we use a similar loss system in the image space as that proposed by Olszewski *et al.* [14], given by the following equations:

$$V_l = S(V_t, T_{t \Rightarrow l}) \tag{1}$$

$$I'_l = D(V_l) \tag{2}$$

$$\mathcal{L}_R = \left\| I'_l - I_l \right\|_1 \tag{3}$$

$$\mathcal{L}_P = \sum_i \left\| VGG_i(I'_l) - VGG_i(I_l) \right\|_2^2 \tag{4}$$

$$\mathcal{L}_{Multi} = \frac{1}{|L_a| + |L_f|} \sum_{l \in \{L_a, L_f\}} (\lambda_R \mathcal{L}_R + \lambda_P \mathcal{L}_P + \lambda_S \mathcal{L}_S + \lambda_A \mathcal{L}_A) \tag{5}$$

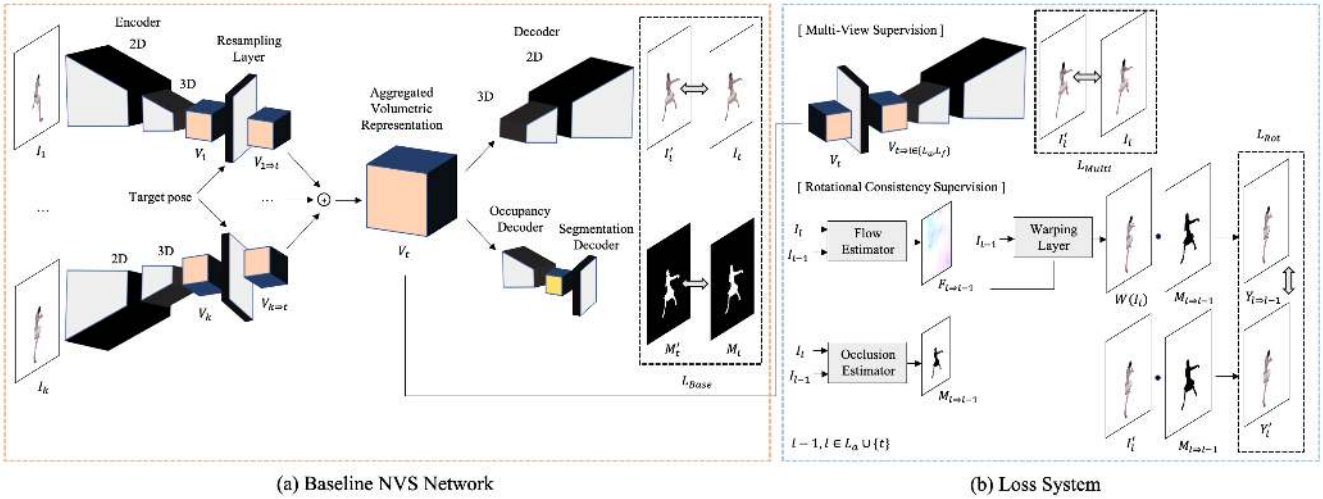

Figure 2: Overview. Our network is built upon the 3D-structure aware NVS network (a) [14]. In multi-view supervision (upper part - (b)), additional target views are generated starting from the same latent volumetric representation, and the loss is computed between ground-truth and synthesized images. In rotational consistency supervision (bottom part - (b)), the backward optical flow between two adjacent ground-truth views I_l and I_{l-1} is first obtained. Next, the loss is computed by comparing the warped ground-truth image I_l and the synthesized view I'_{l-1}. Both images are applied with a binary occupancy mask computed starting from I_l and I_{l-1}.

$\mathcal{L}_R$ is the the pixel-wise L_1 reconstruction loss between ground-truth and reconstructed image, while $\mathcal{L}_P$ is the L_2 loss in the feature space of the VGG-19 network. VGG_i indicates the output of the i^{th} layer of the VGG-19 network. $\mathcal{L}_S$ and $\mathcal{L}_A$ are the structural similarity and adversarial loss, respectively [14, 20, 29].

