Neural Re-Rendering of Humans from a Single Image

Kripasindhu Sarkar[1]   Dushyant Mehta[1]   Weipeng Xu[2]
Vladislav Golyanik[1]   Christian Theobalt[1]

[1]MPI for Informatics, SIC   [2]Facebook Reality Labs

Abstract. Human re-rendering from a single image is a starkly under-constrained problem, and state-of-the-art algorithms often exhibit undesired artefacts, such as over-smoothing, unrealistic distortions of the body parts and garments, or implausible changes of the texture. To address these challenges, we propose a new method for neural re-rendering of a human under a novel user-defined pose and viewpoint, given one input image. Our algorithm represents body pose and shape as a parametric mesh which can be reconstructed from a single image and easily reposed. Instead of a colour-based UV texture map, our approach further employs a learned high-dimensional UV feature map to encode appearance. This rich implicit representation captures detailed appearance variation across poses, viewpoints, person identities and clothing styles better than learned colour texture maps. The body model with the rendered feature maps is fed through a neural image-translation network that creates the final rendered colour image. The above components are combined in an end-to-end-trained neural network architecture that takes as input a source person image, and images of the parametric body model in the source pose and desired target pose. Experimental evaluation demonstrates that our approach produces higher quality single-image re-rendering results than existing methods.

Keywords: Neural Rendering, Pose Transfer, Novel View Synthesis

1 Introduction

Algorithms to realistically render dressed humans under controllable poses and viewpoints are essential for character animation, 3D video, or virtual and augmented reality, to name a few. Over the past decades, computer graphics and vision have developed impressive methods for high-fidelity artist-driven and reconstruction-based human modelling, high-quality animation, and photo-realistic rendering. However, these often require sophisticated multi-camera setups, and deep expertise in animation and rendering, and are thus costly, time-consuming and difficult to use. Recent advances in monocular human reconstruction and neural network-based image synthesis open up a radically different approach to the problem, neural re-rendering of humans from a single image.

Project webpage: `gvv.mpi-inf.mpg.de/projects/NHRR/`

Fig. 1: Given an image of a person, our neural re-rendering approach allows synthesis of images of the person in different poses, or with different clothing obtained from another reference image.

Given a single reference image of a person, the goal is to synthesise a photo-real image of this person in, for instance, a user-controlled new pose, modified body proportions, the same or different garments, or a combination of these.

There has been tremendous progress in monocular human capture and re-rendering [33, 23, 45, 2, 21, 12, 19, 6, 25] towards this goal. However, owing to the starkly underconstrained nature of the problem, true photo-realism under all possible conditions has not yet been achieved. Methods frequently exhibit unwanted over-smoothing and a lack of details in the rendered image, unrealistic distortions of body parts and garments, or implausible texture alterations.

We, therefore, propose a new algorithm for monocular neural re-rendering of a dressed human under a novel user-defined pose and viewpoint, which has starkly improved visual quality, see Figures 1, 3, 6, 7. We take inspiration from recent work on neural rendering of general scenes with a continuous [48] or a multi-dimensional feature representation with implicit [50] or explicit [47] occlusion handling that are learned from multi-view images or videos.

Our algorithm represents body pose and shape with the SMPL parametric human surface model [29], which can be easily reposed. Instead of modelling appearance as explicit colour maps, *e.g.,* learned colour-based UV texture maps on the body surface [33, 12], we employ a learned high-dimensional UV feature map to encode appearance. This rich implicit representation learns the detailed appearance variation across poses, viewpoints, person identities and clothing styles. Given a single image of a person, we predict pixel correspondences to the SMPL [29] mesh using DensePose [37]. We then extract partial UV texture maps based on the observed body regions and use a neural network to convert it to a complete UV feature map, with a d-dimensional feature per texel. The UV feature map is then rendered in the desired target pose and passed through a neural image translation network that creates the final rendered image. These components are combined in an end-to-end trained neural architecture. In quantitative experiments and a user study to judge the qualitative results, we show that the visual quality of our results improves over the current state of the art.

To summarise, our **contributions** are as follows:

– A new end-to-end trainable method that combines monocular parametric
  3D body modelling, a learned detail-preserving neural-feature based body
  appearance representation, and a neural network based image-synthesis net-
  work to enable highly realistic human re-rendering from a single image;
– state-of-the-art results on the DeepFashion dataset [27] which are confirmed
  with quantitative metrics, and qualitatively with a user study.

2   Related Work

While our proposed approach relates to many sub-fields of visual computing,
for brevity we only elaborate on the immediately relevant work on human body
re-enactment and neural rendering methods for object and scene rendering.

2.1   Classical Methods for Novel View Synthesis

Earlier methods for image-based 3D reconstruction and novel view synthesis rely
on traditional concepts of multi-view geometry, explicit 3D shape and appear-
ance reconstruction, and classical computer graphics or image-based rendering.
Methods based on light fields use ray space representations or coarse multi-view
geometry models for novel view synthesis [22, 11, 4]. To achieve high quality,
dense camera arrays are required, which is impractical. Other algorithms cap-
ture and operate on dense depth maps [60], layered depth images [41], 3D point
clouds [1, 26, 40], meshes [32, 52], or surfels [36, 5, 55] for dynamic scenes. Multi-
view stereo can be combined with fusion algorithms operating with implicit
geometry and achieving more temporally consistent reconstructions over short
time windows [9, 34, 13]. Dynamic scene capture and novel view synthesis were
also shown with a low number of RGB or RGB-D cameras [57, 49, 15, 58]. While
reconstruction is fast and feasible with fewer cameras, the coarse approximate
geometry often leads to rendering artefacts.

