International Journal of Computer Vision (2023) 131:644–658
https://doi.org/10.1007/s11263-022-01722-5

Bipartite Graph Reasoning GANs for Person Pose and Facial Image Synthesis

Hao Tang[1] · Ling Shao[2] · Philip H. S. Torr[3] · Nicu Sebe[4]

Received: 12 September 2021 / Accepted: 9 November 2022 / Published online: 8 December 2022

Abstract
We present a novel bipartite graph reasoning Generative Adversarial Network (BiGraphGAN) for two challenging tasks: person pose and facial image synthesis. The proposed graph generator 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 long-range cross relations between the source and target pose in a bipartite graph, which mitigates some of the 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 a person's shape and appearance in an interactive way. To further capture the change in pose of each part more precisely, we propose a novel part-aware bipartite graph reasoning (PBGR) block to decompose the task of reasoning the global structure transformation with a bipartite graph into learning different local transformations for different semantic body/face parts. Experiments on two challenging generation tasks with three public datasets demonstrate the effectiveness of the proposed methods in terms of objective quantitative scores and subjective visual realness. The source code and trained models are available at https://github.com/Ha0Tang/BiGraphGAN.

Keywords GANs · Bipartite graph reasoning · Person pose synthesis · Facial expression synthesis

1 Introduction

In this paper, we focus on translating a person image from one pose to another and a facial image from one expression to another, as depicted in Fig. 1a. Existing person pose and facial image generation methods, such as Ma et al. (2017); Ma and Sun (2018); Siarohin et al. (2018); Tang et al. (2019c); AlBahar and Huang (2019); Esser et al. (2018); Zhu et al. (2019); Chan et al. (2019); Balakrishnan and Zhao (2018); Zanfir et al. (2018); Liang et al. (2019); Liu et al. (2019); Tang et al. (2019c); Zhang et al. (2020) typically rely on convolutional layers. However, due to the physical design of convolutional filters, convolutional operations can only model local relations. To capture global relations, existing methods such as Zhu et al. (2019); Tang et al. (2019c) inefficiently stack multiple convolutional 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 and target pose.

Rather than relying solely on convolutions/Transformers in the coordinate space to implicitly capture the cross relations between the source pose and the target pose, we propose to construct a latent interaction space where global or long-range (can also be understood as long-distance, which means that the distance between the same joint on the source pose and the target pose very long) reasoning can be performed directly. Within this interaction space, a pair of source and target joints that share similar semantics (e.g., the source left-hand and the target left-hand joints) are represented by a single mapping, instead of a set of scattered coordinate-specific mappings. Reasoning the relations of multiple different human joints is thus simplified to modeling those between the corresponding mappings in the

Communicated by Martin Fergie.

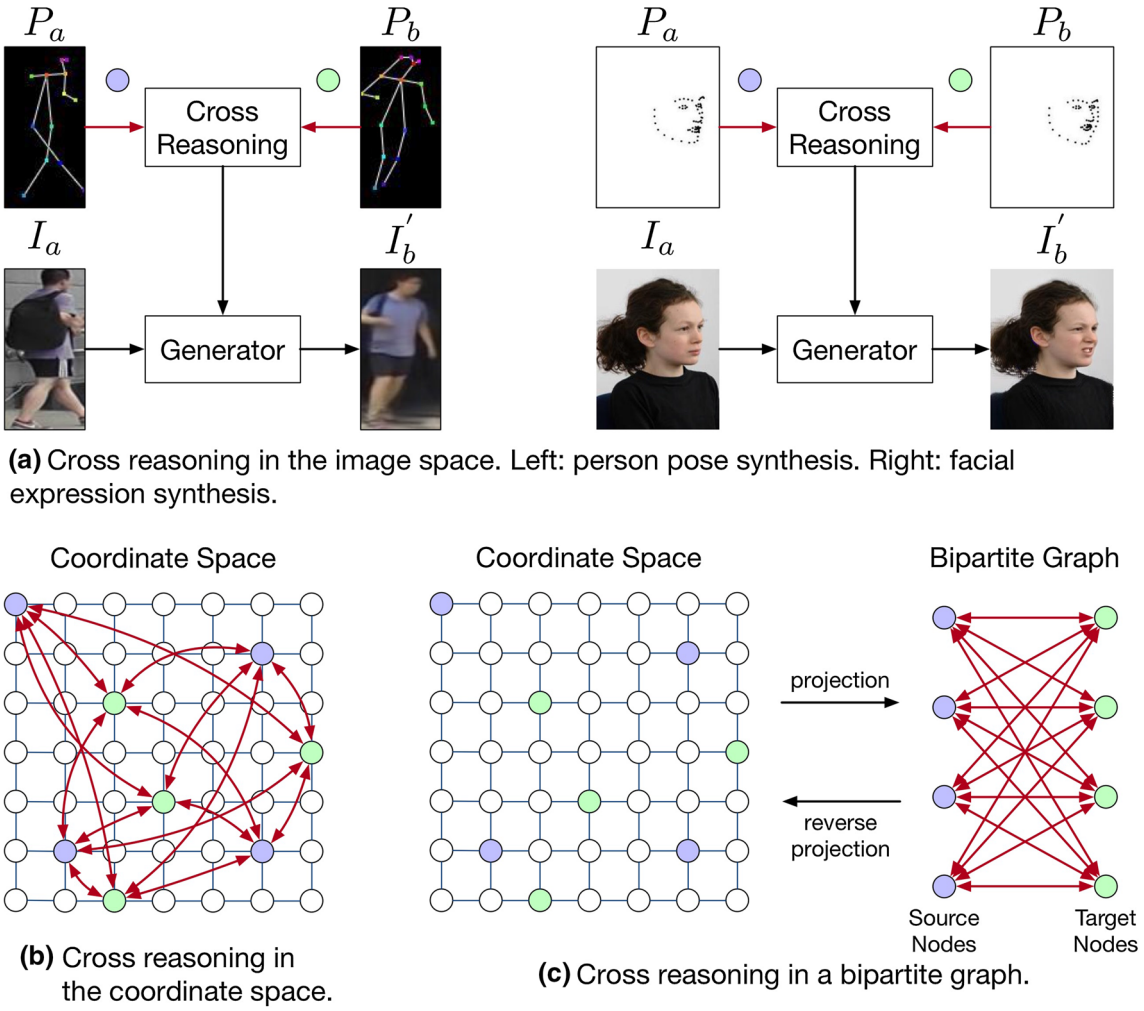
✉ Hao Tang
hao.tang@vision.ee.ethz.ch

[1] Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland

[2] Terminus AI Lab, Terminus Group, Beijing, China

[3] Department of Engineering Science, University of Oxford, Oxford, UK

[4] Department of Information Engineering and Computer Science (DISI), University of Trento, Trento, Italy



(a) Cross reasoning in the image space. Left: person pose synthesis. Right: facial expression synthesis.

(b) Cross reasoning in the coordinate space.

(c) Cross reasoning in a bipartite graph.

Fig. 1 Illustration of our motivation. We propose a novel BiGraphGAN (Fig. (c)) to capture the long-range cross relations between the source pose P_a and the target pose P_b in a bipartite graph. The node features from both the 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

interaction space. We thus build a bipartite graph connecting these mappings within the interaction space and perform relation reasoning over the bipartite graph. After the reasoning, the updated information is then projected back to the original coordinate space for the generation task. Accordingly, we design a novel bipartite graph reasoning (BGR) to efficiently implement the coordinate-interaction space mapping process, as well as the cross-relation reasoning by graph convolution network (GCNs).