3.3 Rotational consistency supervision

Besides being able to reconstruct high-quality complex shapes given a target view t, we also want our network to generate *spatially consistent* results across different views. Our rotational loss is therefore designed to minimize inconsistencies between synthesized views, and is formulated as the warping error computed between the additional views introduced in the previous section.

Particularly, we consider two adjacent views $l-1$ and l both belonging to $L_a \cup \{t\}$, with t being the actual target view. We first compute $W_{l\Rightarrow l-1}$, the function warping an image according to the ground-truth backward flow $F_{l\Rightarrow l-1}$ between the ground-truth images I_l and I_{l-1} [6]. Next, we compute $M_{l\Rightarrow l-1}$, the binary occlusion mask between the I_l and $W_{l\Rightarrow l-1}(I_{l-1})$ [11]:

$$M_{l\Rightarrow l-1} = e^{-\alpha \|I_l - W_{l\Rightarrow l-1}(I_{l-1})\|_2^2} \tag{6}$$

We use a bi-linear sampling layer to warp images and set $\alpha = 50$ (with pixel range between $[0, 1]$) [8]. We apply this occlusion mask to both the generated image I'_l and the warped ground-truth $W_{l\Rightarrow l-1}(I_{l-1})$ to get Y'_l and $Y_{l\Rightarrow l-1}$, and compute the loss between l and $l-1$ as in the following equations:

$$Y'_l = M_{l\Rightarrow l-1} \cdot I'_l \tag{7}$$

$$Y_{l\Rightarrow l-1} = M_{l\Rightarrow l-1} \cdot W_{l\Rightarrow l-1}(I_{l-1}) \tag{8}$$

$$\mathcal{L}_{R_1} = \left\|Y'_l - Y_{l\Rightarrow l-1}\right\|_1 + \left\|Y'_{l-1} - Y_{l-1\Rightarrow l}\right\|_1 \tag{9}$$

$$\begin{aligned}\mathcal{L}_{R_2} = {} & [1 - SSIM(Y'_l, Y_{l\Rightarrow l-1})] \\ & + [1 - SSIM(Y'_{l-1}, Y_{l-1\Rightarrow l})]\end{aligned} \tag{10}$$

Applying the occlusion mask on I'_l and $W_{l\Rightarrow l-1}(I_{l-1})$ allows to calculate the loss only on those pixels that are visible given the backward flow. $\mathcal{L}_{R_1}$ and $\mathcal{L}_{R_2}$ represent the L_1 reconstruction loss and the structural similarity loss ($SSIM$ is considered normalized) between the warped ground-truth and the generated image. We consider in this case both backward flows $l \Rightarrow l-1$ and $l-1 \Rightarrow l$. By comparing the synthesized image with the warped ground-truth, we encourage the network to learn the inherent rotational consistency of the ground-truth images. The final rotational loss is given by:

$$\mathcal{L}_{Rot} = \frac{1}{|L_a|} \sum_{(l-1,l)\in L_a \cup \{t\}} \lambda_{R_1}\mathcal{L}_{R_1} + \lambda_{R_2}\mathcal{L}_{R_2} \tag{11}$$

3.4 Overall loss

We train our network to minimize the following loss function:

$$\mathcal{L}_{Total} = \lambda_{Base}\mathcal{L}_{Base} + \lambda_{Multi}\mathcal{L}_{Multi} + \lambda_{Rot}\mathcal{L}_{Rot}, \tag{12}$$

where $\mathcal{L}_{Base}$ is original the loss presented in [14], $\mathcal{L}_{Multi}$ is the multi-view supervision loss, and $\mathcal{L}_{Rot}$ is the rotational loss. The

