Pose-Guided Person Image Synthesis in the Non-Iconic Views

Chengming Xu, Yanwei Fu, Chao Wen, Ye Pan, Yu-Gang Jiang, *Member, IEEE*, and Xiangyang Xue, *Member, IEEE*

***Abstract*— Generating realistic images with the guidance of reference images and human poses is challenging. Despite the success of previous works on synthesizing person images in the iconic views, no efforts are made towards the task of pose-guided image synthesis in the non-iconic views. Particularly, we find that previous models cannot handle such a complex task, where the person images are captured in the non-iconic views by commercially-available digital cameras. To this end, we propose a new framework – Multi-branch Refinement Network (MR-Net), which utilizes several visual cues, including target person poses, foreground person body and scene images parsed. Furthermore, a novel Region of Interest (RoI) perceptual loss is proposed to optimize the MR-Net. Extensive experiments on two non-iconic datasets, Penn Action and BBC-Pose, as well as an iconic dataset – Market-1501, show the efficacy of the proposed model that can tackle the problem of pose-guided person image generation from the non-iconic views. The data, models, and codes are downloadable from https://github.com/loadder/MR-Net.**

***Index Terms*— Image processing, image generation.**

I. Introduction

SYNTHESIZING images with bespoke human poses has recently attracted pervasive research attention in computer vision, multimedia, and graphics communities [5], [23], [24], [31], [33]. Broadly speaking, pose-guided person image synthesis can be applied in many scenarios, including virtual environment rendering, photography editing, character animation, physics-based simulation, and motion control, etc. Furthermore, the forged person images can also be utilized in the applications of video generation [10] and video completion [5].

Recently, extensive works have been conducted in synthesizing iconic person images. Here we inherit the definition of "iconic" in [19]. Particularly, as shown in Fig. 1 (a), the person images of Market-1501 dataset are in the iconic views: high-quality person instances in the center of images, but lacking important contextual information and non-canonical viewpoints. Previous works [5], [23], [24], [31], [33] perform fairly well on such iconic person image datasets, *e.g.*, Market-1501 and Deep Fashion [21]. These datasets, in general, are in very simple scenes, mostly street views which are not required to predict, or backgrounds with single color. Moreover, persons are either standing or walking and not occluded by objects. However, extending previous models to non-iconic person images would lead to unstable training and generated results with low quality, as shown in Fig. 2.

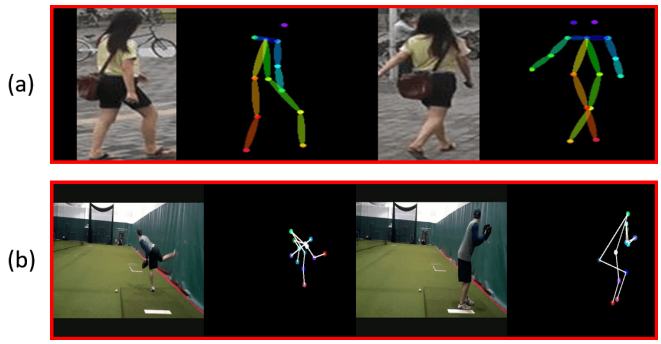

Fig. 1. (a) iconic person images: person instances with walking/standing poses near the center of a neatly composed photo in a simple scene. (b) non-iconic person images: person instances with arbitrary poses in a clutter scene.

Rather than only synthesizing iconic person images, we tackle the novel task of synthesizing person images of target poses in the non-iconic view, as defined in Fig. 2. In such a novel task, the model needs to synthesize both realistic human body and consistent background. Nevertheless, the interaction between objects and persons in non-iconic person images intrinsically makes the image synthesis much more complex than iconic person images. Furthermore, our novel task is even challenged by the casual nature of unprocessed photos. Such photos capture the human body with varying poses at any position in the daily life and work of everyone, and thus typically featured with poor lighting, low resolution, occlusion, camera shake, varying human actions/poses, and cluttered scenes of background noise. Critically the underlying challenges existed in the pose-guided human synthesis from the non-iconic view, can be broadly characterized as below,

(1) **Cluttered scene.** The images of iconic person datasets are captured in the simple scenes, *e.g.*, a specified street view.

Manuscript received October 17, 2019; revised April 11, 2020 and July 26, 2020; accepted August 30, 2020. Date of publication September 18, 2020; date of current version September 24, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jiaying Liu. *(Corresponding author: Yanwei Fu.)*

Chengming Xu and Yanwei Fu are with the Shanghai Key Laboratory of Intelligent Information Processing, School of Data Science, Fudan University, Shanghai 200433, China (e-mail: 18110980002@fudan.edu.cn; yanweifu@fudan.edu.cn).

Chao Wen is with the Academy for Engineering and Technology, Fudan University, Shanghai 200433, China (e-mail: cwen18@fudan.edu.cn).

Ye Pan is with the John Hopcroft Center for Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Disney Research, Glendale, CA 91201 USA (e-mail: whitneypanye@sjtu.edu.cn).

Yu-Gang Jiang and Xiangyang Xue are with the School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: ygj@fudan.edu.cn; xyxue@fudan.edu.cn).

Digital Object Identifier 10.1109/TIP.2020.3023853

Fig. 2. Overview of the task. Given a reference image where the person instance is in the non-iconic view, our task aims at synthesizing person images with the pose from target images. Our results are better than those synthesized images by DVG, PN-GAN and PG^2.

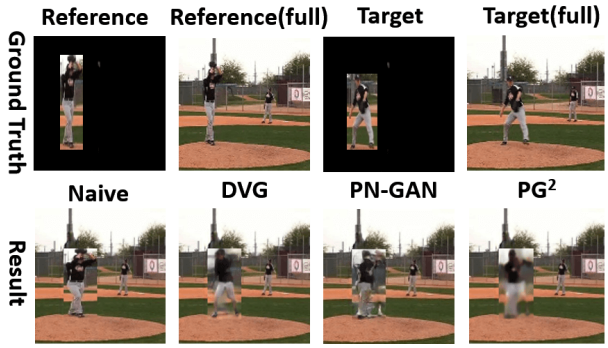

Fig. 3. Inconsistent background by naive methods. We use the ground-truth bounding box given in Penn Action. The naive methods (even further processed by previous methods) produce inconsistent background. More ablation study in Sec. IV-D.

This greatly and unrealistically simplifies the synthesis task, since image generator does not need to explicitly model the image scene. However, this is not the case of actual daily scenes, where a *non-iconic* image can be captured at cluttered scene. For example, Figure 1 (b) shows the person in the non-iconic view may be in the cluttered scenes with varying pose scales and pose shape.

(2) **Arbitrary person poses.** There are plenty of standing or walking postures in the 'iconic' Market-1501 and Deep Fashion dataset, as illustrated in Fig. 1 (a). Most of the previous works could produce remarkable good results on these iconic views, while they may be failed in the novel task of image synthesis from non-iconic views, as illustrated in Fig. 2. Intrinsically, the key challenge comes from the fact that the task of non-iconic view synthesis has to handle the generation of person images from arbitrary human poses, *e.g.*, the bent (Fig. 2) or distorted (Fig. 1 (b)) poses.

(3) **Inconsistent background.** One may ask whether our task can be addressed by any naive solutions, such as, cropping, inpainting and pasting image patches of persons. For example, one simple solution is to firstly crop the person regions in reference and target images respectively, which are pasted into reference images with some necessary inpainting operations. Unfortunately, such a solution could visually generate inconsistent background, as illustrated in Fig. 3.

To tackle person synthesis in the non-iconic views, we propose a novel model – Multi-branch Refinement Network (MR-Net), which makes the use of visual cues such as human poses, and parsed images inspired from the scenes in reference images, and poses in target image respectively. These visual cues are systematically integrated into the proposed pose-guided multi-branch encoders, with the key insights of operating separate background and pose-guided human body foreground to extract different information. Critically, the background branch grasps the holistic visual cues and maintains the cluttered scenes in the synthesized images. The pose-guided branch learns how to rotate and shift the person instances to the ideal position referring to the target poses.

Formally, our model is composed of the key ingredients in Fig. 4. We get the foreground person body and background scene from the reference images with a pre-trained image segmentation model such as Mask-RCNN [11]. The multi-branch encoders are then utilized in encoding the background image, target pose, and foreground person image. All the extracted information is combined through a decoder. Essentially, the shortcut connection, dilated convolution and non-local block are employed here to improve the capability of our model. Moreover, to handle the problem resulted from non-iconic views and arbitrary poses, we propose a useful loss–RoI perceptual loss in addition to the common reconstruction and perceptual losses, which could help efficiently optimize the MR-Net to better synthesize the non-iconic person images. Furthermore, by adopting a refinement strategy, the synthesized images of MR-Net are repeatedly refined to update the result and make the background more consistent.

