Text-guided Eyeglasses Manipulation with Spatial Constraints

Jiacheng Wang, Ping Liu*, Jingen Liu, Wei Xu*

***Abstract*—Virtual try-on of eyeglasses involves placing eyeglasses of different shapes and styles onto a face image without physically trying them on. While existing methods have shown impressive results, the variety of eyeglasses styles is limited and the interactions are not always intuitive or efficient. To address these limitations, we propose GlassesCLIP, a text-guided eyeglasses manipulation method with spatial constraints, which allows for control of the eyeglasses shape and style based on a binary mask and text, respectively. Specifically, we introduce a mask encoder to extract mask conditions and a modulation module that enables simultaneous injection of text and mask conditions. This design allows for fine-grained control of the eyeglasses' appearance based on both textual descriptions and spatial constraints. Our approach includes a disentangled mapper and a decoupling strategy that preserves irrelevant areas, resulting in better local editing. We employ a two-stage training scheme to handle the different convergence speeds of the various modality conditions, successfully controlling both the shape and style of eyeglasses. Extensive comparison experiments and ablation analyses demonstrate the effectiveness of our approach in achieving diverse eyeglasses styles while preserving irrelevant areas.**

***Index Terms*—Eyeglasses virtual try-on, Text-guided face attributes manipulation, Generative adversarial network.**

I. Introduction

VIRTUAL try-on technology allows individuals to virtually add fashion items to their personal images, facilitating the assessment of suitability and appeal. Specifically, virtual try-on for eyeglasses enables users to experiment with a variety of styles without physical presence, *e.g.,* sunglasses, metal glasses, and eyeglasses of different colors. This approach is highly efficient and convenient, as it eliminates the need for individuals to physically try on numerous eyeglasses to determine the appropriate style and shape. Moreover, virtual try-on technology for eyeglasses has gained popularity due to the emergence of short videos, as it enables users to apply various styles to recorded videos for enhanced visual effects, commonly referred to as "eyeglasses special effects" in different applications.

Several approaches have been proposed to improve eyeglasses virtual try-on technology [1]–[8]. These approaches can be broadly classified into two categories: 3D-based and 2D-based methods. 3D-based methods [1]–[5] employ 3D eyeglasses models to align the eyeglasses with the face. In contrast, 2D-based methods [6]–[10] directly edit input images using generative adversarial networks (GANs) [11] to produce the desired eyeglasses, which generally rely on the learned latent space to perform eyeglasses manipulation. However, a limitation of these methods is that the introduction of new eyeglasses styles (*e.g.,* sunglasses, metal glasses, and eyeglasses of different colors) typically requires either building new 3D eyeglasses models [1], [2] for 3D-based methods or recalculation of editing directions in the latent space [7], [8] for 2D-based methods. This can be inefficient and inconvenient to scale.

One alternative solution to this issue is to leverage text input as a means of conveniently scaling diverse eyeglasses styles. Vision-Language models, such as Contrastive Language-Image Pre-training (CLIP) [12], have enabled various methods for arbitrary text-guided image manipulation [13]–[20]. StyleCLIP [13], in particular, is a pioneering work in this area, leveraging the powerful text-to-image alignment capabilities of CLIP to manipulate original images based on text descriptions. Subsequently, many methods [14]–[20] have improved StyleCLIP to achieve better image editing results. However, these methods are primarily designed for general face manipulation and are rarely used for diverse eyeglasses manipulation. To ensure an intuitive and convenient eyeglasses virtual try-on process, it is imperative to conduct further research on controlling eyeglass styles through text prompts.

As we directly applying these methods to eyeglass manipulation tasks, they suffers from several drawbacks, as illustrated by some results from StyleCLIP [13] shown in Fig. 1 (refer to Fig. 5 for more comparisons). Firstly, it can be difficult to express the degree of editing precisely, particularly for the spatial configuration (*e.g.*, sizes, shapes) of eyeglasses, which is critical for eyeglasses virtual try-on. For instance, using the same text prompt of "big glasses" may produce different

Manuscript received 4 May 2023; revised 16 July 2023; accepted 20 September 2023. This work is supported by A*STAR Career Development Funding Award (Grant No: 222D800031). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yuxin Peng. (Corresponding author: Ping Liu, Wei Xu.)

J. Wang and W. Xu are with the Hubei Key Laboratory of Smart Internet Technology, School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China, e-mail: jiacheng@hust.edu.cn, xuwei@hust.edu.cn.

P. Liu is with the Center for Frontier AI Research (CFAR), Research Agency for Science, Technology and Research (A*STAR), Singapore 138634, e-mail: pino.pingliu@gmail.com.

J. Liu is with Disney Streaming Advanced Research, USA, email: jingen.liu@gmail.com.

* means co-corresponding author.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org provided by the author. The material includes selection of spatial constraint, preliminary knowledge, implementation details, detailed quantitative results, more ablation analysis, and failure examples. This material is 5 pages in size.

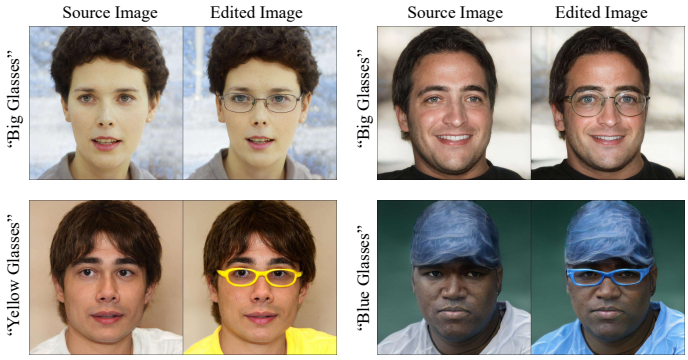


Fig. 1. Issues existing on text-guided virtual try-on for eyeglasses. The first row depicts that eyeglasses in the editing results are of different sizes based on the same text prompt of "big glasses". The second row shows an entangled phenomenon that irrelevant areas are significantly modified, especially in the cloth region.

editing results with eyeglasses of varying sizes, indicating an ambiguity challenge that size words such as "small", "medium", and "large" may not be consistent with the absolute pixel sizes of items in images [21]. Secondly, modifying the eyeglasses region conditioned on text prompts often leads to noticeable changes in unrelated areas, such as the clothing region, which is a common entangled phenomenon in many existing methods [13]–[15]. In this manner, it can be challenging to succinctly modify specific areas to accommodate the given condition while maintaining the consistency of unrelated regions, which requires specific design.

To address the ambiguity challenge and disentanglement challenge, we propose GlassesCLIP, a text-guided eyeglass manipulation method with spatial constraints. Inspired by HairCLIP [22], we harness the remarkable expression capacity of textual descriptions to manipulate diverse styles of eyeglasses using a single model. However, different from HairCLIP, our approach integrates spatial constraints through the use of eyeglasses masks to handle the ambiguity challenge. Specifically, the mask is automatically generated by an off-the-shelf face parser on the FFHQ [23] and CelebA-HQ [24] datasets, which are then used to construct aligned data pairs based on the original face pose. To address the disentanglement challenge, we view the editing process as a two-step procedure, with one step concentrating on concisely altering eyeglasses and the other on maintaining the consistency of unrelated areas. As a result, we specifically design the following components and training strategy to solve the eyeglass editing challenges.

Firstly, we introduce a mask encoder comprising multiple convolutional blocks to extract mask conditions, which are then employed to control the spatial configuration of eyeglasses for more precise manipulation. Simultaneously, we design a modulation module to allow for the simultaneous injection of text and mask conditions. Utilizing the binary eyeglasses mask and straightforward text prompts, we can effortlessly regulate the shape and style of eyeglasses within a single model.

Secondly, to address the entangled issue, we propose GlassMapper, which consists of an editing mapper and a disentangled mapper to enable more distinct and disentangled editing directions. The editing mapper is responsible for controlling the eyeglasses style, while the disentangled mapper preserves irrelevant areas. We achieve this by employing a simple yet effective decoupling strategy, which involves truncating the gradient flow from the disentangled mapper to the editing mapper.

Thirdly, we adopt a two-stage training scheme to address the varying convergence speeds of different modalities. This enables us to sequentially equip the GlassMapper with the ability to modify eyeglasses, resulting in a stable training process. After completing the two-stage training, we can simultaneously manipulate the style and shape of eyeglasses based on different modality conditions.

