Human Appearance Transfer

Mihai Zanfir[2*] Alin-Ionut Popa[2*] Andrei Zanfir[2] Cristian Sminchisescu[1,2]
{mihai.zanfir, alin.popa, andrei.zanfir}@imar.ro cristian.sminchisescu@math.lth.se
[1]Department of Mathematics, Faculty of Engineering, Lund University
[2]Institute of Mathematics of the Romanian Academy

Abstract

*We propose an automatic person-to-person appearance transfer model based on explicit parametric 3d human representations and learned, constrained deep translation network architectures for photographic image synthesis. Given a single source image and a single target image, each corresponding to different human subjects, wearing different clothing and in different poses, our goal is to photo-realistically transfer the appearance from the source image onto the target image while preserving the target shape and clothing segmentation layout. Our solution to this new problem is formulated in terms of a computational pipeline that combines (1) 3d human pose and body shape estimation from monocular images, (2) identifying 3d surface colors elements (mesh triangles) visible in both images, that can be transferred directly using barycentric procedures, and (3) predicting surface appearance missing in the first image but visible in the second one using deep learning-based image synthesis techniques. Our model achieves promising results as supported by a perceptual user study where the participants rated around 65% of our results as good, very good or perfect, as well in automated tests (Inception scores and a Faster-RCNN human detector responding very similarly to real and model generated images). We further show how the proposed architecture can be profiled to automatically generate images of a person dressed with different clothing transferred from a person in another image, opening paths for applications in entertainment and photo-editing (*e.g. *embodying and posing as friends or famous actors), the fashion industry, or affordable online shopping of clothing.*

1. Introduction

People are of central interest in images and video, so understanding and capturing their pose and appearance from visual data is critically important. While problems like detection or 2d pose estimation have received considerable attention and witnessed significant progress recently, appearance modeling has been less explored comparatively, especially for bodies and clothing, in contrast to faces. One setback is that people are extremely sensitive to invalid human appearance variations and immediately spot them. This is to a large extent true for faces, as people are sharply tuned to fine social signals expressed as subtle facial expressions, but also stands true for human body poses, shapes and clothing. This makes it difficult to capture and possibly re-synthesize human appearance in ways that pass the high bar of human perception. While the realistic 3d human shape and appearance generation, including clothing, has been a long standing goal in computer graphics, with impressive studio results that occasionally pass the Turing test, these usually require complex models with sophisticated layering, manual interaction, and many cameras, which makes them difficult to use at large scale. For this purpose, flexible methods, that can be learned from data and can synthesize realistic human appearance are of obvious value. Arguably, even more im-

Figure 1. **Bidirectional transfer automatically produced by our method.** We transfer from one image (first column) to the second (second column) and vice-versa, with automatically generated images shown on the third and fourth columns.

*Authors contributed equally

portant would be methods that can be controlled by image evidence in some way. For instance one may not just aim to generate plausible human shape and appearance in isolation – hard as this may be – but also condition on specific elements of pose and appearance in a given image in order to synthesize new ones based on it.

In this paper we formulate a new problem called *human appearance transfer*. Given a single source and a single target image of a person, each with different appearance, possibly different body shape and pose, the goal is to transfer the appearance of the person in the first image into the one of the person of the target image while preserving the target clothing and body layout. The problem is challenging as people are in different poses and may have different body shapes. A purely image warping or image to image translation approach would not easily generalize due to the large number of degrees of freedom involved in the transformation, *e.g*. the effect of articulation, depth and body shape on appearance. We provide a first solution that relies on fitting state-of-the-art 3d human pose and body models to both the source and the target images, transferring appearance using barycentric methods for commonly visible vertices, and learning to color the remaining ones using deep image synthesis techniques with appropriately structured 2d and 3d inputs. Example images, perceptual user studies, Inception scores [26], and the response of a state-of-the-art person detector confirm that the generated images of humans are perceptually plausible.

2. Related Work

Our work relies on 2d human detection and body part labeling [2, 24, 9], 3d pose estimation [24, 1, 29], parametric 3d human shape modeling [5, 30, 20, 18], procedures devoted to the semantic segmentation of clothing [28, 6, 17, 8], as well as image translation and synthesis methods [11, 3, 21, 14, 34, 3, 31].

