Human Pose Manipulation and Novel View Synthesis using Differentiable Rendering

Guillaume Rochette[1], Chris Russell[2], Richard Bowden[1]
[1] University of Surrey, [2] AWS Tübingen

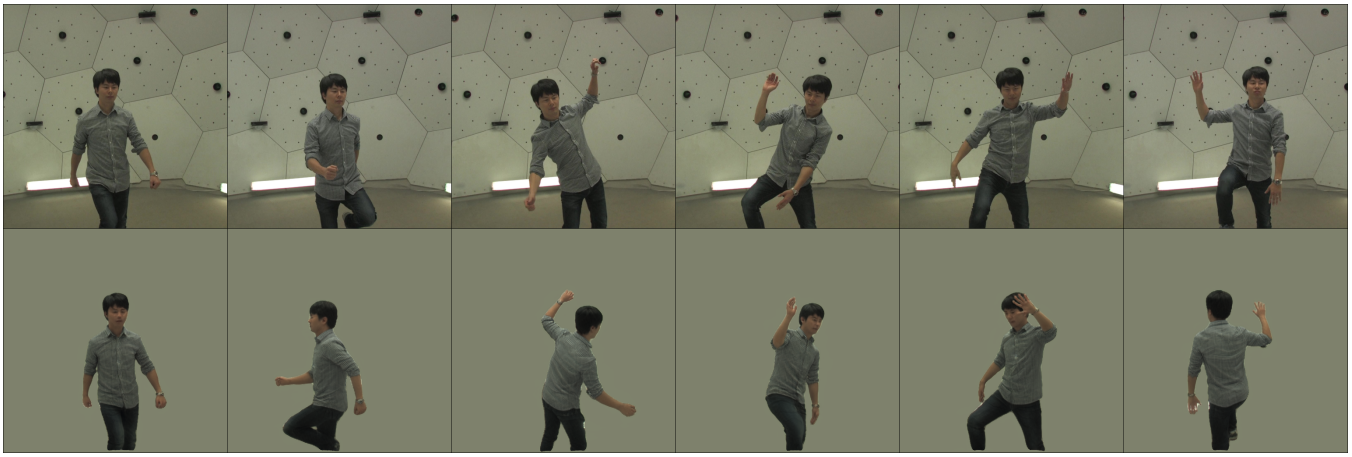

Fig. 1. From a previously unseen input image *(top)*, we estimate the pose and appearance of the subject to render a novel view of the scene *(bottom)*.

***Abstract*— We present a new approach for synthesizing novel views of people in new poses. Our novel differentiable renderer enables the synthesis of highly realistic images from any viewpoint. Rather than operating over mesh-based structures, our renderer makes use of diffuse Gaussian primitives that directly represent the underlying skeletal structure of a human. Rendering these primitives gives results in a high-dimensional latent image, which is then transformed into an RGB image by a decoder network. The formulation gives rise to a fully differentiable framework that can be trained end-to-end.**

We demonstrate the effectiveness of our approach to image reconstruction on both the Human3.6M and Panoptic Studio datasets. We show how our approach can be used for motion transfer between individuals; novel view synthesis of individuals captured from just a single camera; to synthesize individuals from any virtual viewpoint; and to re-render people in novel poses. Code and video results are available at https://github.com/GuillaumeRochette/HumanViewSynthesis.

I. Introduction

We present a new three-step approach for novel view synthesis and motion transfer (see Fig. 1). We first infer the pose and appearance information of a subject from an image. The pose is then transferred to a novel view along with the appearance, from which we render diffuse primitives, using a realistic camera model, onto a high-dimensional image of the foreground of the scene. These diffuse Gaussian primitives are semantically meaningful, simplify optimization and explicitly disentangle pose and appearance. Finally, we use an encoder-decoder architecture to generate novel realistic images. Leveraging multi-view data, we train this framework end-to-end, optimizing both image reconstruction quality and pose estimation. While our approach is generic and can be applied to many tasks, we focus on human pose estimation and novel view synthesis across large view changes.

978-1-6654-3176-7/21/$31.00

Novel view synthesis is a fundamentally ill-posed problem, that we decompose into two parts. The first involves solving an inverse graphics problem, requiring a deep understanding of the scene, while the second relies on image synthesis to generate realistic images using this understanding of the scene. Differentiable rendering is an exciting area of research and offers a unified solution to these two problems. It allows the fusion of expressive graphical models, that capture the underlying physical logic of systems, with learning from gradient-based optimization. Current approaches to differentiable rendering try to directly reason about the world by aligning a well-behaved and smooth approximate model with an underlying non-smooth, and highly nonconvex image. While such approaches guarantee that gradients exist and that a local minimum can be found via continuous optimization, this is a catch-22 situation – the more detailed the model, the better it can be aligned to represent the image; however, the more non-convex the problem becomes, the harder it is to obtain correct alignment.

Disentangling shape and appearance of an image is fundamental for general 3D understanding and lies at the heart of our idea. If we focus on synthesizing novel views of humans, the localization of human joints in the three-dimensional space can be seen as a first step towards human body shape estimation. If we simultaneously estimate the appearance of these body parts, then once we have extracted pose and appearance, we can transfer the information to a novel view and synthesize the output image.

