Improving Key Human Features for Pose Transfer

Victor-Andrei Ivan[1], Ionut Mistreanu[1], Andrei Leica[1], Sung-Jun Yoon[2], Manri Cheon[2], Junwoo Lee[2] and Jinsoo Oh[2]

[1]Arnia Software, Europe House, 47-53 Lascar Catargiu Bvd Bucharest, Romania
[2]LG Electronics, 19, Yangjae-daero 11-gil, Seocho-gu, Seoul, Republic of Korea
{*victor.ivan, ionut.mistreanu, andrei.leica*}*@arnia.ro,*
{*sungjun.yoon,manri.cheon,junwoo.lee,jinsoo.oh* }*@lge.com*

Abstract

It is still a great challenge in the Pose Transfer task to generate visually coherent images, to preserve the texture of clothes, to maintain the source identity and to realistically generate key human features such as the face or the hands. To tackle these challenges, we first conduct a study to obtain the most robust conditioning labels for this task and the baseline method [44] that we choose. We then improve upon the baseline by including deep source features from an Auto-encoder through an Attention mechanism. Finally we add region discriminators that are focused on key human features, thus obtaining results competitive with the state-of-the-art.

1. Introduction

Human image synthesis aims to create believable and photorealistic images and videos of people. It has a vast potential for fashion retail, video character animation, movie or game making, etc.

Due to recent advancements in GANs [8], the Pose Transfer and Motion Transfer tasks have attracted much attention. The pose transfer task attempts to generate people in a certain position given an appearance representation of a person and a target pose, usually represented through keypoints, heatmaps, or semantic masks. Motion Transfer aims to achieve the same result while preserving temporal coherence of adjacent frames at the same time.

Recent methods tackle the human generation problem in a number of ways. In the Pose Transfer case, Siarohin *et al*. [34] embed the target pose and the source human, and apply affine transformations to the human's body parts in the feature space to fit the target pose. Men *et al*. [28] embed each body part of the source person into a style vector and feed the style through ADAIN [11] layers into the generator network. Tang *et al*. [37] model the source and the target pose in a bipartite graph to capture long-range relations. CoCosNet [44] is a more versatile, state-of-the-art network built for general appearance-based semantic image synthesis which can also be applied in the Pose Transfer case. It builds correspondences between the source image and the target pose and feeds them into a generator network through SPADE blocks [30].

In the Motion Transfer case, two main approaches can be found: models that learn a mapping between poses and an identity [5, 40] and one/few-shot models [7, 35, 39].

Chan *et al*. [5] learns to generate two consecutive images, for the second image using the previously generated image. Vid2Vid [40], Few-Shot Vid2Vid [39], FOMM [35] are warping based approaches. Vid2Vid and Few-Shot Vid2Vid warp the source image onto the target pose in the image space, generate an intermediary image and combine the 2 using a predicted mask, whereas FOMM warps the source image in the feature space and generates a single final image. Gafni *et al*. [7] generate videos by performing Pose Transfer on the semantic label map of the source person, and learn to generate realistic textures on the different semantic parts.

The aforementioned works show great results, but maintaining the source identity and realism of key human features, like the face, the hands and the hair, is still a great challenge. In this work we focus on the Pose Transfer task. We build our framework upon the state-of-the-art architecture presented in [44] and attempt to improve the ability of maintaining the face identity and key human features.

To achieve the above we first conduct a study to find the best conditioning for the Pose Transfer task in the setting of the chosen baseline model. The chosen pose conditioning also helps transferring realistic textures from the source

image to target pose through texture transfer using Densepose [9]. Similar to Liu *et al*. [21, 22], given that, when there are large pose differences, gaps are left in the transferred textures, we add an auto-encoder [31] that learns the reconstruction of the source image. Attention [38] modules are used to combine features in the generator with visually meaningful features from the auto-encoder. We use region discriminators to enforce realism on regions of interest similar to the local-global discriminators in Liu *et al*. [21, 22].

Our work can be summarized as follows:

- We improve the identity preservation capacity of the baseline method by transferring real textures from the source image and by employing attention modules between the generator features and the features of a reconstruction auto-encoder
- We address the problem of generating key human features, such as the face and the hands, by employing region-specific discriminators
- We conduct a study on different types of pose conditionings and choose the most appropriate one for our task

2. Related work

Deep Generative Models have made remarkable achievements in image generation. As the original GAN [8] model was only able to synthesize low-resolution images, the follow-ups improved upon it with higher resolutions [15] and increased realism [16, 17].

