GroupDiff: Diffusion-based Group Portrait Editing

Yuming Jiang[1*], Nanxuan Zhao[2✉], Qing Liu[2], Krishna Kumar Singh[2], Shuai Yang[3], Chen Change Loy[1], and Ziwei Liu[1]

[1] College of Computing and Data Science, Nanyang Technological University
[2] Adobe Research
[3] Wangxuan Institute of Computer Technology, Peking University
https://github.com/yumingj/GroupDiff

Abstract. Group portrait editing is highly desirable since users constantly want to add a person, delete a person, or manipulate existing persons. It is also challenging due to the intricate dynamics of human interactions and the diverse gestures. In this work, we present **GroupDiff**, a pioneering effort to tackle group photo editing with three dedicated contributions: **1) Data Engine:** Since there are no labeled data for group photo editing, we create a data engine to generate paired data for training. The training data engine covers the diverse needs of group portrait editing. **2) Appearance Preservation:** To keep the appearance consistent after editing, we inject the images of persons from the group photo into the attention modules and employ skeletons to provide intra-person guidance. **3) Control Flexibility:** Bounding boxes indicating the locations of each person are used to reweight the attention matrix so that the features of each person can be injected into the correct places. This inter-person guidance provides flexible manners for manipulation. Extensive experiments demonstrate that GroupDiff exhibits state-of-the-art performance compared to existing methods. GroupDiff offers controllability for editing and maintains the fidelity of the original photos.

Keywords: Group Photo Editing · Diffusion Models

1 Introduction

Have you ever been bothered by the challenge of capturing the perfect group photo, especially during significant occasions like team-building events, family reunions, or friend gatherings? Group portrait photos become a crucial means of preserving cherished memories. Imagine this scenario, an individual must unexpectedly leave an important reunion party before the group photo is taken due to an emergency. How valuable would it be to have a tool that seamlessly

*This work was done when Yuming Jiang was a research intern at Adobe Research.
✉Corresponding Author.

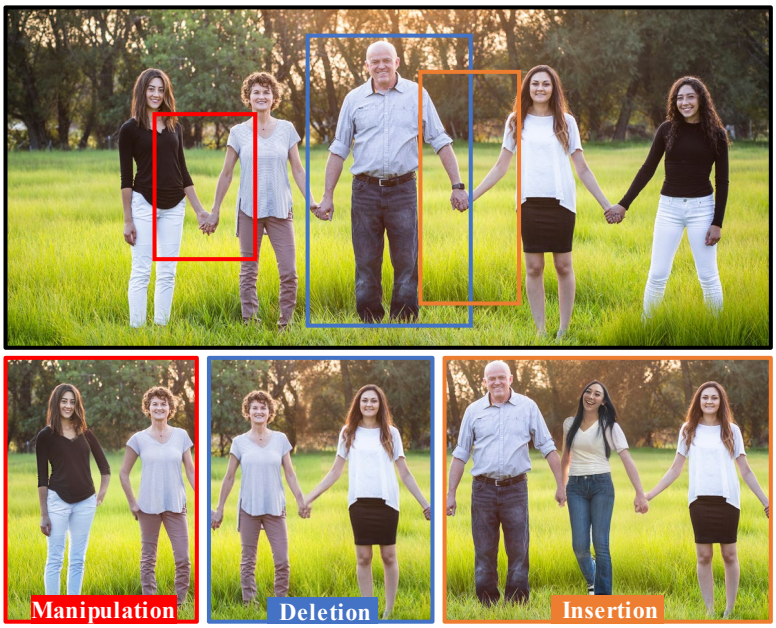

Fig. 1: Applications enabled by our GroupDiff. Given a group photo, we can (a) **manipulate the existing person**, (b) **remove a person**, and (c) **insert a person**.

adds the missing person to the photo later? Group portrait editing addresses precisely this need and more. However, as a broad topic covering various operations, group portrait editing is not an easy task with many factors that need to be considered, such as human identity, interaction, and diverse gestures. It is a difficult and tedious process even for experts with design experience and knowledge.

However, the manipulation of group photos is challenging for two reasons: 1) the generation of the single-person image is difficult because of the complicated human structures, and 2) the generation of human interaction regions. How to generate individual persons and their interactions naturally is an unsolved problem. In this work, we take the initial step to tackle this hard problem. More specifically, we target the unique problem in group portrait editing, *i.e.*, the synthesis of human interactions when we perform person insertion, person removal, and person manipulation, as shown in Fig. 1. As personal features like facial expressions and pose orientation can be adjusted by single human editing methods [3, 31, 32, 34–36, 46, 47], our work builds on the assumption that both the initial group portrait photo and the inserted person are with satisfied facial expressions and facing orientations.

To facilitate a more general setting, we aim to develop a unified framework handling three main group photo editing tasks (*i.e.*, manipulation of existing per-

son, person removal, and person insertion). We propose to perform group photo editing by generating reasonable interaction regions. Given an initial group portrait photo with masked regions indicating the intersection areas to be edited, and reference image(s) of the nearby person(s), we learn to seamlessly hallucinate the human interaction in a natural way while maintaining the identities of individuals. With the well-established generation power of diffusion models [44, 50], we choose to build our method on top of the pre-trained diffusion model [44].

