Free-Viewpoint RGB-D Human Performance Capture and Rendering

Phong Nguyen-Ha[1*], Nikolaos Sarafianos[2],

Christoph Lassner[2], Janne Heikkilä[1], and Tony Tung[2]

[1] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[2] Meta Reality Labs Research, Sausalito
https://www.phongnhhn.info/HVS_Net

Abstract. Capturing and faithfully rendering photorealistic humans from novel views is a fundamental problem for AR/VR applications. While prior work has shown impressive performance capture results in laboratory settings, it is non-trivial to achieve casual free-viewpoint human capture and rendering for unseen identities with high fidelity, especially for facial expressions, hands, and clothes. To tackle these challenges we introduce a novel view synthesis framework that generates realistic renders from unseen views of any human captured from a single-view and sparse RGB-D sensor, similar to a low-cost depth camera, and without actor-specific models. We propose an architecture to create dense feature maps 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 that our method generates high-quality novel views of synthetic and real human actors given a single-stream, sparse RGB-D input. It generalizes to unseen identities, and new poses and faithfully reconstructs facial expressions. Our approach outperforms prior view synthesis methods and is robust to different levels of depth 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 [42,45,62]. However, synthesizing novel views of humans in motion requires methods to handle dynamic scenes with various deformations which is a challenging task [62,67]; especially in those regions with fine details such as the face or the clothes [46,50,63,66]. In addition, prior work usually relies on a large amount of cameras [5,42], expensive capture setups [51], or inference time on the order of several minutes per frame. This work aims to tackle these challenges using a compact, yet effective formulation.

We propose a novel **H**uman **V**iew **S**ynthesis **Net**work (**HVS-Net**) that generates high-fidelity rendered images of clothed humans using a commodity 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) the ability to handle dynamic

*This work was conducted during an internship at Meta Reality Labs Research.

Fig. 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.

behavior of humans in unseen poses as opposed to animating humans using the same poses seen at training, iii) the ability to handle occlusions (either from objects or self-occlusion), iv) capturing facial expressions and v) the generation of high-fidelity images in a live setup given a single-stream, sparse RGB-D input (similar to a low-cost, off-the-shelf depth camera).

HVS-Net 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). 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 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 create 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 image 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. We train the entire pipeline end-to-end using photometric losses between the generated and ground-truth pair of images. In addition, we use stereoscopic rendering to encourage view-consistent results between close-by viewpoints. To train HVS-Net, we rely on high-quality synthetic scans of hu-

mans 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 [31,32,51]. In summary, the contributions of this work are:

- A robust sphere-based synthesis network that generalizes to multiple identities without per-human optimization.
- A refinement module that enhances the self-occluded regions of the initial estimated novel views. This is accomplished by introducing a novel yet simple approach to establish dense surface correspondences for the clothed human body that addresses key limitations of DensePose which is usually used for this task.
- State-of-the-art results on dynamic humans wearing various clothes, or accessories and with a variety of facial expressions 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 [56,60]. 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. [70] similarly use a multi-view capture setup for viewpoint interpolation. These approaches interpolate between recorded images or videos. Ballan et al. [4] coin the term 'video-based rendering': they use it to interpolate between hand-held camera views of performances. The strong generative capabilities of neural networks enable further extrapolation and relaxation of constraints [18,22,28,41]. Zhou et al. [69] introduce Multi-Plane Images (MPIs) for viewpoint synthesis and use a model to predict them from low-baseline stereo input and [17,57] improve over the original baseline and additionally work with camera arrays and light fields. Broxton et al. [8] extend the idea to layered, dynamic meshes for immersive video experiences whereas Bansal et al. [5] use free camera viewpoints, but multiple cameras. With even stronger deep neural network priors, [64] performs viewpoint extrapolation from a single view, but for static scenes, whereas [62,67] can work with a single view in dynamic settings with limited motion. Bemana et al. [6] 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 [38,39] 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 the 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 [29]. 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 [1,15,35,70]. Relightables [21] 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 [27] 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. [8] 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. Many recent works apply neural radiance fields [42,65] to render humans at novel views. Li et al. [36] use a similar multi-view capture system to train a dynamic Neural Radiance Field. Kwon et al. [31] learn generalizable neural radiance fields based on a parametric human body model to perform novel view synthesis. However, this method fails to render high-quality cloth details or facial expressions of the human. Both of these systems use multiple cameras and are unable to transmit performance in real-time. Given multi-view input frames or videos, recent works on rendering animate humans from novel views show impressive results [46,50,51,66]. However such methods can be prohibitively expensive to run ([46] runs at 1 minute/frame) and cannot generalize to unseen humans but instead create a dedicated model for each human that they need to render.

Human View Synthesis using RGB-D. A few methods have been published recently that tackle similar scenarios: LookingGood [40] re-renders novel viewpoints of a captured individual given a single RGB-D input. However, their capture setup produces dense geometry which makes this a comparatively easy task: the target views do not deviate significantly from the input views. A recent approach [48] uses a frontal input view and a large number of calibration images to extrapolate novel views. This method relies on a keypoint estimator to warp the selected calibrated image to the target pose, which leads to unrealistic results for hands, occluded limbs, or for large body shapes.

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: [25,37,54]. 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 [33,64]. Recent works on point-based rendering [2,30] use a point renderer implemented in OpenGL, then use a neural network image space to create novel views. Ruckert et al. [55] use purely pixel-sized points and finite differences for optimization. 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. [16] 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 [19] take view synthesis through warping to its extreme: they solely use warps to create novel views or synthesize gaze. Recent methods [45,53,61] use 3D proxies together with warping and a CNN to generate novel

Fig. 2. *Comparison of 3D point cloud transformations.* From a single RGB-D input, we obtain the warped image using: a depth-based warping transformation [34,40], the neural point-based renderer SynSin [64] and the neural sphere-based Pulsar renderer [33]. The novel image warped by Pulsar is significantly denser.

