Disentangled Representation Learning for Controllable Person Image Generation

Wenju Xu, Chengjiang Long, Yongwei Nie and Guanghui Wang, *Senior Member, IEEE*.

Abstract—In this paper, we propose a novel framework named DRL-CPG to learn disentangled latent representation for controllable person image generation, which can produce realistic person images with desired poses and human attributes (*e.g.* pose, head, upper clothes, and pants) provided by various source persons. Unlike the existing works leveraging the semantic masks to obtain the representation of each component, we propose to generate disentangled latent code via a novel attribute encoder with transformers trained in a manner of curriculum learning from a relatively easy step to a gradually hard one. A random component mask-agnostic strategy is introduced to randomly remove component masks from the person segmentation masks, which aims at increasing the difficulty of training and promoting the transformer encoder to recognize the underlying boundaries between each component. This enables the model to transfer both the shape and texture of the components. Furthermore, we propose a novel attribute decoder network to integrate multi-level attributes (*e.g.* the structure feature and the attribute representation) with well-designed Dual Adaptive Denormalization (DAD) residual blocks. Extensive experiments strongly demonstrate that the proposed approach is able to transfer both the texture and shape of different human parts and yield realistic results. To our knowledge, we are the first to learn disentangled latent representations with transformers for person image generation.

Index Terms—Disentangled representation, Transformer, controllable person synthesize.

I. INTRODUCTION

Deep generative adversarial network [1], [2] has recently drawn increasing attention due to its impressive performance in image/video synthesis [3], [4], which shows great potential in dealing with MultiMedia applications [5], [6], [7]. For instance, exploring synthesized features has been proven to be an effective way to improve the performance of deep neural networks [8], [9], [10]. Manipulating facial images [11], [12], [13] and translating human faces into anime [14], [15] have become popular in social media applications. More recently, researchers have attempted to synthesize human images that can be controlled by user inputs [16], [17]. We can imagine that using synthesized digital humans for broadcasting, advertising, and educating will be promising. However, this is still an open problem that needs more endeavor and will significantly impact multimedia society.

W. Xu is with AMAZON, Palo Alto, CA 94301 USA (e-mail: xuwenju@amazon.com).

C. Long is with the META Reality Labs, Burlingame, CA, USA. Email: cjfykx@gmail.com.

Y. Nie is with the South China University of Technology, Guangzhou, Guangdong, China, Email: nieyongwei@scut.edu.cn.

G. Wang is with the Department of Computer Science, Toronto Metropolitan University, 350 Victoria St, Toronto, ON M5B 2K3. Email: wangcs@ryerson.ca.
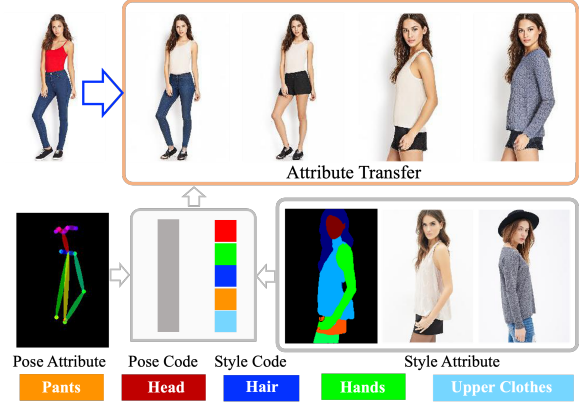

Fig. 1: Given a target pose and source person images with semantic masks, the goal of this paper is to design a unified approach for controllable person image generation and attribute transfer.

Controllable person image generation aims to synthesize a person image conditioned on the given pose and attributes at the component-level from source person images with the corresponding semantic masks, preserving attributes like person identity, cloth color, cloth texture, background *etc.*, as shown in Figure 1. This topic has attracted great attention due to its potentially wide applications in movie composition, image editing, person re-identification, virtual clothes try-on, and so on.

ADGAN [18] was proposed as the first work for controllable person attribute editing based on the semantic mask to separate each component. Although it achieves success in controllable image editing, the synthesized images are not realistic. In particular, separation in image level does not guarantee the disentanglement of the encoded attributes. Moreover, editing person attributes by simply replacing the entangled semantic representations tends to create artifacts or unrealistic results.

To solve the above issues of previous work, we propose a novel and unified framework, termed as DRL-CPG, for controllable person synthetic image generation. As shown in Figure 2, the framework consists of two major parts, *i.e.* an attribute encoder with transformers learned to generate disentangled representation and an attribute decoder that integrates the structure features and attribute representations for controllable person image generation. In contrast to ADGAN [18] which encodes each component into latent code directly, we introduce transformers [19] in the attribute encoder to generate an intermediate representation set for each component and

select the corresponding one as the final representation of this component. Note that a global receptive field via self-attention provided in the transformer encoder is essential for the disentangled representation of a person's image as small components often share similar textures and context with their surroundings. Then the transformer decoder integrates learnable component queries with the transformer encoder output to generate the attribute latent representation for the subsequent attribute decoder to conduct attribute transfer.

To enable an efficient and robust way for attribute editing, we introduce a random component mask-agnostic strategy, *i.e.*, randomly removing several component masks from the entire person mask. The more component masks are removed, the more difficult to distinguish or recognize. Obviously, this treatment is to increase the difficulty for our model to learn how to encode the parts without masks to separate each other into disentangled semantic representations. It requires our model to robustly recognize the underlying boundaries between different components and be able to transfer both the shape and texture of components.

Motivated by human's easy-to-hard learning process, we adopt the curriculum learning strategy [20] and start from a relatively easy step, then gradually increase the difficulty levels to the complete mask-agnostic step. Unlike NTED [21] relied on explicitly learned correlation matrices for feature extraction and distribution, we observe that the generated latent attribute representations are well clustered with the component mask-agnostic strategy. This easy-to-hard learning process not only enables our model to recognize the component boundary but also provides a way to preserve semantic completeness when distributing the extracted neural textures to different target poses. The learned attribute encoder can generate disentangled attribute representations for components in the person image. This significantly increases the flexibility of our model compared to PISE [22] which requires to prediction of a parsing mask and injects texture code into different parts of the predicted mask for person generation. Without requiring predicted masks, our model avoids the issues, *e.g.* , holes and artifacts in the generations raised by misclassified regions in the predicted masks. Our model is also able to adaptively transfer the shape of the components to fit the source person, which is not possible for mask prediction based methods without sufficient guidance to reshape the predicted masks.

Regarding our attribute decoder, it can adaptively integrate the structure feature and the attribute representation for person image generation. Note that the pose map only provides the structural connection between different joints, while it does not contain any structural information within the local region, making it difficult to synthesize rich local structures. Inspired by SPADE [23] and AdaIN [24], [25], we design Dual Adaptive Denormalization (DAD) residual blocks to explore the rich structure information for attribute transfer to ensure high-quality person image generation.

In summary, the novelty of our proposed DRL-CPG is mainly reflected in a novel encoder with transformers and a curriculum learning with a random mask-agnostic strategy to enforce the encoder explored to learn better representation from hard examples. This strategy creates a challenging learning task that requires holistically understanding each component region given the guidance of semantic mask and transferring the learned understanding to encode components without masks as guidance. Thus it ensures our model robustly recognizes the underlying boundaries between different components and is able to localize each component. As a result, our model learning disentangled representations of attended regions works for both pose transfer and attribute transfer with well-preserved texture details and consistent component shapes. Extensive experimental evaluations strongly demonstrate that our proposed DRL-CPG yields more realistic results that are more faithful to the inputs than other state-of-the-art methods in both pose transfer and component attribute transfer.