However, our task still poses several unique challenges. Firstly, collecting a large set of paired training data with before-after editing is infeasible. While several works [25, 53] take the synthetic way to generate the data, they mainly work on single-person editing without considering complex human compositions and interactions. How to generate the training data fitting the natural distribution of group portrait editing is a difficult problem. Secondly, maintaining the appearance of individuals during the editing process is a formidable challenge. Users do not want the appearance to be changed, such as clothing be changed after editing. Besides, group photos often include multiple persons, and how to correctly map the reference appearance to the target person is a remaining problem. Lastly, providing users with the flexibility to specify the desired interactions according to their preferences is essential.

To this end, we propose GroupDiff, a new approach for tackling group portrait editing tasks. For the issue of lacking large-scale before-after editing data, we propose a training data generation engine that mimics the editing requirements encountered in real-world applications of group photo editing. Many factors have been considered, such as the interaction ways, the body parts' appearances, and the poses. To maintain the appearance of the persons including the inserted person and the persons in the original photo, we propose an appearance preservation diffusion model which contains person-aware modules customized on attention layers. By taking the person(s) that the user wants to be kept as reference image(s), this model takes both inter-person and intra-person features when injecting the references into the editing process. This module can tackle the challenge posed by multiple persons through an indicator matrix, indicating the locations of each person and thus precise attention guidance. Additionally, we take the skeleton as an intermediate representation throughout our framework design, ensuring flexibility and controllability for users. Experiments show that our proposed GroupDiff achieves state-of-the-art performance in the task of group portrait editing. Our contributions are summarized as follows:

- We work on a challenging task of Group Portrait Editing and make the first effort to solve it in a unified framework by generating interaction regions.
- We propose a novel framework (**GroupDiff**) that includes person-aware modules specifically designed to maintain the appearance of reference persons, considering both inter-person and intra-person features.
- We introduce a simple yet effective data synthesis method to get paired data for GroupDiff training.
- We demonstrate the flexibility of our model on various applications such as person insertion, person removal, and interaction editing.

2 Related Work

Single-Human Editing. There are two major directions for single-human editing: facial editing and full-body editing. For facial editing, the target is to manipulate the facial attributes, *e.g.*, age manipulation [57, 62], hair synthesis [41, 60] and expression manipulation [23, 26, 27, 46, 47, 56]. As for full-body editing, it mainly includes two directions: pose transfer and virtual try-on. The objective of pose transfer [3, 31, 32, 34–36] is to seamlessly transpose the visual characteristics of an individual from one pose to another. Albahar *et al*. [1] introduces a pose-conditioned StyleGAN framework in which the source image is warped to align with the target pose. The warping is operated according to the correspondence between the source pose and the target pose. The warped image is then employed to modulate the features. Try-on tasks aim to transfer the clothing of the person from one image to the other image. Traditional Try-on methods [5, 11, 19, 59] mainly focus on warping the clothing to the target person. Baldrati *et al*. [4] and Zhu *et al*. [66] propose to address the problem under the setting of diffusion models. Kulal *et al*. [25] study the single human editing task under the human-environment interactive setting. They propose a method to insert a person into the place. Different from previous works, our paper focuses on group portrait editing, where the interactions and compositions among persons need to be carefully handled.

Inpainting. Image inpainting is an ill-posed problem, which aims to fill in the masked regions using meaningful content. Patch-based methods [8, 13] perform the inpainting by finding the nearest matched patch and then placing them into the holes in a proper way. GAN-based methods [20, 29, 30, 38, 43] often use the high-level information extracted by the networks and then use the extracted features to guide the inpainting. Recently, since the emergence of newer generative models, *e.g*. diffusion models and VQGAN models [9, 54], researchers propose to solve the task using languages [39, 55, 58]. RealFill [53] learns LoRA weights [18] to inpaint the background regions from reference images. However, these methods are not designed for human-specific inpainting tasks, and they do not consider the specific human priors, such as human poses, and do not have modules specifically for human appearance preservation.

Diffusion Models. Diffusion models have demonstrated the capability of generating images [15, 49, 51]. Recently, since the emergence of Stable Diffusion [44], it has dominated the field of image generation, especially for text-to-image generation [2, 67]. Apart from texts, ControlNet [64] proposes to introduce multi-modal controls, such as skeleton, sketch, depth maps etc. Stable Diffusion is also applied to some image editing tasks [23, 26, 27]. Some editing methods [14, 37] achieve image editing by modifying the cross-attention maps. Diffusion models have been also widely applied in multiple tasks including video generation [12, 17, 48], image enhancement [16, 45] and image composition [6, 33, 52, 63] To take advantage of the generation power of diffusion models, our work also builds on top of this learning framework.

3 Preliminaries

Stable Diffusion is based on diffusion models, a kind of probabilistic model. Diffusion models generate images by iterative denoising from a noise map, which is normally sampled from the Gaussian distribution. Given a Gaussian noise map $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the reverse diffusion process for sampling is expressed as follows:

$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{1}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \coloneqq \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)), \tag{2}$$

where θ is often parameterized using a trained network, and t is the denoising timestep.