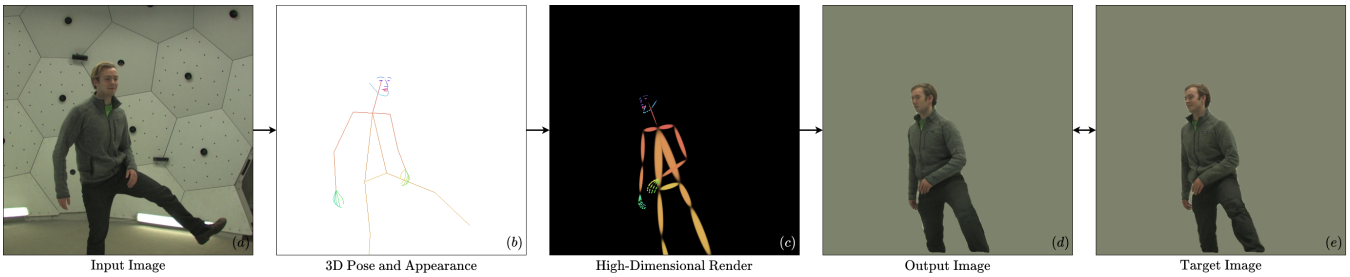

Fig. 2. From an input image I_1^*, we infer the 3D joint locations P_1 and estimate appearance vectors a_1 for each limb (here, represented using an RGB palette). We transfer the pose from the input view to the novel view and render the primitives along with their appearances onto the high-dimensional latent image J_2. From the latent image, we synthesize the output image I_2, which should match the target image I_2^*.

We present a novel rendering primitive that allows us to take full advantage of recent progress in 3D human pose estimation and encoder-decoder networks to gain a higher degree of control over actor rendering, allowing us to synthesize known individuals from arbitrary poses and views. The key insight that allows this to take place is the use of diffuse Gaussian primitives to move from sparse 3D joint locations into a dense 2D feature image via a novel density renderer. An encoder-decoder network then maps from the feature space into RGB images (see Fig. 2). By exploiting the simplicity of the underlying representation, we can generate novel views while 'puppeting' the actors, re-rendering them in novel poses, as shown in Fig. 5.

Section II gives an overview of the literature related to human pose estimation and novel view synthesis. Section III presents our approach for estimating the pose and synthesizing novel views of humans and our novel differentiable rendering. Section IV contains experiments performed on the Panoptic Studio and Human3.6M datasets, evaluating pose estimation accuracy and image reconstruction quality, as well as motion transfer between humans.

II. Related Work

1) Human Pose Estimation: Locating body parts in an image is inherently hard due to variability of human pose, and ambiguities arising from occlusion, motion blur, and lighting. Additionally, estimating depth is a challenging task that is difficult even for humans due to perspective ambiguities.

Approaches to the 2D human pose estimation problem leverage recent advances in deep learning and large amounts of labeled data [19]. Convolutional Pose Machines [30], [45] iteratively refine joint location predictions using the previous inference results. Cao *et al.* [2] improved on this using part affinity fields and enabling multiperson scene parsing.

While 2D human pose estimation is robust to in-the-wild conditions, it is not the case for most 3D approaches, mainly due to the limitations of available data. Large datasets use either Mo-Cap data, such as HumanEva [34] or Human3.6M [10], or rely on a high number of cameras for 3D reconstruction, such as Panoptic Studio [13]. These datasets are captured in tightly controlled environments and lack diversity, limiting the generalization of models. Some approaches use multiobjective learning to overcome the lack of variability in the data, by either solving jointly for the 2D and 3D pose estimation tasks [18], or fusing 2D detection maps with 3D image cues [40]. Recent trends revolve around learning 3D human pose with less constrained sources of data. Inspired by Pose Machines [29], [45], Tome *et al.* [41] trained a network with only 2D poses, by iteratively predicting 2D landmarks, lifting to 3D, and fusing 2D and 3D cues. Martinez *et al.* [24] presented a simple 2D-to-3D residual densely-connected model that outperformed complex baselines relying on image data. Drover *et al.* [5] proposed an adversarial framework, based on [24], randomly reprojecting the predicted 3D pose to 2D with a discriminator judging the realism of the pose. Chen *et al.* [4] refined [5] by adding cycle-consistency constraints.

2) Image Synthesis: Generative adversarial networks [7] are excellent at synthesising high-quality images, but typically offer little control over the generative process. Progressively growing adversarial networks [14] was demonstrated to stabilize training and produce high-resolution photo-realistic faces. Style modulation [15], [16] enabled the synthesis of fine-grained details, as well as a higher variability in the generated faces. They can be conditioned to generate images from segmentation masks or sketches [11], [43]. Ma *et al.* [23] proposed an adversarial framework allowing synthesis of humans in arbitrary poses, by conditioning image generation on an existing image and a 2D pose. Chan *et al.* [3], built on [43], and proposed combining the motion of one human with the appearance of another, while preserving temporal consistency. However, it is limited between two fixed and similar viewpoints, due to the absence of 3D reasoning.