Note that the previous Market-1501 and Deep Fashion datasets only have the person instances in the iconic views. In contrast, we propose to use two novel non-iconic datasets – Penn Action [51] and BBC-Pose [7]. As shown in Fig. 1, the scenes are more complex; and persons have diverse poses. On these benchmark datasets, we show the efficacy of our proposed MR-Net.

Contributions. We highlight our contributions as follows. (1) For the first time, we provide a new and more challenging perspective of addressing pose-guided human synthesis in a non-iconic view. (2) To address this challenging task, we propose a framework – MR-Net. Our network includes several new components: pose-guided multi-branch encoder, RoI perceptual loss, and the newly proposed refinement strategy. Particularly, the pose-guided multi-branch encoder is a network structure introduced here to parse the foreground and background visual cues; and RoI perceptual loss is a novel loss function proposed for efficiently training the MR-Net. The

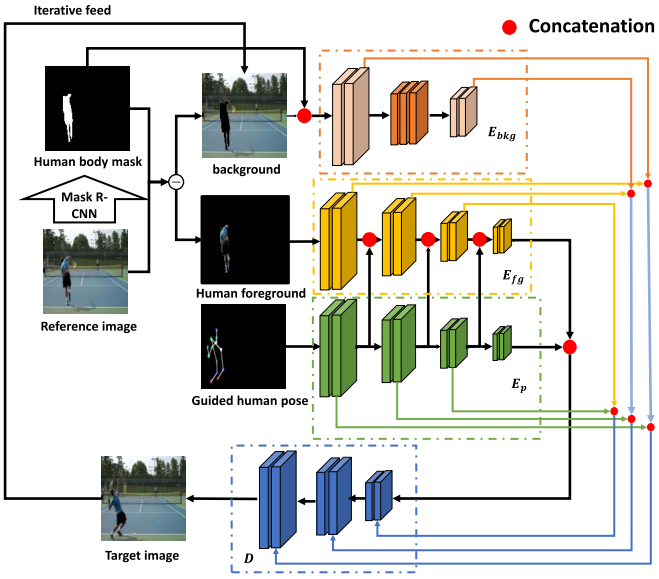

Fig. 4. Overall framework of our proposed model.

refinement step is a learning strategy introduced to iteratively update the synthesized image and improve its visual quality. Extensive experiments and ablation studies show the efficacy of the proposed MR-Net that, even with more complicated scenes and diverse person poses, the proposed model can still perform well in generating person images in the non-iconic views.

II. Related Work

A. Image Generation

Previous approaches can be categorized into three categories: (1) Variational Auto-encoder (VAE) [18], [39]: it provides a probabilistic manner for describing the observation in the latent space; and then samples from the probability distribution to generate new samples. (2) Auto-Regressive Model [27], [28]: it learns an explicit distribution, which is controlled by a prior knowledge imposed by the model structures. (3) Generative Adversarial Network (GAN) [2], [4], [25], [32], [34]: this type of networks utilizes a generator for image generation, while the networks are optimized as a min-max game between the generator and discriminator, and thus can iteratively improve the performance of both models. Rather than generating images from noise or any prior distributions, this paper focuses on the task of person image generation with the input of target poses. Particularly, the person in the image is in the non-iconic view.

B. Image-to-Image Translation

This is an umbrella term over many related tasks, such as image style transfer [15], [16], [55], image inpainting [20], [30], [44] and image super-resolution [8], [46]. With the input of a reference image and meta information, this task aims to produce the corresponding images with predefined types. Most of these works are built upon the existing GANs and auto-encoder models. Especially, as a non-trivial extension of auto-encoders, the U-net [36] with skip connections, can not only accelerate network training but also improve the generalization ability of models. Different from previous image-to-image translation works in transforming the holistic image styles, our task integrates arbitrary target poses to help synthesize the images where persons may appear in the cluttered scenes from the non-iconic views.

C. Loss Functions

For all tasks with the images as output, the objective functions related to image content have the following two forms: (1) Pixel-level reconstruction loss [22], [45]. This type of loss models the problems as pixel-wise regression; therefore it may potentially lose the semantic structure information contained in images. (2) Perceptual loss. This method treats the images as the compositions of visual perceptions, including the high-level features of deep learning models [16], [47], traditional features like HOG [41] and others. Although this type of objective functions could keep the conformation of images, it mostly focuses on the global information, which thus is not the best choice for our task. Typically, our synthesized images have both background and human body foreground with various scales and target poses. To this end, we introduce a region-of-interest perceptual loss.

D. Pose-Guided Image Synthesis

Recent advances in conditional generative networks enable the user-guided image synthesis learned from large-scale datasets. Previous works [5], [23], [31] use the GANs to simultaneously reconstruct the target images and train the generator in an adversarial manner. Dong *et al.* [9] adopt a warping module, which uses information from learned human parsing. However, these works mostly target at pose-guided iconic person image synthesis, where the popular poses are standing and walking. Ma *et al.* [24] learn a disentangled representation of images by training different auto-encoders for poses, human foreground, and background, and they further use the trained decoders to sample pose conditioned images. In their work, the background encoder is trained only for reconstruction, where the target position of the human body is the same as that in the input image. Therefore it might easily fail on non-iconic images. In contrast, our model does not only explicitly learn the representations of scene and background, but also employs a novel loss function in better learning the representation. Balakrishnan *et al.* [3] consider segmenting background and foreground to improve the image quality. However, it is still about person synthesis in the iconic view. Moreover, some methods, *e.g.*, DeformGAN [37], require the input of an affine matrix between any two body joints. Unfortunately, some joints are missed from the source body in many datasets, *e.g.*, Penn Action. Thus it is difficult to make a good estimation of transformation, and it also negatively affects the synthesis. Compared with that, our model does not have such restrictions. In addition to solving the non-iconic task, our model also has to deal with the visual difference between human poses and human bodies in the images; thus,

quite different from previous methods, we introduce the multi-branch structure which gradually transfers the source body with pose information.

The human shape guided transformation is also discussed in other related topics. For example in the task of person re-identification, Qian *et al.* [33] follow PG^2 to directly augment the gallery data by generating pose-guided images. Miao *et al.* [26] consider disentangled features for each joint which are then aggregated to enhance the global feature for human images. [40], [54] explore part-based person feature for re-identification. Zheng *et al.* [52] entitle model to more generalize ability with pose-invariant features. Moreover, such a technique is also important to the application of clothes try-on [13], [14].

III. Methodology

Our goal is to synthesize a new human image in a complex background with a specified pose. In particular, given a reference image $\mathbf{I}_r$ which provides the person image in a complex background with target image and target pose image denoted as $\mathbf{I}_t$ and $\mathbf{I}_{tp}$ respectively where $\mathbf{I}_{tp}$ comes from the target image $\mathbf{I}_t$[1], we can synthesize a new image $\hat{\mathbf{I}}_s$. Note that all the images have the same size, *i.e.*, $\mathbf{I}_t, \mathbf{I}_{tp}, \hat{\mathbf{I}}_s, \mathbf{I}_r \in \mathbb{R}^{w\times h\times 3}$. Essentially, we require that $\hat{\mathbf{I}}_s$ should (1) still has the same person identity and complex background as $\mathbf{I}_r$ while keeping the correct pose as $\mathbf{I}_{tp}$, and (2) be genuine enough to be considered as a natural image.

Overview. We propose a novel architecture —Multi-branch Refinement Network (MR-Net) which treats these several elements separately. We overview the architecture in Fig. 4. Given a reference image $\mathbf{I}_r$, we first segment the human body out by pre-trained Mask-RCNN to produce the person body image $\mathbf{I}_h$, background image $\mathbf{I}_{bkg}$, and corresponding mask $\mathbf{M}$. And $\mathbf{M}$ is the binary mask: its 0 value indicates the background; 1, foreground. Note that segmentation module is not a module of our MR-Net, nor a contribution in this paper. Then, we input the target pose $\mathbf{I}_{tp}$ and $\mathbf{I}_h$ into pose-guided encoder $\mathbf{E}$ with two pathways, *i.e.*, target pose encoder $\mathbf{E}_p$ and foreground encoder $\mathbf{E}_{fg}$, and input the $\mathbf{I}_{bkg}$ and $\mathbf{M}$ into the background encoder $\mathbf{E}_{bkg}$, whose output is then passed through decoder $\mathbf{D}$ via skip connections. Finally, after several convolutions and up samplings, the decoder $\mathbf{D}$ outputs the target image $\hat{\mathbf{I}}_t$.