Traditional diffusion model [15] operates on the pixel space. Due to the high computational cost of generating high-resolution images, the generative capability of the diffusion model is limited to low-resolution images. Stable Diffusion [44] proposes to use the VAE to project the image from pixel space to latent space. With the downscaled latent representations, the model can synthesize high-resolution images. Also, Stable Diffusion introduces texts as a condition and supports text-to-image generation. In Stable Diffusion, θ is parameterized as a UNet with multiple self-attention modules and cross-attention modules. In each block, a self-attention module is followed by a cross-attention module. In the cross-attention module, the features f are updated as follows:

$$f' = \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \tag{3}$$

$$K = \phi_K(f_t), V = \phi_V(f_t), Q = \phi_Q(f), \tag{4}$$

where f_t is the text embedding, ϕ_K, ϕ_V, and ϕ_Q are the linear layers to project the features into keys, values, and queries respectively.

4 GroupDiff

In this paper, we propose a unified framework, named GroupDiff, for group portrait editing. We formulate the problem as a conditional inpainting task. As shown in Fig. 2, when inserting a person, we need to inpaint the interaction regions and the missing human parts. As for removing a person, if the original person shakes their hands, we also need to inpaint the interaction region and the removed part. To better guide the synthesis of interaction regions and offer controllability to users, we introduce the skeleton as an additional condition. For person addition, with pose controls, we can explicitly manipulate the pose of the inserted person and the neighboring person, so that their interactions are natural. As for person removal, we also need to change the pose of the existing person to make it to be natural, if there are some human interactions between the removed

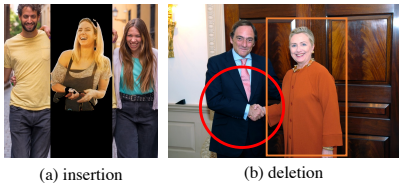

Fig. 2: Illustration of Common Editing Requests. (a) When we insert a person who only has a half-body picture, we need to adjust the interactions and make the lower part of her body complete. (b) When we are to remove a person from a group photo, we need to change the interactions and inpaint the removed region.

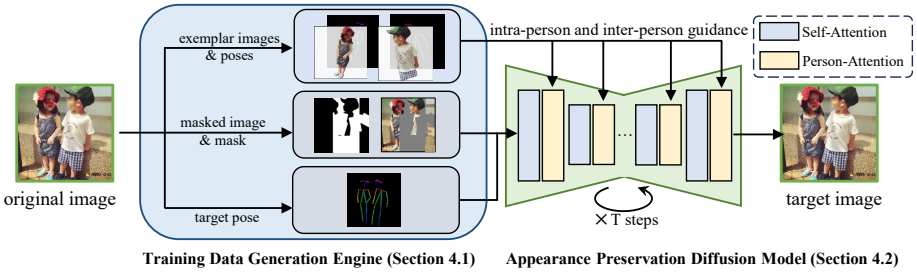

Fig. 3: Overview of GroupDiff. Starting from a group photo from the dataset, we first use the training data generation pipeline (Sec. 4.1) to generate paired data. Then the synthesized pair is fed into the Appearance Preservation Diffusion Model (Sec. 4.2), where inter-person and intra-person guidance are employed to preserve the identities.

person and existing persons. The overview of our proposed GroupDiff is shown in Fig. 3. We tackle the problem of group photo editing from two perspectives. From the data side, we propose a comprehensive training data generation engine to synthesize paired data, which will be introduced in Sec. 4.1. From the model side, we take the stable diffusion as the backbone [44]. Except for the noisy map, our model also takes an input image masked with the region to be filled, the corresponding mask, and the target skeleton map as inputs. These inputs are concatenated together before feeding them to the network. The masked region is filled with gray color. The skeleton map offers the user an option to specify the desired human interaction. As the masked region often covers the area that should remain the same after editing, such as the clothing and skin color, our model also takes the reference images of persons as conditions. Each reference person is given as a separate image. The reference images can be the inserted person or nearby persons around the interaction regions. We inject the reference images through intra-person and inter-person guidance to let the model learn to keep the appearance details after editing.

4.1 Training Data Generation Engine

The training of diffusion-based models requires large-scale paired data. However, there is no such dataset for group photo editing. Given the difficulties of con-

ducting group portrait photo editing, collecting a large amount of before-after editing data is infeasible. Therefore, finding a way to generate pairwise data to support the training is demanded. Since we pose our task as an inpainting task, we start from the existing group photo dataset [28, 40, 65]. Considering the real use cases, we design a comprehensive training data generation engine.

Person Interaction. Person interactions encompass a diverse set of gestures that form the essence of social dynamics including physical interactions, such as hugging, shaking hands, and raising arms, and many others. These free-form interactions increase difficulties in approximating the real distribution. Besides, different editing operations may incur different challenges. For example, when adding a person into the crowd, the surrounding context also should be completed, such as the area above the person's head. To enforce our model to better generate diverse interactions under different conditions, we use a hierarchical way for synthesizing paired data by randomly masking corresponding regions in group photo images. The coarser scale targets scenarios of large modification such as person insertion and removal, which requires inpainting a large region(Fig. 4). The finer scale targets scenarios of small modification such as interaction manipulation, which needs more fine-grained controls of masked regions(Fig. 5).