Modeling the human appearance is a vast topic that has been approached on several fronts. One is through modifications of real images [10, 22], although the results are not entirely realistic. Computer graphics pipelines are also used, either in a mixed reality setting - where a moderately realistic graphics model is rendered in a real scene in a geometrically correct manner [10] – or *e.g*. by fitting a SCAPE model to real images [4]. In the former, the graphics character is still not photo-realistic; in the latter, clothing geometry is lacking. Detailed, accurate human shape estimation from clothed 3d scan sequences [33] can produce very good results but the acquisition setup is considerably more involved. Models to realistically capture complex human appearance including clothing in a laboratory setup, based on multiple cameras and relatively simple backgrounds appear in [15]. Procedures directed to the realistic acquisition of clothing exist [23, 33], but rely on an existing set of 3d models of garments and a 3d scanning device.

The methodology reviewed in the previous paragraph achieves excellent results under the application constraints it was designed for. However, some requires manual interaction, multiple cameras, simple backgrounds, specialized scanners, or complex modeling setups. In contrast, we aim at automatic appearance modeling in situations where one has no control on the acquisition setup and is given a minimal number of images (one or two). The idea is to exploit precise, but inherently limited in coverage, geometric estimation methods for the human pose and shape, and complement them with learning techniques, in order to achieve photo-realistic appearance transfer for specific images. There is relatively little research focusing on human appearance generation based on a combination of geometric and learning methods. One notable exception is the recent work by [13] which is able to generate realistic images of people given their silhouette, pose and clothing segmentation. The method relies on a variational auto-encoder [12] and a GAN [7, 11] for realistic image generation. However, it is not obvious how this model would perform inference for the appearance given an image, or how can it condition on a particular appearance and photo-realistically transfer it to a new pose. The human appearance transfer between two monocular images falls out of the domain of applicability of models like [13], and is the new problem defined and confronted in this research.

3. Human Appearance Transfer

Given a pair of RGB images – source and target, denoted by I_s and I_t, each containing a person –, the main objective of our work is to transfer the appearance of the person from I_s into the body configuration of the person from I_t, resulting in a new image $I_{s \Rightarrow t}$.[1] Our proposed pipeline is shown in fig. 2 and details are given in the next sections.

3.1. 3D Human Pose and Body Shape Fitting

Human Detection & Body Parts Segmentation. To detect each person and infer critical semantic and geometric information, each image is fed through the Deep Multitask Human Sensing (DMHS) network [24], a state-of-the-art predictor for body part labeling (semantic segmentation) and 3d pose estimation. DMHS is a multi-stage architecture, in which each stage refines the output from the previous stage, producing a tuple (J, B, R), where $J \in \mathbb{R}^{18\times 2}$ is the set of 2d body joint configurations, $B \in \mathbb{R}^{w\times h\times 24}$ is the body part labeling map, and $R \in \mathbb{R}^{17\times 3}$ is the 3d body joint configuration of the person detected in an image.

3d Body Shape Fitting. We use the prediction of DMHS with the fitting method of [32] in order to estimate the hu-

[1] The procedure is symmetric, as we can transfer in both directions.

