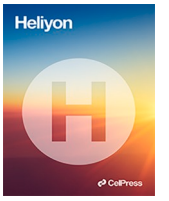
Heliyon 10 (2024) e23568

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Virtual group photography based on scene and person perception

Zhenquan Shi [a], Hongming Lu [b,e], Xinyue Sun [b], Jing Guo [a], Yongzhen Ke [b,c,*], Kai Wang [b], Fan Qin [d]

[a] School of Software Engineering, Tiangong University, Tianjin, China
[b] School of Computer Science and Technology, Tiangong University, Tianjin, China
[c] National Demonstration Center for Experimental Engineering Training Education(Tiangong University), Tianjin, China
[d] Business School, Nankai University, Tianjin, China
[e] Fitow (Tianjin) Detection Technology Co. Ltd, China

ARTICLE INFO

Keywords:
Virtual group photo
Person image generation
Multi-task learning
Pose transfer
Person perception

ABSTRACT

Group photos have become indispensable in various gathering scenarios, such as family reunions, friends' gatherings, competitions, conferences, store openings, and school graduation ceremonies. The researchers tried automatically adding people who could not participate in the group photo. However, the current research on generating the pose or position of the person by context prediction of the group photo ignores the individual attributes (such as height and body shape) of the target person and does not consider the pose and boundary of the person at the same time. To address these issues, we propose a virtual group photography model that combines the global context of a group photo and the individual attributes of the target person. The model is divided into two stages. The first stage is to predict the person's position, pose, and boundary in the new group photo based on the context of the input group photo and the person's characteristics. The second stage generates new group photos based on the first stage's pose and boundary results. The experimental results show that our method can significantly improve the harmony and authenticity of the synthesis of people in group photos and synthesize the characters that should exist in the group photo, which is very suitable for the field of group photos.

1. Introduction

The synthesis of people in group photos is developed from research in image synthesis, which has been very popular in recent years. This research mainly applies to scenes where important people are missing in group photo images, and it is also intended to replace complex manual P-picture work and improve the harmony of composite images.

Many valuable properties in group photo synthesis have an important impact on the harmony of synthesized images. These properties include the interaction between different characters in the group photo (such as posture and expression), clothing attributes (color, style, style, and type), scene attributes (dynamic or static), and the layout of key objects in the scene (with or without photography props), as well as the attributes of the target person (gender, age, height, and weight) and social relationships (friends, couples, teachers, students, or rivals). In addition to the above attributes, many valuable attributes may not be mentioned, which will more or less affect the harmony of a composite photo image.

In recent years, some researchers have also made important contributions to group photo synthesis [1,2], specifically conducted

* Corresponding author. School of Computer Science and Technology, Tiangong University, Tianjin, China.
E-mail address: keyongzhen@tiangong.edu.cn (Y. Ke).

https://doi.org/10.1016/j.heliyon.2023.e23568
Received 13 June 2023; Received in revised form 10 October 2023; Accepted 6 December 2023
Available online 13 December 2023

targeted exploration research on one or several of the attributes mentioned above, and proposed many inspiring ideas. For example, they respectively considered the influence of the interaction of character boundaries and poses in the group photo image on the harmony of the group photo image, which greatly promoted its synthesis.

Based on scene and person perception, we propose a virtual group photography method that simultaneously predicts the target person's position, pose, and boundary. Our method considers not only the context of the person in the group photo but also the attributes of the target person(gender, age, height, and weight). As shown in Fig. 1, our method is divided into two stages: prediction and image generation. In the position, pose, and boundary prediction stages, to make the predicted human poses and boundaries roughly consistent in morphology, we improve the multi-task learning method based on the basic architecture of pix2pixHD [3], combine the individual attributes of the characters at the same time predict poses and boundaries so that the predicted poses and boundaries are more consistent with the characteristics of the target person. In the image generation stage, we improve the basic architecture of the ADGAN [4] network. In addition to individual character attributes and predicted character poses, the single-person boundary predicted in the previous stage is used to constrain the character pose transfer.

Our method makes the characters inserted into the group photo not only have good interaction with other characters of the group photo in terms of poses and boundaries but also retain the inherent attributes of the source characters to a large extent. The experimental results show that our method is perfectly suitable for virtual group photography and greatly improves the harmony and authenticity of the composite image of group photos.

In summary, the contributions of this paper are as follows.

1. We propose a virtual group photography method that decomposes a complex problem into two relatively simple sub-networks.
2. Based on the context of the person in the group photo and the attributes of the target person, we propose a method for simultaneously predicting the pose and boundary of the target person.
3. We propose a pose transfer method constrained by the boundaries of the target character.

2. Related Works.

With the continuous development and optimization of the image synthesis field, the photo synthesis field has gradually shown new application potential, and many related researchers are exploring this field. In the early development of this field, researchers first explored the image synthesis of characters and scenes and only needed to consider the fusion of foreground characters and background images, gradually transitioning to the image synthesis of two-person group photos. It is also necessary to consider the problem of two-person interaction in the scene; until now, it has developed into a multi-person group photo synthesis; based on many previous considerations, the problem of multi-person interaction has been further considered.

In 2018, Yogesh Singh Rawat et al. [5] introduced the concept of color energy and proposed a Spring-Electric graphical model for obtaining visually balanced layouts and dynamic visual elements. This model can be based on the composition and color distribution in the background image To provide users with suggestions for taking visually balanced photos, such as providing users with real-time suggestions on the arrangement, position, and relative size of people in the group photo, aiming to provide real-time help for users to take high-quality group photos. In 2019, Mihai Zanfir et al. [6] proposed a HUSC (Human Body and Scene Synthesis) framework for realistically synthesizing people with different appearances in novel poses and scenes. The model first estimates the scene image's ground plane, then places the target 3D human body model in the 3D scene on the geometric level. It is hoped that the position, size ratio, occlusion, ground, and other issues will be considered on the physical level, and then the target 3D human body model will be adjusted according to the source image. The 3D human body is rendered and seamlessly synthesized into the scene through the scene synthesis network. In 2020, Shuchen Weng et al. [7] proposed a MISC framework for synthesizing character images with various conditions under different backgrounds for conditional image generation and image synthesis. They mainly injected multiple related