On a coarser scale, we first identify different persons through labeled bounding boxes. We assume that the interaction regions are often near the boundary of each bonding box. During training, we randomly select single or multiple bounding boxes and mask out the region near the boundary. Suppose the coordinate of the top-left point of the bounding box is (x_1, y_1) and the right-bottom point is (x_2, y_2). Then the width of the bounding box is $w = x_2 - x_1$. Then the left boundary region of the bounding box can be denoted as the rectangle with $(x_1 - r \cdot w, y_1)$ as the top-left corner and $(x_1 + r \cdot w, y_2)$ as the right-bottom corner. The right boundary of the bounding box can be obtained in a similar way. We simply treat the boundary regions as the interaction regions. When synthesizing data, we will randomly mask the left boundary, right boundary, or both boundaries. Empirically, we randomly sample r from $[0.1, 0.2]$. As shown in the second image in Fig. 4, we mask both boundary regions. Sometimes, we extend the mask outside the bounding box to cover the whole column of the boundary region as shown in the third image of Fig. 4. That is to set the top-left coordinate as $(x_1 - r \cdot w, 0)$ and the right-bottom coordinate as $(x_1 + r \cdot w, h)$, where h is the height of the image. To avoid facial regions being masked, we use human parsing to unmask the facial regions. To make the shape of the mask closer to the users' inputs, we also augment the shape of the mask into a brush-like one.

On a finer scale, we introduce the skeleton to control the person interaction, especially for the hand and arm positions. Involving the skeleton helps users to better specify the desired human interaction. The correctness of the skeleton is important to ensure the data generation quality, we use the state-of-the-art skeleton detection method [7] for extraction. A naive way of generating the mask is to segment out the arm and hand regions with a tight bounding box. However, this simple solution has a critical problem as the mask shape itself will leak the skeleton information. For example, in Fig. 5, if we strictly follow the skeleton, we

Fig. 4: Coarse Level Training Data Generation for Person Interaction. At the coarse level, we generate masks according to the bounding boxes of persons.

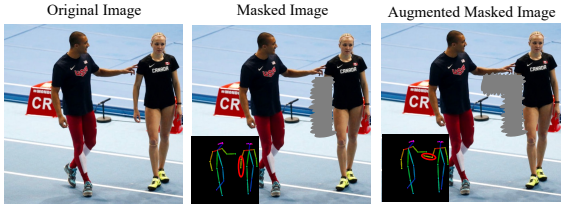

Fig. 5: Fine Level Training Data Generation for Person Interaction. At the fine level, we generate masks according to the skeleton and augmented skeleton.

can obtain the masked image as shown in the second column. In this case, the model may overlook the skeleton condition during training because the masked image itself has leaked the skeleton information. As a result, the obtained model fails to generate images conditioned on skeletons at inference time. Therefore, we adopt data augmentations to avoid this issue. Except for the tight bounding box mask around arms and hands, we randomly rotate the arms or hands in a different direction. As shown in the third column of Fig. 5, we randomly rotate the right arm and obtain the augmented skeleton. We mask both arm regions according to the original skeleton and augmented skeleton.

Body completion. In certain scenarios, our model also needs to complete the body part(s) of a person. For example, when there are occlusions among different persons, removing one person needs the body completion of the surrounding people. When inserting a person, if the reference image only contains a partial body, in this case, the model needs to learn to complete the rest parts. The data synthesis for the body completion is straightforward. We randomly mask out the lower body parts. Still, (x_1, y_1) represents the top-left point of the bounding box and (x_2, y_2) represents the right-bottom point. We randomly mask the region from $(x_1, y_1 + r \cdot (y_2 - y_1)$ to (x_2, y_2), where r is sampled from $[0.5, 0.9]$. Some examples are shown in Fig. 6.

4.2 Person-Aware Attention for Appearance Preservation

Our model needs to preserve the appearance of persons after editing. Since the content cannot be revealed if it is covered with masks during inference, we take the persons that users want to maintain the appearance out as reference images. To ensure details are encoded into the model accurately, inspired by previous works [6, 25], we choose to inject reference image(s) through the cross-attention module. The reference images during the training are obtained as follows: 1)

Fig. 6: Training Data Generation for Body Completion. We randomly mask the lower parts of the body so that the model has the capability of body completion.

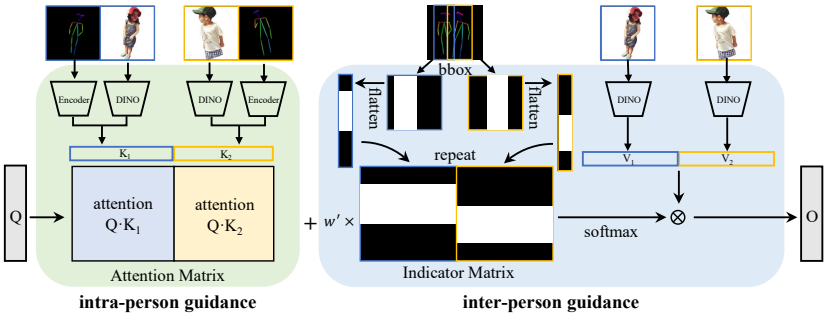

Fig. 7: Overview of Person-Aware Attention Module for Appearance Preservation. We employ intra-person guidance and inter-person guidance to better preserve the appearance. The intra-person guidance concatenates the pose features and image features as keys. The inter-person guidance is provided by using the information deduced from the positions of the persons. The indicator matrix is employed to assign higher weights to corresponding values in the attention matrix.

segment the person(s), and 2) fill the background with white. However, feeding all reference images through concatenation to the model like previous methods does not work in our case. This can cause a mismatch problem between the reference and target persons because the attention matrix is applied among query features and features of all reference images. The model will confuse which part to attend if multiple reference images are provided. We thus introduce a person-aware cross-attention mechanism for preserving content details.

