Coordinate-based Texture Inpainting for Pose-Guided Human Image Generation

Artur Grigorev [1,2] Artem Sevastopolsky [1,2] Alexander Vakhitov [1] Victor Lempitsky [1,2]
[1] Samsung AI Center, Moscow, Russia
[2] Skolkovo Institute of Science and Technology (Skoltech), Moscow, Russia
{a.grigorev,a.sevastopol,a.vakhitov,v.lempitsky}@samsung.com

Abstract

We present a new deep learning approach to pose-guided resynthesis of human photographs. At the heart of the new approach is the estimation of the complete body surface texture based on a single photograph. Since the input photograph always observes only a part of the surface, we suggest a new inpainting method that completes the texture of the human body. Rather than working directly with colors of texture elements, the inpainting network estimates an appropriate source location in the input image for each element of the body surface. This correspondence field between the input image and the texture is then further warped into the target image coordinate frame based on the desired pose, effectively establishing the correspondence between the source and the target view even when the pose change is drastic. The final convolutional network then uses the established correspondence and all other available information to synthesize the output image. A fully-convolutional architecture with deformable skip connections guided by the estimated correspondence field is used. We show state-of-the-art result for pose-guided image synthesis. Additionally, we demonstrate the performance of our system for garment transfer and pose-guided face resynthesis.

1. Introduction

Learning human appearance from a single image (one-shot human modeling) has recently become an area of high research interest. One interesting kind of the problem, which has a number of potential applications in augmented reality and retail, is pose-guided image generation [20]. Here, the task is to resynthesize the view of a person from a new viewpoint and in a new pose, given a single input image. The progress in this problem benefits from the recent advances in human pose estimation and deep generative convolutional networks (ConvNets). A particular challenging setup considers humans wearing complex clothing, such as encountered in fashion photographs.

In this work we suggest a new approach for pose-guided person image generation. The approach is based on a pipeline that includes two deep generative ConvNets. The first convolutional network to estimate the texture of the human body surface from a small part of this texture (texture completion/inpainting). This texture is then warped to the new pose to serve as an input to the second convolutional network that generates the new view.

One novelty of the approach lies in the texture estimation part (Figure 1), where the challenge is to utilize the natural symmetries of the human body. This task is non-trivial since the part of the texture that is known changes from one input image to another. As a result, straightforward image-to-image translation approaches result in very blurred textures, where the colors predicted at unknown locations are effectively averaged over very large number of input locations.

To solve this problem, we suggest a new method for texture completion, which we call *coordinate-based texture inpainting*, and which results in a significant boost of the visual quality output for the entire pipeline. The method is based on a simple idea. Rather than working directly with colors of texture elements, the inpainting network works with coordinates of the texture elements in the source view. These values are analyzed by the inpainting network and then extended into the unknown part of the texture, so that each unknown texture element gets assigned a coordinate in the source view. Thus, a correspondence between source pixels and all points on the body surface is estimated. Using the estimated correspondence, the colors of each texture element can be transferred from the source view. The inpainting thus happens in the coordinate-space, while the extraction of colors from the source image, which generates the final texture, happens *after* the inpainting. As a result, the inpainted textures retain high-frequency details from the source images.

Given the detailed texture generated by the coordinate-based inpainting process, the next step of the pipeline warps both the color texture and the source image coordinate maps according to the target pose (which similarly to [22] is defined by the DensePose [11] descriptor). The final stage of

