Reposing Humans by Warping 3D Features

Markus Knoche István Sárándi Bastian Leibe
RWTH Aachen University, Germany
{knoche,sarandi,leibe}@vision.rwth-aachen.de

Abstract

We address the problem of reposing an image of a human into any desired novel pose. This conditional image-generation task requires reasoning about the 3D structure of the human, including self-occluded body parts. Most prior works are either based on 2D representations or require fitting and manipulating an explicit 3D body mesh. Based on the recent success in deep learning-based volumetric representations, we propose to implicitly learn a dense feature volume from human images, which lends itself to simple and intuitive manipulation through explicit geometric warping. Once the latent feature volume is warped according to the desired pose change, the volume is mapped back to RGB space by a convolutional decoder.

Our state-of-the-art results on the DeepFashion and the iPER benchmarks indicate that dense volumetric human representations are worth investigating in more detail.

1. Introduction

The ability to freely change a human's pose in an image opens the door to a variety of applications, from generating large crowds or performing stunts in filmmaking to data augmentation for human-centric computer vision tasks. State-of-the-art approaches to this problem employ fully-convolutional neural networks. However, convolutional features tend to strongly depend on the input pixels near the center of the receptive field and CNNs often fail to move information over large distances. This makes person reposing difficult when input and target pose differ strongly, as the appearance information of the various body parts needs to move to different places compared to their position in the input image. To tackle this, many recent approaches apply some form of explicit transformations. Some warp 2D features such that they become aligned to the target pose, which is also specified in 2D [2, 32, 9, 3, 23, 6]. We argue that this is insufficient to capture 3D human pose changes.

Mesh-based approaches fit a 3D body model to the input, infer the texture and render the mesh in the target pose [38, 17]. While capturing the 3D aspect, this has the downside that a specific human might not be captured well by a general model, for example due to uncommon hairstyles and spacious clothing.

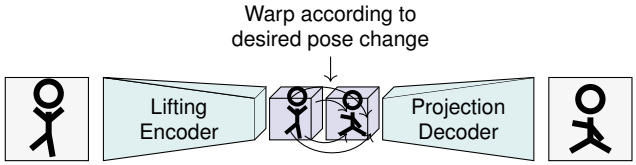

Figure 1. Our encoder network implicitly learns a volumetric representation of the input person, such that 3D feature warping can be applied in the middle of the architecture to achieve reposing.

Inspired by recent volumetric approaches for related tasks [25, 24], we propose a novel reposing method, illustrated in Fig. 1, which warps 3D volumetric CNN-features without requiring an explicit mesh model. Using only a 2D image as input, our model implicitly learns a latent volumetric representation of the input person. This representation is then warped using 3D transformations based on input and target pose to align it to the target pose. We process the warped features along with 3D target pose heatmaps with a decoder, to synthesize the reposed image.

By ablation, we show the benefits of the two 3D aspects of our work: first the 3D warping, and second, representing the target pose in 3D. Overall, our method achieves state-of-the-art scores on the commonly used DeepFashion and the newer iPER benchmarks.

2. Related Work

Image generation methods have come a long way since the introduction of generative adversarial networks (GAN) [5]. Building on Isola *et al.*'s image-conditioned GAN [12], Ma *et al.* were first to tackle pose-conditioned person image generation [20]. They feed the image and 2D target pose heatmap through two stages: the first is trained with a pixel-wise L_1 loss, the second with an adversarial loss. Lakhal *et al.* [10] use two encoders in both stages, distinguishing between aligned and misaligned input in Stage I and between pose and images in Stage II. Similar subdivisions are used in other works [31, 40].