II. RELATED WORK

Person image synthesis. Benefiting from the success in image synthesis [26], [27], [16], [17], many works have focused on synthesizing person images. PG2 [28] firstly proposed a two-stage GAN architecture [1] to generate person images. Esser *et al.* [29] leveraged a variational autoencoder combined with a conditional U-Net [30] to model the inherent shape and appearance. Siarohin *et al.* [31] used a U-Net based generator with deformable skip connections to handle the pixel-to-pixel misalignments between different poses. Zhu *et al.* [32] introduced cascaded Pose-Attentional Transfer Blocks to progressively guide the person image synthesis. [33], [34] utilized a bidirectional strategy for synthesizing person images in an unsupervised manner. SMIS [35] proposed to generate diverse person images with semantic grouping and injection. SPG [36] and PISE [22] deal with the human generation task by predicting semantic parsing masks as guidance. CASD [37] and NTED [21] introduces the attention based style distribution module that learns representative features. ADGAN [18] introduced a controllable way to synthesize person images that allow for attribute editing. However, it lacks efficient ways to learn disentangled representation for efficient person editing. Our method overcomes these challenges with a novel encoder architecture and a better training strategy.

Disentangled representation learning. Generating disentangled representations [38], [39], [12], [40], [41], [42] is essential for tasks involving attribute editing. InfoGAN [43] applies information regularization to obtain interpretable latent representations. StyleGAN [27] is able to synthesize impressive images in high-resolution by integrating adaptive instance normalization layers [24] in a new generator architecture that learns disentangled latent representations. DPIG [44] learns pose and appearance representations separately. Our model is the first attempt to learn disentangled latent semantic representations with transformers for controllable person image generation.

Visual Transformers. Transformer has achieved impressive success in object detection [45], [46] and semantic segmentation [47], [48], [49]. Transformer [19] introduces a new attention mechanism that has been successfully applied in various vision tasks [50], [51], [52]. Similar to non-local neural networks [53], [54], [19], the transformer directly works on sequences of image patches to aggregate information. ViT [55]
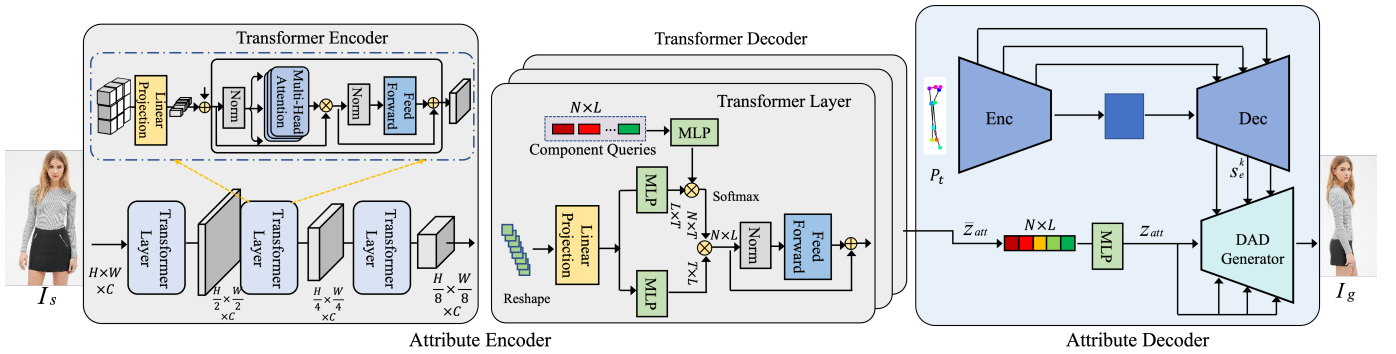
Fig. 2: The framework of our proposed DRL-CPG. It consists of an attribute encoder with transformers and an attribute decoder for pose and style transfer. The attribute encoder is trained with curriculum learning starting from a relatively easy step gradually to a complex complete mask-agnostic step, which can encode a person into latent attribute space robustly based on semantic masks. At the testing stage, the learned encoder can generate meaningful component attribute representations of a source person image. With the latent component attributes extracted, the attribute decoder with dually adaptive denormalization (DAD) ResBlocks integrates multiple attributes with the given pose to generate the desired image and achieve pose and attribute transfer.
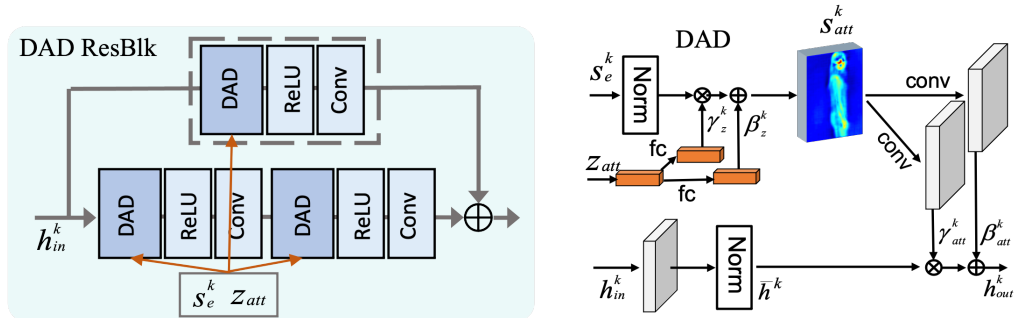

Fig. 3: Detailed views of (left) a DAD ResBlk; (right) a Dual adaptive denormalization layer.

cuts the image into small patches and globally attends all the patches at every transformer layer. The PVT [56] introduces a versatile backbone for dense prediction. Our model takes transformer layers to extract disentangled representations for controllable person generation.

Curriculum learning. Inspired by the human learning process, Bengio *et al.* [20] proposed curriculum learning which starts from a relatively easy task and gradually increases the difficulty of training. It benefits both performance improvement and speed of convergence in various deep learning tasks such as weakly supervised object detection [57], image captioning [58], [59], and video application [60]. We exploit curriculum learning by scheduling the difficulty according to our proposed random mask-agnostic strategy.

III. PROPOSED APPROACH

We propose a novel framework DRL-CPG to synthesize person images with user-controlled human attributes, such as pose, head, upper clothes, and pants. As illustrated in Figure 2, our DRL-CPG takes a collection of source person image I_s and the corresponding semantic mask M_s to provide component attributes, and a target keypoint-based pose P_t to provide the target pose attribute. The framework consists of

two components, *i.e.* , an attribute encoder with transformer trained in a manner of curriculum learning, and an attribute decoder with dual adaptive denormalization for attribute transfer.

A. Attribute Encoder with Transformers

To learn disentangled representation for controllable person generation, we propose a novel attribute encoder with transformers to generate the intermediate representation set. The attribute encoder flattens region components in source images and supplements them with a positional encoding before passing it into a transformer encoder. A transformer decoder then takes as input of the encoded feature map, and searches for matching with a small fixed number of learned component embeddings, which we call "*component queries*". The overall attribute encoder architecture consists of a transformer encoder and a transformer decoder that generates the final feature representations.