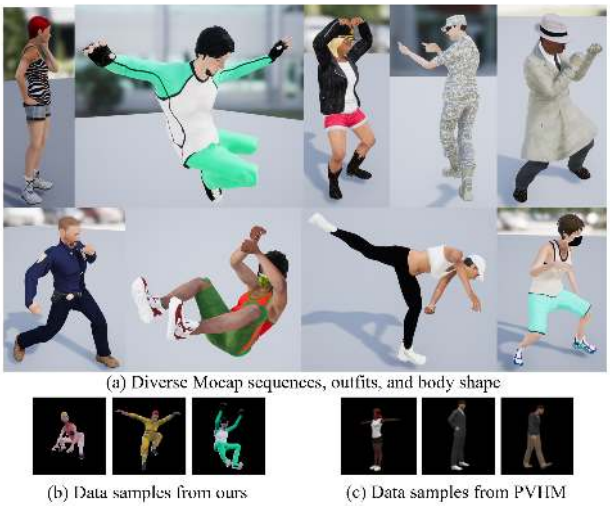

Figure 3: Overview of our MVHA dataset.

balancing weights $\lambda_{Base}, \lambda_{Multi}, \lambda_{Rot}$ are all set to 1 throughout the experiments.

4 MULTI-VIEW HUMAN ACTION (MVHA) DATASET

In this section, we introduce the Multi-View Human Action (MVHA) dataset we collected to support the development of novel human-specific NVS approaches [10] (Figure 3). Compared with previous similar datasets that only provide 4 captured viewpoints [7], ours provide 54 different viewpoints for each unique model (18 azimuth and 3 elevations).

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 and hair geometry. Each subject is dressed with a different outfits including a variety of garments with multiple colors and textures, ranging from casuals, sports, to uniforms, and shoes. Also, each subject features diverse fine-grained details including the makeup, mustache, beard, hats, glasses and mask (see Figure 3 (a)). Compared to existing human datasets [7, 18, 28, 35] with relatively simple outfits and details (see a sample from the PVHM dataset in Figure3 (c)), our synthetic human dataset (Figure3 (b)) allows to train/test on more difficult NVS task applied to humans.

Mocap sequences. We gather 40 diverse Mocap sequences from Adobe Mixamo [13]. The sequences range from static everyday motions (e.g. answering phone, searching pockets, clapping) to very complex and dynamic motion patterns including dancing, sports, fighting (e.g. YMCA dance, aerial, bouncing a basketball, kicking, punching) as in Figure 3 (a) and (b).

Camera, lights and background. A 3D rendering software is used to apply the 40 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 ensures the whole body is always in view. Every sequence is rendered from 54 camera views - 18 azimuth at 20 degree intervals, and 3 elevations with 10 degree intervals. For every rendered view, we provide the final RGB image and associated binary segmentation mask.

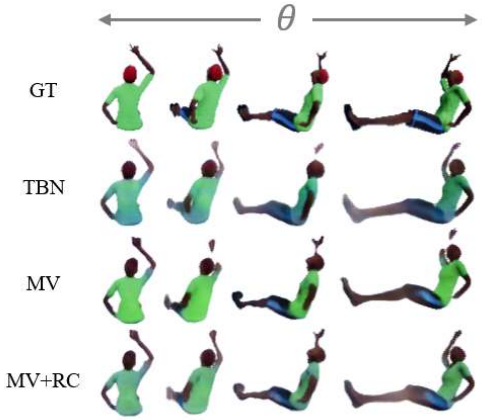

Figure 4: Ablation study (seen model / unseen pose). The proposed multi-view supervision (third row) allows the network to generate higher quality results than TBN (second row), both in terms of reconstructed shape and color (e.g., the blue stripe). With the rotational loss (fourth row), the network can generate finer details across the adjacent views (e.g., right arm).

Methods	$L_1 \downarrow$	SSIM↑	RL-$L_1 \downarrow$	RL-SSIM↑
TBN (Baseline)	.111	.705	.054	.685
Ours (MV)	.096	.742	.050	.707
Ours (MV+RC)	**.087**	**.752**	**.042**	**.722**

Table 1: Ablation study (seen model / unseen pose). We report L_1 (lower is better) and SSIM (higher is better), and the corresponding components for the rotational loss (Equations 9 and 10).

