XingGAN for Person Image Generation

Hao Tang[1,2], Song Bai[2], Li Zhang[2], Philip H.S. Torr[2], and Nicu Sebe[1,3]

[1]University of Trento (`hao.tang@unitn.it`)
[2]University of Oxford   [3]Huawei Research Ireland

Abstract. We propose a novel Generative Adversarial Network (Xing-GAN or CrossingGAN) for person image generation tasks, *i.e.*, translating the pose of a given person to a desired one. The proposed Xing generator consists of two generation branches that model the person's appearance and shape information, respectively. Moreover, we propose two novel blocks to effectively transfer and update the person's shape and appearance embeddings in a crossing way to mutually improve each other, which has not been considered by any other existing GAN-based image generation work. Extensive experiments on two challenging datasets, *i.e.*, Market-1501 and DeepFashion, demonstrate that the proposed XingGAN advances the state-of-the-art performance both in terms of objective quantitative scores and subjective visual realness. The source code and trained models are available at https://github.com/Ha0Tang/XingGAN.

Keywords: Generative Adversarial Networks (GANs), Person Image Generation, Appearance Cues, Shape Cues

1 Introduction

The problem of person image generation aims to generate photo-realistic person images conditioned on an input person image and several desired poses. This task has a wide range of applications such as person image/video generation [41,9,2,11,19] and person re-identification [45,28]. Exiting methods such as [21,22,31,45,35] have achieved promising performance on this challenging task. For example, Zhu *et al.* [45] recently proposed a conditional GAN model that comprises a sequence of pose-attentional transfer blocks. Wherein, each block transfers certain regions it attends to and progressively generates the desired person image.

Although [45] performed an interesting exploration, we still observe unsatisfactory aspects and visual artifacts in the generated person images due to several reasons. First, [45] stacks several convolution layers to generate the attention maps of the shape features, then the generated attention maps are used to attentively highlight the appearance features. Since convolutional operations are building blocks that process one local neighborhood at a time, this means that they cannot capture the joint influence between the appearance and the shape features. Second, the attention maps in [45] are only produced by using one single modality, *i.e.*, the pose, leading to insufficiently accurate correlations
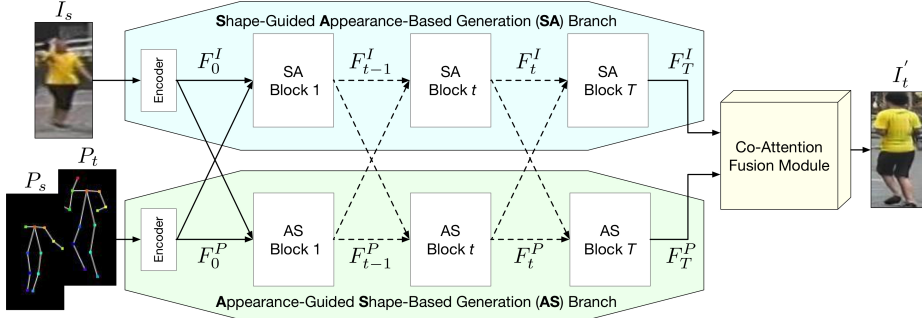
Fig. 1: Overview of the proposed Xing generator. Both the Shape-guided Appearance-based generation (SA) and the Appearance-guided Shape-based generation (AS) branches consist of a sequence of SA and AS blocks in a crossing way. All these components are trained in an end-to-end fashion so that the SA branch and AS branch can benefit from each other to generate more shape-consistent and appearance-consistent person images.

for both modalities (*i.e.*, the pose and the image modality), and thus misguiding the image generation.

Based on these observations, we propose a novel Generative Adversarial Network (XingGAN or CrossingGAN), which consists of a Xing generator, a shape-guided discriminator, and an appearance-guided discriminator. The overall framework is shown in Fig. 1. The Xing generator consists of three parts, *i.e.*, a Shape-guided Appearance-based generation (SA) branch, an Appearance-guided Shape-based generation (AS) branch, and a co-attention fusion module. Specifically, the proposed SA branch contains a sequence of SA blocks, which aim to progressively update the appearance representation under the guidance of the shape representation, while the proposed AS branch contains a sequence of AS blocks, which aim to progressively update the shape representation under the guidance of the appearance representation. We also present a novel crossing operation in both SA and AS blocks to capture the joint influence between the image modality and the pose modality by creating attention maps jointly produced by both modalities. Moreover, we introduce a co-attention fusion model to better fuse the final appearance and shape features to generate the desired person images. We present an appearance-guided discriminator and a shape-guided discriminator to jointly judge how likely is that the generated image contains the same person in the input image and how well the generated image aligns with the targeted pose, respectively. The proposed XingGAN is trained in an end-to-end fashion so that the generation branches can enjoy the mutually improved benefits from each other.

We conduct extensive experiments on two challenging datasets, *i.e.*, Market-1501 [44] and DeepFashion [20]. Qualitative and quantitative results demonstrate that XingGAN achieves better results than state-of-the-art methods, regarding both visual fidelity and alignment with targeted person poses.

To summarize, the contributions of our paper are three-fold:

- We propose a novel XingGAN (or CrossingGAN) for person image generation. It explores cascaded guidance with two different generation branches, and aims at progressively producing a more detailed synthesis from both person shape and appearance embeddings.
- We propose SA and AS blocks, which effectively transfer and update person shape and appearance features in a crossing way to mutually improve each other, and are able to significantly boost the quality of the final outputs.
- Extensive experiments clearly demonstrate the effectiveness of XingGAN, and show new state-of-the-art results on two challenging datasets, *i.e.*, Market-1501 [44] and DeepFashion [20].