We review the literature for human pose transfer and its application to pose-guided image and video generation.

Pose-guided image to image generation. Pose transfer refers to the problem of synthesizing human images with a novel user-defined pose.

Recent work is mainly based on the conditioned generative adversarial networks (CGAN) [2, 25]. Their key technical idea is to combine the source image along with the target pose as inputs and generate a realistic image by GANs. The differences among those approaches usually stand in conditioning label formats, network architectures, warping strategies, and adversarial losses.

The conditioning pose is often captured by 2D keypoints [25, 42], parametric meshes [21, 22] and semantic masks [7, 13, 43]. Many works also use Densepose [9], which is the projection of the SMPL [24] model with UV parameterization in the image coordinates, as conditioning input. This enables direct warping of pixels of the input image to the spatial locations at the output with target pose [26, 29, 33].

Some works [1, 6, 34] warp the source features onto the target pose, like skeleton or parsing map. Besides, Li *et al*. [18] propose to learn a transformation flow from 2D key points and warp the deep features based on the learned transformations. Zhang *et al*. [43] generate a human parsing map in the target pose and use gated convolutions to deform the images. In [13] the authors disentangle the shape and style by using parsing maps and joint global and local region-wise encoding and normalization. Other works [29, 33] extract the textures from the source image and perform in-painting to aid the pose transfer.

Pose-guided video generation. Since the methods for pose transfer are designed to output a single image, their application to a sequence of poses to perform video generation can exhibit temporal inconsistencies. To mitigate this problem, many methods enforce explicit temporal constraints in their algorithms. Chan *et al*. [5] predict the person image in two consecutive frames. Yang *et al*. [41] condition the temporally coherent semantics on a generative adversarial network.

Recent video generation approaches have leveraged optical flow prediction [40], local affine transformation and grid-based warping field [35], body parts transformation [46], and future frame prediction [3] to generate realistic videos.

Liu *et al*. [20] learns to predict a dynamic texture map that allows rendering physical effects, *e.g*. pose-dependent clothing deformation, to enhance the visual realism on the generated person.

3. Method

3.1. Overview

We aim to learn a Pose Transfer network that generates realistic textures and pays attention to human features such as the face and the hands. Besides generating these features in a realistic way, we also aim to maintain the source identity.

In the following we will present the baseline that we compare ourselves to and that we build upon. Then we will present our changes to the baseline to improve the results. The entire architecture can be seen in Figure 1.

3.2. Baseline

We choose the state-of-the-art image-to-image translation network, CoCosNet [44], as our baseline. For the Pose Transfer task, the network takes as inputs a source image and a target label represented by a stick figure obtained from 18 keypoints, with each limb and keypoint color-coded on a black background.

The inputs are fed into a Cross-Domain Correspondence module that embeds a source image and a target label into a shared domain and then performs cross-correlation between the two. The resulting correlation is used to warp the source image onto the target label's semantics.