In this paper, we propose a novel bipartite graph reasoning Generative Adversarial Network (BiGraphGAN), which consists of two novel blocks, i.e., a bipartite graph reasoning (BGR) block and an interaction-and-aggregation (IA) block. The BGR block aims to efficiently capture the long-range cross relations between the source pose and the target pose in a bipartite graph (see Fig. 1c). Specifically, the BGR block first projects both the source pose and target pose feature from the original coordinate space onto a bipartite graph. Next, the two features are represented by a set of nodes to form a fully connected bipartite graph, on which long-range cross relation reasoning is performed by GCNs. To the best of our knowledge, we are the first to use GCNs to model the long-range cross relations for solving both the challenging person pose and facial image generation tasks. After reasoning, we project the node features back to the original coordinate space for further processing. Moreover, to further capture the change in pose of each part more precisely, we further extend the BGR block to the part-aware bipartite graph reasoning (PBGR) block, which can capture the local transformations among body parts.

Meanwhile, the IA block is proposed to effectively and interactively enhance a person's shape and appearance fea-

tures. 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 person pose generation datasets, i.e., Market-1501 (Zheng et al., 2015) and DeepFashion (Liu et al., 2016), demonstrate that the proposed BiGraphGAN and BiGraphGAN++ generate better person images than several state-of-the-art methods, i.e., PG2 (Ma et al., 2017), DPIG (Ma & Sun, 2018), Deform (Siarohin et al., 2018), C2GAN (Tang et al., 2019c), BTF (AlBahar & Huang, 2019), VUNet (Esser et al., 2018), PATN (Zhu et al., 2019), PoseStylizer (Huang et al., 2020), and XingGAN (Tang et al., 2020b).

Lastly, to evaluate the versatility of the proposed BiGraphGAN, we also investigate the facial expression generation task on the Radboud Faces dataset (Langner et al., 2010). Extensive experiments show that the proposed method achieves better results than existing leading baselines, such as Pix2pix (Isola et al., 2017), GPGAN (Di et al., 2018), PG2 (Ma et al., 2017), CocosNet (Zhang et al., 2020), and C2GAN (Tang et al., 2019c).

The contributions of this paper are summarized as follows:

- We propose a novel bipartite graph reasoning GAN (BiGraphGAN) for person pose and facial image synthesis. The proposed BiGraphGAN aims to progressively reason the pose-to-pose and pose-to-image relations via two novel blocks.
- We propose a novel bipartite graph reasoning (BGR) block to effectively reason the long-range cross relations between the source and target pose in a bipartite graph, using GCNs.
- We introduce a new interaction-and-aggregation (IA) block to interactively enhance both a person's appearance and shape feature representations.
- We decompose the process of reasoning the global structure transformation with a bipartite graph into learning different local transformations for different semantic body/face parts, which captures the change in pose of each part more precisely. To this end, we propose a novel part-aware bipartite graph reasoning (PBGR) block to capture the local transformations among body parts.
- Extensive experiments on both the challenging person pose generation and facial expression generation tasks with three public datasets demonstrate the effectiveness of the proposed method and its significantly better performance compared with state-of-the-art methods.

Some of the material presented here appeared in Tang and Bai (2020). The current paper extends (Tang & Bai, 2020) in several ways:

- More detailed analyses are presented in the "Introduction" and "Related Work" sections, which now include very recently published papers dealing with person pose and facial image synthesis.
- We propose a novel PBGR block to capture the local transformations among body parts. Equipped with this new module, our BiGraphGAN proposed in Tang and Bai (2020) is upgraded to BiGraphGAN++.
- We present an in-depth description of the proposed method, providing the architectural and implementation details, with special emphasis on guaranteeing the reproducibility of our experiments. The source code is also available online.
- We extend the experimental evaluation provided in Tang and Bai (2020) in several directions. First, we conduct extensive experiments on two challenging tasks with three popular datasets, demonstrating the wide application scope of the proposed BiGraphGAN and BiGraphGAN++. Second, we also include more state-of-the-art baselines (e.g., PoseStylizer (Huang et al., 2020) and XingGAN Tang et al. (2020b)) for the person pose generation task, and observe that the proposed BiGraphGAN and BiGraphGAN++ achieve better results than both methods. Lastly, we conduct extensive experiments on the facial expression generation task, demonstrating both quantitatively and qualitatively that the proposed method achieves much better results than existing leading methods such as Pix2pix (Isola et al., 2017), GPGAN (Di et al., 2018), PG2 (Ma et al., 2017), CocosNet (Zhang et al., 2020), and C2GAN (Tang et al., 2019c).

2 Related Work

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have shown great potential in generating realistic images (Shaham, 2019; Karras et al., 2019; Brock et al., 2019; Zhang et al., 2022a, b; Tang & Sebe, 2021b; Tang et al., 2020c). For instance, Shaham et al. proposed an unconditional SinGAN (Shaham, 2019) which can be learned from a single image. Moreover, to generate user-defined images, several conditional GANs (CGANs) (Mirza & Osindero, 2014) have recently been proposed. A CGAN always consists of a vanilla GAN and external guidance information such as class labels (Po-Wei et al., 2019; Choi et al., 2018; Zhang et al., 2018a; Tang et al., 2019a), text descriptions (Xu et al., 2022; Tao et al., 2022), segmentation maps (Tang et al., 2019d; Park et al., 2019; Tang et al., 2020d; Liu et al., 2020; Wu et al., 2022; Songsong et al., 2022; Tang & Shao, 2022; Ren et al., 2021; Tang & Sebe, 2021a; Tang et al., 2020a), attention maps (Kim et al., 2020; Tang et al., 2019e; Mejjati, 2018; Tang et al., 2021), or human skeletons (AlBahar & Huang, 2019; Balakrishnan & Zhao, 2018; Zhu et al., 2019; Tang et al., 2018, 2020b).

In this work, we focus on the person pose and facial expression generation tasks, which aim to transfer a person image from one pose to another and a facial image from one expression to another, respectively.

Person Pose 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 and target pose is the key to solving this. However, existing methods, such as Balakrishnan and Zhao (2018); AlBahar and Huang (2019); Esser et al. (2018); Chan et al. (2019); Zanfir et al. (2018); Liang et al. (2019); Liu et al. (2019), are built by stacking several convolutional layers, which can only leverage the relations between the source pose and the target pose locally. For instance, Zhu et al. Zhu et al. (2019) proposed 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.

Facial Expression Generation aims to translate one facial expression to another (Tang et al., 2019b, c; Pumarola et al., 2020; Choi et al., 2018). For instance, Choi et al. (2018) proposed a scalable method that can perform facial expression-to-expression translation for multiple domains using a single model. Pumarola et al. (2020) introduced a GAN conditioning scheme based on action unit (AU) annotations, which describes in a continuous manifold the anatomical facial movements defining a human expression. Finally, Tang et al. (2019c) proposed a novel Cycle in Cycle GAN (C2GAN) for generating human faces and bodies.

Unlike existing person pose and facial expression generation methods, which model the relations between the source and target poses in a localized manner, we show that the proposed BGR block can bring considerable performance improvements in the global view.

Graph-Based Reasoning Graph-based approaches have been shown efficient at reasoning relations in many computer vision tasks such as semi-supervised classification (Kipf & Welling, 2017), video recognition (Wang & Gupta, 2018), crowd counting (Chen et al., 2020), action recognition (Yan et al., 2018; Peng et al., 2020), face clustering (Wang et al., 2019; Yang et al., 2019), and semantic segmentation (Chen & Rohrbach, 2019; Zhang et al., 2019).

In contrast, 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 two novel BiGraphGAN and BiGraphGAN++ frameworks that reason and model the long-range cross relations between different features of the source and target pose in a bipartite graph. Then, the cross relations are further used to guide the image generation process (see Fig. 1). This idea has not been investigated in existing GAN-based person image generation or even image-to-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., the appearance discriminator D_a and shape discriminator D_s). An illustration of the proposed graph generator G is shown in Fig. 2. It contains three main parts, i.e., a sequence of bipartite graph reasoning (BGR) blocks modeling the long-range cross relations between the source pose P_a and the target pose P_b, a sequence of interaction-and-aggregation (IA) blocks interactively enhancing both the 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.