Transformer encoder. Our transformer encoder is adopted from PVT [56]. Given an input image of size $(H, W, 3)$, along with fixed positional encodings to compensate for the missing spatial information, the transformer encoder generates a $\left(\frac{H}{8}, \frac{W}{8}, C\right)$ feature map, which is then flattened into a
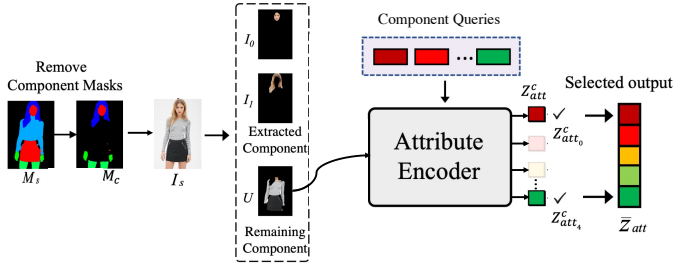
Fig. 4: Attribute encoder works at random component mask-agnostic strategy.

sequence of $(\frac{H}{8} \times \frac{W}{8}, C)$. The H, W, C refer to the height, width, and channel dimensions. For simplification, we denote the size of this flattened feature as (T, C), where $T = \frac{H}{8} \times \frac{W}{8}$. The encoder is composed of stacked transformer layers, each of which consists of a multi-head self-attention module and a feed-forward network.

Transformer decoder. The goal of our transformer decoder is to take input a set of learnable component embedding as "*component queries*", denoted by $E_c \in R^{N \times L}$, the encoded feature as key and value, denoted by $E_f \in R^{T \times C}$, and output a new component embedding. L stands for the feature dimension and N refers to the number of components. This can be formulated as below:

$$\text{Qattn}(E_c) = \text{softmax}((E_c W_\theta)^T (E_f W_\phi)), \quad (1)$$
$$Z^c_{att} = (\text{Qattn}(E_c)(E_f W_z))W_c, \quad (2)$$

where each W represents the parameters of MLP and Qattn is a query attention map used to highlight different individual components. Finally, the output $Z^c_{att} \in R^{N \times L}$ is taken as the intermediate feature with the dimension of L. We take the positioning index so that each item in Z^c_{att} learns representative information for a specific person component. This is different from most previous transformers working at the instance level. We use the transformers to learn localizable features within objects (different parts in the same instance).

B. Random Component Mask-Agnostic

In order to learn the texture and shape representation of each component, existing methods take the semantic mask to separate each component at the pixel level and let the encoder learn the representation of each separated component. This tends to learn entangled representation since the component shape is correlated with other components. For instance, the length of the uncovered arm is directly determined by the upper cloth. As a result, the original cloth item that is not completely removed causes problems in learning the latent representation.

To address this, we propose a random component mask-agnostic strategy to train the model, which truly eliminates the correlation of each component and promotes the model to learn disentangled representation. The workflow is shown in Figure 4. Given a semantic mask M_s containing N attribute regions, *e.g.*, head, upper clothes, skirt, and pants, we first randomly generate a set of indexes indicating the need to remove the component masks from M_s. After removing the component

masks, we extract components based on the remaining component masks M_c. This creates two types of components, denoted by $\{I_c, U\}$, where $I_c = M_c * I_s$ given that M_c is not removed. U is termed as the mask-agnostic component obtained as $U = I - \sum I_c$, where superscript c refers to the c-th component. Then we treat these two types of components equally and feed them into the attribute encoder. For each component, the encoder produces $Z^c_{att} \in R^{N \times L}$. We then pick out an item from Z^c_{att} as the representation of each individual component. This can be described as $Z^c_{att_j} = S_j(Z^c_{att}) \in R^{1 \times L}$, where $S_j(\cdot)$ represents the feature selection that takes the j-th row as output. Finally, we concatenate the selected representations to get $\bar{Z}_{att} = \text{concat}(Z^c_{att_j}) \in R^{N \times L}$. $\bar{Z}_{att}$ is the final latent representation of the person image I_s.

Our random component mask-agnostic strategy introduces different learning tasks in terms of recognition difficulty. If all the component masks are removed, we feed the complete person image into the attribute encoder. The attribute encoder needs to recognize each component without supervision. This is the most difficult task. If parts of the component masks are removed, the remaining component masks are used in extracting components. If all the component masks are available, each component will be separated at the image level. This is an easy task for the attribute encoder to produce representations for each extracted component. According to different levels of difficulty, we further introduce the curriculum learning strategy to train our model.

Curriculum learning. Let the number of available component masks k indicate the difficulty of a task. $k = N$ indicates an easy task where a complete semantic mask is available to separate different components. While $k = 0$ indicates a hard task where no semantic mask is available to separate each component. As we can see, the random component mask-agnostic strategy increases the difficulty in learning latent representations and thus enhances the encoder to recognize each component with or without component masks. To further alleviate the training difficulty in early steps, we adopt a curriculum learning scheme[20]. At the early training stage, our network is given fully separated components (*i.e.* $k = N$). After α epochs, we start to take in the random component mask-agnostic strategy to randomly remove component masks (*i.e.* $k < N$).

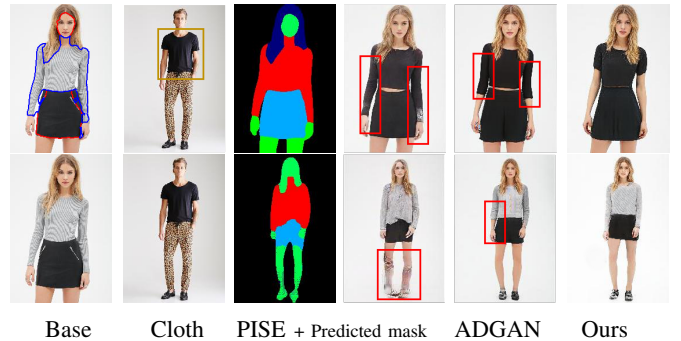Base   Cloth   PISE + Predicted mask   ADGAN   Ours

Fig. 5: Comparison between different methods. (Top) The results of component attribute transfer; (Bottom) The results of pose transfer.

Performance preview. To demonstrate how well our approach works, we compare three performances on person generation task in Figure 5. In this case, we try to transfer the upper cloth from the source person to the base. We observe that our DRL-CPG yields better person images with a short sleeve that is consistent with the source cloth, while ADGAN fails to create a consistent sleeve style and PISE creates a long sleeve style following the mask estimated by itself. The predicted problematic mask will introduce holes and artifacts into the generated person. Our model is robust to recognize the underlying boundaries between different components and is able to transfer both the shape and texture of components.

C. Attribute Decoder

The attribute decoder is to generate person images conditioned on semantic attributes and the target pose. This network consists of a pose-guided structure completion network and a dual adaptive denormalization generator. In the beginning, we take a multilayer perceptron (MLP) module to recombine the disentangled features into a more global representation Z_{att}.

Pose-guided structure completion network. The previous models generate person images conditioned on the pose, which is defined by several landmarks indicating the positions of each joint. However, this lacks details on the structure, making the person image synthesis an ill-posed problem. In order to create the structural details, we propose to generate multi-level feature maps based on the input pose. Specifically, we feed the target pose P_t into a UNet-like structure and select the intermediate feature maps S_e^k from the k_{th} level feature map of the UNet decoder. These features reflect the spatial structure indicated by the target pose. They are further utilized to guide the person image generation.