3) Novel View Synthesis: Novel view synthesis is the task of generating an image of a scene from a previously unseen perspective. Most approaches rely on an encoder-decoder architecture, where the encoder solves an image understanding problem via some latent representation, which is subsequently used by the decoder for image synthesis. Tatarchenko *et al.* [39] presented an encoder-decoder using an image and a viewpoint as input to synthesize a novel image with depth information. Park *et al.* [26] refined it by adding a second stage hallucinating missing details. Rather than synthesizing a novel view of images, Zhou *et al.* [48] learnt the displacement of pixels between views. Inspired by Grant *et al.* [8], Sitzmann *et al.* [37] used voxels to represent the 3D

structure of objects and to deal with occlusion. However, these approaches have difficulties dealing with large view changes, due to the unstructured underlying latent representation.

To account for larger view changes, Worrall *et al*. [46] structured their latent space to be directly parameterized by azimuth and elevation parameters. Rhodin *et al*. [31] proposed a framework that disentangles pose, as a point cloud, and appearance, as a vector, for novel view synthesis and is later reused for 3D human pose estimation. By explicitly structuring the latent space to handle geometric transformations, such models are able to handle larger view changes. However, these approaches learn mappings to project 3D structures onto images, which may be undesirable as the mapping is specific to the nature of the rendered object.

4) Differentiable Rendering: Classical rendering pipelines, such as mesh-based renderers, are widely used in graphics. They are good candidates for projecting information onto an image, however, they suffer from several major limitations. Firstly, some operations such as rasterization are discrete and therefore not naturally differentiable. As a substitute, one can use hand-crafted functions to approximate gradients [17], [21]. However, surrogate gradients can add instability to the optimization process. Liu *et al*. [20] proposed a natively differentiable formulation by defining a 'soft-rasterization' step. Another problem with classical approaches is that representing objects as polygonal meshes for rendering purposes creates implicit constraints on the inverse graphics problem, as it requires the movement of vertices to preserve local neighbourhoods and importantly preserve nonlocal constraints of what is the *inside* and *outside* of the mesh. This is challenging to optimize and requires strong regularization. Shysheya *et al*. [33] proposed learning texture maps for each body part to render novel views of humans, using 3D skeletal information as input.

5) Neural Rendering: Embedding a scene implicitly as a Neural Radiance Function (NeRF) is an emerging and promising approach. Mildenhall *et al*. [25] showed that it was possible to optimize an internal volumetric function using a set of input views of the scene. Using sinusoidal activation functions, Sitzmann *et al*. [36] enabled such networks to learn more complex spatial and temporal signals and their derivatives. By combining neural radiance fields and the SMPL articulated body model, Su *et al*. [38] proposed a model capable of generating human representations in unseen poses and views. A range of dynamic NeRF variants have sprung up recently [27], [28], [42], that simultaneously estimate deformation fields alongside the neural radiance function. These approaches are shape agnostic, each model embeds the appearance of a single individual, and they are designed to run on short sequences and render novel views from a camera position similar to those previously captured. In contrast, we explicitly train the model to learn the appearance of multiple individuals (29), allowing rendering from arbitrary views of a single frame captured separately, or for motion transfer from one individual to another.

III. Methodology

We jointly learn pose estimation as part of the view synthesis process, using our Gaussian-based renderer, as shown in Fig. 2. From the input image I_1^*, our model encodes two modalities, the three-dimensional pose P_1 relative to input camera, and the appearance a_1 of the subject. The pose P_1 is transferred to a new viewpoint using camera extrinsic parameters $(R_{1\to2}, t_{1\to2})$. From the pose P_2, seen from a novel orientation, we derive the location μ_2 and shape Σ_2 of the primitives, which are used, along with their appearance a_1 and the intrinsic parameters and distortion coefficients of the second camera (K_2, D_2), for the rendering of the subject in a high-dimensional image F_2. This feature image F_2 is enhanced in an image translation module to form the output image I_2, which closely resembles the target image I_2^*.

A. Extracting Human Pose and Appearance from Images

We model the human skeleton as a graph $G = (P, E)$ with N vertices and M edges, where $P \in \mathbb{R}^{N\times3}$ denotes the joint locations and $E = \{(i, j)|i, j \in [1..N]\}$.

1) Inferring the Pose: We use OpenPose [2], an off-the-shelf detector, to infer the 2D pose from the input image $I_1^* \in \mathbb{R}^{H_I\times W_I\times3}$,

$$I_1^* \to p_1^*, c_1^* \tag{1}$$

Here $p_1^* \in \mathbb{R}^{N\times2}$ refers to the 2D joints locations and $c_1^* \in \mathbb{R}^N$ to their respective confidence values.
We use a simple fully-connected network [24] to infer the 3D pose from the 2D pose,

$$p_1^*, c_1^* \to \bar{P}_1 \tag{2}$$

Where $\bar{P}_1 \in \mathbb{R}^{(N-1)\times3}$ refers to the 3D pose relative to the root joint.
We compute the root joint location $P_{1,\text{root}}$ in camera coordinates by finding the optimal depth which minimize the error between the re-projected 3D pose $\bar{P}_1$ and the 2D pose p_1^* (see Appendix A). We obtain the pose in camera coordinates $P_1 \in \mathbb{R}^{N\times3}$,

$$P_1 = [P_{1,\text{root}}|P_{1,\text{root}} + \bar{P}_1] \tag{3}$$

2) Estimating the Appearance: From the input image I_1^*, we use a ResNet-50 [9] to infer high-dimensional appearance vectors a_1 used to produce the primitives for rendering,

$$I_1^* \to a_1 \tag{4}$$

Where $a_1 \in \mathbb{R}^{M\times A}$ describe the appearances of each of the edges in the human skeleton as seen from the input view. These high-dimensional vectors allow the transfer of the subject's appearance from one view to the other without considering pose configuration of the subject, and disentangling pose and appearance.

