Human View Synthesis using a Single Sparse RGB-D Input

Phong Nguyen[1*], Nikolaos Sarafianos[2], Christoph Lassner[2], Janne Heikkila[1], Tony Tung[2]
[1] University of Oulu, Finland [2] Reality Labs Research, Sausalito

Figure 1. *Overview.* We present a Human View Synthesis model that predicts novel views of humans from a single-view, sparse RGB-D input. Our method renders high quality novel views of both, synthetic and real humans at 1K resolution without per-subject fine tuning.

Abstract

Novel view synthesis for humans in motion is a challenging computer vision problem that enables applications such as free-viewpoint video. Existing methods typically use complex setups with multiple input views, 3D supervision or pre-trained models that do not generalize well to new identities. Aiming to address these limitations, we present a novel view synthesis framework to generate realistic renders from unseen views of any human captured from a single-view sensor with sparse RGB-D, similar to a low-cost depth camera, and without actor-specific models. We propose an architecture to learn dense features in novel views obtained by sphere-based neural rendering, and create complete renders using a global context inpainting model. Additionally, an enhancer network leverages the overall fidelity, even in occluded areas from the original view, producing crisp renders with fine details. We show our method generates high-quality novel views of synthetic and real human actors given a single sparse RGB-D input. It generalizes to unseen identities, new poses and faithfully reconstructs facial expressions. Our approach outperforms prior human view synthesis methods and is robust to different levels of input sparsity.

1. Introduction

Novel view synthesis of rigid objects or dynamic scenes has been a very active topic of research recently with impressive results across various tasks [43, 48, 65, 68]. However, synthesizing novel views of humans in motion requires methods to handle dynamic scenes with various deformations which is a challenging task [66, 72]; especially in those regions with fine details such as the face or the clothes [49, 51, 71]. In addition, prior work usually relies on a large amount of cameras [4, 43], expensive capture se-

*This work was conducted during an internship at RL Research.

tups [52], or inference time on the order of several minutes per frame. This work aims to tackle each of these challenges using a simple, compact yet effective formulation.

We propose a novel framework that generates high-fidelity rendered images of clothed humans using a single sparse RGB-D sensor. The challenging requirements that we impose are: i) generalization to new subjects at test-time as opposed to models trained per subject, ii) ability to handle dynamic scenes of humans in unseen poses as opposed to animating humans using the same poses seen at training, iii) ability to handle occlusions either from objects or from self-occlusions, iv) capturing facial expressions and v) generation of high-fidelity images in a live setup given a sparse RGB-D input (*i.e*. similar to a low-cost depth camera).

Our method takes as input a single sparse RGB-D image of the upper body of a human and a target camera pose and generates a high-resolution rendering from the target viewpoint (see Fig. 1). A few methods have been published recently that tackle a similar scenario: LookingGood [41] re-renders novel viewpoints of a captured individual given a single RGB-D input. However, their capture setup produces dense geometry which results in a relatively easy task: the target views do not deviate significantly from the input views. Given multi-view input frames or videos, recent works on rendering animatable humans from novel views show impressive results [49, 51, 52, 71]. However such methods can be prohibitively expensive to run (*e.g*., [49] runs at 1 minute/frame) and cannot generalize to unseen humans but instead create a single model for each human that they need to render.

The first key differentiating factor of our proposed approach compared to previous approaches is that we utilize depth as an additional input stream. While the input depth is sparse and noisy it still enables us to utilize the information seen in the input main view and hence simplifying the synthesis of novel views. To account for the sparseness of the input, we opted for a sphere-based neural renderer that uses a learnable radius to achieve a denser warped image compared to simply performing geometry warping from one view to the other. When combined with an encoder-decoder architecture and trained end-to-end, our approach is able to synthesize novel views of unseen individuals and to in-paint areas that are not visible from the main input view. However, we observed that while this approach works well with minimal occlusions it has a hard time generating high-quality renderings when there are severe occlusions, either from the person moving their hands in front of their body or if they're holding various objects. Thus, we propose to utilize a single additional occlusion-free input and warp it to the target novel view by establishing accurate dense correspondences between the two inputs. A compact network can be used for this purpose, which is sufficient to refine the final result and generate the output prediction.

To train our approach, we rely on high-quality synthetic scans of humans that we animated and rendered from various views. A key finding of our work is that it generalizes very well to real data captured by a 3dMD scanner system with a level of detail in the face or the clothes that are not seen in prior works. In summary, the contributions of this work are:

- A robust sphere-based synthesis network that generalizes to multiples identities without per-human optimization.
- A refinement module that enhances the self-occluded regions of the initial estimated novel views.
- State-of-the-art results on dynamic humans of both, synthetic and real-captured data.

2. Related Work

View synthesis for dynamic scenes, in particular for humans, is a well-established field that provides the basis for this work. Our approach builds on ideas from point-based rendering, warping, and image-based representations.

View Synthesis. For a survey of early image-based rendering methods, we refer to [58, 63]. One of the first methods to work with video in this field is presented in [9] and uses a pre-recorded performance in a multi-view capturing setup to create the free-viewpoint illusion. Zitnick et al. [77] similarly uses a multi-view capture setup for viewpoint interpolation. These approaches interpolate between recorded images or videos, and Ballan et al. [3] coin the term 'video-based rendering': they use it to interpolate between handheld camera views of performances.