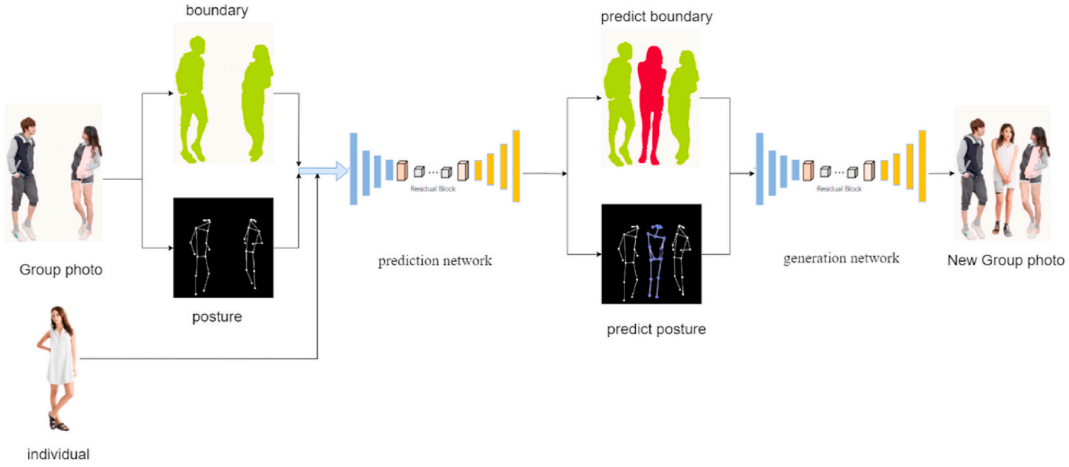


Fig. 1. Overall idea.

Conditions, proposed a space-adaptive image synthesis method, and generated a realistic image. From the results, it performs well in some scenes but still lacks harmony, although some human body color attributes can be customized.

In 2020, Yujia Wang et al. [8] proposed a new method of automatically synthesizing the poses of virtual characters based on the user's poses so that the virtual characters can match each other from the perspective of visual aesthetics to form realistic and vivid photos. The method is divided into two stages. The first stage is the learning stage. By learning the poses in the existing single-person and double-person photo data sets, the model has an aesthetic evaluation standard for the poses. The second stage is the optimization stage. Continuously evaluate and adjust the posture of the virtual character until a posture that best matches the user's posture and has the highest aesthetic score is optimized. It is also the first time that the interaction between two characters has been considered in the image synthesis of group photos, and some exploration research on the image synthesis of multi-person group photos has been derived based on his ideas. In 2020, the Facebook team [2] was the first to generate a semantic map of a new character through the context of other people in the scene and render the character image through the appearance component. Although this paper opened up a new research idea, there are some defects that the style of the generated character semantic map is too different from that of the source character (such as hair length, body size, and head shape). As a result, the character's own identity is largely lost. Although the rendered character image is consistent with the source character in components such as the face and clothes, the overall image is quite different from the source character, making the generated character image look fake and lack a sense of reality. In order to fill the previous deficiencies, in 2022, Prasun Roy et al. [1] proposed to change the semantic map of predicting new characters through the context of the characters in the scene into a posture framework for predicting new characters. In this way, pose transfer can be performed through the predicted pose. The character image after posture migration is closer to the source character image and more realistic. However, the idea of this article also has limitations; that is, posture transfer only predicts human images based on the model trained on the images in the training set, resulting in the characters being universal (the body shape is closer to the common images in the training set). If the character attributes in the source image are relatively distinct (tall, short, fat, thin), then the inherent attributes of the character will inevitably be lost, making the target character in the final synthesized image different from the source character and lacking a sense of reality.

2. Methodology

Our proposed method can predict both the pose and the boundary based on the person's context in the group photo and the inherent properties of the target person. As shown in Fig. 2, our model includes two sub-networks: position, pose, boundary prediction network, and person image generation network. The input to our model is group photo images and single-person images. Multi-person and single-person boundaries and poses are obtained through instance segmentation and pose estimation. The features Y of boundaries and poses of multiple people are extracted through multi-task learning. Subsequently, the encoded features X of the single person's boundary and pose are fused with the abovementioned features to obtain new features and then output the predicted boundary and pose through decoding. Finally, the predicted character boundaries are used as constraints to guide the generation of character images in the image generation sub-network. We will introduce the position, pose, and boundary prediction sub-network in Section 3.1 and the person image generation sub-network in Section 3.2 in detail.

2.1. Position, pose, boundary prediction network

2.1.1. Boundary, posture encoding and decoding

In character boundary and pose encoding and decoding, the character boundary is derived from the multi-semantic label map given and divided in the MHP dataset [9]. We merged them and re-encoded them to construct the character boundary map dataset for

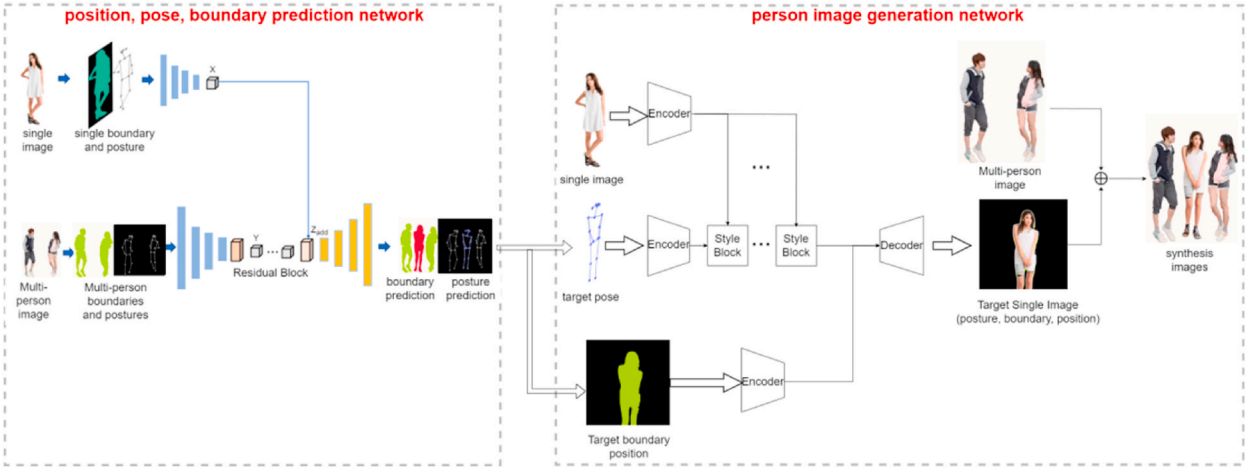


Fig. 2. The proposed framework.

predicting the boundaries of new characters. In addition, we randomly remove one of the multiple characters in the group photo, keep the remaining characters, and combine their boundaries to form a multi-person boundary map. Pose key points are extracted by the human key point detection method [10] for multiple and single-person pose key points, and each pose is represented as a set of K body joint positions (key points), denoted by $P \in R^{K\times H\times W}$. It means that we construct a binary heatmap of K channels for each individual pose P_i, where each channel corresponds to a specific key point, each key point is encoded as 1 in the corresponding position of the respective channel, and the other positions are 0. The multi-person pose is aggregated into n heat maps, denoted as P_n, where each channel contains the spatial encoding of specific key points of n posed participants. Both input bounds and poses are resized to a spatial dimension of 512*512.