In summary, our contributions are as follows:

- We propose GlassesCLIP, a text-guided eyeglasses manipulation method with spatial constraints, to achieve diverse and flexible eyeglasses virtual try-on. Our method controls eyeglasses shape and style based on a simple binary eyeglasses mask and text description, respectively. This approach accommodates a wide range of eyeglasses shapes and common styles within a single model, offering a more convenient and intuitive mode of interaction.
- To support multiple modalities of information in a single model, we propose a new modulation module that combines the binary eyeglasses mask and natural language descriptions. Furthermore, we employ a two-stage training scheme to stabilize the training process, sequentially equipping our method with the ability to modify eyeglasses.
- To better preserve irrelevant areas, we introduce a disentangled mapper and a simple decoupling strategy, resulting in better local editing. This allows for more accurate and precise control over the eyeglasses style and shape, while also maintaining the integrity of the surrounding areas in the image.
- Extensive quantitative and qualitative experiments on the CelebA-HQ dataset [24] are conducted to demonstrate the superiority of our approach. On average, we outperform the state of the arts [14] by 6.28%, 12.89%, and 3.32% on *Structure Similarity Index Measure* (SSIM) [25], *Peak Signal to Noise Ratio* (PSNR) [26], and *Identity Discrepancy Scores* (IDS), respectively.

II. Related Work

A. Eyeglasses Virtual Try-on

The goal of eyeglasses virtual try-on is to add eyeglasses with a specific style to the face of an image or video, which is opposite to eyeglasses removal [27], [28]. This can be accomplished through two main categories of methods: 3D-based eyeglasses virtual try-on and 2D-based eyeglasses virtual try-on. 3D-based eyeglasses virtual try-on methods [1]–[5] typically add eyeglasses to a face image by aligning the 3D models of eyeglasses and face. For example, Milanova et al. [1] proposed a markerless eyeglasses virtual try-on system that overlays the 3D eyeglasses model over the face image according to the estimated head pose. Similarly, Feng et al. [2]

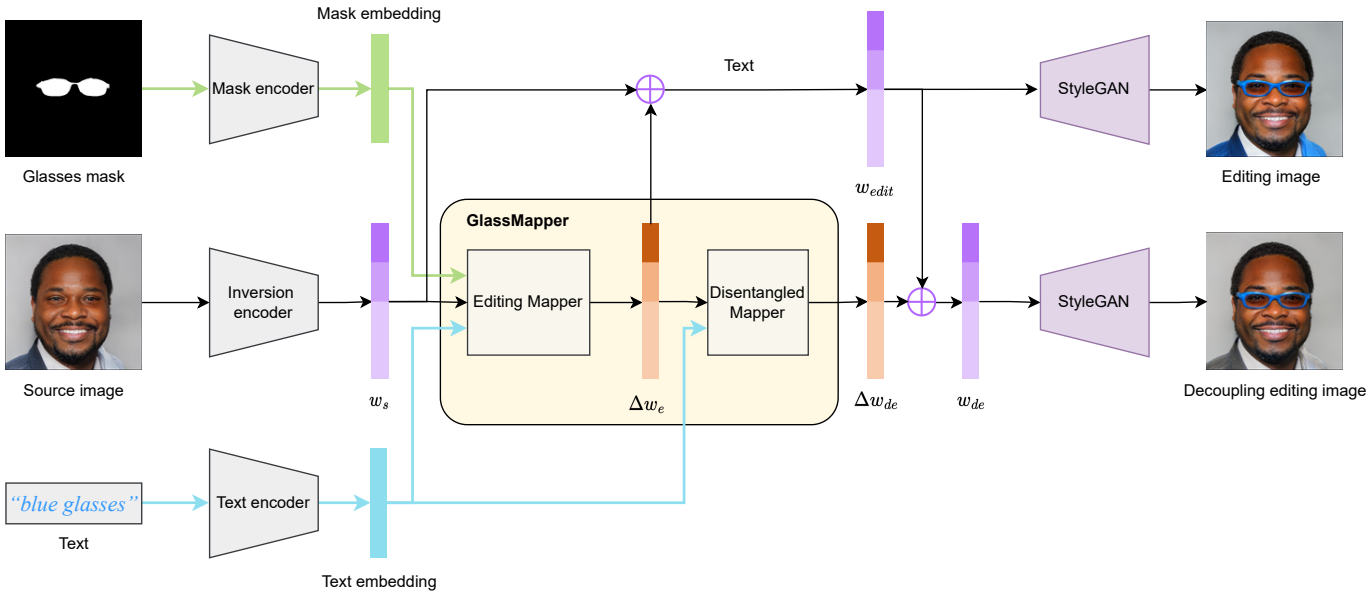


Fig. 2. An overview of GlassesCLIP. Our approach takes the source latent code w_s, mask embedding e_m, and text embedding e_t as input through a StyleGAN inversion encoder, mask encoder, and CLIP text encoder, respectively. Later, the GlassMapper predict an editing latent code w_{edit} and a decoupling editing latent code w_{de}, which are then applied to the StyleGAN generator to generate the editing image I_{edit} and the decoupling editing image I_{de}. Specifically, I_{edit} achieves the desired manipulation but unavoidably suffers significant modification in unrelated regions. In contrast, I_{de} inherits the impressive manipulation capability while preserving irrelevant areas well. The details of GlassMapper are provided in Section III-A.

achieved 3D eyeglasses virtual try-on by combining 3D face reconstruction and pose estimation techniques. While there are many real-time eyeglasses virtual try-on systems based on 3D models, each change of eyeglasses style requires a new 3D eyeglasses model, which can be difficult to collect and scale.

2D-based eyeglasses virtual try-on methods [6]–[10] directly edit the given images to obtain the desired eyeglasses. Some methods [7]–[9] use pre-trained GANs, specifically StyleGAN [23], [29], and learn an editing direction in their latent space to control the generated eyeglasses style. Others [6], [10] mainly train an encoder and a generator from scratch, where the encoder extracts the latent representation of the original image and the generator achieves arbitrary eyeglasses style manipulation based on different attribute constraints. However, the variations of generated eyeglasses in these methods are limited, *e.g.*, only black plastic eyeglasses or sunglasses can be added in most cases. To handle this limitation, a concurrent work GlassGAN [30] focuses on multi-style eyeglasses virtual try-on using pre-trained StyleGAN [23], [29]. They explore the editing directions targeted toward eyeglasses in an unsupervised setting, *i.e.*, solving an eigen-problem. However, they need to subjectively assign human-interpretable attributes to the explored editing directions, such as size, position, squareness, roundness, cat-eye appearance, and thickness, which are limited to express the spatial consistency of generated eyeglasses.

B. Text-guided Image Manipulation

With the development of generative models, there has been a significant amount of research on facial image synthesis and manipulation. The objective of facial image synthesis is to produce realistic and high-fidelity portraits. Unconditional methods [23], [29] emphasize the diversity and quality of the resulting images, whereas conditional methods [31]–[37] prioritize the ability to control the generated images by utilizing textual descriptions [31], [32], sketches [33]–[35], or semantic maps [36], [37]. Moreover, facial image manipulation [38]–[42] endeavors to alter facial attributes in accordance with user predilections, including images [40], [42], sketches [38], and textual descriptions [39].

In pursuit of a more convenient and intuitive means of manipulation interaction, text-guided image manipulation aims to modify source images according to simple text descriptions, while preserving irrelevant areas. With the powerful capabilities of CLIP [12], StyleCLIP [13] proposed a CLIP loss to make the editing results consistent with the given text. Later, to better preserve irrelevant areas and learn a more disentangled editing direction, several works [14], [16] proposed disentangled loss functions, and [17] learned an attention mask to limit editing areas. Given that the original CLIP loss proposed in StyleCLIP is vulnerable to adversarial samples, [15], [19], [20] modified the CLIP loss using data augmentation or contrastive learning to improve robustness.