Instead of sending raw pixels to the cross-attention module, we first encode each reference image I_i using DINOv2 [42] into a feature embedding $f_{I_i} \in \mathbb{R}^{256\times768}$, which has been validated to be effective for image generation and editing tasks [6]. If we follow the conventional cross-attention operation, after concatenating features together as $F = [f_{I_1}, f_{I_2}, ..., f_{I_N}]$, we obtain the value embedding $V \in \mathbb{R}^{(256\cdot N)\times d}$ and key embedding $K \in \mathbb{R}^{(256\cdot N)\times d}$ through linear layers individually to compute the output feature O as:

$$O = \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \tag{5}$$

$$K = \phi_K(F), V = \phi_V(F), \tag{6}$$

where d denotes feature dimension, and the query embedding $Q \in \mathbb{R}^{HW \times d}$ comes from the output of the last layer. H and W refer to the spatial resolution of features from the last layer (a.k.a self-attention layer). Based on this formulation, we introduce intra-person and inter-person guidances to solve our problem.

Intra-Person Guidance. For the intra-person guidance, it is essential to let the model know the correspondences between pixel appearances with body components for each reference image. Then the model can adaptively learn the mapping to restore accurate appearances. We introduce the skeleton pose P_{I_i} of exemplar images I_i as the intra-person indicators as shown in Fig. 7. Similarly, we first extract pose features p_{I_i} through a separately learned encoder, and compute the key embedding with an MLP after concatenating the image and pose features:

$$\hat{K} = \phi_K([K_1, K_2, ..., K_N]), \tag{7}$$

$$K_n = \text{MLP}([f_{I_n}, p_{I_n}]), \tag{8}$$

where $[\cdot]$ is the concatenation operation and $\text{MLP}(\cdot)$ is the layers to process the concatenated features. The attention matrix is obtained as follows:

$$M_{attn} = Q\hat{K}^T \in \mathbb{R}^{HW \times 256 \cdot N}. \tag{9}$$

It should be noted that we only introduce the pose information into keys and do not embed the pose information into values. The motivation lies in that pose information is an indicator of the body component and it should not be fused with the appearance information in keys.

Inter-Person Guidance. For inter-person guidance, as shown in Fig. 7, we introduce the indicator mask to specify the locations of each reference person. The indicator masks $\{m_i \in \mathbb{R}^{H \times W}\}$ are obtained by bounding boxes around each person. The bounding boxes cover the full body. For each reference image, we reshape the corresponding indicator mask m_i into a flattened tensor (whose shape is $\mathbb{R}^{HW}$) and repeat it into a matrix so that it has the shape of $\mathbb{R}^{HW \times 256}$. The indicator matrix $M_{ind} \in \mathbb{R}^{HW \times 256 \cdot N}$ is obtained by concatenating the matrices of all reference images. We add the indicator matrix to the attention matrix before applying the softmax operation. With this operation, when calculating the final output, the corresponding location will have higher weights to attend to the correct reference features. The operation is computed as:

$$O = softmax(\frac{M_{attn} + w' \cdot M_{ind}}{\sqrt{d}})V, \tag{10}$$

where the value embedding V is obtained from the DINO feature of reference images without involving pose information. In practice [2], we multiply the indicator matrix by a scale factor w', which is obtained as follows:

$$w' = w \cdot log(1 + \sigma) \cdot max(M_{attn}), \tag{11}$$

where w is the user-specified factor, and σ corresponds to the noise level at different diffusion steps.

5 Experiments

5.1 Implementation Details

Our method is implemented using PyTorch. The model is trained using eight NVIDIA A100 GPUs. We use the Adam optimizer to optimize the model. The beta is set as the default value. We set the learning rate as 10^{-4}. The learning rate schedular is LambdaLinearScheduler. The number of warm-up steps is set as 2500. The batch size is set as 4 per GPU, and the global batch size is 32. The model is initialized with the Stable-Diffusion v1.5 Inpainting model. The original Stable Diffusion Inpainting model has 9 channels as input. We concatenate the skeleton as an additional input, and thus the number of channels is 12. For the first 9 channels of the first layer, we inherit the weights of Stable Diffusion Inpainting model and initialize the remaining three channels with zero weights. The model is trained for 80 epochs. The resolution of the image we use for training is 512×512. We use the LV-MHP-v2 dataset [28, 40, 65], which is composed of 25, 403 group photo images. With each image, there are human parsing annotations. For human parsing, we directly use the annotations provided in the dataset. We use MMPose [7] to predict the skeleton. When editing group photos at inference time, we adjust the lighting using the off-the-shelf model [24].