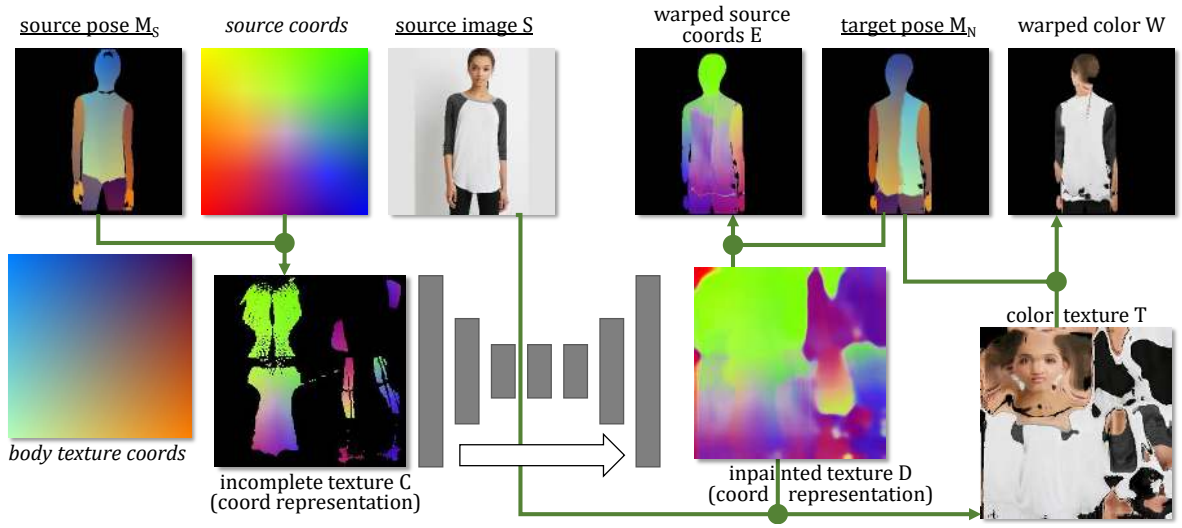

Figure 1. Coordinate-based texture inpainting. The scheme depicts the first (out of the two) part of our pipeline. Given the source pose (estimated by DensePose and converted to SMPL format), we rasterize the source coordinates of the known texture elements (e.g. by warping the source coordinate meshgrid). The resulting map is completed using deep convolutional network (gray) into a complete body texture, where for each texel a corresponding pixel coordinate in the source image is assigned. This correspondence map is then used to estimate the color texture. The second warping transforms the estimated texture maps into the target coordinate frame using the target pose, on which the resynthesis is conditioned (*Data known at test time is underlined. 2D meshgrid arrays that define colormaps in the plot are in italic. Warping transforms are shown using green arrows, where the side connections correspond to the warp coordinates and straight arrows point from the data being warped*).

the pipeline takes the warped images along with the pose information and maps it to the target image using a deep fully-convolutional encoder-decoder architecture with skip connections. The input image is used in this translation network, while the warped source image coordinates obtained during the texture inpainting process, are used to route the deformable skip connections [28].

Our contribution is thus two-fold. First, we suggest the new texture completion method that allows to retain high-level texture details even under large uncertainty. Secondly, we present a pose-guided person image generation pipeline that utilizes this method in two ways (to inpaint texture and to guide deformable skip connections) in order to generate new views with high realism and abundant texture details. Our method is evaluated on the popular Deep Fashion dataset [18], where it obtains good results outperforming prior art. Furthermore, we additionally demonstrate the efficacy of coordinate-based texture inpainting idea on the face texture inpainting task for in-the-wild new view synthesis of faces, using the 300-VW dataset [26]. As a coda, we show that a small modification of our approach can successfully be used to perform garment transfer (virtual try-on) with convincing results.

2. Related Work

Warping-based resynthesis. There is a strong interest in using deep convolutional networks for generating realistic images [10, 4]. In the resynthesis case, when new images are generated by the change of geometry and appearance of the input images, it has been shown that using warping modules greatly enhances the quality of the re-synthesized images [7, 38]. The warping modules in this case are based on the differentiable (backward) grid sampler layer, which was first introduced as a part of Spatial Transformer Networks (STN) [14]. A large number of follow-up works on resynthesis reviewed below have relied on backward sampler. Here we revisit this building block and advocate the use of forward warping module.

Neural human resynthesis. Neural-based systems for transforming an input view of a person into a new view with modified pose has been suggested recently. The initial works [20, 21, 5] used encoder-decoder type of architectures in order to perform resynthesis. More recent works use warping models that redirect either raw pixels or intermediate activations of the source view [30, 28, 36, 22]. Our approach falls into this category and is most related to [22], as it utilizes the DensePose parameterization [11] within the network, and to [30] as we use the idea of deformable skip connections from [30]. We compare our results to [22, 30] and additionally to [5] extensively.

Texture completion. Image inpainting based on deep convolutional networks is attracting increasing attention at

the moment. Special variants of convolutional architectures adapted to the presence of gaps in the input data include Sheppard Networks [24], Sparsity-Invariant CNNs [31], networks with Partial Convolutions [17], networks with Gated Convolutions [35]. We use the latter variant for our texture inpainting network. Learning body texture inpainting has two specific parts that distinguish it from generic image inpainting. First, complete textures may not be easily available and it is desirable to devise a method that can be trained from partial images. Secondly, textures are spatially aligned and possess symmetry structures that can be exploited, which calls for special-purpose algorithms. We are aware of only a few works which address these challenges specifically. Thus, UV-GAN [3] utilizes the main axial symmetry of a face by passing an image and its flipped copy to an inpainting ConvNet. The system in [36] estimates a matrix that corresponds to the probabilities of SMPL model vertices to have similar colors, and use it to color vertices with unobserved colors.

Garment transfer. We also show that a small modification of our approach can be used to transfer clothing from the photograph of one person to the photograph of a different person in a different pose. Most existing works that utilize neural networks can only handle very limited amount of deformation between the source image and the target view [12, 15, 32]. The only work that we are aware of that can handle similar amount of pose change is SwapNet [23], which however only present results at low resolution. We perform comparison to [23] in the experimental section.