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.

3 HVS-Net Methodology

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 a 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: 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 an unoccluded snapshot of the same person, estimates the dense correspondences between the predicted novel view and occlusion-free input view, and then refine the predicted result.

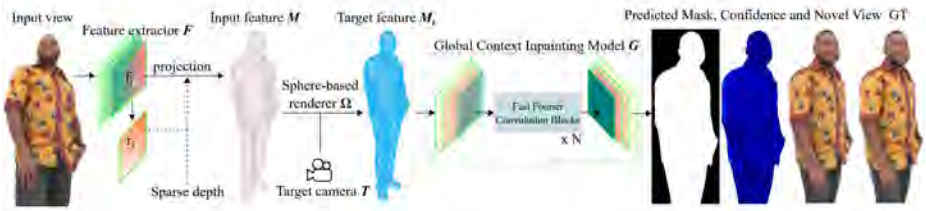

Fig. 3. *Sphere-based view synthesis network architecture.* The feature predictor F learns radius and feature vectors of the sphere set S. We then use the sphere-based differentiable renderer Ω to densify the learned input features M and warp them to the target camera T. The projected features M_t are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

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 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 rendering human subjects, these "gaps" are of limited size, and we can address the problem to a certain extent by 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 [34], point-based rendering [64] and sphere-based rendering [33]. Depth based warping [34] 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 [64] introduces a global radius-per-point parameter which allows to produce a somewhat denser images. Since it uses the same radius for all points, 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 us 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 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 an as-dense-as-possible projection of features into the target view, which is the basis for the following steps.

Global context inpainiting model. Next, the projected features are 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 in 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 [58].

Photometric Losses. The sphere-based view synthesis network S not only predicts an RGB image I_p of the target view, but also a foreground mask I_m and a confidence map I_c which can be used for compositing and error correction, respectively. We then multiply the predicted foreground mask and confidence map with the predicted novel image: $I_p = I_p * I_m * I_c$. However, an imperfect mask I_m may bias the network towards unimportant areas. Therefore, we predict a confidence mask I_c as a side-product of the G network to dynamically assign less weight to "easy" pixels, whereas "hard" pixels get higher importance [40].

All of the aforementioned model components are 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 [20] 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. The two losses lead to high-quality reconstruction results for single images. However, we note that stereoscopic rendering of novel views requires matching left and right images for both views. Whereas the above losses lead to *plausible* reconstructions, they do not necessarily lead to sufficiently consistent reconstructions for close-by 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 [26] 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:

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

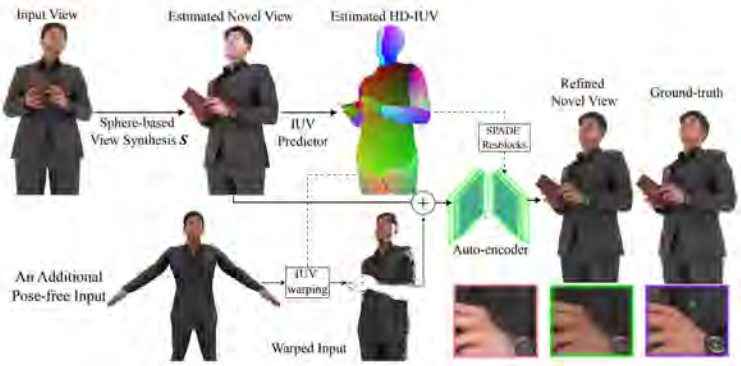

Fig. 4. *IUV-based image refinement.* Using an additional occlusion-free input, we refine the initial estimated novel view by training the Enhancer network E. We infer the dense correspondences of both, predicted novel view and occlusion-free image, using a novel *HD-IUV* module. The occlusion-free image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

3.2 Handling Occlusions

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 (c.t. 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 (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 (e.g. , a selfie in a practical application) to the target camera and enhance the initial estimated novel view to tackling 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 [24,43,44,59]. One could use DensePose [43] for this task but the estimated IUV (where I reflects the body part) predictions cover only the naked body instead of the clothed human and are inaccurate as they are trained based on sparse and noisy human annotations. Instead, we build our own IUV predictor and train it on synthetic data for which we can obtain accurate ground-truth correspondences. With pairs of synthetic RGB images and ground-truth dense surface correspondences, we train a UNet-like network that provides dense surface (i.e. IUV) estimates *for each pixel* of the clothed human body. For each pixel p in the foreground image, we predict 3-channeled (RGB) color p' which represents the correspondence (the colors in such a representation are unique which makes subsequent warping easy). Thus, we treat the whole problem as a multi-task classification problem where each task (predictions for the I, U, and V channels) is trained with the following set of losses: a) multi-task classification loss for each of the 3 channels (per-pixel classification label) and b) silhouette loss. In Fig. 5 we show that, unlike DensePose, the proposed HD-IUV module accurately establishes fine level correspondences for the face and hand regions

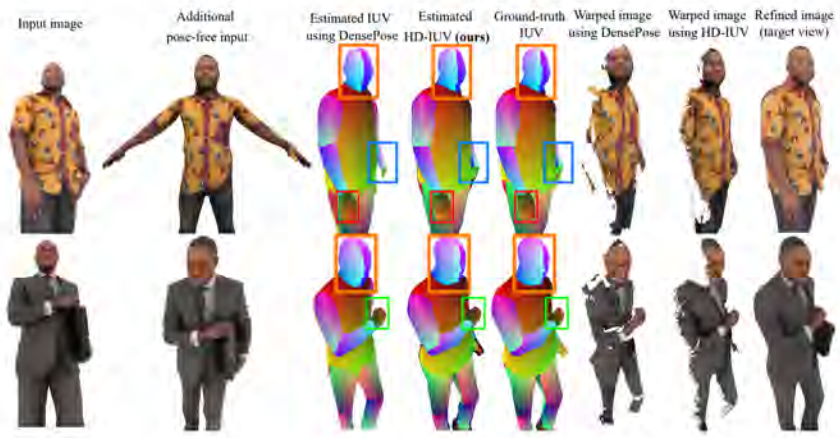

Fig. 5. *Dense correspondence visualization.* Texture warping with DensePose results in inaccurate and distorted images in the target view due to incorrect IUV estimates (enhanced by the fact that it targets the naked body). Our proposed HD-IUV representation covers the human body including clothing, captures facial and hand details with high accuracy, and results in less distorted renderings in the target view. We stack this warped image with the initially estimated target-view synthesized image and provide it as input for the Enhancer network to obtain the final results.