Figure 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 initial appearance code F_0^i, the initial source shape code F_0^{pa}, and the initial target shape code F_0^{pb}. Note that we use the same shape encoder to learn both P_a and P_b, i.e., the two shape encoders used for learning the two different poses share weights.

3.1 Pose-to-Pose Bipartite Graph Reasoning

The proposed BGR block aims to reason the long-range cross 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 GCNs and outputs new shape codes. It contains two symmetrical branches (i.e., the B2A branch and A2B branch) because a bipartite graph is bidirectional. As shown in Fig. 1c, each source node is connected to all the target nodes; at the same time, each target node is connected to all the source nodes. In the following, we describe the detailed modeling process of the B2A branch. Note that the A2B branch is similar.

From Coordinate Space to Bipartite-Graph Space Firstly, we reduce the dimension of the source shape code F_{t-1}^{pa} with the function $\varphi_a\left(F_{t-1}^{pa}\right) \in \mathbb{R}^{C \times D_a}$, where C is the number of feature map channels, and 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 the function $\theta_b\left(F_{t-1}^{pb}\right) = H_b^{\intercal} \in \mathbb{R}^{D_b \times C}$, where D_b is the

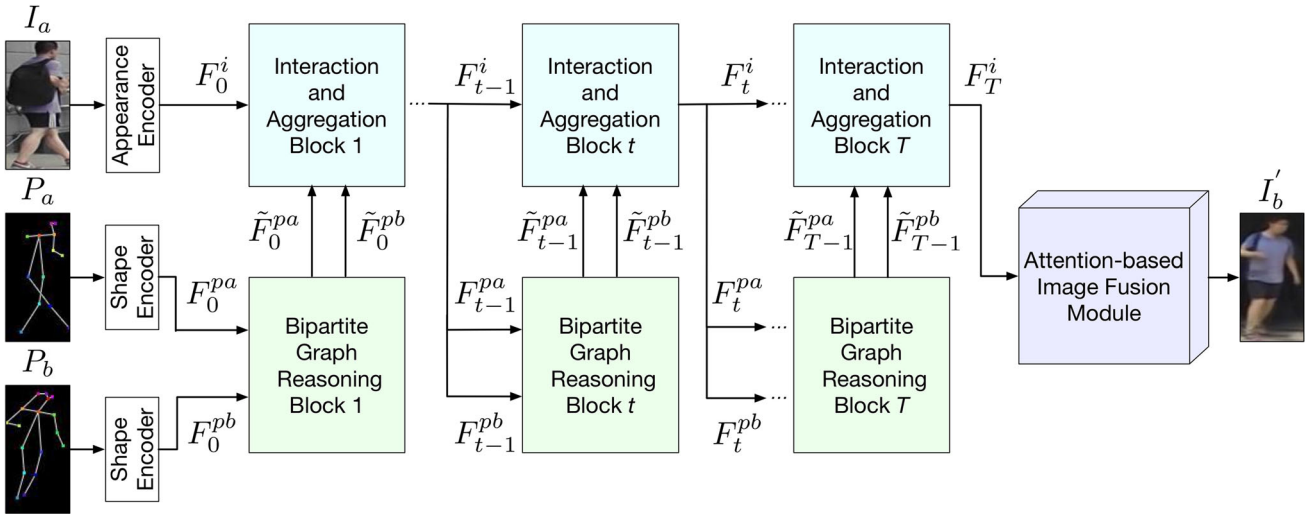


Fig. 2 Overview of the proposed graph generator, which consists of a sequence of bipartite graph reasoning (BGR) blocks, a sequence of interaction-and-aggregation (IA) blocks, and an attention-based image fusion (AIF) module. The BGR blocks aim to reason the long-range cross relations between the source pose and the target pose in a bipartite graph. The IA blocks aim to interactively update a person's appearance and shape feature representations. The AIF module aims to selectively generate the final result via an attention network. The symbols $F^i = \left\{F_j^i\right\}_{j=0}^{T}$, $F^{pa} = \left\{F_j^{pa}\right\}_{j=0}^{T-1}$, $F^{pb} = \left\{F_j^{pb}\right\}_{j=0}^{T-1}$, $\tilde{F}^{pa} = \left\{\tilde{F}_j^{pa}\right\}_{j=0}^{T-1}$, and $\tilde{F}^{pb} = \left\{\tilde{F}_j^{pb}\right\}_{j=0}^{T-1}$ denote the appearance codes, the source shape codes, the target shape codes, the updated source shape codes, and the updated target shape codes, respectively

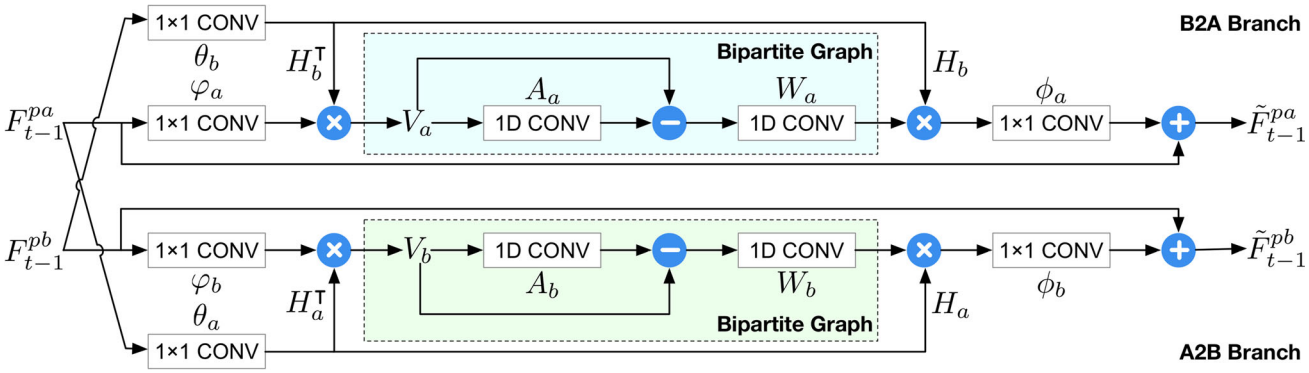


Fig. 3 Illustration of the proposed bipartite graph reasoning (BGR) block t, which consists of two branches, i.e., B2A and A2B. Each branch aims to model cross-contextual information between shape features F_{t-1}^{pa} and F_{t-1}^{pb} in a bipartite graph via GCNs

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:

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

where both functions $\theta_b(\cdot)$ and $\varphi_a(\cdot)$ are implemented using a 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 cross relations between the nodes of the target pose F_{t-1}^{pb} and the source pose F_{t-1}^{pa} (see Fig. 1c).

Cross Reasoning with a 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 long-range cross relations between the nodes from both the source and target poses, which can be formulated as:

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

where $W_a \in \mathbb{R}^{D_a \times D_a}$ denotes the trainable edge weights. We follow (Chen & Rohrbach, 2019; Zhang et al., 2019) and use Laplacian smoothing (Chen & Rohrbach, 2019; Li et al., 2018) 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 the adjacency matrix A_a and the

weights W_a, and then train them by gradient descent in an end-to-end manner.

From Bipartite-Graph Space to Coordinate Space After the cross-reasoning, the new updated 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 (He et al., 2016). This process can be expressed as:

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

where we reuse the projection matrix H_b and apply 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 as 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 behind the proposed BGR block was inspired by the GloRe unit proposed in Chen and Rohrbach (2019). The main difference is that the GloRe unit reasons the relations within the same feature map via a standard graph, while the proposed BGR block reasons the cross 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 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. The IA block aims to simultaneously and interactively enhance F_t^i, F_t^{pa} and F_t^{pb}. Specifically, the shape codes are first concatenated and then fed into two convolutional layers to produce the attention map A_{t-1}. Mathematically,

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

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

Appearance and shape features are crucial to generate the final person image since the appearance feature mainly focus on the texture and color information of clothes, and the shape feature mainly focus on the body orientation and size information. Thus, we propose the "Appearance Code Enhancement" to learn and enhance useful person appearance feature, while the "Shape Code Enhancement" to learn and enhance useful person shape feature. Having both "Appearance Code Enhancement" and "Shape Code Enhancement" together can generate realistic person image.