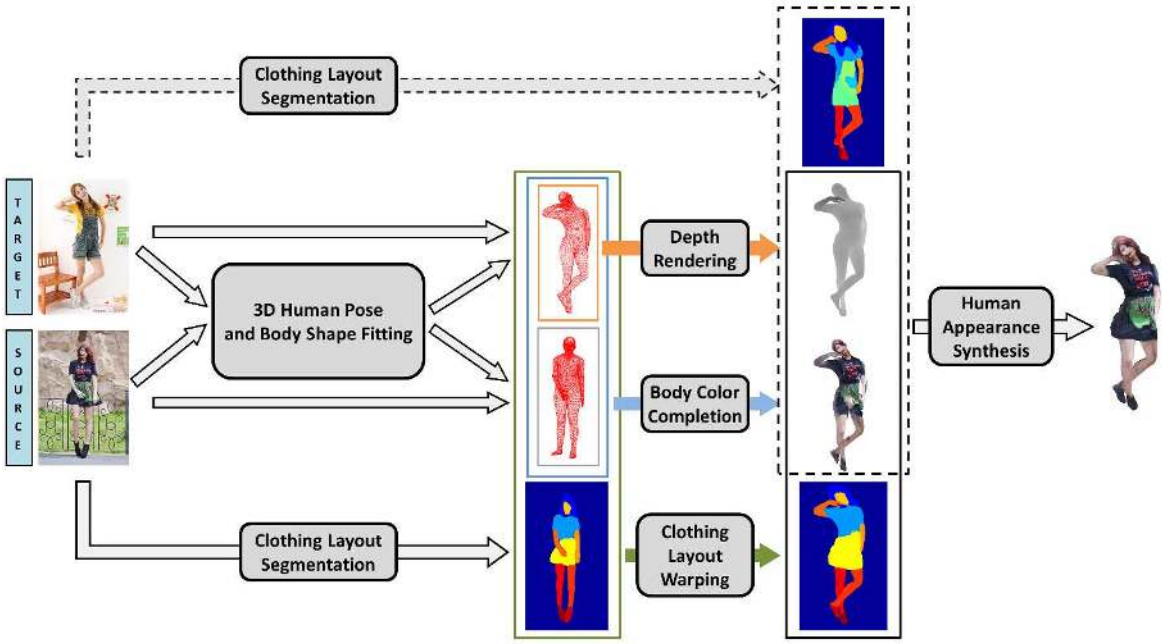

Figure 2. **Human appearance transfer pipeline.** Given only a single source and a single target image, each containing a person, with different appearance and clothing, and in different poses, our goal is to photo-realistically transfer the appearance from the source image onto the target image while preserving the target shape and clothing segmentation layout. The problem is formulated in terms of a computational pipeline that combines *(i)* 3d human pose and body shape fitting from monocular images shown in fig. 3, together with *(ii)* identifying 3d surface colors corresponding to mesh triangles visible in both images, that can be transferred directly using barycentric procedures, *(iii)* predicting surface appearance missing in the target image but visible in the source one using deep learning image synthesis techniques – these will be combined using the Body Color Completion Module detailed in fig. 5. The last step, *(iv)*, takes the previous output together with the clothing layout of the source image warped on the target image (Clothing Layout Warping) and synthesizes the final output. If the clothing source layout is similar to the target, we bypass the warping step and use the target clothing layout instead.

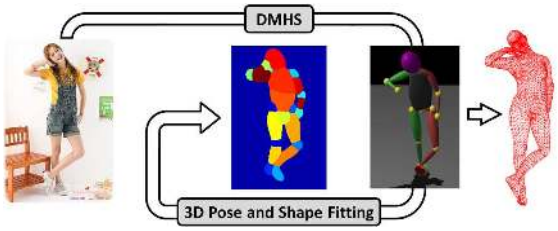

Figure 3. **3d human pose and body shape fitting module.** Given an image of a person, 2d joint locations, semantic body part segmentation and 3d pose are obtained using a multitask deep neural network model (DMHS). 3d estimates are then refined by non-linear optimization, in order to align an articulated human body mesh (SMPL) with the semantic segmentation layout from DMHS. The model produces image fits which tightly cover the body layout and are essential for good quality appearance transfer.

man 3d body shape and pose from an image. The representation is based on the 3d SMPL body model [18]. A commonly used pipeline for fitting the model [1] relies on minimizing a sparse cost error over detected 2d human joint locations. However, the method of [32] utilizes all information available in the tuple (J, B, R) and the image. The 3d pose is initialized using R and refined so that each body part of the model aligns with the corresponding semantic segmentation labels in the image, based on DMHS estimates J and B (fig. 3). Please notice the difference between our body shape fitting procedure and the one of [1], illustrated in fig. 4. The head and arm orientations of our estimate are closer to the perceived one, due to a superior DMHS initialization (as opposed to a canonical T-pose) and the use of dense body parts semantic image segmentation labels during fitting.

The estimated 3d body model consists of a fixed number of $N_v = 6890$ vertices and a set of $N_f = 13776$ faces, $F \in \mathbb{N}^{N_f \times 3}$, that forms the triangle mesh. We define the fitted 3d body shape model as a tuple $S = (C, D, M)$, where $C \in \mathbb{R}^{N_v \times 3}$ encodes the RGB color at each vertex position, $D \in \mathbb{R}^{w \times h}$ encodes the disparity map of the fitted 3d mesh with respect to the image and $M \in \{0, 1\}^{N_v}$ encodes the visibility of each vertex. Given the source and target images I_s and I_t, we obtain human pose and shape estimates as mesh structures S_s and S_t, respectively.

Figure 4. **Comparison of 3d human pose and shape estimation results** of our employed method (green), that combines the 2d and 3d predictions of a deep multitask neural network with pose refinement, with [1] (red). As our model, illustrated in fig. 3, is initialized using an image-sensitive 3d pose predictor (as opposed to fixed initialization) and is fitted to the full semantic person layout (the complete semantic segmentation of the person into body parts, as opposed to only a sparse set of human body joints), it tends to better cope with the angle and the orientation of the head and arms, as well as the body proportions.

3.2. Body Color Completion

We are first interested in estimating the pixel colors for the projected visible surface of S_t, denoted as $I_{s\to t}$, using the pixel colors on the projected visible surface of S_s.