2   Related Work

Generative Adversarial Networks (GANs) [8] consist of a generator and a discriminator where the goal of the generator is to produce photo-realistic images so that the discriminator cannot tell the generated images apart from real images. GANs have shown the capability of generating photo-realistic images [3,14,30]. However, it is still hard for vanilla GANs to generate images in a controlled setting. To fix this limitation, Conditional GANs (CGANs) [23] have been proposed.

Image-to-Image Translation aims to learning the translation mapping between target and input images. CGANs have achieved decent results in pixel-wise aligned image-to-image translation tasks [12,34,1]. For example, Isola *et al.* propose Pix2pix [12], which adopts CGANs to generate the target domain images based on the input domain images, such as photo-to-map, sketch-to-image, and night-to-day. However, pixel-wise alignment is not suitable for person image generation tasks due to the shape deformation between the input person image and target person image.

Person Image Generation. To remedy this, several works started to use poses to guide person image generation [21,22,31,7,35,45]. For example, Ma *et al.* first present PG2 [21], which is a two-stage model to generate the target person images based on an input image and the target poses. Moreover, Siarohin *et al.* propose PoseGAN [31], which requires an extensive affine transformation computation to deal with the input-output misalignment caused by pose differences. Zhu *et al.* propose Pose-Transfer [45], which contains a sequence of pose-attentional transfer blocks to generate the target person image progressively. Besides the aforementioned supervised methods, several works focus on solving this task in an unsupervised setting [27,33]. For instance, Pumarola *et al.* propose an unsupervised framework [27] to generate person images, which induces some geometric errors as revealed in their paper.

Note that the aforementioned methods adopt human keypoints or skeleton as pose guidance, which are usually extracted by using OpenPose [4]. In addition, several works adopt DensePose [24], 3D pose [18], and segmented pose [6] to generate person images because they contain more information about body depth and part segmentation, producing better results with more texture details. However, the keypoint-based pose representation is much cheaper and more flexible

than the DensePose, 3D pose, segmented pose representations, and can be more easily applied to practical applications. Therefore, we favor keypoint-based pose representation in this paper.

Image-Guidance Conditioning Schemes. Recently, there were proposed many schemes to incorporate the extra guidance (*e.g.*, human poses [21,45], segmentation maps [25,36,37], facial landmarks [35,42], *etc*) into an image-to-image translation model, which can be divided into four categories, *i.e.*, input concatenation [35,40,43], feature concatenation [21,22,7,18,16,17], one-way guidance-to-image interaction [31,25,10,26], two-way guidance-and-image interaction [45,1,5].

The most straightforward way of conditioning the guidance is to concatenate the input image and the guidance along the channel dimension. For example, C2GAN [35] takes the input person image and the targeted poses as input to output the corresponding targeted person images. Instead of concatenating the guidance and the image at the input, several works [21,22,7] concatenate their feature representations at a certain layer. For instance, PG2 [21] concatenates the embedded pose feature with the embedded image feature at the bottleneck fully connected layer. Another more general scheme is to use the guidance to guide the generation of the image. For example, Siarohin *et al.* [31] first learn an affine transformation between the input and the target pose, then they use it to 'move' the feature maps between the input image and the targeted image. Unlike existing one-way guidance-to-image interaction schemes that allow information flow only from the guidance to the input image, a recent scheme, *i.e.*, two-way guidance-and-image interaction, also considers the information flow from the input image back to the guidance [45,1]. For example, Zhu *et al.* [45] propose an attention-based GAN model to simultaneously update the person's appearance and shape features under the guidance of each other, and show that the proposed two-way guidance-and-image interaction strategy leads to better performance on person image generation tasks.

Contrary to the existing two-way guidance-and-image interaction schemes [45,1] that allow both the image and guidance to guide and update each other in a local way, we show that the proposed cross-conditioning strategy can further improve the performance of person image generation tasks.

3 Xing Generative Adversarial Networks

We start by presenting the details of the proposed XingGAN (Fig. 1) consisting of three parts, *i.e.*, a Shape-guided Appearance-based generation (SA) branch modeling the person shape representation, an Appearance-guided Shape-based generation (AS) branch modeling the person appearance representation, and a Co-Attention Fusion (CAF) module for fusing these two branches. In the following, we first present the design of the two proposed generation branches, and then introduce the co-attention fusion module. Lastly, we present the proposed two discriminators, the overall optimization objective and implementation details.

The inputs of the proposed Xing generator are the source image I_s, the source pose P_s, and the target pose P_t. The goal is to translate the pose of the
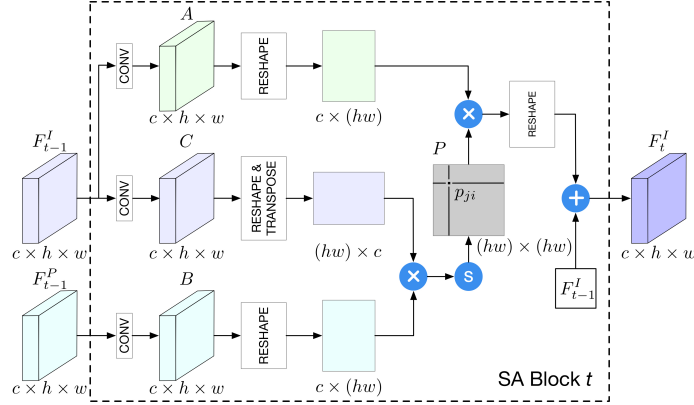
Fig. 2: Structure of the proposed SA block which takes the previous appearance code F_{t-1}^I and the previous shape code F_{t-1}^P as input and obtains the appearance code F_t^I in a crossed non-local way. The symbols $\oplus$, $\otimes$ and $\text{\textcircled{s}}$ and $\text{\textcircled{c}}$ denote element-wise addition, element-wise multiplication, Softmax activation, and channel-wise concatenation, respectively.

person in the source image I_s from the source pose P_s to the target pose P_t, thus synthesizing a photo-realistic person image I_t'. In this way, the source image I_s provides the appearance information and the poses (P_s, P_t) provide the shape information to the Xing generator for synthesizing the desired person image.