2.2   Neural Rendering of Scenes and Objects

Recently, neural rendering approaches have shown promising results for scenes
and objects. Image-based rendering (IBR) methods reconstruct scene geometry
with classical techniques and use it to render novel views [8, 7]. Lack of obser-
vations can cause high uncertainty in novel views. On the other hand, neural
rendering approaches [48, 47, 51, 66] can generate higher-quality results by lever-
aging collections of training data. Many applications of neural rendering have
been recently shown, ranging from synthesising view-dependent effects [66, 51]
to learning the shape and appearance priors from sparse data [39, 56].
    Only a few works on neural scene representation and rendering can handle
dynamic scenes [19, 28]. Some methods combine explicit dynamic scene recon-
struction and traditional graphics rendering with neural re-rendering [31, 19, 18,
51]. Thies *et al.* [50] combine *neural textures* with the classical graphics pipeline
for novel view synthesis of static objects and monocular video re-rendering. Their

technique requires a scene-specific geometric proxy which has to be reconstructed before the training. Instead of more complex joint reasoning of the geometry and appearance needed from the intermediate representation by neural rendering approaches such as that of Sitzmann *et al.* [48], for our human-specific application scenario the coarse geometry is handled by the posable SMPL mesh, with a feature map similar to the Thies *et al.* [50] capturing clothing appearance, which includes fine-scaled geometry, and clothing textures.

Several approaches address related problems such as generating images of humans in new poses [62, 3, 30, 33, 35], or body re-enactment from monocular videos [6], which are discussed next.

2.3   Human Re-enactment and Novel View Rendering

Recent work on photo-realistic human body re-enactment and novel view rendering can be sub-classified along various dimensions.

Object-agnostic approaches [45, 44] model deformable objects directly in the image space. Siarohin *et al.* [45] learn keypoints in a self-supervised manner and capture deformations in the vicinity of the keypoints using affine transforms. Features extracted from the source image are deformed to the target using the predicted transformations and passed on to a generator. Additional predictions of dis-occluded regions indicate to the generator the regions which have to be rendered based on the context. Zhu et al. [65] leverage geometric constraints and optical flow for synthesising novel views of humans from a single image.

Object-specific techniques have the same core components as above, *i.e.*, colour or feature transformation from source to target, occlusion reasoning or inpainting, and photo-realistic image generation from the warped feature or colour image. The key difference is that the feature transformation, occlusion reasoning, and inpainting are guided by an underlying object model, which, in our case, is a parametric human body mesh. Kim *et al.* [19] achieve full control over the head pose and facial expressions in photo-realistic renderings of a target actor by an adversarial training with a performance of the target actor. DensePose Transfer [33] uses direct texture transfer from the input image to the SMPL model, inpaints the occluded regions of the texture and renders it in a new pose. This image is blended with the image resulting from direct conditional generation from the input image, input Densepose, and target Densepose. Zablotskaia *et al.* [59] generate subsequent video frames of human motion and use a direct warping guided by the reference frame, previously generated frame, and DensePose representations of the past and future frames. Their method does not rely on an explicit UV texture map. ClothFlow [14] implicitly captures the geometric transformation between the source and target image by estimating dense flow. Chan*et al.* [6] learn a subject-specific puppeteering system using a video of the subject such that all parts of the subject's body are seen in advance. The GAN-based rendering is driven by 2D pose extracted from the target subject. Zhou *et al.* [63] also learn a personalised model using piecewise affine transforms of the part-segmented source image for modelling pose changes, generating the person image in front of a clean background plate, with a second stage fusing a

given background image with the generated person's image. In contrast to Liu *et al.* [25], we transfer appearance from source to target image using a UV feature map. Instead of directly predicting missing regions of the UV texture map, coordinate-based inpainting [12] predicts correspondence between all regions on the UV texture map and the input image pixels. This results in more texture details in body regions that become dis-occluded when re-posing. As shown in Sec. 4, our UV feature map based approach yields results of much higher quality in comparisons. Shysheya *et al.* [42] explicitly model the body texture and implicitly handle the shape. In contrast, while we explicitly handle the coarse shape, we use a UV feature map to model the fine-scaled shape and clothing texture implicitly. Lazowa *et al.* [21] propose an approach for reconstruction of textured 3D human models from a single image. Similar to our approach, it extracts a partial UV texture map using DensePose but inpaints the UV texture map using a GAN based supervision directly applied to the texture map. Additionally — and similar to Alldieck *et al.* [2] — it predicts a displacement map on top of the SMPL mesh to capture clothing details not present in the SMPL model. Our approach does not explicitly model clothing details.

In contrast to existing methods, we propose a new end-to-end trainable method that combines monocular parametric 3D body modeling [33, 12], a learned neural detail-preserving surface feature representation [50], and a neural image-synthesis network for highly realistic human re-rendering from a single image.

