Bipartite Graph Reasoning GANs for Person Image Generation

Hao Tang[12]
hao.tang@unitn.it
Song Bai[2]
songbai.site@gmail.com
Philip H.S. Torr[2]
philip.torr@eng.ox.ac.uk
Nicu Sebe[13]
sebe@disi.unitn.it

[1] DISI
University of Trento
[2] Department of Engineering Science
University of Oxford
[3] Huawei Research Ireland

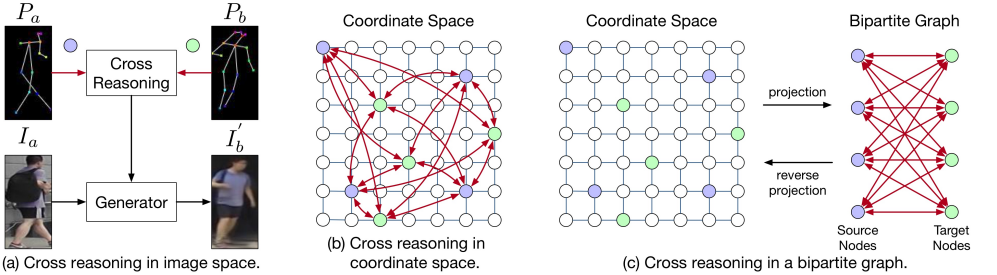
Abstract

We present a novel Bipartite Graph Reasoning GAN (BiGraphGAN) for the challenging person image generation task. The proposed graph generator mainly consists of two novel blocks that aim to model the pose-to-pose and pose-to-image relations, respectively. Specifically, the proposed Bipartite Graph Reasoning (BGR) block aims to reason the crossing long-range relations between the source pose and the target pose in a bipartite graph, which mitigates some challenges caused by pose deformation. Moreover, we propose a new Interaction-and-Aggregation (IA) block to effectively update and enhance the feature representation capability of both person's shape and appearance in an interactive way. Experiments on two challenging and public datasets, *i.e.*, Market-1501 and DeepFashion, show the effectiveness of the proposed BiGraphGAN in terms of objective quantitative scores and subjective visual realness. The source code and trained models are available at https://github.com/Ha0Tang/BiGraphGAN.

1 Introduction

In this paper, we mainly focus on translating a person image from one pose to another as depicted in Fig. 1 and 2. Existing person image generation methods such as [1, 2, 4, 8, 16, 18, 20, 21, 29, 32, 41, 45] always rely on building convolution layers. Due to the physical design of convolutional filters, convolution operations can only model local relations. To capture global relations, existing methods such as [32, 45] inefficiently stack multiple convolution layers to enlarge the receptive fields to cover all the body joints from both the source pose and the target pose. However, none of the above-mentioned methods explicitly consider modeling the cross relations between the source pose and the target pose.

In this paper, we propose a novel Bipartite Graph Reasoning GAN (BiGraphGAN), which mainly consists of two novel blocks, *i.e.*, Bipartite Graph Reasoning (BGR) block and Interaction-and-Aggregation (IA) block. The BGR block aims to efficiently capture the crossing long-range relations between the source pose and the target pose in a bipartite graph (see Fig. 1). Specifically, the BGR block first projects both the source pose feature and the



(a) Cross reasoning in image space. (b) Cross reasoning in coordinate space. (c) Cross reasoning in a bipartite graph.

Figure 1: Illustration of our motivation. We propose a novel BiGraphGAN (Fig. (c)) for capturing crossing long-range relations between the source pose P_a and the target pose P_b in a bipartite graph. The node features from both source and target poses in the coordinate space are projected into the nodes in a bipartite graph, thereby forming a fully-connected bipartite graph. After cross-reasoning the graph, the node features are projected back to the original coordinate space for further processing.

target pose feature in the original coordinate space onto a bipartite graph. Next, both source and target pose features are represented by a set of nodes to form a fully-connected bipartite graph, on which crossing long-range relation reasoning is performed by Graph Convolution Networks (GCNs). To the best of our knowledge, we are the first to explore GCNs to model the crossing long-range relations for solving the challenging person image generation task. After reasoning, we project the node features back to the original coordinate space for further processing.

Also, the proposed IA block is proposed to effectively and interactively enhance person's shape and appearance features. We also introduce an Attention-based Image Fusion (AIF) module to selectively generate the final result using an attention network. Qualitative and quantitative experiments on two challenging datasets, *i.e.*, Market-1501 [44] and DeepFashion [19], demonstrate that the proposed BiGraphGAN generates better person images than several state-of-the-art methods, *i.e.*, PG2 [20], DPIG [21], Deform [29], C2GAN [32], BTF [1], VUnet [8] and PATN [45].