Several works [18], [31], [43], [44] achieved arbitrary text-guided image manipulation without inference-time optimization or restricting to single attribute editing. Specifically, [18], [31], [43] focused on bridging the latent space of StyleGAN [23], [29] and CLIP [12], while [44] injected text conditions into an attention decoder to obtain corresponding editing directions. With the development of diffusion models, several works [45]–[47] combined CLIP with diffusion models [48]–[50] to handle the limitations of GAN inversion [45] and extend the image manipulation domain [46], [47]. HairCLIP [22] is another recent work that focuses on the diversity of

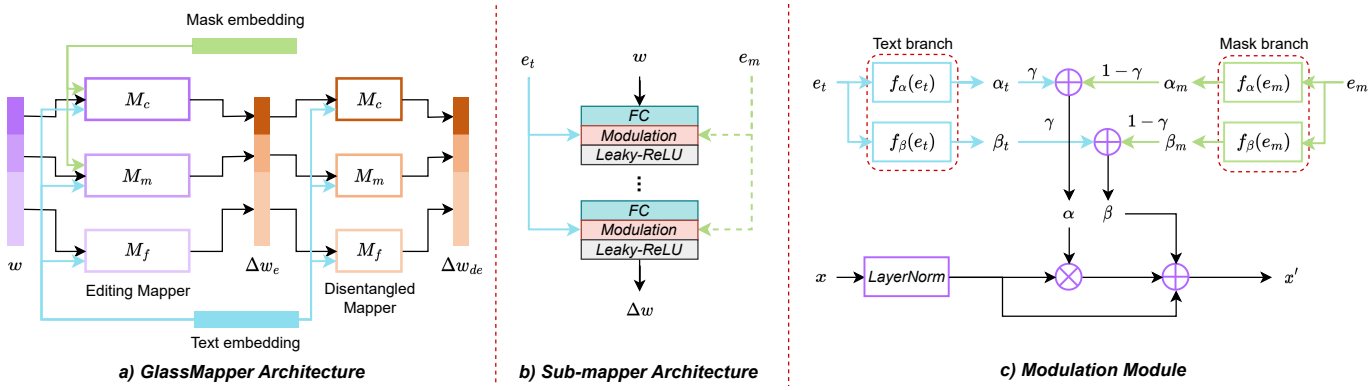


Fig. 3. The architecture of GlassMapper. a) GlassMapper is composed of an editing mapper and a disentangled mapper, each of which is divided into three sub-mappers. b) Each sub-mapper consists of 5 blocks, which are stacked with a fully connected layer, a modulation module, and a leaky ReLU activation layer. c) The modulation module is responsible for modulating the input $x \in \mathbb{R}^{512}$ based on the mask condition $e_m \in \mathbb{R}^{512}$ and text condition $e_t \in \mathbb{R}^{512}$.

specific attributes, specifically hair attributes, which achieves control of 44 hairstyles and 12 hair colors simultaneously.

In this study, we employ straightforward textual cues to control the eyeglasses style for a more user-friendly and intuitive interface. Furthermore, to precisely regulate the spatial configuration of the eyeglasses without introducing undue intricacy, we suggest the implementation of a binary eyeglasses mask as a spatial constraint.

III. Method

Firstly, in Section III-A, we will provide a detailed description of our GlassesCLIP, including the conditional information encoder, the GlassMapper, and the decoupling strategy. Section III-B will contain a thorough examination of the loss functions implemented during the training process. Finally, in Section III-C, we will present the two-stage training scheme and its corresponding objective functions.

A. Framework

Like previous works such as StyleCLIP [13] and HairCLIP [22], our work utilizes a pre-trained StyleGAN to generate high-quality face images. As the $\mathcal{W}+$ space has good semantic decoupling properties, we leverage it to learn a GlassMapper that predicts editing directions to control the shape and style of eyeglasses, given the text prompt and automatically generated mask.

As shown in Fig. 2, we start by obtaining the source latent code $w_s \in \mathcal{W}+$, the mask embedding e_m, and the text embedding e_t through a StyleGAN inversion model, a mask encoder, and a CLIP text encoder, respectively, given a real image I_{src}, a eyeglasses mask m, and a text description t.

The obtained embeddings of different modalities are then sent to the GlassMapper, which consists of an editing mapper and a disentangled mapper. The editing mapper predicts an editing direction Δw_e based on e_m and e_t while the disentangled mapper predicts a more decoupling editing direction Δw_{de} based on e_t, given the calculated editing direction Δw_e.

Finally, the editing latent code $w_{edit} = w_s + \Delta w_e$ and the decoupling editing latent code $w_{de} = w_{edit} + \Delta w_{de}$ are fed into StyleGAN to generate the editing image I_{edit} and the decoupling editing image I_{de}. Specifically, I_{edit} successfully achieves the desired manipulation but unavoidably suffers significant modification in unrelated regions. In contrast, I_{de} inherits the impressive manipulation capability of I_{edit} while preserving irrelevant areas to the greatest extent possible.

Conditional Information Encoder: For text descriptions, a pre-trained CLIP [12] text encoder is used to extract an expressive text embedding $e_t \in \mathbb{R}^{512}$. However, for eyeglasses masks, there is no pre-trained eyeglasses mask encoder available. Therefore, a convolutional neural network, similar to that used in MaskGAN [51], is employed to extract mask conditions. The mask encoder consists of 5 convolutional blocks, each comprising a convolutional layer, an instance normalization layer, and a ReLU activation layer. Given an eyeglasses mask, the mask encoder extracts the corresponding mask embedding $e_m \in \mathbb{R}^{512}$. These mask and text embeddings are then used to condition the GlassMapper and control the shape and style of eyeglasses in the manipulated image I_{edit} and decoupling editing image I_{de}.

GlassMapper Architecture: In Fig. 3(a), we can see that GlassMapper consists of two distinct mappers: an editing mapper and a disentangled mapper. The editing mapper is specifically designed to manipulate the shape and style of eyeglasses, while the disentangled mapper is responsible for preserving areas that are not relevant to the eyeglasses.

Studies [13], [32] have demonstrated that different layers of StyleGAN control different levels of semantic information in the generated image. For instance, shallow layers control coarse-grained information like head pose and facial expressions, while deep layers control fine-grained information like color and micro details. Therefore, we divide the GlassMapper into three sub-mappers, namely M_c, M_m, and M_f, each controlling different semantic levels in the generated image. M_c and M_m are responsible for controlling coarse-grained information like eyeglasses shape, while M_f controls fine-grained information such as eyeglasses color. In accordance with this, we split the source latent code w_s into three parts: w_c, w_m, and w_f, which represent different semantic information.

The editing mapper is purposefully engineered to manipulate the eyeglasses, without being tasked with preserving the ir-

relevant areas. Each sub-mapper of the editing mapper consists of 5 blocks, and each block consists of a fully connected layer, a modulation module, and a leaky relu activation layer, as shown in Fig. 3(b). Since we want to control eyeglasses shape based on the mask, we inject the mask conditions into M_c and M_m. On the other hand, we use text to control eyeglasses style, which involves fine-grained information like eyeglasses color, so we inject text conditions into all sub-glasses mappers. Therefore, the editing mapper can be expressed as:

$$\begin{aligned} &M\left(w_s, e_t, e_m\right) \\ &= \left(M_c\left(w_c, e_t, e_m\right), M_m\left(w_m, e_t, e_m\right), M_f\left(w_f, e_t\right)\right). \end{aligned} \tag{1}$$

Disentangled mapper is similar to the editing mapper but with two differences: 1) The number of blocks is set to 2, less than the editing mapper, according to a simpler goal that reduces the modification in the irrelevant areas; 2) Considering that large modifications in the irrelevant areas are always accompanied by the change of eyeglasses style, we only keep the text branch in the modulation module, ensuring the capability of dealing with different text descriptions. Therefore, the disentangled mapper can be expressed as:

$$\begin{aligned} &M\left(\Delta w, e_t\right) \\ &= \left(M_c\left(\Delta w_c, e_t\right), M_m\left(\Delta w_m, e_t\right), M_f\left(\Delta w_f, e_t\right)\right). \end{aligned} \tag{2}$$

To incorporate both text conditions and mask conditions into M_c and M_m, we modify the original modulation module by introducing a mask branch and a text branch to calculate corresponding scale and bias parameters $\boldsymbol{\alpha_m}$, $\boldsymbol{\beta_m}$, $\boldsymbol{\alpha_t}$, and $\boldsymbol{\beta_t}$, as shown in Fig. 3(c). These parameters are then fused using a weight γ to obtain the final scale parameter $\boldsymbol{\alpha}$ and bias parameter $\boldsymbol{\beta}$. Finally, the input $\boldsymbol{x}$ is modulated using $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as follows:

$$\begin{aligned} \boldsymbol{\alpha} &= (1-\gamma) * \boldsymbol{\alpha_m} + \gamma * \boldsymbol{\alpha_t}, \\ \boldsymbol{\beta} &= (1-\gamma) * \boldsymbol{\beta_m} + \gamma * \boldsymbol{\beta_t}, \\ \boldsymbol{x'} &= (1+\boldsymbol{\alpha}) \frac{\boldsymbol{x}-\mu_x}{\sigma_x} + \boldsymbol{\beta}, \end{aligned} \tag{3}$$

where μ_x and σ_x denote the mean and standard deviation of x, respectively.

