Exploring Dual-task Correlation for Pose Guided Person Image Generation

Pengze Zhang[1], Lingxiao Yang[1], Jianhuang Lai[1,2,3] and Xiaohua Xie[1,2,3*]
[1]School of Computer Science and Engineering, Sun Yat-Sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
zhangpz3@mail2.sysu.edu.cn, {yanglx9, stsljh, xiexiaoh6}@mail.sysu.edu.cn

Abstract

Pose Guided Person Image Generation (PGPIG) is the task of transforming a person image from the source pose to a given target pose. Most of the existing methods only focus on the ill-posed source-to-target task and fail to capture reasonable texture mapping. To address this problem, we propose a novel Dual-task Pose Transformer Network (DPTN), which introduces an auxiliary task (i.e., source-to-source task) and exploits the dual-task correlation to promote the performance of PGPIG. The DPTN is of a Siamese structure, containing a source-to-source self-reconstruction branch, and a transformation branch for source-to-target generation. By sharing partial weights between them, the knowledge learned by the source-to-source task can effectively assist the source-to-target learning. Furthermore, we bridge the two branches with a proposed Pose Transformer Module (PTM) to adaptively explore the correlation between features from dual tasks. Such correlation can establish the fine-grained mapping of all the pixels between the sources and the targets, and promote the source texture transmission to enhance the details of the generated target images. Extensive experiments show that our DPTN outperforms state-of-the-arts in terms of both PSNR and LPIPS. In addition, our DPTN only contains 9.79 million parameters, which is significantly smaller than other approaches. Our code is available at: https://github.com/PangzeCheung/Dual-task-Pose-Transformer-Network.

1. Introduction

Pose Guided Person Image Generation (PGPIG) aims to generate person images with arbitrary given poses. It has various applications such as e-commerce, film special effects, person re-identification [5–7, 19, 34, 35, 40, 41], etc. Due to the significant changes in texture and geometry during the pose transfer, PGPIG is still a challenging task.

Figure 1. Visual comparison of our method with other approaches, including vanilla CNN based [22], attention based [45], optical flow based [24], and parsing map based [21] method. Compared with other methods, our model can generate more realistic images.

Driven by the improvement of generative models, e.g., Generative Adversarial Networks (GANs) [8] and Variational Autoencoders (VAEs) [17], PGPIG has made great progress. However, early works [4, 22] are built on vanilla Convolutional Neural Networks (CNNs), which lack the capability to perform complex geometry transformations [13] (see Fig. 1 (c)). To tackle this problem, attention mechanisms [30, 45] and optical flow [18, 24, 29] are applied to improve spatial transformation abilities. Some methods [21, 39] introduce additional labels such as human parsing maps to provide semantic guidance for pose variations. However, the above mentioned methods solely focus on training the generator G on the **Source-to-Target Task** that transforms the source image $\boldsymbol{x_s}$ from the source pose $\boldsymbol{p_s}$ to the target pose $\boldsymbol{p_t}$: $G(\boldsymbol{x_s}, \boldsymbol{p_s}, \boldsymbol{p_t}) = \tilde{\boldsymbol{x}}_{\boldsymbol{t}}$. This is an ill-posed problem, making it arduous to train a robust generator. Moreover, the existing methods cannot well capture the reasonable texture mapping between the source and target

*Corresponding Author

images, especially when the person undergoes large pose changes. Therefore, those methods often produce unrealistic images, as shown in Fig. 1 (d)-(f).

In this paper, we seek to utilize an auxiliary task [26] to improve the ill-posed source-to-target transformation. Here, we instantiate the auxiliary task as the **Source-to-Source Task**, which reconstructs the source image guided by source pose: $G(\boldsymbol{x_s}, \boldsymbol{p_s}, \boldsymbol{p_s}) = \tilde{\boldsymbol{x}}_{\boldsymbol{s}}$. We observe that simultaneously learning the dual tasks (i.e., source-to-target task and source-to-source task) has the following two benefits: (1) Compared with the source-to-target task, the pixel-aligned source-to-source task is easier to learn because it does not require complex spatial transformations. By sharing weights between the dual tasks, the source-to-source task can not only exploit its knowledge to assist the source-to-target task, but also stabilize the training of the whole network. (2) Since the intermediate features in dual tasks are associated with their generated images $\tilde{\boldsymbol{x}}_{\boldsymbol{s}}$ and $\tilde{\boldsymbol{x}}_{\boldsymbol{t}}$ respectively, we can further explore the correlation between the dual tasks to establish the texture transformation from the sources to the targets. In this way, the natural source textures can be readily disseminated to enhance the details of the generated target image.

Based on these ideas, we propose a novel Dual-task Pose Transformer Network (DPTN) for PGPIG. The architecture of DPTN is shown in Fig. 2. Specifically, our DPTN is of a Siamese structure, incorporating two branches: a self-reconstruction branch for the auxiliary source-to-source task and a transformation branch for the source-to-target task. These two branches share partial weights, and are trained simultaneously with different loss functions. By this means, the knowledge learned by the source-to-source task can directly assist the optimization of the source-to-target task. To explore the correlation between the dual tasks, we bridge the two branches with a novel Pose Transformer Module (PTM). Our PTM consists of several Context Augment Blocks (CABs) and Texture Transfer Blocks (TTBs). CABs first selectively gather the information of the source-to-source task. Then TTBs gradually capture the fine-grained correlation between the features from the dual tasks. With the help of such correlation, TTBs can productively promote the texture transmission from the real source image to the source-to-target task, enabling the synthetic image to preserve more source appearance details. (see Fig. 1 (g)). In sum, the main contributions are:

- We propose a novel Dual-task Pose Transformer Network (DPTN), which introduces an auxiliary task (i.e., source-to-source task) by Siamese architecture and exploits its knowledge to improve the PGPIG.
- We design a Pose Transformer Module (PTM) to explore the dual-task correlation. Such correlation can not only establish the fine-grained mapping between the sources and the targets, but also effectively guide the source texture transmission to further refine the feature in the source-to-target task.
- Results on the two benchmarks, i.e., DeepFashion [20] and Market-1501 [43], have demonstrated that our method exhibits superior performance on PSNR and LPIPS [42]. Moreover, our model only contains 9.79 million parameters, which is relatively 91.6% smaller than the state-of-the-art method SPIG [21].