Appearance Code Enhancement After obtaining A_{t-1}, the appearance F_{t-1}^{i} is enhanced by:

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

where $\otimes$ denotes the element-wise product. By multiplying with the attention map A_{t-1}, the new appearance code F_t^i at certain locations can be either preserved or suppressed.

Shape Code Enhancement As the appearance code gets updated through Eq. (5), the shape code should also be updated to synchronize the change, i.e., update where to sample and put patches given the new appearance code. Therefore, the shape code update should incorporate the new appearance code. Specifically, we concatenate F_t^i, F_{t-1}^{pa} and F_{t-1}^{pb}, and pass them through two convolutional layers to obtain the updated shape codes F_t^{pa} and F_t^{pb} by splitting the result along the channel axis. This process can be formulated as:

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

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

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

4 Part-Aware BiGraphGAN

The proposed part-aware bipartite graph reasoning GAN (i.e., BiGraphGAN++) employs the same framework as BiGraphGAN, presented in Fig. 2, with the only difference being that we need to replace the BGR block from Fig. 2 with the new PBGR from Fig. 4.

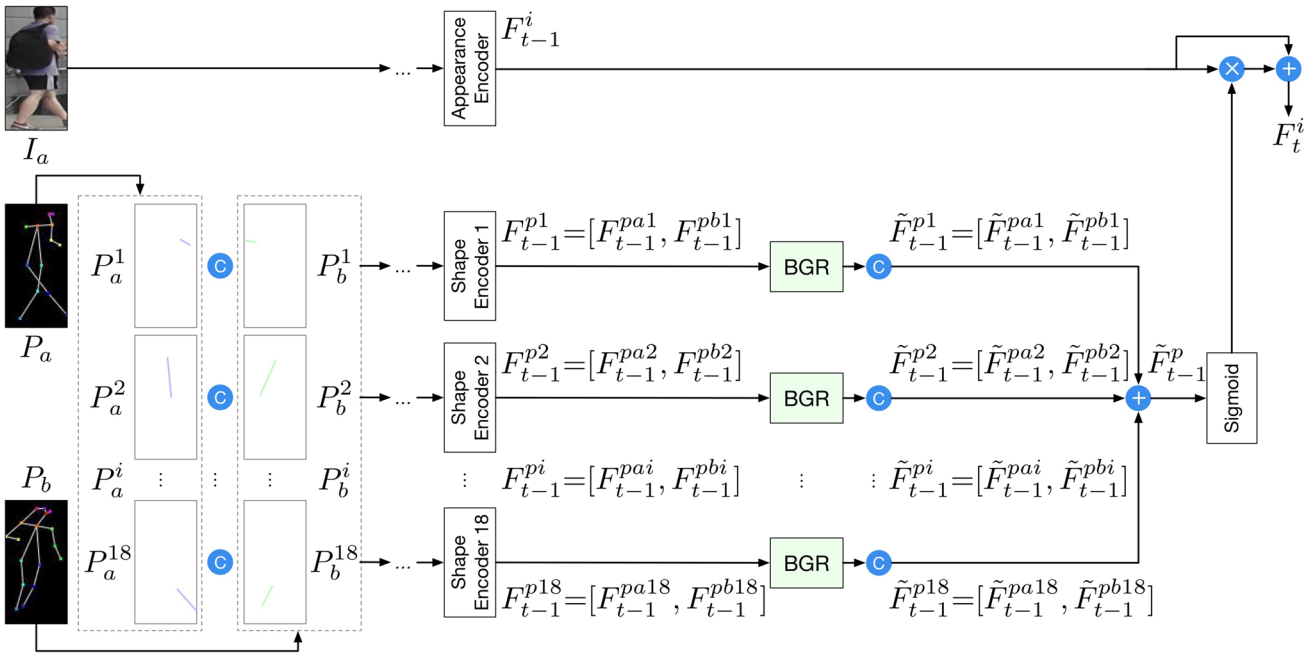


Fig. 4 Illustration of the proposed PBGR block t, which consists of 18 branches. Each branch aims to model local transformations between each source sub-pose F_{t-1}^{pai} and each target sub-pose F_{t-1}^{pbi} in a bipartite graph via a BGR block presented in Fig. 3. Note that the shape encoders can share network parameters, so that no extra parameters are introduced, and the speed of training and testing is not significantly slow down

4.1 Part-Aware Bipartite Graph Reasoning

The framework of the proposed PBGR block is shown in Fig. 4. Specifically, we first follow OpenPose (Cao et al., 2017) and decompose the overall source pose P_a and target pose P_b into 18 different sub-poses (i.e., $\{P_a^i\}_{i=1}^{18}$, and $\{P_b^i\}_{i=1}^{18}$) based on the inherent connection relationships among them. Then the corresponding source and target sub-poses are concatenated and fed into the corresponding shape encoder to generate high-level feature representations.

Consider the t-th block given in Fig. 4. Each source and target sub-pose feature representation can be represented as F_{t-1}^{pai} and F_{t-1}^{pbi}, respectively. Then, the feature pair $\left[F_{t-1}^{pai}, F_{t-1}^{pbi}\right]$ is fed to the i-th BGR block to learn the local transformation for the i-th sub-pose, which can ease the learning and capture the change in pose of each part more precisely. Next, the updated feature representations $\tilde{F}_{t-1}^{pai}$ and $\tilde{F}_{t-1}^{pbi}$ are concatenated to represent the local transformation of the i-th sub-pose, i.e., $\tilde{F}_{t-1}^{pi}=\left[\tilde{F}_{t-1}^{pai}, \tilde{F}_{t-1}^{pbi}\right]$. Finally, we combine all the local transformations from all the different sub-poses to obtain the global transformation between the source pose P_a and target pose P_b, which can be expressed as follows:

$$\tilde{F}_{t-1}^{p} = \tilde{F}_{t-1}^{p1} + \tilde{F}_{t-1}^{p2} + \cdots + \tilde{F}_{t-1}^{pi} + \cdots + \tilde{F}_{t-1}^{p18}. \tag{8}$$

4.2 Part-Aware Interaction and Aggregation

The proposed part-aware IA block aims to simultaneously enhance $\tilde{F}_{t-1}^{p}$ and F_{t-1}^{i}. Specifically, the pose feature $\tilde{F}_{t-1}^{p}$ is fed into a Sigmoid activation layer to produce the attention map A_{t-1}. Mathematically,

$$A_{t-1} = \sigma\left(\tilde{F}_{t-1}^{p}\right), \tag{9}$$

where $\sigma(\cdot)$ denotes the element-wise Sigmoid function. By doing so, A_{t-1} provides important guidance for understanding the spatial deformation of each part region between the source and target poses, specifying which positions in the source pose should be sampled to generate the corresponding target pose.

Appearance Code Enhancement After obtaining A_{t-1}, the appearance F_{t-1}^{i} is enhanced by:

$$F_{t}^{i} = A_{t-1} \otimes F_{t-1}^{i} + F_{t-1}^{i}, \tag{10}$$

where $\otimes$ denotes an element-wise product.

Shape Code Enhancement Next, we concatenate F_{t}^{i} and F_{t-1}^{pi}, and pass them through two convolutional layers to obtain the updated shape codes F_{t}^{pai} and F_{t}^{pbi} by splitting the result along the channel axis. This process can be formulated

Table 1 Quantitative comparison of different methods on Market-1501 and DeepFashion for person pose generation. For all metrics, higher is better.