Face resynthesis. Our approach is related to a number of very recent face resynthesis works that operate by warping the input image into the output image. These works include deforming autoencoders [27] and X2Face [34]. An older class of works going back to the seminal Blanz and Vetter morphable model [1] estimate face texture from its fragment using a parametric model.

3. Methods

Problem formulation. Our goal is to synthesize the new view of the person N from the source view S. The resynthesis progresses by estimating the texture T. Below, we use the indexing $[x, y]$ to denote locations in the image frame (both the source and the new view), and we use the indexing $[u, v]$ to denote locations in the texture. We refer to source and target image elements and locations as *pixels*, and to texture elements and positions as *texels*.

The texture is linked with the source and the new views, and following [22] we assume that both for the source and the new view a mapping from a subset of the pixels covering the body (excluding hair and loose clothing) to the body texture positions is known. We thus assume that for each pixel $[x, y]$ in the source image (respectively in the new image) exists a mapping $M_S[x, y]$ (respectively $M_N[x, y]$) that associates with $[x, y]$ a position $[u, v] = [M_S^1[x, y], M_S^2[x, y]]$ (respectively, $[u, v] = [M_N^1[x, y], M_N^2[x, y]]$) on the texture. For pixels $[x, y]$ that do not fall within the projection of the human body, the mappings M_N and M_S are undefined.

We assume that $M_S[x, y]$ and $M_N[x, y]$ are given and our goal is thus to estimate the new unknown view N given its body texture mapping $M_N[x, y]$, as well as the known source view S and its body texture mapping M_S.

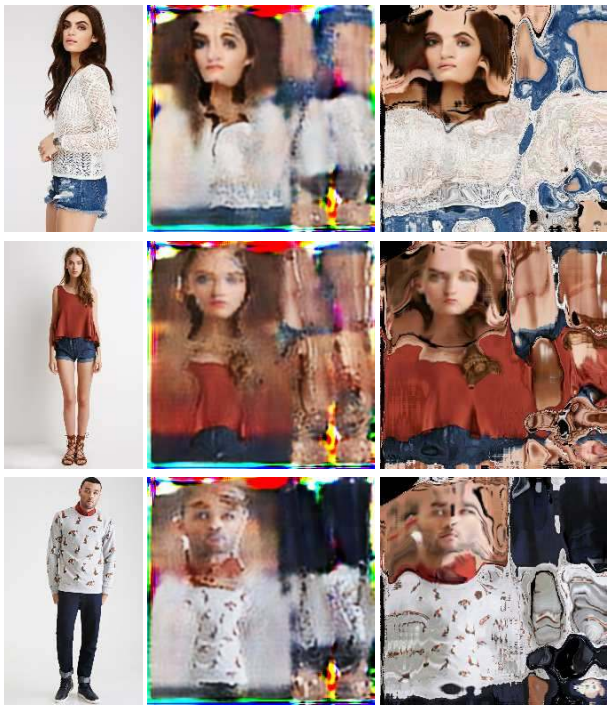

Figure 2. Body surface textures estimated using color-based inpainting (middle) and coordinate-based inpainting (right) for the inputs on the holdout set (left). Both inpaintings are generated using deep networks that were trained end-to-end with a variety of standard losses. Coordinate-based inpainting generates textures with more details leading to better final resynthesis results.

Texture map format and output conditioning. We use the SMPL texture format [19]. To make our approach comparable with [22], we estimate the mappings M_S and M_N based on DensePose [11], and then convert them to SMPL coordinates using a predefined mapping (provided with the DensePose). Thus, unlike [22], we use a single body texture during transfer. The information that is used to encode the source and the target pose is however exactly the same (the DensePose encoding), making the methods directly comparable.

Coordinate-based texture inpainting. The first step of our pipeline estimates the complete body surface texture from the source image S, and the mapping M_S. We first rasterize the source image coordinates over texture using warping. In more detail, we use scattered interpolation with bilinear kernel, so that each source pixel $[x, y]$ is rasterized at position $[M_S^1[x, y], M_S^2[x, y]]$. Unlike [22], we rasterize not the color values, but the values x and y themselves (in other words we apply scattered interpolation to the meshgrid array). The result of this warping step is the source coordinate map C, which for each texture element (texel) $[u, v]$ defines a corresponding location $[x, y] = [C^1[u, v], C^2[u, v]]$ in the source image. Since only a part of a human body can be visible in the source photograph, for a big part of texels, the source image location is undefined. When passing C into the network, we set the unknown values to a negative constant (-10), and also provide the network with the mask $C'[u, v]$ of known texels.