3   Method

Given an image I_s of a person, we synthesise a new image of the person in a different target body pose. Our approach comprises of four distinct steps. The first step uses DensePose [37] to predict dense correspondences between the input image I_s and the SMPL model. This allows a *UV texture map T_s* to be extracted for the visible regions. The second step uses a U-Net [38] based network, which we term *FeatureNet*, to construct the full *UV feature map F_s* from the partial RGB UV texture map T_s. F_s contains a d-dimensional feature representation for all texels, both visible and occluded in the source image. The third step takes a target pose as input, and 'renders' the UV feature map F_s to produce a d dimensional *Feature image $R_{s \to t}$*. The fourth step uses a generator network based on Pix2PixHD [54], which we term *RenderNet*, to generate a photorealistic image $I_{s \to t}$ of the reposed person, from the input Feature image. The overview of our pipeline is shown in Fig. 2.

3.1   Extracting a Partial UV Texture Map from the Input Image

The pixels of the input image are transformed into UV space through matches predicted with DensePose. We use the ResNet-101 based variant of DensePose for predicting the correspondences for the body regions visible in the image. The network is pre-trained on COCO-DensePose dataset and provides 24 body segments and their part-specific U, V coordinates of SMPL model. For easier mapping, the
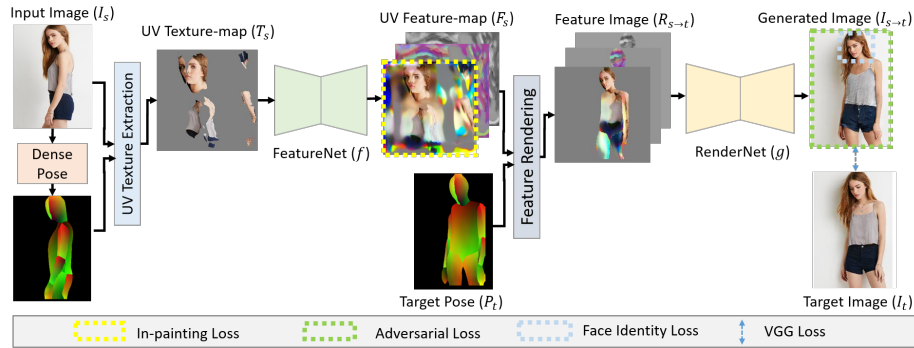
Fig. 2: **Pipeline Overview**: Given a source image I_s, we extract the UV texture map T_s of an underlying parametric body mesh model for the body regions visible in the image. FeatureNet converts the partial UV texture map to a full UV feature map, which encodes a richer 16D representation at each texel. Given a new pose P_t, the parametric body mesh can be re-posed and textured with the UV feature map to produce an intermediate Feature Image $R_{s\to t}$. RenderNet converts the intermediate 16-channel feature image to a realistic image.

24 part-specific UV maps are combined to form a single UV Texture map T_s in the format provided in SURREAL dataset [53] through a pre-computed lookup table. Note that one could putatively use monocular 3D pose estimation methods (*e.g.,* [17]) to compute SMPL parameters, and subsequently, the DensePose of the input image. However, frequent misalignments of the predictions with the end-effector positions in the image lead to significant artefacts in the UV texture map for hands and feet in that case and thus such an approach is not advised [2].

3.2   Generating the Full UV Feature Map

The partial (on account of occlusion) texture map T_s is converted to a full UV feature map F_s using a U-Net-like convolutional network f, which we term *FeatureNet*. That is,

$$F_s = f(T_s).$$

RenderNet comprises of four down-sampling blocks followed by four up-sampling blocks. Therefore, a partial input texture of the spatial dimension of 256×256 is transformed into a spatial dimension of 16×16 in the middle-most layer. Each downsampling block consists of two convolutions followed by maxpool operation. For up-sampling blocks, we use bilinear upsampling followed by two convolutions. The final convolutional layer produces a d-dimensional (channel) *UV feature map* which is used subsequently for rendering a feature image. The first three channels of the UV feature map can be supervised to in-paint the input partial UV texture map T_s, thus having a small subset of the feature channels resemble the classical colour texture map. Our experiments use 16 feature channels.

3.3   Intermediate Feature Image Rendering

The SMPL mesh can be reposed using a target pose P_t, which can be extracted from a target image I_t, or obtained from a different source. In our case, when given a target image I_t, we directly obtain the DensePose output, which is equivalent to the reposed SMPL model. Given the source feature map F_s, we render the SMPL mesh through the DensePose output P_t to produce a d-dimensional *Feature Image $R_{s \to t}$*. That is,

$$R_{s \to t} = r(F_s, P_t).$$

Note that this feature rendering operation r can be conveniently implemented by differentiable sampling. In our experiments, we use bilinear sampling for this operation. The feature image $R_{s \to t}$, which captures the target pose and the source appearance is then used as input to the subsequent translation network.

