Novel View Synthesis of Humans using Differentiable Rendering

Guillaume Rochette[1], Chris Russell[2], and Richard Bowden[1],

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.

Index Terms—Novel View Synthesis, Neural Rendering, Human Pose Estimation, Computer Graphics, Computer Vision, Biometrics.

✦

1 INTRODUCTION

W E present a new three-step approach for novel view synthesis and motion transfer, that preserves the biometric identity of individuals (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.

Scenes in the real world are three-dimensional, yet we observe them as images formed by projecting the world onto a two-dimensional plane. Generating a new synthetic image of a scene from one or many is termed *novel view synthesis*. However, 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 using 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.

Unlike mesh-based renderers, which represent objects as sets of polygons, we treat them as a collection of diffuse three-dimensional shapes. In the case of human reconstruction, these shapes correspond to semantically meaningful joint and limb locations (see Fig. 2). Such a formulation is naturally tractable, and overcomes the aforementioned limitations of mesh-based renderers, as it removes the constraints imposed on the nature of objects, and the discretisation induced by the edges of the mesh. As a result, it enables a disentanglement of the pose and appearance of the object, and is simpler than mesh-based differentiable renderers.

A well-known property of neural decoders [1], [2] is that they fit to images in a coarse-to-fine manner; first recovering the low-frequency components that coarsely approximate the image and then recovering the fine details. A consequence of this design choice is that, as the first and second stages remain smooth and easy to optimize, the overall formulation is naturally self-adapting, remaining coarse, approximately convex and easy to optimize when rendering does not closely follow the image. However, in the latter stages, when it is more detailed and susceptible to the choice of initialization, the primitives are already close to a good location.

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.

- [1] *University of Surrey,* [2] *AWS Tübingen.*
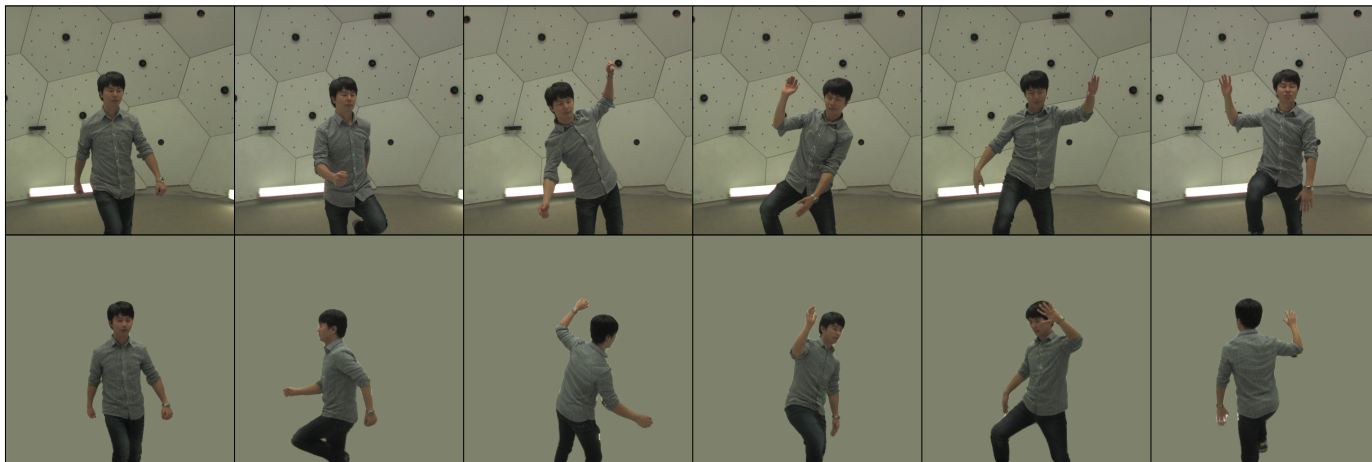- *Code: https://github.com/GuillaumeRochette/HumanViewSynthesis*

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)*

Synthesizing a novel view of the scene is simply the act of projecting the scene from the three-dimensional world into an image. This generative process, also known as rendering in computer graphics, is crucial, as the degree of realism of the generated image relies on it. In addition to the objective of photo-realism, another aim of novel view synthesis is to faithfully reflect information contained in the original view. This requires accurate localization of the human joints, since errors in localization are likely to induce large errors when transferred to other views.

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. 7.

This manuscript revises and extends [3] adding additional details to methodology including aspects of architectural design and further details on the differentiable rendering. Furthermore we include additional explanation on the relationship between the rendered and synthesised images and provide additional experimental validation.

Section 2 gives an overview of the literature related to human pose estimation and novel view synthesis. Section 3 presents our approach for estimating the pose and synthesizing novel views of humans and our novel differentiable renderer. Section 4 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.