The strong generative capabilities of neural networks enable further extrapolation and relaxation of constraints [22, 26, 31, 42]. Zhou et al. [75] introduce Multi-Plane Images (MPIs) for viewpoint synthesis and use a model to predict them from low-baseline stereo input and [21, 59] improve over the original baseline and additionally work with camera arrays and light fields. Broxton et al. [7] extend the idea to layered, dynamic meshes for immersive video experiences whereas Bansal *et al*. [4] use free camera viewpoints, but multiple cameras.

With even stronger deep neural network priors, [69] performs viewpoint extrapolation from a single view, but for static scenes, whereas [66, 72] can work with a single view in dynamic settings with limited motion. Bemana et al. [5] works in static settings but predicts not only the radiance field but also lighting given varying illumination data. Chibane et al. [14] trade instant depth predictions and synthesis for the requirement of multiple images. Alternatively, volumetric representations [39, 40] can also being utilized for capturing dynamic scenes. All these works require significant computation time for optimization, multiple views or offline processing for the entire sequence.

3D & 4D Performance Capture. While aforementioned works are usually scene-agnostic, employing prior knowledge can help in the viewpoint extrapolation task: this has been well explored in the area of 3D & 4D Human Performance Capture. A great overview of the development of the *Virtualized Reality* system developed at CMU in the 90s is presented in [32]. It is one of the first such systems and uses multiple cameras for full 4D capture. Starting from this work, there is a continuous line of work refining and improving over multi-view capture of human performances [15, 18, 36, 60, 77]. Recently, Relightables [25] uses again a multi-camera system and adds controlled lighting to the capture set up, so that the resulting reconstructed performances can be replayed in new lighting conditions. The authors of [30, 67] take a different route: they find a way to use bundle adjustment for triangulation of tracked 3D points and obtain results with sub-frame time accuracy. Broxton et al. [7] is one of the latest systems for general-purpose view interpolation and uses a multi-view capture system to create a layered mesh representation of the scene. Li et al. [37] use a similar multi-view capture system to train a dynamic Neural Radiance Field. All of these systems use multiple cameras and are unable to transmit performance in real-time.

Point-based Rendering. We assume a single input RGB-D sensor as a data source for our method. This naturally allows us to work with the depth data in a point-cloud format. To use this for end-to-end optimization, we build on top of ideas from differentiable point cloud rendering. Some of the first methods rendered point clouds by blending discrete samples using local blurring kernels: [28, 38, 55]. Using the differentiable point cloud rendering together with convolutional neural networks naturally enables the use of latent features and a deferred rendering layer, which has been explored in [33, 69]. Aliev et al. [1] use a point renderer implemented in OpenGL, then use a neural network image space to create novel views. Ruckert et al. [56] use purely pixel-sized points and finite differences for optimization. Dai et al. [16] use a layered intermediate representation during the point cloud projection. Bui et al. [8] specifically address the resolution problem of point clouds and proposes a neural network for densification of point cloud renderings. We are directly building on these methods and use the Pulsar renderer [33] in our method together with an additional model to improve the point cloud density.

Warping Representations. To correctly render occluded regions, we warp the respective image regions from an unoccluded posture to the required posture. Debevec et al. [19] is one of the first methods to use 'projective texture-mapping' for view synthesis. Chaurasia et al. [11] uses depth synthesis and local warps to improve over image-based rendering. The authors of [23, 76] take view synthesis through warping to its extreme: they solely use warps to create novel views or synthesize gaze. Recent methods [48, 53, 54, 64] use 3D proxies together with warping and a CNN to generate novel views. All these methods require either creation of an explicit 3D proxy first, or use of image-based rendering. Instead, we use the dynamic per-frame point cloud together with a pre-captured, unoccluded image to warp necessary information into the target view during online processing.

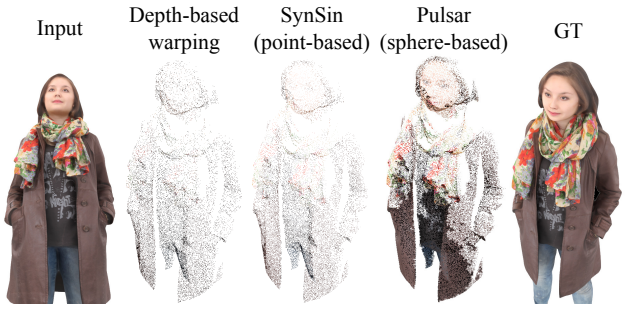

Figure 2. *Comparison of 3D point cloud transformations.* From a single RGB-D input, we obtain the warped image using: a depth-based warping transformation [35,41], neural point-based renderer SynSin [69] and neural sphere-based Pulsar renderer [33]. The novel image warped by Pulsar is significantly denser.