3.4   Creating a Photo-Realistic Rendering

In the final step, the feature image $R_{s \to t}$ is translated to a realistic image $I_{s \to t}$ using a translation network g similar to Pix2Pix, which we term *RenderNet*:

$$I_{s \to t} = g(R_{s \to t}).$$

RenderNet comprises of (a) 3 down-sampling blocks, (b) 6 residual blocks, (c) 3 up-sampling blocks and finally (d) a convolution layer with Tanh activation that gives the final output. The discriminator for adversarial training of RenderNet also uses the multiscale design of Pix2PixHD [54]. In our experiments, we use a three scale discriminator network for adversarial training.

3.5   Training Details and Loss Functions

During training, our system takes pairs of images (I_s, I_t) of the same person (but in different poses) as input. Partial texture T_s extracted from the source image I_s is passed through the above-mentioned operations to produce the generated output $I_{s \to t}$. That is,

$$I_{s \to t} = g \circ r \circ f(T_s, P_t).$$

Note that all operations g, r and f are differentiable. We train the entire system *end-to-end* and optimise the parameters of FeatureNet ($g(\cdot)$) and RenderNet ($f(\cdot)$). We use the combination of the following loss functions in our system:

 – **Perceptual Loss**: We use a perceptual loss based on the VGG Network [16] — the difference between the activations on different layers of the pre-trained VGG network [46] applied on the generated image $I_{s \to t}$ and ground truth image target image I_t.

$$L_p = \sum \frac{1}{N^j} |p^j(I_{s \to t}) - p^j(I_t)|.$$

Here, p^j is the activation and N^j the number of elements of the j-th layer in the ImageNet pre-trained VGG network.

- **Adversarial Loss**: We use a multiscale discriminator D of Pix2PixHD [54] for enforcing adversarial loss L_{adv} in our system. The multiscale discriminator D is conditioned on both the generated image and rendered feature image.
- **Face Identity Loss**: We use a pre-trained network to ensure that the extracted UV feature map and RenderNet preserve the *face identity* on the cropped face of the generated and the ground truth image.

$$L_{face} = |N_{face}(I_{s \to t}) - N_{face}(I_t)|.$$

Here, N_{face} is the pre-trained SphereFaceNet [24]
- **Intermediate in-painting loss**: To mimic classical colour texture map, we enforce a loss L_{tex} on the first three channels of the output of the in-painting network. This loss is set to the sum of 1) l_1 distance of the visible part of the partial source texture and generated texture and 2) l_1 distance of the visible part of the partial target texture and generated texture.

The final loss on the generator is then

$$L_G = \lambda_{VGG}L_p + \lambda_{face}L_{face} + \lambda_{tex}L_{tex} + \lambda_{GAN}L_{adv}.$$

The conditional discriminator D is updated every step enforcing binary cross-entropy loss on real and fake images. We train the networks end-to-end using Adam optimiser [20] with an initial learning rate of 2×10^4, β_1 as 0.5 and no weight decay. The loss weights are set empirically to $\lambda_{GAN} = 1, \lambda_{VGG} = 10, \lambda_{face} = 5, \lambda_{tex} = 1$. For speed, we pre-compute DensePose on the images and directly read them as input.

During testing, the system takes as input a single image of a person and a target Densepose. The target pose can be extracted by DensePose RCNN on the image of the source person in a different pose (used in the experiments on DeepFashion dataset), or alternatively it can be obtained by reposing the SMPL mesh of the source body. In many cases, the actor can be a completely different person (see Fig. 5 and 7). The neural texture is then rendered using the given target Densepose which is followed by the translation network to generate a realistic image of the source person in the target pose.

4   Experimental Results

4.1   Experimental Setup

Datasets We use the In-shop Clothes Retrieval Benchmark of DeepFashion dataset [27] for our main experiments. The dataset comprises of 52,712 images of fashion models with 13,029 different clothing items in different poses. For training and testing, we consider the split provided by Siarohin *et al.* [43], which is also used by other related works [33, 12]. We also show qualitative results of our method with Fashion dataset [59]. Fashion dataset has 500 training and 100 test videos, each containing roughly 350 frames. The videos are single person sequences, containing different people catwalking in different clothes.
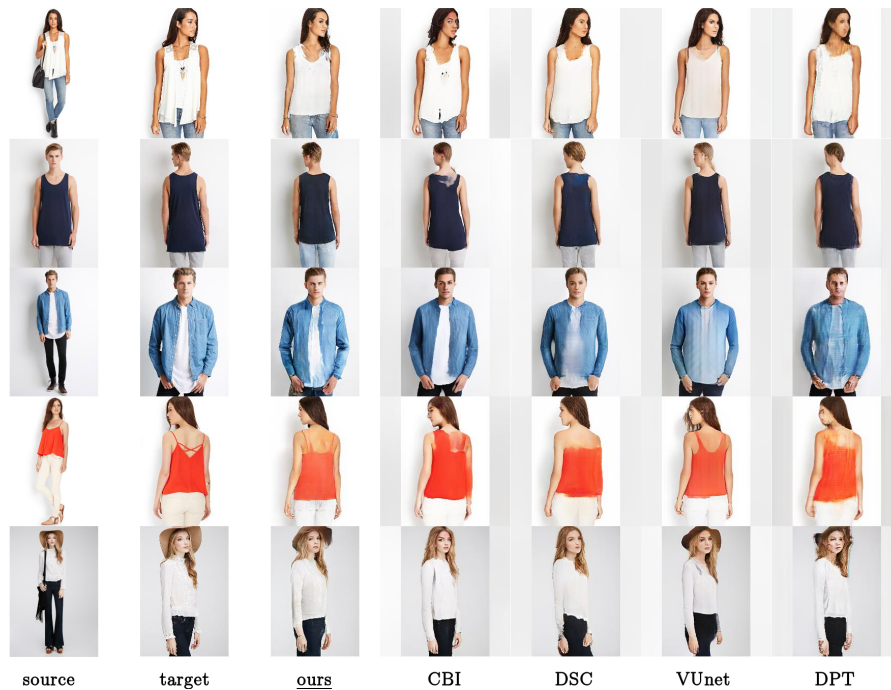
Fig. 3: Results of our method, CBI [12], DSC [43], VUnet [10] and DPT [33]. Our approach produces higher quality renderings than the competing methods.