Barycentric transfer. We begin by defining the common set of visible vertices $\Lambda_{s\wedge t} = \{i | M_s(i) = 1 \wedge M_t(i) = 1, 1 \leq i \leq N_v\}$ and select the corresponding mesh faces $F(\Lambda_{s\wedge t})$. For each face $f \in F(\Lambda_{s\wedge t})$, we project it on I_s and $I_{s\to t}$. For each pixel location in the projection of f on I_s, we find its corresponding pixel location in the projection on $I_{s\to t}$ using barycentric triangle coordinates. Finally, we copy the color information from one location to another.

Vertex Color Completion Network. The remaining set of visible vertices in S_t, $\Lambda_{t\setminus s} = \{i | M_t(i) = 1 \wedge M_s(i) = 0, 1 \leq i \leq N_v\}$ needs to be colored. We rely on learning the implicit correlations among various body parts in order to propagate appearance information from the already colored vertex set C_s to $\Lambda_{t\setminus s}$. Such correlations, effectively forms of pattern completion, are learned automatically from training data using a neural network.

Learning for Mesh Color Completion. We are given as inputs the color set C_s, the visibility mask $M_s \in \{0,1\}^{N_v \times N_v}$, and a binary mask that encodes the vertices we wish to color $M_{\Lambda_{t\setminus s}} \in \{0,1\}^{N_v \times N_v}$, *i.e.* visibility values are replicated along columns. The output is represented by the predicted colors $C^\star_{\Lambda_{t\setminus s}} \in \mathbb{R}^{3\times N_v}$. We define two weight matrices $W_1 \in \mathbb{R}^{N_v \times N_v}$ and $W_2 \in \mathbb{R}^{N_v \times N_v}$. Our network optimizes over these two matrices with the loss L defined as the Euclidean distance between the prediction $C^\star_{\Lambda_{t\setminus s}}$ and the ground-truth colors $C_{\Lambda_{t\setminus s}}$:

$$W_1^\star = \text{softmax}(W_1 \odot M_s) \quad (1)$$

$$W_2^\star = \text{softmax}(W_2 \odot (M_s \vee M_{\Lambda_{t\setminus s}})) \quad (2)$$

$$C^\star_{\Lambda_{t\setminus s}} = (C_s^T \times W_1^\star) \times W_2^\star \quad (3)$$

$$L(W_1, W_2) = \|C_{\Lambda_{t\setminus s}} - C^\star_{\Lambda_{t\setminus s}}\| \quad (4)$$

where the softmax function is applied column-wise. Intuitively, any visible target vertex, without color, will have it assigned to the weighted mean (softmax function) of all the available colored vertex set, with weights encoded in matrices W_1 and W_2. We interpolate the predicted vertex colors $C^\star_{\Lambda_{t\setminus s}}$ from the learned model over the corresponding mesh faces $F(\Lambda_{t\setminus s})$, project the faces, and obtain the missing regions in $I_{s\to t}$.

Generating Training Samples. The training data consists of inputs, each being a subset of colored vertices from a mesh, and outputs that represent different subsets of vertices from the same mesh. In practice, given any monocular image, once we fit the 3d model, we can generate any possible input-output split over the visible vertices. However, our inputs tend to be structured, consisting of subsets of vertices seen in the source and target mesh as well as their difference set. To ensure a similar distribution, we take the inputs and outputs to be sets of visible vertices in the intersection of the source and target mesh (we assume intersection is non-trivial), and choose outputs in their difference sets, respectively. Two training examples can thus be generated, symmetrically, for every pair of images of people, with different appearance and in a different pose.

The drawback of this procedure is that, at training time, the network has to predict colors from a smaller intersection set of colored vertices (*i.e.* $\Lambda_{s\wedge t}$), whereas at test time, it can use the fully colored set of vertices from the source mesh C_s .

3.3. Clothing Layout Warping

We use the model of [6] to estimate the clothing layout for target and source images, L_t and L_s, defined over a set of 20 clothing labels. Given the clothing layout source L_s, we want to transform it into the pose of the target image, $L_{s\Rightarrow t}$. We start by collecting clothing label information for each visible vertex in the source mesh S_s. We propagate the labeling on the entire mesh by using a geodesic nearest-neighbor approach. For labels not on the source