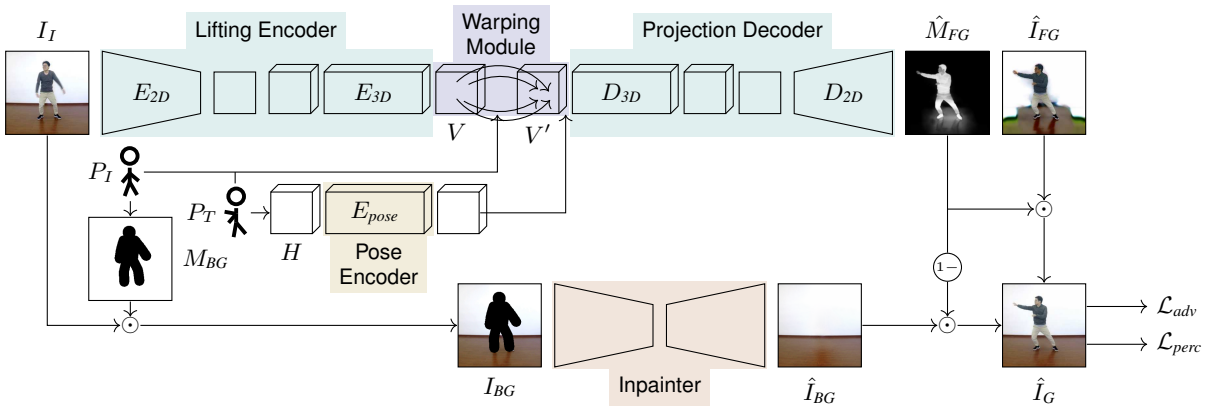

Figure 2. Generator architecture. The foreground stream learns 3D features from a 2D image and applies 3D feature warping. The result is combined with the target pose and projected to an RGBA image. Alpha blending with an inpainted background yields the final output.

The misalignment of input and target is tackled by explicit warping in many works. Siarohin *et al*. [32] use affine warps on the skip connections of a U-net architecture [28]. They mask out features corresponding to bodyparts based on the input pose and warp them to align with the target pose. An extension is proposed in [9], adding self-attention layers, spectral normalization and a relativistic discriminator to the architecture. In [2], a similar transformation is applied directly on the input image, using learned soft masks.

Some methods warp based on body part segmentation or dense pose [1] to encode input and target pose instead of using keypoint representations [3, 23, 6]. This tells the network the exact shape of the target person, making the task simpler, but a dense target pose is not available in general.

The 3D mesh-based approach of [38] fits a body model to the given person, back-projects pixels onto the mesh and transforms the mesh to the target pose. Unseen texture is inpainted with a neural net. Other methods [17, 15] use meshes to compute a transformation flow from input to target pose, which is used to transform network features in 2D.

A line of works in 3D human pose estimation [25, 35, 19, 30] has shown that it is feasible to predict depth-related information from images in a volumetric representation (in that case volumetric body joint heatmaps), by a tensor reshaping operation. We take this as inspiration to predict volumetric feature maps of humans in our work.

Human reposing can be viewed as a generalization of novel-view synthesis (NVS) from rigid to articulated pose. As volumetric prediction has also been successfully applied for NVS [34, 24], we take this as further motivation to investigate the usefulness of a similar representation in reposing.

In contrast to the volumetric approach, a sparse 3D representation is used in [27] to learn NVS. The encoder outputs an appearance feature vector and a 3D point cloud representing the pose. After rotating the point cloud, the decoder transforms both back to an image from another view. The implicitly learned point cloud is given to a shallow human pose estimation network, thereby reducing the amount of labeled pose estimation data needed. Similar to our method, an implicitly learned 3D structure is explicitly transformed, however, instead of a point cloud which only represents the pose we transform volumetric features which also contain appearance information.

3. Method

Given an input image I_I of a person and a target pose P_T, we aim at generating an image $\hat{I}_G$ of the person in pose P_T. We use a two-stream generator network to tackle this problem, where the first stream reposes the person using our novel volumetric feature warping approach, while the second inpaints missing parts of the background. To utilize the volumetric warping, the model has to estimate the depth of different bodyparts such that it can lift the corresponding features accordingly to a 3D volume. This is learned implicitly from the 3D warping, we neither give depth information about the input pose to our model nor do we apply any explicit supervision with respect to the input pose.