3) Changing the Viewpoint: To create an image of the world as seen from another viewpoint, we transfer the pose using the extrinsic parameters $(R_{1\to2}, t_{1\to2})$,

$$P_2 = R_{1\to2} \times P_1 + t_{1\to2} \tag{5}$$

B. Differentiable Rendering of Diffuse Gaussian Primitives

We render a simplified skeletal structure of diffuse primitives, directly obtained from the 3D pose. The intuition underlying our new renderer is straightforward. Each primitive can be understood as an anisotropic Gaussian defined by its location μ and shape Σ, and the rendering operation is the process of integrating along each ray. Occlusions are handled by a smooth aggregation step. Our renderer is differentiable with respect to each input parameter, as the rendering function is itself a composition of differentiable functions.

1) From Pose to Primitives: From the subject's pose, we compute the location and shape of the primitives,

$$P_2 \rightarrow \mu_2, \Sigma_2 \tag{6}$$

Where $\mu_2 \in \mathbb{R}^{M\times 3}$ refers to the locations, while $\Sigma_2 \in \mathbb{R}^{M\times 3\times 3}$ refers to the shapes. For clarity, we drop the subscripts indicating the viewpoint.

For each edge (i, j), we create a primitive at the midpoint between two joints, with an anisotropic shape aligned with the vector between the two joints,

$$\mu_{ij} = \frac{P_i + P_j}{2} \tag{7}$$

$$\Sigma_{ij} = R_{ij} \times \Lambda_{ij} \times R_{ij}^{\intercal} \tag{8}$$

Where,

$$R_{ij} = f_R(P_j - P_i, e) \tag{9}$$

$$\Lambda_{ij} = \text{diag}(||P_j - P_i||_2, w_{ij}, w_{ij}) \tag{10}$$

Here, f_R calculates the rotation between two non-zero vectors (see Appendix B) and w_{ij} loosely represents the width of the limb.

2) Rendering the Primitives: Modelling the scene as a collection of diffuse primitives, we render a high-dimensional latent image J_2,

$$J_2 = \mathcal{R}_{\alpha,\beta}(\mu_2, \Sigma_2, a_1, b, K_2, D_2) \in \mathbb{R}^{H_R\times W_R\times A} \tag{11}$$

Where $\mathcal{R}$ is the rendering function; $\alpha > 0$ is a coefficient scaling the magnitude of the shapes of the primitives; $\beta > 1$ is a background blending coefficient; $b \in \mathbb{R}^A$ describes the appearance of the background; and $K_2 \in \mathbb{R}^{3\times 3}$ and $D_2 \in \mathbb{R}^K$ refers to the intrinsic parameters and distortion coefficients.

To simplify notation, we drop the subscripts indicating the viewpoint, and retain the following letters for subscripts: $i \in [1..H_R]$ refers to the height of the image, $j \in [1..W_R]$ refers to the width of the image, and $k \in [1..M]$ refers to each of the M primitives.

We define the rays r_{ij} as unit vectors originating from the pinhole, distorted by the lens, and passing through every pixel of the image,

$$r_{ij} = \frac{u(K^{-1} \times p_{ij}, D)}{||u(K^{-1} \times p_{ij}, D)||_2} \tag{12}$$

Where u is a fast fixed-point iterative method [1] finding an approximate solution to undistorting the rays, and $p_{ij} = \begin{pmatrix} j & i & 1 \end{pmatrix}^{\intercal}$ is a pixel on the image plane.

Let F_{ijk} be the integral of a single primitive (μ_k, Σ_k) along the ray r_{ij},

$$F_{ijk} = \int_0^{+\infty} e^{-\Delta^2(z\cdot r_{ij}, \mu_k, \alpha\cdot\Sigma_k)} dz \tag{13}$$

See Appendix C.1 for the analytical solution.

For each ray r_{ij}, we define a smooth rasterisation coefficient λ_{ijk} for each primitive (μ_k, Σ_k). In a nutshell, this coefficient smoothly favours one primitive, and discounts the others, based on their proximity to the ray r_{ij}. See Appendix C.2 for details.

The background is treated as the $(M+1)^{\text{th}}$ primitive, with unique properties. Its density F_{ijM+1} and smooth rasterisation coefficient λ_{ijM+1} are detailed in Appendix C.3.

We derive the weights ω_{ijk} quantifying the influence of each primitive (μ_k, Σ_k) (including the background) onto each ray r_{ij}, such that $\forall k \in [1..M+1]$,

$$\omega_{ijk} = \frac{\lambda_{ijk} \cdot F_{ijk}}{\sum_{l=1}^{M+1} \lambda_{ijl} \cdot F_{ijl}} \tag{14}$$

Finally, we render the image by combining the weights with their respective appearance,

$$J_{ij} = \sum_{k=1}^{M+1} \omega_{ijk} \cdot a_k \tag{15}$$

C. Synthesizing the Output Image

The intermediate rendered image is feature-based and not photorealistic due to a small number of primitives. While we could render any image with a sufficient number of primitives, increasing the number not only increases realism, but also the computational cost and the difficulty in optimising the overall problem. Instead, we render a simplified skeletal structure (μ_2, Σ_2) with a high-dimensional appearance a_1.