while capturing the whole clothed human and thus making it applicable for such applications. Once this model is pre-trained, we merge it with the rest of the pipeline and continue the training procedure by using the initially estimated novel view I_p ad an input to an encoder-decoder architecture that contains three prediction heads (for the I, U, and V channels). An in-depth discussion on the data generation, network design, and training is provided in the supplementary material.

Warping Representations and View Refinement. The predicted *HD-IUV* in isolation would not be useful for the task of human view synthesis. However, when used along with the occlusion-free RGB input, it allows us to warp all visible pixels to the human in the target camera T and obtain a partial warped image I_w. For real applications this occlusion-free input can be a selfie image—there are no specific requirements to the body pose for the image. In Fig. 5 we compare DensePose results with the proposed HD-IUV module. DensePose clearly produces less accurate and more distorted textures.

In the next step, we stack I_p and I_w and pass the resulting tensor to a refinement module. This module addresses two key details: a) it learns to be robust to artifacts that are originating either from the occluded regions of the initially synthesized novel view as well as texture artifacts that might appear due to the fact that we rely on HD-IUV dense correspondences for warping and b) it is capable of synthesizing crisper results in the occluded regions as it relies on both the initially synthesized image as well as the warped image to the target view based on HD-IUV. The refinement module is trained using the photometric loss $\mathcal{L}_{photo}$ between the refined novel images and ground truths. All details regarding training and image warping, as well as the full network architecture, can be found in the supplementary material.

4 Experiments

Datasets. The proposed approach is trained solely on synthetic data and evaluated quantitatively and qualitatively on both, synthetic and real data. For training, we use the RenderPeople dataset [52], which has been used extensively [3,7,10,23,47,49,59] for human reconstruction and generation tasks. Overall, we use a subset of 1000 watertight meshes of people 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 appearances and object interaction, all of these meshes are static—the coverage of the pose space is lacking. Hence, we augment the dataset by introducing additional pose variations: we perform non-rigid registration for all meshes, rig them for animation and use a set of pre-defined motions to animate them. 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 of 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 does not include object interactions, but is generally noisier and has complex facial expressions. To summarize: our training set comprises 950 static scans in their original pose and ∼10000 posed scans after animation. Our test set includes 50 static unseen identities along with 1000 animated renders and 3000 frames of two humans captured with a 3dMD full-body scanner.

Novel Viewpoint Range. We assume a scenario with a camera viewpoint at a lower level in front of a person (e.g. , the camera sitting on a desk in front of the user). This is a more challenging scenario than LookingGood [40] or Volumetric Capture [48] use, but also a realistic one: it corresponds to everyday video conference settings. At the same time, the target camera is moving freely in the frontal hemisphere around the person (Pitch & Roll: $[-45^o, 45^o]$, $L_x : [-1.8m, 1.8m]$, $L_y : [1.8m, 2.7m]$, $L_z : [0.1m, 2.7m]$ in a Blender coordinate system). Thus, the viewpoint range is significantly larger per input view than in prior work.

Baselines. 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 it with LookingGood [40]. Since there is no available source code of LookingGood, we reimplemented the method for this comparison and validated in various synthetic and real-world settings that this implementation is qualitatively equivalent to what is reported in the original paper (we include comparison images in the supp.mat.). 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 [64], 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. While there are several recently proposed methods in the topic of human-view synthesis; almost all are relying on either proprietary data captured in lab environ-

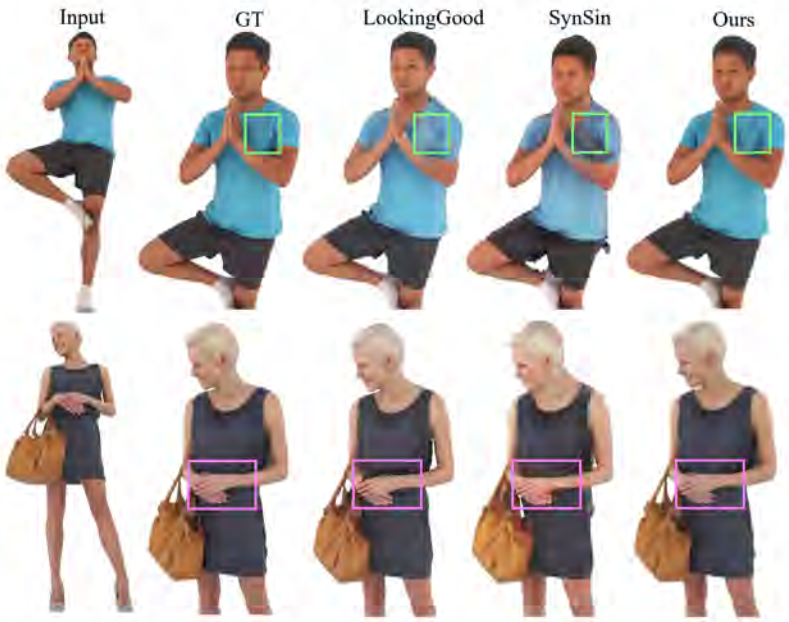

Fig. 6. *Qualitative comparison.* Examples of generated novel views by HVS-Net and state-of-the-art methods on the test set of the RenderPeople [52] dataset. As opposed to all other methods, LookingGood [40] uses dense input depth.

ments [48], multi-view input streams [31,36,51,66] and most importantly none of these works can generalize to new human identities (or for the case of Neural Body [51] not even new poses) at testing time which our proposed HVS-Net can accomplish. Furthermore, inferring new views in a real-time manner is far from solved for most these works. In contrast, our method focuses more on a practical approach of single view synthesis, aiming to generalize to new identities and unseen poses while being fast at inference time. Hence we stick to performing quantitative comparisons against LookingGood [40] and SynSin [64] and we do not compare it with NeRF-based approaches [31,36,51,66] as such comparisons are not applicable.