5 EXPERIMENTS

In this section, we thoroughly evaluate the effectiveness of our method on our MVHA dataset [10], as well as the PVHM [35] and ShapeNet [1] datasets. First, we compare with the baseline method TBN [14] on our MVHA dataset to show the ablative contributions of each proposed component. Also, we present benchmarking results on the PVHM dataset to provide a comparison with state-of-the-art NVS methods for human models. Finally, experiments on the ShapeNet dataset confirm that our method performs well for rigid and symmetric objects like cars and chairs. On all datasets, our approach consistently outperforms other competing methods, both quantitatively and qualitatively.

For evaluation, we use the L_1 and SSIM losses to measure the quality of the synthesized view with respect to the ground-truth, and the L_1 and SSIM components of the Rotational Loss (RL) in Equations 9 and 10 to measure the rotational consistency between

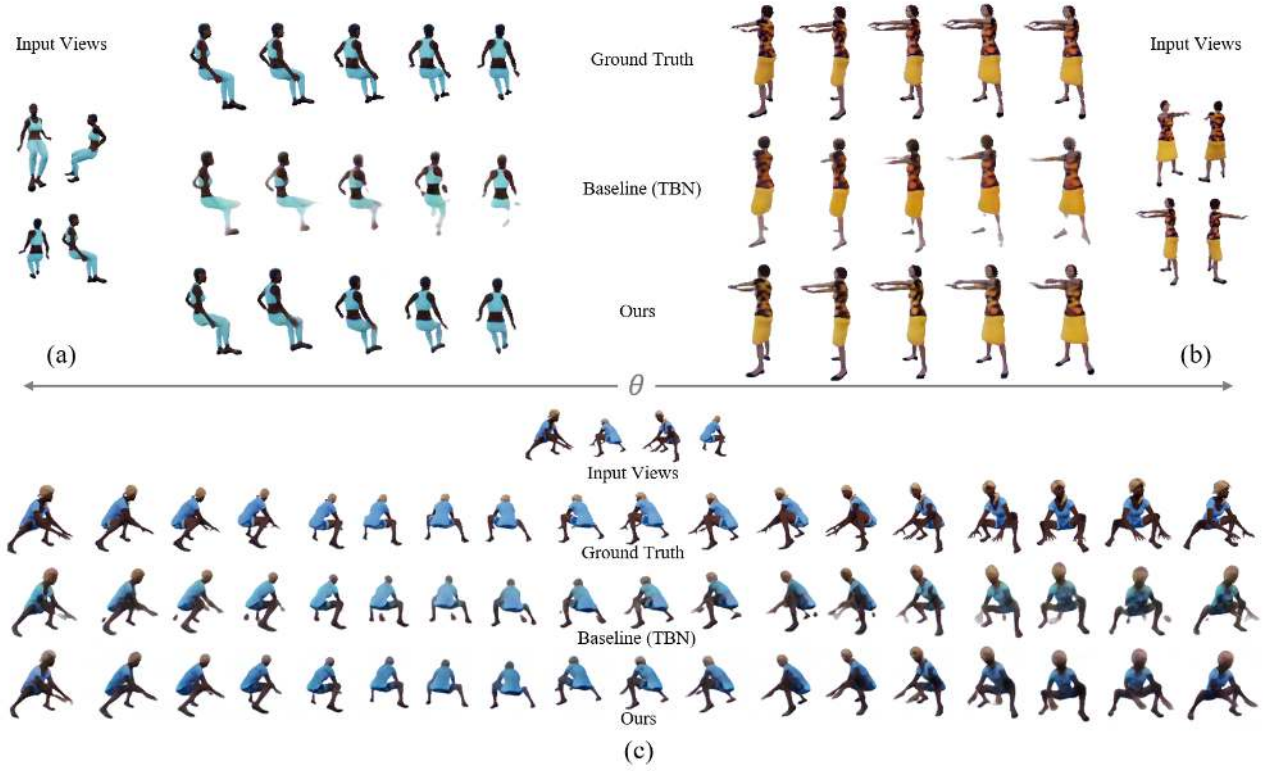

Figure 5: Comparison with TBN on generalization (4 input views). (a) and (b) are results generated from models seen during training with unseen poses. (c) reports 360-degree rotation results from an unseen model with an unseen pose. Independently of the testing scenario, our method is able to generate fine-grained details and better consistency between adjacent views. In both (a), (b) and (c), the arm/leg of the model appears and disappears across views for TBN. Our network is able to successfully generalize even for an unseen model with a challenging unseen pose as in (c).