The contributions of this paper are summarized as follows,

- We propose a novel Bipartite Graph Reasoning GAN (BiGraphGAN) for person image generation. The proposed BiGraphGAN aims to progressively reason the pose-to-pose and pose-to-image relations via two novel proposed blocks.
- We propose a novel Bipartite Graph Reasoning (BGR) block to effectively reason the crossing long-range relations between the source pose and the target pose in a bipartite graph by using Graph Convolutional Networks (GCNs). Moreover, we present a new Interaction-and-Aggregation (IA) block to interactively enhance both person's appearance and shape feature representations.
- Extensive experiments on two challenging datasets, *i.e.*, Market-1501 [44] and DeepFashion [19], demonstrate the effectiveness of the proposed BiGraphGAN and show significantly better performance compared with state-of-the-art approaches.

2 Related Work

Generative Adversarial Networks (GANs) [9] have shown the potential to generate realistic images [3, 11, 28]. For instance, Shaham *et al.* propose an unconditional SinGAN [28]

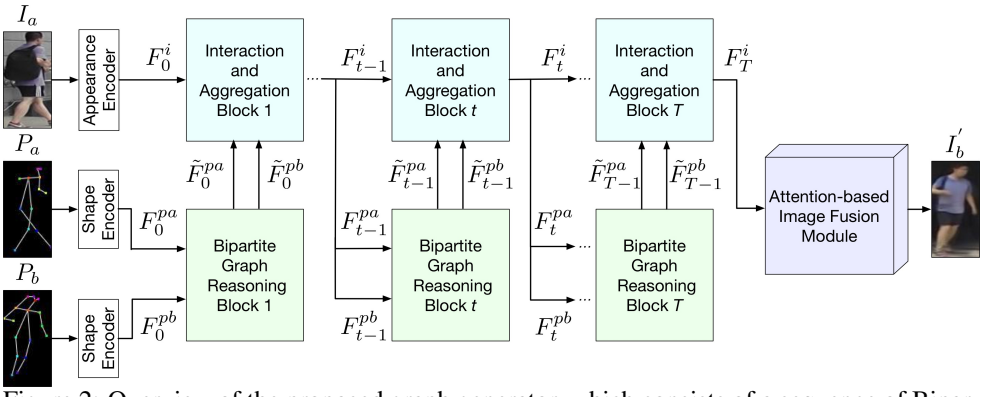


Figure 2: Overview of the proposed graph generator, which consists of a sequence of Bipartite Graph Reasoning (BGR) blocks, a sequence of Interaction-and-Aggregation (IA) blocks and an Attention-based Image Fusion (AIF) module. BGR blocks aim to reason the crossing long-range relations between the source pose and the target pose in a bipartite graph. IA blocks aim to interactively update person's appearance and shape feature representations. AIF module aims to selectively generate the final result via an attention network. The symbols $F^i{=}\{F_j^i\}_{j=0}^{T}$, $F^{pa}{=}\{F_j^{pa}\}_{j=0}^{T-1}$, $F^{pb}{=}\{F_j^{pb}\}_{j=0}^{T-1}$, $\tilde{F}^{pa}{=}\{\tilde{F}_j^{pa}\}_{j=0}^{T-1}$, and $\tilde{F}^{pb}{=}\{\tilde{F}_j^{pb}\}_{j=0}^{T-1}$ denote the appearance codes, the source shape codes, the target shape codes, the updated source shape codes, and the updated target shape codes, respectively.

which can be learned from a single image. Moreover, to generate user-defined images, Conditional GAN (CGAN) [23] has been proposed recently. A CGAN always consists of a vanilla GAN and external guide information such as class labels [7, 39, 42], segmentation maps [17, 24, 33, 36], attention maps [12, 22, 34], and human skeleton [1, 2, 31, 35, 45]. In this work, we mainly focus on the challenging person image generation task, which aims to transfer a person image from one pose to another one.

Person Image Generation is a challenging task due to the pose deformation between the source image and the target image. Modeling the long-range relations between the source pose and the target pose is the key to solving this challenging task. However, existing methods such as [1, 2, 4, 8, 16, 18, 20, 21, 29, 32, 41, 45] built through the stacking of convolutional layers, which can only leverage the relations between the source pose and the target pose locally. For instance, Zhu *et al.* [45] propose a Pose-Attentional Transfer Block (PATB), in which the source and target poses are simply concatenated and then fed into an encoder to capture their dependencies.