A. Pose-Guided Multi-Branch Encoders

We propose to use the multi-branch encoder structure in parsing and encoding the information of the foreground person body, and background scene differently. Particularly, the person instances should be dramatically changed guided by the target poses, while the required update on the background scene is relatively subtle. Therefore, our key lies in separating the tasks into two parts, *i.e.*, (1) background images and the mask of foreground human body with the background encoder ($\mathbf{E}_{bkg}$) (2) target pose and foreground human body with the pose-guided encoder ($\mathbf{E}$).

[1]On Market-1501, $\mathbf{I}_{tp}$ is detected by OpenPose [6]; on Penn Action and BBC-Pose, it is manually annotated [51].

1) Background Encoder: We learn the $\mathbf{E}_{bkg}$ to encode the background image $\mathbf{I}_{bkg}$ with the corresponding mask $\mathbf{M}$. Particularly, inspired by recent inpainting work [49], we utilize dilated convolution [48] to increase the receptive field.

2) Pose-Guided Encoder: It includes the foreground encoder $\mathbf{E}_{fg}$ processing the foreground person image $\mathbf{I}_h$, and target pose encoder $\mathbf{E}_p$ which learns to extract features from target poses. Specifically, in the forward pass, the intermediate features from the layers of $\mathbf{E}_p$ are used to guide the learning process of $\mathbf{E}_{fg}$. As shown in Fig. 4, the features of $\mathbf{E}_p$ are further channel-wise concatenated with the corresponding output features of the same layers in $\mathbf{E}_{fg}$.

3) Decoder D: The outputs of three encoders are fused to pass through decoder $\mathbf{D}$. The outputs of $\mathbf{E}_{fg}$ and $\mathbf{E}_p$ are concatenated and then used as input of $\mathbf{D}$. Note that the shortcut connections are utilized to connect the shallower layers of three encoders, as well as the output of $\mathbf{E}_{bkg}$, to the homologous layers of $\mathbf{D}$. To improve the image quality, we utilize non-local blocks [42] in the decoder.

B. Loss Functions and Training Strategy

We define the loss function $\mathcal{L}$ as follows,

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \cdot \mathcal{L}_{perc} + \lambda_2 \cdot \mathcal{L}_{roi} \tag{1}$$

where λ_1 and λ_2 are two coefficients of each loss term.

1) Weighted Reconstruction Loss: The aim of this loss is to encourage the pixels of generated images to match the target ones; especially, the areas where person instances appear in reference and target images. Therefore, we adopt the common pixel-level reconstruction loss, *i.e.*,

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_j \left[\left\| \hat{\mathbf{I}}_s^{(j)} - \mathbf{I}_t^{(j)} \right\|^2 + \left| \hat{\mathbf{I}}_s^{(j)} - \mathbf{I}_t^{(j)} \right| \odot (\mathbf{M}_r + \mathbf{M}_t) \right] \tag{2}$$

where the $\mathbf{M}_r$ and $\mathbf{M}_t$ denote the binary mask matrices of bounding box of human body on reference and target image, individually: value 1 indicates pixels within the bounding box; 0, otherwise. $\odot$ refers the element-wise product. N is the number of images optimized. Since Eq (2) is computed on each batch, N thus equals the batch size.

2) Perceptual Loss: This loss [16] is introduced to push the high-level features of generated images towards the target images; it measures the perceptual difference in content and style between images, which is defined as,

$$\mathcal{L}_{perc} = \frac{1}{N}\sum_j \sum_l \left\| \Phi^l\left(\hat{\mathbf{I}}_s^{(j)}\right) - \Phi^l\left(\mathbf{I}_t^{(j)}\right) \right\|^2 \tag{3}$$

The function $\Phi^l(\cdot)$ is the output of the l layer in the visual perceptual network [16], *i.e.*, the VGG-19 [38] network pre-trained on ImageNet-2012 dataset.

3) RoI Perceptual Loss: Despite the fact that perceptual loss is widely used in synthesizing realistic images by the generative models [15], [16], [55], we realize that perceptual loss alone is not enough for the visual quality of synthesized person body parts. In contrast to previous pose-guided person image generation in the iconic views [23], [24], [31], [33], our

work has to handle the cases of generating person images of complex poses in cluttered scenes. To this end, we propose an extension– Region of Interest (RoI) perceptual loss. It particularly optimizes the specific image regions, *i.e.*, the regions of person bodies, as follows,

$$\mathcal{L}_{roi} = \frac{1}{N}\sum_{j}\sum_{l}\left\|\phi_{\mathbf{M}_t}\left(\Phi^l\left(\hat{\mathbf{I}}_s^{(j)}\right)\right) - \phi_{\mathbf{M}_t}\left(\Phi^l\left(\mathbf{I}_t^{(j)}\right)\right)\right\|^2 \tag{4}$$

where $\phi_{\mathbf{M}_t}(\cdot)$ is the RoI pooling layer in [35] with the bounding box $\mathbf{M}_t$; In addition to help to localize the regions of interest, the RoI pooling layer $\phi(\cdot)$ also helps to keep the output feature dimension size the same among all samples; thus the loss can be computed via mini-batch, no matter what size the human body is. Particularly, compared to pose mask loss proposed in [23], our RoI perceptual loss term can be applied to the features maps beyond the raw images; and thus it is more flexible. Critically, the RoI perceptual loss term does not require the extra processing over human poses to generate the masks; and this can avoid the failure cases when some of the key joints are missing.

4) Refinement Strategy: We introduce iterative refinement strategy to update our model. Specifically, as shown in Fig. 4, we treat the generated images $\hat{\mathbf{I}}_{s,k}$ as the new background images, and iteratively update the background encoder $\mathbf{E}_{bkg}$, where we use the sub-index to denote the $k-$ th round update. In principle, this strategy could help our model gradually fill the missing region of the background images, thus refining the output image and accelerating the training process.

IV. Experiments

A. Datasets and Settings

Our model is evaluated on three datasets. (1) *Market-1501*. It is a person re-identification dataset [53], composed of 32668 images and 1501 persons captured from six disjoint surveillance cameras. We follow the previous setting [56], which contains 263,632 training pairs and 12,000 testing pairs. The image size is 128×64. (2) *Penn Action*. It contains 2326 videos with 15 action classes [51]. Each frame is annotated by human poses. We gather a training set of 57,925 training pairs of images and 4,634 testing pairs which are uniformly sampled from all of the videos. (3) *BBC-pose*. It has 20 TV broadcast videos [7]; totally 40K frames of different hosts with different gestures. The reference and target frames are also sampled in the same video. In the train/test split, we have 14942 frames as training (80% total frames); and 3736 frames as test set. Quite different from the other two datasets, the images of this dataset has the dynamic background – news being broadcast. This introduces additional challenges in synthesizing images. Note that Penn Action, and BBC-pose have lots of person images in non-iconic views, while most person images in Market-1501 are in iconic-views; also DeepFashion [21] is not used which has a pure white color background. Images of Penn Action and BBC-Pose are all resized to 128×128.

1) Network Architectures: The detailed architecture of our network is listed as,. *Pose Encoder, Foreground Encoder*: K5C32 - Down - K5C32 - K5C64 - Down - K5C64 - K5C128 - Down - K5C128 - K5C256 - Down - K5C256; *Background Encoder*: K5C32 - K5C64 - Down - K3D2C64-K3D4C64 - K3D8C64 - K5C128 - Down - Up(32×32); *Decoder*: K1C256 - K5C128 - Up2 - NL128 - K5C128 - K5C64 - Up2 - NL64 - K5C64 - K5C32 - Up2 - K5C32-K5C3 - Up2. where K denotes kernel size; D denotes dilation rate; Down denotes 2×2 max-pooling; Up($\cdot$) denotes upsampling to target size; Up2 denotes 2×2 upsampling; C, channel number; NL, non-local block with designated channel number. All convolutional layers are with 1 stride, and followed with ReLU and batch normalization, except that the last convolutional layer of decoder uses tanh activation and no batch normalization.

