Coordinate-based Texture Inpainting for Pose-Guided Image Generation

Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, Victor Lempitsky

Samsung AI Center, Moscow
Skolkovo Institute of Science and Technology (Skoltech)
{a.grigorev, a.sevastopol}@partner.samsung.com, {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 using a fully-convolutional architecture with deformable convolutions. 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 applications in augmented reality and retail, is pose-guided image generation [21]. 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 together with warping and resampling models. The approach uses 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. Once such correspondence is estimated, 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

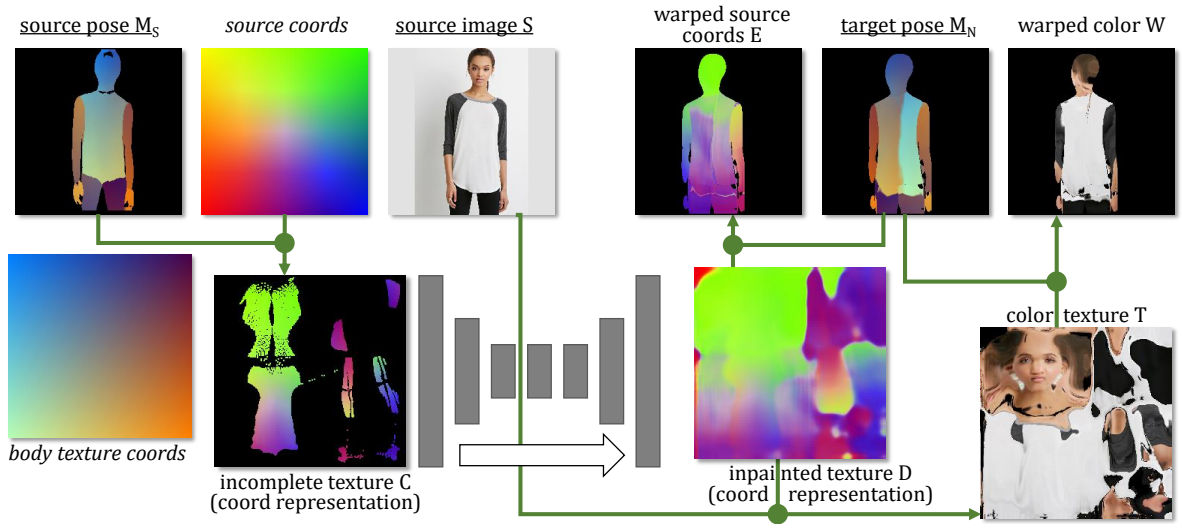

Figure 1. Coordinate-based texture inpainting. The scheme depicts the first (out of the two) part of our pipeline. Given the source pose (estimated by DensePose and converted to SMPL format), we rasterize the source coordinates of the known texture elements (e.g. by warping the source coordinate meshgrid). 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*).

according to the target pose (which similarly to [23] is defined by the DensePose [12] descriptor). The final stage of 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. Optionally, the input image can be 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 [29].

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 [19], 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 [27]. As a coda, we show that a small modification of our approach can successfully be used to perform garment transfer (virtual redress) with convincing results.

2. Related Work

Warping-based resynthesis. There is a strong interest in using deep convolutional networks for generating realistic images [11, 5]. 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 [8, 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) [15]. 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 [21, 22, 6] 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 [31, 29, 37, 23]. Our approach falls into this category and is most related to [23], as it utilizes the DensePose parameterization [12] within the network. We therefore compare extensively our results to [23].

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 [25], Sparsity-Invariant CNNs [32], networks with Partial Convolutions [18], networks with Gated Convolutions [36]. 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 [4] utilizes the main axial symmetry of a face by passing an image and its flipped copy to an inpainting ConvNet. The system in [37] 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 [13, 16, 33]. The only work that we are aware of that can handle similar amount of pose change is SwapNet [24], which however only present results at low resolution.

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 [28] and X2Face [35]. An older class of works going back to the seminal Blanz and Vetter morphable model [2] 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 [23] 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) 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.