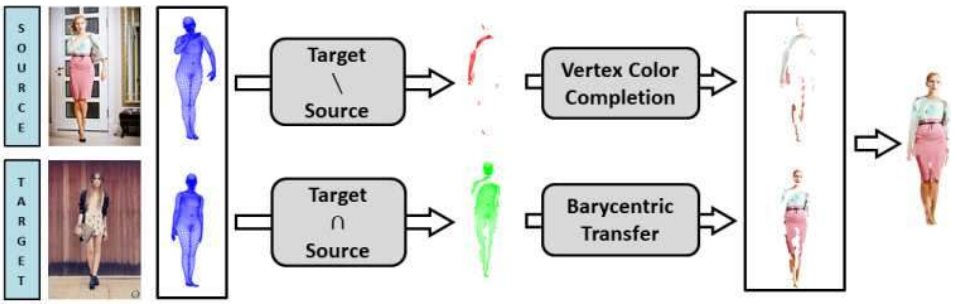

Figure 5. **Body Color Completion Module.** Given a Source (S) and a Target (T) image, with 3d human pose and shape fitted automatically and represented using 3d meshes, we compute two sets of mesh vertices: ones that are visible in both images (T ∩ S), and ones that are only visible in the target T\S. We propagate (copy) the color of the intersection set using barycentric transfer and predict the difference set using a vertex color completion network. The network is trained to predict a subset of vertex colors from other subsets of vertex colors. Training data for this network can be readily available by sampling any 'cut' through the set of visible vertices obtained by fitting a 3d model to a single image. However, in practice we make training more typical by using targets that come from S\T in pairs of images.

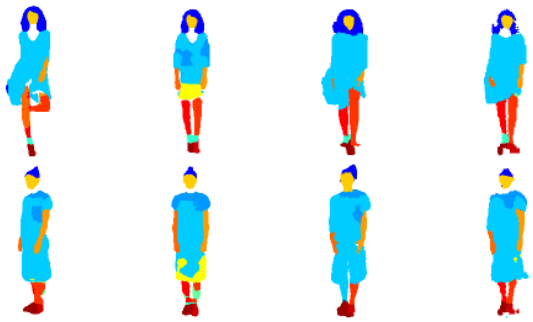

Figure 6. **Clothing layout warping.** From left to right: source image clothing layout L_s, target image clothing layout L_t, input of the clothing layout synthesis network $L_{s\to t}$, final output of the warping network $L_{s\Rightarrow t}$

Figure 7. **Human appearance transfer with clothing layout warping.** From left to right: source image, target image, clothing warped $L_{s\Rightarrow t}$ and RGB data, $I_{s\Rightarrow t}$, generated using the human appearance synthesis module

mesh, we collect the nearest-neighbour source vertex projections, vote on a displacement and translate the labeling accordingly. Thus, we obtain a rough estimate of $L_{s\Rightarrow t}$, which will denote by $L_{s\to t}$. We gather additional image data from the web, consisting of source-target image pairs depicting the same person wearing the same clothes, but in different poses. On this dataset of $\sim 1,500$ training pairs, we train an image to image translation network which outputs $L_{s\Rightarrow t}$ given $L_{s\to t}$ and the disparity map D_t.

3.4. Human Appearance Synthesis

The previously described prediction $I_{s\to t}$ captures the appearance transfer only as covered by our 3d body models. Hence, clothing layers (*e.g.* skirt, jacket) or hair which fall outside the coverage of the human body model are not transferred during the process. To achieve a higher perceptual quality for the generated image, we further refine our prediction using a Human Appearance Synthesis (HAS) network adapted based on ideas in [3]. This method performs multi-resolution refinement and was originally used in synthesizing photographic images conditioned on semantic segmentation layouts. Instead, we train the network to predict an image I_t given three types of inputs: a predicted semantic layout of clothing L_t or $L_{s\Rightarrow t}$, the disparity map D_t, and $I_{s\to t}$. The output of this HAS network, $I_{s\Rightarrow t}$, represents our final, refined result.

4. Experiments

For all of our experiments we use the Chictopia10k dataset [16]. The images in this dataset depict different people, under both full and partial viewing, captured frontally. The high variability in color, clothing, illumination and pose makes this dataset suitable for our task. There are $17,706$ images available together with additional ground truth clothing segmentations. We do not use the clothing labels provided, but only the figure-ground segmentation such that we can generate training images cropped on the human silhouette.