The overall boundary, pose encoder-decoder network is improved based on the basic architecture of pix2pixHD [3]. The first important improvement is integrating the features of single-person boundaries and poses in the feature stage before its decoder, aiming to predict the new person's boundary and pose closer to the target person. The second important improvement is to perform multi-task learning on the basic architecture, sharing the same network in both tasks. In this way, the predicted human pose and boundary are consistent in morphology.

2.1.2. Feature fusion

In the feature fusion stage, we fuse the 512-dimensional single-person features after the single-person boundary or pose encoding with the 512-dimensional multi-person features after the residual module. Since we aim to predict poses or boundaries within the same image, we sum single-person and multi-person features in each dimension. This operation neither increases parameters nor increases computational complexity [11]. It enables the model to learn the contextual information of the person in the group photo and the target person's information.

The method can be expressed as:

$$Z_{add} = \sum_{i=1}^{c} (X_i + Y_i) * K_i \tag{1}$$

Among them, X_i and Y_i represent the channels of two different input X(single-person feature) and Y(multi-person feature) respectively, K_i represent the channel after fusion, and $*$ represent convolution.

2.1.3. Training and loss

The discriminator in this task refers to the multi-scale discriminator design of pix2pixHD [3], which aims to distinguish real images and synthetic images of different scales, thereby enhancing the capabilities of the discriminator. Specifically, we downsample real and synthetic images by a factor of 2 and 4, create image pyramids of 3 scales, and then train three discriminators, D1, D2, and D3, on three different scales to distinguish real images from synthetic images, as shown in Equation (2):

$$\min_{G} \max_{D_1,D_2,D_3} \sum_{k=1,2,3} \mathscr{L}_{\mathrm{GAN}}(G, D_k) \tag{2}$$

Among them, the $\mathscr{L}_{\mathrm{GAN}}(G, D)$ function is given by the following formula (3). For our task, the goal of generator G is to predict the identity of the new person from the context of the group photo boundary or pose, and the goal of discriminator D is to distinguish whether the predicted boundary or pose is true or false. Specifically, the training data set is given as a set of corresponding image pairs $\{(s_i, t_i, x_i)\}$, where s_i is the boundary map or pose map of a group photo of multiple people, t_i is the boundary map or pose map of the missing person image, and x_i is the corresponding ground truth boundary map or pose map. Conditional GANs aim to model the conditional distribution of ground-truth boundaries and pose maps, given input boundary maps and pose maps, by adversarial training as follows:

$$\mathscr{L}_{\mathrm{GAN}}(G, D_k) = \mathbb{E}_{(\mathbf{s},\mathbf{x})}[\log D_k(\mathbf{s}, \mathbf{x})] + \mathbf{s}, \mathbf{x}\mathbb{E}_{\mathbf{s}}[\log (1 - D_k(\mathbf{s}, G(\mathbf{s}, \mathbf{t}))]\mathbf{s}, \mathbf{t}))]\mathbf{s}, G \tag{3}$$

Likewise, the discriminator uses feature matching loss to improve the GAN network loss. This loss makes training more stable. The feature matching loss $\mathscr{L}_{\mathrm{FM}}(G, D_k)$ is as follows:

$$\mathscr{L}_{\mathrm{FM}}(G, D_k) = \mathbb{E}_{(\mathbf{s},\mathbf{x})} \sum_{i=1}^{T} \frac{1}{N_i} \left[\| D_k^{(i)}(\mathbf{s}, \mathbf{x}) - \mathbf{s}, \mathbf{x} D_k^{(i)}(\mathbf{s}, G(\mathbf{s}, \mathbf{t}))\mathbf{s}, \mathbf{t}))\mathbf{s}, G \|_1 \right] \tag{4}$$

Where T is the total number of layers, and N_i is the number of elements in each layer.

2.2. Person image generation network under boundary and pose constraints

At present, the method of character generation based on the target pose is becoming mature. However, since it only has the target pose as a constraint, it is not completely suitable for our task. We improve the pose transfer model [4] by adding the person boundary as a necessary constraint. In this way, the generated characters not only have the target posture but also fit perfectly the boundary of the target human body, promoting the harmony of the composition of the group photo. The model's input is a single-person image, target pose, and target boundary. In pose transfer, the encoded character boundary is constrained to the feature stage before character decoding, and the decoded character image is obtained under the target boundary and pose constraints.

2.2.1. Posture transfer

As shown in Fig. 3, our pose transfer network improves the basic style transfer model [4] (inside the dashed box) by adding predicted character boundaries as guidance. Based on the strategy of decomposing human images into components to participate in pose transfer, the predictions in the first stage of the fusion network are fused before decoding. Character pose boundary constraints are used for pose transfer. In this way, the generated characters not only have the target posture but also perfectly fit the boundary of the target human body, promoting the harmonious synthesis of subsequent group photos.

2.2.2. Boundary constraints

Since we need to constrain the character feature map before decoding, the 2 - channel human body boundary map (512*512) is encoded into a 256-channel human body boundary map (128* 128) through 3 layers of convolution, and then constrain the 256-channel human body feature map (128*128) on each channel. The human body features and boundary features on each channel are constrained by element-by-element point multiplication, as shown in formula (5):

$$H_m^i = H_i \odot M_i \tag{5}$$

where $\odot$ represents element-wise multiplication, H_i and M_i represents the human body feature matrix and boundary feature matrix of each channel, respectively, to obtain the human body features under the boundary constraints in each channel H_m^i.

2.2.3. Character synthesis

After obtaining the target person image after the target boundary and pose constraints, we need to synthesize it into the group photo image according to the predicted target person boundary in the first stage. To assemble the final output, we use a Gaussian blur weight mask for alpha (transparency) blending, as shown in formula(6):

$$M_P = g(\sigma) * \mathbf{1}_{B_P} \tag{6}$$

where * denotes a 2D convolution operator, g(σ) is a discrete 2D Gaussian kernel, and $\mathbf{1}_{B_p}$ is an indicator function equal to 1 if the corresponding pixel belongs to the human body bounding box BP and is equal to 0 in other positions. Finally, the results are combined into formula(7) to get the final result:

$$\widetilde{G}_t = (\mathbf{1} - M_P)\widetilde{I}_m + M_p\widetilde{I}_h \tag{7}$$

Where, $\widetilde{I}_m$ represents a group photo image of multiple people, $\widetilde{I}_h$ represents a target person image, and $\widetilde{G}_t$ represents a combined group