3. Proposed method

The goal of our method is to create realistic novel views of a human captured by a single RGB-D sensor (with sparse depth, similar to low-cost RGB-D camera), as faithful and fast as possible. We assume that the camera parameterization of the view to generate is known. Still, this poses several challenges: 1) the information we are working with is incomplete, since not all regions that are visible from the novel view can be observed by the RGB-D sensor; 2) occlusion adds additional regions with unknown information; 3) even the pixels that are correctly observed by the original sensor are sparse and exhibit 'holes' when viewed from a different angle. We tackle the aforementioned problems using an end-to-end trainable neural network with two components. First, given an RGB-D image parameterized as its two components RGB I_v and sparse depth D_v taken from the input view v, a sphere-based view synthesis model S produces dense features of the target view and renders the resulting RGB image from the target camera view using a global context inpainting network G (see Sec. 3.1). However, this first network can not fully resolve all occlusions (even assuming it would be trained 'perfectly'): information from fully occluded regions is missing (*e.g.* rendering a pattern on a T-shirt that is occluded by a hand). To account for such cases, we optionally extend our model with an enhancer module E (see Sec. 3.2). It uses information from a prior, an unoccluded snapshot of the same person,

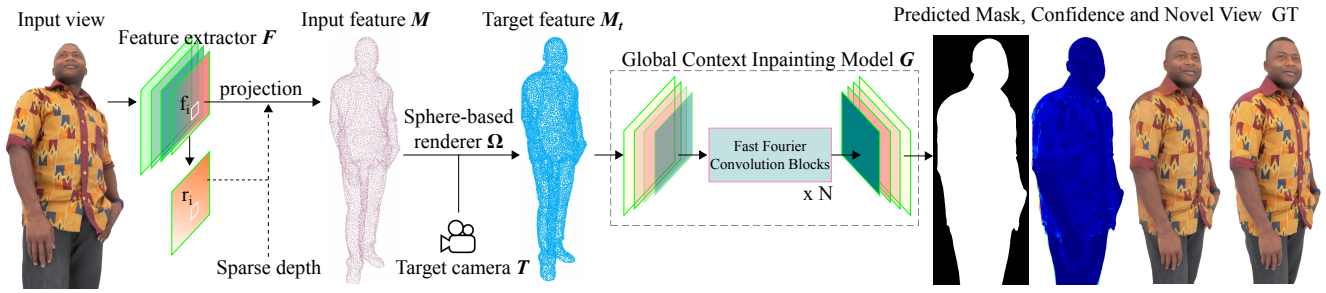

Figure 3. *Architecture of the sphere-based view synthesis network.* The feature predictor F learns radius and feature vectors of the sphere set S. We then use the sphere-based differentiable renderer Ω to densify the learned input features M and warp them to the target camera T. The projected features M_t are passed through the global context inpainting module G to generate the foreground mask, confidence map and novel image. Brighter colors of the confidence map indicate lower confidence.

estimates the dense correspondences between the predicted novel view and prior view, and then refines the predicted result.

3.1. Sphere-based View Synthesis

The goal of this first part of our pipeline is to render a sparse RGB-D view of a human as faithfully as possible from a different perspective. Of the aforementioned artifacts, it can mostly deal with the inherent sparsity of spheres that is caused due to the depth foreshortening: from a single viewpoint in two neighboring pixels, we only get a signal at their two respective depths—no matter how much they differ. This means that for every two pixels that have a large difference in depth and are seen from the side, large gaps occur. For humans, these 'gaps' are of limited size, and we can address them to a certain extent by using a model using a sphere-based renderer for view synthesis.

Sphere-based renderer. Given the depth of every pixel from the original viewpoint as well as the camera parameters, these points can naturally be projected into a novel view. This makes the use of depth-based warping or of a differentiable point- or sphere-renderer a natural choice for the first step in the development of the view synthesis model. The better this renderer can transform the initial information into the novel view, the better; this projection step is automatically correct (except for sensor noise) and not subject to training errors.

In Fig. 2, we compare the density of the warped images from a single sparse RGB-D input using three different methods: depth-based warping [35], point-based rendering [69] and sphere-based rendering [33]. Depth based warping [35] represents the RGD-D input as a set of pixel-sized 3D points and thus, the correctly projected pixels in the novel view are very sensitive to the density of the input view. The widely-used differentiable point-based renderer [69] introduces a global radius-per-point parameter which allows to produce a somewhat denser images. This comes, however, with a trade-off: if the radius is selected too large, details in dense regions of the input image are lost; if the radius is selected too small, the resulting images get sparser in sparse regions. The recently introduced, sphere-based Pulsar renderer [33] not only provides the option to use a per-sphere radius parameter, but it also provides gradients for these radiuses, which enables to set them dynamically. As depicted in Fig. 2, this allows us to produce denser images compared to the other methods. Fig. 3 shows an overview of the overall architecture of our method. In a first step, we use a shallow set of convolutional layers F to encode the input image I_v to a d-dimensional feature map $M = F(I_v)$. From this feature map, we have to create a sphere representation that can be rendered using the Pulsar renderer. This means that we have to find position p_i, feature vector f_i, and radius r_i for every sphere $i \in 1, .., N$ when using N spheres (for further details about the rendering step, we refer to [33]). The sphere positions p_i can trivially be inferred from camera parameters, pixel index and depth for each of the pixels. We choose the features f_i as the values of M at the respective pixel position; we infer r_i by passing M to another convolution layer with a sigmoid activation function to bound its range. This leads to a relatively dense projection of features into the target view, which is the basis for the following steps.