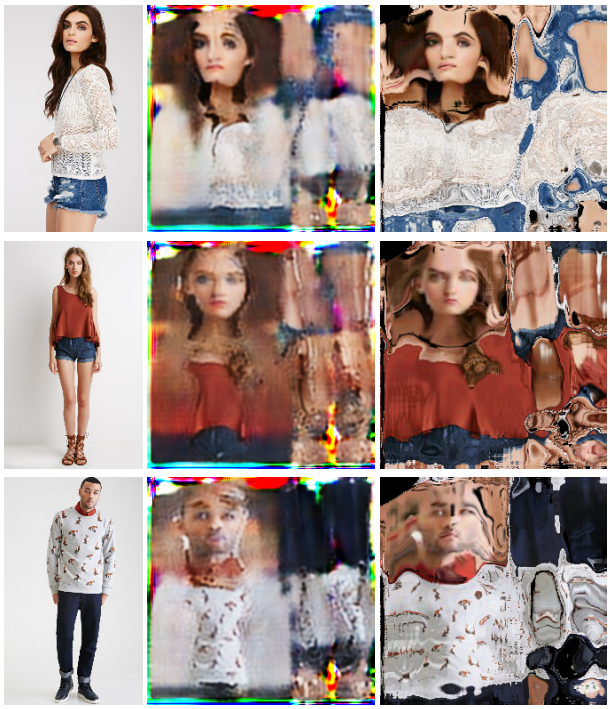

Figure 2. Body surface textures estimated using color-based inpainting (middle) and coordinate-based inpainting (right) for the inputs on the holdout set (left). Both inpaintings are generated using deep networks that were trained end-to-end with a variety of standard losses. Coordinate-based inpainting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

Texture map format and output conditioning. We use the SMPL texture format [20]. To make our approach comparable with [23], we estimate the mappings M_S and M_N based on DensePose [12], and then convert them to SMPL coordinates using a predefined mapping (provided with the DensePose). Thus, unlike [23], 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 [23], 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 [36] 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 [15] is used to sample the source image at fractional locations. More formally, the result of the backward warping is defined as:

$$T[u, v] = \sum_{x,y} S[x, y]\, \Delta(x, y, D^1[u, v], D^2[u, v])\,, \tag{3}$$

where the bilinear kernel Δ is defined as follows:

$$\Delta(k, l, m, n) = \max(1-|m-k|, 0)\, \max(1-|n-l|, 0)\,, \tag{4}$$

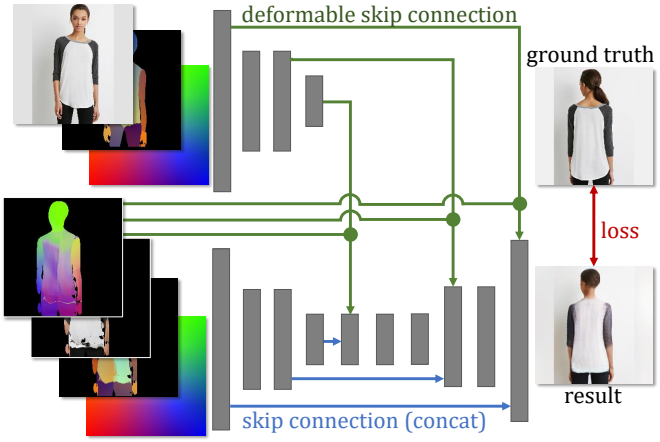

Figure 3. Final resynthesis. The second (of the two) part of our pipeline that takes the maps computed by the inpainting stage and map them to the final output image. Two separate encoders are used for maps aligned with the source pose (source pose, source image, meshgrid) and for maps aligned with the 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.

so that for each $[u, v]$ the summation in (3) is taken over $x = \{\lfloor D^1[u, v] \rfloor, \lceil D^1[u, v] \rceil\}$ and $y = \{\lfloor D^2[u, v] \rfloor, \lceil D^2[u, v] \rceil\}$.

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*), including [23, 4, 37]. 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.

New view resynthesis. Similarly to [23], 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:

$$W[x,y] = T\left[M_N^1[x,y], M_N^2[x,y]\right] , \quad (5)$$

$$E[x,y] = D\left[M_N^1[x,y], M_N^2[x,y]\right] , \quad (6)$$

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 (6) 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 [17] 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 [29] 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.

Loss functions. To measure the success of the reconstruction, we use the linear combination of the same loss functions that were used in [23]. Namely, we combine the reconstruction ℓ_1 loss, the perceptual loss [17] based on the VGG-19 network [30], the style loss [9] based on the same network, the adversarial loss [11] based on the patch GAN discriminator [14].

On top of that, we add an *identity loss* term that regularizes the training of the inpainting network f by penalizing 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.