Dually adaptive denormalization generator. Our generator integrates two types of representation, Z_{att} and S_e, to generate a person image I_g. As discussed, the Z_{att} is a one-dimensional semantic representation, while S_e is a spatial feature. We propose a novel Dually Adaptive Denormalization (DAD) layer, as shown in Figure 3, to integrate them in a more adaptive fashion. Let $h_{in}^k \in R^{C^k \times H^k \times W^k}$ denote one activation map that is fed into the k_{th} layer, with C^k being the number of channels and $H^k \times W^k$ being the spatial dimensions. The output is calculated by

$$h_{out}^k = \gamma_{att}^k \otimes \frac{h_{in}^k - \mu^k}{\sigma^k} + \beta_{att}^k, \qquad (3)$$

where $\mu^k \in R^{C^k}$ and $\sigma^k \in R^{C^k}$ are the means and standard deviations of the channel-wise activations within h_{in}^k; and γ_{att}^k and β_{att}^k are two modulation parameters used to inject semantic information into the normalized activations. Typically, these two parameters are calculated from two-dimensional semantic segmentation maps. Since our spatial features S_e^k contain only structural information, we adopt another modulation operator to adaptively adjust its effective regions, and inject the semantic information learned in Z_{att} into corresponding regions. This produces a semantic feature map S_{att}^k. Formally, this can be described as

$$S_{att}^k = \gamma_z^k \otimes \frac{S_e^k - \mu_e^k}{\sigma_e^k} + \beta_z^k, \qquad (4)$$

where μ_e^k and σ_e^k are the means and standard deviations of the channel-wise activations within S_e^k; and γ_z^k and β_z^k are two modulation parameters both mapped from Z_{att} with two full connection layers. Two convolutional layers are used to generate γ_{att}^k and β_{att}^k based on S_{att}^k.

The DAD Residual Block (DAD ResBlk) is designed as a combination of "DAD+Relu+Conv" with a residual connection [61]. With the attribute representation Z_{att} and the structure feature S_e, we cascade DAD residual blocks to generate the target person I_g.

D. Loss Functions

The joint loss function is formulated with an adversarial loss $\mathcal{L}_{adv}$, a reconstruction loss $\mathcal{L}_{rec}$, a perceptual loss $\mathcal{L}_{per}$ [62], and a contextual loss $\mathcal{L}_{ctx}$ [63], [18] as

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{ctx}\mathcal{L}_{ctx} \qquad (5)$$

where we set $\lambda_{rec} = 1$, $\lambda_{per} = 5$ and $\lambda_{ctx} = 1$ in our experiments.

Adversarial loss. We employ an adversarial loss $\mathcal{L}_{adv}$ with discriminators D_p and D_t to help the generator G synthesize the target person image with visual textures similar to the reference one, as well as following the target pose. It penalizes for the distance between the distribution of real pair (P_t, I_t) and the distribution of fake pair (P_t, I_g) containing generated images

$$\mathcal{L}_{adv} = \mathbb{E}[\log(D_t(I_s, I_t)D_p(P_t, I_t))] + \mathbb{E}[\log((1 - D_t(I_s, I_g)) \\ (1 - D_p(P_t, I_g)))] \qquad (6)$$

Reconstruction loss. We define a reconstruction loss as the pixel-level L_1 distance between the target image I_t and the generated image

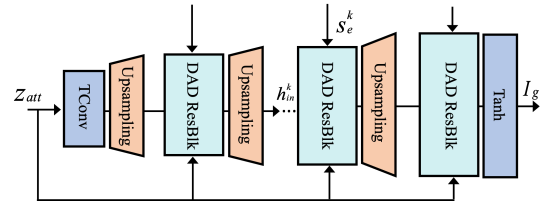$$\mathcal{L}_{rec} = ||G(I_s, P_t) - I_t||_1. \qquad (7)$$

Fig. 6: Architecture of our DAD generator. TConv refers to the TransposeConvolutional layer.

IV. Experiments

To evaluate the effectiveness of our DRL-CPG, we compete it with two types of person generation methods including (a) pose transfer method *i.e.* PG^2 [28], DPIG [44], Def-GAN [31], PATN [32] and SPG [36]; (b) both pose transfer and component attribute transfer method *i.e.* ADGAN [18] and PISE [22]. Following the data pre-processing manner in [32], [18], from the DeepFashion [64] dataset we take 101,966 pairs of images for training and 8,750 pairs for testing. We also evaluate the performance on the Market1501 dataset [65] which contains 12,936 training images and 19,732 testing images. For each image, we acquire the semantic map of a person image with the human parser [66]. All the images are with a resolution of 256×256.
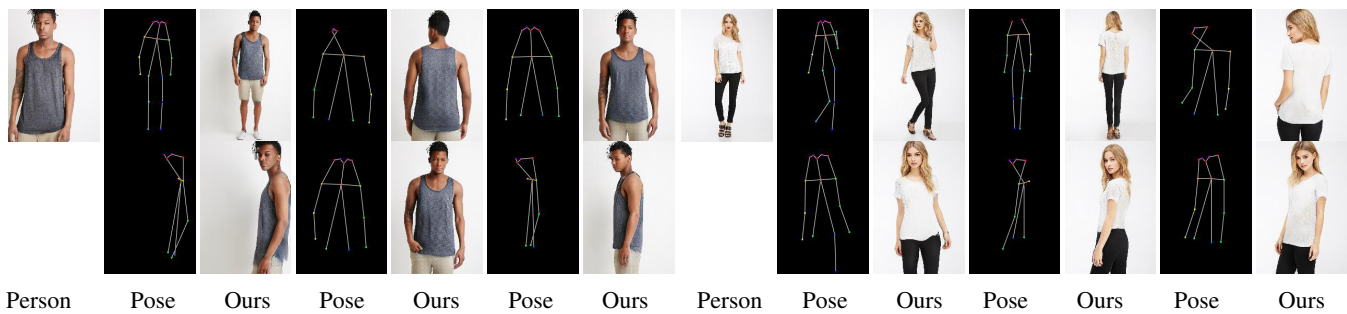
| Person | Pose | Ours | Pose | Ours | Pose | Ours | Person | Pose | Ours | Pose | Ours | Pose | Ours |

Fig. 7: Results of synthesizing person images in arbitrary poses.

| Base | GT | DPIG | Def-GAN | PATN | ADGAN | SPG | PISE | CASD | NTED | Ours |

Fig. 8: Qualitative comparison with baseline methods.

| Person | Pose | ADGAN | PISE | SPG | Ours |

Fig. 9: Swap pose transfer results.

Method	CAT	FID↓	SSIM↑	LPIPS ↓
PG^2 [28]	N	23.202	0.773	0.259
DPIG [44]	N	21.323	0.745	0.246
Def-GAN [31]	N	18.475	0.760	0.2330
PATN [32]	N	20.739	0.773	0.2533
SPG [36]	N	12.243	0.790	0.2105
ADGAN [18]	Y	14.460	0.772	0.2256
PISE [22]	Y	13.610	0.778	0.2059
CASD [37]	Y	13.939	0.768	0.2174
NTED [21]	Y	9.216	0.781	0.1961
DRL-CPG (Ours)	Y	13.514	0.792	0.2027

TABLE I: Quantitative comparison on the DeepFashion dataset. CAT is component attribute transfer.