Simple Decoupling Strategy: Generating eyeglasses with diverse shapes and styles while preserving irrelevant regions presents a significant challenge. Our experimental observations suggest that this objective may compromise the diversity and realism of the generated eyeglasses. To address this issue, we propose a simple yet effective decoupling strategy comprising two operations: 1) We truncate the gradient flow from the disentangled mapper to the editing mapper, thereby liberating the editing mapper from the responsibility of preserving unrelated areas; 2) Based on the outcomes of the editing mapper, we then use the source image as a reference to minimize alterations in unrelated areas while simultaneously preserving the eyeglasses region. In summary, with the decoupling strategy, the editing mapper mainly focuses on controlling the eyeglasses style while the disentangled mapper mainly focuses on preserving the irrelevant areas, without affecting each other.

B. Loss Function

In order to meet our objective of controlling the shape and style of eyeglasses in the manipulated image based on mask and text conditions, while preserving the irrelevant areas, we have devised several loss functions. The details are explained below.

Shape Consistency Loss: We propose a shape consistency loss that utilizes a face parser network [51] to guide our model in generating eyeglasses that are consistent with the given mask. Specifically, we first obtain the segmentation label S_{src} of the source image I_{src} using the face parser network P. Next, we combine S_{src} and the eyeglasses mask $\boldsymbol{m}$ to create the target segmentation label S_{tar}, which classifies pixels in the mask region as eyeglasses (excluding the eyes category) while leaving the others unchanged. The combination is defined as follows:

$$\left(S_{tar}\right)_{ij} = \begin{cases} N_{glasses}, & \text{if } m_{ij} = 1 \,\&\, \left(S_{src}\right)_{ij} \neq N_{eyes} \\ \left(S_{src}\right)_{ij}, & \text{otherwise,} \end{cases} \tag{4}$$

where $N_{glasses}$ and N_{eyes} denote the category number of eyeglasses and eyes, respectively.

To ensure that the generated eyeglasses have the desired shape conditioned on $\boldsymbol{m}$, we then minimize the cross-entropy loss between S_{tar} and the predicted probability of the manipulated image I_{edit}. The shape consistency loss is defined as follows:

$$\mathcal{L}_{sc} = CE\left(P\left(I_{edit}\right), S_{tar}\right), \tag{5}$$

where CE is cross-entropy loss.

Classification Loss: To avoid the undesirable phenomenon where GlassMapper relies on shortcuts that inconspicuous and mutilated eyeglasses may match the given mask and achieves a low shape consistency loss, we propose an eyeglasses classification loss. GlassMapper leverages this eyeglasses classifier[1] to distinguish between images with and without eyeglasses. The eyeglasses classification loss is expressed as:

$$\mathcal{L}_{cls} = Cg\left(I_{edit}\right), \tag{6}$$

where Cg denotes the eyeglasses classifier. Intuitively, if the classification score decreases, the present probability of eyeglasses will increase, *i.e.*, eyeglasses in the manipulated image will be more fidelity and complete.

CLIP-NCE Loss: To perform eyeglasses style manipulation conditioned on the text description, we adopt the CLIP-based Noise Contrastive Estimation (CLIP-NCE) loss [15], which encourages the generated eyeglasses to be semantically similar to the input text description. It is defined as:

$$\begin{aligned} \mathcal{L}_{NCE} = &-\log \frac{e^{(Q \cdot K_T^+ / \tau)}}{e^{(Q \cdot K_T^+ / \tau)} + \sum_{K^-} e^{(Q \cdot K^- / \tau)}} \\ &-\log \frac{e^{(Q \cdot K_I^+ / \tau)}}{e^{(Q \cdot K_I^+ / \tau)} + \sum_{K^-} e^{(Q \cdot K^- / \tau)}}, \end{aligned} \tag{7}$$

where τ is the temperature and is set to 1.0. Q, K_T^+, K_I^+ and K^- denote as query, text positive pairs, image positive pairs, and negative pairs, respectively.

[1] https://github.com/apnkv/eyeglasses_on_photo

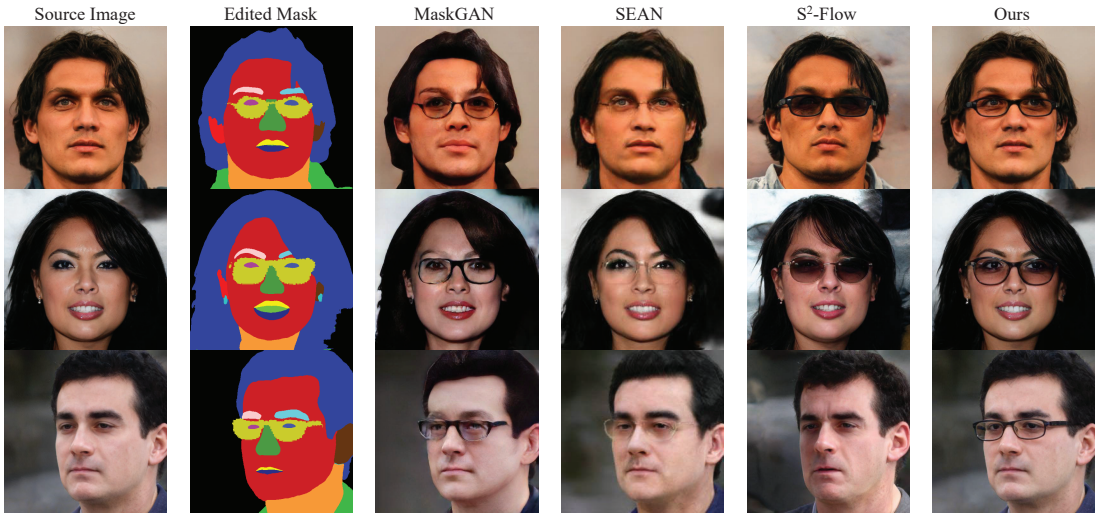


Fig. 4. Qualitative comparisons to semantic-based editing methods: MaskGAN [51], SEAN [36], S^2-Flow [54]. Our approach successfully generates natural and high-fidelity eyeglasses consistent with the given semantic mask, along with a slight loss of identity information.

Latent Loss and ID Loss: Similar to StyleCLIP [13], we adopt L2 distance to regularize the latent code and constrain the image domain; we also adopt cosine similarity to evaluate the facial feature similarity and preserve the identity information. The formulations are defined as follows:

$$\mathcal{L}_{norm} = \|w_{edit} - w_s\|_2 , \tag{8}$$

$$\mathcal{L}_{id} = 1 - \cos\left(R\left(I_{edit}\right), R\left(I_{src}\right)\right), \tag{9}$$

where R denotes ArcFace [52] network for face recognition.

Background Loss: To preserve the irrelevant areas, we directly compute pixel-wise mean square error between manipulated image I_{edit} and source image I_{src}, which is expressed as:

$$\mathcal{L}_{bg} = \|(I_{edit} - I_{src}) * (P_{ng}\left(I_{edit}\right) \cap (1 - m))\|_2 , \tag{10}$$

where $P_{ng}()$ denotes non-eyeglasses areas mask.

Disentangled Loss: When editing the eyeglasses color, we observe a consequent change in cloth color. As the $W+$ space of StyleGAN exhibits strong semantic decoupling properties, we hypothesize that the entanglement issue may arise from the global guidance provided by CLIP, which may bind the color attribute to other objects, rather than eyeglasses. In other words, CLIP tends to identify an entangled editing direction in which clothing also conforms to color attributes, making it challenging to fully utilize the decoupling properties of $W+$ space. Therefore, as the RGB color space is not linearly correlated with human visual perception, we convert all the images from RGB color space to LAB [53] color space and define the disentangled loss as:

$$\begin{aligned}\mathcal{L}_{disentangle} =& \lambda_g \|(I_{de} - I_{edit}) * P_g\left(I_{edit}\right)\|_2 \\ &+ \lambda_c \|(I_{de} - I_{src}) * P_c\left(I_{src}\right)\|_2 ,\end{aligned} \tag{11}$$

where $P_g()$ denotes eyeglasses area mask and $P_c()$ denotes cloth area mask. λ_g and λ_c are set to 4, 5 respectively. In this way, we achieve desired eyeglasses style in I_{de} under the supervision of I_{edit}, and preserve the irrelevant area under the supervision of I_{src}.

C. Training Scheme

Intuitively, we can divide the whole task into two sub-tasks, one of which is to control the shape of eyeglasses based on the mask, and the other is to control the style of eyeglasses based on the text. As we observed that the training convergence speed of each sub-task is distinct, roughly 20 times the difference, it is challenging to train our framework end to end based on the mask and text simultaneously.