2. Related Works

Pose guided person image generation. Ma *et al.* [22] generated the fake image in a coarse-to-fine manner. Esser *et al.* [4] combined the VAE and U-net [25] to disentangle pose and appearance of the person. However, these methods are based on vanilla CNNs, which cannot handle the complex deformation. To address this problem, Zhu *et al.* [45] proposed a Pose Attention Transfer Network (PATN) to optimize the appearance by pose relation. Furthermore, Tang *et al.* [30] added more crossing ways between the pose and appearance into PATN. Nevertheless, these attention based methods do not explicitly learn the spatial transformation between different poses, losing many source textures.

To boost the texture transformation, Li *et al.* [18], Ren *et al.* [24] and Tabejamaat *et al.* [29] proposed to introduce the warping operations to PGPIG. They first estimated the dense optical flow, and then generated images by warping the source image feature. Nevertheless, under the large pose change and occlusion, these methods tended to produce inaccurate optical flow, resulting in unsatisfied images.

Besides, both Zhang *et al.* [39] and Lv *et al.* [21] utilized additional human parsing labels to improve the PGPIG. They first predicted the target parsing maps, and then output person images with the help of semantic information. However, the target parsing maps estimated by these methods are often unreliable, which will mislead the generation of the synthetic images. Moreover, pixel-wise annotations are hard to collect, which limits their applications.

In summary, all the above methods only focus on the source-to-target task, and cannot accurately capture the texture mapping between the source and the target images. Contrary to them, we show that introducing the auxiliary source-to-source task through a Siamese structure and simultaneously exploring the dual-task correlation can further improve the performance of PGPIG.

Dual-task learning. Dual-task learning is a popular learning framework for Natural Language Processing (NLP) [9, 36, 37], which utilizes the different tasks to improve the learning progress. For example, [9] leveraged the closed-loop of the English-to-French translation and French-to-English translation to enhance each other, making it possible to train translation models without paired

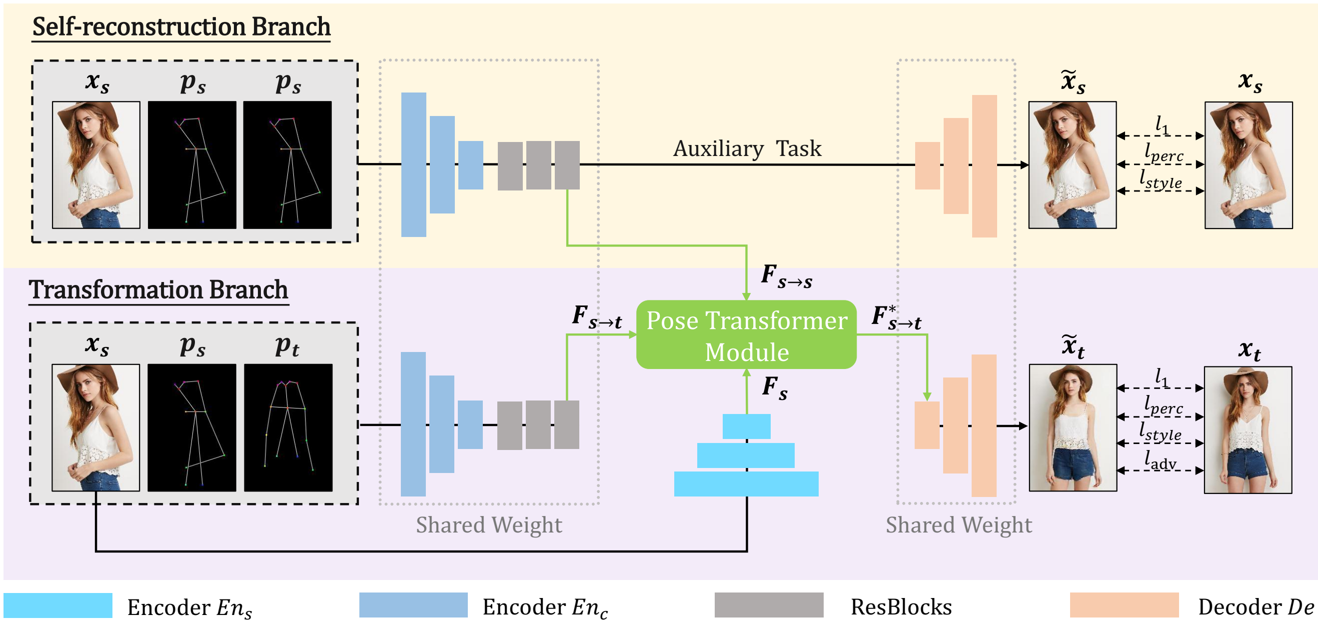

Figure 2. Overview of our model. It contains a self-reconstruction branch for auxiliary source-to-source task, and a transformation branch for source-to-target task. These two branches share partial weights and are communicated by a pose transformer module.

data. Different from these methods, our dual tasks refer to the source-to-source task and the source-to-target task. We have verified that the source-to-source learning can promote the training of the source-to-target task in PGPIG.

Transformers in vision tasks. Inspired by the success of transformers [33] in NLP, many researchers had applied transformer architecture to computer vision tasks such as image recognition [3, 31], object detection [2, 44], and image generation [12, 14]. Specially, for image generation tasks, Jiang *et al.* [14] built a GAN with a pure transformer-based architecture without convolutions. Hudson *et al.* [12] proposed a GANformer to exchange information between image features and latent variables. However, these GANs were designed for unconditional generation tasks, and were not well suitable for conditional generation tasks with complex space deformation (i.e., PGPIG). In this work, inspired by the core idea of the transformer, we design a novel pose transformer module to explore the dual-task correlation.