Unlike existing methods for modeling the relations between the source and target poses in a localized manner, we show that the proposed Bipartite Graph Reasoning (BGR) block can bring considerable performance improvements in the global view.

Graph-Based Reasoning. Graph-based approaches have shown to be an efficient way to reason relation in many computer vision tasks such as semi-supervised classification [14], video recognition [37], crowd counting [5], action recognition [26, 40] and semantic segmentation [6, 43].

Compared to these graph-based reasoning methods which model the long-range relations within the same feature map to incorporate global information, we focus on developing a novel BiGraphGAN framework that reasons and models the crossing long-range relations between different features of the source pose and target pose in a bipartite graph. Then the

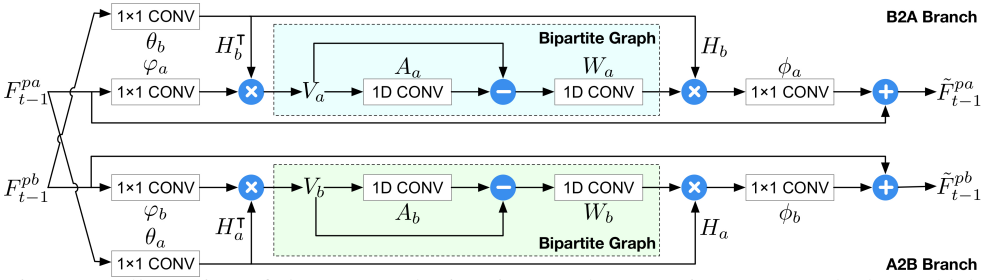


Figure 3: Illustration of the proposed Bipartite Graph Reasoning (BGR) Block t, which consists of two branches, *i.e.*, B2A and A2B. Each of them aims to model cross-contextual information between shape features F_{t-1}^{pa} and F_{t-1}^{pb} in a bipartite graph via Graph Convolutional Networks (GCNs).

crossing relations are further used to guide the image generation process (see Fig. 1). This idea has not been investigated in existing GAN-based image translation methods.

3 Bipartite Graph Reasoning GANs

We start by introducing the details of the proposed Bipartite Graph Reasoning GAN (BiGraphGAN), which consists of a graph generator G and two discriminators (*i.e.*, appearance discriminator D_a and shape discriminator D_s). An illustration of the proposed graph generator G is shown in Fig. 2, which mainly contains three parts, *i.e.*, a sequence of Bipartite Graph Reasoning (BGR) blocks modeling the crossing long-range relations between the source pose P_a and the target pose P_b, a sequence of Interaction-and-Aggregation (IA) blocks interactively enhancing both person's shape and appearance feature representations, and an Attention-based Image Fusion (AIF) module attentively generating the final result $I_b^{'}$. In the following, we first present the proposed blocks and then introduce the optimization objective and implementation details of the proposed BiGraphGAN.

Fig. 2 shows the proposed graph generator G, whose inputs are the source image I_a, the source pose P_a and the target pose P_b. The generator G aims to transfer the pose of the person in the source image I_a from the source pose P_a to the target pose P_b, generating the desired image $I_b^{'}$. Firstly, I_a, P_a and P_b are separately fed into three encoders to obtain the appearance code F_0^i, the source shape code F_0^{pa} and the target shape code F_0^{pb}. Note that we used the same shape encoder to learn both P_a and P_b, *i.e.*, the two shape encoders for learning the two different poses are sharing the weights.

3.1 Pose-to-Pose Bipartite Graph Reasoning

The proposed Bipartite Graph Reasoning (BGR) block aims to reason the crossing long-range relations between the source pose and the target pose in a bipartite graph. All BGR blocks have an identical structure as illustrated in Fig. 2. Consider the t-th block given in Fig. 3, whose inputs are the source shape code F_{t-1}^{pa} and the target shape code F_{t-1}^{pb}. The BGR block aims to reason these two codes in a bipartite graph via Graph Convolutional Networks (GCNs) and outputs new shape codes. The proposed BGR block contains two symmetrical branches (*i.e.*, B2A branch and A2B branch) because a bipartite graph is a bidirectional

graph. As shown in Fig. 1(c), each node in the source nodes connects all the target nodes; at the same time, each node in the target nodes connects all the source nodes. In the following, we mainly describe the detailed modeling process of the B2A branch, and another A2B branch is similar to this.