We split the data in two subsets: $15,404$ images for train-

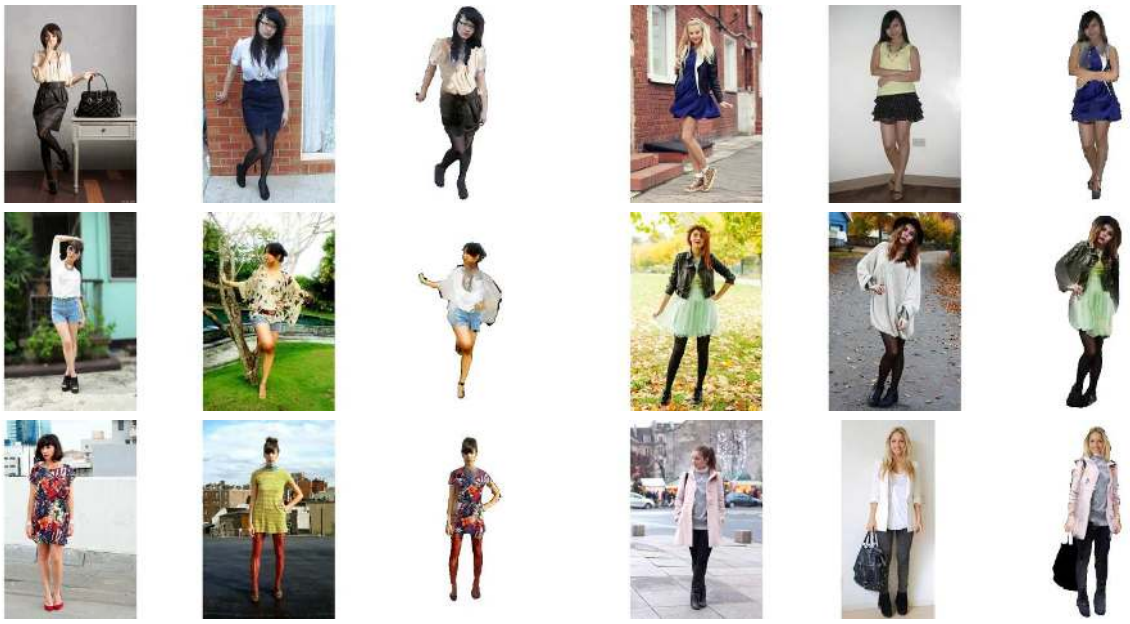

Figure 8. **Sample appearance transfer results from I_s to I_t**, where we preserved the identity of the person from I_t. From left to right: source image, I_s, target image, I_t, generated image with person from I_t using the clothing of person from I_s.

ing and $2,302$ images for testing. We additionally prune some of the images based on the quality of body shape fitting, This is done by applying a soft threshold on the intersection over union (IoU) between the projection of the fitted body model and the foreground mask of the person. For each image in the dataset, we randomly select two other images from its corresponding subset (*i.e.* train or test) to construct image pairs. In the end, we use $28,808$ training pairs and $4,080$ testing pairs.

Appearance Transfer. Results of our Body Color Completion module are shown in fig. 9. Sample results of our Human Appearance Synthesis module are also given in fig. 10. Although we show transfer results in one direction, our method is symmetrical, so we obtain results of similar quality both ways, as shown in fig. 1.

Impact of Components and Failure Modes. Our Human Appearance Synthesis network receives as inputs $I_{s\to t}$, the depth map and the clothing segmentation of the target image. To evaluate the contribution of each of these inputs in the visual quality of the output, we train two additional Human Appearance Synthesis networks under similar conditions, but with different input data: one without the depth map, and the other without both the depth map and the clothing segmentation. In fig. 11, we provide visual results for all three networks. We observe that the best quality is obtained when using the complete network.

Errors occur in our pipeline when the clothing segmentation fails or the 3d body shape fitting does not yield good alignment with the person in the image. Examples are shown in fig. 12.

Figure 9. **Sample results for the body color completion module.** From left to right: source image, target image, appearance data copied using visible mesh triangles on both meshes (source and target), appearance data completed using vertex color completion.

4.1. Identity Preserving Appearance Transfer

We also implement a variation of our model in order to preserve the identity of the target subject during appearance