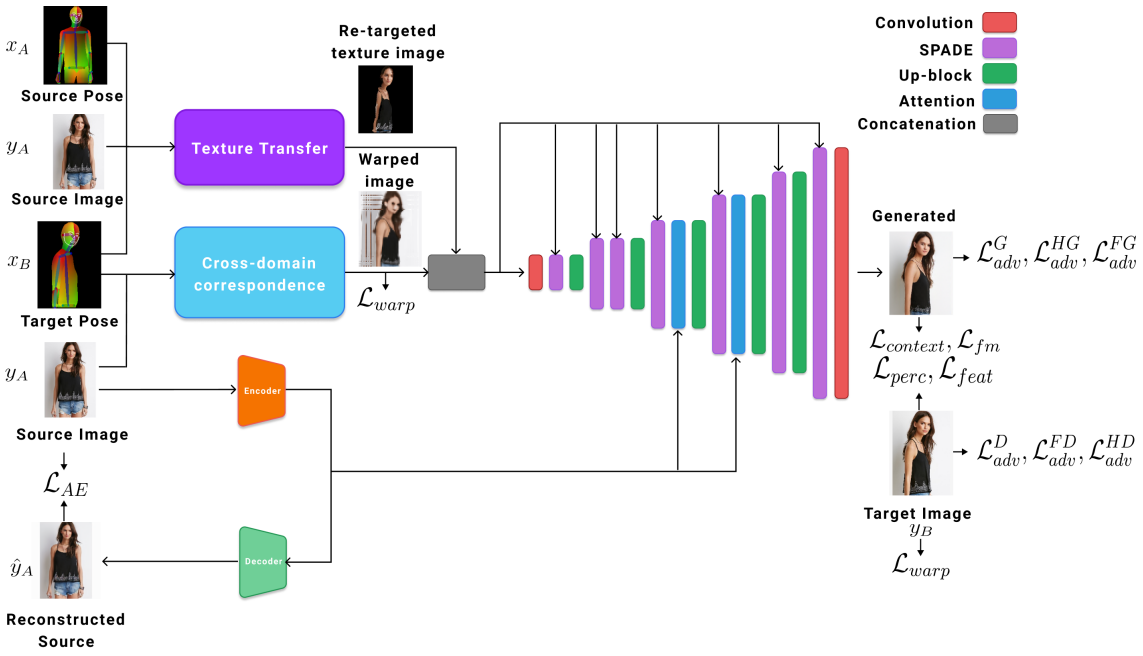

Figure 1. The figure shows the architecture used in this paper.

The result of the warp is then concatenated to the target label and fed as conditioning to the generator through SPADE modules.

The CoCosNet [44] loss function for the Pose Transfer task is:

$$\begin{aligned}\mathcal{L}_{CoCosNet} =& \mathcal{L}^D_{adv} + \mathcal{L}^G_{adv} + \mathcal{L}_{fm} + \mathcal{L}_{domain} \\ &+ \mathcal{L}_{warp} + \mathcal{L}_{feat} + \mathcal{L}_{perc} \\ &+ \mathcal{L}_{context},\end{aligned} \tag{1}$$

where $\mathcal{L}^D_{adv}$ and $\mathcal{L}^G_{adv}$ are hinge adversarial losses [19] for the discriminator and generator respectively. $\mathcal{L}_{fm}$ is the GAN feature matching loss [32]. The domain alignment loss, $\mathcal{L}_{domain}$, from [44] is computed in order to bring the embeddings of the target pose and of the source image to the same domain. The warp loss, $\mathcal{L}_{warp}$, is applied to the warped image and target image in order to train the cross-correspondence module. Two VGG-based [36] perceptual losses [14] are used. The first one, $\mathcal{L}_{feat}$, is an L_1 loss on multiple VGG features between the generated image and the target image. The second one, $\mathcal{L}_{perc}$, is an L_2 loss on the *relu4_2* layer. The contextual loss [27], $\mathcal{L}_{context}$, is used to match statistics between the generated image and the source image. For more details on the losses used we refer the reader to the CoCosNet paper [44].

Although CoCosNet [44] has great results for the Pose Transfer task on the widely used DeepFashion benchmark [23], we notice that there are a few improvements that could be made.

3.3. Improving the results through dense conditioning

The Cross-Domain Correspondence network performs a correlation between the source image and the target label. The target label is represented by a stick figure obtained from keypoints for the body, hands, legs, nose, eyes and ears.

Firstly, our intuition is that a pose representation that takes into account areas of fine detail, *e.g.* the face and the hands, would improve the results. Therefore we add face and hands keypoints using the out-of-the-box Openpose detector [4].

Secondly, the target label is formed of an image showing joints, limbs and distance maps for each joint like in [44]. Therefore, it might be ambiguous for the cross-correlation operation in the Cross-Domain Correspondence network to align a source image with such a sparse representation. Thus, we test the method with a more dense pose representation, using the Densepose architecture [9] to detect IUV maps. The joints and limbs are now drawn on top of the image of the IUV maps similar to [39].