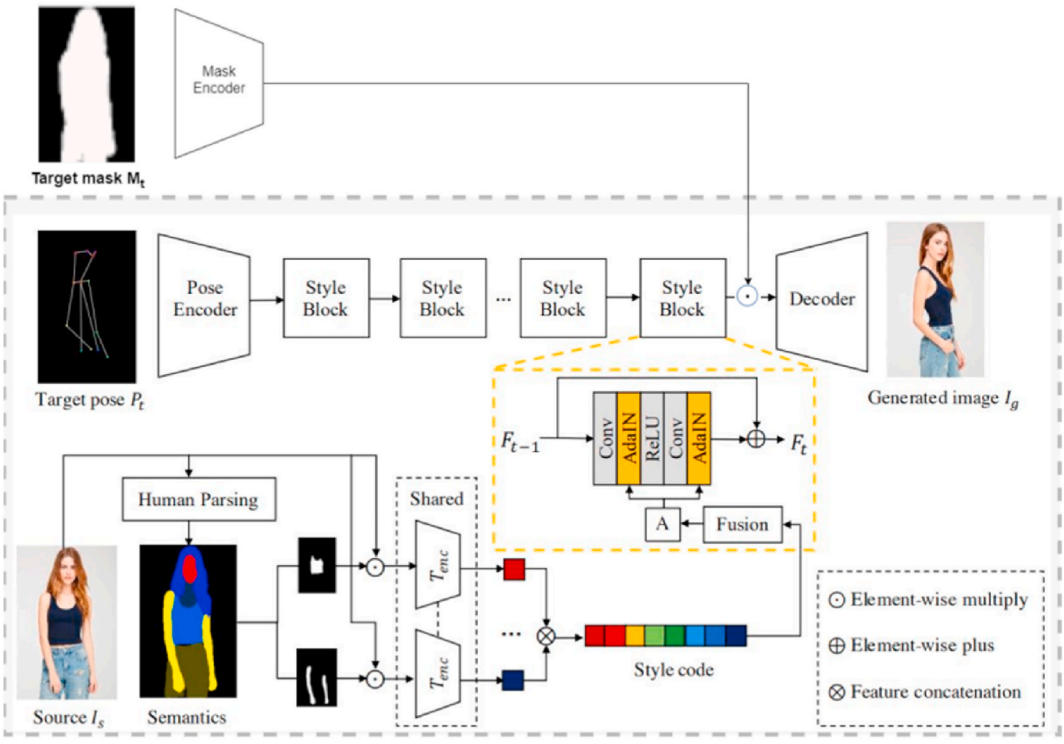


Fig. 3. The improved ADGAN pose transfer model.

photo image.

2.2.4. Training and loss

The discriminator and loss in this section follow the conventional configuration in the pose transfer task [4]. Two discriminators, D_s and D_P, ensure that the generated character image H_g is consistent with the appearance and texture of the source character image H_s and with the pose P_t and boundary M_t of the target character H_t.

Loss functions include conventional adversarial loss function, L1 loss function, perceptual loss function [12], and contextual loss function. Compared with the pose transfer task [4], we add the predicted character boundary to the adversarial loss function and reconstruction loss function as conditional guidance for generation. The overall loss function is shown in formula (8). The adversarial loss function L_{adv} is shown in formula (9). The reconstruction loss function L_{rec} is shown in formula (10). The perceptual loss function [12] $\mathscr{L}_{per}$ is shown in formula (11). The context loss function $\mathscr{L}_{CX}$ is shown in formula (12).

$$\mathscr{L}_{total} = \mathscr{L}_{adv} + \lambda_{rec}\mathscr{L}_{rec} + \lambda_{\text{per}}\mathscr{L}_{per} + \lambda_{CX}\mathscr{L}_{CX} \tag{8}$$

Where λ_{rec}、λ_{per}、λ_{CX} are the weights of the corresponding losses, respectively.

$$\begin{aligned} L_{adv} = {} & \mathbb{E}_{H_s,P_t,M_t,H_t}\left[\log\left(D_s(H_s,H_t)\cdot D_p(P_t,M_t,H_t)\right)\right] + \\ & \mathbb{E}_{H_s,P_t,M_t}\left[\log\left((1-D_s(H_s,G(H_s,P_t,M_t)))\right.\right. \\ & \left.\left.\cdot\left(1-D_p(H_t,G(H_s,P_t,M_t))\right)\right)\right] \end{aligned} \tag{9}$$

$$L_{rec} = \| G(H_s,P_t,M_t) - H_t \|_1 \tag{10}$$

$$\mathscr{L}_{per} = \frac{1}{W_l H_l C_l}\sum_{x=1}^{W_l}\sum_{y=1}^{H_l}\sum_{z=1}^{C_l} \| \varphi_l(H_g)_{x,y,z} - \varphi_l(H_t)_{x,y,z} \|_1 \tag{11}$$

$$\mathscr{L}_{CX} = -\log\left(CX\left(\mathscr{F}^l(H_g),\mathscr{F}^l(H_t)\right)\right) \tag{12}$$

Among them, H_s represents the source character image. H_g represents the generation image. P_t represents the target pose. M_t represents the predicted human boundary. H_t represents the real image with the target pose. $\mathscr{F}^l(H_g)$和$\mathscr{F}^l(H_t)$ respectively represent the features extracted from the pre-trained VGG 19 network and then use CX to measure the similarity between features.

3. Experiments and discussions

3.1. Datasets, evaluation criteria, and experimental environment

Dataset: The position, pose, and boundary prediction network is trained based on the multi-person group photo dataset (MHP-V1) [9], which contains a wealth of multi-person group photo images and semantic labels of each person. Based on it, we merge the semantic tags of the character, and the person's pose key points are extracted using the pose estimation method [10]. Since our model aims to predict new people from the context of people in group photos, we remove the people in each group photo one by one and keep the remaining people to construct our dataset. The dataset contains about 5000 group photo images; each image contains three people on average, and the total contains about 15,000 people. We use these 15,000 single people as the missing people in the group photo to construct the data set, of which 12,000 image is used as training samples and 3000 images are used as testing samples. The network is trained for 200 epochs, and the batch size is 12.

The person image generation network is based on the In-shop clothes retrieval benchmark DeepFashion [13] dataset, which contains 50,000 images of people with rich appearances and poses. Firstly, we use the human body semantic analysis method [14] to semantically segment the human body, segment it into eight components (background, hair, face, left arm, right arm, upper body, left leg, right leg), and use it as the input of human body texture component encoding. We use the person matting method [15] for the corresponding human body boundary. Finally, we use the pose estimation method [10] to extract the person's pose key points. Following the same data configuration in the pose transfer method [4], we randomly select 101966 pairs of images for training and 8750 for testing.

Evaluation Metrics: In our overall experiments, Inception Score (IS) [16] and Structural Similarity (SSIM) [17], which are commonly used in pose transfer tasks, are used to evaluate the authenticity and diversity of generated person images as well as the similarity to real images. Furthermore, the perceptual metrics Fresche Inception Distance (FID) [18] and Learned Perceptual Image Patch Similarity (LPIPS) [19] metrics are used to evaluate the authenticity and consistency of generated images. The synthesis result of the overall group photo is displayed through image comparison to achieve the desired effect.