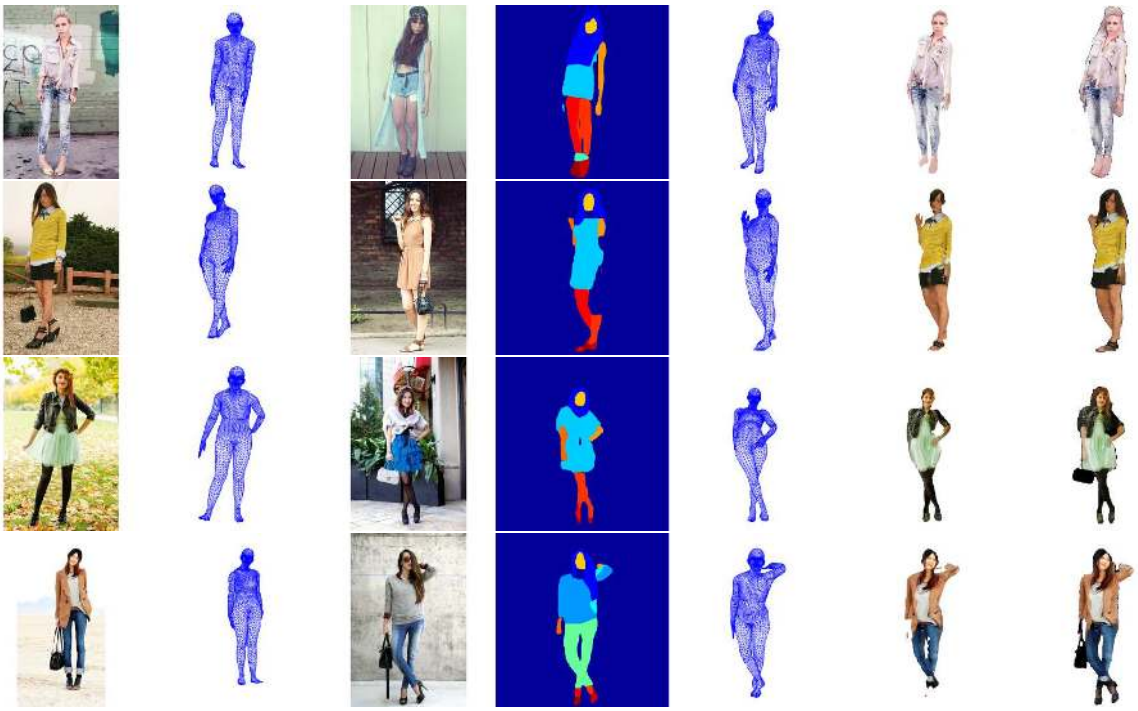

Figure 10. **Sample results for the human appearance transfer pipeline.** From left to right: source image with corresponding fitted 3d body model, target image with its clothing layout and fitted 3d body model, RGB data generated using the Body Color Completion module (*i.e.* $I_{s\rightarrow t}$), RGB data generated using the Human Appearance Synthesis module (*i.e.* $I_{s\Rightarrow t}$).

Figure 11. **Impact of the different model components** in the effectiveness of human appearance transfer. In the third column we show our complete method, in the fourth column we do not use the depth map and in the fifth column we neither use the depth map nor the clothing segmentation. For the first example (first row), notice the difference in quality for the girl's skirt (*i.e.* for the complete network the skirt is fully defined, for the network without depth the skirt becomes blurry and for the network without both depth and clothing, the skirt is nearly in-existent and body detail is blurred). For the second example (second row) notice the improved quality of the generated hair.

transfer. To do so, we redefine the sets of vertices used in the Body Color Completion module. We start by identifying the skin and clothing image pixels of the target and source images by using the set of labels provided by the clothing segmentation model. We place the set of labels defined by hair, face, arms and legs in the skin category and the remaining ones in the clothing category. Then we assign a category (*i.e.* skin/cloth) for each vertex in the body models by inspecting the clothing labeling under their projections in the image. We fix the colors for the pixels/vertices categorized as skin in the target, and perform barycentric transfer only for the intersection of source and target vertices categorized as clothing. The colors for the remaining vertices are predicted as before by our Vertex Color Completion network. Sample results are shown in fig. 8.

4.2. Perceptual User Study and Automated Tests

We perform a perceptual study by asking 20 human subjects to evaluate our results. We present each one with 100 results in the form of source image (I_s), target image (I_t) and our automatically generated appearance transfer image $I_{s\Rightarrow t}$. We ask subjects to evaluate the appearance transfer quality of $I_{s\Rightarrow t}$, by assigning it one of the following scores: very poor (1), poor (2), good (3), very good (4), perfect (5). Finally, we quantified their scores in the form of a normalized histogram, shown in fig. 13 (bottom). The mean score is 2.9, with standard deviation 0.96, suggesting that our transfer is reasonable on average.

We compare our method against recent work [19, 27] independently addressing the problem of pose conditioned

Figure 12. **Several failure modes of our method.** From left to right: source image I_s with corresponding mesh, target image I_t with corresponding mesh and the generated $I_{s \Rightarrow t}$. In the first example the body shape fitting in I_s misses the model's left hand, causing the final output to contain 3 hands, one resulting from barycentric transfer. In the second example, the model has problems dealing with the coat of the person from the source image. In the third example, the issue is caused by the small number of common vertices between the two body shape models, producing appearance mixing and considerable transfer ambiguity.