3. Our Approach

Fig. 2 shows the overall framework of our DPTN. It mainly contains Siamese branches for the dual tasks and a pose transformer module for exploring the dual-task correlation. In the following sections, we will describe each component of DPTN and loss functions in detail.

3.1. Siamese Structure for Dual Tasks

Although the existing PGPIG methods attempt to learn the source-to-target transformation through various approaches, we argue that these methods ignore some essential knowledge without the source-to-source learning, thus limiting their potential improvement. To demonstrate this, we conduct an experiment on a basic network (same structure as the self-reconstruction branch in Fig. 2, including En_c, ResBlocks and De) to explore the impact of the source-to-source learning, and show the results tested on the source-to-source task in Tab. 1. Compared with source-to-target learning, the + source-to-source learning in Tab. 1 only adds the self-reconstruction training, and does not change the basic network structure. It can be seen that there is a significant gap between these two learning schemes. Solely learning from the source-to-target task cannot well reconstruct the source images, and lacks some knowledge of PGPIG. Based on our analysis, in this paper, we add the source-to-source task into PGPIG, and explore its knowledge to assist the source-to-target transformation in the training process.

Table 1. The comparisons of the basic network on whether using the source-to-source learning. Both of the following two results are tested on the source-to-source task.

Learning scheme	PSNR ↑	LPIPS ↓
Source-to-target learning	19.1855	0.1962
+ Source-to-source learning	23.7606	0.1468

To achieve this goal, we construct our DPTN with a Siamese architecture, incorporating two branches: a self-reconstruction branch for source-to-source reconstruction,

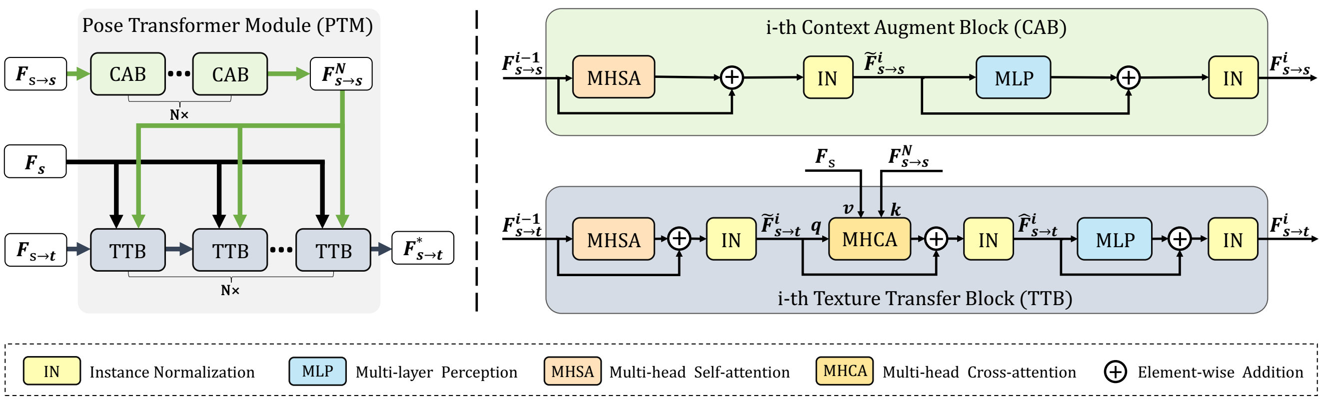

Figure 3. The structure of the Pose Transformer Module (PTM). It contains two types of blocks: Context Augment Block (CAB) and Texture Transfer Block (TTB). The CABs integrate the information of the feature $\boldsymbol{F_{s\to s}}$, while the TTBs transfer the real source image textures from $\boldsymbol{F_s}$ to optimize $\boldsymbol{F_{s\to t}}$ by capturing the correlation between features from the dual tasks.

and a transformation branch for source-to-target generation. As shown in Fig. 2, the two branches share three parts: an encoder En_c, a series of ResBlocks, and a decoder De. In more detail, the encoder first extracts the feature of two types of inputs, including the source-to-target input (the concatenation of $\boldsymbol{x_s}$, $\boldsymbol{p_s}$ and $\boldsymbol{p_t}$) and the source-to-source input (the concatenation of $\boldsymbol{x_s}$, $\boldsymbol{p_s}$ and $\boldsymbol{p_s}$). Then, ResBlocks are applied to gradually perform pose transformation. Outputs of ResBlocks are the feature $\boldsymbol{F_{s\to s}}$ aligned with the source pose, and the transformed feature $\boldsymbol{F_{s\to t}}$ aligned with the target pose. Finally, the De in the self-reconstruction branch accepts the $\boldsymbol{F_{s\to s}}$ to generate fake source image $\widetilde{\boldsymbol{x}}_{\boldsymbol{s}}$, and the De in the transformation branch accepts the refined feature $\boldsymbol{F^*_{s\to t}}$ (output of our pose transformer module) to produce the target generated image $\widetilde{\boldsymbol{x}}_{\boldsymbol{t}}$.

In conclusion, the proposed Siamese architecture has the following advantages: (1) Our encoder, ResBlocks and decoder are shared by the dual tasks, so that the learned knowledge can be easily transferred between these tasks. (2) Introducing self-reconstruction branch does not significantly add extra parameters, as most of our model are reused in different tasks. (3) The Siamese architecture enables the intermediate outputs of the dual tasks close in feature distribution, facilitating the PTM in the next section to explore the dual-task correlation.