the adjacent views. We use the acronyms *MV* and *MV+RC* for our network with multi-view supervision only (Equation 5) and multi-view supervision in combination with rotational consistency (Equation 11), respectively. We use our *MV+RC* model for all experiments unless specified otherwise. The proposed approach has been implemented and trained using the PyTorch framework [17]. Each network was trained on 8 NVIDIA Tesla V100s, with each batch distributed across the GPUs. We trained each model until convergence on the test set, which took approximately 8 days.

5.1 Results on the MVHA dataset

Ablation Study. Our MVHA dataset exhibits highly asymmetric and complex poses, which allows to clearly show the limitations of previous methods. We investigate the ablative impact of the proposed multi-view supervision (MV) and rotational consistency (RC) losses over the baseline TBN method, both quantitatively (Table 2) and qualitatively (Figure 4). Both *MV* and *RC* improve the baseline performance on the per-frame synthesis quality (L_1, *SSIM*) and rotational consistency (*RL-*L_1, *RL-SSIM*). Visual results confirm the gains brought by the proposed supervisions. As seen in Figure 4, views generated by TBN fail to reconstruct body parts and details

Testing Scenario	Method	L1↓	SSIM↑	RL-L1↓	RL-SSIM↑
1. Seen model	TBN	.111	.705	.054	.685
+ Unseen pose	Ours	**.087**	**.752**	**.042**	**.722**
2. Unseen model	TBN	.109	.708	.050	.692
+ Seen pose	Ours	**.094**	**.721**	**.045**	**.705**
3. Unseen model	TBN	.117	.683	.056	.678
+ Unseen pose	Ours	**.097**	**.715**	**.046**	**.697**

Table 2: Quantitative comparison with TBN on our synthetic human dataset on a variety of scenarios in terms of models and poses.

such as the right arm, shoes and trousers of the human model. Using our proposed multi-view supervision loss, our network is able to generate better details (such as the blue stripe of the trousers). It is worth noting that multi-view supervision enforces stronger constraints on the asymmetric and complex shape of the target subject during training, which in turn leads to a better learned volumetric representation. Adding the rotational consistency produces better reconstruction of body parts, *e.g.*, the right arm in Figure 4,

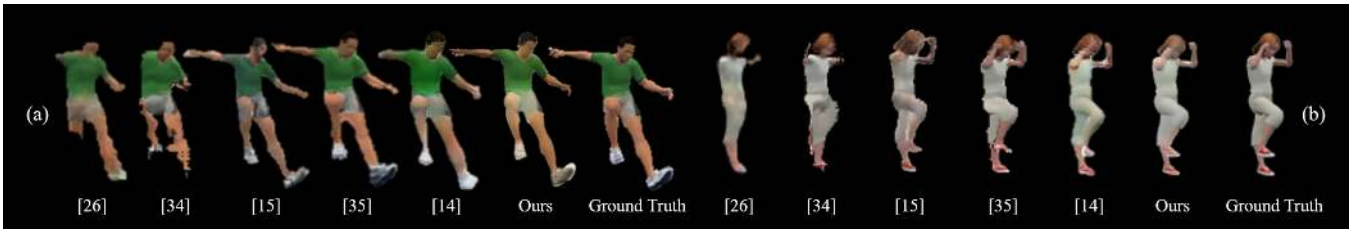

Figure 6: Qualitative results on the PVHM dataset, for our approach and different baselines [14, 15, 26, 34, 35].

by enforcing explicit consistency across multiple adjacent views. The representations learned in one view are shared to the adjacent views, which allows our method to produce higher quality results. Both quantitatively and qualitatively, the best results are achieved when both *MV* and *RC* are used together.