Configuration of Our DRL-CPG Networks. In the transformer encoder, the embedding dim, number of layers, and MLP ratio are set to be 256, 4, and 2, respectively. The dimension of component attribute embedding L is set to 64. For each human image, we separate it into $N = 8$ different semantic components, *i.e.* hair, head, arm, upper cloth, leg, pant, background, and skirt. A basic Conv layer contains a 3×3 convolution operation, a normalization, and a ReLU activation sequentially. Our pose-guided structure completion network consists of 7 Conv layers with stride size 2 and 7 Deconv layers with stride size 2. Our dually adaptive denormalization generator consists of 7 up-sampling layers with stride size 2. Each of them is followed by one DAD ResBlk. The structure of our DAD generator is shown in Figure 6.

Implementation Details. We adopt Adam optimizer [67] to train our model for 100 epochs. The initial learning rate is set to 0.0001 and linearly decayed to 0 after 60 epochs. Without using a curriculum learning strategy (such as Base + MA model), we start to randomly remove component masks (*i.e.* $k < N$) from the first epoch. Under curriculum learning

Method	CAT	FID ↓	SSIM ↑	LPIPS ↓
SPG [36]	N	23.331	0.315	0.2779
ADGAN [18]	Y	26.784	0.261	0.3162
PISE [22]	Y	24.852	0.273	0.3073
DRL-CPG (Ours)	Y	22.517	0.310	0.2789

TABLE II: Quantitative comparison on the Market1501 dataset. CAT indicates component attribute transfer.

Source   Target   ADGAN   PISE   SPG   Ours

Fig. 10: Visual performance comparison on the market-1501 dataset.

Setting	User study ↑		
	Component attribute transfer	Pose transfer	Swap pose transfer
ADGAN	32.63%	20.75%	21.2%
PISE	29.01%	24.34%	27.6%
SPG	N/	28.28%	13.4%
Ours	38.36%	26.63%	37.8%

TABLE III: User study of different methods on person image generation tasks. N/ indicates that the method is not able to perform on this task.

strategy (such as Base + MA + CL model and our DRL-CPG model), our network is given fully separated components (*i.e.* $k = N$) at the early training stage. We then randomly remove component masks (*i.e.* $k < N$) after α epochs. In our experiment, we take $\alpha = 50$. At the testing stage, our DRL-CPG model takes use of all segmentation masks to generate the best person images. We also observe that our trained Base + MA + CL model can generate decent person images and handle component attribute transfer without using segmentation masks to extract components. This proves the effectiveness of our proposed random component mask-agnostic strategy in learning disentangled representation.

Evaluation metrics. We conduct several metrics to evaluate the performance of the human pose transfer task. With the source and target image pairs available, we calculate the metrics scores, including Learned Structural Similarity Index Measure (SSIM [68]), Fréchet Inception Distance (FID [69]) and Learned Perceptual Image Patch Similarity (LPIPS [70]), to compute the distance between the generated images and the corresponding ground-truth images.

A. Pose Transfer

Given a source person image and target pose extracted from another person image, the task of pose transfer is to generate a natural and realistic person in the shape of the target pose while preserving the person's identity.

1) DeepFashion Dataset: We first conduct experiments on the DeepFashion Dataset. In Figure 7, we show some results synthesized by our method. Based on the person image and the target poses, our model generates realistic images. In Figure 8, we compare the performance on the pose transfer task. As we can see, our DRL-CPG produces realistic person images

with a better cloth texture and local structure. To further demonstrate the outperformance of our approach, in Figure 9 we visualize some comparison results where the target poses are from different persons. Our DRL-CPG produces person images maintaining the attributes consistent with the source person, while PISE and SPG wrongly generate female heads when the target poses are adopted from women. Because these two methods only inject the texture code into the estimated masks that determine the final shape of the created person, one of the issues is that the estimated masks are heavily correlated with the poses and thus the generated person preserves features from the target person, such as traces of objects and shapes of components.

In Table I, we list the quantitative results. Our DRL-CPG generates the best person images in terms of SSIM score and preserves the similarity of the texture as proved by the LPIPS values; DRL-CPG achieves comparable FID scores, which indicates our method is able to preserve the shape and texture.

We shall emphasize that although the quantitative improvement in pose transfer performance of the state-of-the-art methods SPG [36] and PISE [22] is marginal, the improvements in swap pose transfer (Figure 11) and attribute transfer tasks are obvious. Our method produces person images maintaining the attributes consistent to the source/target person, while PISE and SPG struggle to generate a person of the same gender as the source person (Figure 9) or maintain the style consistency (Figure 12). Although NTED [21] produces the best FID and LIPIPS scores with the fact that it takes a coarse-to-fine generation strategy to deploy the NTED operations at different scales, in the second row of Figure 8 a bag that does not exist in the input image was created by NTED along with the female model. The similar phenomenon can also be observed in the 8_{th} column of Figure 9, where NTED additionally generated a chair. These reflect that NTED fails to completely disentangle the representation as it opts to introduce nonnegligible residues of the target image into the generated output. By contrast, due to the well-learned latent representation our method performs pose transfer with consistent semantic regions without being affected by artifacts.

2) Market1501 Dataset: We also evaluate our model on the Market1501 dataset. Note that SPG [36] is the state-of-the-art method designed for pose transfer only without learning representations for attribute transfer, while ADGAN [18], PISE [22] and our method work for both pose transfer and attribute transfer. In spite of this, our method achieves comparable performance to SPG. See Table II and Figure 10. These strongly demonstrate the superiority of our method.

| Person | Pose | Source | ADGAN | PISE | SPGNet | CASD | NTED | Ours |

Fig. 11: Performance comparison on swap pose transfer task.

B. Component Attribute Transfer

Component attribute transfer is the replacement of the attribute of a person in the source image with that of another person in the target image while preserving other attributes of the source person. We compare our proposed DRL-CPG with ADGAN and PISE which are able to edit component-level human attributes based on the corresponding semantic mask that provides guidance to separate each component at the image level, and summarize the results in Figure 12. As we can observe, our DRL-CPG generates natural images with new attributes introduced harmoniously while preserving the textures of the remaining components. In contrast, ADGAN

Fig. 12: Component attribute transfer results.

Settings	FID↓	SSIM↑	LPIPS ↓
Base	38.238	0.362	0.3537
Base + MA	15.521	0.772	0.229
Base + MA + CL	14.845	0.781	0.217
Base + MA + CL + Mask (DRL-CPG)	13.514	0.792	0.2027

TABLE IV: Ablation study on pose transfer task.

fails to create a consistent sleeve style. This can be explained by the fact that the representations learned by ADGAN are not well disentangled; PISE generates person images with new attributes but fails to change the shape of the clothes accordingly.

C. User Study

We report the results of a user study comparing our model to ADGAN, PISE and SPG on pose transfer, component attribute transfer, and swap pose transfer tasks. For the tasks, we select 200 pairs from the test subset to synthesize person images. Eight participants are required to make a choice about which output they prefer in terms of the visual quality, and more importantly, the pose and component attribute consistency with ground-truth. Note that SPG does not learn disentangled latent representation for different attribute components. This is partly the reason that it works well on pose transfer tasks but fails to work on tasks requiring disentanglement like swap pose transfer. Inherently, it does not hold the capability of conducting component attribute transfer tasks.