4.2 Comparison with SOTA

We compare our results with four state-of-the-art methods, *i.e.,* Coordinate Based Inpainting (CBI) [12], Deformable GAN (DSC) [43], Variational U-Net (VUnet) [10] and Dense Pose Transfer (DPT) [33]. The qualitative results are shown in Fig. 3. It can be observed that our results show higher realism and better preserve identity and garment details compared to the other methods.

The quantitative results are provided in Table 1. Due to inconsistent reporting (or unavailability) of the metrics for the existing approaches, we computed them ourselves. To this end, we collected the results of 176 testing pairs for each state-of-the-art method (the testing pairs and results were kindly provided by the authors of Coordinate Based In-painting [12]) and used them for this report. We use the following two metrics for comparison - 1) Structural Similarity Index (SSIM) [64] 2) Learned Perceptual Image Patch Similarity (LPIPS) [61]. SSIM is a structure preservation metric widely used in the existing literature. Though it is an excellent metric for assessment of image degradation quality, it often does not reflect human perception [61]. On the other hand, the recently introduced LPIPS claims to capture human judgment better than existing hand-designed metrics. In terms of SSIM, we perform as well as the existing methods, whereas we significantly outperform them on LPIPS metric. Please note that similar to

Table 1: Comparison with state-of-the-art methods, using various perceptual metrics, Structural Similarity Index (SSIM) [64] and Learned Perceptual Image Patch Similarity (LPIPS) [61]. ↑ (↓) means higher (lower) is better.

	SSIM ↑	LPIPS ↓
CBI [12]	0.766	0.178
DSC [43]	0.750	0.214
VUnet [10]	0.739	0.202
DPT [33]	0.759	0.206
Ours	**0.768**	**0.164**
GT	1.0	0.0

other learning-based methods, our approach will struggle with poses that are far from those seen in the training set. However, our method performs well in such scenarios for many cases. Qualitative results on some target poses outside of training dataset distribution are shown in Fig. 5.

4.3   User Study

To assess the qualitative impact of our method, we perform an extensive user study which compares it with two other state-of-the-art pose transfer methods – Coordinate Base Inpainting (CBI) [12] and DensePose Transfer (DPT) [33]. We train on the DeepFashion dataset [27] and generate renderings on the test split. The user study follows several criteria. First, it covers as large a variety of source and target poses. Second, the ratio between the male and female samples reflects the same ratio of the dataset. It also contains failure cases as those shown in Fig. 8 with difficult decisions. In total, we prepare 26 samples containing the source image (explicitly marked as such) and three novel views generated by CBI, DPT and our method (labeled as view A, B or C in randomised order). For each sample, two questions are asked: 1) Which view looks the most like the person in the source image? and 2) Which view looks the most realistic?

The user study was performed with a browser interface, the order of questions is randomised, and 46 anonymous participants submitted their answers. The results are as follows. The first question has been answered in 46.2% of the cases in favour of CBI, and in 53.8% of the cases in favour of our method. In all cases, DPT has always been the last choice. The second question has been answered by 30.8% of the participants in favour of CBI, and by 69.2% of the participants in favour of our approach. Again, DPT was preferred in no case.

The user study shows that our method achieves state-of-the-art quality in preserving the identity, and significantly outperforms the baselines in the realism of the generated images. In 23% of the overall cases, the participants have preferred CBI as the best identity-preserving method and, at the same time, our method as those producing the most realistic renderings. In contrast, there was only one case (3.8%) when our method had been voted as the best identity-preserving and, at the same time, CBI was chosen as the approach producing most realistic renderings.

Fig. 4: Results of different baselines and our full model. *No-Int* has no intermediate loss, *Warp* and *WarpCond* perform translation on warped partial texture and *IP* inpaints full colour-texture followed by translation. Under extreme poses and strong occlusions, our method outperforms all the baselines (see Sec. 4.4).

Fig. 5: Generalisation of our method to new body poses. The images of the target pose are obtained from the internet.

4.4   Ablation Study

To study the advantage of the learned neural texture over other natural choices of texture-based human re-rendering, we created the following three baselines.

IP - This baseline involves two stages. First, we train an inpainting network to generate the full UV texture map from the partial UV texture map extracted from the input image. We use the same in-painting loss function as described in Section 3.5 for training this network. After the convergence, we fix and use this network to generate full colour texture from a partial input texture. This full 3-channel UV texture map is then rendered into an intermediate image, and translated through a trained RenderNet g.