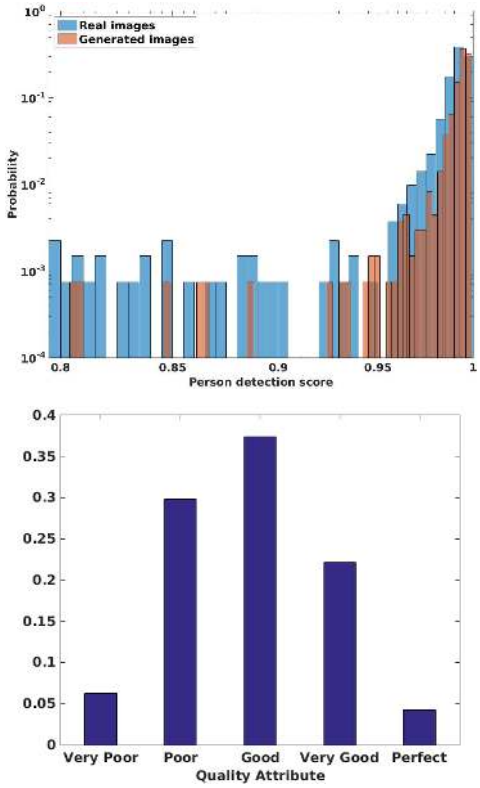

Figure 13. (Top) Distributions of person detection scores given by running Faster R-CNN [25] over two sets of real and automatically generated images, respectively. (Bottom) We conducted a quality survey for a set of 100 automatically generated images by our method (randomly selected), where people were asked to assign a score from 1 (very poor quality) to 5 (perfect quality).

human image generation. Such methods rely solely on information in the image without explicitly inferring 3d body pose. Set aside significant methodological differences, we are additionally able to perform identity-preserving transfer (fig. 8). Our results are more visually pleasing and superior in terms of Inception Scores [26], which are 3.09 [19], 3.35 [27] and **4.13** (Ours).

In order to understand possible difference in terms of image statistics, we also perform an automated test using a state-of-the-art human detector, Faster R-CNN [25]. We compute the human detection scores on two sets containing $2,000$ generated and real images, respectively. In fig. 13 (top) we observe that the two distributions of detection scores are similar, with a dominant mode around value 0.99.

5. Conclusions

Modeling and synthesizing human appearance is difficult due to variability in human body proportions, shape, clothing and poses. However, models that can realistically synthesize complete images of humans under a degree of control (conditioning) on an input image appearance or pose, could be valuable for entertainment, photo-editing, or affordable online shopping of clothing. In this context, we define a new problem entitled *human appearance transfer* where given two images, source and target, of different people with different poses and clothing, we learn to transfer the appearance of the source person on the body layout of the target person. Our solution relies on state-of-the-art 3d human pose and shape estimation based on deep multitask neural networks and parametric human shape modeling, combined with deep photographic synthesis networks controlled by appropriate 2d and 3d inputs. Our image results, backed-up by a perceptual user study, Inception scores, and the response of a state-of-the-art human person detector indicate that the proposed model can automatically generate images of humans of good perceptual quality, and with similar statistics as real human images. We also show how the model can be modified to realistically 'dress' a person shown in one image with clothing captured from a person in another image.

Acknowledgments: This work was supported in part by the European Research Council Consolidator grant SEED, CNCS-UEFISCDI PN-III-P4-ID-PCE-2016-0535, the EU Horizon 2020 Grant No. 688835 DE-ENIGMA, and SSF.

References

[1] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

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

[3] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, October 2017.

[4] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 479–488. IEEE, 2016.

[5] R. Goldenthal, D. Harmon, R. Fattal, M. Bercovier, and E. Grinspun. Efficient simulation of inextensible cloth. *ACM Transactions on Graphics (TOG)*, 26(3):49, 2007.

[6] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, July 2017.

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

[8] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, December 2015.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.

[10] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI*, 2014.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[12] T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015.

[13] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model of people in clothing. In *ICCV*, 2017.

[14] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[15] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *ICCV*, 2017.

[16] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015.

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

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *SIGGRAPH*, 34(6):248, 2015.

[19] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *NIPS*, 2017.

[20] R. Narain, A. Samii, and J. F. O'Brien. Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)*, 31(6):152, 2012.

[21] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, 2017.

[22] L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, June 2012.

[23] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally.

[24] A. Popa, M. Zanfir, and C. Sminchisescu. Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. In *CVPR*, July 2017.

[25] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016.

[27] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018.

[28] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. A high performance crf model for clothes parsing. In *ACCV*, 2014.

[29] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[30] F. Xu, Y. Liu, C. Stoll, J. Tompkin, G. Bharaj, Q. Dai, H.-P. Seidel, J. Kautz, and C. Theobalt. Video-based characters: Creating new human performances from a multi-view video database. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH, pages 32:1–32:10, New York, NY, USA, 2011. ACM.

[31] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *CVPR*, 2017.

[32] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes – The Importance of Multiple Scene Constraints. In *CVPR*, 2018.

[33] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, 2017.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.