Finally, through the nature of the architecture, the CoCosNet [44] generator never sees the source image, only its warped version concatenated with the target label, thus it heavily relies on the quality of the warp and the Cross-Domain Correspondence module. If the warping module does not learn to properly warp fine details like patterns and facial features, the Generator simply learns the dataset statistics and fills out the missing gaps, without maintaining

Figure 2. The figure shows a comparison of the results on the test split between different conditioning for the baseline network. Op stands for the baseline annotation type, 18 body keypoints without face and hands keypoints. Op++ adds face and hands keypoints. Dp adds the Densepose [9] annotation. TT adds the texture transfer.

the source identity. For this we make use of Densepose [9] once more. Densepose [9] predicts correspondences, for 24 body parts, beetween the 2D image and a 3D surface model. Similar to [26, 29, 33], we extract the visible textures from the source image using the IUV representation, and transfer them to the corresponding areas in the target Densepose. The re-targeted texture image is concatenated to the label and the warping result. This can be seen in Figure 1 as Texture Transfer which is not learnable.

In Section 4.1 we show an ablation study on these conditioning methods and the impact on the results.

3.4. Filling in the gaps

Depending on the difference between the source and the target poses, the re-targeted textures will have more or fewer gaps. Thus, the need appears to have an alternative source of

information. Similar to [21, 22] we train an auto-encoder in parallel with the Pose Transfer architecture that learns to reconstruct the source images. The intuition here is that using encoded features from this auto-encoder we can in-paint in the generator's feature space the missing information from the Densepose-based transferred textures. This information exchange, though, cannot be made naively through addition or concatenation, due to pose misalignment. To solve this problem we employ attention modules at different resolutions in the generator architecture that recombine the corresponding features of the two networks. The encoder and decoder of the auto-encoder are 3 layer, convolutional neural networks.

Features from the auto-encoder's encoder, at 32×32 and 64×64 resolution, are fed to the attention modules together with features from the generator of the corresponding resolutions. In the attention module, the two feature maps, of shape $C \times H \times W$ are first passed through two 1×1 convolutions to reduce the channel size $n = 8$ times. The reduced feature maps are reshaped into two matrices of shape $C \times HW$, multiplied and passed through a Softmax layer to give an attention matrix of dimension $HW \times HW$. The attention matrix is multiplied with the auto-encoder's rehsaped feature map which is reshaped back to $C \times H \times W$ and finally scaled and added to the generator's feature map.

In Section 4.1 we show an ablation study on the changes made to the architecture and the impact on the results.

3.5. Training

Inspired by the global-local discriminators in [21, 22], additional to the loss in equation 1, we use a patch discriminator [12] for the face region and one for the hands regions, under the motivation that the human eye pays much attention to these details. The face and hands regions are only small areas in the entire image, thus gradients are not focused on those areas in particular. We notice that adding these discriminators improves the quality considerably. The face and hands discriminators' losses are as follows:

$$\begin{aligned}\mathcal{L}_{adv}^{FD} = &- \mathbb{E}[\min(0, -1 + \mathcal{D}(y_B^{face}))] \\ &- \mathbb{E}[\min(0, -1 - \mathcal{D}(\mathcal{G}(x_B, y_A)^{face}))]\end{aligned} \tag{2}$$

$$\mathcal{L}_{adv}^{FG} = -\mathbb{E}[\mathcal{D}(\mathcal{G}(x_B, y_A)^{face}] \tag{3}$$

$$\begin{aligned}\mathcal{L}_{adv}^{HD} = &- \mathbb{E}[\min(0, -1 + \mathcal{D}(y_B^{hands}))] \\ &- \mathbb{E}[\min(0, -1 - \mathcal{D}(\mathcal{G}(x_B, y_A)^{hands}))]\end{aligned} \tag{4}$$

$$\mathcal{L}_{adv}^{HG} = -\mathbb{E}[\mathcal{D}(\mathcal{G}(x_B, y_A)^{hands}] \tag{5}$$

Where y_B^{face} and y_B^{hands} are face and hand crops from the target image, $G(x_B, y_A)^{face}$ and $G(x_B, y_A)^{hands}$ are crops from the generated image, and x_B and y_A are the target pose and the source image respectively.