Comparison with TBN and generalization capabilities. The ability to generalize to unseen models and poses is essential in the human NVS problem. We consider three generalization scenarios with different levels of complexity: 1) models seen during training and new unseen poses, 2) unseen models with seen poses, and 3) *both* unseen models and unseen poses. We use the same volume resolution ratio as in TBN for fair comparison. Table 2 shows that our method can generalize not only for unseen poses (Scenario 1), but also for new human models (Scenario 2 and 3). The per-view quality scores (L_1, *SSIM*) drop only slightly even in the most difficult generalization scenario, and are consistently better than TBN. It is worth noting that our method tested on unseen models/unseen poses (L_1: 0.097, *RL*-L_1: 0.046) outperforms TBN tested on the simplest case of seen models/unseen poses (L_1: 0.111, *RL*-L_1: 0.054).

Qualitative results confirm the quantitative ones, as shown in Figures 1 and 5. Results for Scenario 1 (seen models/unseen poses) are shown in Figure 1 and Figure 5(a)-(b). Our approach successfully generates fine details (e.g., arms/legs) in a multi-view consistent manner. In contrast, TBN results in missing body parts (such as arms and hands), and the generated views are inconsistent among each other, *e.g.*, some body parts appear in one view, but disappear in another. Even for the most challenging Scenario 3 (unseen models/unseen poses), shown in Figure 5(c), our method is able to successfully generalize to complex and asymmetric human shapes, despite the presence of heavy occlusions, *e.g.*, between arms and legs. On the other hand, TBN shows incomplete reconstruction of the hands and inconsistent generation of the legs.

5.2 Results on PVHM and ShapeNet datasets

To conclude the analysis, we also report quantitative and qualitative results on the PVHM and ShapeNet datasets. These results allows us to show that the proposed approach can work also beyond the MVHA dataset we collected and for rigid objects as well.

Quantitative and qualitative results for the PVHM dataset are reported in Table 3 and Figure 6, respectively. We show the effetiveness of our approach by comparing our view generation quality with diverse well-know NVS approaches. Specifically, we perform comparisons with pixel regression method (Tatarchenko *et al.* [26]), flow-based approach (Zhou *et al.* [34], Park *et al.* [15], and Zhu *et al.* [35]), and TBN.

Method	MSE↓	SSIM↑
Tatarchenko *et al.* [26]	96.83	.9488
Zhou *et al.* [34]	131.6	.9527
Park *et al.* [15]	85.35	.9519
Zhu *et al.* [35]	72.86	.9670
Olszewski *et al.* [14]	70.34	.9695
Ours	**66.85**	**.9749**

Table 3: Quantitative results on the PVHM dataset [35], for our method and several baseline methods.

For numerical comparison, we use the MSE and SSIM to quantitatively compare the different approaches, as proposed by Zhu *et al.* [35]. Our approach results in consistent better performance, by a large margin, compared to the other baseline methods (Table 3), providing state-of-the-art results on PVHM.

In terms of qualitative results (Figure 6), Tatarchenko *et al.* [26] show blurry results and incomplete reconstruction of human structure due to the challenging nature of pixel regression out of empty state. Flow-based methods [15, 34, 35] present artifacts on the regions that are not visible from the source view. Specifically, Zhou *et al.* [34] and Park *et al.* [15] show clear artifacts on the arms and legs in both examples (a) and (b). Park *et al.* [15] present artifacts around the face region in (a) due to the hallucination failure of invisible areas from the source view, while Zhu *et al.* [35] show artifacts on the legs in (a), and foot and left arm in (b). Similar inconsistencies can be noted for Olszewski *et al.* [14]. Our method generates views that are much closer to the ground-truth both in terms of pose and fine-grained details (e.g, arms, legs, hands), for both examples (a) and (b).