Global context inpainiting model. The projected features have to be converted to the final image. This remains a challenging problem, since several 'gaps' in the re-projected feature images M_t cannot be avoided. To address this, we design an efficient encoder-decoder based inpainting model G to produce the final renders. The encoding bottleneck severely increases the receptive field size of the model, which in turn allows it to correctly fill more of the missing information. Additionally, we employ a series of Fast Fourier Convolutions (FFC) [13] to take into account the image-wide receptive field. The model is able to hallucinate missing pixels much more accurately compared to regular convolution layers [61]. The architecture of the G module is described in detail in the supplementary material.

Photometric Losses. The sphere-based view synthesis network S not only predicts an RGB image of the target view, but also a foreground mask and a confidence map which can be used for compositing and error correction, respectively. It is trained end-to-end using the photometric loss $\mathcal{L}_{photo}$, which is defined as: $\mathcal{L}_{photo} = \mathcal{L}_i + \mathcal{L}_m$. $\mathcal{L}_i$ is the combination of an ℓ_1, perceptual [12] and hinge GAN [24] loss between the estimated new view I_p and the ground-truth image I_{GT}. $\mathcal{L}_m$ is the binary cross entropy loss between the predicted and ground-truth foreground mask. We found that this loss encourages the model to predict sharp contours in the novel image.

These two losses lead to high quality reconstruction results for single images. However, we find that stereoscopic rendering of novel views requires matching left and right images for both views and whereas the above losses lead to *plausible* reconstructions, they do not necessarily lead to sufficiently consistent reconstructions for closeby viewpoints. We found a two-step strategy to address this issue: 1) Instead of predicting a novel image of a single viewpoint, we train the model to predict two nearby novel views. To obtain perfectly consistent depth between both views, we use the warping operator W from Jaderberg *et al*. [29] to warp the predicted image and the depth from one to the nearby paired viewpoint. 2) In the second step, we define a multi-view consistency loss $\mathcal{L}_c$ as:

$$\mathcal{L}_c = ||I_p^L - W(I_p^R)||_1, \tag{1}$$

where I_p^L and I_p^R are predicted left and right novel views. With this, we define the photometric loss as follows:

$$\mathcal{L}_{photo} = \mathcal{L}_i + 0.5 \times \mathcal{L}_m + 0.5 \times \mathcal{L}_c. \tag{2}$$

3.2. Enhancer Network E

The sphere-based view synthesis network S predicts plausible novel views with high-quality. However, if the person is holding an object such as a wallet in Fig. 4 or if their hands are obstructing large parts of their torso then the warped transformation will result in missing points in this region (as discussed in Fig 2). This leads to low-fidelity texture estimates for those occluded regions when performing novel view synthesis with a target camera that is not close to the input view. Hence, to further enhance the quality of the novel views, we introduce two additional modules: *i)* an *HD-IUV* predictor D to predict dense correspondences between an RGB image (*i.e*., render of a human) and the 3D surface of a human body template, and *ii)* a refinement module R to warp an additional occlusion-free input to the target camera and enhance the initial estimated novel view to tackle the self-occlusion issue.

HD-IUV Predictor D. We first estimate a representation that maps an RGB image of a human to the 3D surface of a body template [45–47]. We build upon DensePose [45] in terms of how the surface correspondences are established as well as the supervisions employed but instead we train on synthetic data since DensePose predictions tend to be noisy and not very accurate. We feed the initially estimated novel view I_p to an encoder-decoder architecture which contains three prediction heads (for the I, U and V channels).

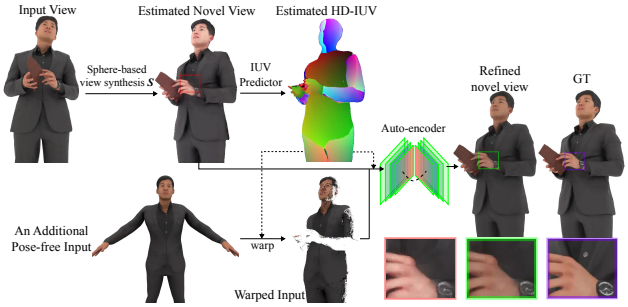

Figure 4. *IUV-based image refinement.* Using an additional occlusion-free input, we refine the initial estimated novel view by training the Enhancer E network. We infer the dense correspondences of both predicted novel view and occlusion-free image using a novel *HD-IUV* module. The occlusion-free image is warped to the target view and then refined by an auto-encoder. The refined novel view shows better result on the occluded area compared to the initial estimated.

Warping Representations and View Refinement. The predicted *HD-IUV* in isolation would not be useful for the task of novel-view synthesis of humans. However when used along with a single occlusion-free RGB input, we are able to warp all visible pixels to the human in the target camera T and obtain a partial warp image I_w. We then feed both I_p and I_w to another refinement module to fix the rendering artifacts of the occluded regions at the target viewpoint. This refinement module is trained using the photometric loss $\mathcal{L}_{photo}$ between the refined novel images and ground-truths. For all training and warping details, and the detailed network architecture, we refer the reader to the supplementary material.