2 RELATED WORK
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 have gradually shifted from energy-based model fitting methods [4], [5] using hand-crafted features to a pattern recognition task that leverages recent advances in deep learning and large amounts of labeled data [6]. Convolutional Pose Machines [7], based on the work of Ramakrishna *et al.* [8], iteratively refine joint location predictions using the previous inference results. Cao *et al.* [9] improved on this using part affinity fields and enabling multi-person 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 [10] or Human3.6M [11], or rely on a high number of cameras for 3D reconstruction, such as Panoptic Studio [12]. These datasets are captured in tightly controlled environments and lack diversity, limiting the generalization of models. Inspired by the structure of the human skeleton, Zhou *et al.* [13] proposed kinematic layers to produce physically plausible poses, while Sun *et al.* [14] designed a compositional loss function to backpropagate a supervision signal aware of human structure. Chen and Ramanan [15] proposed an example-based method, inspired by [16], that matched the 2D pose with a set of known 3D poses, rather than estimating pose directly from images. Bogo *et al.* [17] fitted the parametric 3D body model SMPL [18] by minimizing the re-projection error of the model using 2D landmarks inferred by DeepCut [19]. 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 [20], or fusing 2D detection maps with 3D image cues [21]. Recent trends revolve around learning 3D human pose with less constrained sources of data. Inspired by Pose Machines [7], [16], Tome *et al.* [22] trained a network with only 2D poses, by iteratively predicting 2D landmarks, lifting to 3D, and fusing 2D and 3D cues. Martinez *et al.* [23] presented a simple 2D-to-3D residual densely-connected model that outperformed complex baselines relying on image data. Exploiting unla-
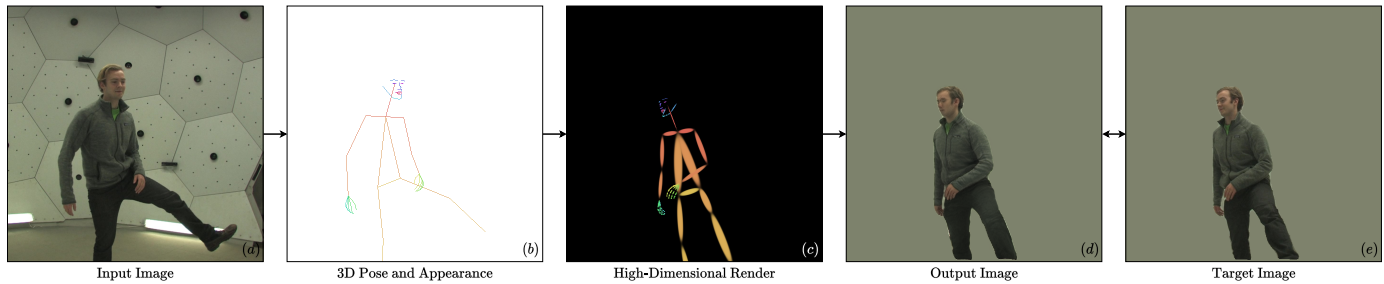
| Input Image | 3D Pose and Appearance | High-Dimensional Render | Output Image | Target Image |

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^*.

beled images from multiple cameras and some annotated images, Rhodin *et al.* [24] used geometric constraints to show the benefits of augmenting the training with unlabeled data. Drover *et al.* [25] proposed an adversarial framework, based on [23], randomly reprojecting the predicted 3D pose to 2D with a discriminator judging the realism of the pose. Chen *et al.* [26] refined [25] by adding cycle-consistency constraints.

Image Synthesis

Generative adversarial networks [27] are excellent at synthesising high-quality images, but typically offer little control over the generative process. Progressively growing adversarial networks [28] was demonstrated to stabilize training and produce high-resolution photo-realistic faces. Style modulation [29], [30] 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 [31], [32]. Ma *et al.* [33] 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.* [34], built on [32], and proposed combining the motion of one human with the appearance of another, while preserving temporal consistency. However, the approach was limited to the generation between two fixed and similar viewpoints, due to the absence of 3D reasoning. Such works open up the possibility of 'deepfaking' somebody's appearance, while another human drives the motion.

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.* [35] presented an encoder-decoder using an image and a viewpoint as input to synthesize a novel image with depth information. Park *et al.* [36] refined it by adding a second stage hallucinating missing details. Rather than synthesizing a novel view of images, Zhou *et al.* [37] learnt the displacement of pixels between views. Inspired by Grant *et al.* [38], Sitzmann *et al.* [39] 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.* [40] structured their latent space to be directly parameterized by azimuth and elevation parameters. Rhodin *et al.* [41] 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.