From Coordinate Space to Bipartite-Graph Space. Firstly, we reduce the dimension of the source shape code F_{t-1}^{pa} with function $\varphi_a(F_{t-1}^{pa})\in\mathbb{R}^{C\times D_a}$, where C is the number of feature map channels, D_a is the number of nodes of F_{t-1}^{pa}. Then we reduce the dimension of the target shape code F_{t-1}^{pb} with function $\theta_b(F_{t-1}^{pb})=H_b^{\intercal}\in\mathbb{R}^{D_b\times C}$, where D_b is the number of nodes of F_{t-1}^{pb}. Next, we project F_{t-1}^{pa} to a new feature V_a in a bipartite graph using the projection function H_b^T. Therefore we have,

$$V_a = H_b^{\intercal}\varphi_a(F_{t-1}^{pa}) = \theta_b(F_{t-1}^{pb})\varphi_a(F_{t-1}^{pa}), \tag{1}$$

where both functions $\theta_b(\cdot)$ and $\varphi_a(\cdot)$ are implemented using 1×1 convolutional layer. This results in a new feature $V_a\in\mathbb{R}^{D_b\times D_a}$ in the bipartite graph, which represents the crossing relations between the nodes of the target pose F_{t-1}^{pb} and the source pose F_{t-1}^{pa} (see Fig. 1(c)).

Cross Reasoning with Graph Convolution. After projection, we build a fully-connected bipartite graph with adjacency matrix $A_a\in\mathbb{R}^{D_b\times D_b}$. We then use a graph convolution to reason the crossing long-range relations between the nodes from both source and target poses, which can be formulated as,

$$M_a = (\mathrm{I}-A_a)V_aW_a, \tag{2}$$

where $W_a\in\mathbb{R}^{D_a\times D_a}$ denotes the trainable edge weights. We follow [6, 43] and use Laplacian smoothing [6, 15] to propagate the node features over the bipartite graph. The identity matrix I can be viewed as a residual sum connection to alleviate optimization difficulties. Note that we randomly initialize both adjacency matrix A_a and the weights W_a, and then train both by gradient descent in an end-to-end manner.

From Bipartite-Graph Space to Coordinate Space. After the cross-reasoning, the updated new feature M_a is mapped back to the original coordinate space for further processing. Next, we add the result to the original source shape code F_{t-1}^{pa} to form a residual connection [10]. This process can be expressed as,

$$\tilde{F}_{t-1}^{pa} = \phi_a(H_bM_a) + F_{t-1}^{pa}, \tag{3}$$

where we reuse the projection matrix H_b and perform a linear projection $\phi_a(\cdot)$ to project M_a back to the original coordinate space. Therefore, we obtain the new source feature $\tilde{F}_{t-1}^{pa}$, which has the same dimension with the original one F_{t-1}^{pa}.

Similarly, the A2B branch outputs the new target shape feature $\tilde{F}_{t-1}^{pb}$. Note that the idea of the proposed BGR block is inspired by the GloRe unit proposed by [6]. The main difference is that the GloRe unit reasons the relations within the same feature map via a standard graph, but the proposed BGR block reasons the crossing relations between feature maps of different poses using a bipartite graph.

3.2 Pose-to-Image Interaction and Aggregation

As shown in Fig. 2, the proposed Interaction-and-Aggregation (IA) block receives the appearance code F_{t-1}^{i}, the new source shape code $\tilde{F}_{t-1}^{pa}$ and the new target shape code $\tilde{F}_{t-1}^{pb}$ as

inputs. IA block aims to simultaneously and interactively enhance F_t^i, F_t^{pa} and F_t^{pb}. Specifically, both shape codes firstly concatenated and fed into two convolutional layers to produce the attention map A_p. Mathematically,

$$A_p = \sigma(\text{Conv}(\text{Concat}(\tilde{F}_{t-1}^{pa}, \tilde{F}_{t-1}^{pb}))), \tag{4}$$

where $\sigma(\cdot)$ denotes the element-wise Sigmoid function.

Appearance Code Enhance. After obtaining A_p, the appearance F_{t-1}^i is enhanced by,

$$F_t^i = A_p \otimes F_{t-1}^i + F_{t-1}^i, \tag{5}$$

where $\otimes$ denotes element-wise product. By multiplying with the attention map A_p, the new appearance code F_t^i at certain locations can be either preserved or suppressed.

Shape Code Enhance. Next, we concatenate F_t^i, F_{t-1}^{pa} and F_{t-1}^{pb}, and go through two convolutional layers to obtain the updated shape code F_t^{pa} and F_t^{pb} by splitting the result along the channel axis. This process can be performed by,

$$F_t^{pa}, F_t^{pb} = \text{Conv}(\text{Concat}(F_t^i, \tilde{F}_{t-1}^{pa}, \tilde{F}_{t-1}^{pb})). \tag{6}$$

In this way, both new shape codes F_t^{pa} and F_t^{pb} can synchronize the changes caused by the new appearance code F_t^i.