Shape-Guided Appearance-Based Generation. The proposed Shape-guided Appearance-based generation (SA) branch consists of an image encoder and a series of the proposed SA blocks. The source image I_s is first fed into the image encoder to produce the appearance code F_0^I, as shown in Fig. 1. The encoder consists of two convolutional layers in our experiments. The SA branch contains several cascaded SA blocks which progressively update the initial appearance code F_0^I to the final appearance code F_T^I under the guidance of the AS branch. As we can see in Fig. 1, all SA blocks have an identical network structure. Consider the t-th block in Fig. 2, whose inputs are the appearance code $F_{t-1}^I \in \mathbb{R}^{c \times h \times w}$ and the shape code $F_{t-1}^P \in \mathbb{R}^{c \times h \times w}$. The output is the refined appearance code $F_t^I \in \mathbb{R}^{c \times h \times w}$. Specifically, given the appearance code F_{t-1}^I, we first feed it into a convolution layer to generate a new appearance code C, where $C \in \mathbb{R}^{c \times h \times w}$. Then we reshape C to $\mathbb{R}^{c \times (hw)}$, where $n{=}hw$ is the number of pixels. At the same time, the SA block receives the shape code F_{t-1}^P from the AS branch, which is also fed into a convolution layer to produce a new shape code $B \in \mathbb{R}^{c \times h \times w}$ and then reshape to $\mathbb{R}^{c \times (hw)}$. After that, we perform a matrix multiplication between the transpose of C and B, and apply a Softmax layer to produce a correlation matrix $P \in \mathbb{R}^{(hw) \times (hw)}$,

$$p_{ji} = \frac{\exp(B_i C_j)}{\sum_{i=1}^n \exp(B_i C_j)}, \tag{1}$$

where p_{ji} measures the impact of the i-th position of B on the j-th position of the appearance code C. In this crossing way, the SA branch can capture more joint

influence between the appearance code F_{t-1}^I and shape code F_{t-1}^P, producing a richer appearance code F_t^I.

Note that Eq. (1) has a close relationship with the non-local operator proposed by Wang *et al.* [38]. The major difference is that the non-local operator in [38] computes the pairwise similarity within the same feature map to incorporate global information, whereas the proposed crossing way computes the pairwise similarity between different feature maps, *i.e.*, the person appearance and shape feature maps.

After that, we feed F_{t-1}^I into a convolution layer to produce a new appearance code $A \in \mathbb{R}^{c \times h \times w}$ and reshape it to $\mathbb{R}^{c \times (hw)}$. We then perform a matrix multiplication between A and the transpose of P and reshape the result to $\mathbb{R}^{c \times h \times w}$. Finally, we multiply the result by a scale parameter α and conduct an element-wise sum operation with the original appearance code F_{t-1}^I to obtain the refined appearance code $F_t^I \in \mathbb{R}^{c \times h \times w}$,

$$F_t^I = \alpha \sum_{i=1}^{n} (p_{ji} A_i) + F_{t-1}^I, \tag{2}$$

where α is 0 in the beginning and but is gradually updated. By doing so, each position of the refined appearance code F_t^I is a weighted sum of all positions of the shape code F_{t-1}^P and the previous appearance code F_{t-1}^I. Thus, it has a global contextual view between F_{t-1}^P and F_{t-1}^I, and it selectively aggregates useful contexts according to the correlation matrix P.

Appearance-Guided Shape-Based Generation. In our preliminary experiments, we observe that only the SA generation branch is not sufficient to learn such a complex deformable translation process. Intuitively, since the shape features can guide the appearance features, we believe the appearance features can also be used to guide the shape features in turn. Therefore, we also propose an Appearance-guided Shape-based generation (AS) branch. The proposed AS branch mainly consists of a pose encoder and a sequence of AS blocks, as shown in Fig. 1. The source pose P_s and target pose P_t are first concatenated along the channel dimension and then fed into the pose encoder to produce the initial shape representation F_0^P. The pose encoder has the same network structure as the image encoder. Note that to capture the dependency between the two poses, we only adopt one pose encoder.

The AS branch contains several cascaded AS blocks, which progressively update the initial shape code F_0^P to the final shape code F_T^P under the guidance of the SA branch. All AS blocks have the same network structure, as illustrated in Fig. 1. Consider the t-th block in Fig. 3, whose inputs are the shape code $F_{t-1}^P \in \mathbb{R}^{c \times h \times w}$ and the appearance code $F_{t-1}^I \in \mathbb{R}^{c \times h \times w}$. The output is the refined shape code $F_t^P \in \mathbb{R}^{c \times h \times w}$.