The first learnable module of our pipeline is the inpainting network $f(C, C'; \phi)$ with learnable parameters ϕ that takes an incomplete coordinate map C in the texture space along with the mask of known texels, and outputs a completed and corrected source correspondence map D, where for each $[u, v]$ the corresponding location in the source image is defined:

$$D = f(C, C'; \phi)\,. \tag{1}$$

The mapping f has a fully-convolutional structure. The task of the network is to learn the symmetries typical for human body and human dress, such as the left-right symmetry between body parts as well as less obvious symmetries. E.g. the network has a chance to learn that many clothings have repeated textures, so that if a guess needs to be made about the texture of the back from the front view, the best the network can do is to copy the frontal texture. Since the network f deals with the inpainting task, we utilize the recently proposed gated convolution layers [35] instead of standard convolutional layers. We use an hourglass (without skip-connection) architecture with 14 convolutional layers and 2.8 millions of parameters.

Given the estimated source correspondence map D, we can obtain the completed texture by sampling the original image using the locations prescribed by D:

$$T[u, v] = S[D^1[u, v], D^2[u, v]]\,. \tag{2}$$

where the bilinear sampling operator [14] is used to sample the source image at fractional locations.

It is interesting to compare the way our approach (*coordinate-based inpainting*) obtains the complete texture with the way the texture is obtained by other texture inpainting approaches (*color-based inpainting*), e.g. [22, 3, 36]. In the case of the color-based inpainting, the sampling (2) and the inpainting operation (1) are swapped, i.e. the colors are first sampled from the source image to the texture leading to an incomplete color texture and then the incomplete color texture is inpainted using a learnable convolutional architecture. As we have compared the two approaches, we have found that due to a very high uncertainty and multimodality of the texture inpainting task, the color-based inpainting produces the textures with very blurred details as compared to the coordinate-based inpainting (see Fig. 6). As will be shown in the experiments, when embedded into end-to-end resynthesis pipeline, considerably better results are obtained with coordinate-based inpaintings.

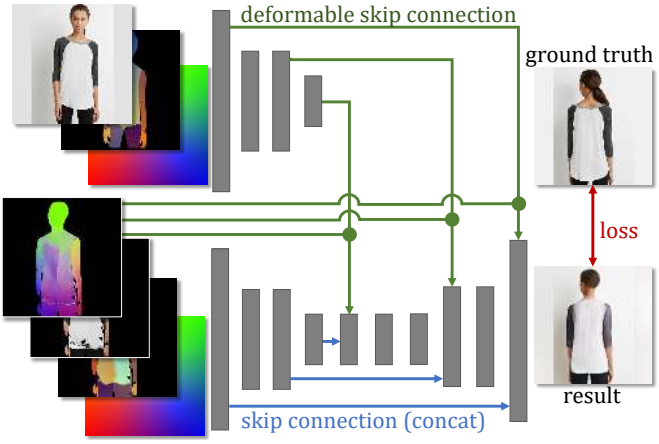

Figure 3. Final resynthesis. The second (of the two) part of our pipeline that takes the maps computed by the inpainting stage and map them to the final output image. Two separate encoders are used for maps aligned with the source pose (source pose, source image, meshgrid) and for maps aligned with the target pose (target pose, warped color texture, warped source coordinate map, meshgrid). The network has a U-Net type architecture (with intermediate residual blocks). Deformable skip connections are used to pass the activations of the source coordinate encoder to the joint decoder. The estimated correspondence map between the target and the source image is used to guide the deformable skip connections. Standard loss functions computed between the output of the pipeline and the ground truth target image in each pair are used for learning.

New view resynthesis. Similarly to [22], in order to resynthesize the target view, we warp the obtained color texture T as well as the coordinate-based texture map D to the new image frame, using the backward bilinear warping:

$$W[x, y] = T\left[M_N^1[x, y], M_N^2[x, y]\right]\,, \tag{3}$$

$$E[x, y] = D\left[M_N^1[x, y], M_N^2[x, y]\right]\,, \tag{4}$$

where W and E are the new maps containing RGB color and the source view location for each body pixel of the target view. The values for non-body pixels are undefined (set to zeros in practice). The warping (4) effectively estimates the correspondence between the target and the source views.

The final stage of our pipeline is a single convolutional network g that converts (translates) the maps W, E, as well as the input maps S, M_S, and M_N into an output image N. We first consider a straightforward architecture that takes all five maps, together with the meshgrid defined over the image frame as an input and uses the architecture of [16] with added skip-connections to synthesize the output image.

One caveat is that the input maps S, M_S are not in any ways aligned with the target new image, which is known to cause problems. As a more advanced variant (Figure 3), we have used the deformable skip connections [28] idea. Towards this end, we use a separate encoder part for the two maps S and M_S concatenated with a separate meshgrid. When passing the activations of this encoder into the decoder, we use the warp field E and its downsampled versions to do bilinear resampling of the activations. In the experiments, we compare both variants of the architecture and find that deformable skip connections considerably boost the performance of our pipeline.