3.2. Pose Transformer Module

In our Siamese structure, we have already obtained feature $\boldsymbol{F_{s\to s}}$ aligned with the source pose $\boldsymbol{p_s}$, and $\boldsymbol{F_{s\to t}}$ aligned with the target pose $\boldsymbol{p_t}$ respectively. However, since the vanilla CNN based transformation branch (i.e., source-to-target) is hard to handle complex space deformation, $\boldsymbol{F_{s\to t}}$ tends to lose many source appearance details, as shown in Fig. 7. To tackle this problem, we propose a novel Pose Transformer Module (PTM), which can further refine $\boldsymbol{F_{s\to t}}$ via capturing the pixel-wise source-to-target correspondence between the features from dual tasks. Our PTM is built upon the Multi-Head Attention (MHA) mechanism. To be self-contained, we briefly introduce MHA as follows:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = softmax(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d_k})\boldsymbol{V}, \quad (1)$$

$$\boldsymbol{head_i} = Attention(\boldsymbol{Q}\boldsymbol{W^i_q}, \boldsymbol{K}\boldsymbol{W^i_k}, \boldsymbol{V}\boldsymbol{W^i_v}), \quad (2)$$

$$MHA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = concat(\boldsymbol{head_1}, ..., \boldsymbol{head_h}). \quad (3)$$

The $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ are queries, keys and values. $\boldsymbol{W^i_q}$, $\boldsymbol{W^i_k}$, $\boldsymbol{W^i_v}$ are learnable parameters. h is the number of attention heads. d_k is the dimension of the keys. In particular, when $\boldsymbol{Q} = \boldsymbol{K}$, the MHA functions as Multi-Head Self-Attention (MHSA); otherwise it acts as Multi-Head Cross-Attention (MHCA).

The proposed PTM is shown in Fig. 3. Unlike traditional vision transformer [3], our PTM adopts a new architecture to explore the relation among triple features (i.e., feature from source-to-source task, source-to-target task and source image texture), making it more suitable for PGPIG. In general, PTM contains two types of blocks: Context Augment Block (CAB) and Texture Transfer Block (TTB), which can be formulated as: $\boldsymbol{F^N_{s\to s}} = CAB(...CAB(\boldsymbol{F_{s\to s}})...)$, $\boldsymbol{F^*_{s\to t}} = TTB(...TTB(\boldsymbol{F_{s\to t}}, \boldsymbol{F^N_{s\to s}}, \boldsymbol{F_s})..., \boldsymbol{F^N_{s\to s}}, \boldsymbol{F_s})$. The superscript denotes the index of the feature. N is the number of the blocks. $\boldsymbol{F_{s\to s}} = \boldsymbol{F^0_{s\to s}}$, $\boldsymbol{F_{s\to t}} = \boldsymbol{F^0_{s\to t}}$, and $\boldsymbol{F^*_{s\to t}} = \boldsymbol{F^N_{s\to t}}$. $\boldsymbol{F_s}$ is the source image feature obtained by an additional encoder En_s. In the proposed PTM, the CABs gradually integrate the information of the feature $\boldsymbol{F_{s\to s}}$ from the self-reconstruction branch and produce $\boldsymbol{F^N_{s\to s}}$. Then, each TTB combines three kinds of features: the source image texture feature $\boldsymbol{F_s}$, the integrated source-to-source feature $\boldsymbol{F_{s\to s}}$, and the previous TTB output $\boldsymbol{F_{s\to t}}$. This combination is achieved by an MHCA module to capture the correlation among all inputs. Next, we will present the structure of the CAB and TTB respectively.

3.2.1 Context Augment Block

The structure of the i-th CAB is shown in the right-top of Fig. 3. It first applies an MHSA unit with residual connection to adaptively enhance the contextual representation of the input feature $\boldsymbol{F}^{i-1}_{s\to s}$:

$$\tilde{\boldsymbol{F}}^{i}_{s\to s} = IN(\boldsymbol{F}^{i-1}_{s\to s} + MHSA(\boldsymbol{F}^{i-1}_{s\to s}, \boldsymbol{F}^{i-1}_{s\to s}, \boldsymbol{F}^{i-1}_{s\to s})), \tag{4}$$

where IN is the instance normalization [32]. Then a Multi-Layer Perceptron (MLP) module with multiple fully connected layers is used to increase the capacity in CAB:

$$\boldsymbol{F}^{i}_{s\to s} = IN(\tilde{\boldsymbol{F}}^{i}_{s\to s} + MLP(\tilde{\boldsymbol{F}}^{i}_{s\to s})). \tag{5}$$

After N CABs, we obtain the final refined feature $\boldsymbol{F}^{N}_{s\to s}$ and add this feature into each TTB for source-to-target task.

3.2.2 Texture Transfer Block

The structure of the i-th TTB is shown in the right-bottom of Fig. 3. First, MHSA is applied to selectively focus on the key information of the transformation branch feature $\boldsymbol{F}^{i-1}_{s\to t}$:

$$\tilde{\boldsymbol{F}}^{i}_{s\to t} = IN(\boldsymbol{F}^{i-1}_{s\to t} + MHSA(\boldsymbol{F}^{i-1}_{s\to t}, \boldsymbol{F}^{i-1}_{s\to t}, \boldsymbol{F}^{i-1}_{s\to t})). \tag{6}$$

Then an MHCA unit is employed to build correlation of $\tilde{\boldsymbol{F}}^{i}_{s\to t}$, $\boldsymbol{F}^{N}_{s\to s}$, and $\boldsymbol{F}_{s}$. Specifically, we employ $\tilde{\boldsymbol{F}}^{i}_{s\to t}$ as queries and $\boldsymbol{F}^{N}_{s\to s}$ as keys to calculate the pixel-wise similarity between the sources and the targets. With the aid of such similarity, $\boldsymbol{F}_{s}$ is used as values in MHCA to transmit the real source textures to refine $\tilde{\boldsymbol{F}}^{i}_{s\to t}$. This procedure can be written as:

$$\hat{\boldsymbol{F}}^{i}_{s\to t} = IN(\tilde{\boldsymbol{F}}^{i}_{s\to t} + MHCA(\tilde{\boldsymbol{F}}^{i}_{s\to t}, \boldsymbol{F}^{N}_{s\to s}, \boldsymbol{F}_{s})). \tag{7}$$