To handle the disparate training convergence rate of different modalities, we adopt a two-stage training approach. In the first stage (Stage-I), we train the editing mapper separately for each sub-task to enable preliminary control of the eyeglasses. In the second stage (Stage-II), we jointly train the editing mapper to simultaneously achieve all sub-tasks while introducing and learning the disentangled mapper to preserve the irrelevant areas. With this approach, we are able to control the eyeglasses shape and style based on the mask and text simultaneously, while also preserving the irrelevant areas.

Stage-I: Considering the distinct training convergence speed of each sub-task (*i.e.*, mask-guided eyeglasses shape manipulation and text-guided eyeglasses style manipulation), we separately train the editing mapper to achieve each sub-task. First, we focus on the sub-task of mask-guided eyeglasses manipulation, where we train the editing mapper to control the eyeglasses shape conditioned on the mask. Specifically, we set the weight γ to 0 and exclude the fine mapper, *i.e.*, only the mask encoder and mask branches in the course mapper and medium mapper are trainable. In this way, the objective function is defined as:

$$\mathcal{L} = \lambda_{sc}\mathcal{L}_{sc} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{norm}\mathcal{L}_{norm} + \lambda_{id}\mathcal{L}_{id} + \lambda_{\text{bg}}\mathcal{L}_{\text{bg}}, \tag{12}$$

where λ_{sc}, λ_{cls}, λ_{norm}, λ_{id} and λ_{bg} are set to 3, 0.03, 0.8, 0.1 and 2, respectively.

Later, we focus on the sub-task of text-guided eyeglasses manipulation, where we incorporate text conditions into the editing mapper to modify the eyeglasses style. Specifically, we freeze the parameters of the pre-trained mask encoder and

mask branches and train all the text branches in the editing mapper. Moreover, we also set the weight γ to 0.5, which means the mask conditions and text conditions contribute to the editing results equally. Finally, the objective function is:

$$\mathcal{L} = \lambda_{nce}\mathcal{L}_{nce} + \lambda_{norm}\mathcal{L}_{norm} + \lambda_{id}\mathcal{L}_{id}, \tag{13}$$

where λ_{nce}, λ_{norm}, λ_{id}, are set to 0.3, 0.8, and 0.2, respectively.

Stage-II: After Stage-I, the editing mapper has preliminary capabilities to achieve each sub-task separately, while performing poorly on the whole task. In other words, given a mask and text simultaneously, editing results would be overly modified based on the text, and also inconsistent with the mask. Therefore, we jointly train the editing mapper based on the mask and text simultaneously, *i.e.*, all the parameters in the editing mapper are trainable at this stage. To preserve the irrelevant areas, we also introduce and learn a disentangled mapper to predict more disentangled editing directions. Combined with the decoupling strategy, the editing mapper and the disentangled mapper would perform their duties independently of each other. Finally, the total objective function is:

$$\begin{aligned}\mathcal{L} = \lambda_{nce}\mathcal{L}_{nce} + \lambda_{norm}\mathcal{L}_{norm} + \lambda_{id}\mathcal{L}_{id} \\ + \lambda_{\text{bg}}\mathcal{L}_{\text{bg}} + \lambda_{\text{sc}}\mathcal{L}_{\text{sc}} + \lambda_{\text{disentangle}}\mathcal{L}_{\text{disentangle}},\end{aligned} \tag{14}$$

where λ_{nce}, λ_{norm}, λ_{id}, λ_{bg}, λ_{sc}, and $\lambda_{disentangle}$ are set to 0.3, 0.8, 0.2, 5, 4, and 1, respectively.

IV. Experiments

A. Implementation Details

We use FFHQ dataset [23] for training and CelebA-HQ dateset [24] for evaluating. Since we focus on the manipulation of eyeglasses, we split all the datasets based on the presence of eyeglasses. To obtain diverse eyeglasses masks, we utilize a face parser to segment all the images with eyeglasses and get 12,698/1,341 eyeglasses masks in FFHQ/CelebA-HQ datasets. In this way, we construct 12,698 data pairs in FFHQ as the training set and 1,341 data pairs in CelebA-HQ as the test set[2].

We use an 18-layer StyleGAN2 [29] pretrained on FFHQ as our generator and the e4e [55] encoder pretrained on FFHQ as our inversion model. For text input, we select seven common eyeglasses colors(*i.e.*, red, blue, green, yellow, pink, orange, and purple) and two common eyeglasses styles(*i.e.*, metal glasses and sunglasses). As for training, Stage-I is trained for 150,000 iterations, among which we train the mask branches for 145,000 iterations with a base learning rate of 0.005 and train the text branches for 5,000 iterations with a base learning rate of 0.002. Stage-II is trained for 20,000 iterations with a base learning rate of 0.001.

To quantitatively evaluate the performance of our model, we use *Structure Similarity Index Measure* (SSIM) [25], *Peak Signal to Noise Ratio* (PSNR) [26], *Fréchet Inception Distances* (FID) [56], *Identity Discrepancy Scores* (IDS), *mean of class-wise Intersection over Union* (mIoU), *Pixel Accuracy* (PA), and *CLIP Similarity Score* (CLIPScore) as evaluation metrics. Intuitively, we used mIoU and PA to evaluate the consistency between the eyeglasses shape and the mask, and CLIPScore to evaluate the alignment between the eyeglasses style and the text description. SSIM, PSNR, IDS, and FID are used to evaluate the preservation of irrelevant regions and identity information. More implementation details can be found in the supplementary material.

[2]Please note that our constructed dataset mainly consists of face images without eyeglasses, binary eyeglass masks, and some textual descriptions of eyeglass styles. Therefore, the training dataset does not contain samples that represent the eyeglass styles discussed in the introduction.

TABLE I
QUANTITATIVE COMPARISONS TO SEMANTIC-BASED EDITING METHODS. WE ACHIEVE BETTER RESULTS IN MOST CASES, DEMONSTRATING A STRIKING EDITING CAPABILITY OF PRESERVING IRRELEVANT AREAS AND GENERATING HIGH-FIDELITY EYEGLASSES CONSISTENT WITH THE GIVEN SEMANTIC MASK SIMULTANEOUSLY.

Method	SSIM(↑)	PSNR(↑)	IDS(↑)	FID(↓)	mIoU(↑)	PA(↑)
MaskGAN [51]	0.7086	20.6624	0.2823	39.44	0.6926	0.8985
SEAN [36]	0.7365	21.3615	0.4621	**15.30**	0.7467	0.9249
S^2-Flow [54]	0.6824	16.5182	0.6057	21.81	0.5194	0.8400
Ours	**0.8936**	**27.7014**	**0.7009**	27.10	**0.7741**	**0.9584**

TABLE II
QUANTITATIVE COMPARISONS TO TEXT-BASED EDITING METHODS. IN MOST CASES, WE ACHIEVE SIGNIFICANT IMPROVEMENTS, INDICATING THE SUPERIOR CAPABILITY OF PRESERVING IRRELEVANT AREAS AND IDENTITY INFORMATION.

Method	SSIM(↑)	PSNR(↑)	IDS(↑)	FID(↓)	CLIPScore(↑)
StyleCLIP [13]	0.7303	17.0802	0.5250	102.10	24.4427
PPE [14]	0.8077	20.2526	0.5911	107.68	23.8046
CF-CLIP [15]	0.7792	19.2501	0.5193	138.2	24.9713
HairCLIP [22]	0.8298	23.2001	0.6358	138.18	**25.7187**
DeltaEdit [18]	0.7660	18.7950	0.4700	**48.98**	20.6276
Ours	**0.8819**	**26.1910**	**0.6569**	94.56	24.2708

B. Qualitative and Quantitative Comparison

To achieve diverse and flexible eyeglasses virtual try-on, we propose GlassesCLIP to control eyeglasses shape and style based on a simple binary eyeglasses mask and text description, respectively. Some manipulation methods [36], [51], [54] can edit various facial attributes based on the spatial information provided by the semantic map, while others [13]–[15], [18], [22] can edit facial images to align with the given text description. This seems to suggest that existing manipulation methods have the potential to be directly applied to eyeglasses manipulation tasks to achieve arbitrary changes in eyeglasses shape and style. However, eyeglasses manipulation tasks are distinct and challenging, requiring consideration of some special issues, such as maintaining irrelevant areas when changing eyeglasses style, and handling eyeglasses material and reflection effects. Therefore, we chose to compare our method with semantic-based editing methods [36], [51], [54] and text-based editing methods [13]–[15], [18], [22] to discuss the unique challenges of eyeglasses manipulation, demonstrate our effectiveness in addressing these challenges, and achieving superior results.