2) Implementation Details: Totally, our model has 8 convolution layers with kernel size of 5 and stride of 1 combined with 4 Max-pooling layers for down-sample in the pose and human body encoders, and the symmetric structure for the decoder. For the background encoder, we use two convolution layers, with three dilated convolution layers and two normal convolution layers after. Our model is trained by 3 iterations, in order to make a balance between good results and reasonable computational cost. For perceptual loss, we choose to compare feature maps at the layers of relu1_2, relu2_2, relu3_3, relu4_3 and relu5_3. On these layers, the sizes of RoI pooling outputs are set as 30, 30, 15, 7 and 4 respectively, to further facilitate the RoI perceptual loss. Our model is implemented on Pytorch [29], by Adam optimizer [17] with the learning rate of 0.0002. The λ_1, λ_2 is set as 0.5 and 0.05. The dataset and models will be released.

3) Competitors: Several previous models are utilized as our baselines, namely, PG^2 [23], PN-GAN [33], DVG [5], PoseWarp [3], BodyROI7 [24], PATN [56] and Guided Pix2Pix [1]. We conduct experiments of all these models based on open-sourced codes, with parameters and learning rate simply tuned. As discussed in the related work, BodyROI7 and PoseWarp, intrinsically, are not applicable to synthesize person image in the non-iconic view. Thus, BodyROI7 and PoseWarp are only used on Penn Action, which is a typical dataset of a large portion of people in non-iconic views, to support our claim and experiments on these two models are not extended to other datasets.

B. Quantitative Results

We employ Inception Score (IS), Fréchet Inception Distance (FID), Structural Similarity (SSIM), mask-SSIM and Learned Perceptual Image Patch Similarity (LPIPS) to evaluate the performance of each model. Specifically, (1) Inception Score computes distribution certainty and diversity for each generated images by an Inception model. (2) Fréchet Inception Distance (FID) [12] measures distance of two sets of images. (3) SSIM [43] is a metric evaluating the difference of perceived quality between generated and target images. (4) M-SSIM proposed in [23] evaluates the quality of generated human body without considering the background. (5) LPIPS [50] is a newly proposed metric on image quality. Note that, the persons in many images of Penn Action are in non-iconic views: very small person instances in the whole image. This

TABLE I

QUANTITATIVE RESULTS OF ALL MODELS ON PENN ACTION AND BBC-POSE. FOR IS, SSIM AND M-SSIM, HIGHER SCORE MEANS THE BETTER, FOR FID AND LPIPS, THE LOWER THE BETTER

Model	Penn Action					BBC-Pose				
	IS	FID	SSIM	M-SSIM	LPIPS	IS	FID	SSIM	M-SSIM	LPIPS
PG^2 [33]	5.835	61.016	0.588	0.362	0.306	1.361	**81.191**	**0.578**	**0.900**	0.389
PN-GAN [33]	6.861	27.016	0.723	0.429	0.121	1.434	265.224	0.432	0.832	0.397
DVG [5]	6.366	55.035	0.622	0.943	0.154	1.297	137.523	0.496	0.873	0.386
BodyROI7 [24]	3.547	224.439	0.301	0.923	0.452	—	—	—	—	—
PoseWarp [3]	6.643	21.576	0.684	0.442	0.142	—	—	—	—	—
PATN [56]	6.227	27.016	0.731	0.412	0.129	**2.159**	132.884	0.485	0.873	0.356
G-Pix2Pix [1]	2.986	260.059	0.424	0.938	0.534	1.376	140.984	0.451	0.858	0.470
MR-Net	**6.868**	**18.583**	**0.751**	**0.962**	**0.110**	1.751	95.114	0.527	0.892	**0.321**
Ground-Truth	6.997	—	—	—	—	3.693	—	—	—	—

TABLE II

QUANTITATIVE RESULTS OF ALL MODELS ON MARKET-1501. FOR IS, SSIM AND M-SSIM, HIGHER SCORE MEANS THE BETTER, FOR FID AND LPIPS, THE LOWER THE BETTER

Model	Market-1501				
	IS	FID	SSIM	M-SSIM	LPIPS
PN-GAN [33]	2.586	53.481	0.259	0.776	0.316
DVG [5]	2.814	107.567	0.303	0.790	0.331
PATN [56]	2.951	58.996	**0.311**	0.811	0.304
G-Pix2Pix [1]	**3.263**	103.793	0.254	0.758	0.363
MR-Net	2.831	**52.841**	0.302	**0.857**	**0.296**
Ground-Truth	3.612	—	—	—	—

may lead to a whole masked images black. To make a fair comparison, we only evaluate the M-SSIM inside the ground-truth bounding box of persons in each image of Penn Action.

1) Results: The results are summarized in Tab. I and Tab. II. Particularly, we highlight that (1) On non-iconic datsets, our inception scores and FID are better than those of competitors. This indicates that, perceptually, the images synthesized by our models have better quality than those produced by competitors. Our synthesized images are more authentic with diverse content. Meanwhile, our method is comparable to the competitors on iconic dataset, i.e., Market-1501. (2) On almost all datasets, our SSIM and M-SSIM scores are higher than those of competitors. This suggests that our synthesized images can better maintain the overall image structure, particularly, the image scene. Note that, as mentioned before, images in BBC-Pose dataset have dynamic background, i.e., the news being broadcast (also as shown in Fig. 5). Thus the holistic image structure may be changed.

However, we realize that these metrics can only reflect the general image quality, but not the quality of pose changes in the target images. To this end, we give more qualitative results, and subjective user study to show the advantage of our model over the other competitors.

C. Qualitative Results

We present our qualitative results on three datasets in Fig. 5, where we compare four generated images of our model on each dataset and baselines conditioned on the same reference images and target poses.

As shown in Fig. 5(a), we find that generated images by competitors are blurry, or have distorted person poses on Penn Action dataset. This is reasonable, since previous models do not consider the proposed tasks – pose-guided human synthesis in the non-iconic views. Specifically, when human body moves (the first and second row of Fig. 5(a)), the competitors could hardly learn the target pose except that DVG and PATN, which could only generate a blurry profile of the target human body as well. When the position of human body does not change a lot, and with small body scale (the third row of Fig. 5(a)), the specific parts of body such as head produced by competitors are not so vivid as ours. The generated images of BodyROI7 are not only too blurry to be discriminated, but also filled with artifacts. This indicates that this model fails in the non-iconic image synthesis; and we will not compare it in the rest experiments. Other methods can learn nothing from the dataset, thus copying the reference images. Compared to these results, our method could synthesize the correct target human bodies with appropriate identities and background. This demonstrates the efficacy of our model in parsing background and human body to synthesize a new image.

In Fig. 5(c), we compare the results on Market-1501. All models generate much more satisfied results than those on Penn Action, with the only exception – PG^2 which still synthesizes vague images. However, we find that the competitors are failed to synthesize the details in reference images, such as logos on T-shirt (the third row of Fig. 5(c)) and color (the third row) of the clothes; in contrast, our model can still keep such traits of person body in the synthesized image. Moreover, note that in the second row, the target human does not have a bag on his back, while the reference human body has one; therefore, the successfully generated image should also have the shoulder straps, as shown in our result of Fig. 5(c). This further validates that two-pathway structure could help better synthesize person images in the non-iconic views. Furthermore, it can be seen that our results are much more genuine than other baselines, which is compatible with the fact that our M-SSIM scores on this dataset far overpass the competitors.