Specifically, given the shape code F_{t-1}^P, we first feed it into a convolution layer to generate a new shape code H, where $H \in \mathbb{R}^{c \times h \times w}$. We then reshape H to $\mathbb{R}^{c \times (hw)}$. At the same time, the AS block receives the appearance code F_{t-1}^I from the SA branch, which is also fed into a convolution layer to produce a new appearance code E and then reshape it to $\mathbb{R}^{c \times (hw)}$. After that, we perform a
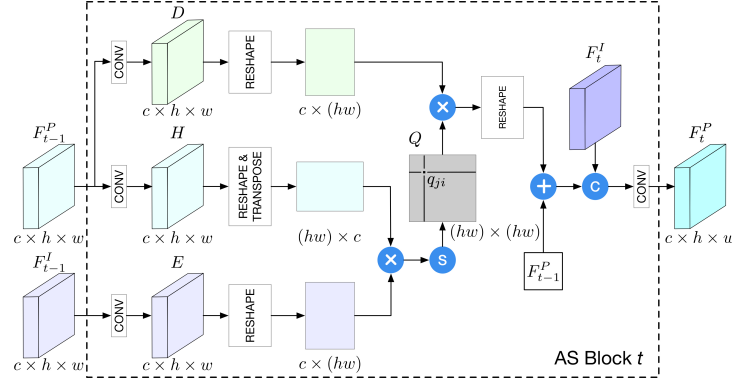
Fig. 3: Structure of the proposed AS block, which takes the previous shape code F_{t-1}^P and the previous appearance code F_{t-1}^I as inputs and obtains the shape code F_t^P in a crossing way. The symbols $\oplus$, $\otimes$ and ⓢ and ⓒ denote element-wise addition, element-wise multiplication, Softmax activation, and channel-wise concatenation, respectively.

matrix multiplication between the transpose of H and E, and apply a Softmax layer to produce another correlation matrix $Q \in \mathbb{R}^{(hw) \times (hw)}$,

$$q_{ji} = \frac{\exp(E_i H_j)}{\sum_{i=1}^{n} \exp(E_i H_j)}, \tag{3}$$

where q_{ji} measures the impact of i-th position of E on the j-th position of the shape code H. $n{=}hw$ is the number of pixels.

Meanwhile, we feed F_{t-1}^P into a convolution layer to produce a new shape code $D \in \mathbb{R}^{c \times h \times w}$ and reshape it to $\mathbb{R}^{c \times (hw)}$. We then perform a matrix multiplication between D and the transpose of Q and reshape the result to $\mathbb{R}^{c \times h \times w}$. Finally, we multiply the result by a scale parameter β and conduct an element-wise sum operation with the original shape code F_{t-1}^P. The result is then concatenated with the appearance code F_t^I and fed into a convolution layer to obtain the updated shape code $F_t^P \in \mathbb{R}^{c \times h \times w}$,

$$F_t^P = \text{Concat}(\beta \sum_{i=1}^{n} (q_{ji} D_i) + F_{t-1}^P, F_t^I), \tag{4}$$

where $\text{Concat}(\cdot)$ denotes the channel-wise concatenation operation and β is a parameter. By doing so, each position in the refined shape code F_t^P is a weighted sum of all positions in the appearance code F_{t-1}^I and previous shape code F_{t-1}^P.

Co-Attention Fusion. The proposed Co-Attention Fusion (CAF) module consists of two parts, *i.e.*, generating intermediate results and co-attention maps. These co-attention maps are used to spatially select from both the intermediate generations and the input image, and are combined to synthesize a final output. This idea of the proposed CAF module comes from the multi-channel attention selection module in SelectionGAN [36]. However, there are three differences: (i)

We use two generation branches to generate intermediate results, *i.e.*, SA branch and AS branch. (ii) Attention maps are generated by the combination of both shape and appearance features, so the model learns more correlations between the two features. (iii) We also produce the input attention map, which aims to select useful content from the input image for generating the final image.

We consider two directions to generate intermediate results. One is generating multiple intermediate image synthesis results from the final appearance code F_T^I, and the other is generating multiple intermediate image synthesis results from the final shape code F_T^P. Specifically, the appearance code F_T^I is fed into a decoder to generate N intermediate results $I^I = \{I_i^I\}_{i=1}^N$, and followed by a Tanh activation function. Meanwhile, the final shape code F_T^P is fed into another decoder to generate another N intermediate results $I^P = \{I_i^P\}_{i=1}^N$, and also followed by a Tanh activation function. Both can be formulated as,

$$
\begin{aligned}
I_i^I &= \mathrm{Tanh}(F_T^I W_i^I + b_i^I), & \text{for } i = 1, \cdots, N \\
I_i^P &= \mathrm{Tanh}(F_T^P W_i^P + b_i^P), & \text{for } i = 1, \cdots, N
\end{aligned}
\tag{5}
$$

where two convolution operations are performed with N convolutional filters $\{W_i^I, b_i^I\}_{i=1}^N$ and $\{W_i^P, b_i^P\}_{i=1}^N$. Thus, the $2N$ intermediate results and the input image I_s can be regarded as the candidate image pool.

To generate the co-attention map which reflects the correlation between the appearance F_T^I and shape F_T^P codes, we first stack both F_T^I and F_T^P along the channel axes, and then feed them into a group of filters $\{W_i^A, b_i^A\}_{i=1}^{2N+1}$ to generate the corresponding $2N+1$ co-attention maps,

$$
I_i^A = \mathrm{Softmax}(\mathrm{Concat}(F_T^I, F_T^P) W_i^A + b_i^A), \quad \text{for } i = 1, \cdots, 2N+1
\tag{6}
$$

where Softmax is a channel-wise Softmax function used for the normalization, and $\mathrm{Concat}(\cdot)$ denotes the channel-wise concatenation operation. Finally, the learned co-attention maps are used to perform a channel-wise selection from each intermediate generation and the input image as follows,

$$
I_t^{'} = (I_1^A \otimes I_1^I) \oplus \cdots (I_{2N}^A \otimes I_{2N}^P) \oplus (I_{2N+1}^A \otimes I_s),
\tag{7}
$$

where $I_t^{'}$ represents the final synthesized person image selected from the multiple diverse results and the input image. $\otimes$ and $\oplus$ denote the element-wise multiplication and addition, respectively.