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 pretrain the network f by minimizing the total loss between the warped images W and the target image N with the background pixels masked out. After the pretraining, we run end-to-end learning, where we minimize the loss between the predicted $\hat{N}$ and the ground truth new view N.

Garment transfer. A slight modification of our architecture allows it to perform garment transfer [13, 16, 33, 24]. 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) [19]. In general, we follow the same splits as used in [29, 23] that include 140,110 training and 8,670 test pairs, where clothing and models do not overlap between train and test sets.

Experimental details. For inpainting (both coordinate-based in our main method and rgb-based in ablation study) we employ hourglass architecture with gated convolutions from [36] which proved effective in image reconstruction tasks with large hidden areas.

Refinement network g is also a hourglass network consisted of encoder that maps images into $256 \times 64 \times 64$ feature tensors, followed by consecutive residual blocks and mirrored by a decoder. Encoder and decoder are also connected via three skip connections. Another encoder copies the structure of the first one. It takes image of a person in a source pose as an input and is connected to the decoder with deformable skip connections, same way as the first encoder with regular ones.

For ablations that do not use deformable skip connections, the second encoder is omitted, while for the baseline method that do not consider texture space at all, the same structure of refinement network is used with out inpainting one.

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

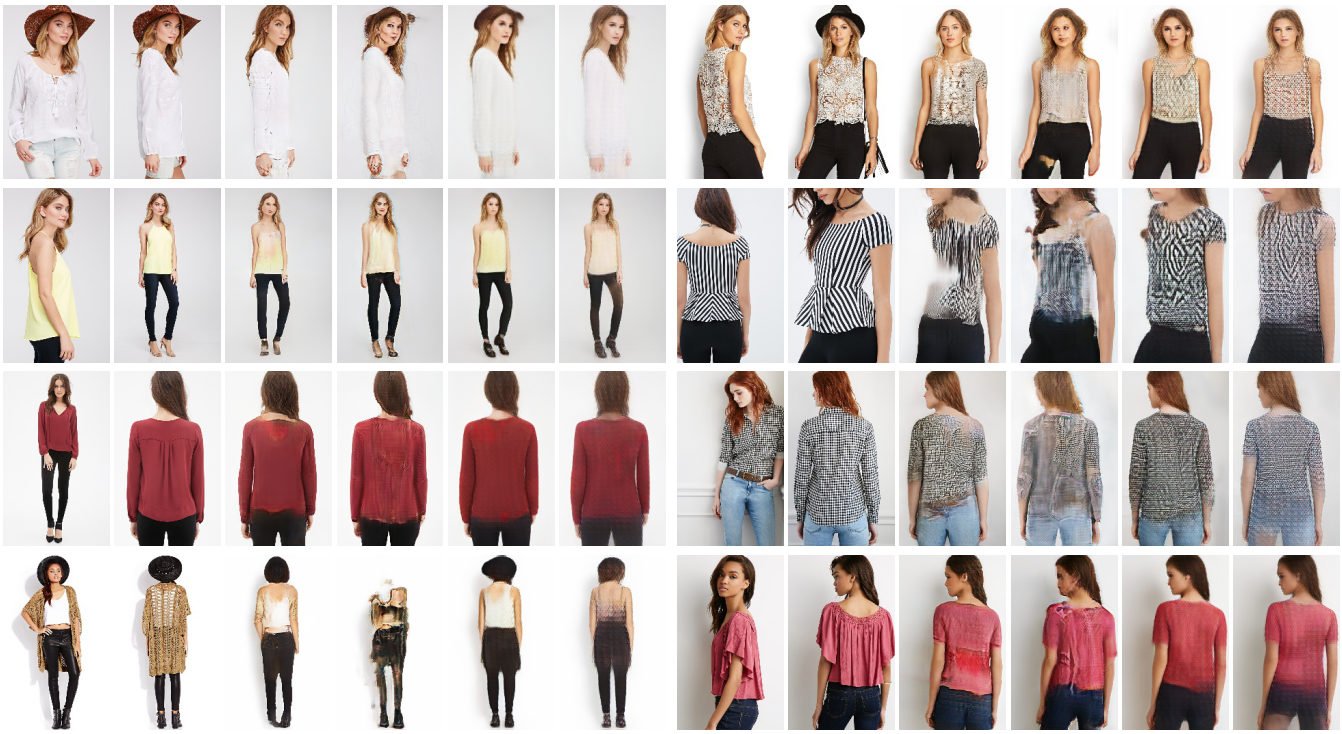

Figure 4. Side-by-side comparison with state-of-the-art (first eight samples from the test set). We show source image (SRC), ground truth in the target pose (GT), deformable GAN [29], our method conditioned on dense pose (Ours-D), and our method conditioned on keypoints (Ours-K). Consistently with the user study on a broader set, our method is more robust and has less artefacts than the state-of-the-art [29, 23] on this subset. *Electronic zoom-in recommended.*

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.