In Fig. 5(b), we have the results on BBC-Pose, which is an easy dataset. PN-GAN gives the worst results; and the original human bodies are still in the generated images, but the target poses can hardly be found. These results show that PG^2 is a very effective model on BBC-Pose that it could generate sharp images due to the discriminator; thus it achieves very high SSIM and M-SSIM scores. Nonetheless,

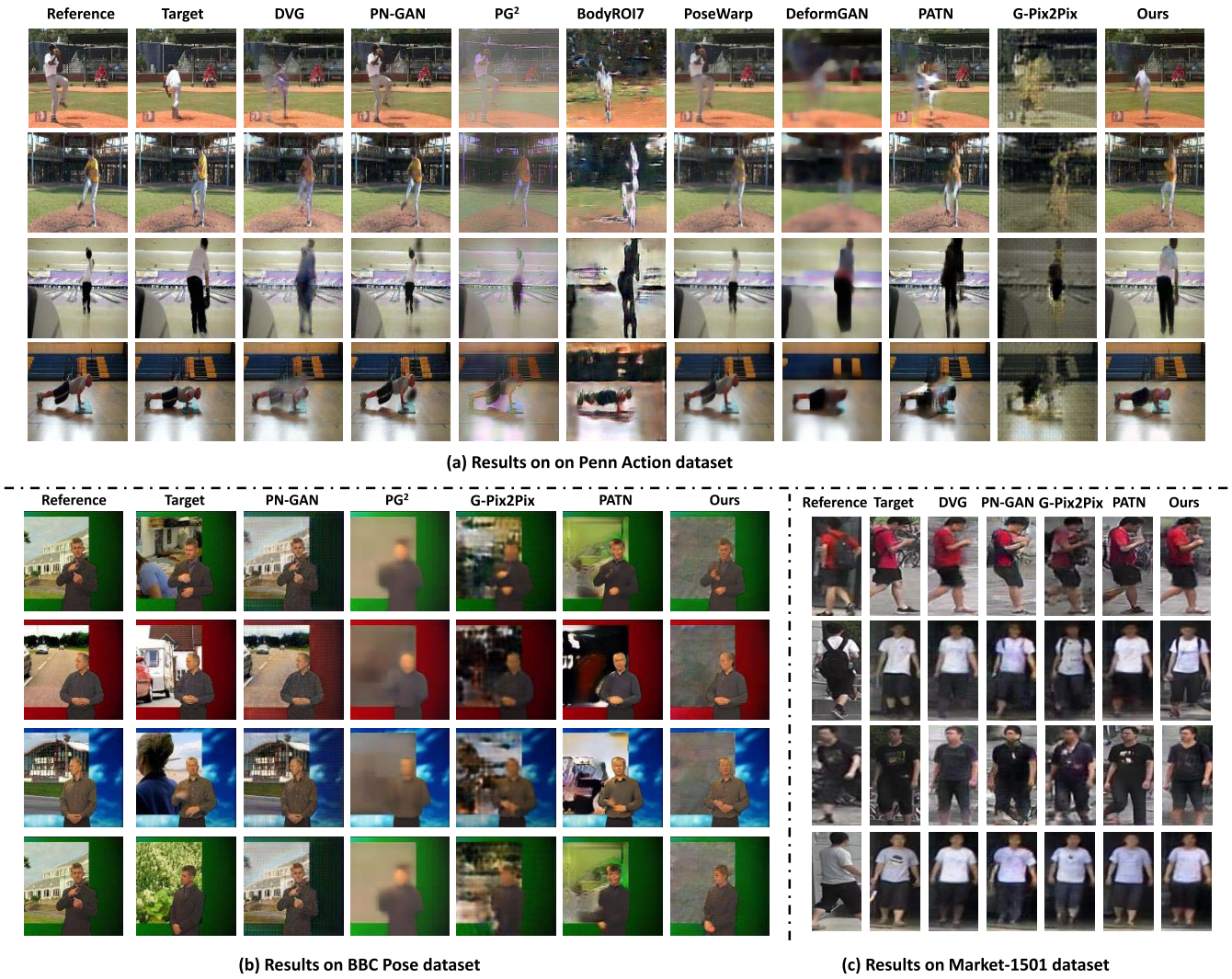

Fig. 5. Qualitative Results on each dataset.

our model could generate not only the accurate human body, but it could also surprisingly restore the wrinkles of cloth and rough structure of faces as well. This further shows the efficacy of our proposed model.

In all cases, our model could generate more sharp background, thanks to our multi-branch structure, which can decompose and learn to store the background, as well as the most useful information in the reference images. Furthermore, the poses in our synthesized images are very accurate. This indicates that the latent embedding of target human pose learned by our main pose-guided encoder is representative enough to help synthesize non-iconic person images. We provide more qualitative results on Penn Action in Fig. 6, which show that our method consistently outperforms other competitors.

D. Ablation Study

1) Are There Any Naive Methods for Pose-Guided Synthesis in the Non-Iconic Views?: No, naive methods do not work here. Particularly, we compare two naive baselines: (a) *union-crop*: We crop out both the bounding boxes of human body in reference images and target images. These patch pairs are used to train all competitors and the rest parts are added to the generated results of each model to get the final results. As shown in Fig. 11. This method cannot learn to synthesize the target poses correctly. (b) *crop-inpaint*: We crop the human body part out for training. During inference, we scale the generated image to the size of the bounding box of corresponding target image and add the generated image and the reference image after cropping together. The DeepFill model [49] trained on Penn Action is utilized to inpaint the cropped region. As visualized in Fig. 12, crop-inpaint will produce the inconsistent background, due to the huge difference between the reference and target image in training period.

2) Number of Refinement Iterations: We train our model with different number of iterations. With more iterations, the model gets higher metrics, which means it can better learn image synthesis, as illustrated in the 4, 5, 9 column of Fig. 8. Particularly, the model trained with one iteration could only generate images with severe distortion, and models with more iterations gradually correct this artifact. Note that in the last two rows, the images generated by 1 and 2 iterations still have black areas from the original human bodies in the reference

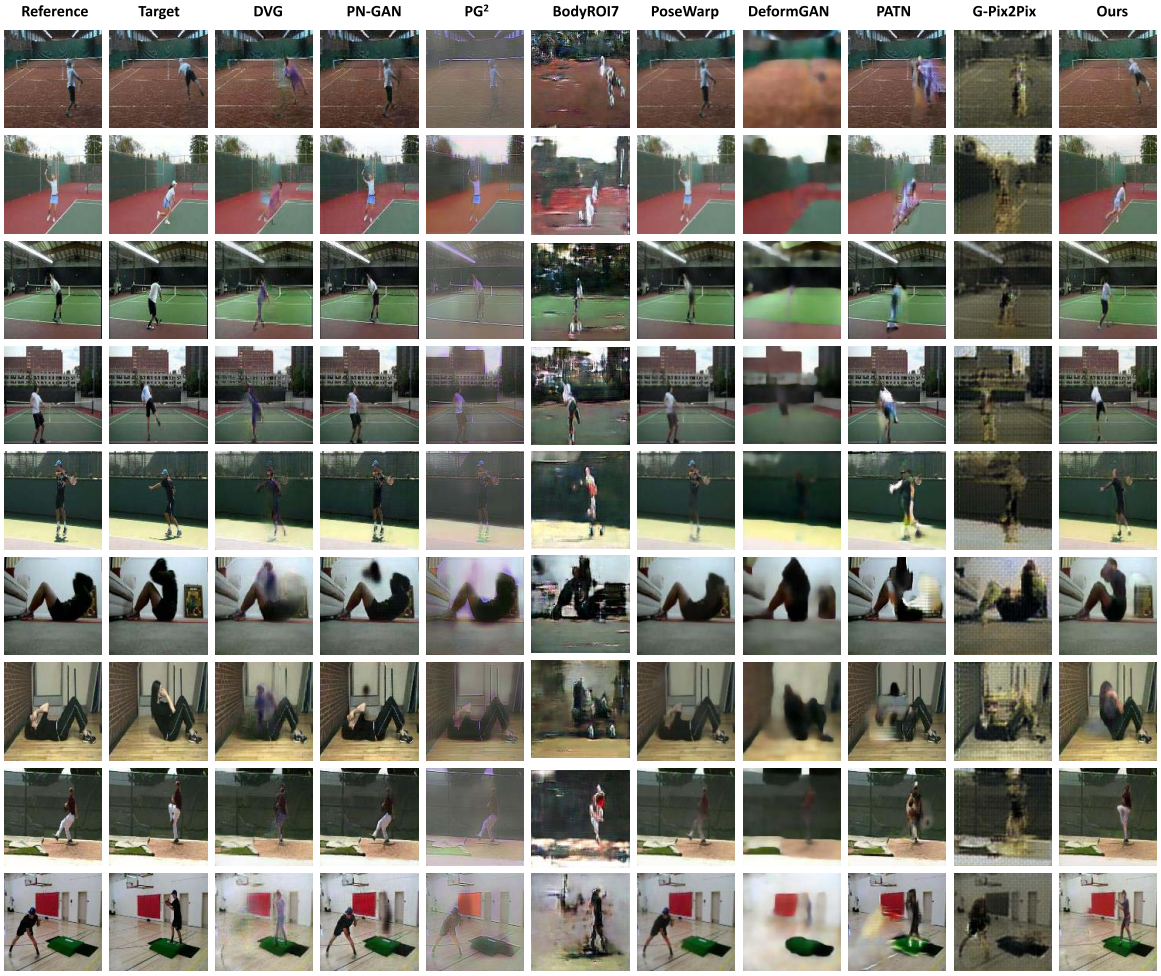

Fig. 6. More synthesized results of our full model on Penn Action.