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 [42], [43]. However, surrogate gradients can add instability to the optimization process. Liu *et al.* [44] 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. A polygonal mesh can be understood as a fixed hyper-graph, where each face corresponds to a hyper-edge that may bind any number of vertices together. Vertices represent points constituting the object in space, whereas edges characterise the relationships between the vertices, and by extension the nature of the object itself. Therefore, although the location of vertices can be differentiated, the existence of connecting edges, which are discrete by nature, cannot. Shysheya *et al.* [45] proposed learning texture maps for each body part to render novel views of humans, using 3D skeletal information as input.

Neural Rendering

Embedding a scene implicitly as a Neural Radiance Function (NeRF) is an emerging and promising approach. Mildenhall
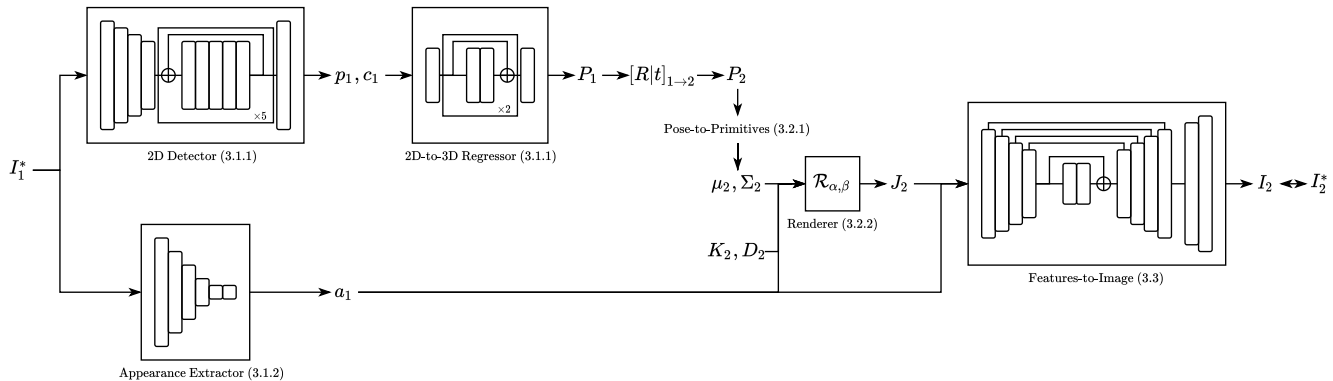
Fig. 3. Architecture of the model. Pose is first localized in the 2D space, then regressed in 3D. Simultaneously appearance vectors are inferred from the input image. Primitives, naturally deducted from the pose, are rendered in the novel viewpoint using the camera intrinsic parameters and distortion coefficients. We synthesize the output image using a high-dimensional feature image alongside with the appearance vectors.

et al. [46] showed that it was possible to optimize an internal volumetric mapping function using a set of input views of the scene. Using sinusoidal activation functions, Sitzmann *et al.* [47] 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.* [48] proposed a model capable of generating human representations in unseen poses and views. A range of dynamic NeRF variants have since sprung up [49], [50], [51], 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. More recently, Weng *et al.* [52] proposed a framework rendering high quality images from novel viewpoints of dancing humans from monocular videos, by learning inverse linear-blend skinning coefficients mapping the subject's estimated pose back to a canonical pose followed by a learnt volumetric function of the colour and density. 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.

3 METHODOLOGY

We jointly learn pose estimation as part of the view synthesis process, using our Gaussian-based renderer, as shown in Fig. 3. 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 J_2. This feature image J_2 is enhanced in an image translation module to form the output image I_2, which closely resembles the target image I_2^*.

3.1 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 $P \in \mathbb{R}^{N \times 3}$ denotes the joint locations and $E = \{(i, j) | i, j \in [1..N]\}$.

3.1.1 Inferring the Pose

We use OpenPose [9], an off-the-shelf detector, to infer the 2D pose from the input image $I_1^* \in \mathbb{R}^{H_I \times W_I \times 3}$,

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

where $p_1 \in \mathbb{R}^{N \times 2}$ 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 [23] to infer the 3D pose from the 2D pose,

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

where $\bar{P}_1 \in \mathbb{R}^{(N-1) \times 3}$ 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 I). We obtain the pose in camera coordinates $P_1 \in \mathbb{R}^{N \times 3}$,