4. Experiments

Datasets. Our proposed approach is trained solely on synthetic data and evaluated quantitatively and qualitatively on both, synthetic as well as real data. For synthetic data, we use the RenderPeople dataset [17], which has been used extensively [2, 6, 10, 27, 34, 50, 57, 62, 74] for human reconstruction and generation tasks. Overall, we use a subset of 1000 watertight meshes of persons wearing a variety of garments and in some cases holding objects such as mugs, bags or mobile phones. Whereas this covers a variety of personal appearance and object interaction, all of these meshes are static—the coverage of the pose space is still lacking. Hence, we augment the dataset by introducing additional pose variations: we perform non-rigid registration for all

Method	RenderPeople (static scans)			RenderPeople (animated scans)			Real 3dMD Data		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
LookingGood† [41]	0.24	0.925	25.32	0.25	0.912	24.53	0.29	0.863	25.12
SynSin† [69]	0.31	0.851	24.18	0.35	0.937	23.64	0.35	0.937	22.18
SynSin [69]	0.52	0.824	22.45	0.55	0.853	20.86	0.65	0.819	19.92
HVS-Net†	**0.14**	**0.986**	**28.56**	**0.17**	**0.958**	27.41	**0.20**	**0.918**	**26.47**
HVS-Net	0.15	**0.986**	28.54	**0.17**	0.955	**27.45**	**0.20**	**0.918**	**26.47**

Table 1. *Quantitative results on synthetic and real-capture images.* For all datasets, the metrics are averaged across all views. Methods with a † symbol are using dense input depth. Both HVS-Net and HVS-Net† achieve best results compared to other view synthesis methods.

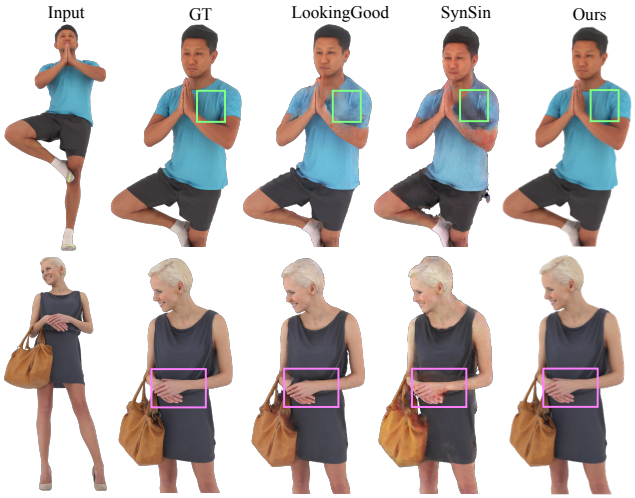

Figure 5. *Qualitative comparison.* Examples of generated novel views by HVS-Net and state-of-the-art methods on the testing set of RenderPeople [17] dataset. While LookingGood [41] uses denser input depth, we use a set of sparse 3D points as input to SynSin [69] and our proposed HVS-Net.

meshes, rig them for animation and use the Mixamo motion capture dataset [44] to animate them in an automatic fasion. The Mixamo dataset provides human animations from which we collect a set of 2,446 sequences covering a wide range of action categories of daily activities and sports.

With this set of meshes *and* animations, we are able to assemble a set of high quality ground-truth RGB-D renders as well as their corresponding IUV maps for 25 views per frame using Blender. We use a 90/10 train/test split based on identities to evaluate whether our model can generalize well to unseen individuals. In addition to the synthetic test set we also assemble a real-world test dataset consisting of 3dMD 4D scans people in motion. The 3dMD 4D scanner is a full body scanner that captures unregistered volumetric point clouds at 60Hz. We use this dataset solely for testing to investigate how well our method handles the domain gap between synthetic and real data. The 3dMD data do not include object interactions, but are generally noisier and have complex facial expressions. To summarize our training set comprised 950 static scans at their original pose and ∼10000 posed scans after animation. Our test-set consisted of 50 static unseen identities along with 1000 animated renders as well as ∼3000 frames of the two humans captured with a 3dMD scan.

Baselines and Metrics. In this evaluation, we compare our approach to two novel view synthesis baselines by comparing the performance in generating single, novel-view RGB images. To evaluate the generalization of HVS-Net, we compare with LookingGood [41]. Since there is no available source code of LookingGood, we reimplemented the method for this comparison. We followed the stereo set up of LookingGood and use a dense depth map to predict the novel views. Furthermore, we compare HVS-Net with the recently proposed view synthesis method SynSin [69], which estimates monocular depth using a depth predictor. To create fair evaluation conditions, we replace this depth predictor and either provide dense or sparse depth maps as inputs directly. We report the PSNR, SSIM, and perceptual similarity (LPIPS) [73] of view synthesis between HVS-Net and other state-of-the-art methods.

4.1. Results

In Tab. 1 and Fig. 5, we summarize the quantitative and qualitative results for samples from the RenderPeople dataset. We first compare the full model HVS-Net against a variant HVS-Net†, which utilizes a dense map as an input. We observe no significant differences between the predicted novel views produced by HVS-Net when trained using either sparse or dense depth input. This confirms the effectiveness of the sphere radius predictor: it makes HVS-Net more robust w.r.t. input point cloud density.