3.3 Attention-Based Image Fusion

At the T-th IA block, we obtain the final appearance code F_T^i. We then feed F_T^i to an image decoder to generate the intermediate result $\tilde{I}_b$. At the same time, we feed F_T^i to an attention decoder to produce the attention mask A_i.

The attention encoder consists of several deconvolutional layers and a Sigmoid activation layer. Thus, the attention encoder aims to generate a one-channel attention mask A_i, in which each pixel value is between 0 to 1. The attention mask A_i aims to selectively pick useful content from both the input image I_a and the intermediate result $\tilde{I}_b$ for generating the final result $I_b^{'}$. This process can be expressed as,

$$I_b^{'} = I_a \otimes A_i + \tilde{I}_b \otimes (1 - A_i), \tag{7}$$

where $\otimes$ denotes element-wise product. In this way, both the image decoder and the attention decoder can interact with each other and ultimately produce better results.

3.4 Model Training

Appearance and Shape Discriminators. We adopt two discriminators for adversarial training. Specifically, we feed image-image pair (I_a, I_b) and (I_a, $I_b^{'}$) into the appearance discriminator D_a to ensure appearance consistency. Meanwhile, we feed pose-image pair (P_b, I_b) and (P_b, $I_b^{'}$) into the shape discriminator D_s for shape consistency. Both discriminators (*i.e.*, D_a and D_s), and the proposed graph generator G are trained in an end-to-end way, aiming to enjoy mutual benefits from each other in a joint framework.

Optimization Objectives. We follow [35, 45] and use the adversarial loss $\mathcal{L}_{gan}$, the pixel-wise $L1$ loss $\mathcal{L}_{l1}$ and the perceptual loss $\mathcal{L}_{per}$ as our optimization objectives,

$$\mathcal{L}_{full} = \lambda_{gan}\mathcal{L}_{gan} + \lambda_{l1}\mathcal{L}_{l1} + \lambda_{per}\mathcal{L}_{per}, \tag{8}$$

Table 1: Quantitative comparison of different methods on Market-1501 and DeepFashion. For all metrics, higher is better. (∗) denotes the results tested on our testing set.

Method	Market-1501					DeepFashion		
	SSIM	IS	Mask-SSIM	Mask-IS	PCKh	SSIM	IS	PCKh
PG2 [20]	0.253	3.460	0.792	3.435	-	0.762	3.090	-
DPIG [21]	0.099	3.483	0.614	3.491	-	0.614	3.228	-
Deform [29]	0.290	3.185	0.805	3.502	-	0.756	3.439	-
C2GAN [32]	0.282	3.349	0.811	3.510	-	-	-	-
BTF [1]	-	-	-	-	-	0.767	3.220	-
PG2* [20]	0.261	**3.495**	0.782	3.367	0.73	0.773	3.163	0.89
Deform* [29]	0.291	3.230	0.807	3.502	**0.94**	0.760	3.362	0.94
VUnet* [8]	0.266	2.965	0.793	3.549	0.92	0.763	**3.440**	0.93
PATN* [45]	0.311	3.323	0.811	**3.773**	**0.94**	0.773	3.209	0.96
BiGraphGAN	**0.325**	3.329	**0.818**	3.695	**0.94**	**0.778**	3.430	**0.97**
Real Data	1.000	3.890	1.000	3.706	1.00	1.000	4.053	1.00

where λ_{gan}, λ_{l1} and λ_{per} control the relative importance of the three objectives. For the perception loss, we follow [35, 45] and use the $Conv1\_2$ layer.

Implementation Details. In our experiments, we follow previous work [35, 45] and represent the source pose P_a and the target pose P_b as two 18-channel heat maps that encode the locations of 18 joints of a human body. Adam optimizer [13] is employed to learn the proposed BiGraphGAN for around 90K iterations with β_1=0.5 and β_2=0.999.

In preliminary experiments, we found that as T increases, the performance is getting better and better. When T is equal to 9, the proposed model achieves the best results, and then the performance begins to decline. Thus we set T=9 in the proposed graph generator. Moreover, λ_{gan}, λ_{l1}, λ_{per} in Eq. (8), and the number of feature map channels C are set to 5, 10, 10, and 128, respectively. The proposed BiGraphGAN is implemented in PyTorch [25].