Metrics. We report the PSNR, SSIM, and perceptual similarity (LPIPS) [68] of view synthesis between HVS-Net and other state-of-the-art methods.

4.1 Results

In Tab. 1 and Fig. 6, 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 [40,64], which do not require per-subject finetuning. Even

Method	RenderPeople (static)			RenderPeople (animated)			Real 3dMD Data		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
LookingGood† [40]	0.24	0.925	25.32	0.25	0.912	24.53	0.29	0.863	25.12
SynSin† [64]	0.31	0.851	24.18	0.35	0.937	23.64	0.35	0.937	22.18
SynSin [64]	0.52	0.824	22.45	0.55	0.853	20.86	0.65	0.819	19.92
HVS-Net (w/o Enhancer)	0.18	0.986	28.54	0.19	0.926	26.24	0.20	0.910	26.25
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 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 the best results compared to other view synthesis methods. We observe a slight drop of performance without using the proposed Enhancer module.

Fig. 7. *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.

though we use dense depth maps as input to LookingGood† [40], 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. 6, LookingGood† [40] also struggles to recover clean and accurate textures of the occluded regions behind the hands of the person. Although both SynSin [64] and HVS-Net utilize the same sparse depth input, the rendered target images are notably different. SynSin [64] 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. 6. 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. 7. 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. As can be seen in Fig. 1 and Fig. 7, HVS-Net is still able to perform novel view synthesis with high quality. Despite using sparse input depth, our method is able to render realistic textures on the clothes of both subjects. In addition, facial expressions such as opening the mouth or smiling

Method Variant	LPIPS↓	SSIM↑	PSNR↑
No Sphere Repres.	0.22	0.934	26.15
No Global Context	0.21	0.954	26.82
No Enhancer	0.18	0.967	27.92
HVS-Net (full)	**0.15**	**0.986**	**28.54**

Input depth (%)	Run-time↑ (fps)	LPIPS↓	SSIM↑	PSNR↑
5	**25**	0.17	0.985	28.27
10	22	0.15	0.986	28.54
25	21	**0.14**	0.986	28.55
100	20	0.14	0.986	**28.56**

Table 2. ***Left****: Ablation study.* Reconstruction accuracy on the RenderPeople testing set. ***Right****: Reconstruction accuracy and inference speed* using different levels of input depth sparsity.

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. 7 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 to the supplementary material.

4.2 Ablation Studies and Discussion

Model Design. Tab. 2 (left) and Fig. 8 summarize the quantitative and qualitative performance for different model variants on the test set of the RenderPeople dataset [52]. HVS-Net without the sphere-based representation does not produce plausible target views (see, for example, 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 a drop in render quality in the occluded region (red box). Using the proposed model architecture, but without the enhancer (5^{th} column of Fig. 8) leads to a loss of detail in texture. In contrast, the full proposed model using the Enhancer network renders the logo accurately. Note that this logo is completely occluded by the human's hands so it is non-trivial to render the logo using a single input image.
Sparse Depth Robustness. In Fig. 9, 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 highlights the importance of the proposed sphere-based rendering component and the enhancer module. As can be seen in Tab. 2 (right), we observe a slight drop of performance when using 5% or 10% of the input maps. To balance between visual quality and rendering speed, 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, a prime target for a method like the one proposed, runtime performance is critical. At test time, HVS-Net generates 1024×1024 images at 21FPS using a single NVIDIA V100 GPU. This speed can be further increased with more efficient data loaders and an optimized implementation that uses the

Fig. 8. *Qualitative ablation study.* Comparison of the ground-truth with predicted novel views by several variants of the proposed HVS-Net.

(a) 5% foreground points (b) 10% foreground points (c) 25% foreground points

Fig. 9. *HVS-Net sparsity robustness.* We randomly sample (a) 5%, (b) 10% and (c) 25% of dense depth points as input depth map and use it as an input for HVS-Net to predict novel views. The text in the T-shirt is reconstructed at high-fidelity with 25% of the depth points utilized.

NVIDIA TensorRT engine. Finally, different depth sparsity levels do not significantly affect the average runtime of HVS-Net, which is a plus compared to prior work.

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 complete 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 based on an novel IUV mapping. 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 and real-world examples.

Acknowledgements: The authors would like to thank Albert Para Pozzo, Sam Johnson and Ronald Mallet for the initial discussions related to the project.