We synthesize the output image of the subject $I_2 \in \mathbb{R}^{H_O\times W_O\times 3}$ using the primitive image $F_2 \in \mathbb{R}^{H_R\times W_R\times A}$, which is fed to an encoder-decoder network. Following StyleGAN2 [16] style mixing and design principles, we use a U-Net encoder-decoder [32]. Where the output resolution is higher than the rendered resolution, the input is upsampled. To recover high-frequency details, we incorporate the appearance a_1 as styles at each stage,

$$F_2, a_1 \rightarrow I_2 \tag{16}$$

From the input image I_1^*, we only estimate information about the foreground subject. As it is impossible to accurately infer a novel view of a static background captured by a static camera, we infer a constant background around the subject, and use a segmentation mask to discard the background information from the groundtruth image I_2^*.

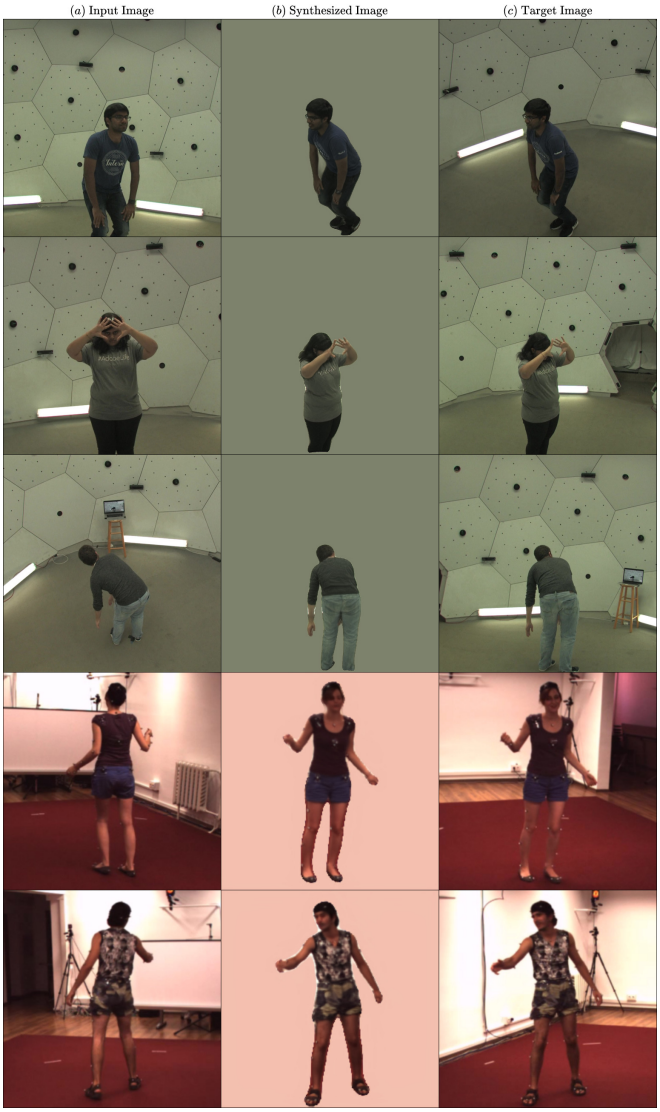

Fig. 3. View synthesis of known subjects in previously unseen poses on Panoptic Studio *(top)* and Human3.6M *(bottom)*. From the input image (a), we extract pose and appearance before rendering the primitives in the novel viewpoint and synthesizing the image (b) which resembles target image (c).

D. Losses

1) Image Reconstruction: We assess the performance of novel view synthesis by measuring the average pixel-to-pixel distance between the image generated by the model I_2 and the target image I_2^* in the pixel space,

$$\mathcal{L}_I = \mathbb{E}_{I_1^*, I_2^*}\left[||I_2^* - I_2||_1^1\right] \tag{17}$$

Which we complement with either the standard perceptual loss [12], $\mathcal{L}_{\phi_{\text{VGG}}}$ with a pretrained VGG network [35], or the fine-tuned LPIPS loss [47], $\mathcal{L}_{\phi_{\text{LPIPS}}}$, to enable the model to synthesize images containing high-frequency detail.

2) Adversarial Framework: To further enhance the realism of the synthesized images, we fine-tune our novel synthesis model in an adversarial framework,

$$\mathcal{L}_A = \mathbb{E}_{I_2^*}\left[\log(D(I_2^*))\right] + \mathbb{E}_{I_1^*}\left[\log(1 - D(I_2))\right] \tag{18}$$

3) Pose Estimation: We require supervision to ensure that the locations of body parts correspond to prespecified keypoints and convey the same semantic meaning,

$$\mathcal{L}_P = \mathbb{E}_{p_1^*, \bar{P}_1^*}\left[c^* \cdot ||\bar{P}_1^* - \bar{P}_1||_2^2\right] \tag{19}$$

Where $c^* \in \mathbb{R}^N$ denotes the confidence values of the points.