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

3.1.2 Estimating the Appearance

From the input image I_1^*, we use a ResNet-50 [53] 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.
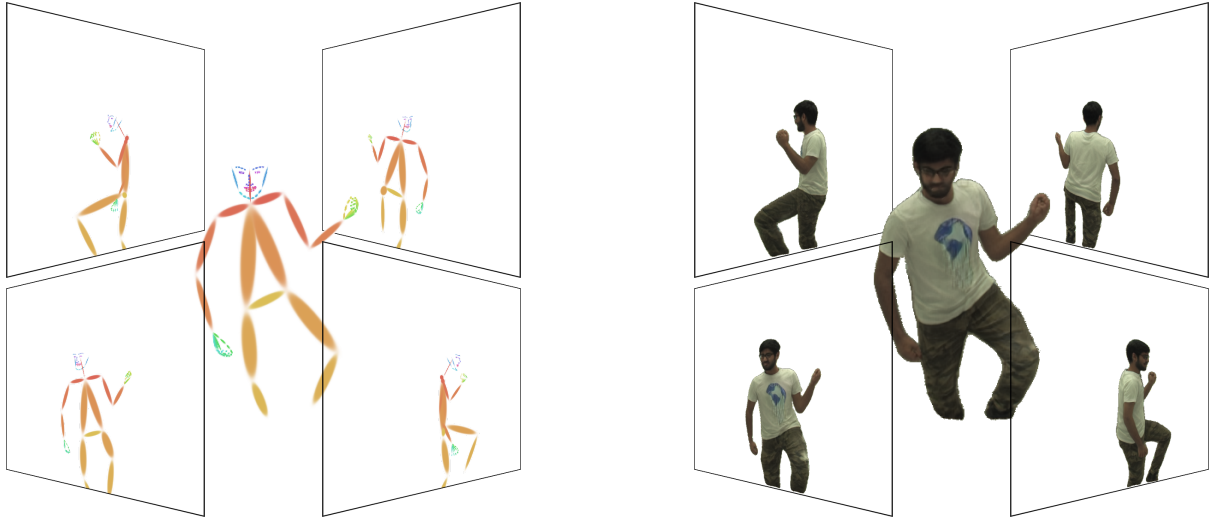
Fig. 4. Visualising the subject from multiple viewpoints. Left: Output of the renderer. Right: Output of the image translation module.

3.1.3 Changing the Viewpoint

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

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

3.2 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.

3.2.1 From Pose to Primitives

From the subject's pose, we compute the location and shape of the primitives,

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

where $\mu_2 \in \mathbb{R}^{M\times3}$ refers to the locations, while $\Sigma_2 \in \mathbb{R}^{M\times3\times3}$ 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,

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

$$\Sigma_{ij} = R_{ij} \times \Lambda_{ij} \times R_{ij}^{\mathsf{T}}, \tag{8}$$

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 II) and w_{ij} loosely represents the width of the limb.

3.2.2 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 $K_2 \in \mathbb{R}^{3\times3}$ 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,

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

where d^{-1} is a fast fixed-point iterative method [54] finding an approximate solution to undistorting the rays, and $p_{ij} = \begin{pmatrix} j & i & 1 \end{pmatrix}^{\mathsf{T}}$ is a pixel on the image plane.

We calculate the integral F_{ijk} of the diffusion of a single primitive (μ_k, Σ_k) along a 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 III-A for the analytical solution.

Objects closer to the camera should occlude those farther from it. We define a smooth rasterisation coefficient λ_{ijk} for each primitive (μ_k, Σ_k), which smoothly favours one primitive, and discounts the others, based on their proximity to the ray r_{ij},

$$\lambda_{ijk} = \frac{1}{1 + (z_{ijk}^*)^4}, \tag{14}$$

where,

$$z^*_{ijk} = \operatorname*{argmax}_z e^{-\Delta^2(z \cdot r_{ij}, \mu_k, \alpha \cdot \Sigma_k)}. \tag{15}$$

The optimal depth z^*_{ijk}, is attained when the squared Mahalanobis distance Δ^2 between the point on the ray $z \cdot r_{ij}$ and the primitive (μ_k, Σ_k) is minimal, which in turns maximises the Gaussian density function. See Appendix III-B for details.

The background is treated as the $(M+1)^{\text{th}}$ primitive, with unique properties: it is colinear with every ray and located after the furthest primitive, its shape is a constant matrix, and its appearance is also a constant. Its density F_{ijM+1} and smooth rasterisation coefficient λ_{ijM+1} are detailed in Appendix III-C.

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 $\forall k \in [1..M+1]$,

$$\omega_{ijk} = \frac{\lambda_{ijk} \cdot F_{ijk}}{\sum_{l=1}^{M+1} \lambda_{ijl} \cdot F_{ijl}}. \tag{16}$$

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

$$J_{ij} = \sum_{k=1}^{M+1} \omega_{ijk} \cdot a_k. \tag{17}$$

3.3 Synthesizing the Output Image

As shown on the left side of Fig. 4, 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 $J_2 \in \mathbb{R}^{H_R \times W_R \times A}$, which is fed to an encoder-decoder network. Following StyleGAN2 [30] style mixing and design principles, we use a U-Net encoder-decoder [55]. 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,

$$J_2, a_1 \rightarrow I_2. \tag{18}$$

From the input image I^*_1, we only estimate information about the foreground subject, as seen on the right side of Fig. 4. 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, as seen on Fig. 5.