References

1. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. TOG (2008) 4
2. Aliev, K.A., Sevastopolsky, A., Kolos, M., Ulyanov, D., Lempitsky, V.: Neural point-based graphics. In: ECCV (2020) 4
3. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: ICCV (2019) 10
4. Ballan, L., Brostow, G.J., Puwein, J., Pollefeys, M.: Unstructured video-based rendering: Interactive exploration of casually captured videos. In: SIGGRAPH (2010) 3
5. Bansal, A., Vo, M., Sheikh, Y., Ramanan, D., Narasimhan, S.: 4d visualization of dynamic events from unconstrained multi-view videos. In: CVPR (2020) 1, 3
6. Bemana, M., Myszkowski, K., Seidel, H.P., Ritschel, T.: X-fields: Implicit neural view-, light- and time-image interpolation. In: SIGGRAPH Asia (2020) 3
7. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3D people from images. In: ICCV (2019) 10
8. Broxton, M., Flynn, J., Overbeck, R., Erickson, D., Hedman, P., Duvall, M., Dourgarian, J., Busch, J., Whalen, M., Debevec, P.: Immersive light field video with a layered mesh representation. TOG (2020) 3, 4
9. Carranza, J., Theobalt, C., Magnor, M.A., Seidel, H.P.: Free-viewpoint video of human actors. TOG (2003) 3
10. Chaudhuri, B., Sarafianos, N., Shapiro, L., Tung, T.: Semi-supervised synthesis of high-resolution editable textures for 3d humans. In: CVPR (2021) 10
11. Chaurasia, G., Duchene, S., Sorkine-Hornung, O., Drettakis, G.: Depth synthesis and local warps for plausible image-based navigation. TOG (2013) 4
12. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: ICCV (2017) 7
13. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. In: NeurIPS (2020) 7
14. Chibane, J., Bansal, A., Lazova, V., Pons-Moll, G.: Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In: CVPR (2021) 3
15. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. TOG (2015) 4
16. Debevec, P., Yu, Y., Borshukov, G.: Efficient View-Dependent Image-Based Rendering with Projective Texture-Mapping. Eurographics Rendering Workshop (1998) 4
17. Flynn, J., Broxton, M., Debevec, P., Duvall, M., Fyffe, G., Overbeck, R., Snavely, N., Tucker, R.: Deepview: View synthesis with learned gradient descent. In: CVPR (2019) 3
18. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: Deep stereo: Learning to predict new views from the world's imagery. In: CVPR (2016) 3
19. Ganin, Y., Kononenko, D., Sungatullina, D., Lempitsky, V.S.: Deepwarp: Photorealistic image resynthesis for gaze manipulation. In: ECCV (2016) 4
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014) 7
21. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., Tang, D., Tkach, A., Kowdle, A., Cooper, E., Dou, M., Fanello, S., Fyffe, G., Rhemann, C., Taylor, J., Debevec, P., Izadi, S.: The relightables: Volumetric performance capture of humans with realistic relighting. TOG (2019) 4
22. Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., LeGendre, C., Luo, L., Ma, C., Li, H.: Deep volumetric video from very sparse multi-view performance capture. In: ECCV (2018) 3
23. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: ARCH: Animatable reconstruction of clothed humans. In: CVPR (2020) 10

24. Ianina, A., Sarafianos, N., Xu, Y., Rocco, I., Tung, T.: BodyMap: Learning full-body dense correspondence map. In: CVPR (2022) 8
25. Insafutdinov, E., Dosovitskiy, A.: Unsupervised learning of shape and pose with differentiable point clouds. In: NeurIPS (2018) 4
26. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: NeurIPS (2015) 7
27. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV (2015) 4
28. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. TOG (2016) 3
29. Kanade, T., Rander, P., Narayanan, P.: Virtualized reality: constructing virtual worlds from real scenes. IEEE MultiMedia (1997) 3
30. Kopanas, G., Philip, J., Leimkühler, T., Drettakis, G.: Point-based neural rendering with per-view optimization. Computer Graphics Forum (2021) 4
31. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. In: NeurIPS (2021) 3, 4, 11
32. Kwon, Y., Petrangeli, S., Kim, D., Wang, H., Park, E., Swaminathan, V., Fuchs, H.: Rotationally-temporally consistent novel view synthesis of human performance video. In: ECCV (2020) 3
33. Lassner, C., Zollhofer, M.: Pulsar: Efficient sphere-based neural rendering. In: CVPR (2021) 4, 5, 6, 7
34. Le, H.A., Mensink, T., Das, P., Gevers, T.: Novel view synthesis from a single image via point cloud transformation. In: BMVC (2020) 5, 6
35. Li, H., Luo, L., Vlasic, D., Peers, P., Popović, J., Pauly, M., Rusinkiewicz, S.: Temporally coherent completion of dynamic shapes. TOG (2012) 4
36. Li, T., Slavcheva, M., Zollhöfer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Lv, Z.: Neural 3d video synthesis. In: CVPR (2021) 4, 11
37. Lin, C.H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3d object reconstruction. In: AAAI (2018) 4
38. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. TOG (2019) 3
39. Lombardi, S., Simon, T., Schwartz, G., Zollhoefer, M., Sheikh, Y., Saragih, J.: Mixture of volumetric primitives for efficient neural rendering. TOG (2021) 3
40. Martin-Brualla, R., Pandey, R., Yang, S., Pidlypenskyi, P., Taylor, J., Valentin, J., Khamis, S., Davidson, P., Tkach, A., Lincoln, P., Kowdle, A., Rhemann, C., Goldman, D.B., Keskin, C., Seitz, S., Izadi, S., Fanello, S.: Lookingood: Enhancing performance capture with real-time neural re-rendering. TOG (2018) 4, 5, 7, 10, 11, 12
41. Meshry, M., Goldman, D.B., Khamis, S., Hoppe, H., Pandey, R., Snavely, N., Martin-Brualla, R.: Neural rerendering in the wild. In: CVPR (2019) 3
42. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 1, 4
43. Neverova, N., Alp Guler, R., Kokkinos, I.: Dense pose transfer. In: ECCV (2018) 8
44. Neverova, N., Novotny, D., Khalidov, V., Szafraniec, M., Labatut, P., Vedaldi, A.: Continuous surface embeddings. In: NeurIPS (2020) 8
45. Nguyen, P., Karnewar, A., Huynh, L., Rahtu, E., Matas, J., Heikkila, J.: Rgbd-net: Predicting color and depth images for novel views synthesis. In: 3DV (2021) 1, 4
46. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. In: ICCV (2021) 1, 4
47. Palafox, P., Sarafianos, N., Tung, T., Dai, A.: SPAMs: Structured implicit parametric models. In: CVPR (2022) 10