Experimental details: Our two-stage model is trained separately on different datasets using supervised approaches. Our methods are implemented based on Python and PyTorch framework, using Tesla V100 and Tesla K80 GPUs for training tasks. The boundaries of the people in the group photo are merged from the human body analysis labels of each person provided by the MHP dataset, and the key points of the people in the group photo are extracted through the OpenPose method [10]. Adam optimizer is used in our experiment; the initial learning rates of the generators of the two parts of the network are 0.0002 and 0.0001, respectively. The weights of the first half λ are all set to 10, and the initial weights of the second half (λ_{rec}, λ_{per} , λ_{cx}) are set to (2.0, 1.0, 1.0).

3.2. Boundary and pose prediction results

3.2.1. Boundary prediction

Fig. 4 shows the results of perceiving the human body boundaries of new people with different people in the group photo. The first column is the input context of the group photo, the second column is our predicted boundary of the group photo, and the third column is the boundary of the real group photo, and the target person is displayed in orange. The boundary of the target person is displayed in orange. It can be seen from the figure that in the group photo scene with different numbers of people, the human boundary and position of the new character can be reasonably predicted to a certain extent. It can maintain interaction with the original characters in the group photo, maintaining harmony as a whole, and at the same time, the body shape and height of the person are consistent with the real person, which greatly improves the authenticity of predicting the person's boundary.

In order to further prove the effectiveness of our model in predicting the boundary of people, we calculated the histogram

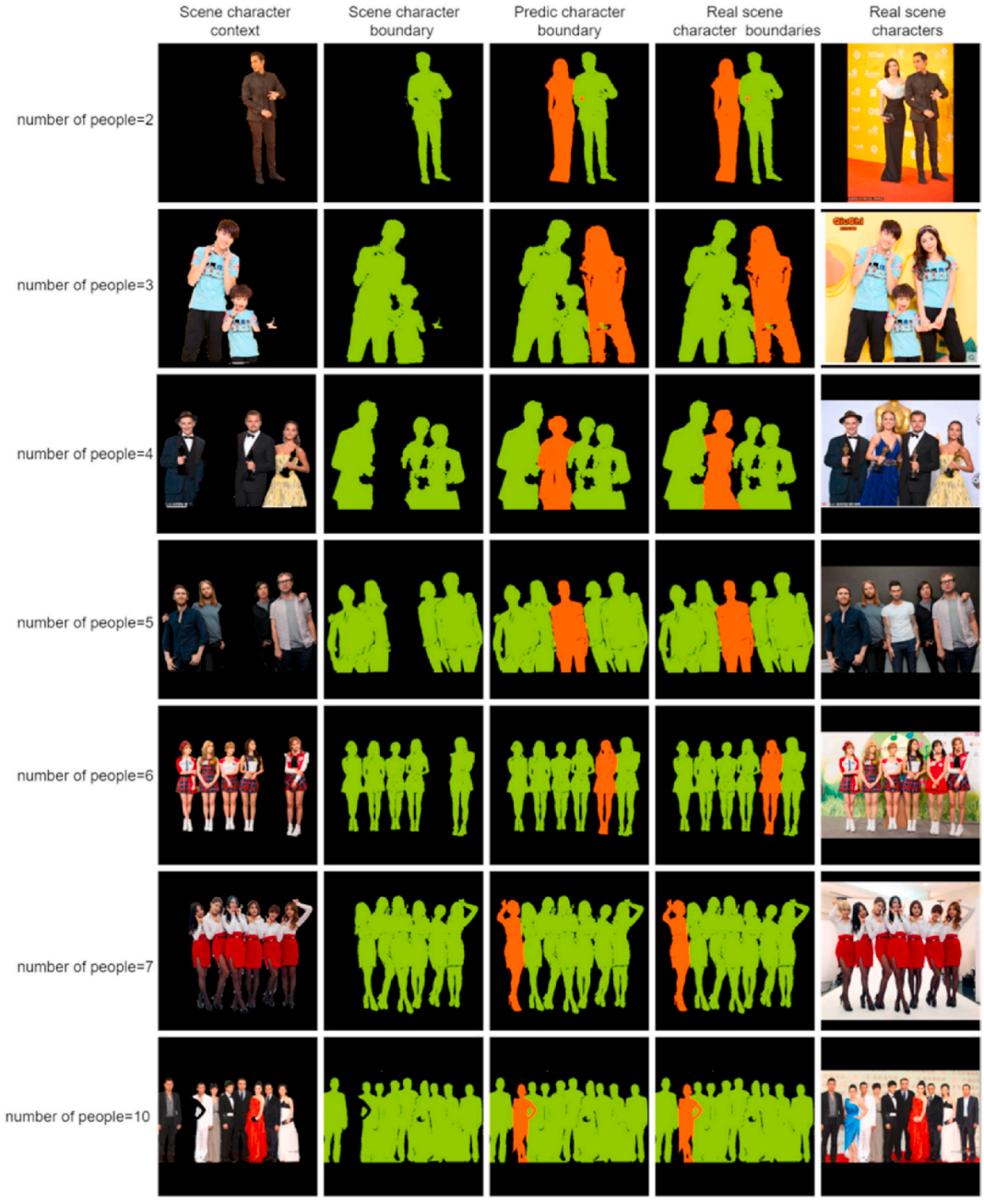


Fig. 4. Boundary prediction results under different photo groups.

coincidence degree between the predicted boundary of the person in the group photo and the boundary image of the real person in the group photo, expressed as the similarity score, to show the gap between our predicted person boundaries and real person boundaries. In addition, to prove the accuracy of the predicted person's position, as shown in Fig. 5, we restricted the predicted target person's boundary and the real target person's boundary with a rectangular frame. Then, we calculated the intersection and union ratio of the two rectangular frames, expressed as an IOU metric. As shown in Table 1, both the Similarly and IOU indicators are at the upper-middle level, and the IOU indicator is relatively high. It shows that our model accurately predicts the direction of the character's position. However, the pose prediction is uncertain, and the similarity with the real boundary is slightly lower, but it does not affect the harmony of the overall group photo.

3.2.2. Posture prediction

Fig. 6 shows the results of perceiving the human body posture of the new person under different group photos. The first column is the input context of the group photo, the second column is our predicted pose, and the third column is the real group photo character pose. It can be seen that the posture and position of the new character can be reasonably predicted to a certain extent. It can maintain interaction with the original characters in the group photo, maintaining harmony. At the same time, the body shape and height of the person are consistent with the real person, which greatly improves the authenticity of the predicted person.