3.4 Losses

3.4.1 Image Reconstruction

We assess the performance of novel view synthesis by measuring the average pixel-to-pixel distance between the
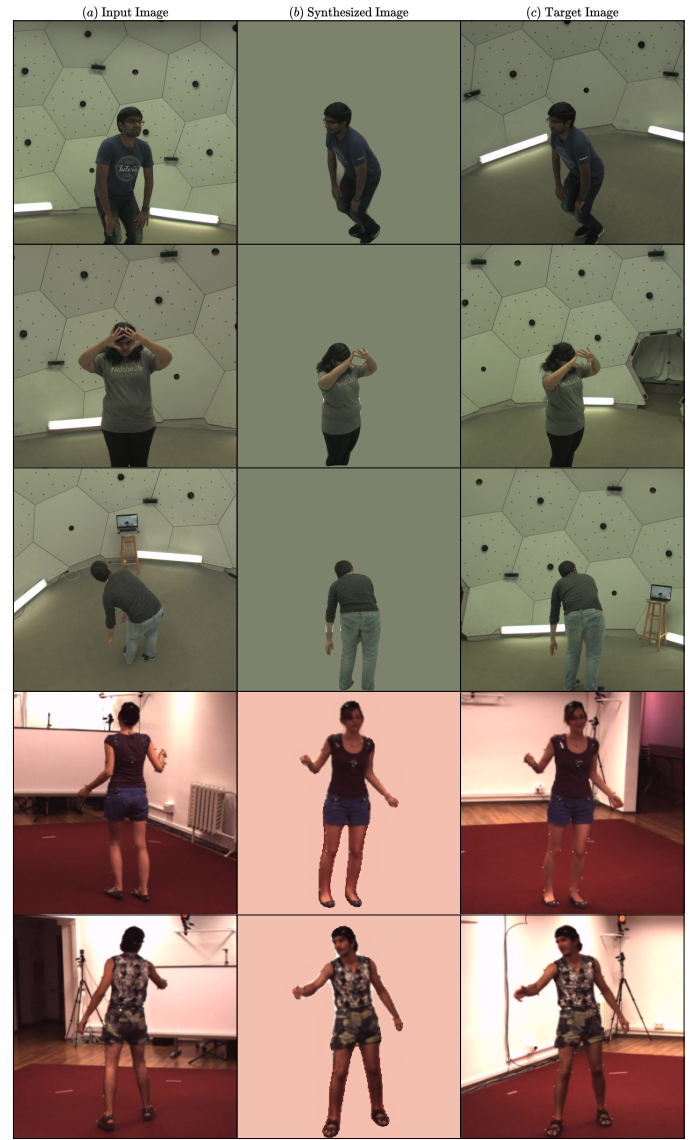

Fig. 5. View synthesis of known subjects in previously unseen poses on Panoptic Studio *(top)* and Human3.6M *(bottom)*. From the input image (a), we extract pose and appearance before rendering the primitives in the novel viewpoint and synthesizing the image (b) which resembles target image (c).

image generated by the model I_2 and the target image I^*_2 in the pixel space,

$$\mathcal{L}_I = \mathbb{E}_{I^*_1, I^*_2} \left[||I^*_2 - I_2||^1_1 \right], \tag{19}$$

which we complement with either the standard perceptual loss [56], $\mathcal{L}_{\phi_{\text{VGG}}}$ with a pretrained VGG network [57], or the fine-tuned LPIPS loss [58], $\mathcal{L}_{\phi_{\text{LPIPS}}}$, to enable the model to synthesize images containing high-frequency detail.

3.4.2 Adversarial Framework

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

$$\mathcal{L}_A = \mathbb{E}_{I^*_2} \left[\log(D(I^*_2)) \right] + \mathbb{E}_{I^*_1} \left[\log(1 - D(I_2)) \right]. \tag{20}$$

3.4.3 Pose Estimation

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

$$\mathcal{L}_P = \mathbb{E}_{p_1, \bar{P}_1^*} \left[c^* \cdot ||\bar{P}_1^* - \bar{P}_1||_2^2 \right], \tag{21}$$

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

3.4.4 Appearance Regularization

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

$$R_a = ||a_1||_2^2. \tag{22}$$

3.4.5 Final Objective

We obtain our final objective,

$$\min_M \max_D \lambda_I \mathcal{L}_I + \lambda_\phi \mathcal{L}_\phi + \lambda_A \mathcal{L}_A + \lambda_P \mathcal{L}_P + \lambda_a R_a, \tag{23}$$