Training procedure. Our complete pipeline includes two convolutional networks, namely the inpainting network f that performs coordinate-based texture completion, and the final network g. Both networks are trained on quadruplets $\{S, M_S, N, M_N\}$. We first train the network f by minimizing the loss comprising two terms: (1) the ℓ_1 difference between the input incomplete texture C and the inpainted texture D, where the difference is computed over texels that are observed in C; (2) the ℓ_1 difference between the inpainted texture D and the incomplete output texture that is obtained by warping the target image N into the texture space using the map M_N, where the difference is computed over texels that are observed in the output image.

After that, we fix f and optimize the weights of network g, where we minimize the loss between the predicted $\hat{N}$ and the ground truth new view N. Here, we combine the perceptual loss [16] based on the VGG-19 network [29], the style loss [8] based on the same network, the adversarial loss [10] based on the patch GAN discriminator [13] and the nearest neighbour loss introduced in [28] (that proved to be a good substitution for l_1 loss used in [22]). While the first network f can be fine-tuned during the second stage, we did not find it beneficial for the resulting image quality.

Garment transfer. A slight modification of our architecture allows it to perform garment transfer [12, 15, 32, 23]. Here, given two views A and B, we want to synthesize a new view, where the pose and the person identity is taken from the view B, while the clothing is taken from view A. We achieve this by taking the architecture outlined above, and additionally conditioning the network g on the masked image N' of the target view, where we mask out all areas except head (including face, hair, hats, and glasses) and hands (including gloves).

The network g is trained on the pairs of views of the same person, and effectively learns to copy heads and hands from N' to N. At test time, we provide the network the identity-specific image N' and the body texture mapping M_N that are both obtained from the image of a different person from the one depicted in the input view. We show that our architecture successfully generalizes to this setting and thus accomplishes the virtual re-dress task.

4. Applications and experiments

4.1. Pose-guided image generation

For the main experiments, we use the DeepFashion dataset (the in-shop clothes part) [18]. In general, we follow the same splits as used in [28, 22] that include 140,110 training and 8,670 test pairs, where clothing and models do not overlap between train and test sets.

Network architectures. For the texture inpainting network f we employ an hourglass architecture with gated convolutions from [35] which proved effective in image reconstruction tasks with large hidden areas. The refinement network g is also a hourglass network that has two encoders that map images by a series of gated convolutions interleaved with three downsampling layers resulting in $256 \times 64 \times 64$ feature tensors. This is followed by consecutive residual blocks and concluded by a decoder. The encoder and the decoder are also connected via three skip connections (at each of three resolutions). The encoder that works with S and M_S is connected to the decoder with deformable skip connections that are guided by the deformation field E. The network f has 2,824,866 parameters, and the network g has 11,382,984 parameters.

Comparison with state-of-the-art. We compare the results of our method (full pipeline) with three state-of-the-art works [22, 28, 5]. We again follow the previous work [22] closely using structural self-similarity (SSIM) along with its' multi-scale version (MS-SSIM) metrics [33] to measure the structure preservation and the inception score (IS) [25] to measure image realism. We also use recently introduced perceptual distance metric (LPIPS) [37] which measures distance between images using a network trained on human judgements (Table 1).

Additionally we perform a user study to compare our results with state-of-the-art based on 80 image pairs from the test set (the indices of the pairs, as well as the results of [22, 28, 5] were kindly provided by the authors of [22]). In the user study, we have shown our results alongside of [22, 28, 5] and asked to pick the variant, which was best fitting the ground truth (target) image. The source image was not shown. The order of presentation was normalized. 50