3.1. Architecture

Our architecture consists of a lifting encoder, a 3D warping module, a projection decoder and a background inpainter as shown in Fig. 2.

The **lifting encoder** maps a 2D input image to 3D volumetric features. The 2D input image I_I is passed to a convolutional network E_{2D} which outputs 2D feature maps $E_{2D}(I_I) \in \mathbb{R}^{H\times W\times D\cdot C}$. A reshape operation splits the channel dimension of the resulting tensor into different depth layers, yielding the feature volume $F \in \mathbb{R}^{H\times W\times D\times C}$. This is similar to how joint heatmaps are estimated in [25], but instead of heatmaps, we produce a latent feature volume. E_{2D} thus learns that different features in its output belong to different depths. To further process these volumetric features, a 3D convolutional network (E_{3D}) is applied to yield $V \in \mathbb{R}^{H\times W\times D\times C}$.

The key element of our approach is our novel **3D warping module**, whose purpose is to shuttle voxel features

to their target location. It gets a feature volume $V \in \mathbb{R}^{H\times W\times D\times C}$, together with the 3D input and target pose $P_I, P_T \in \mathbb{R}^{J\times 3}$ which are given as 3D joint coordinates. The input pose P_I is used to create ten masks $M_i \in \{0,1\}^{H\times W\times D}$, one per bodypart. Masks are generated by drawing capsular shapes between the joints corresponding to that bodypart, *e.g.*, the lower left leg's mask is based on the left ankle and the left knee joints and the mask of the torso depends on the hips and shoulders. We then create ten copies V_i of the feature volume and apply the corresponding mask by voxel-wise multiplication, giving ten volumes, one per body part. Next, we fit a transformation T_i for each body part based on input and target joints. We assume that each part moves rigidly, but as the scale of the person in pixel space may change, we also add a scale parameter. The result is a 7-parameter Helmert transformation, estimated by least squares. When a body part has only two joints, as for leg and arm parts, we use a third joint to specify the rotation around the body part's axis. For example, the left lower arm's motion would only depend on left wrist and left elbow, so we use the left shoulder's position as an anchor. The masked bodypart features are then warped according to the respective transformation using trilinear interpolation and combined using the maximum activation. Given M_i and T_i, the output feature volume of the warping module is

$$V' = \max_i T_i(M_i \odot V).$$

The **target pose encoder** E_{pose} feeds the target pose into our model. Its input are Gaussian volumetric heatmaps $H \in \mathbb{R}^{H\times W\times D\times J}$, one per body joint. The result is concatenated to the warped volumetric features and processed by the projection decoder.

Mirroring the lifting encoder, our **projection decoder** contains two parts D_{3D} and D_{2D}. The 3D convolutional network D_{3D} allows to enhance the warped features and also combines them with the output of the target pose encoder. This volume is reshaped to 2D by combining depth and volumetric channels into a single channel dimension. We then apply the second decoder network D_{2D}, yielding the generated RGB image $\hat{I}_{FG}$ together with a soft mask $\hat{M}_{FG}$.

We apply a **background inpainter** stream, since our warping module masks bodyparts and only copies those to the decoder, so background information is lost. We remove the person from the inpainter's input by using the bodypart masks from our warping module. Pixels not included in any of the projected bodypart masks become part of the background mask M_{BG}. The inpainting itself is performed using PartialConv layers [16]. The final result is a weighted combination (alpha blending) of the inpainted background $\hat{I}_{BG}$ and the generated person $\hat{I}_{FG}$ using $\hat{M}_{FG}$ as the weights.

Architectural details. All our sub-networks except the background inpainter, but including the discriminator, are based on ResNet [7, 8]. We use GroupNorm [37] instead of BatchNorm [11] due to its better performance with small batch sizes. In E_{2D} and D_{2D} we use bottleneck residual blocks to reduce computational cost. Our 3D convolutional networks E_{3D}, D_{3D} and E_{pose} do not use bottlenecks, as the number of features is already comparatively low.