Comparison to Semantic-based Editing Methods: The semantic-based editing methods aim to manipulate the source

Fig. 5. Qualitative comparisons to text-based editing methods: StyleCLIP [13], PPE [14], CF-CLIP [15], HairCLIP [22], and DeltaEdit [18]. Each row demonstrates the editing results of different methods, and the text conditions are listed on the left side of each row. *Combined Image* denotes the combination of the source image and corresponding eyeglasses mask for better visualization. $Ours^{w/o\ mask}$ denotes we ignore the effect of mask condition. Our approach successfully generated proper eyeglasses in all cases along with the least modification in irrelevant areas. It is noteworthy that comparison methods typically produce eyeglasses with identical shapes. In contrast, we can precisely control the shape of the eyeglasses using a binary mask.

image by utilizing a given semantic map, ultimately aligning the edited image with the semantic map. In the context of our task scenario, we aim to use the semantic map to control the shape and size of eyeglasses in the source image, in order to achieve flexible and precise virtual eyeglasses try-on. To evaluate our capability of controlling the eyeglasses shape, we select three semantic-based image editing methods for comparison: MaskGAN [51], SEAN [36], and S^2-Flow [54]. All the following comparison experiments are based on their pre-trained models.

The qualitative results are shown in Fig. 4. As illustrated in Fig. 4, our method demonstrates superior performance in terms of consistency in eyeglasses shape and preservation of identity information. MaskGAN [51] successfully generates eyeglasses consistent with the edited semantic mask, while the editing results are unnatural with a severe loss of identity. On the contrary, SEAN [36] better preserves the irrelevant areas and identity information while generating low-fidelity eyeglasses. Visually, S^2-Flow [54] has a better trade-off between generating high-fidelity eyeglasses and preserving irrelevant areas. However, it fails to generate eyeglasses consistent with the given semantic mask in some cases. Benefiting from the shape consistency loss, we successfully generate natural and high-fidelity eyeglasses consistent with the given binary mask, along with a slight loss of identity information.

Table I presents the results of our quantitative comparison experiment to evaluate our capability of preserving irrelevant areas and generating high-fidelity eyeglasses, compared to other semantic-based editing methods. Compared with other semantic-based editing methods, we achieve a significant improvement in SSIM, PSNR, and IDS metrics, which demonstrates our capability of preserving irrelevant areas and identity information. We also get the best results on mIoU and PA metrics, demonstrating the extraordinary capacity to generate high-fidelity eyeglasses consistent with the given semantic mask. In terms of FID, we still achieve comparable results, indicating that the distribution of editing results is similar to the source images. In total, we demonstrate a striking editing capability of preserving irrelevant areas and generating high-fidelity eyeglasses simultaneously.

Comparison to Text-based Editing Methods: The text-based editing methods aim to manipulate the source image based on a given textual description, ultimately aligning the edited image with the text description. To provide a more

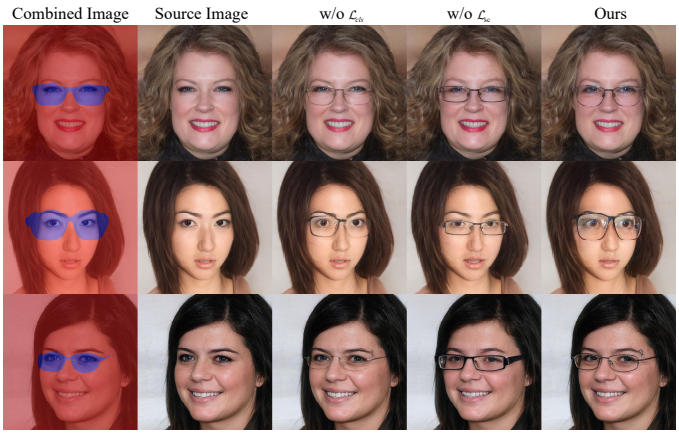


Fig. 6. Ablation analysis of $\mathcal{L}_{cls}$ and $\mathcal{L}_{sc}$. *Combined Image* denotes the combination of the source image and corresponding eyeglasses mask for better visualization. By incorporating both the classification loss ($\mathcal{L}_{cls}$) and the shape consistency loss ($\mathcal{L}_{sc}$), we can generate eyeglasses that are more complete and realistic.

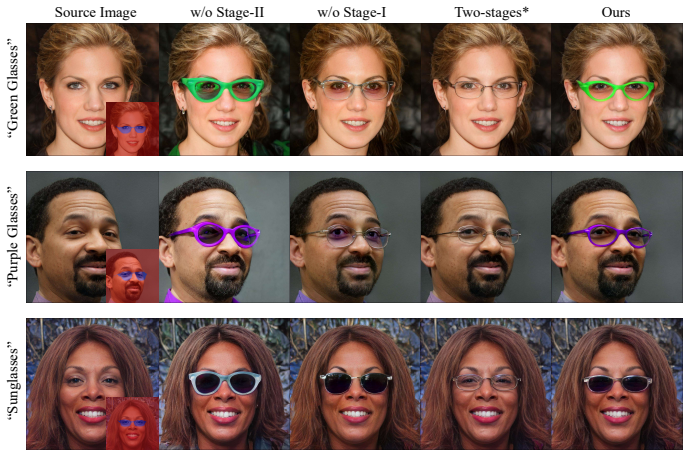


Fig. 7. Ablation analysis of the two-stage training scheme. *Two-stage** denotes we exchange the training orders in Stage-I. Only by using the two-stage training scheme, did we succeed to control the shape and style of the eyeglasses simultaneously.

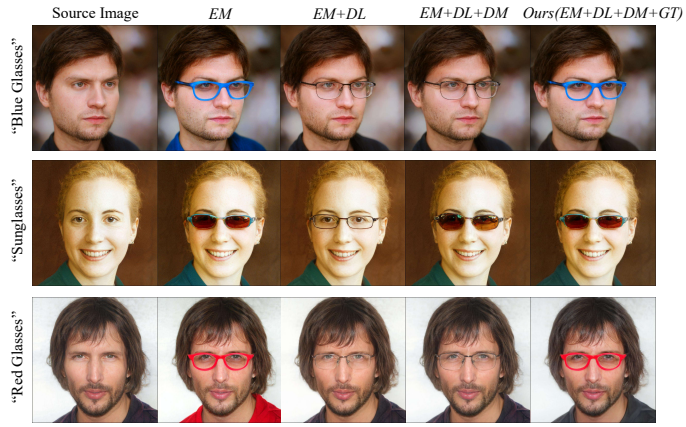


Fig. 8. Ablation analysis of the disentangled mapper and decoupling strategy. *EM, DL, DM,* and *GT* denote the editing mapper, disentangling loss, disentangled mapper, and gradient flow truncation, respectively. It indicates that both the disentangled mapper and decoupling strategy is necessary to preserve the irrelevant areas.

convenient and intuitive interactive process, we employ textual descriptions as conditions for modifying different styles of eyeglasses. We compare our model to four state-of-the-art text-guided image manipulation methods: StyleCLIP [13], PPE [14], CF-CLIP [15], and DeltaEdit[3] [18]. Since these compared methods do not involve mask conditions, we simply focus on the capabilities of modifying the eyeglasses style based on text conditions for a fair comparison, *i.e.*, we do not calculate mIoU and PA during quantitative comparison.

The qualitative results over CelebA-HQ dataset [24] are shown in Fig. 5, our method achieves more natural and realistic results while preserving the irrelevant areas to the greatest extent. StyleCLIP [13] successfully achieves most of the eyeglasses styles besides sunglasses, along with a significant modification on irrelevant areas, *e.g.*, hair, skin, and cloth. Focusing on disentangling editing, PPE [14] reduces changes in most of the irrelevant areas, while still failing in the sunglasses setting. CF-CLIP [15] achieves all the eyeglasses styles we demonstrated, but it may suffer from an excessive modification issue, e.g the lens is also modified in the setting of blue glasses and green glasses. HairCLIP [22] is originally designed for hair style transfer tasks, but with minor modifications, it can also be applied to eyeglass editing. Visually, while Hairclip is capable of achieving various styles of eyeglass editing, it falls short when editing metal eyeglasses, and there are also significant changes in irrelevant areas. DeltaEdit [18], which is designed to perform arbitrary text-guided image manipulation, fails to produce satisfactory results in any of our experimental settings, even when the editing scale is increased. This illustrates the challenge of manipulating eyeglasses based solely on text prompts. Meanwhile, all the comparison methods fail to preserve the cloth region when modifying the eyeglasses color. Thanks to the disentangled mapper and the simple decoupling strategy, we successfully preserve the cloth region in most cases, regardless of whether we incorporate or ignore the spatial information. Moreover, since we control the shape and style of eyeglasses in a single model, we mitigate the excessive modification phenomenon by keeping the balance of each eyeglasses style.