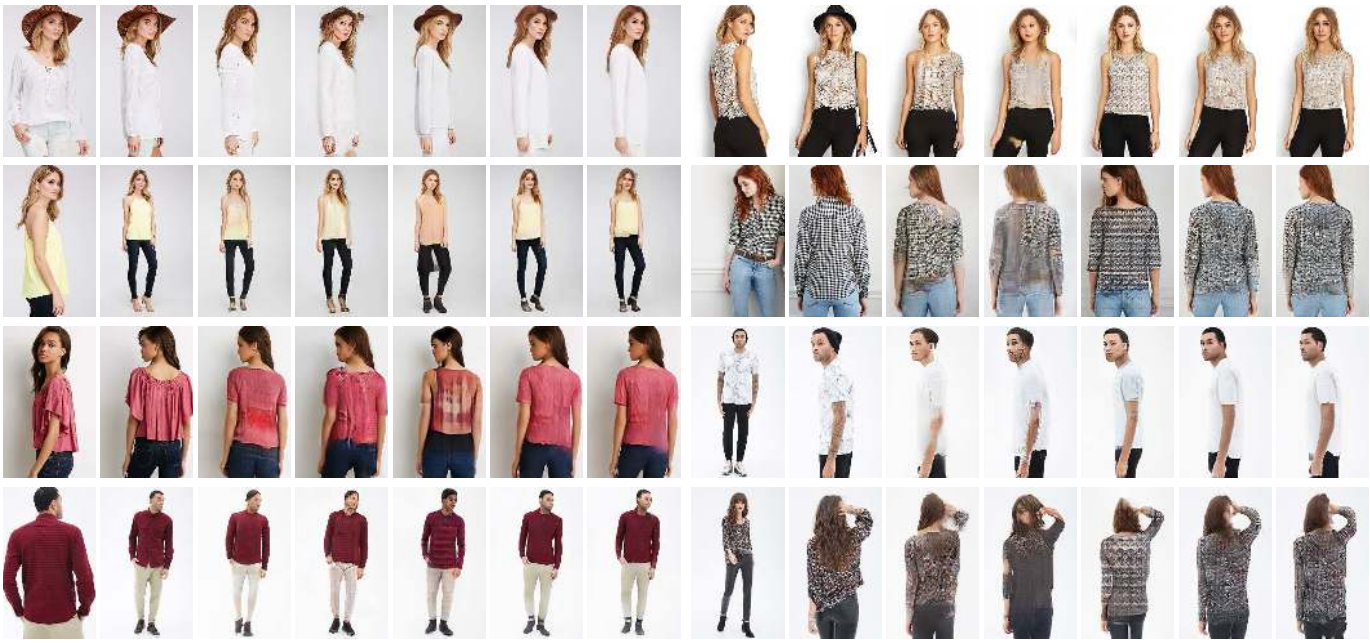

Figure 4. Side-by-side comparison with state-of-the-art (first eight samples from the test set). We show source image (SRC), ground truth in the target pose (GT), deformable GAN [28], our method conditioned on dense pose (Ours-D), and our method conditioned on keypoints (Ours-K). Consistently with the user study on a broader set, our method is more robust and has less artefacts than the state-of-the-art [28, 22] on this subset. *Electronic zoom-in recommended.*

people were involved in the user study. Each of them were to chose more realistic image in each of 80 pairs. In **90%** cases our reconstructions were preferred over those of [22] and in **76.7%** cases cases over [28], while against [5] our results were considered more realistic in **71.6%** cases (approximately 4000 pairs were compared in each of the three cases).

Ablation study. We evaluate the full variant of our approach that is described above, as well as the following ablations. In the *Ours-NoDeform* ablation we do not use the deformable skip-connections in the network f, resulting in a single encoder for W, E, S, M_S, M_N even though some of them (S, M_S) are aligned with the source view, while others (W, E, M_N) are aligned with the target view.

In the *RGB inpainting* ablation we additionally replace coordinate-based inpainting with color-space inpainting, so that the output of the texture inpainting stage is only the color texture T, which is warped according to M_N into the warped texture W aligned with the target view. Since the map E is unavailable in this scenario, no deformable skip-connections are used in this case. Finally, the *No textures* ablation simply uses the maps S, M_S, and M_N as an input to the translation network, ignoring texture estimation step altogether.

We compare the full version of the algorithm in terms of same four metrics: SSIM, MS-SSIM, IS and LPIPS. To ensure superiority of coordinate-based inpainting to color-based we have also performed a user study comparing *Ours-Full* and *RGB inpainting* methods. During this evaluation *Ours-Full* were preferred in **62.7%** cases.

	SSIM↑	MS-SSIM↑	IS↑	LPIPS↓
Ours	**0.791**	**0.810**	4.46	**0.169**
DPT [22]	0.785	0.807	3.61	—
DSC [28]	0.761	—	3.39	—
VUnet [5]	0.753	0.757	**4.55**	0.196

Table 1. Comparison with state-of-the-art. Our approach outperforms the other three in three of the four used metrics, although we found SSIM, MS-SSIM and IS to be much less adequate judgements of visual fidelity than user judgements. Arrows ↑, ↓ tell which value is better for the score larger or smaller, respectively. Since we do not have access to full test set and code of some methods, values for metrics not presented in the respective papers are missing.

Keypoint-guided resynthesis. It can be argued that our method (as well as [22]) has an unfair advantage over [28, 5] and other keypoint-conditioned methods, since DensePose-based conditioning provides more information about the target pose compared to just keypoints (skeleton). To address this argument, we have trained a fully-convolutional network that rasterizes the OpenPose [2]-detected skeleton over a set of maps (one bone per map) and train a network

Figure 5. Examples of garment transfer procedure obtained using a simple modification of our approach. In each triplet, the third image shows the person from the first image dressed into the clothes from the second image.

to predict the DensePose [22] result. We fine-tune our full network, while showing such "fake" DensePose results for the target image, effectively conditioning the system on the keypoints at test time. We add this variant to comparison and observe that the performance of our network in this mode is very similar to the mode with DensePose conditioning (Figure 4).