Optimization Objective. We use three different losses as our full optimization objective, *i.e.*, adversarial loss $\mathcal{L}_{gan}$, pixel loss $\mathcal{L}_{l1}$, and perceptual loss $\mathcal{L}_p$,

$$
\min_G \max_{D_I, D_P} \mathcal{L} = \lambda_{gan} \mathcal{L}_{gan} + \lambda_{l1} \mathcal{L}_{l1} + \lambda_p \mathcal{L}_p,
\tag{8}
$$

where λ_{gan}, λ_{l1} and λ_p are the weights, measuring corresponding contributions of each loss to the total loss $\mathcal{L}$. The total adversarial loss is derived from the appearance-guided discriminator D_I and the shape-guided discriminator D_P, which aims to judge how likely is that $I_t^{'}$ contains the same person in I_s and how well $I_t^{'}$ aligns with the target pose P_t, respectively. The $L1$ pixel loss is

used to compute the difference between the generated image $I_t^{'}$ and the real target image I_t, $i.e.$, $\mathcal{L}_{l1}=||I_t - I_t^{'}||_1$. The perceptual loss $\mathcal{L}_p$ is used to reduce pose distortions and make the generated images look more natural and smooth, $i.e.$, $\mathcal{L}_p=||\phi(I_t) - \phi(I_t^{'})||_1$, where ϕ denotes the outputs of several layers in the pre-trained VGG19 network [32].

Implementation Details. We follow the training procedures of GANs and alternatively train the proposed Xing generator G and two discriminators (D_I, D_P). During training, G takes I_s, P_s and P_t as input and outputs a translated person image $I_t^{'}$ with target pose P_t. Specifically, I_s is fed to the SA branch, and P_s, P_t are fed to the AS branch. For the adversarial training, (I_s, I_t) and (I_s, $I_t^{'}$) are fed to the appearance-guided discriminator D_P for ensuring appearance consistency. (P_t, I_t) and (P_t, $I_t^{'}$) are fed to the shape-guided discriminator D_P for ensuring shape consistency.

Adam optimizer [15] is used to train the proposed XingGAN for around 90K iterations with β_1=0.5 and β_2=0.999. We set T=9 in the proposed Xing generator and N=10 in the proposed co-attention fusion module on both datasets. λ_{gan}, λ_{l1} and λ_p in Eq. (8) are set to 5, 50 and 50, respectively. For the decoders, the kernel size of convolutions for generating the intermediate images and co-attention maps are 3×3 and 1×1, respectively.

4   Experiments

Datasets. We follow [21,31,45] and conduct experiments on two challenging datasets, $i.e.$, Market-1501 [44] and DeepFashion [20]. Images on Market-1501 and DeepFashion are rescaled to 128×64 and 256×256, respectively. To generate human skeletons as training data, we employ OpenPose [4] to extract human joints. In this way, both P_s and P_t consist of an 18-channel heat map encoding the positions of 18 joints of a human body. We also filter out images where no human is detected. Thus, we collect 101,966 training pairs and 8,570 testing pairs on DeepFashion. For Market-1501, we have 263,632 training and 12,000 testing pairs. Note that to better evaluate the proposed XingGAN, the person identities of the training set do not overlap with those of the testing set.

Evaluation Metrics. We follow [21,31,45] and adopt Structure Similarity (SSIM) [39], Inception Score (IS) [29], and their masked versions, $i.e.$, Mask-SSIM and Mask-IS, as the evaluation metrics. Moreover, we adopt the PCKh score proposed in [45] to explicitly assess the shape consistency.

Quantitative Comparisons. We compare the proposed XingGAN with several leading methods, $i.e.$, PG2 [21], DPIG [22], VUnet [7], PoseGAN [31], Pose-Warp [2], CMA [5], C2GAN [35], BTF [1] and Pose-Transfer [45]. Quantitative results measured by SSIM, IS, Mask-SSIM, Mask-IS, and PCKh metrics are shown in Table 1. Note that previous works [21,31] did not release the train/test split, thus we use their well-trained models and re-evaluate their performance on our testing set as in Pose-Transfer [45]. Although our testing set inevitably includes some of their training samples, XingGAN still achieves the best results in terms of SSIM, IS, Mask-SSIM, and Mask-IS metrics on both datasets. For

Table 1: Quantitative results 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 [21]	0.253	3.460	0.792	3.435	-	0.762	3.090	-
DPIG [22]	0.099	3.483	0.614	3.491	-	0.614	3.228	-
PoseGAN [31]	0.290	3.185	0.805	3.502	-	0.756	3.439	-
C2GAN [35]	0.282	3.349	0.811	3.510	-	-	-	-
BTF [1]	-	-	-	-	-	0.767	3.220	-
PG2∗ [21]	0.261	3.495	0.782	3.367	0.73	0.773	3.163	0.89
PoseGAN∗ [31]	0.291	3.230	0.807	3.502	**0.94**	0.760	3.362	0.94
VUnet∗ [7]	0.266	2.965	0.793	3.549	0.92	0.763	3.440	0.93
PoseWarp∗ [2]	-	-	-	-	-	0.764	3.368	0.93
CMA∗ [5]	-	-	-	-	-	0.768	3.213	0.92
Pose-Transfer∗ [45]	0.311	3.323	0.811	3.773	**0.94**	0.773	3.209	**0.96**
XingGAN (Ours)	**0.313**	**3.506**	**0.816**	**3.872**	0.93	**0.778**	**3.476**	0.95
Real Data	1.000	3.890	1.000	3.706	1.00	1.000	4.053	1.00

the PCKh metric, [45] obtains slightly better results than XingGAN. However, we observe that the images generated by XingGAN are more realistic and have less visual artifacts than those generated by [45] (see Fig. 4 and 5).