Results for the chair and car category of ShapeNet are reported in Table 4 and Figure 7. Numerical results (Table 4) show that our approach provide state-of-the-art results on the chair and car category, and outperform all other baselines by a consistent margin. This result clearly shows that, even though the proposed method was designed with the human novel view synthesis in mind, it can be applied to other rigid object categories as well. Qualitative results (Figure 7) indicate that, compared with TBN, our approach can generate higher-quality and finer-grained details, with better consistency across adjacent views.

Overall, the results on PVHM and ShapeNet confirm the superior performance of the proposed approach with respect to state-of-the-art methods on the NVS task.

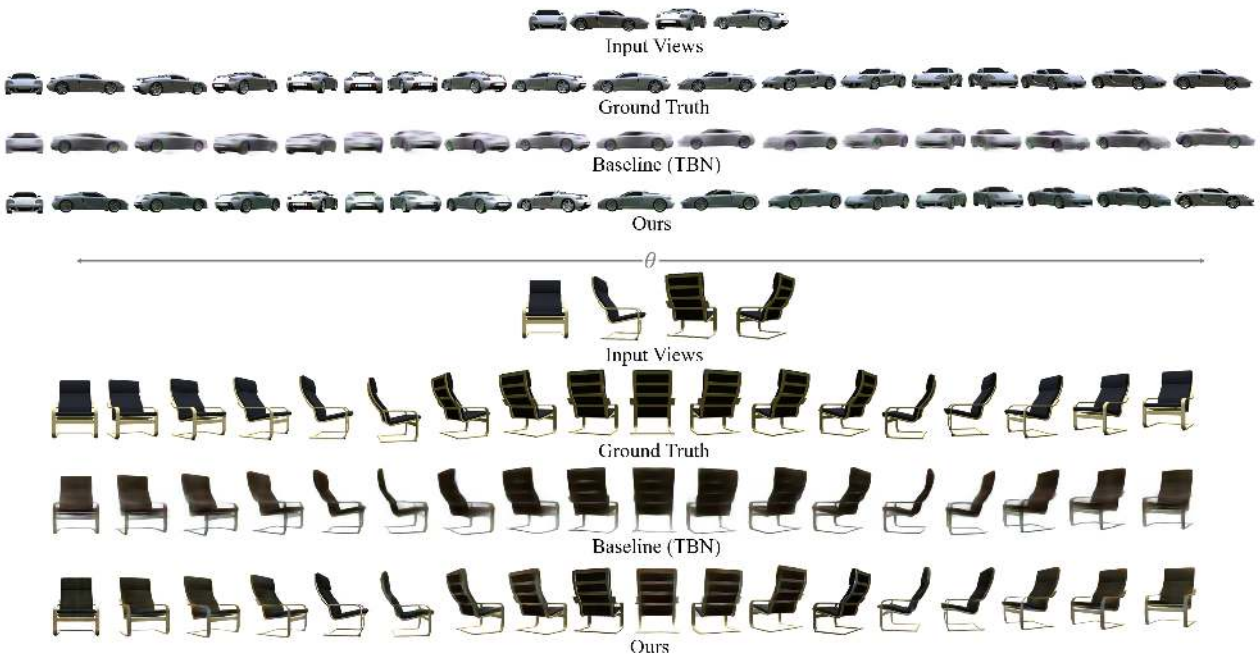

Figure 7: Qualitative results on ShapeNet for 360-degree rotations. Our method generates fine-grained details and better consistency between adjacent views.

Methods	Car				Chair			
	$L_1 \downarrow$	SSIM↑	RL-$L_1 \downarrow$	RL-SSIM↑	$L_1 \downarrow$	SSIM↑	RL-$L_1 \downarrow$	RL-SSIM↑
Tatarchenko *et al.* [26]	.112	.890	-	-	.192	.900	-	-
Zhou *et al.* [34]	.081	.924	-	-	.165	.891	-	-
Sun *et al.* [25]	.062	.946	-	-	.111	.925	-	-
Olszewski *et al.* [14]	.059	.946	.076	.729	.107	.939	.073	.735
Ours	**.054**	**.954**	**.068**	**.748**	**.099**	**.948**	**.067**	**.740**

Table 4: 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.* [14].