In this way, $\hat{\boldsymbol{F}}^{i}_{s\to t}$ carries more real source textures, which will foster the transformation branch to generate more delicate patterns. Finally, similar to CAB, the i-th TTB output $\boldsymbol{F}^{i}_{s\to t}$ is obtained as follows:

$$\boldsymbol{F}^{i}_{s\to t} = IN(\hat{\boldsymbol{F}}^{i}_{s\to t} + MLP(\hat{\boldsymbol{F}}^{i}_{s\to t})). \tag{8}$$

After N time TTB blocks, the final output feature $\boldsymbol{F}^{N}_{s\to t}$ will be fed into the decoder De to generate target image $\tilde{\boldsymbol{x}}_{t}$.

3.3. Loss Functions

Our network contains two branches for the source-to-source task and source-to-target task. Thus, the overall loss function can be simply formulated as:

$$\mathcal{L} = \mathcal{L}_{s\to s} + \mathcal{L}_{s\to t}, \tag{9}$$

where $\mathcal{L}_{s\to s}$ and $\mathcal{L}_{s\to t}$ stand for the loss of the dual tasks respectively. Both of them contain an l_1 loss $\mathcal{L}_{l_1}$, a perceptual loss $\mathcal{L}_{perc}$ and a style loss $\mathcal{L}_{style}$. In addition, we apply an additional adversarial loss $\mathcal{L}_{adv}$ in the source-to-target task to produce more realistic textures. In sum, $\mathcal{L}_{s\to s}$ and $\mathcal{L}_{s\to t}$ can be written as:

$$\mathcal{L}_{s\to s} = \lambda_{l_1}\mathcal{L}^{s}_{l_1} + \lambda_{perc}\mathcal{L}^{s}_{perc} + \lambda_{style}\mathcal{L}^{s}_{style}, \tag{10}$$

$$\mathcal{L}_{s\to t} = \lambda_{l_1}\mathcal{L}^{t}_{l_1} + \lambda_{perc}\mathcal{L}^{t}_{perc} + \lambda_{style}\mathcal{L}^{t}_{style} + \lambda_{adv}\mathcal{L}_{adv}, \tag{11}$$

where λ_{l_1}, λ_{perc}, λ_{style} and λ_{adv} are the loss weights for the dual tasks. Specifically, the l_1 loss penalizes the l_1 distance between the generated image and the ground truth:

$$\mathcal{L}^{d}_{l_1} = \|\boldsymbol{x_d} - \tilde{\boldsymbol{x}}_{\boldsymbol{d}}\|_1, \tag{12}$$

where $d \in \{s, t\}$ represents the source or the target data. The perceptual loss [15] calculates the feature distance:

$$\mathcal{L}^{d}_{perc} = \sum_{i} \|\phi_i(\boldsymbol{x_d}) - \phi_i(\tilde{\boldsymbol{x}}_{\boldsymbol{d}})\|_1, \tag{13}$$

where ϕ_i denotes the i-th feature from VGG network [28]. The style loss [15] compares the style similarity between images:

$$\mathcal{L}^{d}_{style} = \sum_{j} \left\|Gram^{\phi}_{j}(\boldsymbol{x_d}) - Gram^{\phi}_{j}(\tilde{\boldsymbol{x}}_{\boldsymbol{d}})\right\|_1, \tag{14}$$

where $Gram^{\phi}_{j}$ is the Gram matrix of feature ϕ_j. Finally, the adversarial loss with a discriminator D is employed to penalize the distribution difference between the generated target image $\tilde{\boldsymbol{x}}_{\boldsymbol{t}}$ and the ground truth $\boldsymbol{x_t}$:

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D(\tilde{\boldsymbol{x}}_{\boldsymbol{t}}))] + \mathbb{E}[\log D(\boldsymbol{x_t})]. \tag{15}$$

4. Experiments

4.1. Implementation Details

We evaluate our proposed model on two datasets: DeepFashion [20] and Market1501 [43]. The DeepFashion dataset contains 52,712 high quality in-shop clothes images (256×176) with clean backgrounds, while the Market1501 dataset contains 32,668 low-resolution images (128×64) with various illumination and viewpoints. For a fair comparison, we split the datasets with the same setting as [45]. It collects 101,966 training pairs and 8,570 testing pairs for DeepFashion, and 263,632 training pairs and 12,000 testing pairs for Market1501. In addition, the human pose keypoints are extracted from Human Pose Estimator (HPE) [1].

In our experiment, Adam optimizer [16] is adopted to train the proposed DPTN with the learning rate 1e-4. We choose $h = 2$ and $N = 2$ in the PTM on both datasets. For the loss functions in Eq. (10) and Eq. (11), we set $\lambda_{l_1} = 2.5$, $\lambda_{perc} = 0.25$, $\lambda_{style} = 250$, $\lambda_{adv} = 2$.

Table 2. Quantitative comparisons of image quality and model size with several state-of-the-art methods. * denotes the method using additional human parsing labels. The best and second best results are shown in bold and underline respectively.