Method	Market-1501				DeepFashion		
	SSIM ↑	IS ↑	Mask-SSIM ↑	Mask-IS ↑	SSIM ↑	IS ↑	PCKh ↑
PG2 (Ma et al., 2017)	0.253	3.460	0.792	3.435	0.762	3.090	–
DPIG (Ma & Sun, 2018)	0.099	3.483	0.614	3.491	0.614	3.228	–
Deform (Siarohin et al., 2018)	0.290	3.185	0.805	3.502	0.756	3.439	–
C2GAN (Tang et al., 2019c)	0.282	3.349	0.811	3.510	–	–	–
BTF (AlBahar & Huang, 2019)	–	–	–	–	0.767	3.220	–
PG2* (Ma et al., 2017)	0.261	3.495	0.782	3.367	0.773	3.163	0.89
Deform* (Siarohin et al., 2018)	0.291	3.230	0.807	3.502	0.760	3.362	0.94
VUNet* (Esser et al., 2018)	0.266	2.965	0.793	3.549	0.763	3.440	0.93
PATN* (Zhu et al., 2019)	0.311	3.323	0.811	3.773	0.773	3.209	0.96
PoseStylizer* (Huang et al., 2020)	0.312	3.132	0.808	3.729	0.775	3.292	0.96
XingGAN* (Tang et al., 2020b)	0.313	3.506	0.816	**3.872**	0.778	3.476	0.95
BiGraphGAN (Ours)	0.325	3.329	0.818	3.695	0.778	3.430	**0.97**
BiGraphGAN++ (Ours)	**0.334**	**3.592**	**0.822**	3.701	**0.802**	**3.508**	**0.97**
Real Data	1.000	3.890	1.000	3.706	1.000	4.053	1.00

(∗) denotes the results tested on our testing set
Bold values indicate the best results

as:

$$\begin{aligned} F_t^{pi} &= \left[F_t^{pai}, F_t^{pbi}\right] \\ &= \text{Conv}\left(\text{Concat}\left(F_t^i, F_{t-1}^{pi}\right)\right), i = 1, \cdots, 18. \end{aligned} \tag{11}$$

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

5 Model Training

Appearance and Shape Discriminators We adopt two discriminators for adversarial training. Specifically, we feed the image-image pairs (I_a, I_b) and $(I_a, I_b^{'})$ into the appearance discriminator D_{app} to ensure appearance consistency. Meanwhile, we feed the pose-image pairs (P_b, I_b) and $(P_b, I_b^{'})$ into the shape discriminator D_{sha} for shape consistency. Both discriminators (i.e., D_{app} and D_{sha}) and the proposed graph generator G are trained in an end-to-end way, enabling them to enjoy mutual benefits from each other in a joint framework.
Optimization Objectives We follow (Zhu et al., 2019; Tang et al., 2020b) 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:

$$\mathcal{L}_{full} = \lambda_{gan}\mathcal{L}_{gan} + \lambda_{l1}\mathcal{L}_{l1} + \lambda_{per}\mathcal{L}_{per}, \tag{12}$$

where λ_{gan}, λ_{l1}, and λ_{per} control the relative importance of the three objectives. For the perceptual loss, we follow (Zhu et al., 2019; Tang et al., 2020b) and use the $Conv1\_2$ layer.
Implementation Details In our experiments, we follow previous work (Zhu et al., 2019; Tang et al., 2020b) 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. The Adam optimizer (Kingma & Ba, 2015) is employed to learn the proposed BiGraphGAN and BiGraphGAN++ for around 90K iterations with β_1=0.5 and β_2=0.999.

In our preliminary experiments, we found that as T increases, the performance gets better and better. However, when T reaches 9, the proposed models achieve 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 Equation (12), and the number of feature map channels C, are set to 5, 10, 10, and 128, respectively. The proposed BiGraphGAN is implemented in PyTorch (Paszke et al., 2019).

6 Experiments

6.1 Person Pose Synthesis

Datasets We follow previous works (Ma et al., 2017; Siarohin et al., 2018; Zhu et al., 2019) and conduct extensive experiments on two public datasets, i.e., Market-1501 (Zheng et al., 2015) and DeepFashion (Liu et al., 2016). Specifically,

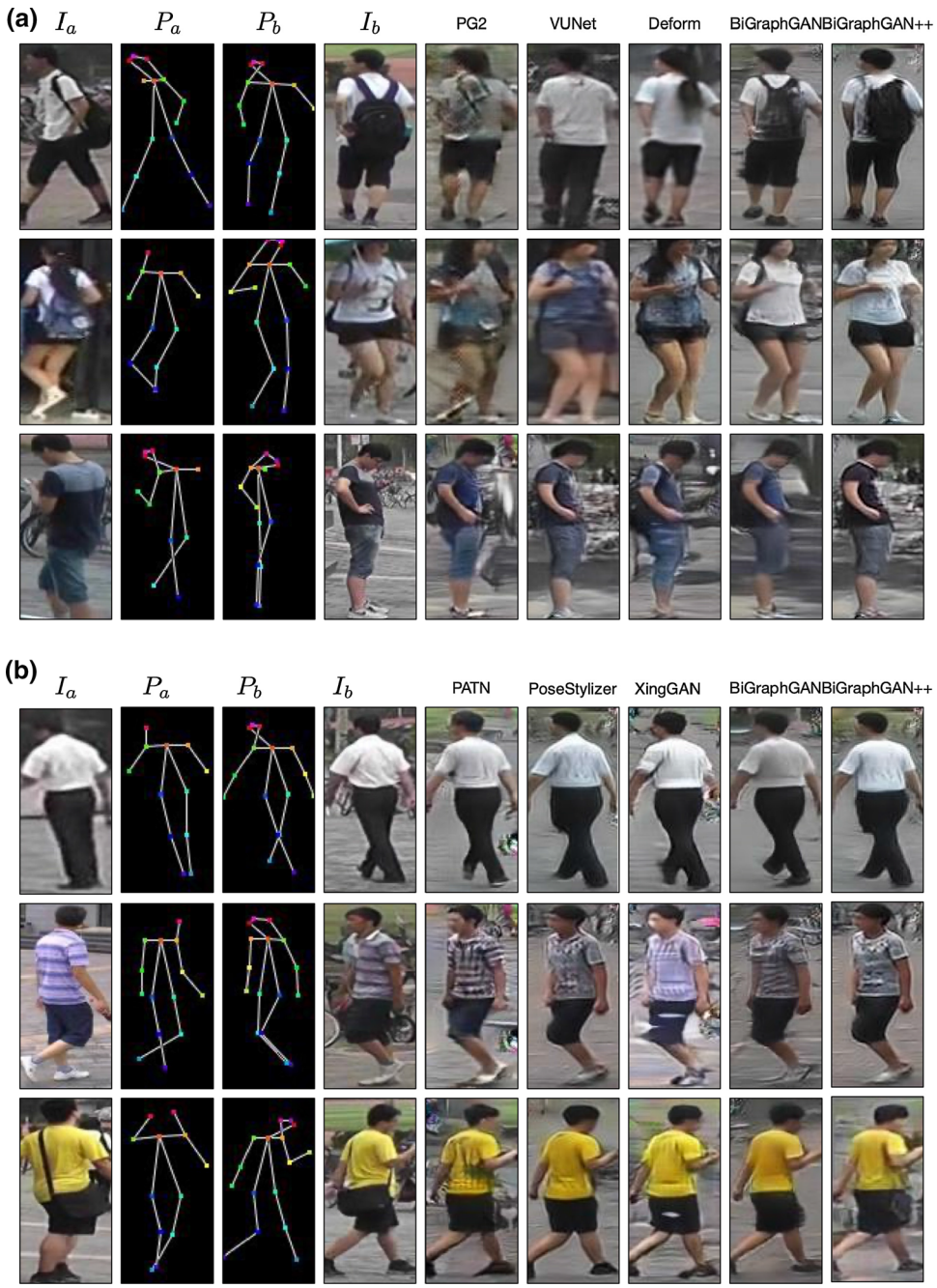


Fig. 5 Qualitative comparisons of person pose generation on Market-1501. **a** From left to right: Source Image (I_a), Source Pose (P_a), Target Pose (P_b), Target Image(I_b), PG2 (Ma et al., 2017), VUNet (Esser et al., 2018), Deform (Siarohin et al., 2018), BiGraphGAN (Ours), and BiGraphGAN++ (Ours). **b** From left to right: Source Image (I_a), Source Pose (P_a), Target Pose (P_b), Target Image(I_b), PATN (Zhu et al., 2019), PoseStylizer (Huang et al., 2020), XingGAN (Tang et al., 2020b), BiGraphGAN (Ours), and BiGraphGAN++ (Ours)

we adopt the training/test split used in Zhu et al. (2019); Tang et al. (2020b) for fair comparison. In addition, images are resized to 128×64 and 256×256 on Market-1501 and DeepFashion, respectively.