3.2. Training

The perceptual loss $\mathcal{L}_{perc}$ [13] compares generated and target image by passing both images through an ImageNet-pretrained VGG net [33] and computing the L_1 loss on multiple feature maps. The adversarial loss $\mathcal{L}_{adv}$ uses a discriminator net as in a classical GAN. The discriminator gets the generated or ground truth image along with the input image and the 3D target heatmap. We jointly optimize a weighted combination of these losses:

$$\mathcal{L}(\theta) = \lambda_{perc}\mathcal{L}_{perc}(\theta) + \lambda_{adv}\mathcal{L}_{adv}(\theta)$$

We use data augmentation with rotation, scaling, translation, horizontal flip and color distortion. We train with the Adam optimizer [14] for 150,000 steps with batch size 2 and learning rate $\alpha = 2 \cdot 10^{-4}$. We set $\lambda_{adv} = 1$, $\lambda_{perc} = 3$.

4. Experiments

4.1. Datasets

Commonly used in related work, the In-shop Clothes Retrieval Benchmark of the DeepFashion dataset (Fashion) [18] has almost 50,000 images and 8,000 sets of clothes.

The newer Impersonator dataset (iPER) [17] contains videos of 30 people and 103 clothing styles in total. Two videos exist per clothing style, filmed from a static camera. In one, the person turns around in an A-pose, the other shows arbitrary movements.

As these benchmark datasets do not supply 3D poses, we apply a 3D human pose estimation network based on [30] to obtain the input and target poses P_I and P_T.

4.2. Evaluation Metrics

Although generated image quality is somewhat subjective, several quantitative metrics have been used in related work to compare methods. The structural similarity index (SSIM) [36] compares patches of the generated image to patches of the ground truth according to luminance, contrast and structure. While also used in some related work, we argue that the Inception score [29] is not suited for this task (we elaborate this in the supplementary). We further use the learned perceptual image patch similarity (LPIPS) [39], which compares deep features between generated image and ground truth, similar to perceptual losses [13].

To evaluate high-level structure, we compare the response of a pretrained 3D pose estimator based on [30], when applied to the generated and the true image. We use the area under the PCK (percentage of correct keypoints) curve (AUC@150mm), a standard pose metric [22].

3D warping	3D target pose	SSIM ↑	$SSIM_{fg}$ ↑	Pose AUC ↑
–	–	0.872	0.566	0.698
–	✓	0.875	0.578	0.749
✓	–	0.877	0.607	0.749
✓	✓	**0.883**	**0.626**	**0.777**

Table 1. Ablation on iPER. $SSIM_{fg}$ only evaluates the foreground.

	iPER		Fashion	
	SSIM ↑	LPIPS ↓	SSIM ↑	LPIPS ↓
PG2, Ma *et al*. [20]	0.854	0.135	0.762	–
SHUP, Balakrishnan *et al*. [2]	0.823	0.099	–	–
DSC, Siarohin *et al*. [32]	0.829	0.129	0.756	–
LWB, Liu *et al*. [17]	0.840	0.087	–	–
SGW, Dong *et al*. [3]	–	–	0.793	–
UPIS, Pumarola *et al*. [26]	–	–	0.747	–
VUNET, Esser *et al*. [4]	–	–	0.786	0.196
BodyROI7, Ma *et al*. [21]	–	–	0.614	–
DPT, Neverova *et al*. [23]	–	–	0.796	–
CTI, Grigorev *et al*. [6]	–	–	0.791	**0.169**
Li *et al*. [15]	–	–	0.778	–
Ours	**0.883**	**0.081**	**0.800**	0.186

Table 2. Comparison to prior work. iPER scores taken from [17].

4.3. Ablation Study

In contrast to prior work on person reposing, we propose to perform two different aspects in 3D: first, we use a 3D target pose and second, we perform 3D feature warping in the center of our model. Architectural differences make it hard to directly compare our results to prior works, so we define ablation models to investigate these two aspects while keeping the exact same architecture otherwise.

To drop the depth information from the 3D target pose heatmaps, we project the pose to the image plane and replicate it to all depth layers. Similarly, to perform warping in 2D, we project the body part masks to the image plane and copy them to all depths and apply 2D affine warpings to all depths independently.