In order to further prove the effectiveness of our model in predicting the poses of people, we also calculate the histogram coincidence degree between the predicted poses of people in a group photo and the real poses of people in a group photo, expressed as a similarity score Similarly, to show the gap between our predicted person boundary and the real person boundary. In addition, in order to prove the accuracy of the predicted person's pose position, we restricted the predicted target person's pose and the real target person's pose with a rectangular frame in a similar way and then calculated the intersection and union ratio of the two rectangular frames, expressed as the IOU index. As shown in Table 2, both the Similarly and IOU indicators are at the upper-middle level, and the IOU indicator is relatively high. It shows that our model is accurate in predicting characters' positions. However, due to the interaction complexity of multi-person poses, the applicability of pose prediction with many people is poor.

3.2.3. Boundary and posture Synchronization prediction

The above experiments demonstrate the effectiveness of our model in predicting character boundaries and poses based on the context of the group photo scene. As shown in Fig. 7, our model can not only reasonably predict the boundaries and poses of new characters, but our model can also use the same model to predict the boundaries and poses of new characters at the same time so that the predicted characters' boundaries and poses are morphologically as close as possible be consistent. At the same time, we further demonstrate the boundary and pose prediction of missing persons in different positions (such as left, center, and right) in a group photo scene with the same number of people to demonstrate the applicability of our model.

3.3. Person image generation under boundary and pose constraints

Fig. 8 shows the comparative results of person image generation under the object boundary and pose constraints in the pose transfer task. Compared with other pose transfer results, it can be seen that our results are consistent with other advanced results in terms of realism. However, our characters are generated under boundary constraints, which is more suitable for our task.

In order to further demonstrate the effect of our character generation, we selected IS, SSIM, FID, and LPIPS to conduct comparative experiments, as shown in Table 3. Among them, the higher the IS and SSIM indicators, the higher the authenticity and diversity of the generated person images; the lower the FID and LPIPS indicators, the higher the consistency and authenticity of the images. Our metric outperforms all other methods on SSIM and LPIPS, which are 0.058 and 0.0956 higher than the best results, respectively. However, it is worse than the optimal result regarding IS and FID indicators, which are 0.04 and 1.325 lower, respectively.

In addition, the pose transfer model guided by the human body boundary can adjust the human body generation under different

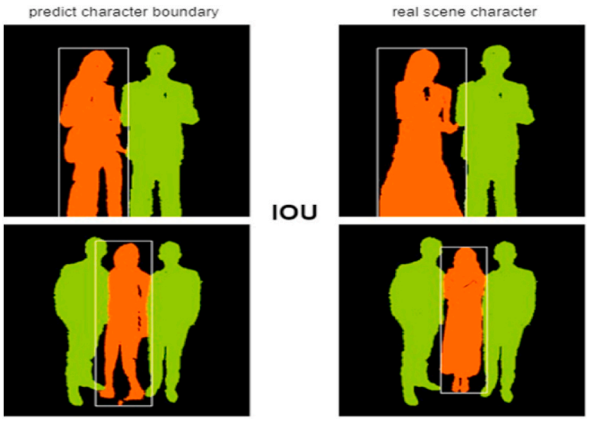


Fig. 5. The example of IOU.

Table 1
Indexes of human body boundary prediction under different photo groups.

Number of people taking photos	Similarly↑	IOU↑
number of people = 2	0.8014	0.8608
number of people = 3	0.7178	0.8225
number of people = 4	0.7379	0.8229
number of people = 5	0.7407	0.8123
number of people = 6	0.7850	0.8308
number of people = 7	0.7613	0.8014
number of people = 10	0.7854	0.8421
Average	0.7614	0.8275

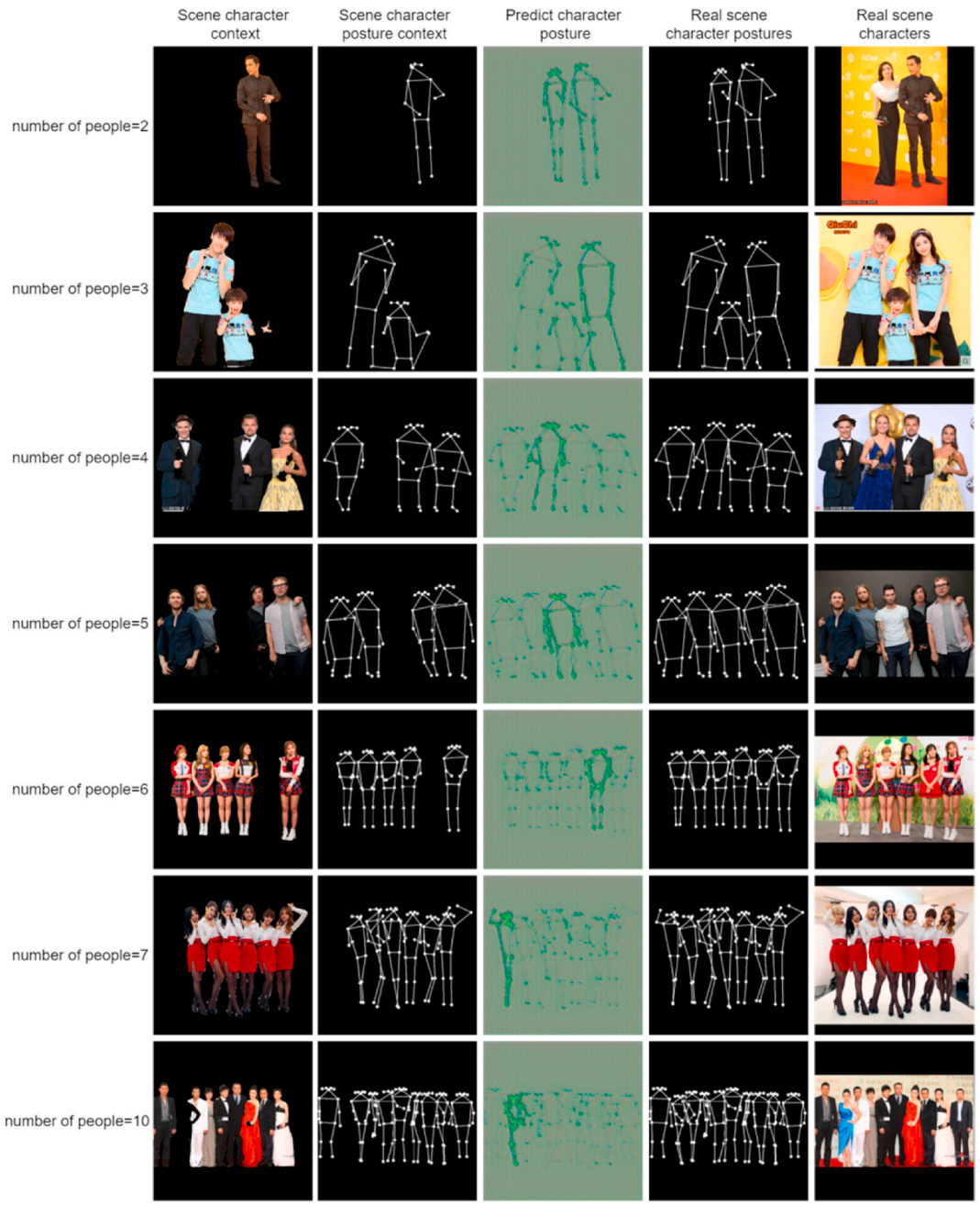


Fig. 6. The posture prediction results under different photo groups.