Since all ablations as well as our own method are optimized with the same loss terms with the same weights, the most direct comparison is in terms of these losses. We provide the values of each of the term in the loss function in Table 2. Our full approach is more successful than baselines in optimizing most of the loss terms and also the quality metrics, suggesting higher power of the coordinate-based inpainting.

Comparison with state-of-the-art. We compare the results of our method (full pipeline) with two state-of-the-art works [23, 29]. We again follow the previous work [23] closely using structural self-similarity (SSIM) metrics [34] to measure the structure preservation and the inception score (IS) [26] to measure image realism (Table 1). Both of these metrics are coarse proxies to the quantities they aim to measure (structure preservations and realism), and empirically we have found these metrics to have very low correlation with actual perceptual quality.

Therefore we perform additional 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 [23, 29] were kindly provided by the authors of [23]). In the user study, we have shown our results alongside of [23, 29] 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 people were involved in the user study. Each of them were to chose more realistic image in each of 80 pairs. In **71.8%** cases our reconstructions were preferred over those of [23], while against [29] our results were considered more realistic in **64.9%** cases (approximately 4000 pairs were compared in each of the two cases).

Keypoint-guided resynthesis. It can be argued that our method (as well as [23]) has unfair advantage over [29]

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.

	SSIM	MS-SSIM	IS
Ours	**0.835**	0.802	2.92
DPT [23]	0.785	**0.807**	**3.61**
DSC [29]	0.761	—	3.39

Table 1. Comparison with state-of-the-art. Our approach outperforms the other two in SSIM and underperforms in IS metrics. Ultimately, we found both metrics to be much less adequate judgements of visual fidelity than user judgements.

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 [3]-detected skeleton over a set of maps (one bone per map) and train a network to predict the DensePose [23] 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 redress). 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, hair, hands) to the input of the network. We use the DensePose coordinates to find the face and the hand parts, and we additionally used the pretrained network [10] to segment 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 and in [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 [7] — 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 [27] 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

Full body	$L_1 \downarrow$	$D_{patch} \uparrow$	$L_{cont} \downarrow$	$L_{style} \downarrow$	SSIM↑	MS-SSIM↑	IS↑
Ours-Full	**0.628**	0.226	**0.547**	**0.469**	**0.776**	0.779	2.66
Ours-NoDeform	0.633	**0.312**	0.563	0.55	**0.776**	**0.781**	2.72
RGB inpainting	0.634	0.262	0.553	0.491	0.771	**0.781**	**2.84**
No textures	0.664	0.250	0.581	0.494	0.735	0.754	2.58

Face	$L_1 \downarrow$	$D_{patch} \uparrow$	$L_{cont} \downarrow$	$L_{style} \downarrow$	SSIM↑	MS-SSIM↑	IS↑
Ours-Full	**0.880**	**0.0260**	**0.737**	0.0464	**0.613**	**0.764**	**1.834**
Ours-NoDeform	0.979	0.0257	0.756	**0.0454**	0.609	0.758	1.819
RGB inpainting	0.986	0.0258	0.779	0.050	0.601	0.755	1.756

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, and values of all employed losses relevant to image realism and reconstruction quality are reported. All losses correspond to the ones used in [23]: L_1 is the reconstruction loss, D_{patch} — patch discriminator loss, L_{cont} — content loss, L_{style} — perceptual loss. Arrows ↑, ↓ tell which value is better for the score — larger or smaller, respectively.