4 Experiments

Datasets. We follow previous works [20, 29, 45] and conduct extensive experiments on two public datasets, *i.e.*, Market-1501 [44] and DeepFashion [19]. Specifically, we adopt the train/test split used in [35, 45] for a fair comparison. In addition, images are resized to 128×64 and 256×256 on Market-1501 and DeepFashion, respectively.

Evaluation Metrics. We follow [20, 29, 45] and employ Inception score (IS) [27], Structure Similarity (SSIM) [38] and their masked versions (*i.e.*, Mask-IS and Mask-SSIM) as our evaluation metrics to quantitatively measure the quality of the generated images by different approaches. Moreover, we employ the PCKh score proposed in [45] to explicitly evaluate the shape consistency of the generated person images.

4.1 State-of-the-Art Comparisons

Quantitative Comparisons. We compare the proposed BiGraphGAN with several leading person image synthesis methods, *i.e.*, PG2 [20], DPIG [21], Deform [29, 30], C2GAN [32], BTF [1], VUnet [8], and PATN [45]. Quantitative comparison results are shown in Table 1, we can see that the proposed method achieves the best results on most metrics such as SSIM,

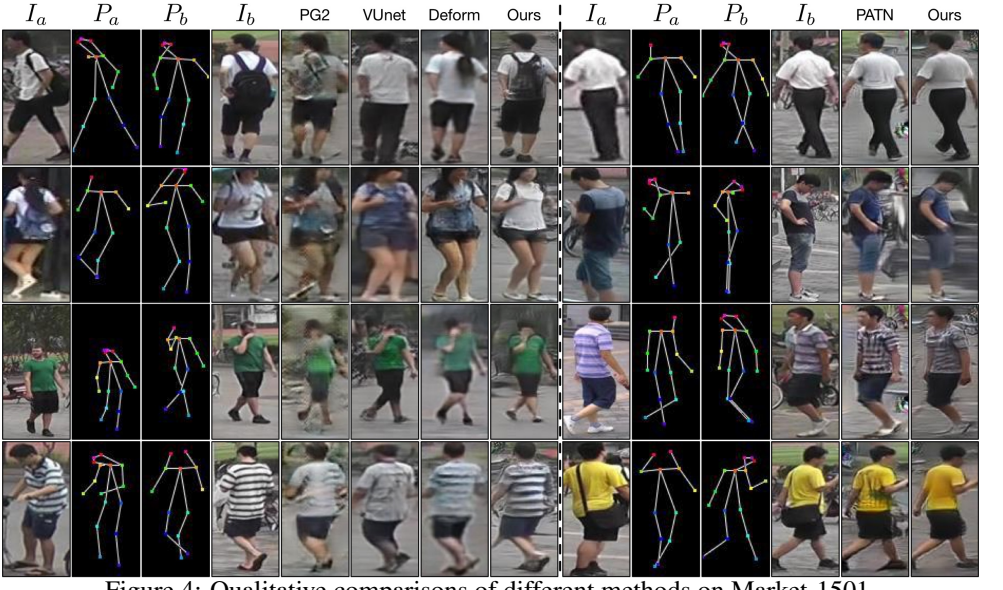


Figure 4: Qualitative comparisons of different methods on Market-1501.

Table 2: Quantitative comparison of user study (%) on Market-1501 and DeepFashion. 'R2G' and 'G2R' represent the percentage of real images rated as fake *w.r.t.* all real images, and the percentage of generated images rated as real *w.r.t.* all generated images, respectively.

Method	Market-1501		DeepFashion	
	R2G	G2R	R2G	G2R
PG2 [20]	11.20	5.50	9.20	14.90
Deform [29]	22.67	50.24	12.42	24.61
C2GAN [32]	23.20	46.70	-	-
PATN [45]	32.23	63.47	19.14	31.78
BiGraphGAN	**35.76**	**65.91**	**22.39**	**34.16**

Mask-SSIM and PCKh on Market-1501, and SSIM and PCKh on DeepFashion. For other metrics such as IS, the proposed method still achieves better results than the most related model PATN on both datasets. These results validate the effectiveness of our method.

Qualitative Comparisons. We also provide visualization comparison results on both datasets in Fig. 4 and 5. As shown in the left of both figures, the proposed BiGraphGAN generates remarkably better results than PG2 [20], VUnet [8] and Deform [29] on both datasets. To further evaluate the effectiveness of the proposed method, we compare the proposed BiGraphGAN with the most state-of-the-art model, *i.e.*, PATN [45], in the right of both figures. We still observe that our proposed BiGraphGAN generates more clear and visually plausible person images than PATN on both datasets.