48. Pandey, R., Keskin, C., Izadi, S., Fanello, S., Tkach, A., Yang, S., Pidlypenskyi, P., Taylor, J., Martin-Brualla, R., Tagliasacchi, A., Papandreou, G., Davidson, P.: Volumetric capture of humans with a single rgbd camera via semi-parametric learning. In: CVPR (2019) 4, 10, 11
49. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in geography optimized for regression analysis. In: CVPR (2021) 10
50. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: ICCV (2021) 1, 4
51. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021) 1, 3, 4, 11
52. RenderPeople: http://renderpeople.com/ 10, 11, 13
53. Riegler, G., Koltun, V.: Free view synthesis. In: ECCV (2020) 4
54. Roveri, R., Rahmann, L., Oztireli, C., Gross, M.: A network architecture for point cloud classification via automatic depth images generation. In: CVPR (2018) 4
55. Rückert, D., Franke, L., Stamminger, M.: Adop: Approximate differentiable one-pixel point rendering. arXiv preprint arXiv:2110.06635 (2021) 4
56. Shum, H., Kang, S.B.: Review of image-based rendering techniques. In: Visual Communications and Image Processing (2000) 3
57. Srinivasan, P.P., Tucker, R., Barron, J.T., Ramamoorthi, R., Ng, R., Snavely, N.: Pushing the boundaries of view extrapolation with multiplane images. In: CVPR (2019) 3
58. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: WACV (2022) 7
59. Tan, F., Tang, D., Mingsong, D., Kaiwen, G., Pandey, R., Keskin, C., Du, R., Sun, D., Bouaziz, S., Fanello, S., Tan, P., Zhang, Y.: Humangps: Geodesic preserving feature for dense human correspondences. In: CVPR (2021) 8, 10
60. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Martin-Brualla, R., Simon, T., Saragih, J., Nießner, M., Pandey, R., Fanello, S., Wetzstein, G., Zhu, J.Y., Theobalt, C., Agrawala, M., Shechtman, E., Goldman, D.B., Zollhöfer, M.: State of the art on neural rendering. Computer Graphics Forum (2020) 3
61. Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., Nießner, M.: IGNOR: Image-guided Neural Object Rendering. In: ICLR (2020) 4
62. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: ICCV (2021) 1, 3
63. Wang, T., Sarafianos, N., Yang, M.H., Tung, T.: Animatable neural radiance fields from monocular rgb-d. arXiv preprint arXiv:2204.01218 (2022) 1
64. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: CVPR (2020) 3, 4, 5, 6, 10, 11, 12
65. Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., Sridhar, S.: Neural fields in visual computing and beyond (2021) 4
66. Xu, H., Alldieck, T., Sminchisescu, C.: H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In: NeurIPS (2021) 1, 4, 11
67. Yoon, J.S., Kim, K., Gallo, O., Park, H.S., Kautz, J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In: CVPR (2020) 1, 3
68. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 11
69. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. TOG (2018) 3
70. Zitnick, C., Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High-quality video view interpolation using a layered representation. TOG (2004) 3, 4

Free-Viewpoint RGB-D Human Performance Capture and Rendering

Phong Nguyen-Ha[1*], Nikolaos Sarafianos[2],
Christoph Lassner[2], Janne Heikkilä[1], and Tony Tung[2]

[1] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[2] Meta Reality Labs Research, Sausalito
https://www.phongnhhn.info/HVS_Net

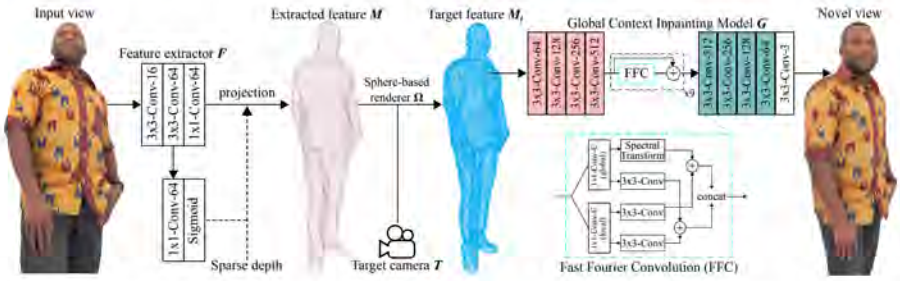

Fig. 1. *Detailed architecture of the sphere-based view synthesis network.* The feature extractor F first use three convolution layers with stride 1 to extract the features of the input view. We then infer the radius of each sphere by passing the learned features through another convolution layer and the sigmoid activation function. The green and red convolution layers of G module scale up and down the feature maps respectively.

In this supplementary material we provide additional details regarding our network designs (Sec. A), as well as implementation details (Sec. B). Additional qualitative evaluations and results are shown in the supplemental video. Finally, we discuss the limitation of our approach (Sec. C).

A Network Designs

In this section, we describe the technical details of two sub-networks of our proposed HVS-Net: a sphere-based view synthesis S and a enhancer model E.

A.1 Sphere-based view synthesis model S

Sphere-based feature warping. The architecture of the sphere-based view synthesis model S is shown in Fig. 1. Instead of directly rendering novel views using the RGB input image, we first passed it through a feature extractor F

*This work was conducted during an internship at Meta Reality Labs Research.