Model	DeepFashion				Market1501				Number of
	SSIM ↑	PSNR ↑	FID ↓	LPIPS ↓	SSIM ↑	PSNR ↑	FID ↓	LPIPS ↓	Parameters ↓
PG2 [22] (NeurIPS'17)	0.7730	17.5324	49.5674	0.2928	0.2704	14.1749	86.0288	0.3619	437.09 M
VU-net [4] (CVPR'18)	0.7639	17.6582	15.5747	0.2415	0.2665	14.4220	44.2743	0.3285	139.36 M
DSC [27] (CVPR'18)	0.7682	18.0990	21.2686	0.2440	<u>0.3054</u>	14.3081	27.0118	0.3029	82.08 M
PATN [45] (CVPR'19)	0.7717	18.2543	20.7500	0.2536	0.2818	14.2622	22.6814	0.3194	41.36 M
DIAF [18] (CVPR'19)	0.7738	16.9004	14.8825	0.2388	0.3052	14.2011	32.8787	0.3059	49.58 M
DIST [24] (CVPR'20)	0.7677	18.5737	**10.8429**	0.2258	0.2808	14.3368	<u>19.7403</u>	0.2815	<u>14.04 M</u>
XingGAN [30] (ECCV'20)	0.7706	17.9226	39.3194	0.2928	0.3044	14.4458	22.5198	0.3058	42.77 M
PISE* [39] (CVPR'21)	0.7682	18.5208	11.5144	<u>0.2080</u>	—	—	—	—	64.01 M
SPIG* [21] (CVPR'21)	<u>0.7758</u>	<u>18.5867</u>	12.7027	0.2102	**0.3139**	<u>14.4894</u>	23.0573	<u>0.2777</u>	117.13 M
Ours	**0.7782**	**19.1492**	<u>11.4664</u>	**0.1957**	0.2854	**14.5207**	**18.9946**	**0.2711**	**9.79** M

4.2. Metrics

Following previous works [21, 45], we adopt Structural Similarity Index Measure (SSIM) [38], Peak Signal-to-Noise Ratio (PSNR), Fréchet Inception Distance (FID) [11] and Learned Perceptual Image Patch Similarity (LPIPS) [42] as evaluation metrics. Moreover, we use rank-k and Mean Average Precision (MAP) to further test the texture consistency between the source images and the generated target images through the state-of-the-art re-identification (re-id) platform FastReID [10]. More precisely, we train a re-id model on the training set. Then we use the generated images as the query set and the real images as the gallery set to calculate the metrics. High rank-k and MAP indicate that the generated images do not lose much source appearance, and can be easily recognized by the current re-id system.

4.3. Comparison with Previous Work

4.3.1 Quantitative Comparison

We compare our method with several state-of-the-art methods, including PG2 [22], VU-net [4], DSC [27], PATN [45], DIAF [18], DIST [24], XingGAN [30], PISE [39] and SPIG [21]. Tab. 2 shows the quantitative results on image quality and model size. As one can see, our method achieves seven best and one second-best results among all compared methods, including PISE and SPIG using additional parsing labels. This verifies the superiority of our DPTN in generating high quality images. In addition, our DPTN only contains 9.79 M parameters, which is 91.6% lower than that of SPIG (117.13 M). It clearly demonstrates the efficiency of our method in modeling pose transformations.

Tab. 3 provides the comparison of the texture consistency on DeepFashion. First, we train the re-id system on the DeepFashion training set. As shown in the last row of Tab. 3, this re-id system achieves 99.08% rank-1 score. Then, the same re-id system is applied to identify the person in the fake images generated by different methods. From the result, we can find that our method surpasses others in all four metrics. In particular, we promote the best rank-1 performance of previous works by 3%. This indicates that the images generated by our DPTN can effectively maintain the discriminative texture of the source person.

Table 3. Quantitative comparisons of texture consistency. The best results are shown in bold.

Methods	rank-1 ↑	rank-5 ↑	rank-10 ↑	MAP ↑
PG2	60.12%	75.44%	81.95%	59.20%
VU-net	73.49%	87.49%	91.97%	72.01%
DSC	94.17%	98.19%	99.08%	90.40%
PATN	74.35%	87.95%	92.37%	73.17%
DIAF	94.87%	98.02%	99.13%	91.45%
DIST	90.84%	96.64%	98.11%	87.56%
XingGAN	59.63%	72.48%	81.19%	58.36%
PISE	90.09%	96.35%	98.02%	87.22%
SPIG	94.43%	98.23%	99.04%	91.60%
Ours	**97.69%**	**99.35%**	**99.63%**	**95.04%**
Real Data	99.08%	99.80%	99.88%	98.40%

4.3.2 Qualitative Comparison

The qualitative comparison results are shown in Fig. 4. For the DeepFashion dataset, the attention based methods PATN and XingGAN tend to generate blurred images (see 1st and 2nd rows). DIAF and DIST attempt to promote the texture transfer by using optical flow. However, in the case of large pose changes, their predicted optical flow fails to represent such complex deformation, resulting in unacceptable results (see 2nd and 3rd rows). PISE and SPIG introduce the additional semantic parsing map to ease the difficulty of PGPIG. Nevertheless, the target parsing maps estimated by these methods are often inaccurate, which will mislead the