Qualitative Comparisons. Results compared with PG2 [21], VUnet [7] and PoseGAN [31] are shown on the left of Fig. 4 and 5. We can see that the proposed XingGAN achieves much better results than PG2, VUnet, and PoseGAN on both datasets, especially at appearance details and the integrity of generated persons. Moreover, to evaluate the effectiveness of XingGAN, we compare it with a stronger baseline, *i.e.*, Pose-Transfer [45]. Results are shown on the right of Fig. 4 and 5. We can see that XingGAN also generates much better person images having fewer visual artifacts than Pose-Transfer. For instance, Pose-Transfer [45] always generates a lot of visual artifacts in the background as shown in Fig. 5.

Human Evaluation. We follow the evaluation protocol of [21,31,45] and recruited 30 volunteers to conduct a user study. Participants were shown a sequence of images and asked to give an instant judgment about each image within a second. Specifically, we randomly select 55 real and 55 fake images (generated by our model) and shuffle them. The first 10 of them are used for practice and the remaining 100 images are used for evaluation. Results compared with PG2 [21], PoseGAN [31], Pose-Transfer [45] and C2GAN [35] are shown in Table 2. We observe that the proposed XingGAN achieves the best results on all measurements compared with the leading methods, further validating that the generated images by our model are more sharp and photo-realistic.

Variants of XingGAN. We conduct extensive ablation studies on Market-1501 [44] to evaluate different components of our XingGAN. XingGAN has four baselines as shown in Table 3: (i) 'SA' means only using the proposed Shape-guided Appearance-based generation branch. (ii) 'AS' means only adopting the
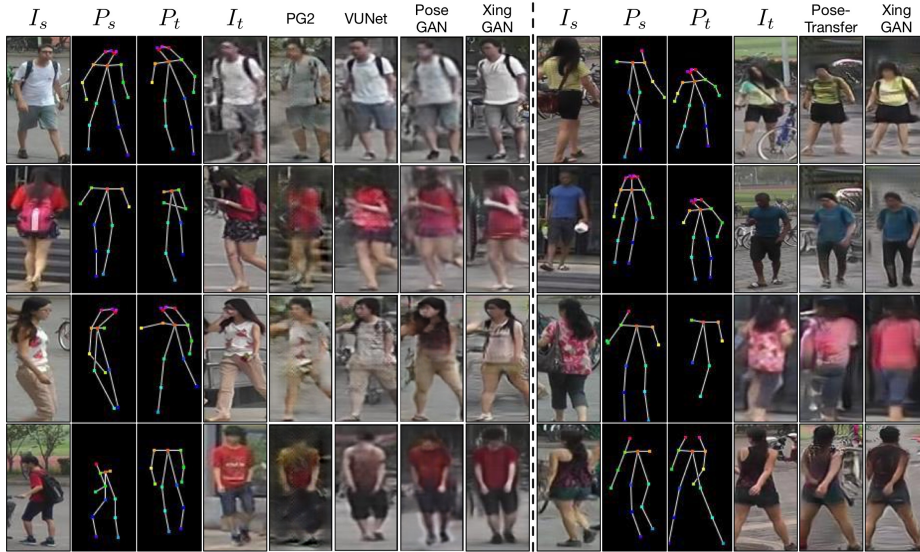
Fig. 4: Qualitative comparison with PG2 [21], VUnet [7], PoseGAN [31] and Pose-Transfer [45] on Market-1501.

Table 2: User study of person image generation (%). R2G means the percentage of real images rated as generated w.r.t. all real images. G2R means the percentage of generated images rated as real w.r.t. all generated images. The results of other methods are reported from their papers.

Method	Market-1501		DeepFashion	
	R2G	G2R	R2G	G2R
PG2 [21]	11.2	5.5	9.2	14.9
PoseGAN [31]	22.67	50.24	12.42	24.61
C2GAN [35]	23.20	46.70	-	-
Pose-Transfer [45]	32.23	63.47	19.14	31.78
XingGAN (Ours)	**35.28**	**65.16**	**21.61**	**33.75**

proposed Appearance-guided Shape-based generation branch. (iii) 'SA+AS' combines both branches to produce the final person images. (iv) 'SA+AS+CAF' is our full model and employs the proposed Co-Attention Fusion module.

Effect of Dual-Branch Generation. The results of the ablation study are shown in Table 3. We see that the proposed SA branch achieves only 0.239 and 0.768 in SSIM and Mask-SSIM, respectively. When we only use the proposed AS branch, the values of SSIM and Mask-SSIM are improved to 0.286 and 0.798, respectively. Thus we conclude that the AS branch is more effective than the SA branch for generating photo-realistic person images. The AS branch takes the person poses as input and aims to learn person appearance representations, while the SA branch takes the person image as input and targets to learn person shape
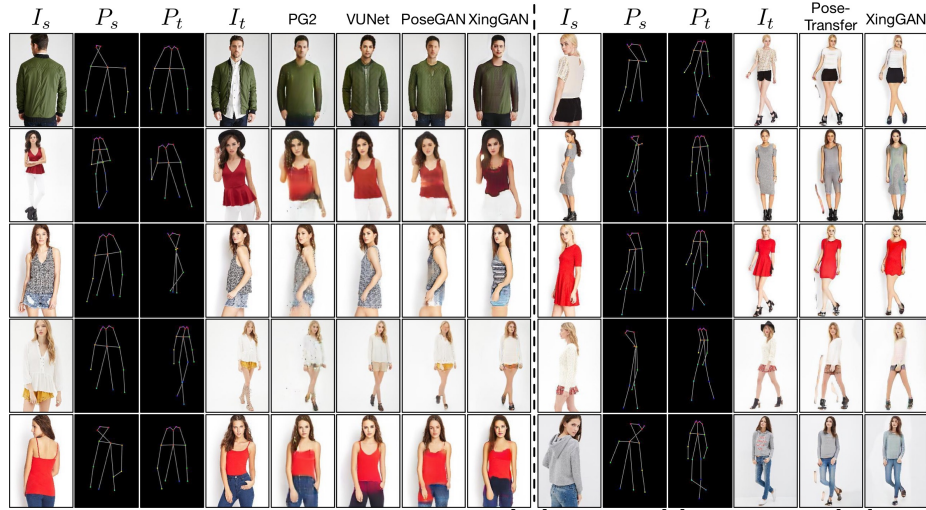
Fig. 5: Qualitative comparison with PG2 [21], VUnet [7], PoseGAN [31] and Pose-Transfer [45] on DeepFashion.