Table 2
Indexes of human body boundary prediction under different photo groups.

Number of people taking photos	Similarly↑	IOU↑
number of people = 2	0.9015	0.9341
number of people = 3	0.8245	0.9245
number of people = 4	0.8374	0.9027
number of people = 5	0.8415	0.9274
number of people = 6	0.8752	0.9358
number of people = 8	0.8351	0.9142
number of people = 10	0.8088	0.9036
Average	0.8463	0.9203

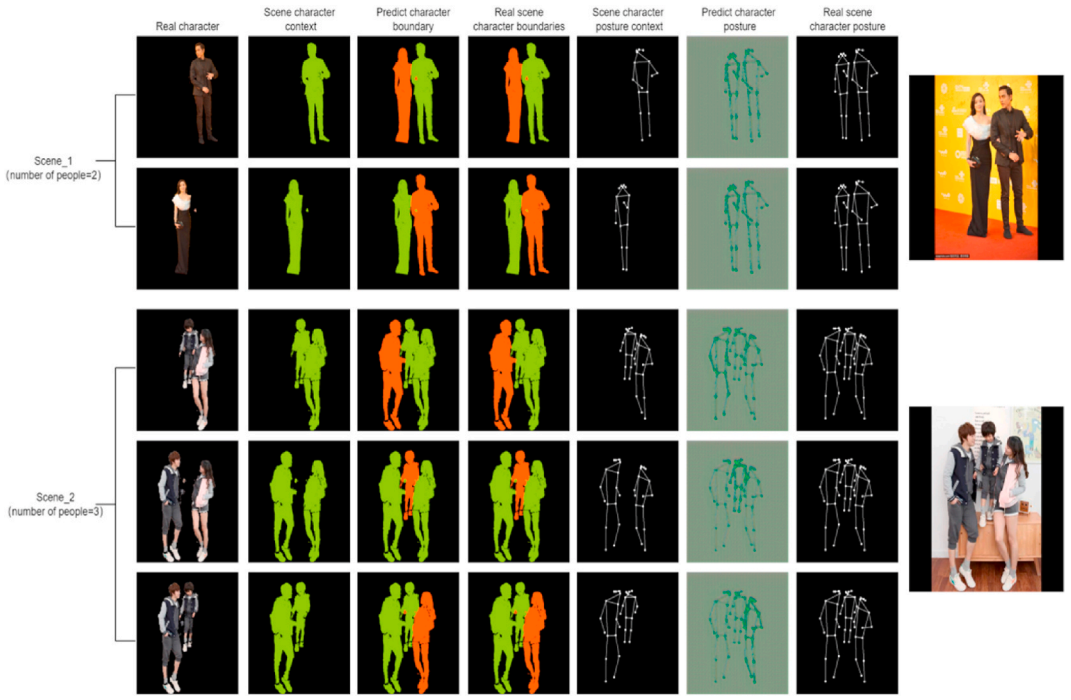


Fig. 7. The boundary and pose prediction results with different numbers of people.

human body boundaries to a certain extent. As shown in Fig. 9, we show the generated person images constrained by offsetting the human body boundary. It turns out that although our model can generate character images under different boundary constraints, the quality of the generated character images has a large room for improvement.

3.4. Character synthesis

Fig. 10 shows the results of perfectly merging the generated target person image into the group photo image according to the predicted boundaries. It can be seen that the fused target person can maintain good interaction with other characters in the group photo while maintaining harmony and consistency in the image as a whole.

4. Conclusions and future work

The research on predicting the target person based on the person's context in the group photo only predicts the semantics or boundary of the target person and does not consider the user's attributes. We propose a new model for perceiving person images by combining the context of the people in the group photo and the user attributes. Our model can not only predict the pose and boundary of the target person simultaneously but also consider the user's attributes, which greatly improves the authenticity and harmony of the composite image of the group photo. The experimental results show that the person images predicted by our model based on the group photo not only have good interaction with the boundary and pose of the person in the group photo but also can greatly improve the authenticity of the target person in the composite image of the group photo. Our method only considers the user's attributes and does not further divide them. Therefore, in future work, we hope to consider further the user's attributes and the social relationship in the

Fig. 8. Qualitative comparison with other state-of-the-art pose transfer methods.

Table 3
Quantitative comparison with other methods on the DeepFashion dataset.

Model	IS↑	SSIM↑	FID↓	LPIPS↓
PATN [1]	3.2	0.762	19.825	0.2538
ADGAN [5]	3.327	0.770	18.374	0.2379
MustGAN [4]	3.689	0.739	15.890	0.2408
GFLA[27]	3.634	0.784	14.058	0.2341
ours	3.649	0.842	15.383	0.1582

group photo and strive to improve the authenticity and harmony of the composite image of the group photo.

Data availability statement

We do not publish the data and models associated with the study. If required by the relevant researcher, we will provide it upon request.

Additional information

No additional information is available for this paper.

CRediT authorship contribution statement

Zhenquan Shi: Conceptualization, Methodology, Visualization, Writing – original draft. **Hongming Lu:** Conceptualization, Methodology, Software, Writing – original draft. **Xinyue Sun:** Software, Validation, Visualization, Writing – review & editing. **Jing Guo:** Data curation, Investigation, Resources. **Yongzhen Ke:** Conceptualization, Formal analysis, Methodology, Project administration, Writing – review & editing. **Kai Wang:** Investigation, Validation. **Fan Qin:** Data curation, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to