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

4 EXPERIMENTS

Implementation

Our framework is implemented in PyTorch and trained end-to-end, with the exception of the 2D detector, which is OpenPose [9]. We use a slightly modified version of [23] to infer the 3D pose from the 2D landmarks and a ResNet-50 [53] to infer the appearance from the input image. In both networks, Group Normalization [59] replaces Batch Normalization [60]. The renderer presented in Section 3.2 is simply a differentiable function without any learnt parameters. To transfer the rendered image into the output image, we design a U-Net encoder-decoder [55], which follows StyleGAN2 [30] style mixing and design principles. We use a patch discriminator [31] following StyleGAN2 design principles. Our models are trained with a batch size of 32, using AdamW [61], 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 target images are obtained with SOLOv2 [62]. 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*. [12] 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. Some sequences include multiple people engaged in social activities, such as playing music, or playing games, while other sequences focus on single individuals performing diverse tasks. To avoid problems with matching people across
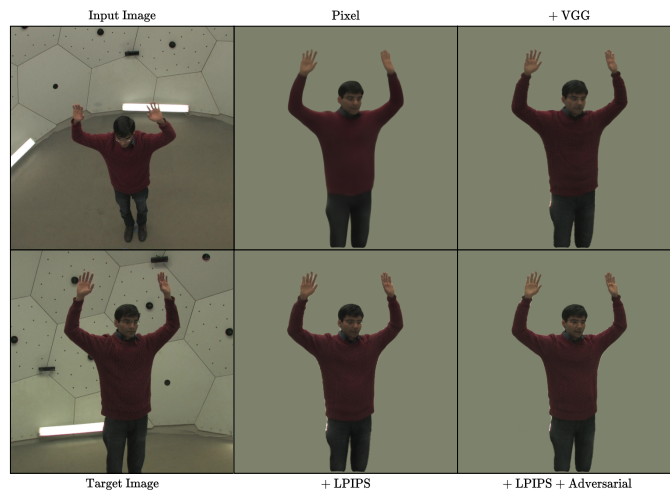
Fig. 6. Effects of the losses on the view synthesis of known subjects in previously unseen poses. There is a clear improvement in reconstruction quality when adding perceptual losses, and the highest-frequency detail appears with the adversarial training.

different viewpoints, we restrict ourselves to single person sequences, and use only images recorded by high-definition cameras. Panoptic Studio presents greater variability in subject 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 [11]. 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 [63], 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. 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 [11] 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). This gives us a total of 6.2M pairs of images for training, 2.2M pairs for validation.

TABLE 1
Comparison of the reconstruction quality depending on the losses.

	Panoptic 256^2			Panoptic 512^2			Human3.6M 256^2		
	LPIPS ↓	PNSR ↑	SSIM ↑	LPIPS ↓	PNSR ↑	SSIM ↑	LPIPS ↓	PNSR ↑	SSIM ↑
$\mathcal{L}_I$	0.0512	**32.34**	**0.9553**	0.0578	**31.67**	**0.9572**	0.1168	**21.47**	**0.8738**
$+ \mathcal{L}_{\phi_{\mathrm{VGG}}}$	0.0363	32.15	0.9540	0.0403	31.49	0.9554	0.0905	21.39	0.8720
$+ \mathcal{L}_{\phi_{\mathrm{LPIPS}}}$	0.0291	31.96	0.9512	0.0323	31.13	0.9521	**0.0792**	21.17	0.8669
$+ \mathcal{L}_{\phi_{\mathrm{LPIPS}}} + \mathcal{L}_A$	**0.0284**	31.91	0.9519	**0.0318**	31.01	0.9509	-	-	-

4.1 Ablation Study

We evaluate the contribution of each loss for the view synthesis task, reporting the PSNR, SSIM [64] and LPIPS [58]. 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 1 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. 6 that the models trained with the LPIPS loss contain more high-frequency detail, and that the adversarial loss enables finer levels of detail to be reached in the 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 the 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.

4.2 Motion Transfer

To demonstrate the versatility of our approach, we apply it to whole body motion transfer. Figure 7 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.

4.3 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. 8, we can create virtual cameras travelling on a spherical orbit around the subject.

Fig. 7. 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.

4.4 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 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. 9 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 in Panoptic Studio. As such, novel view synthesis requires estimating the appearance of completely unseen parts.

Fig. 8. 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.

Fig. 10. Evolution of the synthesis quality and pose estimation against the number of neighbouring input views on Panoptic Studio.

Fig. 9. 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).

Fig. 11. Evolution of the synthesis quality and pose estimation against the number of neighbouring input views on Human3.6M.