Evaluation Metrics We follow (Ma et al., 2017; Siarohin et al., 2018; Zhu et al., 2019) and employ Inception score (IS) (Salimans et al., 2016), structural similarity index measure (SSIM) (Wang et al., 2004), and their masked versions (i.e., Mask-IS and Mask-SSIM) as our evaluation metrics to quantitatively measure the quality of the images generated by different approaches. Moreover, we employ the percentage of correct keypoints (PCKh) score proposed in Zhu et al. (2019) to explicitly evaluate the shape consistency of the person images generated for the DeepFashion dataset.

Quantitative Comparisons We compare the proposed BiGraphGAN and BiGraphGAN++ with several leading person image synthesis methods, i.e., PG2 (Ma et al., 2017), DPIG (Ma & Sun, 2018), Deform (Siarohin et al., 2018), C2GAN (Tang et al., 2019c), BTF (AlBahar & Huang, 2019), VUNet (Esser et al., 2018), PATN (Zhu et al., 2019), PoseStylizer (Huang et al., 2020), and XingGAN (Tang et al.,

Fig. 6 Qualitative comparisons of person pose generation on DeepFashion. **a** From left to right: Source Image (I_a), Source Pose (P_a), Target Pose (P_b), Target Image(I_b), PG2 (Ma et al., 2017), VUNet (Esser et al., 2018), Deform (Siarohin et al., 2018), BiGraphGAN (Ours), and BiGraphGAN++ (Ours). **b** From left to right: Source Image (I_a), Source Pose (P_a), Target Pose (P_b), Target Image(I_b), PATN (Zhu et al., 2019), XingGAN (Tang et al., 2020b), BiGraphGAN (Ours), and BiGraphGAN++ (Ours)

2020b). Note that all of them use the same training data and data augmentation to train the models.

Quantitative comparison results are shown in Table 1. We observe that the proposed methods achieve the best results in most metrics, including SSIM and Mask-SSIM on Market-1501, and SSIM and PCKh on DeepFashion. For other metrics, such as IS, the proposed methods still achieve better scores than the most related model, PATN, on both datasets. These results validate the effectiveness of our proposed methods.

Qualitative Comparisons We also provide visual comparison results on both datasets in Figs. 5 and 6. As shown on the left of both figures, the proposed BiGraphGAN and BiGraphGAN++ generate remarkably better results than PG2 (Ma et al., 2017), VUNet (Esser et al., 2018), and Deform (Siarohin et al., 2018) on both datasets. To further evaluate the effective-

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 (Ma et al., 2017)	11.20	5.50	9.20	14.90
Deform (Siarohin et al., 2018)	22.67	50.24	12.42	24.61
C2GAN (Tang et al., 2019c)	23.20	46.70	–	–
PATN (Zhu et al., 2019)	32.23	63.47	19.14	31.78
BiGraphGAN (Ours)	35.76	65.91	22.39	34.16
BiGraphGAN++ (Ours)	**37.32**	**66.83**	**23.76**	**35.57**

Bold values indicate the best results

Table 3 Quantitative comparison of facial expression image synthesis on the Radboud Faces dataset. For all the metrics except LPIPS, higher is better

Method	AMT ↑	SSIM ↑	PSNR ↑	LPIPS ↓
Pix2pix (Isola et al., 2017)	13.4	0.8217	19.9971	0.1334
GPGAN (Di et al., 2018)	0.3	0.8185	18.7211	0.2531
PG2 (Ma et al., 2017)	28.4	0.8462	20.1462	0.1130
CocosNet (Zhang et al., 2020)	31.3	0.8524	20.7915	0.0985
C2GAN (Tang et al., 2019c)	34.2	0.8618	21.9192	0.0934
BiGraphGAN (Ours)	37.9	0.8644	27.5923	0.0806
BiGraphGAN++ (Ours)	**39.1**	**0.8665**	**29.3917**	**0.0798**

Bold values indicate the best results

ness of the proposed methods, we compare BiGraphGAN and BiGraphGAN++ with the most state-of-the-art models, i.e., PATN (Zhu et al., 2019), PoseStylizer (Huang et al., 2020), and XingGAN (Tang et al., 2020b), on the right of both figures. We again observe that our proposed BiGraphGAN and BiGraphGAN++ generate clearer and more visually plausible person images than PATN, PoseStylizer, and XingGAN on both datasets.

User Study We also follow (Ma et al., 2017; Siarohin et al., 2018; Zhu et al., 2019) and conduct a user study to evaluate the quality of the generated images. Specifically, we follow the evaluation protocol used in Zhu et al. (2019); Tang et al. (2020b) for fair comparison. Comparison results of different methods are shown in Table 2. We see that the proposed methods achieve the best results in all metrics, which further confirms that the images generated by the proposed BiGraphGAN and BiGraphGAN++ are more photorealistic.

6.2 Facial Expression Synthesis

Datasets The Radboud Faces dataset (Langner et al., 2010) is used to conduct experiments on the facial expression generation task. This dataset consists of over 8,000 face images with eight different facial expressions, i.e., neutral, angry, contemptuous, disgusted, fearful, happy, sad, and surprised.

We follow C2GAN (Tang et al., 2019c) and select 67% of the images for training, while the remaining 33% are used for testing. We use the public software OpenFace (Amos et al., 2016) to extract facial landmarks. For the facial expression-to-expression translation task, we combine two different facial expression images of the same person to form an image pair for training (e.g., neutral and angry). Thus, we obtain 5628 and 1407 image pairs for the training and testing sets, respectively.

Evaluation Metrics. We follow C2GAN (Tang et al., 2019c) and first adopt SSIM (Wang et al., 2004), peak signal-to-noise ratio (PSNR), and learned perceptual image patch similarity (LPIPS) (Zhang & Isola, 2018b) for quantitative evaluation. Note that both SSIM and PSNR measure the image quality at a pixel level, while LPIPS evaluates the generated images at a deep feature level. Next, we again follow C2GAN and adopt the amazon mechanical turk (AMT) user study to evaluate the generated facial images.

Quantitative Comparisons To evaluate the effectiveness of the proposed BiGraphGAN, we compare it with several leading facial image generation methods, i.e., Pix2pix (Isola et al., 2017), GPGAN (Di et al., 2018), PG2 (Ma et al., 2017), CocosNet (Zhang et al., 2020), and C2GAN (Tang et al., 2019c).

The results in terms of SSIM, PSNR, and LPIPS are shown in Table 3. We observe that the proposed BiGraphGAN and BiGraphGAN++ achieve the best scores in all three evaluation metrics, confirming the effectiveness of our methods. Notably, the proposed BiGraphGAN is 5.6731 points higher than the current best method (i.e., C2GAN) in the PSNR metric.