6 CONCLUSIONS

We presented in this paper a novel approach for novel view synthesis for humans, which often present complex and asymmetrical poses and rich details. We introduced a novel approach that leverages the concepts of *multi-view supervision* and *rotationally consistency* to generate high-quality results. Using our method in combination with a leading NVS approach [14], we showed that we outperform it both quantitatively and qualitatively in the human novel view synthesis task, for a variety of scenarios and datasets. Our solution is indeed able to produce higher-quality images that can retain fine-grained details with respect to the ground-truth, for complex and asymmetric body shapes. Moreover, results on the MVHA dataset we collected, PVHM and ShapeNet confirm that our approach can produce state-of-the-art results for both human subjects and rigid objects, when compared with several baseline solutions.

ACKNOWLEDGMENTS

YoungJoong Kwon was supported partly by Adobe and partly by National Science Foundation grant 1816148.

REFERENCES

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[2] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*. Springer, 628–644.

[3] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor,

et al. 2018. Neural scene representation and rendering. *Science* 360, 6394 (2018), 1204–1210.
[4] Adobe Fuse. [n.d.]. https://www.adobe.com/products/fuse.html.
[5] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. 2016. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*. Springer, 484–499.
[6] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*. 2462–2470.
[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.
[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*. 2017–2025.
[9] Abhishek Kar, Christian Häne, and Jitendra Malik. 2017. Learning a multi-view stereo machine. In *Advances in neural information processing systems*. 365–376.
[10] YoungJoong Kwon, Stefano Petrangeli, Dahun Kim, Haoliang Wang, Eunbyung Park, Viswanathan Swaminathan, and Henry Fuchs. 2020. Rotationally-Temporally Consistent Novel View Synthesis of Human Performance Video. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
[11] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. 2018. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 170–185.
[12] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *arXiv preprint arXiv:1906.07751* (2019).
[13] Adobe Mixamo. [n.d.]. https://www.mixamo.com.
[14] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. 2019. Transformable Bottleneck Networks. In *The IEEE International Conference on Computer Vision (ICCV)*.
[15] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. 2017. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 3500–3509.
[16] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. *arXiv preprint arXiv:1901.05103* (2019).
[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).
[18] Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2019. 3DPeople: Modeling the Geometry of Dressed Humans. *arXiv preprint arXiv:1904.04571* (2019).
[19] Danilo Jimenez Rezende, SM Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. 2016. Unsupervised learning of 3d structure from images. In *Advances in Neural Information Processing Systems*. 4996–5004.
[20] Karl Ridgeway, Jake Snell, Brett Roads, Richard S Zemel, and Michael C Mozer. 2015. Learning to generate images with perceptual similarity metrics. *arXiv preprint arXiv:1511.06409* (2015).
[21] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. *arXiv preprint arXiv:1905.05172* (2019).
[22] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. 2019. Textured Neural Avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2387–2397.
[23] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. 2019. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2437–2446.
[24] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. 2019. Scene representation networks: Continuous 3D-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*. 1119–1130.
[25] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. 2018. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 155–171.
[26] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. 2015. Single-view to multi-view: Reconstructing unseen views with a convolutional network. *arXiv preprint arXiv:1511.06702* 6 (2015).
[27] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. 2017. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2626–2634.
[28] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 109–117.
[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
[30] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. 2017. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems*. 540–550.
[31] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*. 82–90.
[32] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. 2016. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*. 1696–1704.
[33] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. 2015. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*. 1099–1107.
[34] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. 2016. View synthesis by appearance flow. In *European conference on computer vision*. Springer, 286–301.
[35] Hao Zhu, Hao Su, Peng Wang, Xun Cao, and Ruigang Yang. 2018. View extrapolation of human body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4450–4459.