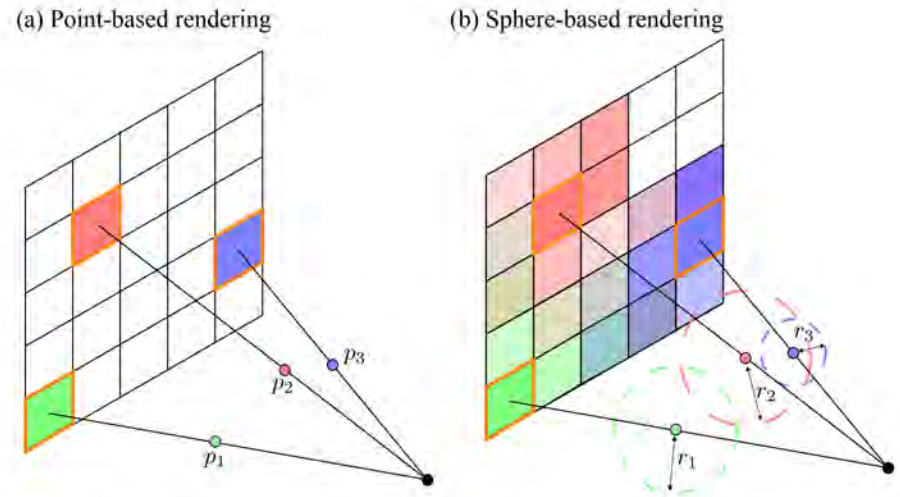

Fig. 2. Visualization of the *rendered features* between (a) point and (b) sphere-based rendering methods. Point-based method [?] can only render pixels (orange boxes) that have valid 3D coordinates. In contrast, sphere-based method [?] uses learned radius r_i of each point p_i to render neighboring pixels which leads to a denser feature map.

which consists of three convolution layers with stride 1 to maintain the spatial resolution. We choose the features f_i as the values of M where there is a valid depth value. We estimate per-sphere radius r_i by passing M to another convolution layer with sigmoid activation function. In Fig. 2, we show the visualization of rendered feature maps from a set of sparse points using point and sphere-based renderers. In case of point-based rendering [?], each 3D point p_i can render a single pixel. Therefore, a large amount of pixels can not be rendered because there is no ray connecting those pixels with valid 3D points. In contrast, the sphere-based neural renderer [?] Ω renders a pixel by blending the colors of any intersected spheres with the given ray. Since we estimate radius r_i of each sphere (dashed circle) using a shallow network, this allows us to render pixels that do not have a valid 3D coordinates. As a result, we obtain a much denser feature maps as can be seen in the Fig. 2 of the main paper. Note that, Ω is fully-differentiable and renders target feature maps very efficiently using PyTorch3D [?].

Global context inpainting model. We render the novel view using a global context inpainting model G. We design the architecture of the G module based on the encoder-decoder structure with skip connections and nine residual blocks are also utilized in the bottleneck.

In each residual block, we replace the regular convolution layers with the recently proposed Fast Fourier Convolution(FFC) [?] which possesses the non-local receptive fields. According to the spectral convolution theorem in Fourier theory, point-wise update in the spectral domain globally affects all features involved in the Fourier transform. The FFC layer splits the input features into

local and global branches. The local branch utilizes conventional convolution layers to obtain local features. In contrast, the global branch includes a Spectral Transform block [**?**] which uses channel-wise Fast Fourier Transform [**?**] to enable image-wide receptive field. The output of both branches are then summed, aggregated before adding to the residuals.

Outputs. The view synthesis model S not only predicts an RGB image I_p of the target view but also a foreground mask I_m and a confidence map I_c. We employ three different 3×3 convolution layers to predict those outputs using the output of the final layer of the G module. Thus, we apply the predicted foreground mask and confidence map to the predicted novel image as follow: $I_p = I_p \times I_m \times I_c$. We train the model S using the photometric loss $\mathcal{L}_{photo}$ as defined in the main paper.

A.2 Enhancer model *E*

Ground-truth Data: We use the RenderPeople dataset [**?**] to train all our models; which comprises of 1000 watertight raw meshes. To obtain IUV ground-truth we first fit an SMPL-like parametric body model to the scans and then perform non-rigid registration for all meshes and rig them for animation. In that way we obtain 1000 rigged models to which we can apply the same IUV map during rendering with an emission shader in Blender Cycles and thus obtain per-pixel perfect IUV ground-truth given an RGB input. This process is depicted in Fig. 3.

HD-IUV predictor D: Now that we have generated pairs of RGB images and ground-truth IUV maps the next step is to train a network that given an RGB image of a human, can establish accurate per-pixel correspondences **for each pixel** corresponding to the clothed human (see Fig.4). Note that the key difference between this approach and what methods such as DensePose [**?**] or CSE [**?**] are doing which is dense correspondence estimates to the unclothed human body. In addition because most approaches are trained on the DensePose-COCO dataset [**?**] which comprises sparse (only ~100 discrete points per human) and noisy annotations such predictions are usually inaccurate and not applicable to our application that targets clothed humans. This is also depicted in Fig. 5 of the main paper where its clear that DensePose IUV estimates result into poor texture warpings.

To train our model which we term as HD-IUV (that stands for High-Definition IUV) we employed an encoder-decoder architecture with four downsampling and upsampling convolution layers along with skip connections between them while the bottleneck comprises 3 residual blocks. This design is justified by the fact that our input-output pairs are always well aligned due to the dense correspondences established by HD-IUV which is not the case with prior work. For HD-IUV, we utilize instance normalization [**?**] and the ReLU activation function in all layers of the network besides the 3 output branches for each task (I, U, V outputs). The UV branches have 256 output channels (since the UV predictions can take any possible value), whereas the I channel has 25 channels which correspond to 24 body parts and background. In all branches a $1 \times$ 1 convolution is applied and

its output is an unnormalized logit that is then fed to the cross-entropy losses. Each task's scores are fed to their respective classification losses which are used to train the network as:

$$L_{IUV} = \lambda_I * L_I + \lambda_U * L_U + \lambda_V * L_V \quad (1)$$

where λ_i, L_i are the respective weighting parameters and loss functions for the I, U, V channels. Framing this problem as a multi-task learning problem (3 tasks) where the U, V and I tasks are $(256D, 256D, 25D)$ per-pixel classification problems respectively, ended up being a very effective approach to enforce strong supervisions for the surface correspondences that other losses we experimented with could not achieve. In addition we employed a silhouette loss to ensure that dense correspondence estimates are provided for each pixel of the foreground clothed human. Finally, using the predicted IUV, we can warp the occlusion-free input image to the target camera using the texture transfer technique[3] from DensePose [**?**].