Although the quantitative improvement in pose transfer is marginal, the improvements in swap pose transfer and attribute transfer tasks are obvious. Our method produces person images maintaining the attributes consistent to the source/target person, while PISE and SPG struggle to generate person of the same gender as the source person and maintain the style consistency (Figure 15). Results are listed in Table III. We can see the results of ours outperform the state of the arts, which further indicates the superiority of our method.

D. Ablation Study

We ablate our training mechanism by training our model with different combinations of strategies. To evaluate the effectiveness of the training strategies assisting in learning disentangled representations, we test the trained model variants

Fig. 13: Ablation study on component attribute transfer task. The results of ADGAN and PISE are listed as references.

Fig. 14: Ablation study on loss function and resolution.

Settings	FID↓	SSIM↑	LPIPS ↓
No-TX	18.374	0.762	0.2403
Small-TX	16.386	0.775	0.2297
Medium-TX	13.514	0.792	0.2027
Large-TX	14.403	0.777	0.2170

TABLE V: Performance of Transformer (TX) at different sizes.

Settings	FID↓	SSIM↑	LPIPS ↓
Full model	13.514	0.792	0.2027
w/o λ_{ctx}	13.724	0.787	0.2070
w/o $\mathcal{L}_{per}$	13.962	0.772	0.2121

TABLE VI: Ablation study on different loss functions.

without masks. In this way, we feed the complete person image into the trained model for person generation tasks. Note our full model is tested with masks to extract the components.

Base model. We train a base model without any strategy, and test the trained model without masks.

Base + MA model. The random component mask-agnostic (MA) strategy is included to train the base model.

Base + MA + CL model. The curriculum learning (CL) strategy is further added to improve performance.

Base + MA + CL + Mask (DRL-CPG) model. Given the previous model variants are tested without masks, we take all the masks to extract components at the testing stage. This is the proposed full model.

We summarize the quantitative results in Table IV and demonstrate the qualitative comparison in Figure 13. At the testing stage, the base model fails to generate realistic images without masks. After taking the random component mask-

Fig. 15: Performance comparison in terms of fidelity.

Source Target ADGAN PISE Base+MA Ours

Fig. 16: Visualize of attribute distribution map of Base + MA + CL + Mask (DRL-CPG) model (top), Base + MA + CL model (middle) and Base model (bottom).

agnostic strategy to train the model, the trained model is able to generate images without masks at the testing stage. This shows the capability of the MA strategy in assisting the transformer encoder to recognize each component and extract meaningful representation from the complete person images. The curriculum learning strategy further improves the performance. Our full model generates the best person images by taking full use of segmentation masks at the testing stage.

Resolution. We train another model with minor modifications on images of 512×352 resolution. As shown in Figure 14, we observe that our model@512 produces faithful synthetic images with well-preserved texture details. It demonstrates that training on high-resolution images consistently learns the model to better distinguish boundaries between different semantic regions and significantly improves the generation quality.

Loss function. We further compare the effect of different loss in Figure 14 and list the quantitative scores in Table VI. As can be seen, w/o contextual loss λ_{ctx} the model fails to achieve the completeness of semantic regions, such as the cloth in the first row of Figure 14. Without making use of the perceptual loss $\mathcal{L}_{per}$, the model tends to create artifacts in the synthesized image, making the generation quality degraded to some degree.

E. Analysis and Discussion

We conduct further experiments for analysis and discussion of the effectiveness of transformers and disentangling latent
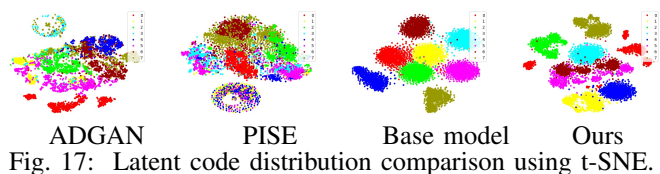

ADGAN PISE Base model Ours
Fig. 17: Latent code distribution comparison using t-SNE.

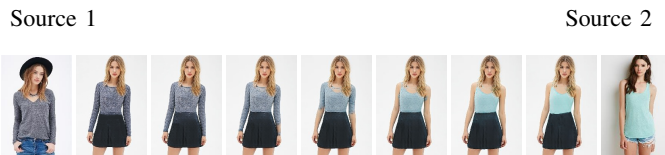Source 1   Source 2

Fig. 18: Cloth Style Interpolation between two source images.

Fig. 19: Comparison between different attribute decoders in terms of loss curve.
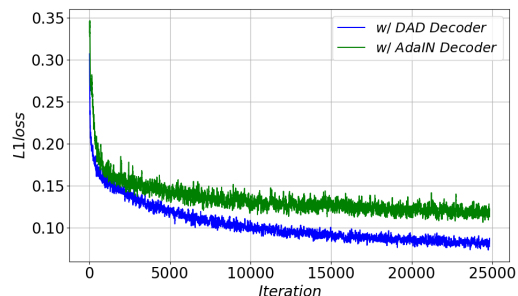

ResBlk$_0$ ResBlk$_8$

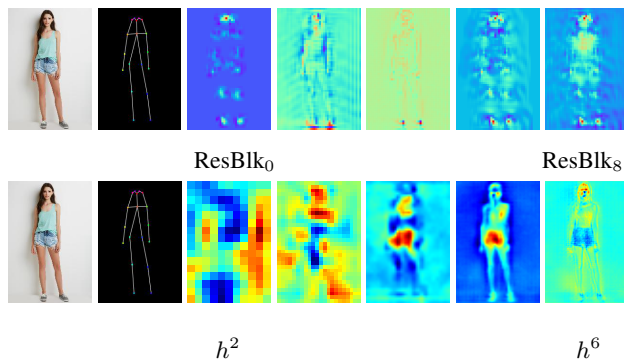h^2 h^6

Fig. 20: Left: input person image and pose; Top: feature maps in Residual Blocks of ADGAN; Bottom: feature maps in Attribute Decoder of DRL-CPG.

representation in the attribute encoder.

Effectiveness of preserving texture details. We show the comparison on the fidelity of cloth texture in Figure 15. Compared to our base+MA model, our proposed strategy enables our method to transfer component shapes and preserve the texture details missed in generated images from ADGAN and PISE.