4) Appearance Regularization: Appearance vectors are an unsupervised intermediate representation. We regularize the squared norm of the appearance vectors,

$$R_a = ||a_1||_2^2 \tag{20}$$

5) Final Objective: We obtain our final objective,

$$\min_M \max_D \lambda_I \mathcal{L}_I + \lambda_\phi \mathcal{L}_\phi + \lambda_A \mathcal{L}_A + \lambda_P \mathcal{L}_P + \lambda_a R_a \tag{21}$$

With, $\lambda_I = \lambda_\phi = 10$, $\lambda_A = \lambda_P = 1$, and $\lambda_a = 10^{-3}$.

IV. Experiments

Our framework is implemented in PyTorch and trained end-to-end, with the exception of the 2D detector, which is OpenPose [2]. Our models are trained with a batch size of 32, using AdamW [22], with a learning rate of $2 \cdot 10^{-3}$ and a weight decay of 10^{-1}. The appearance vectors are 16 dimensional. For the differentiable renderer, we set $\alpha = 2.5 \cdot 10^{-2}$, $\beta = 2$ and $b = 0_A$. The segmentation masks used for cropping the background out of the groundtruth image are obtained with SOLOv2 [44]. Images are loosely cropped around the subject to a resolution of 1080^2 for Panoptic Studio and cropped given a bounding box of the subject for Human3.6M. Input images are resized to a resolution of 256^2, high-dimensional latent images are rendered at a resolution of 256^2, and output images are resized to either 256^2 or 512^2.

Datasets

Panoptic Studio: Joo *et al.* [13] provides marker-less multi-view sequences captured in a studio. There are over 70 sequences captured from multiple cameras, including 31 HD cameras at 30 Hz. We restrict ourselves to single person sequences, and use only images recorded by high-definition cameras. Panoptic Studio presents greater variability in subject's clothing, morphology and ethnicity, with the most significant detail being that it was captured in a marker-less fashion, compared to datasets such as Human3.6M [10]. Its high variability in camera poses is beneficial both for novel view synthesis and when demonstrating the 3D understanding and robustness of the model.

We infer 2D poses by running OpenPose over every view of a sequence, and reconstruct 3D pose following the approach detailed by Faugeras [6], which computes a closed-form initialization followed by an iterative refinement. We remove consistently unreliable 2D estimates and obtain a 117-point body model.

Data is partitioned at the subject level into training, validation, and test sets, with each subject in only one of the sets. Approximately 80% of the frames are used for training, 10% for validation and 10% for testing. This corresponds to 29 subjects for training, 4 for validation, and 4 for testing.

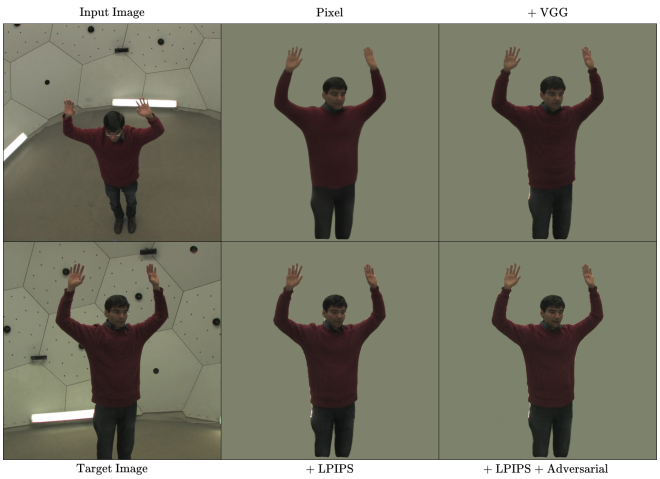

Fig. 4. Effects of the losses on the view synthesis of known subjects in previously unseen poses. There is a clear improvement in reconstruction quality when adding perceptual losses, and the highest-frequency detail appears with the adversarial training.

We require pairs of images for training, therefore we have $|\mathcal{E}| = |\mathcal{F}| \times |\mathcal{V}|_2$ possible pairs, with $\mathcal{F}$ and $\mathcal{V}$, referring to the sets of available frames and views respectively. This gives us a total of 81.6M pairs of images for training, 11.7M pairs for validation and 12.7M for testing.

Human3.6M: Human3.6M [10] is one of the largest publicly available datasets for 3D human pose estimation. It contains sequences of 11 actors performing 15 different scenarios recorded with 4 high-definition cameras at 50 Hz. Mo-Cap is used to provide accurate ground truth. We use the provided 17-joint skeletal model. Following previous work, data is partitioned at the subject level into training (S1, S5, S6, S7 and S8) and validation on two subjects (S9 and S11).