Qualitative Comparisons We also provide qualitative results compared with the current leading models in Fig. 7. We observe that GPGAN performs the worst among all comparison models. Pix2pix can generate correct expressions,

Fig. 7 Qualitative comparisons of facial expression translation on Radboud Faces. From left to right: Source Image (I_a), Source Landmark (P_a), Target Landmark (P_b), Target Image (I_b), Pix2pix (Isola et al., 2017), GPGAN (Di et al., 2018), PG2 (Ma et al., 2017), CocosNet (Zhang et al., 2020), C2GAN (Tang et al., 2019c), BiGraphGAN (Ours), and and BiGraphGAN++ (Ours)

but the faces are distorted. Moreover, the results of PG2 tend to be blurry. Compared with these methods, the results generated by the proposed BiGraphGAN are smoother, sharper, and contain more convincing details.

User Study Following C2GAN (Tang et al., 2019c), we conduct a user study to evaluate the quality of the images generated by different models, i.e., Pix2pix (Isola et al., 2017), GPGAN (Di et al., 2018) , PG2 (Ma et al., 2017), CocosNet (Zhang et al., 2020), and C2GAN (Tang et al., 2019c). Comparison results are shown in Table 3. We observe that the proposed BiGraphGAN achieves the best results, which further validates that the images generated by the proposed model are more photorealistic.

6.3 Ablation Study

We perform extensive ablation studies to validate the effectiveness of each component of the proposed BiGraphGAN on the Market-1501 dataset.

Baselines of BiGraphGAN. The proposed BiGraphGAN has six baselines (i.e., B1, B2, B3, B4, B5, B6), as shown in Table 4 and Fig. 8 (left). B1 is our baseline. B2 uses the proposed B2A branch to model the cross relations from the target pose to the source pose. B3 adopts the proposed A2B branch to model the cross relations from the source pose to the target pose. B4 combines both the A2B and B2A branches to model the cross relations between the source pose and the target pose. Note that both GCNs in B4 share parameters. B5 employs a non-sharing strategy between the two GCNs to model the cross relations. B6 is our full model and employs the proposed AIF module to enable the graph generator to attentively determine which part is most useful for generating the final person image.

Ablation Analysis The results of the ablation study are shown in Table 4 and Fig. 8 (left). We observe that both B2 and B3 achieve significantly better results than B1, proving our initial hypothesis that modeling the cross relations between the source and target pose in a bipartite graph will

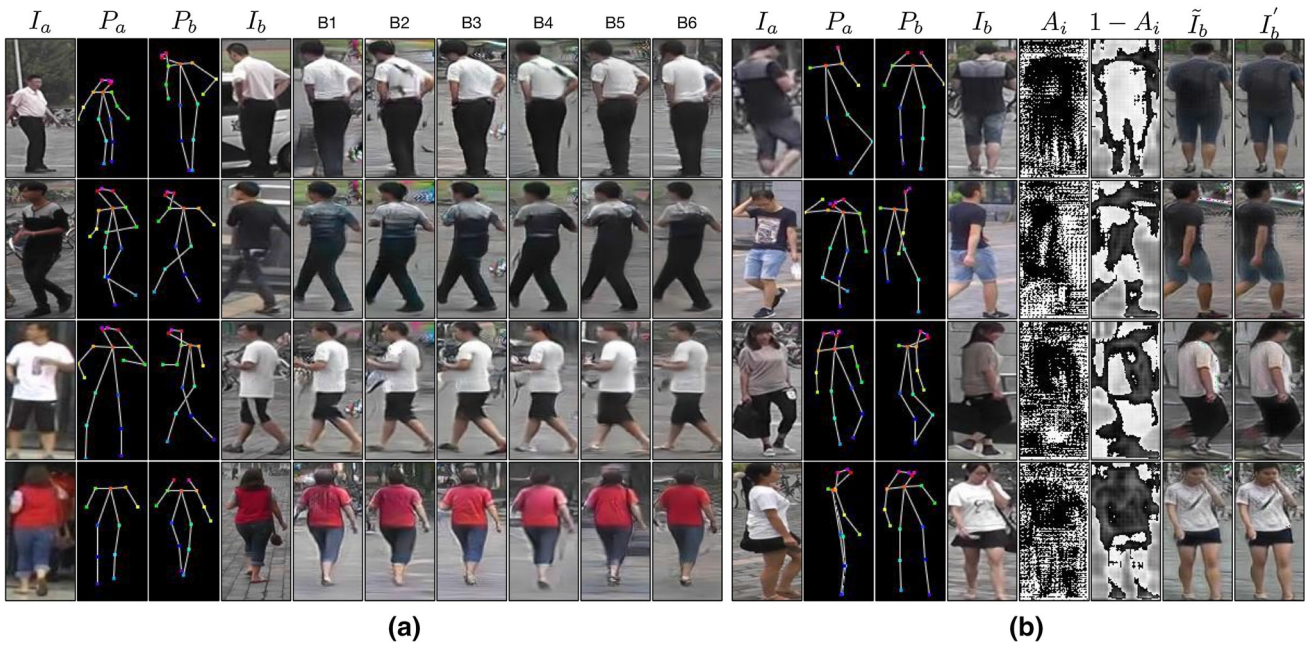


Fig. 8 Qualitative comparison of ablation study on Market-1501. **a** Qualitative comparisons of different baselines of the proposed BiGraphGAN. **b** Visualization of the learned attention masks and intermediate results

Table 4 Ablation study of the proposed BiGraphGAN on Market-1501 for person pose generation. 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**

Bold values indicate the best results

boost the generation performance. In addition, we see that B4 outperforms B2 and B3, demonstrating the effectiveness of modeling the symmetric relations between the source and target poses. B5 achieves better results than B4, which indicates that using two separate GCNs 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. 8 (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 validating our design motivation.

BiGraphGAN versus BiGraphGAN++ We also provide comparison results of BiGraphGAN and BiGraphGAN++ on both Market-1501 and DeepFashion. The results for person pose image generation are shown in Tables 1 and 2. We see that BiGraphGAN++ achieves much better results in most metrics, indicating that the proposed PBGR module does indeed learn the local transformations among body parts, thus improving the generation performance. From the visualization results in Figs. 5 and 6, we can see that BiGraphGAN++ generates more photorealistic images with fewer visual artifacts than BiGraphGAN, on both datasets. The same conclusion can be drawn from the facial expression synthesis task, as shown in Table 3 and Fig. 7. Overall, the proposed BiGraphGAN++ can achieve better reuslts than BiGraphGAN on both challenging tasks, validating the effectiveness of our network design.

7 Conclusion

In this paper, we propose a novel bipartite graph reasoning GAN (BiGraphGAN) framework for both the challenging person pose and facial image generation tasks. We introduce

two novel blocks, i.e., the bipartite graph reasoning (BGR) block and interaction-and-aggregation (IA) block. The former is employed to model the long-range cross relations between the source pose and the target pose in a bipartite graph. The latter is used to interactively enhance both a person's shape and appearance features.

To further capture the detailed local structure transformations among body parts, we propose a novel part-aware bipartite graph reasoning (PBGR) block. Extensive experiments in terms of both human judgments and automatic evaluation demonstrate that the proposed BiGraphGAN achieves remarkably better performance than the state-of-the-art approaches on three challenging datasets. Lastly, we believe that the proposed method will inspire researchers to explore the cross-contextual information in other vision task.

Acknowledgements This work has been partially supported by the EU H2020 project AI4Media (No. 951911).

References

AlBahar, B., & Huang, J.-B. (2019). Guided image-to-image translation with bi-directional feature transformation. In *ICCV*.

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

Balakrishnan, G., Zhao, A., Dalca, A. V., Durand, F., & Guttag J. (2018). In *CVPR: Synthesizing images of humans in unseen poses*.

Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *ICLR*

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*

Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019). Everybody dance now. In *ICCV*.

Chen, X., Bin, Y., Gao, C., Sang, N., & Tang, H. (2020). Relevant region prediction for crowd counting. *Elsevier Neurocomputing*.

Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Yan, S., Feng, J., & Kalantidis, Y. (2019). Graph-based global reasoning networks. In *CVPR*.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*.

Di, X., Sindagi, V. A., & Patel, V. M. (2018). GP-GAN: Gender preserving GAN for synthesizing faces from landmarks. In *ICPR*.

Esser, P., Sutter, E., & Ommer, B. A. (2018). variational u-net for conditional appearance and shape generation. In *CVPR*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Bing, X., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *NeurIPS*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Huang, S., Xiong, H., Cheng, Z.-Q., Wang, Q., Zhou, X., Wen, B., Huan, J., & Dou, D. (2020). Generating person images with appearance-aware pose stylizer. In *IJCAI*.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *CVPR*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *CVPR*.

Kim, J., Kim, M., Kang, H., & Lee, K. (2020). U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR*.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & Knippenberg, A. D. V. (2010). Presentation and validation of the radboud faces database. *Taylor & Francis Cognition and emotion*.

Li, Q., Han, Z., & Wu, X.-M. (2018). Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*.

Liang, D., Wang, R., Tian, X. & Zou, C. (2019). Pcgan: Partition-controlled human image generation. In *AAAI*.

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In CVPR.

Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., & Gao, S. (2019). Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*.

Liu, G., Tang, H., Latapie, H., & Yan, Y. (2020). Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP*.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Luc, G. (2017). Pose guided person image generation. In *NeurIPS*.

Ma, L., Sun, Q. Georgoulis, S., Van Gool, L., Schiele, B., & Fritz, M. (2018). Disentangled person image generation. In *CVPR*.

Mejjati, Y. A., Richardt, C., Tompkin, J., Cosker, D., & Kim, K. I. (2018). In *NeurIPS: Unsupervised attention-guided image-to-image translation*.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Park, T., Liu, M.-Y., Wang, T.-C., & Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *CVPR*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L, et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Peng, W., Shi, J., Xia, Z., & Zhao, G. (2020). Mix dimension in poincaré geometry for 3d skeleton-based action recognition. In *ACM MM*.

Po-Wei, W., Lin, Y.-J.,Chang, C.-H. Chang, E. Y., & Liao, S.-W. (2019). Relgan: Multi-domain image-to-image translation via relative attributes. In *ICCV*.

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2020). Ganimation: One-shot anatomically consistent facial animation. *Springer IJCV, 128*(3), 698–713.

Ren, B., Tang, H., & Sebe, N. (2021). Cascaded cross mlp-mixer gans for cross-view image translation. In *BMVC*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *NeurIPS*.

Shaham, T. R., Dekel, T., & Michaeli, T. (2019). Singan: Learning a generative model from a single natural image. In *ICCV*).

Siarohin, A., Sangineto, E., Lathuilière, S., & Sebe, N. (2018). Deformable gans for pose-based human image generation. In *CVPR*.

Songsong, W., Tang, H., Jing, X.-Y. *Zhao, H.* Qian, J., Sebe, N., & Yan, Y. (2022). Cross-view panorama image synthesis. *IEEE TMM*.

Tang, H., & Sebe, N. (2021). Layout-to-image translation with double pooling generative adversarial networks. *IEEE TIP*.

Tang, H., & Sebe, N. (2021). Total generate: Cycle in cycle generative adversarial networks for generating human faces, hands, bodies, and natural scenes. *IEEE TMM*.

Tang, H., Bai, S., & Sebe, N. (2020). Dual attention gans for semantic image synthesis. In *ACM MM*.

Tang, H., Bai, S., Zhang, L., Torr, P. H. S., & Sebe, N. (2020). In *ECCV: Xinggan for person image generation*.

Tang, H., Bai, S. Torr, P. H. S., & Sebe, N. (2020). In *BMVC: Bipartite graph reasoning gans for person image generation*.

Tang, H., Chen, X., Wang, W., Dan, X. Corso, J. J., Sebe, N. & Yan, Y. (2019). In *FG: Attribute-guided sketch generation*.

Tang, H., Liu, H., & Sebe, N. (2020). Unified generative adversarial networks for controllable image-to-image translation. *IEEE TIP*.

Tang, H., Liu, H., Dan, X., Torr, P. H. S., & Sebe, N. (2021). Attention-gan: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE TNNLS.

Tang, H., Shao, L., Torr, P. H. S. & Sebe, N. (2022). IEEE TPAMI: Local and global gans with semantic-aware upsampling for image generation.

Tang, H., Wang, W., Wu, S., Chen, X., Xu, D., Sebe, N., & Yan, Y. (2019). Expression conditional gan for facial expression-to-expression translation. In *ICIP*.

Tang, H., Wang, W., Xu, D., Yan, Y., & Sebe, N. (2018). Gesturegan for hand gesture-to-gesture translation in the wild. In *ACM MM*.

Tang, H., Xu, D., Liu, G., Wang, W., Sebe, N., & Yan, Y. (2019). Cycle in cycle generative adversarial networks for keypoint-guided image generation. In *ACM MM*.

Tang, H., Xu, D., Sebe, N., Wang, Y., Corso, J. J., & Yan,Y. (2019). Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*.

Tang, H., Xu, D., Yan, Y., Torr, P. H.S., & Sebe, N. (2020). Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*.

Tang, H, Xu, D., Sebe, N., & Yan, Y. (2019). Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *IJCNN*.

Tao, M., Tang, H., Fei, W., Jing, X.-Y., Bao, B.-K., & Xu, C. (2022). Df-gan: A simple and effective baseline for text-to-image synthesis. In *CVPR*.

Wang, X., & Gupta, A. (2018). Videos as space-time region graphs. In *ECCV*.

Wang, Z., Zheng, L., Li, Y., & Wang, S. (2019). Linkage based face clustering via graph convolution network. In *CVPR*.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE TIP, 13*(4), 600–612.

Wu, S., Tang, H., Jing, X.-Y., Qian, J., Sebe, N., Yan, Y., & Zhang, Q. (2022). Cross-view panorama image synthesis with progressive attention gans. *Elsevier PR*.

Xu, Z., Lin, T., Tang, H., Li, F., He, D., Sebe, N., Timofte, R., Van Gool, L., & Ding, E. (2022). Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In *CVPR*.

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.

Yang, L., Zhan, X., Chen, D., Yan, J., Change Loy, C., & Lin, D. (2019). In *CVPR: Learning to cluster faces on an affinity graph*.

Zanfir, M., Popa, A.-I., Zanfir, A., & Sminchisescu, C. (2018). Human appearance transfer. In *CVPR*.

Zhang, J., Chen, J., Tang, H., Sangineto, E., Wu, P., Yan, Y., Sebe, N., & Wang, W. (2022). Unsupervised high-resolution portrait gaze correction and animation. *IEEE TIP*.

Zhang, R., Isola, P., Efros, A., Shechtman, E., & Wang, O. (2018). In *CVPR: The unreasonable effectiveness of deep features as a perceptual metric*.

Zhang, L., Li, X., Arnab, A., Yang, K., Tong, Y., & Torr, P. H. S. (2019). Dual graph convolutional network for semantic segmentation. In *BMVC*.

Zhang, J., Sangineto, E., Tang, H., Siarohin, A., Zhong, Z., Sebe, N., & Wang, W. (2022). 3D-aware semantic-guided generative model for human synthesis. In *ECCV*.

Zhang, J., Shu, Y., Xu, S., Cao, G, Zhong, F., Liu, M., & Qin, X. (2018). Sparsely grouped multi-task generative adversarial networks for facial attribute manipulation. In *ACM MM*.

Zhang, P., Zhang, B., Chen, D., Yuan, L., & Wen, F. (2020). Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*.

Zheng, L., Liyue Shen, L., Tian, S. W., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *ICCV*.

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., & Bai, X. (2019). Progressive pose attention transfer for person image generation. In *CVPR*.