Warp - In this experiment, we warp the incomplete partial UV texture map to the target pose. The reposed incomplete 3-channel intermediate image is then fed to a trained RenderNet g to produce realistic output.
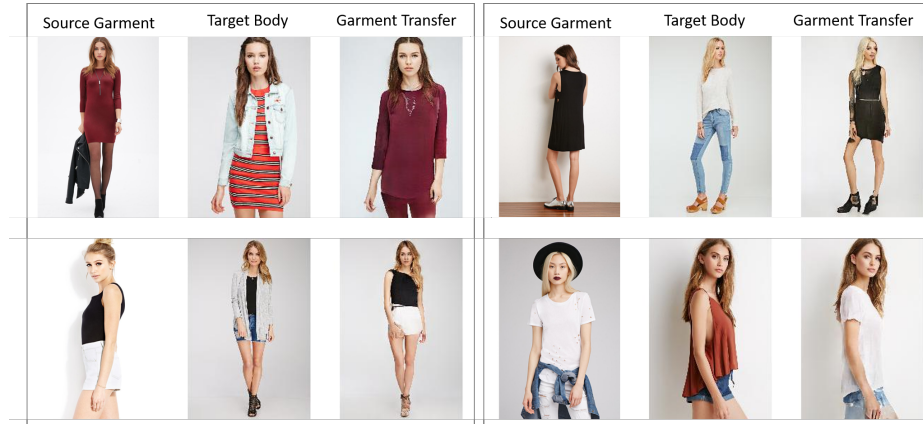
| Source Garment | Target Body | Garment Transfer | Source Garment | Target Body | Garment Transfer |

Fig. 6: Garment Transfer: Our approach can also be used to render garments from a source image onto the person in a target image.

WarpCond - In this experiment, we warp the partial source texture to the target pose P_T similar to the previous experiment. In addition to the reposed incomplete texture, we also condition the generator network g with the target DensePose image. The target DensePose image acts as a cue to the generator when the texture information is missing.

In all these baselines, the architecture of RenderNet g and the losses on the generated image are the same as ours. The only difference is in the number of input channels to RenderNet. Besides, we perform an ablation experiment with an identical pipeline as ours, except we do not enforce any intermediate texture loss – **No-Int**. The qualitative results of all the networks are shown in Fig. 4. It can be seen that our methods using a richer intermediate representation (*full* and *No-Int*) produce more realistic images than the other baselines. *Baseline-IP* performs well but produces smooth output compared to the other methods. Because of the lack of details, *Baseline-Warp* often produces non-realistic output in both face and garment regions. When the incomplete texture information is supervised with additional DesnsePose image (as in *Baseline-WarpCond*), the output is of higher quality. In the presence of strong occlusions, the method fails, as the translation network g is incapable of performing both inpainting and realistic rendering at the same time. In contrast, our methods performed well in all the scenarios. Adding the intermediate texture loss (to mimic real texture) to the part of neural texture helps our network to converge faster. However, over a large number of iterations, the quality of the final result without such intermediate loss (*No-Int*) is similar to that with intermediate loss.

4.5   Garment Transfer

Our method can be naturally extended to perform garment transfer without any further training. Given an image of a person with the source body, we extract

Fig. 7: Motion transfer on the Fashion dataset [59]. Our approach also can generate realistic renderings for a sequence of poses given a single source image.

the partial texture T'_s of the 'body regions' (*e.g.,* face, hands and regions with garments which remain unchanged). We use part indices provided by DensePose to extract the partial texture of the required body parts. Next, we extract the partial texture T'_t of the 'garment regions' of an image with the desired target garments. We make a union of the extracted partial textures $T'_s \cup T'_t$ based on their texel regions and feed it to our pipeline with the pose P_s of the body image. Note that texel occupancies of T'_s and T'_t are mutually exclusive as they are extracted from different body parts. See Fig. 6 for the qualitative results.

4.6 Motion Transfer

Even though we did not train specifically for generating videos, our method can be applied to each frame of a driving video to create motion transfer. To this end, we keep the source image of the imitator fixed and use the pose from the actor of the driving video (for each frame) in our system to create image animation. We perform the experiment on Fashion Dataset [59] and show our results in Fig. 7.

Please refer to the supplemental document and the accompanying video for more results.

5 Discussion

Limitations. Even though we produce high-quality novel views which are preserving the identity and look very realistic, there remain certain limitations for future work to address. Fig. 8 visualises two representative examples which are difficult for our as well as competing methods. In the first row, the head in the source image is only partially visible so that the methods have high uncertainty in the frontal facial view and change the gender to female (*e.g.,* hallucinate long

Fig. 8: Limitations: Even though our method produces better quality results than all competing approaches, it nevertheless has some limitations, which are also shared by all competing methods. The top row highlights failures arising out of biases in the training set, while the bottom row highlights failures owing to fine scaled textures which are not effectively captured by any approach.

hair). In the second case, the source texture is too fine-grained for the methods so that some hallucinate repeating patterns and the other ones generate patterns reminiscent of noise. In this case, our method generates a texture which is neither repetitive nor looks like noise, and which is still far from the reference.