Garment transfer. We also show some qualitative results of the garment transfer (virtual try-on). The garment transfer network was obtained by cloning our complete pipeline in the middle of the training and adding the masked target image (with revealed face and hair) to the input of the network. During training background on ground truth targets is segmented out by the pretrained network [9] resulting in white background on try-on images. We use the DensePose coordinates to find the face part, and we additionally used the same segmentation network [9] to detect hair. As the training progressed, the network has quickly learned to copy the revealed parts through skip-connections, achieving the desired effect. We show examples of garment transfer in Figure 5. We conducted a user-study using 73 try-on samples provided by the authors of [23]. Participants were given quadruplets of images – cloth image, person image, our try-on result and result of [23] and asked to chose which of the try-on images seem more realistic. Since work of [23] produce only 128×128 images, our results were downsampled. Each sample was assessed by 50 people totalling in 3650 cases, of which our method were preferred in **57.1%**.

4.2. Pose-guided face resynthesis

To demonstrate the generality of our idea on texture inpainting, we also apply it to the additional task of face resynthesis. Here, reusing the pipeline used for full body resynthesis, we provide a pair of face images in different poses as a source and a new, unseen view. To estimate the mappings M_S and M_N we use PRNet [6] — a state-of-the-art 3D face reconstruction algorithm which provides a full 3D mesh with a fixed number of vertices (43867 in a publicly available version) and triangles (86906). A fixed precomputed mapping from the vertices numbers to their (u, v) texture coordinates is also provided with PRNet implementation. By processing source and target images with PRNet, we obtain estimated (x, y, z) coordinates of a 3D face mesh which leans on an image, such that (x, y) axes are aligned with image axes. We set $(u, v, 1)$ texture coordinates of each vertex as its (R, G, B) color and render a mesh onto an image via Z-buffer, which leaves pixels only visible on a camera view (those not occluded by different faces of a mesh). Similarly to the full body scenario, the obtained rendering for the source view reflects $M_S[x, y]$ mapping, and rendering for the new view reflects $M_N[x, y]$. The pipeline consists of two networks f and g which follow the same architectures as used for the full body view resynthesis. Provided with a source view image and a new view image, the system transfers facial texture from source image onto a pose of a new view image.

For this subtask, we use 300-VW [26] dataset of continuous interview-style videos of 114 people taken in-the-wild as a source of training data. Duration of each video is typically around 1 minute and the spatial resolution varies from 480 x 360 to 1280 x 720. Despite that original videos were taken in 25-30 fps, we took each sixth frame of a video in order to speed up the data preparation. Images are preliminarily cropped by a bounding box of 3D face found by PRNet with a margin of 10 pixels and bilinearly resized to a resolution of 128 x 128. Dataset was split into train and validation in proportion of 91 and 23 subjects respectively.

	Full body				Face			
	SSIM↑	MS-SSIM↑	IS↑	LPIPS↓	SSIM↑	MS-SSIM↑	IS↑	LPIPS↓
Ours-Full	0.791	0.810	**4.46**	**0.169**	**0.613**	**0.764**	**1.834**	**0.203**
Ours-NoDeform	**0.797**	0.815	3.23	0.198	0.609	0.758	1.819	**0.203**
RGB inpainting	**0.797**	**0.818**	3.02	0.198	0.595	0.745	1.821	0.221
No textures	0.796	0.812	3.295	0.202				

Table 2. Ablation study for both **full body** and **face** resynthesis. For all algorithms, evaluation is performed based on the same set of validation images. Arrows ↑, ↓ tell which value is better for the score — larger or smaller, respectively.