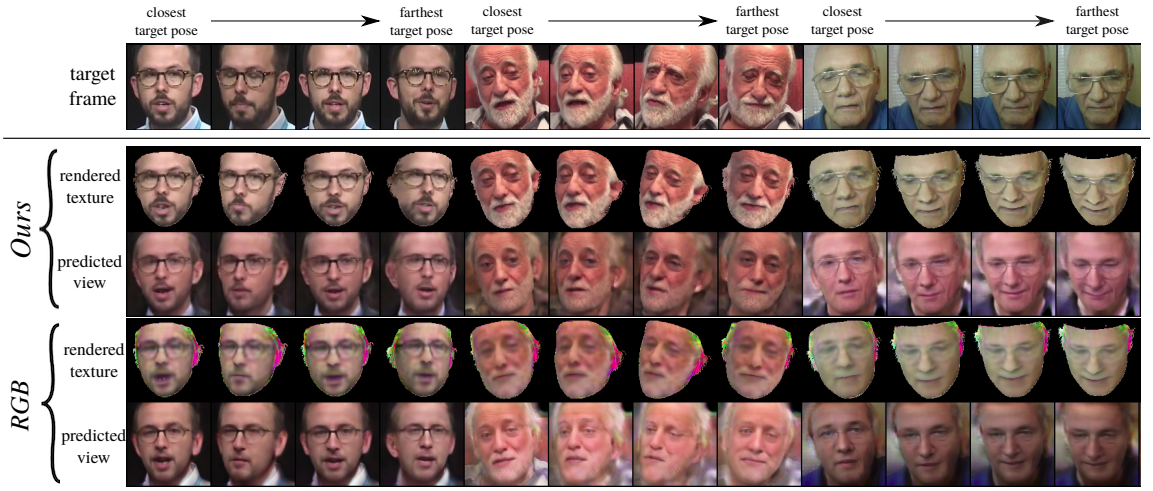

Figure 6. Predictions for several test samples. For each method, we take 3 random subjects, first video frame as a source frame and sample 4 target views according to the 4-quantiles of the pose difference distribution (see testing protocol in Subsection 4.2). For each subject, source frame is identical 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.*

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.

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). In addition to SSIM, MS-SSIM and IS scores, we report values of several losses used during training which represent the similarity between predicted and target image. 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. Discussion

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 [29, 23]. 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 [29] may be more appropriate under limited changes.

References

[1] Anonymous. Supplementary material. 7

[2] V. Blanz and T. 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

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 7

[4] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. Uv-gan: adversarial facial uv map completion for pose-invariant face recognition. In *Proc. CVPR*, pages 7093–7102, 2018. 3, 4

[5] A. Dosovitskiy, J. T. Springenberg, M. Tatarchenko, and T. 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

[6] P. Esser, E. Sutter, and B. 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

[7] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 7

[8] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European Conference on Computer Vision*, pages 311–326. Springer, 2016. 2

[9] L. A. Gatys, A. S. Ecker, and M. 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

[10] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin. Instance-level human parsing via part grouping network. *arXiv preprint arXiv:1808.00157*, 2018. 7

[11] 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*, pages 2672–2680, 2014. 2, 5

[12] R. A. Güler, N. Neverova, and I. Kokkinos. DensePose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[13] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 3, 5

[14] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, pages 5967–5976, 2017. 5

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Proc. NIPS*, pages 2017–2025, 2015. 2, 4

[16] N. Jetchev and U. 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

[17] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*, pages 694–711, 2016. 5

[18] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. *arXiv preprint arXiv:1804.07723*, 2018. 3

[19] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. CVPR*, pages 1096–1104, 2016. 2, 5

[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 3

[21] 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*, pages 405–415, 2017. 1, 2

[22] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[23] N. Neverova, R. A. Güler, and I. Kokkinos. Dense pose transfer. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 3, 4, 5, 6, 7, 8

[24] A. Raj, P. Sangkloy, H. Chang, J. Lu, D. Ceylan, and J. Hays. Swapnet: Garment transfer in single view images. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3, 5

[25] J. S. Ren, L. Xu, Q. Yan, and W. Sun. Shepard convolutional neural networks. In *Proc. NIPS*, pages 901–909, 2015. 3

[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 6

[27] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. 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

[28] Z. Shu, M. Sahasrabudhe, R. Alp Guler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[29] A. Siarohin, E. Sangineto, S. Lathuilière, and N. 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, 7, 8, 9

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

[31] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[32] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 3

[33] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 3, 5

[34] 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*, 13(4):600–612, 2004. 6

[35] O. Wiles, A. Sophia Koepke, and A. 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

[36] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*. 3, 4, 5

[37] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4

[38] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *Proc. ECCV*, pages 286–301, 2016. 2