The auto-encoder is trained in 2 steps. Firstly, the encoder part is trained jointly with the generator through the same losses as the generator. Secondly, we add a reconstruction loss to train the auto-encoder to reconstruct the source image, such that visually relevant features are learned. The second step is performed once every 5 iterations:

$$\mathcal{L}_{AE} = \|y_A - \hat{y}_A\|_1 \tag{6}$$

The final loss used to optimize our model is:

$$\begin{aligned}\mathcal{L} = \min_{G,AE} \max_{D,FD,HD} &\mathcal{L}_{CoCosNet} + \mathcal{L}_{adv}^{FD} + \mathcal{L}_{adv}^{FG} \\ &+ \mathcal{L}_{adv}^{HD} + \mathcal{L}_{adv}^{HG} + \mathcal{L}_{AE}\end{aligned} \tag{7}$$

4. Experiments

Implementation details. The model is trained using person images of resolution 256×256, face crops of resolution 64×64 and hand crops of resolution 32×32 on 4 NVIDIA RTX2080Ti GPUs with a mini-batch size of 4, *i.e.* one image per GPU, for 100 epochs. The face and hands crops are obtained by using the Densepose part mask. The model was trained using the Adam optimizer with learning rate 2e-4, β_1=0.5, β_2=0.999 as in [44]. We train the network on the DeepFashion dataset [23] using the same train/test splits as in [44] for comparison with the baseline, and the train/test splits from [47] for comparison with other Pose Transfer methods.

4.1. Ablation

As explained in section 3.3 we tested a few pose conditioning methods in order to improve the results of the Pose Transfer network. For this purpose we train multiple models and evaluate which conditioning is best. Note that due to resource constraints, we only train the models for 10 epochs. First we train the baseline network with the conditioning used in [44]. Afterwards we successively modify the conditioning by first adding face and hands keypoints, then adding the Densepose annotation, and finally adding the texture transfer. We also add an experiment without face and hands keypoints, but with Densepose and texture transfer. We compute the FID for each of these models. As it can be seen in Table 1, the FID keeps on improving as we add more information in the conditioning label. We choose to use the conditioning that contains face and hands keypoints, Densepose annotation and texture transfer for the following reason: given the FID values and by visually evaluating the results, see Figure 2, we observe that in most cases, the identity and the textures are better preserved. Considering

Figure 3. Figure showing the results on the test split of our method when compared to our baseline. Our model generates faces that are more realistic and closer to the identity. In some cases, *e.g.* rows 4 and 6, the hands look more realistic as well. Due to the dense conditioning label, it can also be noticed that when the target has their hands in their pocket it is reflected in the generated image.

this, we finally opt for the conditioning that includes the texture transfer.

We also run experiments to see whether the auto-encoder with attention and the region-wise discriminators bring improvements to the score. In Table 2. Base stands for the baseline model conditioned on Openpose [4] with hands and face, on Densepose and on the transferred texture. AE + ATT is the Base model with the auto-encoder and the attention added. Ours represents AE + ATT trained with the extra face and hands discriminators. As it can be seen in the table, adding the auto-encoder and the attention improves the FID. When adding the region discriminators, the score only slightly drops, but upon visual inspection, it was observed that the face and hands are generated better.

4.2. Comparison to baseline

We train our full model and make a quantitative comparison to the CoCosNet [44] baseline, using the trained weights offered by the authors. We compute the same metrics as Zhang *et al*. [44]: the Frechét Inception Distance (FID) [10], which measures the distance between two distributions, and the Sliced Wasserstein Distance (SWD) [15], which measures the statistical similarity of patches at different scales. The results are shown in Table 3. The FID and SWD results are shown. As it can be seen our model performs better in the case of both metrics. A visual comparison between our method and the baseline can be seen in Figure 3, showing improvements in key human features and identity.

4.3. Comparison with state-of-the-art methods

We also compare our method to state-of-the-art Pose Transfer methods, BiGraphGAN [37], ADGAN [28], PINet [43] and PISE [13]. We compute the same metrics as in PISE [13]: the FID [10], the Learned Perceptual Image Patch Similarity (LPIPS) [45] to measure the perceptual difference between images and PSNR to measure pixel-level differences. We use the generated images provided by the authors in order to compute the metrics on the test split. In the case of PINet [43], BiGraphGAN [37] and ADGAN [28] the provided images have 176x256 resolution. Therefore, for a fair comparison, we crop all the images to 176x256 for evaluation with PSNR, and pad back to 256x256 with white pixels for all the images in the case of FID and LPIPS. The quantitative results of the comparison can be seen in Table 4, and a visual comparison can be seen in Figure 4.