In Table II, we give the average quantitative comparison results of modifying eyeglasses styles, and more quantitative results are given in the supplementary material. As is evident, we achieve the best SSIM, PSNR and IDS scores, which are 6.28%, 12.89%, and 3.32% higher than the state of the arts [14], demonstrating our capability of preserving the irrelevant areas and identity information. All the quantitative results are consistent with our visual comparison, *i.e.*, our approach can not only generate natural and realistic eyeglasses in the editing results but also preserve the irrelevant areas to the greatest extent. Despite our approach not achieving the best performance in terms of CLIPScore, it is still comparable to the top results, indicating that we are able to preserve irrelevant areas to a significant extent without sacrificing text control capability. We also found that DeltaEdit [18] achieves the best

[3]In the comparison experiments, we multiply their editing scale since there are no significant alterations in the initial configuration.

TABLE III
QUANTITATIVE EXPERIMENTS OF ABLATION ANALYSIS. THOUGH OUR METHOD DID NOT ACHIEVE THE BEST RESULTS IN EACH METRIC, IT WAS COMPARABLE TO THE OPTIMAL RESULTS AND ACHIEVED A BALANCE BETWEEN ALIGNING WITH THE EYEGLASSES MASK AND TEXT DESCRIPTION AND PRESERVING IRRELEVANT REGIONS.

$\mathcal{L}_{cls}$	$\mathcal{L}_{sc}$	S-I	S-II	DL	DM	GT	PA(↑)	mIoU(↑)	SSIM(↑)	PSNR(↑)	IDS(↑)	CLIPScore(↑)
✓							0.9404	0.6977	0.8711	25.5300	0.6109	18.7031
	✓						0.9601	**0.7835**	**0.9182**	**29.0105**	**0.7358**	18.7343
✓	✓						**0.9617**	0.7790	0.9139	28.5235	0.6935	**19.0625**
✓	✓	✓					0.9146	0.6567	0.7746	16.7176	0.5609	**27.5312**
✓	✓		✓				**0.9541**	**0.7550**	0.8688	25.9677	**0.6679**	20.8125
✓	✓	✓*	✓*				0.9522	0.7455	**0.8930**	**27.3941**	0.6652	18.7503
✓	✓	✓	✓				0.9538	0.7529	0.8644	23.8862	0.6463	26.4843
✓	✓	✓	✓	✓			0.9603	0.7776	0.9114	28.5231	0.7073	19.0937
✓	✓	✓	✓	✓	✓		**0.9625**	**0.7849**	**0.9209**	**29.3569**	**0.7282**	19.1093
✓	✓	✓	✓	✓	✓	✓	0.9512	0.7509	0.8736	25.5683	0.6585	**26.3752**

FID scores, but it fails to produce visually appealing results in all our experimental settings. This suggests that FID may not be an appropriate metric for evaluating the manipulation capability of these methods, which is consistent with the findings in [22].

Discussion: Given the limited availability of dedicated and open-source methods specifically designed for eyeglass manipulation, we conducted detailed comparative experiments with existing methods [13]–[15], [18], [22], [36], [51], [54] that address broader issues of face manipulation. It can be observed that the semantic-based editing methods [36], [51], [54] cannot accurately control eyeglass shape conditioned on mask in many cases, while the text-based editing methods [13]–[15], [18], [22] fail with some eyeglasses styles and significantly affects irrelevant areas in some cases. This highlights the uniqueness and challenges of the eyeglass manipulation task, where existing methods cannot be easily transferred to this task and require special design. Our method performs well in aligning with different modal information and preserving irrelevant areas, providing a flexible and diverse glasses virtual try-on framework.

C. Ablation Analysis

To investigate the effects of loss functions, decoupling strategy, and two-stage training scheme on eyeglasses manipulation tasks, we conducted ablation studies that included both quantitative and qualitative analysis. As FID may not be suitable for evaluating manipulation tasks [22], we ignored the FID results in the quantitative experiments. For more intuitive understanding, we summarized the quantitative results in Table III. The specific roles of each component will be analyzed in the following sections, including mask-related loss functions, decoupling strategies, and two-stage training scheme.

Importance of Shape Consistency Loss and Classification Loss: During our two-stage training process, we use several loss functions to control the generated eyeglasses and preserve the irrelevant areas. We employ widely-used loss functions such as $\mathcal{L}_{nce}$, $\mathcal{L}_{id}$, $\mathcal{L}_{norm}$, and $\mathcal{L}_{bg}$, which have proven their effectiveness in aligning the manipulation results with the text description while keeping irrelevant regions unchanged. We introduce $\mathcal{L}_{cls}$ and $\mathcal{L}_{sc}$ to better control the shape and alignment of the eyeglasses.

As shown in the first part of Table III, the absence of $\mathcal{L}_{sc}$ leads to poorer mIoU and PA results, indicating that $\mathcal{L}_{sc}$ plays a crucial role in aligning the eyeglasses shape with the given mask. Note that excluding $\mathcal{L}_{cls}$ seems to get better metrics, such as higher mIoU, SSIM, PSNR, and IDS. However, visual examples in Fig. 6 demonstrate that dropping $\mathcal{L}_{cls}$ sometimes results in mutilated eyeglasses, indicating an insufficient editing magnitude of the original image in the eyeglasses region. Although it better preserves the irrelevant regions, it also yields results with lower fidelity. Therefore, both $\mathcal{L}_{cls}$ and $\mathcal{L}_{sc}$ are critical in generating more complete and realistic eyeglasses that are well-aligned with the given mask. Dropping either of these loss functions leads to noticeable issues, such as mutilated eyeglasses or misaligned eyeglasses.

Two-Stage Training Scheme: To verify the role of our two-stage training scheme, we compare the editing results using different training schemes: 1) We exclude Stage-II, where we do not jointly train our model after the preliminary training of Stage-I; 2) We exclude Stage-I, where we jointly train our model based on the mask and text conditions simultaneously; 3) We adopt another two-stage training scheme, where we exchange the training orders in Stage-I and use * to distinguish it in Table III. In other world, we first train our model based on the text conditions and then based on the mask conditions.

Visual examples in Fig. 7 and metrics results in the second part of Table III demonstrate the roles of each training stage and the positive impact of adopting a two-stage training strategy on the controllability of eyeglass shape and style. Firstly, when Stage-II is not utilized, all the editing results are well-aligned with the text descriptions, resulting in a higher CLIPScore. However, the eyeglass shapes differ from the given mask, resulting in lower mIoU and PA results and indicating a loss of control over the eyeglass shape. Secondly, when Stage-I is not utilized, all the editing results are better consistent with the eyeglass mask, resulting in the best mIoU and PA results. However, we fail to obtain satisfactory results that are well-aligned with the text description, indicating a loss of control over the eyeglass style. Thirdly, when we exchange the training orders of Stage-I, although we have a better ability to preserve irrelevant regions, it is still difficult to achieve the desired eyeglass style consistently with the text. As can be seen, it is difficult for all the variants to simultaneously

control the shape and style of eyeglasses based on mask and text, and aligning the manipulation results with one modality often results in the loss of control ability of the other modality. Nonetheless, our two-stage training scheme can strike a better balance between the control abilities of the two modalities, which leads to a better alignment with the eyeglasses mask and text descriptions.

Impact of Disentangled Mapper and Decoupling Strategy: To better preserve irrelevant areas, we propose a disentangled mapper and a simple decoupling strategy, *i.e.*, truncate the gradient flow computed by the disentangling loss. Therefore, we compare several variants to verify the effect of our disentangled mapper and decoupling strategy, as shown in Fig. 8 and the third part of Table III.

Solely with the editing mapper (i.e. adopting $\mathcal{L}_{cls}$, $\mathcal{L}_{sc}$ and two-stage training scheme), we successfully achieve satisfactory eyeglasses with desired shape and style, but there present significant modifications in the irrelevant areas, especially the cloth region. When including the disentangling loss, we better preserve the irrelevant areas and get higher SSIM, PSNR and IDS results, while failing in generating desired eyeglasses, indicating that the disentangling loss may impair the capability of the editing mapper to generate desired eyeglasses. In addition, when incorporating both disentangling loss and disentangled mapper, although the preservation of irrelevant regions is improved, we still encounter the same issue, where the generated eyeglasses significantly deviate from the given text, and the resulting CLIPScore is much lower. Combining the disentangled mapper and disentangling loss with gradient flow truncation, our approach both obtains diverse eyeglasses styles and preserves the irrelevant areas well, achieving a better balance between the controllability of different modalities (i.e. mask and text) and the preservation of irrelevant regions. In other words, by truncating the gradient flow, the editing mapper mainly focuses on controlling the eyeglasses style while the disentangled mapper mainly focuses on preserving the irrelevant areas, without affecting each other.