Effectiveness of transformer encoder. We provide an ablation study on different model sizes of DRL-CPG. In particular, we investigate three different transformer configurations, the "*Small-TX*", "*Medium-TX*" and "*Large-TX*" models. For the "*Small-TX*" model, the embedding dim, number of layers and MLP ratio are set to be 128, 1, 1, respectively. While those hyperparameters for the "*Medium-TX*" model are 256, 4, 2 and those for the "*Large-TX*" model are 768, 8, 4. "*No-TX*"
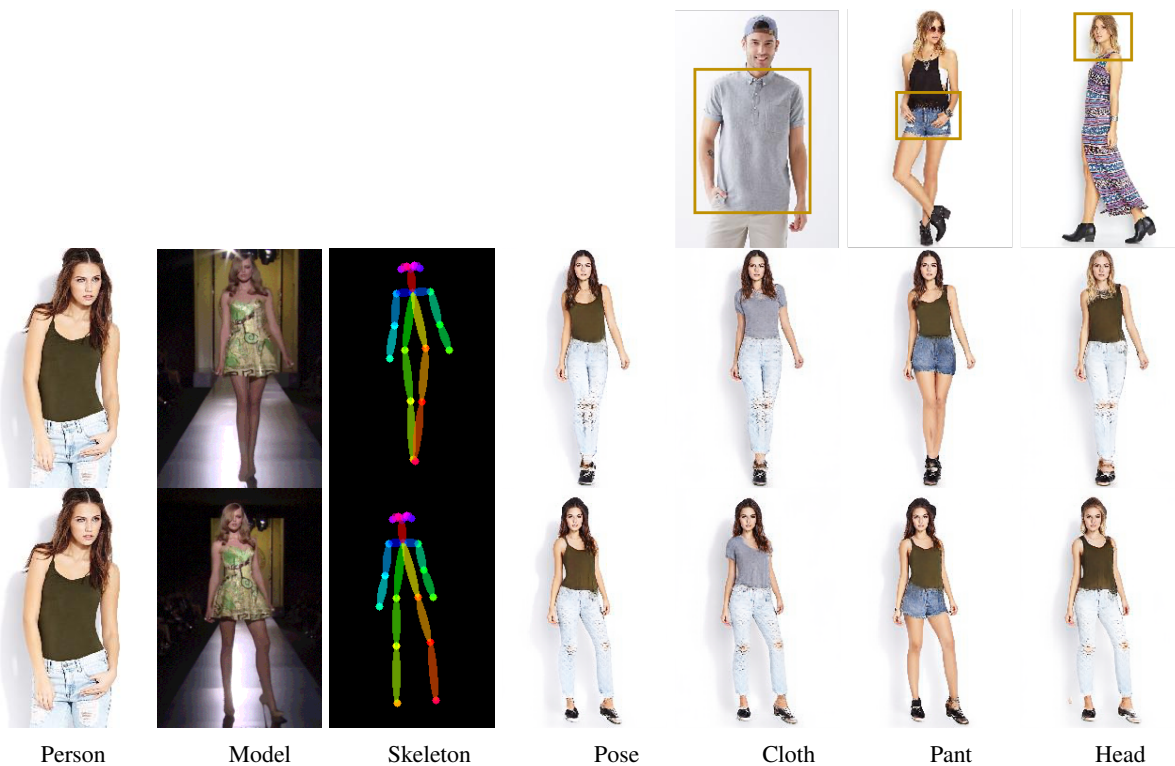
| Person | Model | Skeleton | Pose | Cloth | Pant | Head |

Fig. 21: Video demonstration of Controllable Person Synthetic Image Generation. More detainls are in the supplemental video (DRL-CPG_video.mp4).

indicates a CNN attribute encoder without transformers. As shown in Table V, we observe that with the increase of the transformer size, the performance first becomes better and then turns worse. We conclude that the medium model results in a better performance.

We further apply the trained attribute encoder in our DRL-CPG, Base model, and Base + MA + CL model to a source person image, respectively. As we can observe in Figure 16, our attribute encoder with transformers is robust in extracting main attributes from different components, regardless of whether a semantic mask is provided or not. However, the

attribute encoder in Base model can not obtain a proper attention map on different components. This model variant fails to operate controllable person synthetic image generation without a semantic mask. Again, this observation strongly demonstrates the efficacy of our well-designed DRL-CPG. As we can observe in Figure 16, our attribute encoder with transformers is robust in extracting main attributes from different components, regardless of whether a semantic mask is provided or not.

Effectiveness of disentangling latent representation. To illustrate the encoded latent space learned by different model variants, we show the feature distribution comparison in Figure 17 using t-SNE. With semantic masks available, our encoded features can form more compact and separable clusters than the ADGAN and PISE networks. Thanks to the transformer encoder, our base model trained without any strategy can project all the intermediate features to well-separated regions. However, these latent features do not learn semantic meanings as proven by the results listed in the third column of Figure 13. Therefore, although the distribution of our base model looks well separated, it cannot deal with some hard or ambiguous cases, while our DRL-CPG can handle them well. In Figure 18, we demonstrate the interpolation results. The smooth transition between two clothe styles proves our model learns a well latent structure.

Effectiveness of Attribute Decoder. To evaluate the effectiveness of our proposed DAD-based attribute decoder, we train a variant of our model, which consists of a decoder used by ADGAN. As AdaIN residual block is the main building

| Person | Source | Pose | Style | Pose + style | | |

Fig. 22: Performance comparison in terms of different tasks.

block of ADGAN's decoder, we denote this decoder as AdaIN decoder. Specifically, we take the same encoder to obtain the component attribute representations and feed them to different decoders to synthesize person images. The loss curve for the effects of our DAD-based attribute decoder during training is shown in Figure 19. We observe that our decoder achieves lower L1 loss. Our DAD module integrating semantic representations into different spatial regions outperforms the AdaIN which only focuses on the semantic representations and ignores the spatial information. Thus, our proposed DRL-CPG can synthesis better person images.

Additionally, we compare the feature maps from the residual blocks of ADGAN and the attribute decoder of our model in Figure 20. These feature maps show that our decoder establishes the relation between different joints and gradually constructs the person under guidance, while ADGAN focuses on local regions around joints. and creates unsmooth features.

V. MORE VISUALIZATION RESULTS

Controllable Person Synthetic Image Generation. For the original person image, our proposed DRL-CPG can change its component attributes (*e.g.*, upper clothes, pants, and head) with another person image providing the desired attribute. In Figure 22, we show more results on the challenging task, which requires to transfer multiple attributes into the source person. By learning the disentangled representations, our model successfully transfers multiple attributes into the person, while other baseline methods struggle with reconstructing the attributes.

Video demonstration of Controllable Person Synthetic Image Generation. To further demonstrate the capability of our model in learning disentangled representation, we perform video generation. Given a source person, a video containing a walking motion of a fashion model and desired attributes, *e.g.*, cloth, pant, hair, our model can synthesize a motion for the source person under a series of target poses extracted from the fashion model, and simultaneously sharing component attributes transferred from other person. In Figure 21 we show the synthetic person with different attributes, and list two typical poses extracted from the fashion model as reference. It is strongly recommended to watch the supplemental video (DRL-CPG_video.mp4) for the visualization, which proves the effectiveness of controllable person synthetic image generation. This demonstrates that our model learns a smooth and well-distributed latent space that is constituted of various human attributes of the person images, including pose, upper clothes, pants, head and so on.