Table 3: Quantitative comparison of different variants of the proposed XingGAN on Market-1501. For all metrics, higher is better. 'SA', 'AS' and 'CAF' stand for the proposed SA branch, AS branch and co-attention fusion module, respectively.

Variants of XingGAN	IS	Mask-IS	SSIM	Mask-SSIM
SA	**3.849**	3.645	0.239	0.768
AS	3.796	3.810	0.286	0.798
SA + AS	3.558	3.807	0.310	0.807
SA + AS + CAF (Full)	3.506	**3.872**	**0.313**	**0.816**

representations. Learning the appearance representations is much easier than learning the shape representations since there are shape deformations between the input person image and the desired person image, leading the AS branch to achieve better results than the SA branch.

Next, when adopting the combination of the proposed SA and AS branches, the performance in terms of SSIM and Mask-SSIM further boosts. However, the results in terms of IS and Mask-IS do not decline too much. Moreover, Fig. 6 (*left*) shows some qualitative examples of the ablation study. We observe that the visualization results of 'SA', 'AS', and 'SA+AS' are consistent with the quantitative results. Therefore, both quantitative and qualitative results confirm the effectiveness of the proposed dual-branch generation strategy.

Effect of Co-Attention Fusion. 'SA+AS+CAF' outperforms the 'SA+AS' baseline with around 0.065, 0.003, and 0.009 gain on Mask-IS, SSIM, and Mask-SSIM, respectively. This means that the proposed co-attention fusion model indeed learns more correlations between the appearance and shape representations for generating the targeted person images, confirming our design motiva-
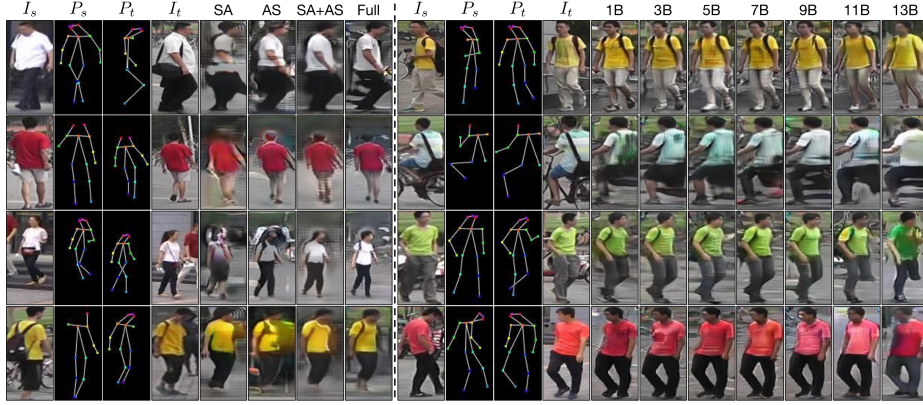
Fig. 6: Ablation study of the proposed XingGAN on Market-1501. (*left*) Results of different variants of the proposed XingGAN. (*right*) Results of varying the number of the proposed Xing blocks. 'B' stands for the proposed Xing Blocks.

Table 4: Quantitative comparison and ablation study of the proposed Xing generator on Market-1501. For all metrics, higher is better.

Method	IS	Mask-IS	SSIM	Mask-SSIM
Xing Generator (1 blocks)	3.378	3.713	0.310	0.812
Xing Generator (3 blocks)	3.241	3.866	**0.316**	0.813
Xing Generator (5 blocks)	3.292	3.860	0.313	0.812
Xing Generator (7 blocks)	3.293	3.871	0.310	0.810
Xing Generator (9 blocks)	3.506	**3.872**	0.313	**0.816**
Xing Generator (11 blocks)	3.428	3.712	0.286	0.793
Xing Generator (13 blocks)	**3.708**	3.679	0.257	0.774
Resnet Generator (5 blocks)	3.236	3.807	0.297	0.802
Resnet Generator (9 blocks)	3.077	3.862	0.301	0.802
Resnet Generator (13 blocks)	3.134	3.731	0.300	0.797
PATN Generator (5 blocks)	3.273	3.870	0.309	0.809
PATN Generator (9 blocks)	3.323	3.773	0.311	0.811
PATN Generator (13 blocks)	3.274	3.797	0.314	0.808

tion. Moreover, the proposed CAF module obviously improves the quality of the visualization results, as shown in the column 'Full' of Fig. 6.

Lastly, we show the learned co-attention maps and the generated intermediate results. These co-attention maps are complementary, which could be qualitatively verified by visualizing the results in Fig. 7. It is clear that they have learned different activated content between the generated intermediate results and the input image for generating the final person images.