5.2 Comparison Methods

SD Inpainting [44]. Stable Diffusion has provided the pretrained weights for inpainting. Here, we use the Stable Diffusion Inpainting (SD Inpainting) model as a baseline method. For person insertion, the interaction regions and the regions surrounding the inserted person are masked. For person removal, we mask the regions of the person to be removed. At inference time, SD Inpainting model is fed with the same masked image and masks as ours. Apart from pretrained SD Inpainting model, we also compare our model with two more baselines, which are adapted from the pretrained SD Inpainting model. The first baseline is the finetuned SD Inpainting model. We use our synthesized data to finetune the SD Inpainting model. The second baseline is the finetuned SD Inpainting model with skeleton control. Similar to our method, we add skeleton control to the SD inpainting model. We concatenate the skeleton with other inputs. Then we finetune the model using our generated data pair.

Paint by Example [61]. Paint by Example is an exemplar-based image editing method. It fills masked regions using exemplar images. The exemplar image is fed into an image encoder, and then the extracted features are fed into the cross-attention module. For person insertion, we feed the model the full mask covering the place where the person is to be inserted and the single-person image as the exemplar image. For person deletion, the mask region is the area where the person is removed, and the exemplar image is a crop of background.

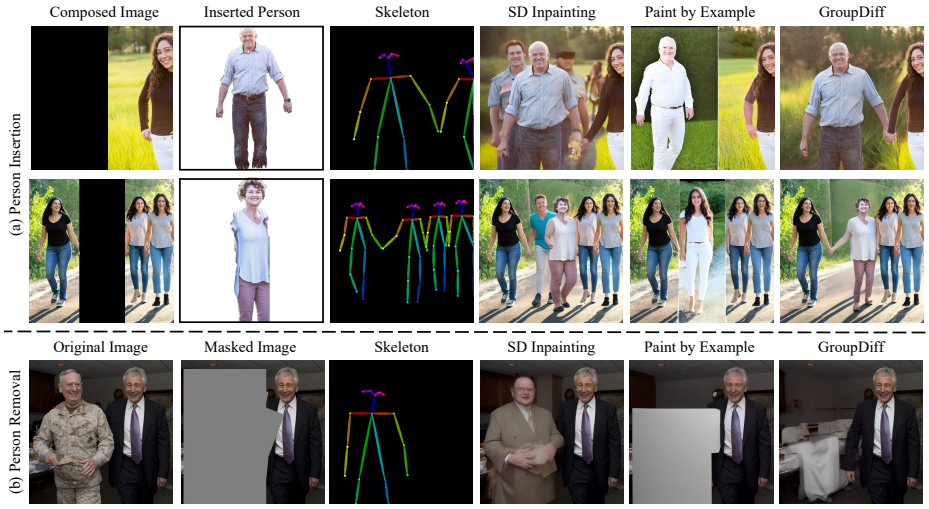

Fig. 8: Qualitative Comparisons. We compare our proposed GroupDiff against SD Inpainting [44] and Paint by Example [61]. (a) We show two examples of person insertion. (b) We show one example of person removal.

5.3 Qualitative Comparisons

Figure 8 shows some qualitative comparisons of our method. The results of person insertion are shown in Fig. 8(a). In the first row, we aim to add the old man to the image and make their hands hold together. According to the original image and the skeleton, we can get the masked image covering the interaction regions and the background regions. For the SD Inpainting model, the model adds a lot of persons behind the old man. Also, due to the lack of skeleton control, the interaction regions are not filled as expected. For Paint by Example, it inserts the old man into the left side of the photo and the obtained image looks to be composed of two split parts. Also, the appearance of the inserted person changed a lot. As for our proposed GroupDiff, it successfully inserts the person harmoniously. The hands of the persons correctly follow the given skeleton mask.

In the second example of Fig. 8(a), a lady is to be inserted into the middle of the image. After insertion, the hands of the lady are supposed to be connected with neighboring persons as indicated by the skeleton mask. Still, the SD Inpainting model tends to add another person into the middle. The interaction regions are not synthesized as expected. For the Paint by Example, a lady is inserted into the photo, however, the identity of the person is not the one indicated in the exemplar image. By contrast, our model successfully inserts the target person into the photo with reasonable interactions.

In Fig. 8(b), we show one example of person removal. We aim to remove the person standing on the left. The mask is obtained according to the bounding box of the person to be removed. After removing the person, we need to adjust the

Table 1: Quantitative Comparisons. We report PSNR, SSIM, FID, CLIP-I and DINO scores on synthesized data. Compared to baseline methods, our method achieves significantly better performance. "Random Mask" and "Mask One Person" refer to two ablation studies on the training data generation pipeline. "w/o inter-person" and "w/o intra-person" are two ablation studies on the person-aware module.

Methods	PSNR ↑	SSIM ↑	FID ↓	CLIP-I ↑	DINO ↑
SD-Inpainting [44]	17.4148	0.7562	24.1276	79.5467	91.2230
Paint-by-Example [61]	14.4848	0.6298	34.8999	74.1095	87.4915
Finetuned SD Inpainting	18.3132	0.7846	20.4833	82.4550	93.1132
Finetuned SD Inpainting with Skeleton	18.5688	0.7868	20.3351	83.3211	93.5158
GroupDiff (Ours)	**19.4569**	**0.8033**	**18.1035**	**86.8967**	**95.0284**

postures of the remaining person. Otherwise, the pose of the remaining person would look weird. Using the SD Inpainting model, there is another person present in the final image. For the Paint by Example model, we feed a part of background as the exemplar image. The masked region is filled with some random colors. The final column shows our result. By removing the person and feeding the adjusted skeleton, our model can synthesize a natural photo.