TABLE III

QUANTITATIVE RESULTS OF ABLATION STUDY. FOR IS, SSIM AND M-SSIM, HIGHER SCORE MEANS THE BETTER, FOR FID AND LPIPS, THE LOWER THE BETTER. ROI INDICATES WHETHER TRAINING WITH ROI PERCEPTUAL LOSS, #ITER. MEANS NUMBER OF ITERATIONS USED, #PATH INDICATES WHETHER USING TWO-PATHWAY STRUCTURE AND N.L. DENOTES WHETHER USING NON-LOCAL BLOCK IN DECODER

Model						Penn Action					BBC-Pose				
λ_1	λ_2	RoI	#Iter.	#Path	N.L.	IS	FID	SSIM	M-SSIM	LPIPS	IS	FID	SSIM	M-SSIM	LPIPS
0.5	0.05	✓	1	✓	✓	6.366	28.717	0.724	0.945	0.110	1.509	149.264	0.433	0.869	0.381
0.5	0.05	✓	2	✓	✓	6.856	32.241	0.701	0.946	**0.106**	1.543	127.649	0.454	0.872	0.350
0.5	0.05	×	3	✓	✓	**6.922**	35.464	0.702	0.944	0.121	1.571	106.981	0.407	0.863	0.338
0.5	0.05	✓	3	×	✓	6.509	40.291	0.691	0.942	0.119	1.708	109.187	0.527	0.892	0.353
0.5	0.05	✓	3	✓	×	6.554	37.236	0.693	0.945	0.116	1.707	120.439	0.496	0.891	0.367
0	0.05	✓	3	✓	✓	5.309	86.257	0.556	0.927	0.287	1.764	129.562	0.454	0.876	0.353
0.1	0.05	✓	3	✓	✓	6.149	34.045	0.636	0.952	0.115	**1.767**	107.019	0.468	0.877	0.348
0.5	0.15	✓	3	✓	✓	6.772	23.542	0.734	0.957	0.153	1.689	103.307	0.479	0.880	0.351
0.5	0.5	✓	3	✓	✓	6.512	39.817	0.629	0.930	0.183	1.719	110.238	0.497	0.882	0.353
0.5	0.05	✓	3	✓	✓	6.868	**18.583**	**0.751**	**0.962**	0.110	1.751	**95.114**	**0.527**	**0.892**	**0.321**

images. The results further get improved over 3 iterations. The qualitative results with more refinement iterations are shown in Fig. 10. It indicates, even with only 3 iterations, our model can still synthesize relatively good results. This further demonstrates the efficacy of our model.

3) Effectiveness of RoI Perceptual Loss: We train our model without RoI perceptual loss. Both the quantitative (Tab. III) and qualitative results (the 6,9 column in Fig. 8) show that the quality of synthesized images gets dropped significantly; and the model cannot learn the identity information of person body.

In addition, we try to validate the efficacy of ROI-perceptual loss by using this term when training PN-GAN on Penn Action and Market-1501. The results are shown in Fig. 13. We find that on Market-1501, the ROI-perceptual can help the model improve the performance. However, on Penn Action,

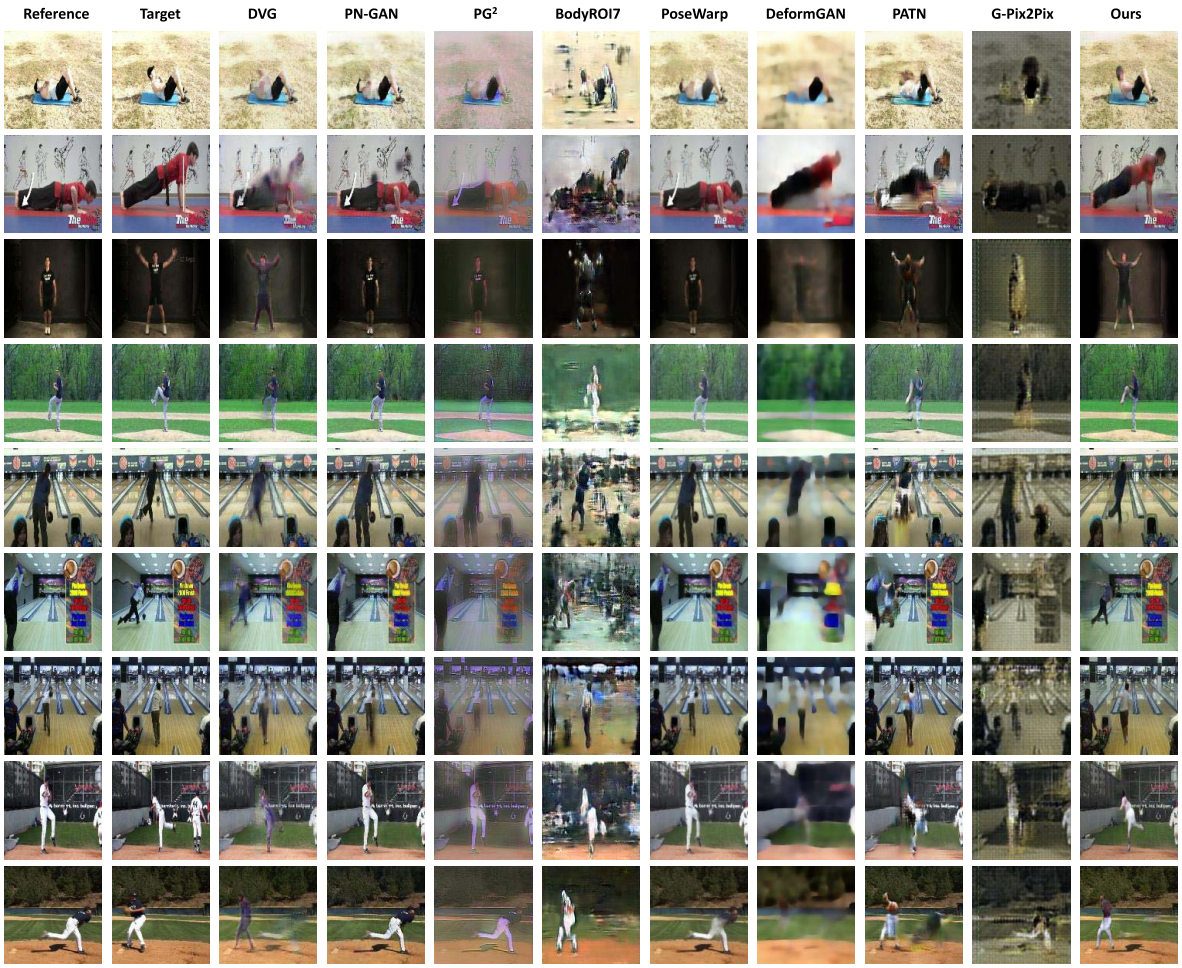

Fig. 7. More synthesized results of our full model on Penn Action.

TABLE IV

QUANTITATIVE RESULTS OF ABLATION STUDY. FOR IS, SSIM AND M-SSIM, HIGHER SCORE MEANS THE BETTER, FOR FID AND LPIPS, THE LOWER THE BETTER. ROI INDICATES WHETHER TRAINING WITH ROI PERCEPTUAL LOSS, #ITER. MEANS NUMBER OF ITERATIONS USED, #PATH INDICATES WHETHER USING TWO-PATHWAY STRUCTURE AND N.L. DENOTES WHETHER USING NON-LOCAL BLOCK IN DECODER

Model						Market1501				
λ_1	λ_2	RoI	#Iter.	#Path	N.L.	IS	FID	SSIM	M-SSIM	LPIPS
0.5	0.05	✓	1	✓	✓	2.639	89.827	0.299	0.631	0.311
0.5	0.05	✓	2	✓	✓	2.831	81.165	0.291	0.788	0.308
0.5	0.05	×	3	✓	✓	2.498	115.307	0.297	0.791	0.330
0.5	0.05	✓	3	×	✓	2.486	87.244	0.286	0.785	0.316
0.5	0.05	✓	3	✓	×	2.597	90.005	0.292	0.788	0.351
0	0.05	✓	3	✓	✓	2.339	175.232	0.203	0.844	0.434
0.1	0.05	✓	3	✓	✓	**2.929**	79.786	0.228	0.851	0.372
0.5	0.15	✓	3	✓	✓	2.548	203.361	0.251	0.847	0.483
0.5	0.5	✓	3	✓	✓	2.659	157.694	0.239	0.806	0.462
0.5	0.05	✓	3	✓	✓	2.831	**52.841**	**0.302**	**0.857**	**0.296**

the generated images are still very similar to the reference images. One possible reason is the low capacity of this model in dealing with images from the non-iconic views. Under such setting, the model cannot learn from the target images, even if we added ROI-perceptual loss to focus on human body. This supports our conclusion that non-iconic setting is not trivial and is worth to be considered.