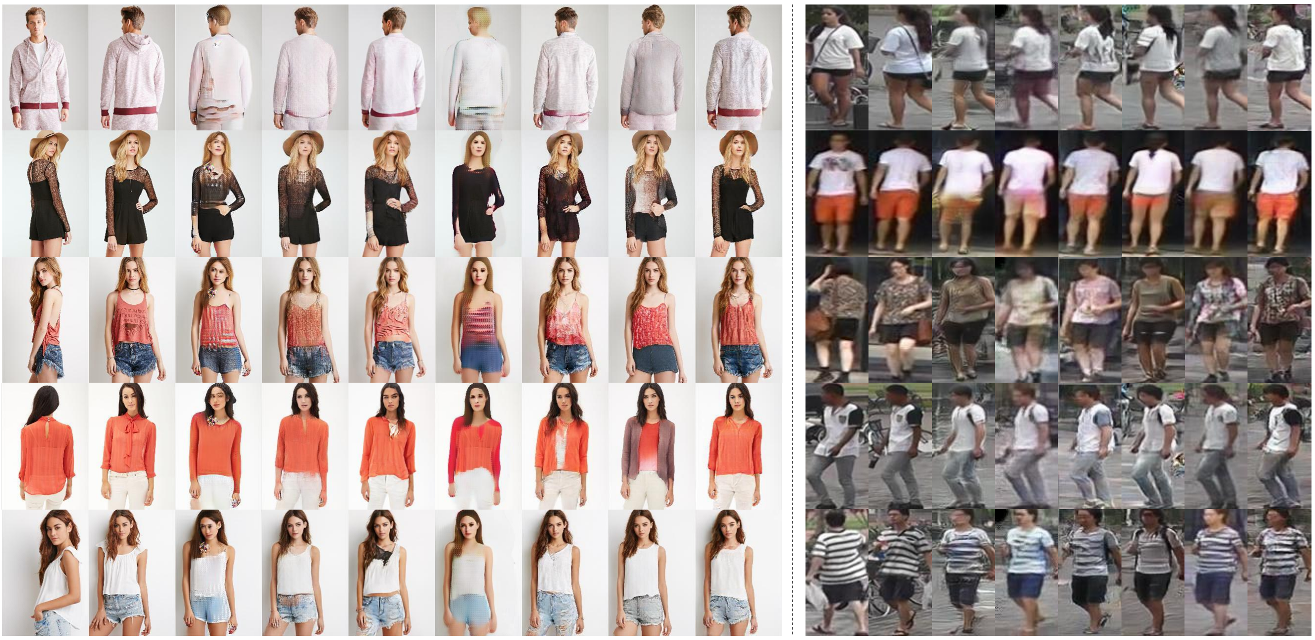

Figure 4. Qualitative comparison with several state-of-the-art methods on DeepFashion (Left) and Market1501 (Right).

generation of the synthetic images. For example, in the 4th row on the left of Fig. 4, PISE and SPIG improperly generate jackets in the synthetic images. Unlike the aforementioned methods, our DPTN optimizes the source-to-target task with the help of the auxiliary source-to-source task, making the generated image more vibrant. On the Market1501 dataset, our DPTN can still generate finer and more vivid textures than other methods. For instance, in the 4th row on the right of Fig. 4, only our method retains the garment pattern of the source image.

4.4. Ablation Study

We conduct a series of experiments on DeepFashion to verify the contribution of each component in our model. The various options for removing the corresponding components from our full model are listed as follows.

The model without Dual-Task Learning (w/o DTL). This model is similar to the existing methods that only focus on the source-to-target task. The entire self-reconstruction branch, including De and the loss function, is removed.

The model without Pose Transformer Module (w/o PTM). This model removes the PTM. In this way, the source-to-target branch will lack the guidance of the dual-task correlation, and will directly produce the target generated image ($\widetilde{x}_t$) from $\boldsymbol{F_{s\rightarrow t}}$.

The model without Contextual Augment Blocks (w/o CABs). This model removes CABs in the PTM. In this way, the feature from the source-to-source task ($\boldsymbol{F_{s\rightarrow s}}$) will be simply fed into TTBs to calculate the dual-task correlation.

Figure 5. Qualitative comparison of the ablation study.

The model without encoder $\boldsymbol{En_s}$ (w/o $\boldsymbol{En_s}$). This model removes the encoder En_s, and directly uses the feature $\boldsymbol{F_{s\rightarrow s}}$ as the value in MHCA.

Full Model (Full). We use our proposed dual-task pose transformer network in this model.

Fig. 5 and Tab. 4 show the qualitative and quantitative results of the ablation study. As shown in Fig. 5, we can see that (1) Compared with the full model, the model w/o DTL is unstable and tends to generate heavy artifacts. This demonstrates the significance of the source-to-source task

Table 4. Quantitative comparisons of ablation study on the DeepFashion dataset. The best results are shown in bold.

Methods	SSIM ↑	PSNR ↑	FID ↓	LPIPS ↓
w/o DTL	0.7713	18.8134	14.7168	0.2143
w/o PTM	0.7755	18.8503	15.5281	0.2195
w/o CABs	0.7760	19.0489	12.0932	0.1989
w/o En_s	0.7778	19.1084	12.6858	0.1976
Full	**0.7782**	**19.1492**	**11.4664**	**0.1957**

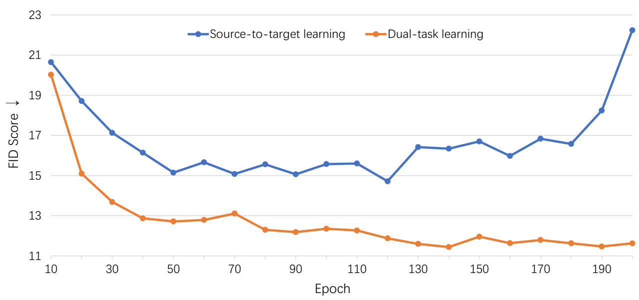

Figure 6. Learning curves of FID score by using source-to-target learning and dual-task learning on DeepFashion dataset.

during the training process. (2) Lack of texture mapping between the sources and the targets, the model w/o PTM cannot well utilize the textures of the real source image, resulting in blurred images. (3) For the model w/o CABs, the information of the source-to-source task is not well integrated, which misleads the source-to-target task to generate unrealistic patterns. (4) The images generated by the model w/o En_s lose many appearance details, verifying the effect of En_s in supplying fine source textures for PTM. (5) Compared with others, our full model can not only generate satisfactory global appearance but also produce realistic local textures. In addition, the quantitative comparison in Tab. 4 further demonstrates the effectiveness of our full model.