5.4 Quantitative Comparisons

As collecting large well-curated test sets is difficult, we conduct the quantitative comparisons on a synthetic dataset that the training has never seen. We use the validation set of LV-MHP-v2 [28], which has no overlap with our synthesized training data. We report the PSNR, SSIM, FID, CLIP-I, and DINO scores in Table 1. Compared with SD Inpainting and Paint by Example, our proposed GroupDiff has significantly better performance. Table 1 shows that the quantitative results improve after we fine-tune the pretrained SD inpainting model. Besides, skeleton control helps to improve the performance. The PSNR and SSIM values increase, while the FID score decreases. However, even with skeleton control, there is still a performance gap between this model and our proposed GroupDiff, which indicates the necessity of our proposed modules.

5.5 Ablation Studies

We report quantitative results for ablation studies in Table 2. We will show visual comparisons in supplementary files.

Training Data Generation Engine. We first investigate the performance of the model with "Random Mask". The ablation model is trained using the randomly masked group photo images. The images are randomly masked regardless of the bounding boxes of the persons and the skeletons. Quantitative results reported in Table 2 demonstrate that the model trained with the random mask has inferior performance. As discussed before, with the random mask, the model can infer the pose from the masked image and thus the capability of conditioning on the skeleton is limited. We also experiment with the "Mask One Person" strategy, we synthesize the data by directly masking one person according to

Table 2: Quantitative Comparisons for Ablation Studies. We conduct ablation studies from two persepctives. "Random Mask" and "Mask One Person" refer to two ablation studies on the training data generation pipeline. "w/o inter-person" and "w/o intra-person" are two ablation studies on the person-aware module.

Methods	PSNR ↑	SSIM ↑	FID ↓
Random Mask	17.3299	0.7727	26.2193
Mask One Person	17.9515	0.7777	22.7869
w/o inter-person	19.2612	0.7997	18.4946
w/o intra-person	<u>19.4236</u>	**0.8034**	<u>18.2274</u>
GroupDiff (Ours)	**19.4569**	<u>0.8033</u>	**18.1035**

the bounding box [25]. Still, by randomly masking a single person, the model has limited capability of inpainting interaction regions. As shown in Table 2, our model has a significantly better performance.

Person-aware Attention. As shown in Table 2, after removing the inter-person guidance, the model will have inferior PSNR, SSIM, and FID scores, which means the performance of the model "w/o inter-person" is inferior. As for removing the intra-person guidance, the model "w/o intra-person" has a better SSIM score but inferior PSNR and FID performance. We show visual examples in the supplementary file to demonstrate the effectiveness of this module.

6 Conclusion

In this work, we introduce a pioneering solution to the intricate challenging problem of group portrait editing. We formulate the task as an inpainting problem, addressing issues such as limited labeled data, appearance preservation, and manipulation flexibility. Our approach includes a data generation pipeline mimicking the real editing scenarios and a person-aware appearance preservation module for consistent editing results. Extensive experiments demonstrate the superiority of our proposed method. With the well-known difficulties of human generation and manipulation even for a single person, there is still ample room for refinement and innovation in our task. We hope our work can inspire the exploration of this important task.

Limitations. Our proposed GroupDiff makes an assumption that the inserted person image has correct facial expression and facing orientations. If a person image from the side view is given, our method cannot insert the person into the group photo where all people are facing front. Besides, in the current model, the lighting of the inserted person is changed as a preprocessing step. Thus, the ability to adjust the lighting is capped by the capability of the pretrained model. More powerful models can be used to better adjust the lighting [10, 21, 22].

Potential Negative Impact. The proposed method can be used to generate fake images by putting a lot of people who don't know each other in a group photo. It may be negatively used to fabricate certain facts.

Acknowledgement. This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221- 0012), and NTU NAP.

References

1. Albahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., Huang, J.B.: Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. ACM Transactions on Graphics (TOG) **40**(6), 1–11 (2021)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
3. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8340–8348 (2018)
4. Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. arXiv preprint arXiv:2304.02051 (2023)
5. Chen, C.Y., Chen, Y.C., Shuai, H.H., Cheng, W.H.: Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7513–7522 (2023)
6. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
7. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose` (2020)
8. Darabi, S., Shechtman, E., Barnes, C., Goldman, D.B., Sen, P.: Image melding: Combining inconsistent images using patch-based synthesis. ACM Transactions on graphics (TOG) **31**(4), 1–10 (2012)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
10. Guerreiro, J.J.A., Nakazawa, M., Stenger, B.: Pct-net: Full resolution image harmonization using pixel-wise color transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5917–5926 (2023)
11. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7543–7552 (2018)
12. Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., Wood, F.: Flexible diffusion modeling of long videos. arXiv preprint arXiv:2205.11495 (2022)
13. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Transactions on Graphics (ToG) **26**(3), 4–es (2007)
14. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
16. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR (2022)

17. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
19. Huang, Z., Li, H., Xie, Z., Kampffmeyer, M., Liang, X., et al.: Towards hard-pose virtual try-on via 3d-aware global correspondence learning. Advances in Neural Information Processing Systems **35**, 32736–32748 (2022)
20. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (ToG) **36**(4), 1–14 (2017)
21. Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: Ssh: A self-supervised framework for image harmonization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4832–4841 (2021)
22. Ke, Z., Sun, C., Zhu, L., Xu, K., Lau, R.W.: Harmonizer: Learning to perform white-box image and video harmonization. In: European Conference on Computer Vision. pp. 690–706. Springer (2022)
23. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: CVPR. pp. 2426–2435 (2022)
24. Konstantin Sofiiuk, Polina Popenova, A.K.: Foreground-aware semantic representations for image harmonization. arXiv preprint arXiv:2006.00809 (2020)
25. Kulal, S., Brooks, T., Aiken, A., Wu, J., Yang, J., Lu, J., Efros, A.A., Singh, K.K.: Putting people in their place: Affordance-aware human insertion into scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
26. Kwon, G., Ye, J.C.: Diffusion-based image translation using disentangled style and content representation. In: ICLR (2023)
27. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. In: ICLR (2023)
28. Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Sim, T., Yan, S., Feng, J.: Multiple-human parsing in the wild. arXiv preprint arXiv:1705.07206 (2017)
29. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European conference on computer vision (ECCV). pp. 85–100 (2018)
30. Liu, H., Jiang, B., Song, Y., Huang, W., Yang, C.: Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 725–741. Springer (2020)
31. Liu, L., Xu, W., Habermann, M., Zollhöfer, M., Bernard, F., Kim, H., Wang, W., Theobalt, C.: Neural human video rendering by learning dynamic textures and rendering-to-video translation. arXiv preprint arXiv:2001.04947 (2020)
32. Liu, L., Xu, W., Zollhoefer, M., Kim, H., Bernard, F., Habermann, M., Wang, W., Theobalt, C.: Neural rendering and reenactment of human actor videos. ACM Transactions on Graphics (TOG) **38**(5), 1–14 (2019)
33. Lu, S., Liu, Y., Kong, A.W.K.: Tf-icon: Diffusion-based training-free cross-domain image composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2294–2305 (2023)
34. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. arXiv preprint arXiv:1705.09368 (2017)

35. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 99–108 (2018)
36. Men, Y., Mao, Y., Jiang, Y., Ma, W.Y., Lian, Z.: Controllable person image synthesis with attribute-decomposed gan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5084–5093 (2020)
37. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. arXiv preprint arXiv:2211.09794 (2022)
38. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)
39. Ni, M., Li, X., Zuo, W.: Nuwa-lip: Language-guided image inpainting with defect-free vqgan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14183–14192 (2023)
40. Nie, X., Feng, J., Xing, J., Yan, S.: Generative partition networks for multi-person pose estimation. arXiv preprint arXiv:1705.07422 (2017)
41. Olszewski, K., Ceylan, D., Xing, J., Echevarria, J., Chen, Z., Chen, W., Li, H.: Intuitive, interactive beard and hair synthesis with generative models. In: CVPR. pp. 7446–7456 (2020)
42. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
43. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
45. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE TPAMI (2022)
46. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: CVPR. pp. 9243–9252 (2020)
47. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
48. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
49. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
50. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
51. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
52. Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., Aliaga, D.: Objectstitch: Object compositing with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18310–18319 (2023)

53. Tang, L., Ruiz, N., Qinghao, C., Li, Y., Holynski, A., Jacobs, D.E., Hariharan, B., Pritch, Y., Wadhwa, N., Aberman, K., Rubinstein, M.: Realfill: Reference-driven generation for authentic image completion. arXiv preprint arXiv:2309.16668 (2023)
54. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in Neural Information Processing Systems pp. 6306–6315 (2017)
55. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., et al.: Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18359–18369 (2023)
56. Wang, W., Alameda-Pineda, X., Xu, D., Fua, P., Ricci, E., Sebe, N.: Every smile is unique: Landmark-guided diverse smile generation. In: CVPR. pp. 7083–7092 (2018)
57. Wang, W., Cui, Z., Yan, Y., Feng, J., Yan, S., Shu, X., Sebe, N.: Recurrent face aging. In: CVPR. pp. 2378–2386 (2016)
58. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023)
59. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23550–23559 (2023)
60. Xing, J., Nagano, K., Chen, W., Xu, H., Wei, L.y., Zhao, Y., Lu, J., Kim, B., Li, H.: Hairbrush for immersive data-driven hair modeling. In: Proceedings of the 32Nd Annual ACM Symposium on User Interface Software and Technology. pp. 263–279 (2019)
61. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. arXiv preprint arXiv:2211.13227 (2022)
62. Yang, H., Huang, D., Wang, Y., Jain, A.K.: Learning face age progression: A pyramid architecture of gans. In: CVPR. pp. 31–39 (2018)
63. Zhang, B., Duan, Y., Lan, J., Hong, Y., Zhu, H., Wang, W., Niu, L.: Controlcom: Controllable image composition using diffusion model. arXiv preprint arXiv:2308.10040 (2023)
64. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
65. Zhao, J., Li, J., Cheng, Y., Sim, T., Yan, S., Feng, J.: Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 792–800 (2018)
66. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4606–4615 (2023)
67. Zhu, Y., Li, Z., Wang, T., He, M., Yao, C.: Conditional text image generation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14235–14245 (2023)