VI. CONCLUSION

In this paper, we have proposed a novel framework with transformers to learn disentangled representation with transformers for controllable person image synthesis. Our model learns well disentangled latent representations via the proposed random component mask-agnostic strategies and is able to operate on human editing. While very promising results have been achieved in all experiments, the texture details of the synthesized images are not entirely realistic. We plan to

synthesize images in high-resolution and improve the quality of our future work.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680. 1, 2

[2] R. Huang, W. Xu, T.-Y. Lee, A. Cherian, Y. Wang, and T. Marks, "Fx-gan: Self-supervised gan learning via feature exchange," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 3194–3202. 1

[3] W. Xu, C. Long, R. Wang, and G. Wang, "Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6383–6392. 1

[4] W. Xu, C. Long, and Y. Nie, "Learning dynamic style kernels for artistic style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 083–10 092. 1

[5] Z. Zhai, J. Zhao, C. Long, W. Xu, S. He, and H. Zhao, "Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 086–22 095. 1

[6] W. Xu and D. Choi, "Direct visual-inertial odometry and mapping for unmanned vehicle," in *International Symposium on Visual Computing*. Springer, 2016, pp. 595–604. 1

[7] J. Yu, Y. Nie, C. Long, W. Xu, Q. Zhang, and G. Li, "Monte carlo denoising via auxiliary feature guided self-attention." *ACM Trans. Graph.*, vol. 40, no. 6, pp. 273–1, 2021. 1

[8] W. Xu, S. Keshmiri, and G. Wang, "Adversarially approximated autoencoder for image generation and manipulation," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2387–2396, 2019. 1

[9] Y. Yang, X. Zhang, M. Yang, and C. Deng, "Adaptive bias-aware feature generation for generalized zero-shot learning," *IEEE Transactions on Multimedia*, 2021. 1

[10] W. Xu, K. Shawn, and G. Wang, "Stacked wasserstein autoencoder," *Neurocomputing*, vol. 363, pp. 195 – 204, 2019. 1

[11] S. Karaoğlu, T. Gevers *et al.*, "Self-supervised face image manipulation by conditioning gan on face decomposition," *IEEE Transactions on Multimedia*, vol. 24, pp. 377–385, 2021. 1

[12] K. Zhang, W. Luo, L. Ma, W. Ren, and H. Li, "Disentangled feature networks for facial portrait and caricature generation," *IEEE Transactions on Multimedia*, vol. 24, pp. 1378–1388, 2021. 1, 2

[13] W. Xu, S. Keshmiri, and G. Wang, "Toward learning a unified many-to-many mapping for diverse image translation," *Pattern Recognition*, vol. 93, pp. 570 – 580, 2019. 1

[14] B. Li, Y. Zhu, Y. Wang, C.-W. Lin, B. Ghanem, and L. Shen, "Anigan: Style-guided generative adversarial networks for unsupervised anime face generation," *IEEE Transactions on Multimedia*, vol. 24, pp. 4077–4091, 2021. 1

[15] W. Xu and G. Wang, "A domain gap aware generative adversarial network for multi-domain image translation," *IEEE Transactions on Image Processing*, vol. 31, pp. 72–84, 2021. 1

[16] P. Hu, E. S.-L. Ho, and A. Munteanu, "3dbodynet: fast reconstruction of 3d animatable human body shape from a single commodity depth camera," *IEEE Transactions on Multimedia*, vol. 24, pp. 2139–2149, 2021. 1, 2

[17] H. Tang and N. Sebe, "Total generate: Cycle in cycle generative adversarial networks for generating human faces, hands, bodies, and natural scenes," *IEEE Transactions on Multimedia*, vol. 24, pp. 2963–2974, 2021. 1, 2

[18] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5084–5093. 1, 2, 5, 6, 7

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008. 1, 2

[20] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48. 2, 3, 4

[21] Y. Ren, X. Fan, G. Li, S. Liu, and T. H. Li, "Neural texture extraction and distribution for controllable person image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 535–13 544. 2, 6, 7

[22] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, "Pise: Person image synthesis and editing with decoupled gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7982–7990. 2, 5, 6, 7

[23] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346. 2

[24] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510. 2

[25] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *arXiv preprint arXiv:1610.07629*, 2016. 2

[26] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189. 2

[27] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410. 2

[28] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Advances in neural information processing systems*, 2017, pp. 406–416. 2, 5, 6

[29] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8857–8866. 2

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. 2

[31] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable gans for pose-based human image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3408–3416. 2, 5, 6

[32] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2347–2356. 2, 5, 6

[33] S. Song, W. Zhang, J. Liu, and T. Mei, "Unsupervised person image generation with semantic parsing transformation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2357–2366. 2

[34] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Unsupervised person image synthesis in arbitrary poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8620–8628. 2

[35] Z. Zhu, Z. Xu, A. You, and X. Bai, "Semantically multi-modal image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5467–5476. 2

[36] Z. Lv, X. Li, X. Li, F. Li, T. Lin, D. He, and W. Zuo, "Learning semantic person image generation by region-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 806–10 815. 2, 5, 6, 7

[37] X. Zhou, M. Yin, X. Chen, L. Sun, C. Gao, and Q. Li, "Cross attention based style distribution for controllable person image synthesis," in *European Conference on Computer Vision*. Springer, 2022, pp. 161–178. 2, 6

[38] Y. Xiao, D. Yu, X. Wang, L. Jin, G. Wang, and Q. Zhang, "Learning quality-aware representation for multi-person pose regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 2

[39] S. Wang, L. Li, Y. Ding, and X. Yu, "One-shot talking face generation from single-speaker audio-visual correlation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 2

[40] W. Xu, G. Wang, A. Sullivan, and Z. Zhang, "Towards learning affine-invariant representations via data-efficient cnns," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[41] G. Zhou, Y. Zhao, F. Guo, and W. Xu, "A smart high accuracy silicon piezoresistive pressure sensor temperature compensation system," *Sensors*, vol. 14, no. 7, pp. 12 174–12 190, 2014. 2

[42] W. Xu, Y. Wu, W. Ma, and G. Wang, "Adaptively denoising proposal collection for weakly supervised object localization," *Neural Processing Letters*, vol. 51, no. 1, pp. 993–1006, 2020. 2

[43] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, 2016, pp. 2172–2180. 2

[44] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 5, 6

[45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229. 2

[46] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020. 2

[47] E. Xie, W. Wang, W. Wang, P. Sun, H. Xu, D. Liang, and P. Luo, "Segmenting transparent object in the wild with transformer," *arXiv preprint arXiv:2101.08461*, 2021. 2

[48] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021. 2

[49] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *arXiv preprint arXiv:2012.15840*, 2020. 2

[50] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," *arXiv preprint arXiv:2012.00364*, 2020. 2

[51] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," *arXiv preprint arXiv:2102.04378*, 2021. 2

[52] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on visual transformer," *arXiv preprint arXiv:2012.12556*, 2020. 2

[53] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803. 2

[54] Y. Jung, D. Kim, S. Woo, K. Kim, S. Kim, and I. S. Kweon, "Hide-and-tell: Learning to bridge photo streams for visual storytelling," *arXiv preprint arXiv:2002.00774*, 2020. 2

[55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 2

[56] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *IEEE ICCV*, 2021. 3

[57] M. Shi and V. Ferrari, "Weakly supervised object localization using size estimates," in *European Conference on Computer Vision*. Springer, 2016, pp. 105–121. 3

[58] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 290–298. 3

[59] X. Dong, C. Long, W. Xu, and C. Xiao, "Dual graph convolutional networks with transformer and curriculum learning for image captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2615–2624. 3

[60] J. Xiao, L. Li, D. Xu, C. Long, J. Shao, S. Zhang, S. Pu, and Y. Zhuang, "Explore video clip order with self-supervised and curriculum learning for video applications," *IEEE Transactions on Multimedia*, 2020. 3

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 5

[62] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711. 5

[63] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 768–783. 5

[64] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104. 5

[65] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Computer Vision, IEEE International Conference on*, 2015. 5

[66] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for

human parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 932–940. 5

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 6

[68] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 7

[69] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6629–6640. 7

[70] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018. 7