4.5. Effect of dual-task learning on training stability

To explore the influence of dual-task learning on training stability, following [23], we visualize the learning curves of FID score under the source-to-target learning and dual-task learning in Fig. 6. We can see that the FID score of the DPTN with solely source-to-target learning plateaus around 50 epochs, while the DPTN with dual-task learning continues to improve even afterward. This verifies that the knowledge brought by the source-to-source learning can effectively assist the learning of the source-to-target task. In addition, compared with dual-task learning, the DPTN with solely source-to-target learning tends to collapse after 130 epochs. This shows that by sharing partial weights between the dual tasks, the training of the easier source-to-source task can stabilize the training of the whole network to a certain extent, so as to better optimize the source-to-target task.

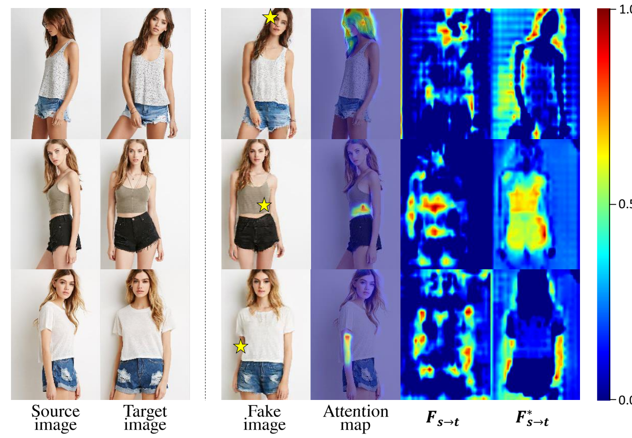

Figure 7. Visualization of the attention weight in MHCA and the heatmaps of $F_{s\to t}$ and $F^*_{s\to t}$. The yellow star on the fake image represents the query position.

4.6. Visualization of PTM

To explore how the PTM works in our framework, we also visualize the attention weight in MHCA as well as the heatmaps of the $F_{s\to t}$ and $F^*_{s\to t}$ in Fig. 7. As one can see, the attention weight obtained in the PTM can accurately focus the area related to the query position. This verifies that our PTM can effectively explore the pixel-wise transformation between the sources and targets. In addition, compared with the heatmap of the $F_{s\to t}$, the $F^*_{s\to t}$ produced by PTM contains more appearance cues. This manifests that our PTM can transfer the natural source textures to refine $F_{s\to t}$, and facilitate the source-to-target task to generate more realistic details.

5. Conclusions

In this paper, we propose a novel Dual-task Pose Transformer Network (DPTN) for PGPIG. Unlike most of the existing methods only focusing on the source-to-target task, our DPTN introduces an auxiliary task (i.e., source-to-source task) by a Siamese architecture, and exploits its knowledge to assist the source-to-target learning. Moreover, we carefully design a Pose Transformer Model (PTM) to explore the correlation between the dual tasks. Such correlation can be employed as a strong guidance for transferring source textures to the target generated image. Both the quantitative and qualitative results show that the proposed DPTN can improve upon prior PGPIG methods.

Acknowledgments. This project is supported by the Key-Area Research and Development Program of Guangdong Province (2019B010155003), and the National Natural Science Foundation of China (62072482).

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, July 2017. 5

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 4

[4] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 8857–8866, June 2018. 1, 2, 6

[5] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning view-specific deep networks for person re-identification. *IEEE Transactions on Image Processing*, 27(7):3472–3483, 2018. 1

[6] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590, 2020. 1

[7] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Resolution-aware knowledge distillation for efficient inference. *IEEE Transactions on Image Processing*, 30:6985–6996, 2021. 1

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 1

[9] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 820–828, 2016. 2

[10] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 6

[11] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6629–6640, 2017. 6

[12] Drew A Hudson and C. Lawrence Zitnick. Generative adversarial transformers. *International Conference on Machine Learning (ICML)*, 2021. 3

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2017–2025, 2015. 1

[14] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 2021. 3

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711, 2016. 5

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[17] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 1

[18] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 3688–3697, June 2019. 1, 2, 6

[19] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE Transactions on Image Processing*, 30:6392–6407, 2021. 1

[20] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, June 2016. 2, 5

[21] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 10806–10815, June 2021. 1, 2, 6

[22] 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 (NeurIPS)*, pages 405–415, 2017. 1, 2, 6

[23] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. 8

[24] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 7687–7696, 2020. 1, 2, 6

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 2

[26] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 2

[27] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 3408–3416, June 2018. 6

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[29] Mohsen Tabejamaat, Farhood Negin, and Francois Bremond. Guided flow field estimation by generating independent patches. *British Machine Vision Conference (BMVC)*, 2021. 1, 2

[30] Hao Tang, Song Bai, Li Zhang, Philip H. S. Torr, and Nicu Sebe. Xinggan for person image generation. In *European Conference on Computer Vision (ECCV)*, pages 717–734, 2020. 1, 2, 6

[31] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 3

[32] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 4105–4113, July 2017. 5

[33] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6000–6010, 2017. 3

[34] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8933–8940, 2019. 1

[35] Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. P2snet: Can an image match a video for person re-identification in an end-to-end way? *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2777–2787, 2018. 1

[36] Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, Cheng Xiang Zhai, and Tie Yan Liu. Multi-agent dual learning. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[37] Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. Dual transfer learning for neural machine translation with marginal distribution regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5553–5560, 2018. 2

[38] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[39] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 7982–7990, June 2021. 1, 2, 6

[40] Quan Zhang, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. Seeing like a human: Asynchronous learning with dynamic progressive refinement for person re-identification. *IEEE Transactions on Image Processing*, 31:352–365, 2022. 1

[41] Quan Zhang, Jianhuang Lai, and Xiaohua Xie. Learning modal-invariant angular metric by cyclic projection network for vis-nir person re-identification. *IEEE Transactions on Image Processing*, 30:8019–8033, 2021. 1

[42] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 586–595, June 2018. 2, 6

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 2, 5

[44] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[45] 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 (CVPR)*, pages 2342–2351, 2019. 1, 2, 5, 6