4) Effectiveness of Pose-Guided Encoder and Non-Local Block: The pose-guided encoder is replaced by a single encoder whose structure is the same as $\mathbf{E}_{fg}$, except that the input is the concatenation of pose image and reference human body image. The results of model with and without two-pathway structure (as shown in the 1,4 row in Fig. 8) indicate that, for some complex poses, the two-pathway structure

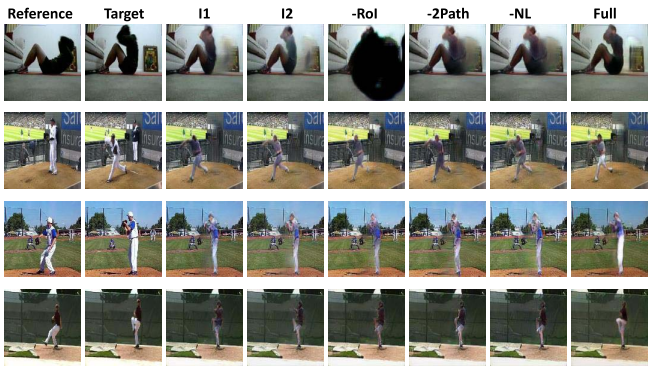

Fig. 8. Comparison of the different methods with the same reference images and target poses. "I1" indicates the image with one iteration; RoI denotes the RoI perceptual loss; 2Path denotes two-pathway structure; NL denotes non-local block.

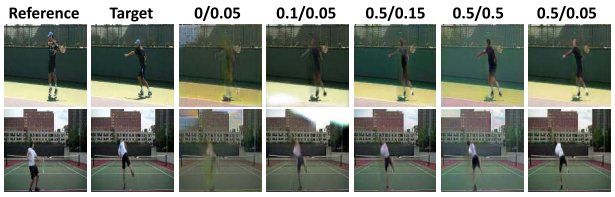

Fig. 9. Comparison of the different settings of hyper-parameter λ_1 and λ_2 with the same reference images and target poses. "0/0.05" denotes $\lambda_1 = 0, \lambda_2 = 0.05$ and the others are the same.

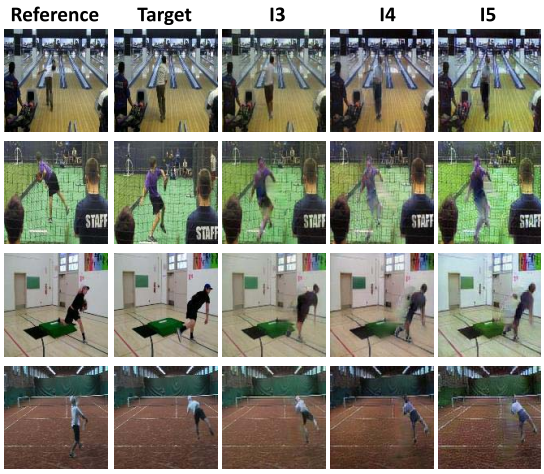

Fig. 10. Qualitative comparison with more number of refinements. "I3" indicates the Image with 3 iterations.

could indeed help the model focus on corresponding parts of reference person bodies. Moreover, the results prove that even without the non-local blocks, our model still outperforms the baselines on Penn Action in Inception Score and M-SSIM, and the non-local blocks could further boost the performance.

5) Effectiveness of Hyper-Parameter λ_1 and λ_2: We try to train our model with different sets of λ_1, λ_2. The qualitative results are shown in Fig. 9 which results reflect that (a) It is still necessary to use original perceptual loss. (b) As λ_2 increases, the performance slight drops. We think these are

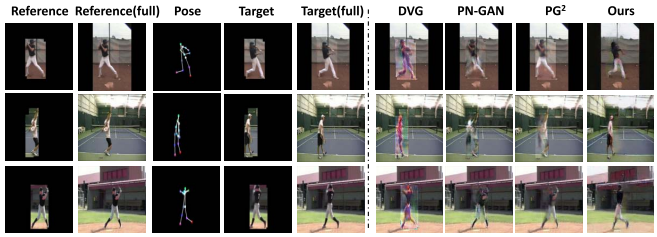

Fig. 11. Competitors are trained by cropped images and the remain parts are added to the generated results. Our results are produced by our framework.

TABLE V

USER STUDY RESULTS ON PENN ACTION AND MARKET-1501. S, AU. AND AC. INDICATES THE SHARPNESS, AUTHENTICITY AND ACCURACY RESPECTIVELY. HIGHER IS BETTER FOR ALL MEASURES

Model	Penn Action			Market-1501		
	S	Au	Ac	S	Au	Ac
PG²	0.67	0.17	0.24	0.64	0.42	0.82
PN-GAN	0.80	0.12	0.31	0.64	0.35	0.87
DVG	0.62	0.14	0.85	0.70	0.36	0.87
PATN	0.74	0.10	0.61	0.72	0.35	0.88
G-Pix2Pix	0.16	0.04	0.14	0.75	0.26	0.79
Ours	0.74	0.36	0.88	0.75	0.40	0.91
Real data	0.90	0.49	—	0.76	0.49	—

consistent with our design since the ROI-perceptual loss is proposed to improve the specific quality for regions of human body. If its scale is comparable to or more than that of the perceptual loss, then the model will stress on the human body but omit the background, thus resulting in worse performance.

6) User Study: We conduct user study here. Fifteen subjects unfamiliar with the project were recruited for the study. We randomly picked 50 generated images with their ground truth. During the user study, the participants are asked to evaluate these three metrics: (a) Sharpness: the participants are presented with images (could be generated by different methods or real) and are asked to judge whether the images are clear enough, which can be defined as "clear enough to discriminate what this image is about". (b) Authenticity: we provide participants with image pairs containing a generated image and a corresponding image, and ask them to decide which one is fake. For the first question, the participants are asked to give a true or false choice and for the latter they need to choose one image. (c) Accuracy: we present generated images with corresponding target images, and ask participants to judge the accuracy, which means whether one generated image is showing the same pose as the target image. We report the mean scores of all metrics in Tab. V. These results confirm our conclusion that, compared to baselines, our model can have better results on the pose-guided person image synthesis both in iconic and non-iconic views. The generated images are both much more authentic and accurate than the images produced in other models.

7) Failure Cases: We provide two failure cases to further discuss this task. In Fig. 14, the target image needs the extra information which is not shown in the reference image. Even in such a case, our model can still be able to recover the general shape of the target pose. This actually reflects the

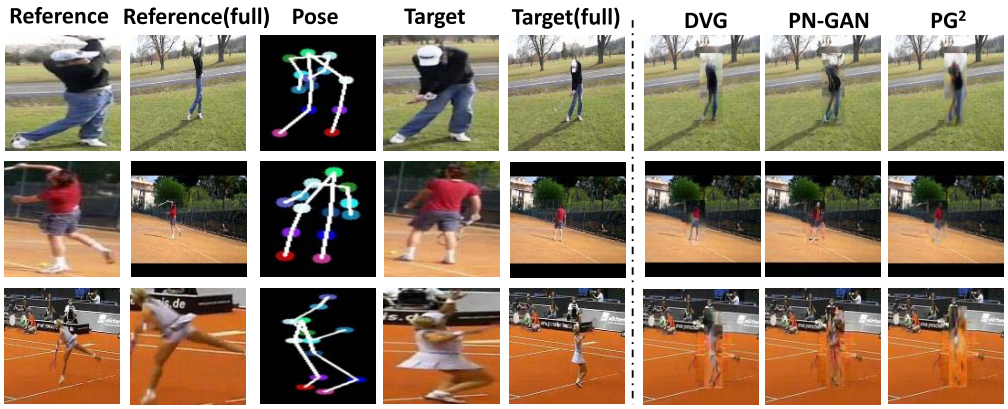

Fig. 12. Competitors are trained by cropped images and the cropped out area is inpainted by DeepFill.