The results on iPER (Tab. 1) show that both of our 3D enhancements improve the scores compared to the 2D baseline and the results get even better when they are combined. This is supported by the qualitative results (Tab. 3). In the first row, the 2D pose models wrongly generate the right hand in front of the body, while the second row shows that a combination of both 3D aspects achieves the best results.

4.4. Comparison to Prior Work

Our model achieves state-of-the-art scores on both datasets (Tab. 2). Comparison to [17] on iPER (Tab. 3) shows that our model is able to transform the features of the left arm independently from the body features. In the upper row the hand correctly appears behind the body and the blue jacket in the lower row does not have a white stain as residue from the arm color. On Fashion (Tab. 4), our model generates the overlapping arms of the right person better than the 2D feature warping approach of [32].

input image	target pose	LWB [17]	2D	3D pose	3D warp	3D both (ours)
						

Table 3. Comparison to a mesh-based method and ablation models.

input image	target pose	DSC [32]	ours	input image	target pose	DSC [32]	ours
							

Table 4. Comparison with a 2D feature warping method. The target image is not used as input, only its pose.

Our architecture decreases the spatial size of the feature maps, which has the result that fine details are lost in some cases, which is also visible in the generated results. The buttons on the shirt in the first row of Tab. 3 are missing in two ablation models and replaced by a zipper and shirt pockets in the other two.

5. Conclusion

We presented a novel architecture for person reposing, which relies on 3D warping of implicitly learned volumetric features. Different from prior work, our approach is neither limited by approximating 3D motion with 2D transformations nor is an explicit 3D human mesh model required.

The ablation study and the comparison to related approaches showed that our method outperforms 2D warping methods by a significant margin. This indicates that volumetric representations and 3D warping are a promising way to tackle reposing and we expect that more sophisticated neural rendering techniques could further improve results.

Acknowledgments. This work was funded, in parts, by ERC Consolidator Grant project "DeeViSe" (ERC-CoG-2017-773161) and a Bosch Research Foundation grant. Some of the experiments were performed with computing resources granted by RWTH Aachen University under projects "thes0618" and "rwth0479".

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *CVPR*, 2018.

[2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018.

[3] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-GAN for pose-guided person image synthesis. In *NIPS*, 2018.

[4] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational U-net for conditional appearance and shape generation. In *CVPR*, 2018.

[5] Ian Goodfellow et al. Generative adversarial nets. In *NIPS*, 2014.

[6] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *CVPR*, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

[9] Yusuke Horiuchi, Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Spectral normalization and relativistic adversarial training for conditional pose generation with self-attention. In *MVA*, 2019.

[10] Mohamed Ilyes Lakhal, Oswald Lanz, and Andrea Cavallaro. Pose guided human image synthesis by view disentanglement and enhanced weighting loss. In *ECCV*, 2018.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[15] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *CVPR*, 2019.

[16] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018.

[17] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, 2019.

[18] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

[19] Diogo C. Luvizon, David Picard, and Hedi Tabia. 2D/3D pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018.

[20] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017.

[21] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018.

[22] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017.

[23] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *ECCV*, 2018.

[24] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3d representations from natural images. 2019.

[25] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017.

[26] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *CVPR*, 2018.

[27] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3D human pose estimation. In *ECCV*, 2018.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[29] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NIPS*, 2016.

[30] István Sárándi, Timm Linder, Kai O. Arras, and Bastian Leibe. Metric-scale truncation-robust heatmaps for 3D human pose estimation. In *Int. Conf. on Automatic Face and Gesture Recognition*, 2020.

[31] Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. Multistage adversarial losses for pose-based human image synthesis. In *CVPR*, 2018.

[32] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable GANs for pose-based human image generation. In *CVPR*, 2018.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[34] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. DeepVoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019.

[35] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018.

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *Trans. Image Proc.*, 2004.

[37] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018.

[38] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *CVPR*, 2018.

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[40] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019.