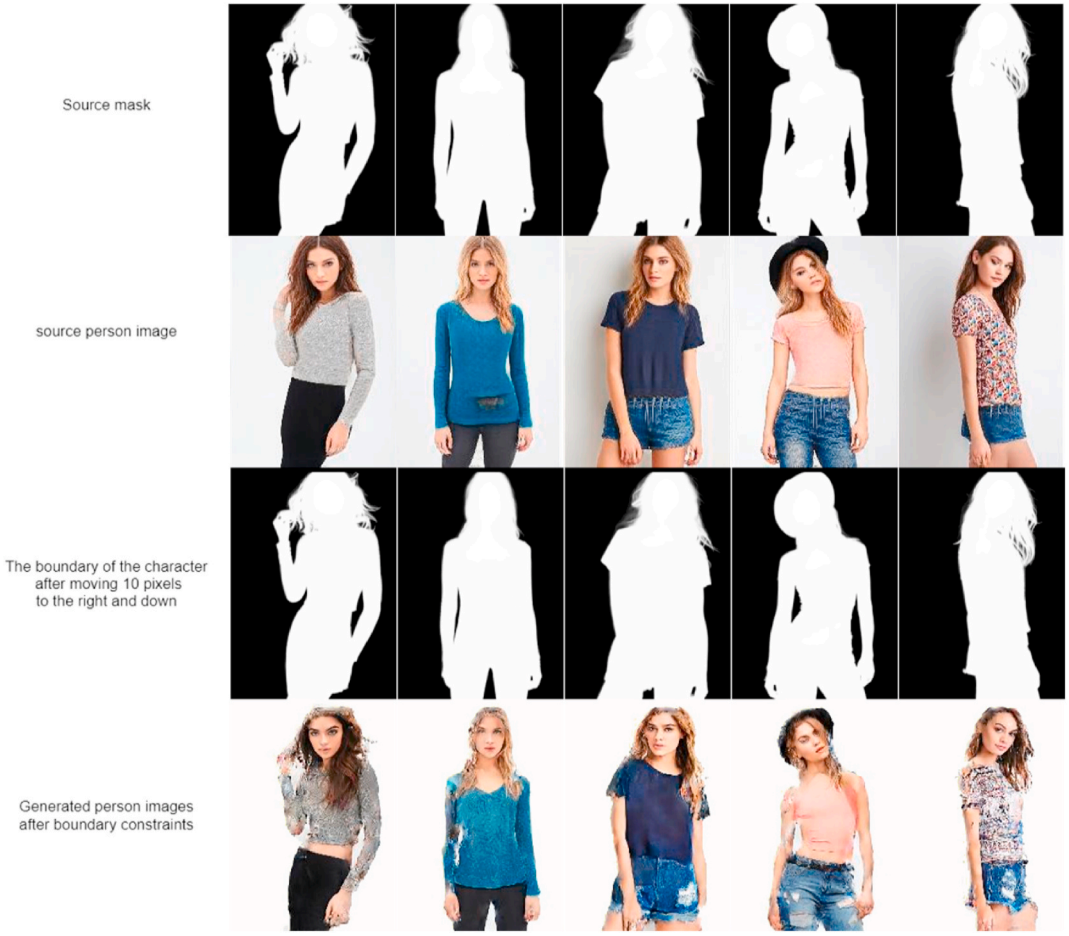


Fig. 9. Person image generation under boundary constraints.

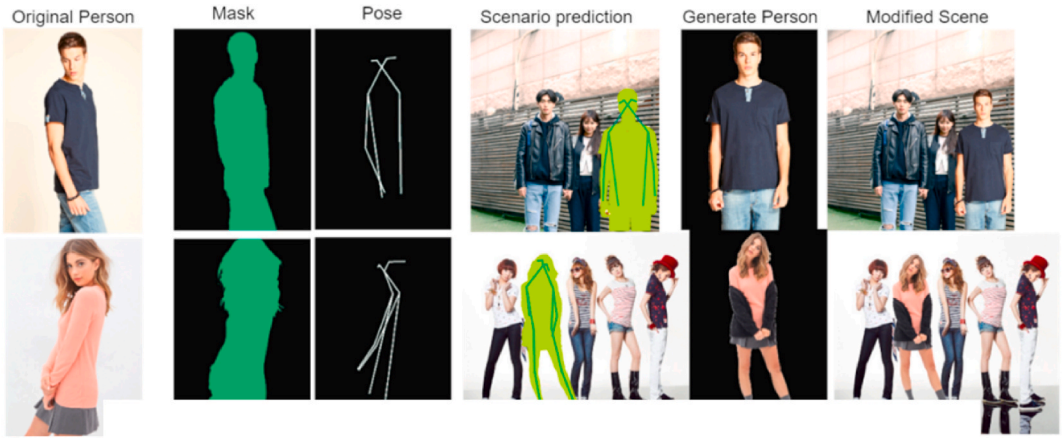


Fig. 10. Character synthesis according to the predicted boundary.

influence the work reported in this paper.

References

[1] P. Roy, S. Ghosh, S. Bhattacharya, et al., Scene Aware Person Image Generation through Global Contextual Conditioning[J], 2022 arXiv preprint arXiv: 2206.02717.
[2] O. Gafni, L. Wolf, Wish you were here: context-aware human generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020 [C].
[3] T.C. Wang, M.Y. Liu, J.Y. Zhu, et al., High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs[J], 2017.
[4] Y. Men, Y. Mao, Y. Jiang, et al., Controllable Person Image Synthesis with Attribute-Decomposed GAN[J], 2020.
[5] Y.S. Rawat, M. Song, M.S. Kankanhalli, A spring-electric graph model for socialized group photography, IEEE Trans. Multimed. 20 (3) (2017) 754–766.
[6] M. Zanfir, E. Oneata, A. Popa, et al., Human Synthesis and Scene Compositing: Proceedings of the AAAI Conference on Artificial Intelligence, 2020 [C].
[7] S. Weng, W. Li, D. Li, et al., Misc: multi-condition injection and spatially-adaptive compositing for conditional person image synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020 [C].
[8] Y. Wang, S. Hou, B. Ning, et al., Photo stand-out: photography with virtual character. Proceedings of the 28th ACM International Conference on Multimedia, 2020 [C].
[9] J. Li, J. Zhao, Y. Wei, et al., Multiple-human parsing in the wild[J]. arXiv preprint arXiv:1705.07206 (2017).
[10] C. Zhe, T. Simon, S.E. Wei, et al., Real-time multi-person 2D pose estimation using Part Affinity fields[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR), 2017.
[11] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition[C], Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 770–778.
[12] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual Losses for Real-Time Style Transfer and Super-Resolution[J], Springer, Cham, 2016.
[13] Z. Liu, P. Luo, S. Qiu, et al., DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations: Computer Vision & Pattern Recognition, 2016 [C].
[14] G. Ke, X. Liang, D. Zhang, et al., Look into person: Self-supervised StructureSensitive learning and a new benchmark for human parsing[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR), 2017.
[15] Z. Ke, K. Li, Y. Zhou, et al., Is a green screen really necessary for real-time person matting? arXiv preprint arXiv:2011.11961 (2020).
[16] T. Salimans, I. Goodfellow, W. Zarem Ba, et al., Improved Techniques for Training GANs[J], 2016.
[17] W. Zhou, A.C. Bovik, H.R. Sheikh, et al., Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004).
[18] M. Heusel, H. Ramsauer, T. Unterthiner, et al., GANs trained by a two TimeScale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).
[19] R. Zhang, P. Isola, A.A. Efros, et al., The unreasonable effectiveness of deep features as a perceptual metric[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR), 2018.