User Study. We also follow [20, 29, 45] and conduct a user study to evaluate the quality of the generated images. Specifically, we follow the evaluation protocol used in [45] for a fair comparison. Comparison results of different methods are shown in Table 2, we can see that the proposed method achieves the best results on all metrics, which further validates that the generated images by the proposed BiGraphGAN are more photo-realistic.

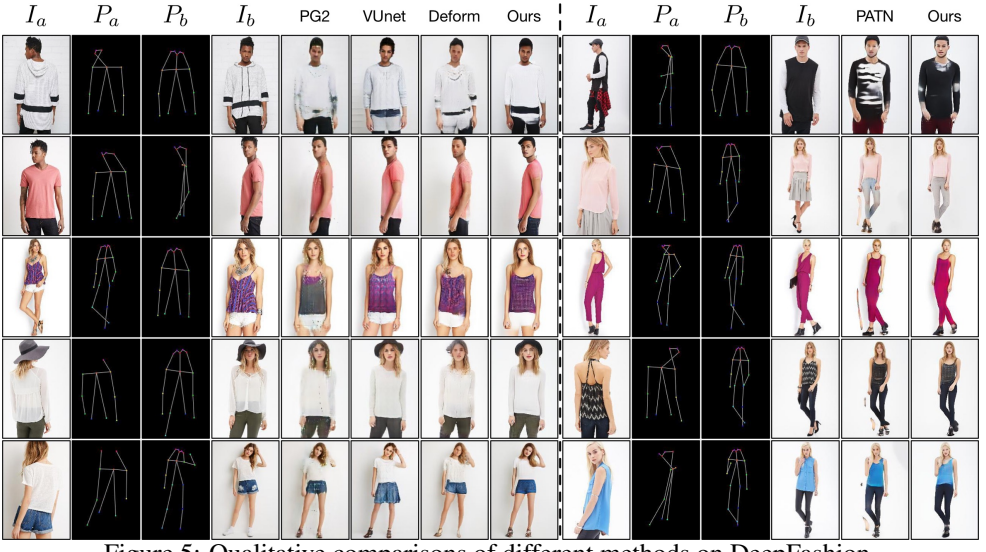


Figure 5: Qualitative comparisons of different methods on DeepFashion.

Table 3: Ablation study of the proposed BiGraphGAN on Market-1501. For both metrics, higher is better.

Baselines of BiGraphGAN	SSIM ↑	Mask-SSIM ↑
B1: Our Baseline	0.305	0.804
B2: B1 + B2A	0.310	0.809
B3: B1 + A2B	0.310	0.808
B4: B1 + A2B + B2A (Sharing)	0.322	0.813
B5: B1 + A2B + B2A (Non-Sharing)	0.324	0.813
B6: B5 + AIF	**0.325**	**0.818**

4.2 Ablation Study

Baselines of BiGraphGAN. We perform extensive ablation studies to validate the effectiveness of each component of the proposed BiGraphGAN on Market-1501. The proposed BiGraphGAN has 6 baselines (*i.e.*, B1, B2, B3, B4, B5, B6) as shown in Table 3 and Fig. 6(*left*). B1 is our baseline. B2 uses the proposed B2A branch for modeling the crossing relations from the target pose to the source pose. B3 adopts the proposed A2B branch to model the crossing relations from the source pose to the target pose. B4 uses the combination of both A2B and B2A branches to model the crossing relations between the source pose and the target pose. Note that both GCNs in B4 are sharing the parameters. B5 employs a non-sharing strategy between the two GCNs to model the crossing relations. B6 employs the proposed AIF module to make the graph generator attentively select which part is more useful for generating the final person image.

Ablation Analysis. The results of the ablation study are shown in Table 3 and Fig. 6(*left*). We observe that both B2 and B3 achieve significantly better results than B1, which proves our initial motivation that modeling the crossing relations between the source pose and the target pose in a bipartite graph will boost the generation performance. In addition, we see