A. Ablation Study

We evaluate the contribution of each loss for the view synthesis task, reporting the LPIPS, PSNR and SSIM. For each sequence, we sample two subsequences representing 10% of the whole sequence, for validation and testing, and use the remaining frames for training. We select models by looking at the validation error, and report the results on the test set. As expected, Table I shows that, the models trained with the LPIPS loss achieve a better LPIPS score, while the models trained with the pixel and VGG losses have lower PSNR and SSIM error. However, when we look closely at the qualitative examples, we can see on Fig. 4 that the models trained with the LPIPS loss contain much more high-frequency detail, and that the adversarial loss enables to reach finer levels of details in images. We notice a sharp gap in performance between the models trained on Panoptic Studio and Human3.6M, which we attribute to the granularity of skeletal model, on Panoptic Studio, the skeletal model provides detailed information about the hands and face, while on Human3.6M, face and hands are represented by a single point. More examples are given in the supplementary material.

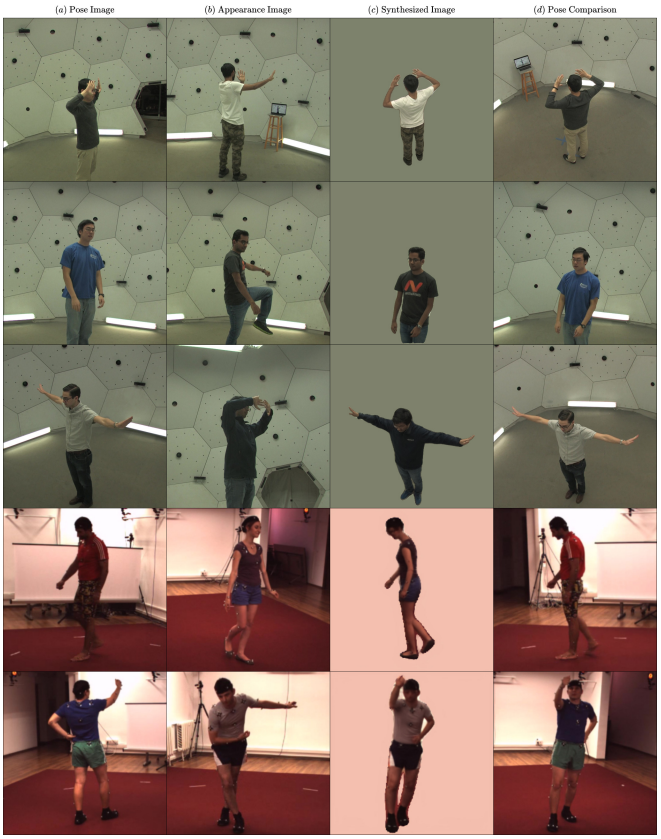

Fig. 5. Motion transfer on Panoptic Studio and Human3.6M. We extract pose from an previously unseen subject (a) and appearance from a known subject (b), and synthesize the combination into a novel view (c), while (d) shows the similarity in terms of pose.

B. Motion Transfer

To demonstrate the versatility of our approach, we apply it to whole body motion transfer. Figure 5 shows how our approach can be used for motion and view transfer from one individual to another. Given an unseen person (a) from the test partition, we estimate their pose from a new viewpoint, and extract the appearance of an individual (b) whose style was learned during training. Since our framework naturally disentangles pose and appearance, without further training, we can combine them and render the image from a novel viewpoint and obtain (c). We provide (d) as a visual comparison to show that our network is able to extract and render 3D information faithfully. Despite small errors in the pose, caused by ambiguities in the pose image (a), we reconstruct a convincing approximation of a different person in the same pose. As such, this work represents an initial step towards full-actor synthesis from arbitrary input videos.

C. Synthesizing Images from Unseen Viewpoints

As our model is trained on multiple views, it can generate realistic looking images of a subject in unseen poses from virtual viewpoints, *e.g.* where cameras do not actually exist. As seen in Fig. 6, we can create virtual cameras travelling on a spherical orbit around the subject. More videos examples are given in the supplementary material.

TABLE I

COMPARISON OF THE RECONSTRUCTION QUALITY DEPENDING ON THE LOSSES.

	Panoptic@256			Panoptic@512			Human3.6M@256		
	LPIPS ↓	PNSR ↑	SSIM ↑	LPIPS ↓	PNSR ↑	SSIM ↑	LPIPS ↓	PNSR ↑	SSIM ↑
$\mathcal{L}_I$	0.0400	**36.78**	**0.9758**	0.0471	**35.67**	**0.9719**	0.0909	**25.42**	**0.9241**
+ $\mathcal{L}_{\phi_{\text{VGG}}}$	0.0239	36.51	0.9749	0.0280	35.35	0.9704	0.0613	25.29	0.9229
+ $\mathcal{L}_{\phi_{\text{LPIPS}}}$	0.0165	35.88	0.9712	0.0193	34.59	0.9667	**0.0481**	24.87	0.9179
+ $\mathcal{L}_{\phi_{\text{LPIPS}}}$ + $\mathcal{L}_A$	**0.0160**	35.75	0.9703	**0.0184**	34.44	0.9660	-	-	-

Fig. 6. Given an input image (a) of a known subject in an unknown pose, we synthesize novel views from non-existing viewpoints on a spherical orbit $(b-f)$ on Panoptic Studio and Human3.6M.