4.5 Refining View Synthesis by Fusing the Pose and Appearance from Multiple Input Viewpoints

Our model is able to synthesize high quality images. However, it is still subject to errors, especially when the 3D pose is incorrect. Recall that the 3D pose is recovered in a two-step approach, first the 2D pose is inferred from the input image using an off-the-shelf detector [9], which is then passed to a 2D-to-3D regression model [23] (see Fig 3). This top-down approach is error-prone, and error in the 2D landmarks are likely to contaminate the 3D skeleton. Additionally, the appearance extracted from a single viewpoint will not be
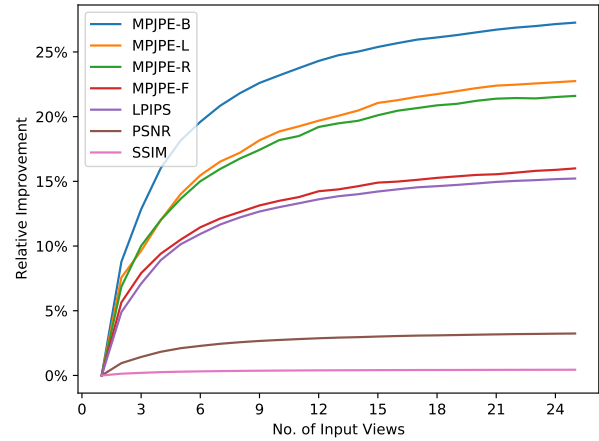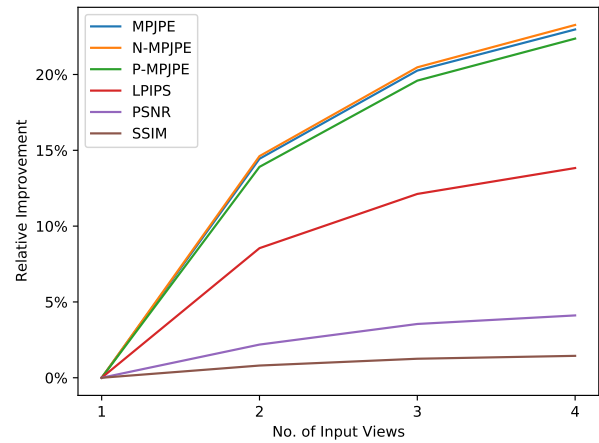
sufficient unless the change in viewpoint between the input and output view is small. Therefore a question naturally arises, does fusing the information from multiple input views help enhance the image synthesis quality? To answer this we propose to infer the pose and appearance from neighbouring viewpoints, rotate the inferred poses back into the original input view, and average the poses and appearances. We re-write Eqs. (1) and (2),

$$I_i^* \rightarrow p_i, c_i \rightarrow \bar{P}_i \quad \forall i = 1..N, \qquad (24)$$

$$\bar{P}_1 = \frac{1}{N} \sum_{i=1}^{N} R_{i \rightarrow 1} \times \bar{P}_i. \qquad (25)$$

As well as modify Eq. (4),

$$I_i^* \rightarrow a_i \quad \forall i = 1..N, \qquad (26)$$

$$a_1 = \frac{1}{N} \sum_{i=1}^{N} a_i. \qquad (27)$$

We evaluate the evolution of both the pose estimation and synthesis quality over the test partition while increasing the number of neighbouring views presented to the model. We report the LPIPS, PSNR and SSIM metrics for image reconstruction quality, for the model trained on Panoptic Studio [12], we report the MPJPE for the four main body regions, body, left and right hands and face, as the error magnitudes between the various region are very different, and for the model trained on Human3.6M [11] we report the MPJPE, N-MPJPE and P-MPJPE.

Figs. 10 and 11 show that as the quality of the 3D pose improves, the image synthesis does as well, which is expected as errors and ambiguities in estimating the 3D pose in the input view will in turn contaminate and worsen the output image synthesis process. Fusing pose information from multiple views yields the largest improvements for the body pose, because it is the least "rigid" region, while hands and face have a much bigger prior. While all the curves are monotonically increasing, we observe diminishing returns to adding neighbouring views. Roughly one third of the maximal improvement possible is already reached by adding the information of a single neighbouring view, and two thirds of the improvement in the pose error, and in turn the synthesis error, is realised with only 5 neighbouring views. Moreover, we notice that the absolute improvement of the reconstruction quality is much less significant with respect to the PSNR and SSIM metrics than for the LPIPS, which we attribute to the shallowness and frailty of these metrics [58], since the PSNR relies on the assumption that the pixels forming the image are independent from each others and the SSIM, while overcoming the pixel-wise independence assumption assumes a Gaussian prior on the pixel distribution while operating in the RGB colour space.

5 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.

Our model allows the synthesis of high quality images of known subjects, and while it can be further fine-tuned on unknown subjects, it is not currently possible to generalise over the space of all human appearances. An important direction for future work would be to investigate the usage of additional biometric cues in order to improve the degree of realism of the reconstructed images, as well as to provide a better structure of the appearance space, which is a first step towards the generalisation of unseen subjects.

While we trained our framework with in-the-lab data, we think it could be applied with little changes to in-the-wild environments, since the off-the-shelf 2D pose estimator is robust to cluttered environments, and while the appearance extractor might struggle for extracting meaningful features, we should be able to overcome this by removing the complex background, using an instance segmentation tool as a pre-processing step to the input image.

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.