Fig. 13. PN-GAN is trained with ROI-perceptual loss on Penn Action and Market-1501.

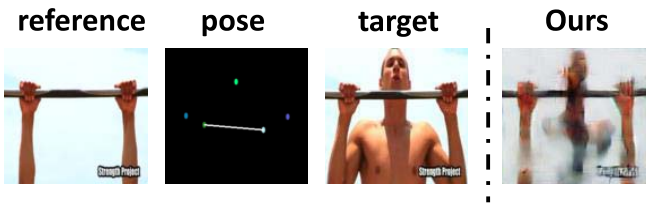

Fig. 14. A failure case of our model caused by incomplete reference information.

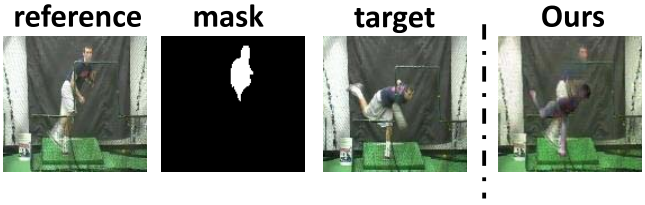

Fig. 15. A failure case of our model caused by inaccurate human mask.

efficacy of our model. In Fig. 15, the mask predicted by Mask-RCNN may not be accurate enough, which results in the low quality of the generated image. However, our strategy is still

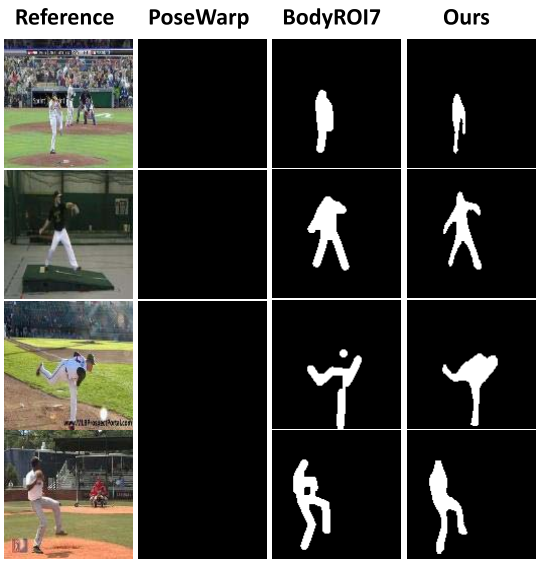

Fig. 16. Visualization of different strategies to produce masks.

the best among the existing methods adopting masks to split foreground and background, including PoseWarp, which utilizes a network to learn the masks, and BodyROI7 which uses morphological erosion and dilation to transfer joint heatmaps to masks. To further show this point, we visualize some samples with masks produced by these methods in Fig. 16. We find that PoseWarp cannot learn any useful information to produce good mask, so that this model is failed to predict all target poses. As a comparison, non-learnable methods, i.e., BodyROI7 and ours are more suitable. However, since only fixed parameter can be used in BodyROI7 for those images in non-iconic views, the changeable body scale would lead to the worse performance.

V. Conclusion

This paper presents a new model targeting pose-guided person image synthesis in the non-iconic views. We propose the MR-Net in parsing and encoding the information of foreground person body, and background scene separately. RoI perceptual loss and iterative refinement strategy are proposed

here. Experiments on three datasets demonstrate the effectiveness of our model over the competitors.

REFERENCES

[1] B. Albahar and J.-B. Huang, "Guided image-to-image translation with bi-directional feature transformation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9016–9025.

[2] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[3] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8340–8348.

[4] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," 2017, *arXiv:1703.10717*. [Online]. Available: http://arxiv.org/abs/1703.10717

[5] H. Cai, C. Bai, Y.-W. Tai, and C.-K. Tang, "Deep video generation, prediction and completion of human action sequences," 2017, *arXiv:1711.08682*. [Online]. Available: http://arxiv.org/abs/1711.08682

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*. [Online]. Available: http://arxiv.org/abs/1812.08008

[7] J. Charles, T. Pfister, M. Everingham, and A. Zisserman, "Automatic and efficient human pose estimation for sign language videos," *Int. J. Comput. Vis.*, vol. 110, no. 1, pp. 70–90, Oct. 2014.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[9] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-gan for pose-guided person image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 474–484.

[10] J. He, A. Lehrmann, J. Marino, G. Mori, and L. Sigal, "Probabilistic video generation using holistic attribute control," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 452–467.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[13] S. C. Hidayati, C.-C. Hsu, Y.-T. Chang, K.-L. Hua, J. Fu, and W.-H. Cheng, "What dress fits me best?: Fashion recommendation on the clothing style for personal body shape," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, 2018, pp. 438–446.

[14] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng, "FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 275–283.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[16] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 694–711.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[18] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[19] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2014, pp. 740–755.

[20] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," 2018, *arXiv:1804.07723*. [Online]. Available: http://arxiv.org/abs/1804.07723

[21] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[23] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 406–416.

[24] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 99–108.

[25] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2813–2821.

[26] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 542–551.

[27] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," 2016, *arXiv:1601.06759*. [Online]. Available: http://arxiv.org/abs/1601.06759

[28] A. V. D. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 4797–4805.

[29] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. Autodiff Workshop 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017.

[30] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[31] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Unsupervised person image synthesis in arbitrary poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8620–8628.

[32] G.-J. Qi, "Loss-sensitive generative adversarial networks on lipschitz densities," 2017, *arXiv:1701.06264*. [Online]. Available: http://arxiv.org/abs/1701.06264

[33] X. Qian *et al.*, "Pose-normalized image generation for person re-identification," 2017, *arXiv:1712.02225*. [Online]. Available: http://arxiv.org/abs/1712.02225

[34] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2015, pp. 234–241.

[37] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3408–3416.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[39] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3738–3746.

[40] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 5, 2019, doi: 10.1109/TPAMI.2019.2938523.

[41] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "HOGgles: Visualizing object detection features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1–8.

[42] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[44] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 341–349.

[45] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[46] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2014, pp. 372–386.

[47] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," 2015, *arXiv:1506.06579*. [Online]. Available: http://arxiv.org/abs/1506.06579

[48] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122
[49] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.
[50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[51] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2248–2255.
[52] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.
[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[54] R. Zhou, X. Chang, L. Shi, Y.-D. Shen, Y. Yang, and F. Nie, "Person reidentification via multi-feature fusion with adaptive graph learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1592–1601, May 2020.
[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017, *arXiv:1703.10593*. [Online]. Available: http://arxiv.org/abs/1703.10593
[56] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2347–2356.

Chengming Xu received the bachelor's degree in computer science from Fudan University in 2018, where he is currently pursuing the Ph.D. degree in statistics advised by Prof. Y. Fu.

Yanwei Fu received the Ph.D. degree from the Queen Mary University of London in 2014 and the M.Eng. degree from the Department of Computer Science and Technology, Nanjing University, China, in 2011. He held a postdoctoral position at Disney Research, Pittsburgh, PA, USA, from 2015 to 2016. He is currently a tenure-track Professor with Fudan University. His research interests are image and video understanding, and life-long learning.

Chao Wen received the B.E. degree in computer science from the University of Electronic Science and Technology of China, in 2018. He is currently pursuing the master's degree with the Academy for Engineering and Technology, Fudan University. His research is focused on 3D computer vision and human image synthesis.

Ye Pan received the B.Sc. degree in communication and information engineering from Purdue/UESTC in 2010 and the Ph.D. degree in computer graphics from the University College London (UCL) in 2015. She is currently an Associate Professor with Shanghai Jiao Tong University. Her research interests include AR/VR, avatars/characters, 3D animations, HCI, and computer graphics. She was selected as a Forbes China 30 Under 30 In Science 2019.

Yu-Gang Jiang (Member, IEEE) is currently a Professor of computer science with Fudan University, China, where he also serves as the Dean of the School of Computer Science and the School of Software. His research is focused on multimedia, computer vision, and robust and trustworthy AI. He is also an Associate Editor of ACM TOMM.

Xiangyang Xue (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in communication engineering from Xidian University, Xi'an, China, in 1989, 1992, and 1995, respectively. He is currently a Professor of computer science with Fudan University, Shanghai, China. His research interests include computer vision, multimedia information processing, and machine learning.