Effect of The Xing Generator. The proposed Xing generator has two important network designs. One is the carefully designed Xing block, consisting of two sub-blocks, *i.e.*, SA block and AS block. The Xing blocks jointly model both shape and appearance representations in a crossing way and enjoying the
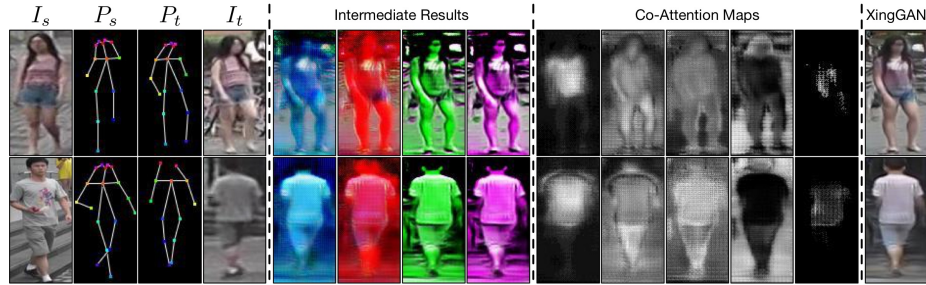
Fig. 7: Visualization of intermediate results and co-attention maps generated by the proposed XingGAN on Market-1501. We randomly show four intermediate results, the corresponding four co-attention maps and the input attention map. Attention maps are normalized for better visualization.

mutually improved benefits from each other. The other one is the cascaded network design, which deals with the complex and deformable translation problem progressively. Thus, we further conduct two experiments, one is to show the advantage of the progressive generation strategy by varying the number of the proposed Xing blocks, and the other is to explore the advantage of the Xing block by replacing it with the residual block [13] and PATB [45] resulting in two generators named Resnet generator and PATN generator in Table 4, respectively.

Quantitative and qualitative results are shown in Table 4 and Fig. 6 (*right*). We observe that the proposed Xing generator with 9 blocks works the best. However, increasing the number of blocks further reduces generation performance. This could be attributed to the proposed Xing block. Only a few blocks are needed to capture the useful appearance and shape representations and the connection between them. Thus, we adopt 9 Xing blocks as default in our experiments for both datasets. Moreover, we see that the proposed Xing generator with only 5 Xing blocks outperforms both ResNet and PATN generators with 13 blocks on most metrics, which further certifies that our Xing generator has a good appearance and shape modeling capabilities with a very few blocks.

5   Conclusions

We propose a novel XingGAN for the challenging person image generation task. It uses cascaded guidance with two different generation branches, and learns a deformable translation mapping from both person shape and appearance features. Moreover, we propose two novel blocks to effectively update person shape and appearance features in a crossing way. Extensive experiments based on human judgments and automatic evaluation metrics show that XingGAN achieves new state-of-the-art results on two challenging datasets. Lastly, we believe that the proposed blocks and the XingGAN framework can be easily extended to address other GAN-based generation and even multi-modality fusion tasks.

References

1. AlBahar, B., Huang, J.B.: Guided image-to-image translation with bi-directional feature transformation. In: ICCV (2019)
2. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: CVPR (2018)
3. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: ICLR (2019)
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
5. Chi, L., Tian, G., Mu, Y., Tian, Q.: Two-stream video classification with cross-modality attention. In: ICCV Workshops (2019)
6. Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., Yin, J.: Soft-gated warping-gan for pose-guided person image synthesis. In: NeurIPS (2018)
7. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: CVPR (2018)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
9. Grigorev, A., Sevastopolsky, A., Vakhitov, A., Lempitsky, V.: Coordinate-based texture inpainting for pose-guided human image generation. In: CVPR (2019)
10. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
11. Ilyes Lakhal, M., Lanz, O., Cavallaro, A.: Pose guided human image synthesis by view disentanglement and enhanced weighting loss. In: ECCV (2018)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
16. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: ECCV (2018)
17. Li, Y., Huang, J.B., Ahuja, N., Yang, M.H.: Joint image filtering with deep convolutional networks. IEEE TPAMI **41**(8), 1909–1923 (2019)
18. Li, Y., Huang, C., Loy, C.C.: Dense intrinsic appearance flow for human pose transfer. In: CVPR (2019)
19. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: ICCV (2019)
20. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016)
21. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: NeurIPS (2017)
22. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: CVPR (2018)
23. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
24. Neverova, N., Alp Guler, R., Kokkinos, I.: Dense pose transfer. In: ECCV (2018)

25. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)
26. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)
27. Pumarola, A., Agudo, A., Sanfeliu, A., Moreno-Noguer, F.: Unsupervised person image synthesis in arbitrary poses. In: CVPR (2018)
28. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. In: ECCV (2018)
29. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NeurIPS (2016)
30. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: ICCV (2019)
31. Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable gans for pose-based human image generation. In: CVPR (2018)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
33. Song, S., Zhang, W., Liu, J., Mei, T.: Unsupervised person image generation with semantic parsing transformation. In: CVPR (2019)
34. Tang, H., Wang, W., Xu, D., Yan, Y., Sebe, N.: Gesturegan for hand gesture-to-gesture translation in the wild. In: ACM MM (2018)
35. Tang, H., Xu, D., Liu, G., Wang, W., Sebe, N., Yan, Y.: Cycle in cycle generative adversarial networks for keypoint-guided image generation. In: ACM MM (2019)
36. Tang, H., Xu, D., Sebe, N., Wang, Y., Corso, J.J., Yan, Y.: Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In: CVPR (2019)
37. Tang, H., Xu, D., Yan, Y., Torr, P.H., Sebe, N.: Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In: CVPR (2020)
38. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
40. Xian, W., Sangkloy, P., Agrawal, V., Raj, A., Lu, J., Fang, C., Yu, F., Hays, J.: Texturegan: Controlling deep image synthesis with texture patches. In: CVPR (2018)
41. Yang, C., Wang, Z., Zhu, X., Huang, C., Shi, J., Lin, D.: Pose guided human video generation. In: ECCV (2018)
42. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: ICCV (2019)
43. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM TOG **9**(4) (2017)
44. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
45. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: CVPR (2019)