ACKNOWLEDGMENTS

REFERENCES

[1] P. Chakrabarty and S. Maji, "The spectral bias of the deep image prior," *arXiv:1912.08905*, 2019. 1

[2] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[3] G. Rochette, C. Russell, and R. Bowden, "Human pose manipulation and novel view synthesis using differentiable rendering," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021. 2

[4] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International journal of computer vision*, 2005. 2

[5] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2

[6] 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. 2

[7] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[8] 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. 2

[9] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4, 7, 9

[10] 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. 2

[11] 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. 2, 7, 10

[12] 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. 2, 7, 10

[13] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *European Conference on Computer Vision*, 2016. 2

[14] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *IEEE International Conference on Computer Vision*, 2017. 2

[15] C.-H. Chen and D. Ramanan, "3d human pose estimation= 2d pose estimation+ matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[16] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3d human pose from 2d image landmarks," in *European Conference on Computer Vision*, 2012. 2

[17] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European Conference on Computer Vision*, 2016. 2

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics*, 2015. 2

[19] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[20] 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. 2

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

[22] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[23] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *IEEE International Conference on Computer Vision*, 2017. 2, 3, 4, 7, 9

[24] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. Fua, "Learning monocular 3d human pose estimation from multi-view images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[25] D. Drover, C.-H. Chen, A. Agrawal, A. Tyagi, and C. Phuoc Huynh, "Can 3d pose be learned from 2d projections alone?" in *European Conference on Computer Vision*, 2018. 3

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

[27] 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. 3

[28] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv:1710.10196*, 2017. 3

[29] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[30] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3, 6, 7

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 7

[32] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[33] 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. 3

[34] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *IEEE International Conference on Computer Vision*, 2019. 3

[35] 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. 3

[36] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg, "Transformation-grounded image generation network for novel 3d view synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[37] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *European Conference on Computer Vision*, 2016. 3

[38] E. Grant, P. Kohli, and M. van Gerven, "Deep disentangled representations for volumetric reconstruction," in *European Conference on Computer Vision*, 2016. 3

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

[40] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Interpretable transformations with encoder-decoder networks," in *IEEE International Conference on Computer Vision*, 2017. 3

[41] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3d human pose estimation," in *European Conference on Computer Vision*, 2018. 3

[42] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[43] M. M. Loper and M. J. Black, "Opendr: An approximate differentiable renderer," in *European Conference on Computer Vision*, 2014. 3

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

[45] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[46] 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:2003.08934*, 2020. 4

[47] V. Sitzmann, J. N. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *arXiv:2006.09661*, 2020. 4

[48] S.-Y. Su, F. Yu, M. Zollhoefer, and H. Rhodin, "A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose," in *Advances in Neural Information Processing Systems*, 2021. 4

[49] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Deformable neural radiance fields," *arXiv:2011.12948*, 2020. 4

[50] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4

[51] 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:2012.12247*, 2020. 4

[52] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman, "Humannerf: Free-viewpoint rendering of moving people from monocular video," *arXiv:2201.04127*, 2022. 4

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 4, 7

[54] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000. 5

[55] 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. 6, 7

[56] 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. 6

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014. 6

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6, 8, 10

[59] Y. Wu and K. He, "Group normalization," in *European Conference on Computer Vision*, 2018. 7

[60] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015. 7

[61] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv:1711.05101*, 2017. 7

[62] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *arXiv:2003.10152*, 2020. 7

[63] O. Faugeras, *Three-dimensional computer vision: a geometric viewpoint*. MIT press, 1993. 7

[64] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004. 8

Guillaume Rochette received the M.Sc. degree in applied mathematics and computer science in 2017 from the National Institute of Applied Sciences, Rouen, France. He is currently pursuing the Ph.D. degree in computer vision and machine learning at the University of Surrey, Guildford, United Kingdom, with the Cognitive Vision Group within the Centre for Vision, Speech and Signal Processing. His research interests include deep learning applied to human-centered computer vision, human pose estimation and novel view synthesis.

Chris Russell is a senior applied scientist at Amazon Web Services where he works on the problems of ethical machine learning and 3D computer vision. He continues to work closely with the Governance of Emerging Technology programme he co-founded at the Oxford Internet Institute. He is a member of the BMVA and an ELLIS fellow.

Richard Bowden is Professor of computer vision and machine learning at the University of Surrey where he leads the Cognitive Vision Group within the Centre for Vision, Speech and Signal Processing. His research centres on the use of computer vision to locate, track, and understand humans. He has been Associate Editor for Image and Vision Computing since 2009 and from 2013-2019 IEEE Pattern Analysis and Machine Intelligence. He was a member of the British Machine Vision Association (BMVA) executive committee and a company director for seven years. He is a member of the BMVA, a fellow of the Higher Education Academy, a senior member of the IEEE and a Fellow of the International Association of Pattern Recognition.