D. Learning Novel View Synthesis of Unknown Subjects from a Monocular Sequence

While our model generalizes well to the pose estimation task, the number of subjects (29 and 5) in both datasets are insufficient for the model to generalize over the space of all human appearances. However, using a monocular sequence of an unseen subject, we can quickly retrain both the appearance and image synthesis modules to the new individual. Our model is able to produce novel views of an unseen individual from a single camera. As we see in Fig. 7 appearance fine-tuning shows a clear visual improvement, however it is more effective on Human3.6M than Panoptic Studio. This is a consequence of subjects facing in one direction only for single sequences in Panoptic Studio. As such, novel view synthesis requires estimating the appearance of completely unseen parts.

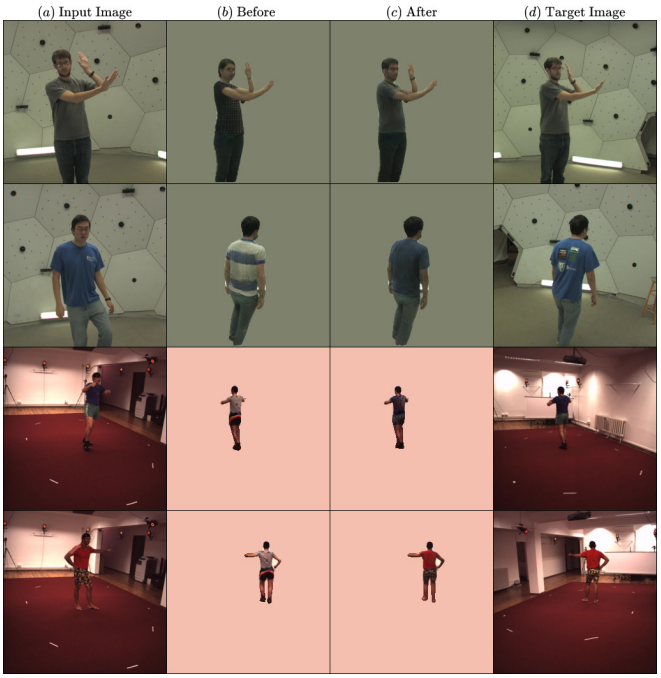

Fig. 7. We fine-tune the appearance and image synthesis module on a monocular sequence of a previously unseen subject. Given an input image (a) appearance from an unknown subject, we synthesize novel views before (b) and after fine-tuning (c).

V. CONCLUSION

We have presented a novel 3D renderer highly suited for human reconstruction and synthesis. By design, our formulation encodes a semantically and physically meaningful latent 3D space of parts while our novel feature rendering approach translates these parts into 2D images giving rise to a robust and easy to optimize representation. A encoder-decoder architecture allows us to transfer style and move from feature images back into the image space. We illustrated the versatility of our approach on multiple tasks: semi-supervised learning; novel view synthesis; and style and motion transfer, allowing us to puppet one person's movement from any viewpoint.

Although we focus on human reconstruction, we believe our approach of rendering a small number of tractable and semantically meaningful primitives as feature images to be useful in a wider scope. Potential applications include semantic mapping, reconstruction and dynamic estimation of the poses of articulated objects and animals, and novel view synthesis of rigid objects. To this end, we will release the complete source code.

VI. ACKNOWLEDGEMENTS

This work received funding from the SNSF Sinergia project 'SMILE II' (CRSII5 193686), the European Union's Horizon2020 research and innovation programme under grant agreement no. 101016982 'EASIER' and the EPSRC project 'ExTOL' (EP/R03298X/1). This work reflects only the authors view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[3] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[4] C.-H. Chen, A. Tyagi, A. Agrawal, D. Drover, S. Stojanov, and J. M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[5] D. Drover, C.-H. Chen, A. Agrawal, A. Tyagi, and C. Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[6] O. Faugeras and O. A. FAUGERAS. *Three-dimensional computer vision: a geometric viewpoint*. MIT press, 1993.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

[8] E. Grant, P. Kohli, and M. van Gerven. Deep disentangled representations for volumetric reconstruction. In *European Conference on Computer Vision*, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[10] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 2016.

[13] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[14] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[15] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[16] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[17] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, 2014.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014.

[20] S. Liu, T. Li, W. Chen, and H. Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[21] M. M. Loper and M. J. Black. Opendr: An approximate differentiable renderer. In *European Conference on Computer Vision*, 2014.

[22] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[23] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *Advances in neural information processing systems*, 2017.

[24] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[25] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *arXiv preprint arXiv:2003.08934*, 2020.

[26] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 2017.

[27] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Deformable neural radiance fields. *arXiv preprint arXiv:2011.12948*, 2020.

[28] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[29] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European conference on computer vision*, 2012.

[30] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *European Conference on Computer Vision*, 2014.

[31] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[33] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, et al. Textured neural avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[34] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 2010.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] V. Sitzmann, J. N. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661*, 2020.

[37] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[38] S.-Y. Su, F. Yu, M. Zollhoefer, and H. Rhodin. A-nerf: Surface-free human 3d pose refinement via neural rendering. *arXiv preprint arXiv:2102.06199*, 2021.

[39] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision*, 2016.

[40] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[41] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[42] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Lassner, and C. Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. *arXiv preprint arXiv:2012.12247*, 2020.

[43] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[44] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen. Solov2: Dynamic and fast instance segmentation. *arXiv preprint arXiv:2003.10152*, 2020.

[45] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[46] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Interpretable transformations with encoder-decoder networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[47] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[48] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European conference on computer vision*, 2016.