5. Conclusion

In this paper we present a Pose Transfer method built on top of CoCosNet [44] that achieves better performance in terms of maintaining the source identity and generating human features more realistically. This is achieved by using a dense conditioning augmented by textures transferred from the source image to the target pose by using Densepose [9]. Additionally, similar to *et al*. [21, 22], we add a network that learns visually coherent features in an unsupervised way and combines the learned features with the generator's features through an attention mechanism and we region discriminators for important areas in the images, *i.e*. the face and the hands. We obtain results comparable with state-of-the-art methods.

Conditioning	FID ↓
Op	23.56
Op++	21.84
Dp + Op++	19.32
Dp + Tt + Op	18
Dp + Tt + Op++	17.41

Table 1. The FID results for the baseline model trained with different label conditioning. Op stands for Openpose [4] annotation without hands and face keypoints. Op++ adds face and hands keypoints to the annotation. Dp stands for Densepose annotation. Tt stands for texture transfer.

Model	FID ↓
Base	17.41
AE + ATT	16.95
Ours	17.25

Table 2. The table shows the FID results for the baseline model with the chosen conditioning (Base), for the Base model + the auto-encoder and attention modules (AE + ATT) and for our final model (Ours) which stands for AE + ATT with the face and hands disciminators.

Conditioning	FID ↓	SWD ↓
CoCosNet [44]	14.4	17.2
Ours	12.5	13

Table 3. The FID and SWD results for the baseline model and our model.

Conditioning	FID ↓	LPIPS ↓	PSNR ↑
BiGraph [37]	25.54	0.2021	**31.43**
ADGAN [28]	16.69	0.1973	31.34
PINet [43]	16.01	0.1924	31.36
PISE [13]	13.98	0.192	**31.43**
Ours	**12.74**	**0.1758**	31.15

Table 4. The FID, LPIPS and PSNR results for pose transfer methods and our model.

Figure 4. Figure showing the results on the test split of our method when compared to BiGraph [37], ADGAN [28], PINet [43] and PISE [13].

References

[1] Badour AlBahar and Jia-Bin Huang. Guided image-to-image translation with bi-directional feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[2] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. Contextvp: Fully context-aware video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[6] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 472–482, Red Hook, NY, USA, 2018. Curran Associates Inc.

[7] Oran Gafni, Oron Ashual, and Lior Wolf. Single-shot freestyle dance reenactment, 2020.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

[9] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc.

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

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

[13] Zhang Jinsong, Li Kun, Lai Yu-Kun, and Yang Jingyu. PISE: Person image synthesis and editing with decoupled gan. In *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

[18] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[19] Jae Hyun Lim and Jong Chul Ye. Geometric gan, 2017.

[20] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 05 2020.

[21] Wen Liu, Zhixin Piao, Min Jie, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[22] Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan with attention: A unified framework for human image synthesis, 2020.

[23] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

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

[26] Liqian Ma, Zhe Lin, Connelly Barnes, Alexei A. Efros, and Jingwan Lu. Unselfie: Translating selfies to neutral-pose portraits in the wild. In *ECCV*, 2020.

[27] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[28] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Computer Vision and Pattern Recognition (CVPR), 2020 IEEE Conference on*, 2020.

[29] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[31] D. E. Rumelhart and J. L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. 1987.

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[33] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *European Conference on Computer Vision (ECCV)*, 2020.

[34] Aliaksandr Siarohin, Stéphane Lathuilière, Enver Sangineto, and Nicu Sebe. Appearance and pose-conditioned human image generation using deformable gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[35] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[37] Hao Tang, Song Bai, Philip HS Torr, and Nicu Sebe. Bipartite graph reasoning gans for person image generation. In *BMVC*, 2020.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[39] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[41] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[42] Lingbo Yang, Pan Wang, Chang Liu, Zhanning Gao, Peiran Ren, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Xiansheng Hua, and Wen Gao. Towards fine-grained human pose transfer with detail replenishing network. *CoRR*, abs/2005.12494, 2020.

[43] Jinsong Zhang, Xingzi Liu, and Kun Li. Human pose transfer by adaptive hierarchical deformation. *Computer Graphics Forum*, 39(7):325–337, 2020.

[44] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020.

[45] 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.

[46] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Dance dance generation: Motion transfer for internet videos. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1208–1216, 2019.

[47] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019.