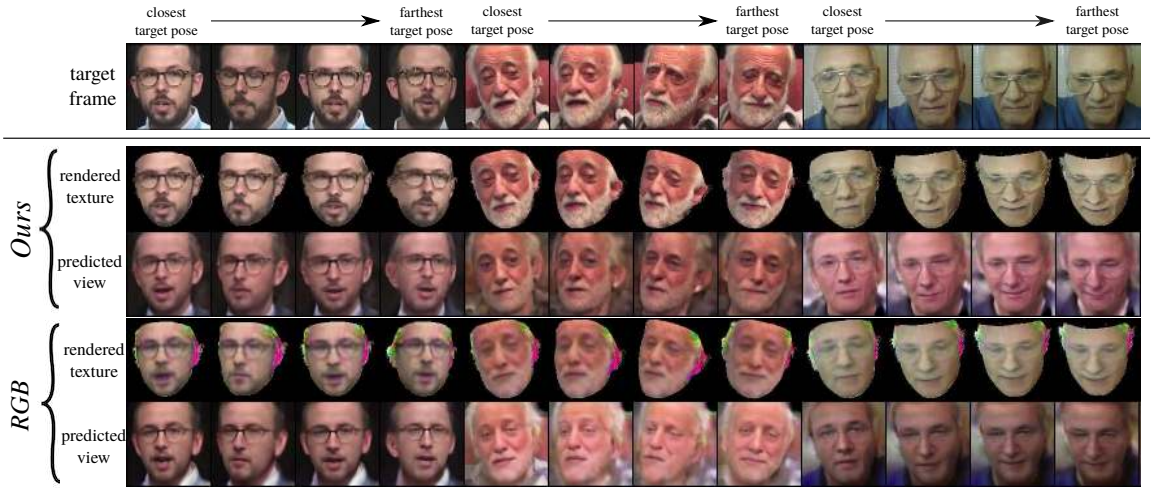

Figure 6. Predictions for several test samples. For each method, we take 3 random subjects, first video frame as a source frame and sample 4 target views according to the 4-quantiles of the pose difference distribution (see testing protocol in Subsection 4.2). For each subject, source frame is identical to the leftmost target frame shown. In the figure, *rendered texture* refers to the result of warping an inpainted texture onto a new view coordinates, and *predicted view* is a final algorithm output containing the result of texture transfer. Note the differences in sharpness between textures in *Ours* and in *RGB inpainting* and visual quality of their predicted views. *Electronic zoom-in recommended.*

New view resynthesis. Table 2 contains the results of the ablation study, in which we compare three investigated versions of the method (see Subsection 4.1). The reported values were computed for a subset of 1356 hold out images, collected by a following procedure. For each of 23 videos in the validation set, each 120^{th} frame of a video was selected as a source frame. Then, pose orientations of 3D models provided by PRNet were collected for all frames of the video, and angles between pose vector of a source frame 3D model and 3D models of all other frames were calculated. 4 target frames were selected for each source frame as the closest to all of the 4-quantiles of the angles cosine distribution. This way, we test the ability of a model to generalize on target poses both near and far from a source pose (Fig. 6).

5. Conclusion

We have present a new deep learning approach to pose-guided image synthesis. The approach works by estimating the texture of the human body, while a new method for coordinate-based texture inpainting allows to reconstruct detail-rich textures. The reconstructed textures are then used by final resynthesis. The user study suggests that the approach performs well and outperforms state-of-the-art methods [28, 22, 5]. We note that for smaller variation of pose, the mapping and estimation of the full texture may be unnecessary, and therefore more direct warping approaches such as [28] may be more appropriate under limited changes.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 3

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 6

[3] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: adversarial facial uv map completion for pose-invariant face recognition. In *Proc. CVPR*, pages 7093–7102, 2018. 3, 4

[4] Alexey Dosovitskiy, Jost Tobias Springenberg, Maxim Tatarchenko, and Thomas Brox. Learning to generate chairs, tables and cars with convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):692–705, 2017. 2

[5] Patrick Esser, Ekaterina Sutter, and Bjrn Ommer. A variational u-net for conditional appearance and shape generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 5, 6, 8

[6] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 7

[7] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European Conference on Computer Vision*, pages 311–326. Springer, 2016. 2

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 5

[9] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. *arXiv preprint arXiv:1808.00157*, 2018. 7

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2, 5

[11] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 3, 5

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, pages 5967–5976, 2017. 5

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proc. NIPS*, pages 2017–2025, 2015. 2, 4

[15] Nikolay Jetchev and Urs Bergmann. The conditional analogy GAN: swapping fashion articles on people images. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 2287–2292, 2017. 3, 5

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*, pages 694–711, 2016. 5

[17] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. *arXiv preprint arXiv:1804.07723*, 2018. 3

[18] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. CVPR*, pages 1096–1104, 2016. 2, 5

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 3

[20] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 405–415, 2017. 1, 2

[21] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[22] Natalia Neverova, Riza Alp Güler, and Iasonas Kokkinos. Dense pose transfer. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 3, 4, 5, 6, 7, 8

[23] Amit Raj, Patsorn Sangkloy, Huiwen Chang, Jingwan Lu, Duygu Ceylan, and James Hays. Swapnet: Garment transfer in single view images. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3, 5, 7

[24] Jimmy SJ Ren, Li Xu, Qiong Yan, and Wenxiu Sun. Shepard convolutional neural networks. In *Proc. NIPS*, pages 901–909, 2015. 3

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 5

[26] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 50–58, 2015. 2, 7

[27] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[28] Aliaksandr Siarohin, Enver Sangineto, Stphane Lathuilire, and Nicu Sebe. Deformable gans for pose-based human image generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 5, 6, 8

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[30] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[31] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 3

[32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-

video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 3, 5

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[34] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[35] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*. 3, 4, 5

[36] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[38] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. View synthesis by appearance flow. In *Proc. ECCV*, pages 286–301, 2016. 2