Future Extension of Our UV Feature Maps. Instead of sampling RGB textures from the input image to construct a partial UV texture map, learned CNN based features could be used to construct a more informative partial UV feature map, which putatively captures off-geometry details not modelled by the SMPL mesh. Then FeatureNet would convert this partial UV feature map to a full UV feature map. Another alternative would be to use displacement map prediction similar to prior work [21, 2] to capture off-geometry details.

6   Conclusion

In this work, we present an approach for human image synthesis, which allows us to change the camera view and the pose and garments of the subject in the source image. Our approach uses a high-dimensional UV feature map to encode appearance as an intermediate representation, which is then re-posed and translated using a generator network to achieve realistic rendering of the input subject in a novel pose. We qualitatively and quantitatively demonstrate the efficacy of our proposed approach at better preserving identity and garment details compared to the other competing methods. Our system, trained once for pose guided image synthesis, can be directly used for other tasks such as garment transfer and motion transfer.

Acknowledgements

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105112 (2011)
2. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: International Conference on Computer Vision (ICCV) (2019)
3. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.V.: Synthesizing images of humans in unseen poses. Computer Vision and Pattern Recognition (CVPR) (2018)
4. Buehler, C., Bosse, M., McMillan, L., Gortler, S.J., Cohen, M.F.: Unstructured lumigraph rendering. In: SIGGRAPH (2001)
5. Carceroni, R.L., Kutulakos, K.N.: Multi-view scene capture by surfel sampling: From video streams to non-rigid 3d motion, shape and reflectance. International Journal of Computer Vision (IJCV) **49**(2), 175–214 (2002)
6. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: International Conference on Computer Vision (ICCV) (2019)
7. Chaurasia, G., Duchêne, S., Sorkine-Hornung, O., Drettakis, G.: Depth synthesis and local warps for plausible image-based navigation. ACM Transactions on Graphics **32** (2013)
8. Debevec, P., Yu, Y., Borshukov, G.: Efficient view-dependent image-based rendering with projective texture-mapping. Eurographics Workshop on Rendering (1998)
9. Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., et al.: Fusion4d: Real-time performance capture of challenging scenes. ACM Trans. Graph. **35**(4) (2016)
10. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: Computer Vision and Pattern Recognition (CVPR). pp. 8857–8866 (2018)
11. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: SIGGRAPH. p. 4354 (1996)
12. Grigor'ev, A.K., Sevastopolsky, A., Vakhitov, A., Lempitsky, V.S.: Coordinate-based texture inpainting for pose-guided human image generation. Computer Vision and Pattern Recognition (CVPR) pp. 12127–12136 (2019)
13. Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q., Liu, Y.: Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. ACM Trans. Graph. **36**(4) (2017)
14. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
15. Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., Legendre, C., Luo, L., Ma, C., Li, H.: Deep volumetric video from very sparse multi-view performance capture. In: European Conference on Computer Vision (ECCV). pp. 351–369 (2018)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision (ECCV). pp. 694–711 (2016)
17. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR) (2018)
18. Kim, H., Elgharib, M., Zollöfer, Michael Seidel, H.P., Beeler, T., Richardt, C., Theobalt, C.: Neural style-preserving visual dubbing. ACM Transactions on Graphics (TOG) **38**(6), 178:1–13 (2019)

19. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (TOG) **37** (2018)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
21. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. International Conference on 3D Vision (3DV) pp. 643–653 (2019)
22. Levoy, M., Hanrahan, P.: Light field rendering. In: SIGGRAPH. p. 3142 (1996)
23. Liu, L., Xu, W., Zollhoefer, M., Kim, H., Bernard, F., Habermann, M., Wang, W., Theobalt, C.: Neural rendering and reenactment of human actor videos. ACM Transactions on Graphics (TOG) (2019)
24. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Computer Vision and Pattern Recognition (CVPR). pp. 212–220 (2017)
25. Liu, W., Piao, Z., Jie, M., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: International Conference on Computer Vision (ICCV) (2019)
26. Liu, Y., Dai, Q., Xu, W.: A point-cloud-based multiview stereo algorithm for free-viewpoint video. IEEE Transactions on Visualization and Computer Graphics (TVCG) **16**(3), 407–418 (2010)
27. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Computer Vision and Pattern Recognition (CVPR). pp. 1096–1104 (2016)
28. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. ACM Trans. Graph. (SIGGRAPH) **38**(4) (2019)
29. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
30. Ma, L., Sun, Q., Georgoulis, S., van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. Computer Vision and Pattern Recognition (CVPR) (2018)
31. Martin Brualla, R., Lincoln, P., Kowdle, A., Rhemann, C., Goldman, D., Keskin, C., Seitz, S., Izadi, S., Fanello, S., Pandey, R., Yang, S., Pidlypenskyi, P., Taylor, J., Valentin, J., Khamis, S., Davidson, P., Tkach, A.: Lookingood: Enhancing performance capture with real-time neural re-rendering. ACM Transactions on Graphics (TOG) **37** (2018)
32. Matsuyama, T., Xiaojun Wu, Takai, T., Wada, T.: Real-time dynamic 3-d object shape reconstruction and high-fidelity texture mapping for 3-d video. IEEE Transactions on Circuits and Systems for Video Technology **14**(3), 357–369 (2004)
33. Neverova, N., Güler, R.A., Kokkinos, I.: Dense pose transfer. European Conference on Computer Vision (ECCV) (2018)
34. Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., Kim, D., Davidson, P.L., Khamis, S., Dou, M., et al.: Holoportation: Virtual 3d teleportation in real-time. In: Annual Symposium on User Interface Software and Technology. p. 741754 (2016)
35. Pandey, R., Tkach, A., Yang, S., Pidlypenskyi, P., Taylor, J., Martin-Brualla, R., Tagliasacchi, A., Papandreou, G., Davidson, P., Keskin, C., Izadi, S., Fanello, S.: Volumetric capture of humans with a single rgbd camera via semi-parametric learning. In: Computer Vision and Pattern Recognition (CVPR) (2019)