In a next step, we evaluate HVS-Net against the current top performing single view human synthesis methods [41, 69], which do not require per-subject finetuning. Even though we use dense depth maps as input to LookingGood† [41], the method still struggles to produce realistic results if the target pose deviates significantly from the input viewpoint. In the 1^{st} row of Fig. 5,

Figure 6. *Generalization to real-world examples.* Our method generalizes well to real-world 4D data and shows robustness w.r.t to different target poses. These results are produced using HVS-Net, trained solely on synthetic data without further fine-tuning.

LookingGood† [41] also struggles to recover clean and accurate textures of the occluded regions behind the hands of the person. Although both SynSin [69] and HVS-Net, utilize the same sparse depth input, the rendered target images are notably different. Synsin [69] not only performs poorly on the occluded regions but also produces artifacts around the neck of the person, visible in the 2^{nd} row of Fig. 5. In contrast, our method is not only able to render plausible and realistic novel views, but creates them also faithful w.r.t. the input views. Notice that HVS-Net is able to predict fairly accurate hair for both subjects given very little information.

In a last experiment, we test the generalization ability of our method on real-world 4D data, shown in Fig. 6. Being trained only on synthetic data, this requires generalization to novel identity, novel poses, and bridging the domain gap. In the 4D scans, the subjects are able to move freely within the capture volume. We use a fixed, virtual 3D sensor position to create the sparse RGB-D input stream for HVS-Net. The input camera is placed near the feet of the subjects and is facing up. This is a realistic scenario, for example, for VR applications in which an outside-in sensor is used to track a subject in a limited movement area—however it complicates view synthesis, since information about the hair, upper arms and shoulders is not readily available. As can be seen in Fig. 1 and Fig. 6, HVS-Net is still able to perform novel view synthesis with high quality. Despite using sparse input depth, our method is able to recover realistic textures on the clothes of both subjects. In addition, facial expressions such as opening the mouth or smiling are also well-reconstructed, despite the fact that the static or animated scans used to train our network did not have a variety of facial expressions. The quality of the results obtained in Fig. 6 demonstrates that our approach can render high-fidelity novel views of real humans in motion. We observe that the generated novel views are also temporally consistent across different target view trajectories. For additional results and video examples, we refer the reader to the supplementary material.

	LPIPS↓	SSIM↑	PSNR↑
No sphere representation	0.22	0.934	26.15
No global context reasoning	0.21	0.954	26.82
No HE-Net	0.18	0.967	27.92
HVS-Net (full)	**0.15**	**0.986**	**28.54**

Table 2. *HVS-Net architecture ablation study.* Reconstruction accuracy on novel view synthesis on the RenderPeople testing set.

4.2. Ablation Studies

Model Design. Tab. 2 and Fig. 7 summarize the quantitative and qualitative performance for different model variants on the test set of the RenderPeople dataset [17]. HVS-Net without the sphere-based representation does not produce plausible target views which can be obvious if one looks at the rendered face which is blurry compared to the full model. This is due to the high level of sparsity of the input depth, which leads to a harder inpainting problem for the neural network that addresses this task. Replacing the Fast Fourier Convolution residual blocks of the global context inpainting model with regular convolution layers leads to an improvement in the quality of the rendered face, but the occluded region (red box) behind the hands is rendered in an unrealistic way. Using the proposed model architecture, but without the enhancer (visible in the 3^{rd} column of Fig. 7), improves the rendering of the occluded region to a certain extent, however some details remain unrecognizable. In contrast, the full proposed model using the enhancer is able to render the logo accurately.

Sparse Depth Robustness. In Fig. 8, we show novel view synthesis results using different levels of sparsity of the input depth maps. We first randomly sample several versions of the sparse input depth and HVS-Net to process them.

Our method is able to maintain the quality of view synthesis despite strong reductions in point cloud density. This

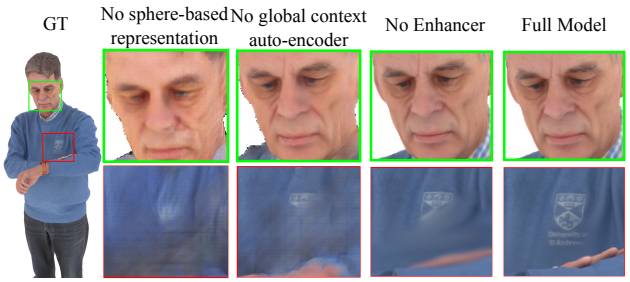

Figure 7. *Qualitative ablation study.* Comparison of the ground-truth with predicted novel views by HVS-Net without the sphere-based representation, without Fast Fourier Convolution, without the Enhancer module and the full model.

highlights the importance of the proposed sphere-based rendering component and the enhancer module. We still observe a slight drop of performance when using 5% or 10% of the input maps. Thus, we suggest that using 25% of the input depth data is sufficient to achieve similar results compared to using the full data.

Inference Speed. For AR/VR applications, it is important to synthesize novel views of humans as fast as possible. This was they key reason why we did not opt for complex architectures such as vision transformers [20] as their inference speed would be prohibitively expensive. During testing, HVS-Net infers 1024×1024 resolution novel images at 21 fps using a single GPU NVIDIA V100. Note that this speed can be further increased with more efficient data loaders as well as optimized models that use the NVIDIA TensorRT engine. Finally we also observed that different levels of depth sparsity do not affect the average runtime of HVS-Net which is a plus compared to prior work.