Refinement module In this section, we utilize the warped image I_w from previous step to enhance the initially estimated target view I_p using a refinement module R. Based on the predicted confidence of the view synthesis network, we combine both images as follows: $\hat{I} = I_p + (1 - I_c) * I_w$ where $\hat{I}$ is fed to a encoder-decoder network for the refinement purposes. In this work, we try to generate humans at the novel viewpoints so rendering realistic human body parts is required. We observe that the predicted semantic I contains valuable information about the semantic information of the human in the target camera. Therefore, we use the SPADE normalization [**?**] to inject the semantics I to the decoder of the refinement module. As can be seen in the qualitative results, the refined image is photo-realistic compared to the ground-truth image. Note that, we use the same discriminator with [**?**] to perform adversarial training between both before and after refined images and the ground-truth novel views.

Discussion Here we discuss the effectiveness of our proposed HD-IUV over DensePose [**?**] representations to refine the target views. As can be seen in the Fig. 8 of the main paper, our Enhancer model can handle heavy occlusions using just a single photo. We emphasize that the HD-IUV representation is crucial for this refinement step because we can obtain pixel-aligned warped images at the target viewpoints compared to the ground-truth data. Therefore our warped images have higher quality compared to those produced by DensePose.

B Implementation Details

The models were trained with the Adam optimizer using a 0.004 learning rate for the discriminator, 0.001 for both the view synthesis model R and the enhancer module E and momentum parameters (0, 0.9). The input/output of our method are 1024×1024. We implement HVS-Net in PyTorch and the training across our large-scale dataset with all identities and views took 2 days to converge on 4 NVIDIA V100 GPUs.

[3] Texture Transfer Using Estimated Dense Coordinates

C 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 systems, 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 [**?**] 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.

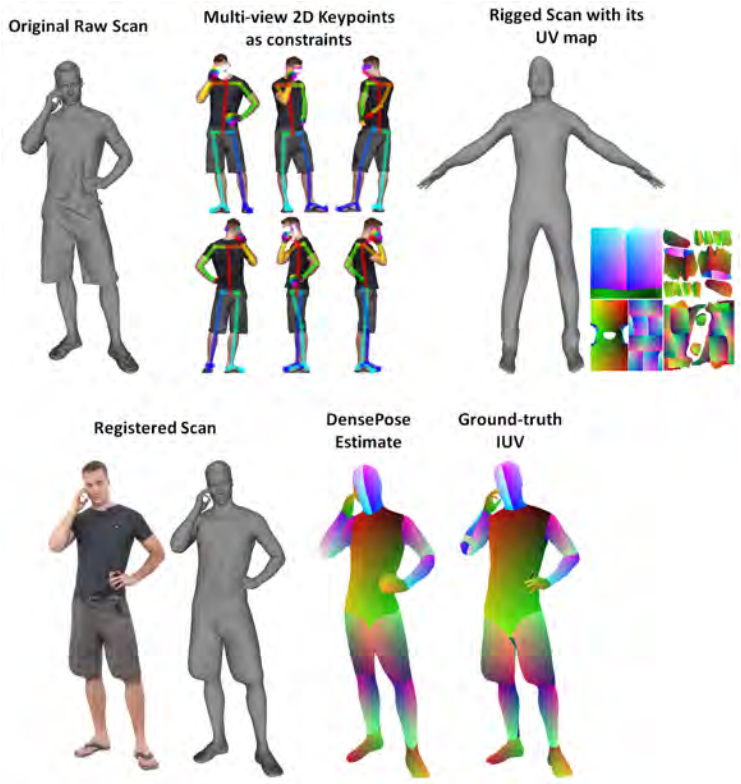

Fig. 3. *Process for IUV ground-truth generation* Given a raw synthetic scan of a clothed human (top left) we perform non-rigid registration with 2D keypoints as additional constraints (top-middle) and obtain the registered scan to the body template (bottom left) and the rigged scan (top right) which is animation ready. Using the corresponding UV map we can now obtain accurate IUV ground-truth (bottom right) that we use to train the proposed HD-IUV model. We provide the corresponding DensePose estimate to demonstrate the stark difference between the two in terms of quality as well as coverage.

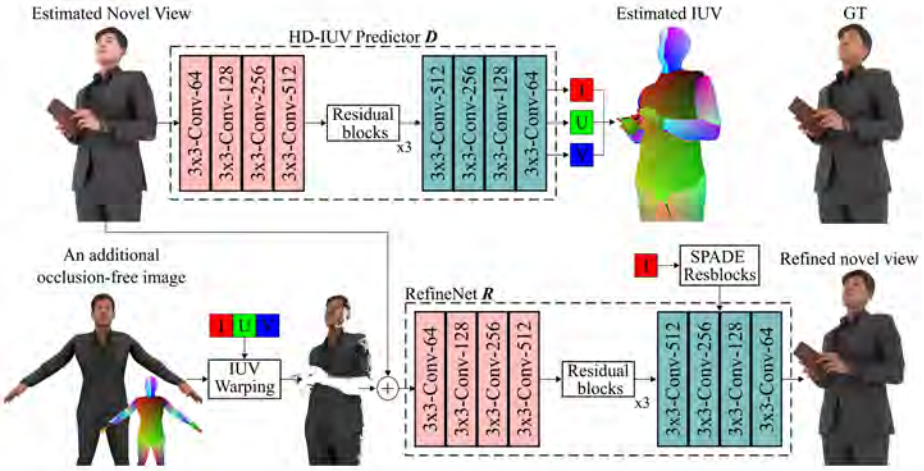

Fig. 4. *IUV-based image refinement.* Using an additional occlusion-free input, we refine the initial estimated novel view by training the Enhancer E network. We infer the dense correspondences of both predicted novel view and occlusion-free image using a novel *HD-IUV* module. The occlusion-free image 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.