36. Pfister, H., Zwicker, M., van Baar, J., Gross, M.: Surfels: Surface elements as rendering primitives. In: SIGGRAPH. p. 335342 (2000)
37. Rieza Alp Gueler, Natalia Neverova, I.K.: Densepose: Dense human pose estimation in the wild. In: Computer Vision and Pattern Recognition (CVPR) (2018)
38. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241 (2015)
39. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. International Conference on Computer Vision (ICCV) (2019)
40. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016)
41. Shade, J., Gortler, S., He, L.w., Szeliski, R.: Layered depth images. In: SIGGRAPH. p. 231242 (1998)
42. Shysheya, A., Zakharov, E., Aliev, K.A., Bashirov, R., Burkov, E., Iskakov, K., Ivakhnenko, A., Malkov, Y., Pasechnik, I., Ulyanov, D., Vakhitov, A., Lempitsky, V.: Textured neural avatars. In: Computer Vision and Pattern Recognition (CVPR) (2019)
43. Siarohin, A., Lathuilière, S., Sangineto, E., Sebe, N.: Appearance and pose-conditioned human image generation using deformable gans. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019)
44. Siarohin, A., Lathuilire, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: Computer Vision and Pattern Recognition (CVPR) (2019)
45. Siarohin, A., Lathuilire, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: Conference on Neural Information Processing Systems (NeurIPS) (2019)
46. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
47. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhöfer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: Computer Vision and Pattern Recognition (CVPR) (2019)
48. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
49. Tao, Y., Zheng, Z., Guo, K., Zhao, J., Quionhai, D., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performance with inner body shape from a depth sensor. In: Computer Vision and Pattern Recognition (CVPR) (2018)
50. Thies, J., Zollhöfer, M., Niener, M.: Deferred neural rendering: image synthesis using neural textures. ACM Transactions on Graphics (TOG) **38** (2019)
51. Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., Nießner, M.: Image-guided neural object rendering. In: International Conference on Learning Representations (ICLR) (2020)
52. Tung, T., Nobuhara, S., Matsuyama, T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In: International Conference on Computer Vision (ICCV). pp. 1709–1716 (2009)
53. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Computer Vision and Pattern Regognition (CVPR) (2017)

54. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Computer Vision and Pattern Recognition (CVPR) (2018)
55. Waschbüsch, M., Würmlin, S., Cotting, D., Sadlo, F., Gross, M.: Scalable 3d video of dynamic scenes. The Visual Computer **21**(8), 629–638 (2005)
56. Xu, Z., Bi, S., Sunkavalli, K., Hadap, S., Su, H., Ramamoorthi, R.: Deep view synthesis from sparse photometric images. ACM Trans. Graph. **38**(4), 76:1–76:13 (2019)
57. Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., Li, J., Dai, Q., Liu, Y.: Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In: International Conference on Computer Vision (ICCV). pp. 910–919 (2017)
58. Yu, T., Zheng, Z., Zhong, Y., Zhao, J., Dai, Q., Pons-Moll, G., Liu, Y.: Simulcap: Single-view human performance capture with cloth simulation. In: Computer Vision and Pattern Recognition (CVPR) (2019)
59. Zablotskaia, P., Siarohin, A., Sigal, L., Zhao, B.: Dwnet: Dense warp-based network for pose-guided human video generation. In: British Machine Vision Conference (BMVC) (2019)
60. Zhang, L., Curless, B., Seitz, S.M.: Spacetime stereo: shape recovery for dynamic scenes. In: Computer Vision and Pattern Recognition (CVPR) (2003)
61. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Computer Vision and Pattern Recognition (CVPR) (2018)
62. Zhao, B., Wu, X., Cheng, Z.Q., Liu, H., Jie, Z., Feng, J.: Multi-view image generation from a single-view. In: ACM International Conference on Multimedia. p. 383391 (2018)
63. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Dance dance generation: Motion transfer for internet videos. In: International Conference on Computer Vision Workshops (ICCVW) (2019)
64. Zhou Wang, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
65. Zhu, H., Su, H., Wang, P., Cao, X., Yang, R.: View extrapolation of human body from a single image. In: Computer Vision and Pattern Recognition (CVPR) (2018)
66. Zhu, J.Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J., Freeman, B.: Visual object networks: image generation with disentangled 3d representations. In: Conference on Neural Information Processing Systems (NeurIPS). pp. 118–129 (2018)