4.3. Discussion

Limitations. Despite producing appealing results on real-world data, the proposed method is trained solely on synthetic data. It manages to bridge the domain gap remarkably well, however we believe its performance could be further improved by integrating real-world data into the training set.

However, gathering such data is not trivial: generating (close to) noise-free point clouds for training requires elaborate multi-view capture system, possibly enhanced with controlled lighting to simulate varying lighting conditions. A way to circumvent this partially is to train on a large-scale synthetic dataset [70] and then fine-tuning on a smaller-scale real-world dataset. This, at least, reduces the amount of data that has to be captured. Another limitation we identified is that the warped image used as input to the enhancer model has lower quality compared to the initial estimated novel view. This is independent of the quality of the IUV mapping and is an inherent problem of the differentiable warping operation. Improving this operation could be a promising direction for future work that could increase the upper bound in quality for the novel view synthesis of fine structures in occlusion scenarios.

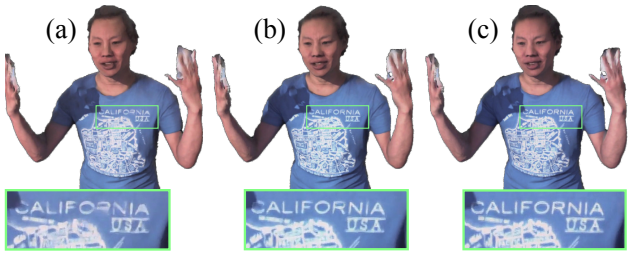

Figure 8. *HVS-Net sparsity robustness.* We randomly sample (a) 5%, (b) 10% and (c) 25% of foreground points as a input depth map and feed it to HVS-Net to predict novel views.

Broader Impact. View synthesis of human subjects presents challenges worth discussing that are not present in other view synthesis scenarios: we need to ensure that such methods perform equally well across different demographics. Unlike real-world data, which is usually captured in specific geographic regions, synthetic data is better suited to avoid biases in this regard, since it is easier to engineer it to equally represent humanity as a whole. Such an effort would enable the proposed method to work well across clothing garments, genders and skin tones. While the design of our method is well suited to be used in this setting, we did not explicitly take dataset bias of the RenderPeople dataset into account in our evaluations.

5. Conclusion

We presented HVS-Net, a method that performs novel view synthesis of humans in motion given a single, sparse RGB-D source. HVS-Net uses a sphere-based view synthesis model that produces dense features of the target view; these are then utilized along with an autoencoder to fill-in the missing details of the target viewpoints. To account for heavily occluded regions, we propose an enhancer module that uses an additional unoccluded view of the human to provide additional information and produce high quality results. Using losses that encourage consistency across views, our method generates hiqh quality results not only for single views, but also for stereoscopic views and across time. Our approach generates high-fidelity renders at new views of unseen humans in various new poses and can faithfully capture and render facial expressions that were not present in training. This is especially remarkable, since we train HVS-Net only on synthetic data; yet it achieves high-quality results across synthetic as well as a real-world dataset.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020. 3

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 5

[3] Luca Ballan, Gabriel J Brostow, Jens Puwein, and Marc Pollefeys. Unstructured video-based rendering: Interactive exploration of casually captured videos. In *ACM SIGGRAPH 2010 Papers, SIGGRAPH 2010*, 2010. 2

[4] Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 4d visualization of dynamic events from unconstrained multi-view videos. In *CVPR*, 2020. 1, 2

[5] Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. X-fields: Implicit neural view-, light- and time-image interpolation. *ACM Trans. Graph.*, 39(6), Nov. 2020. 2

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3D people from images. In *ICCV*, 2019. 5

[7] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Trans. Graph.*, 39(4), July 2020. 2, 3

[8] Giang Bui, Truc Le, Brittany Morago, and Ye Duan. Point-based rendering enhancement via deep learning. *Vis. Comput.*, 34(6–8):829–841, June 2018. 3

[9] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3):569–577, July 2003. 2

[10] Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, and Tony Tung. Semi-supervised synthesis of high-resolution editable textures for 3d humans. In *CVPR*, 2021. 5

[11] Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Trans. Graph.*, 32(3), July 2013. 3

[12] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1520–1529. IEEE Computer Society, 2017. 5

[13] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4479–4488. Curran Associates, Inc., 2020. 4

[14] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[15] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), July 2015. 3

[16] Peng Dai, Yinda Zhang, Zhuwen Li, Shuaicheng Liu, and Bing Zeng. Neural point cloud rendering via multi-plane projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7830–7839, 2020. 3

[17] RenderPeople Dataset. *http://renderpeople.com/*. 5, 6, 7

[18] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *ACM Trans. Graph.*, 27(3):1–10, Aug. 2008. 3

[19] Paul Debevec, Yizhou Yu, and George Borshukov. Efficient View-Dependent Image-Based Rendering with Projective Texture-Mapping. *Eurographics Rendering Workshop*, 4(11):105–116, 1998. 3

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 8

[21] John Flynn, Michael Broxton, Paul Debevec, Matthew Duvall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 2362–2371, 2019. 2