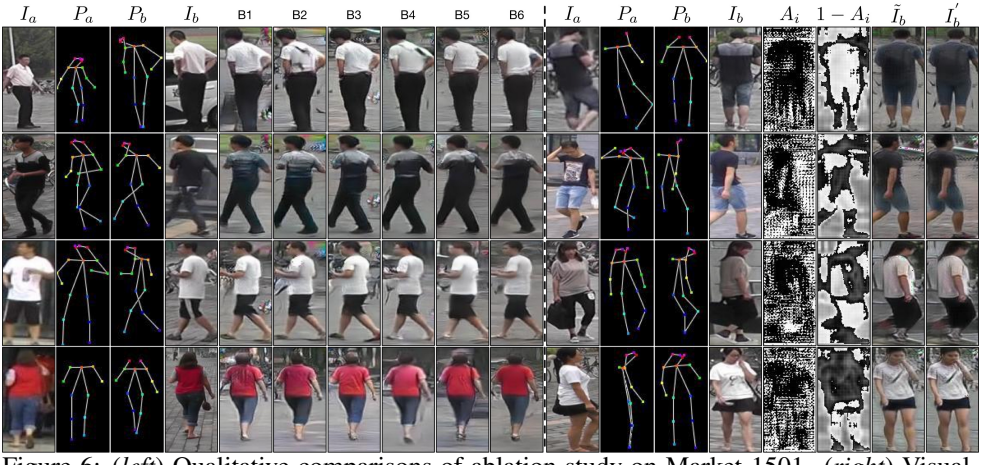


Figure 6: (*left*) Qualitative comparisons of ablation study on Market-1501. (*right*) Visualization of the learned attention masks and intermediate results.

that B4 performs better than B2 and B3, demonstrating the effectiveness of modeling the symmetric relations between the source and target poses. B5 achieves better results than B4, which means that two GCNs are constructed separately to model the symmetric relations will improve the generation performance in the joint network. B6 is better than B5, which clearly proves the effectiveness of the proposed attention-based image fusion strategy.

Moreover, we show several examples of the learned attention masks and intermediate results in Fig. 6(*right*) We can see that the proposed module attentively selects useful content from both the input image and intermediate result to generate the final result, thus verifying our design motivation.

5 Conclusions

In this paper, we propose a novel Bipartite Graph Reasoning GAN (BiGraphGAN) framework for the challenging person image generation task. We introduce two novel blocks, *i.e.*, Bipartite Graph Reasoning (BGR) block and Interaction-and-Aggregation (IA) block. The first is employed to model the crossing long-range relations between the source pose and the target pose in a bipartite graph. The second block is used to interactively enhance both person's shape and appearance features. Extensive experiments of both human judgments and automatic evaluation demonstrate that the proposed BiGraphGAN achieves remarkably better performance than the state-of-the-art approaches.

Acknowledgment

This work has been partially supported by the Italy-China collaboration project TALENT, the Royal Academy of Engineering under the Research Chair and Senior Research Fellowships scheme, EPSRC/MURI grant EP/N019474/1 and FiveAI.

References

[1] Badour AlBahar and Jia-Bin Huang. Guided image-to-image translation with bi-directional feature transformation. In *ICCV*, 2019.

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

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019.

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019.

[5] Xinya Chen, Yanrui Bin, Changxin Gao, Nong Sang, and Hao Tang. Relevant region prediction for crowd counting. *Elsevier Neurocomputing*, 2020.

[6] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, 2019.

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

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

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

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

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[12] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*, 2020.

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

[14] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[15] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*, 2018.

[16] Dong Liang, Rui Wang, Xiaowei Tian, and Cong Zou. Pcgan: Partition-controlled human image generation. In *AAAI*, 2019.

[17] Gaowen Liu, Hao Tang, Hugo Latapie, and Yan Yan. Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP*, 2020.

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

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

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

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

[22] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *NeurIPS*, 2018.

[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[26] Wei Peng, Jingang Shi, Zhaoqiang Xia, and Guoying Zhao. Mix dimension in poincaré geometry for 3d skeleton-based action recognition. In *ACM MM*, 2020.

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

[28] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *ICCV*, 2019.

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

[30] Aliaksandr Siarohin, Stéphane Lathuilière, Enver Sangineto, and Nicu Sebe. Appearance and pose-conditioned human image generation using deformable gans. *IEEE TPAMI*, 2019.

[31] Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. Gesturegan for hand gesture-to-gesture translation in the wild. In *ACM MM*, 2018.

[32] Hao Tang, Dan Xu, Gaowen Liu, Wei Wang, Nicu Sebe, and Yan Yan. Cycle in cycle generative adversarial networks for keypoint-guided image generation. In *ACM MM*, 2019.

[33] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, 2019.

[34] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *IJCNN*, 2019.

[35] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *ECCV*, 2020.

[36] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*, 2020.

[37] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[39] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *ICCV*, 2019.

[40] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

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

[42] Jichao Zhang, Yezhi Shu, Songhua Xu, Gongze Cao, Fan Zhong, Meng Liu, and Xueying Qin. Sparsely grouped multi-task generative adversarial networks for facial attribute manipulation. In *ACM MM*, 2018.

[43] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip HS Torr. Dual graph convolutional network for semantic segmentation. In *BMVC*, 2019.

[44] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

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