[22] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deep stereo: Learning to predict new views from the world's imagery. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5515–5524, 2016. 2

[23] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor S. Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 311–326. Springer, 2016. 3

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 5

[25] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6), Nov. 2019. 3

[26] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *ECCV*, 2018. 2

[27] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 5

[28] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. In *Proceedings of the IEEE International Conference on Neural Information Processing Systems (NeurIPS)*, pages 2802–2812, 2018. 3

[29] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2017–2025, Cambridge, MA, USA, 2015. MIT Press. 5

[30] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3334–3342, 2015. 3

[31] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Trans. Graph.*, 35(6), Nov. 2016. 2

[32] T. Kanade, P. Rander, and P.J. Narayanan. Virtualized reality: constructing virtual worlds from real scenes. *IEEE MultiMedia*, 4(1):34–47, 1997. 3

[33] Christoph Lassner and Michael Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1440–1449, 2021. 3, 4

[34] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *3DV*, 2019. 5

[35] Hoang-An Le, Thomas Mensink, Partha Das, and Theo Gevers. Novel view synthesis from a single image via point cloud transformation. In *BMVC*, 2020. 3, 4

[36] Hao Li, Linjie Luo, Daniel Vlasic, Pieter Peers, Jovan Popović, Mark Pauly, and Szymon Rusinkiewicz. Temporally coherent completion of dynamic shapes. *ACM Trans. Graph.*, 31(1), Feb. 2012. 3

[37] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis. *CoRR*, abs/2103.02597, 2021. 3

[38] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *AAAI Conference on Artificial Intelligence*, 2018. 3

[39] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019. 2

[40] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics*, 40(4):1–13, 2021. 2

[41] Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, Adarsh Kowdle, Christoph Rhemann, Dan B Goldman, Cem Keskin, Steve Seitz, Shahram Izadi, and Sean Fanello. Lookingood: Enhancing performance capture with real-time neural re-rendering. *ACM Trans. Graph.*, 37(6), Dec. 2018. 2, 3, 6, 7

[42] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 6871–6880, 2019. 2

[43] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[44] Mixamo. *Mixamo Dataset*. https://www.mixamo.com/. 6

[45] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *ECCV*, 2018. 5

[46] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. *arXiv preprint arXiv:2011.12438*, 2020. 5

[47] Natalia Neverova, Artsiom Sanakoyeu, Patrick Labatut, David Novotny, and Andrea Vedaldi. Discovering relationships between object categories via universal canonical maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 404–413, 2021. 5

[48] Phong Nguyen, Animesh Karnewar, Lam Huynh, Esa Rahtu, Jiri Matas, and Janne Heikkila. Rgbd-net: Predicting color and depth images for novel views synthesis. *arXiv preprint arXiv:2011.14398*, 2020. 1, 3

[49] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *International Conference on Computer Vision*, 2021. 1, 2

[50] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, 2021. 5

[51] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 1, 2

[52] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2

[53] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, 2020. 3

[54] Gernot Riegler and Vladlen Koltun. Stable View Synthesis. *Proceedings of the IEEE Conference on Computer Vi-*

sion and Pattern Recognition, pages 12216—-12225, 2021. 3

[55] Riccardo Roveri, Lukas Rahmann, Cengiz Oztireli, and Markus Gross. A network architecture for point cloud classification via automatic depth images generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4184, 2018. 3

[56] Darius Rückert, Linus Franke, and Marc Stamminger. Adop: Approximate differentiable one-pixel point rendering. *arXiv preprint arXiv:2110.06635*, 2021. 3

[57] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, 2020. 5

[58] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In King N. Ngan, Thomas Sikora, and Ming-Ting Sun, editors, *Visual Communications and Image Processing 2000*, volume 4067, pages 2 – 13. International Society for Optics and Photonics, SPIE, 2000. 2

[59] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 175–184, 2019. 2

[60] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 3

[61] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 4

[62] Feitong Tan, Danhang Tang, Dou Mingsong, Guo Kaiwen, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, Sean Fanello, Ping Tan, and Yinda Zhang. Humangps: Geodesic preserving feature for dense human correspondences. In *CVPR*, 2021. 5

[63] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B Goldman, and M. Zollhöfer. State of the art on neural rendering. *Computer Graphics Forum*, 39(2):701–727, 2020. 2

[64] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. IGNOR: Image-guided Neural Object Rendering. *International Conference on Learning Representations*, 2020. 3

[65] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 1

[66] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video, 2020. 1, 2

[67] Minh Vo, Srinivasa G. Narasimhan, and Yaser Sheikh. Spatiotemporal bundle adjustment for dynamic 3d reconstruction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1710–1718, 2016. 3

[68] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 1

[69] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6 2020. 2, 3, 4, 6, 7

[70] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone, 2021. 8

[71] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

[72] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5335–5344. Computer Vision Foundation / IEEE, 2020. 1, 2

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

[74] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G Narasimhan, and Minh Vo. TexMesh: Reconstructing detailed human texture and geometry from RGB-D video. In *ECCV*, 2020. 5

[75] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4), July 2018. 2

[76] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European Conference on Computer Vision*, 2016. 3

[77] C. Zitnick, Sing Bing Kang, Matt Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph.*, 23:600–608, 08 2004. 2, 3