Overall, while our approach did not achieve the best performance across all metrics, it demonstrated comparable performance to the optimal results and achieved a balance between the alignment of different modalities and the preservation of irrelevant regions: Firstly, $\mathcal{L}_{cls}$ and $\mathcal{L}_{sc}$ improved the completeness and alignment of the eyeglasses, demonstrating a positive impact on improving the controllability of the mask; Later, two-stage training scheme effectively addressed the mutual interference between the two modalities and achieved a good balance between aligning with the eyeglasses mask and text description; Finally, the disentangled mapper and decoupling strategy maximally preserved the irrelevant regions without sacrificing the control ability of both mask and text. All the experimental results indicate that each component plays a crucial role in eyeglass manipulation tasks and effectively addresses some challenges in this field.

D. Exploring the Potential of Advanced Models in Eyeglasses Manipulation

With the rapid advancements in diffusion models, large text-to-image models [50], [59], [60] have demonstrated remarkable generation capabilities, producing stunning and high-quality images. Building upon these impressive generation abilities, researchers have explored their potential in the field of image editing, aiming to achieve more versatile and comprehensive image manipulation. Notable studies in this area include Prompt-to-Prompt [57] and SEGA [58]. Specifically, Prompt-to-Prompt utilizes the semantic relationship in the cross-attention layers between pixels and tokens to implement three editing modes: word swap for local editing, new phrase addition for global editing, and attention re-weighting for extent controlling. SEGA demonstrates how text interacts with the diffusion process and propose an approach to discover semantic directions in a single forward pass, allowing for local and global editing in composition and style. Given the notable progress in advanced manipulation models, we were motivated to investigate their potential in the specific task of eyeglasses virtual try-on.

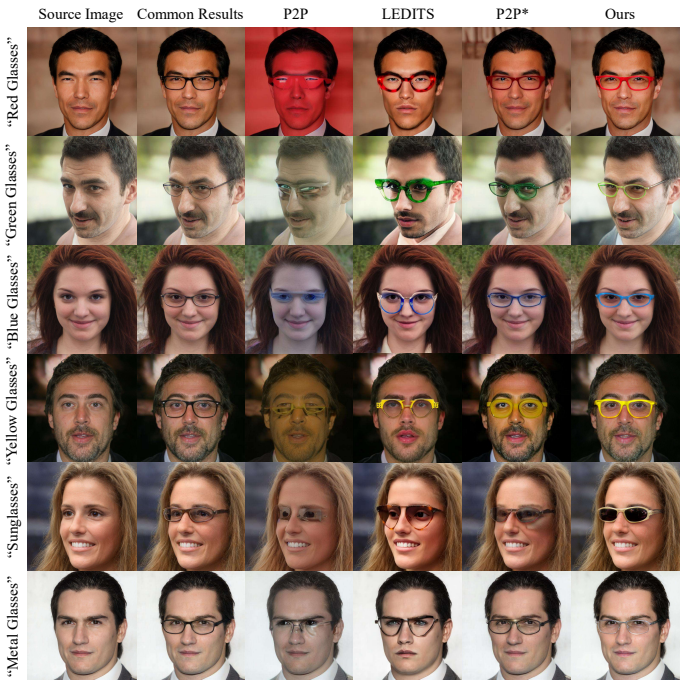


Fig. 9. Comparison with the advanced manipulation models. Existing advanced manipulation models [57], [58] fall short in fully addressing the specific and challenging task of eyeglass virtual try-on, while our method achieves superior results.

In order to achieve real image editing, we utilized Prompt-to-Prompt (combined with null-text inversion [61]) and LEDITS [62] (an expanded version of SEGA) in our experimental setting for eyeglasses manipulation. As illustrated in Fig. 9, when eyeglasses-related phrases were directly added in the text, Prompt-to-Prompt [57] always resulted in global editing and it was difficult to achieve the desired eyeglass styles. Compared to Prompt-to-Prompt, LEDITS [62] achieved better results; however, the generated eyeglasses were still unrealistic and resulted in a loss of identity information. To further explore the local editing ability of Prompt-to-Prompt, we used a simpler experimental setting, denoted as *P2P**, where we edited the existing eyeglasses in the source images, rather than

TABLE IV
INFERENCE TIME ANALYSIS. OUR APPROACH CAN ACHIEVE FLEXIBLE AND DIVERSE EYEGLASS VIRTUAL TRY-ON UNDER THE CONTROL OF BOTH MASK AND TEXT INPUTS WITHOUT INCURRING SIGNIFICANT INFERENCE TIME COSTS.

	mask control	text control	retraining per style	inference time (sec)
MaskGAN [51]	✓		no	**0.0199±0.0796**
SEAN [36]	✓		no	0.3602±0.0530
S^2-Flow [54]	✓		no	0.5378±0.1058
PPE [14]		✓	yes	0.0952±0.0278
StyleCLIP [13]		✓	yes	0.0976±0.0210
CF-CLIP [15]		✓	yes	0.1273±0.0272
HairCLIP [22]		✓	no	0.1171±0.0073
DeltaEdit [18]		✓	no	0.0994±0.0006
Ours	✓	✓	no	0.1168±0.0069

directly generating eyeglasses with a specified style[4]. As can be seen, under such a simpler experimental setting, Prompt-to-Prompt can maintain irrelevant areas and identity information well, but the edited eyeglasses were not entirely satisfactory, such as sunglasses, metal eyeglasses, and yellow eyeglasses. In contrast, our method can not only directly generate diverse and high-quality eyeglasses but also maintain irrelevant areas and identity information to a great extent.

Undoubtedly, advanced manipulation models [57], [58], [63] have extraordinary editing capabilities and can perform various types of editing on different types of images. However, for editing tasks that involve adding small objects, especially for eyeglass virtual try-on, existing methods are still far from satisfactory. In other words, eyeglass virtual try-on is a specific and challenging task that cannot be fully addressed by these advanced manipulation models. Although our method are not as versatile as these models, it can achieve better results for the specific task of eyeglass virtual try-on.

E. Inference Time Analysis

In this section, we conducted a comprehensive performance evaluation of our proposed GlassesCLIP and most comparison methods, specifically focusing on the inference time required for editing a single image. Specifically, we utilized an RTX 1080 GPU and measured the overall duration of the model, from input to output, for each test sample. We then computed the average inference time based on these measurements.

As demonstrated in Table IV, while our method may not be the fastest in terms of inference speed, it still achieves excellent results that are comparable to other methods [13]–[15], [18], [22]. It is important to note that our approach can manipulate both the eyeglass shape using the mask input and the eyeglass style using the text input, without necessitating retraining a new model for each style, showcasing the efficiency of our approach. In summary, following a single training session, our approach enables flexible and diverse eyeglass try-on, controlled by both mask and text inputs, without incurring significant inference time costs.

[4]Common results with regular eyeglasses were generated by our approach, where the differences compared to the source images may be visually negligible.

V. CONCLUSION

Focusing on eyeglasses virtual try-on, we propose GlassesCLIP, a text-guided eyeglasses manipulation method with spatial constraints, to control the eyeglasses shape and style based on the mask and text, which is intuitive and effective. Thanks to the new modulation module and two-stage training scheme, the mask conditions and text conditions control the eyeglasses simultaneously in a decoupling way, which is a more flexible and controllable interaction way. Furthermore, by utilizing the simple but effective decoupling strategy, we preserve the irrelevant areas to a great extent, resulting in better local editing. Quantitative and Qualitative Comparison and ablation analysis demonstrate the superiority of our approach to achieve diverse and realistic eyeglasses and preserve the irrelevant areas in the editing results.

Future research may focus on streamlining the training process, which currently necessitates separate training for each modality followed by joint training. Another potential avenue for improvement is the generalization to more challenging cases, such as highly regular eyeglasses shapes or unusual eyeglasses styles. The collection of additional extreme eyeglasses masks and unusual eyeglasses prompts may provide a viable solution.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers and the Associate Editor for their helpful suggestions and valuable comments. This work is supported by A*STAR Career Development Funding Award (Grant No: 222D800031).

REFERENCES

[1] M. Milanova and F. Aldaeif, "Markerless 3d virtual glasses try-on system," in *New Approaches for Multidimensional Signal Processing*, 2021, pp. 99–111.
[2] Z. Feng, F. Jiang, and R. Shen, "Virtual glasses try-on based on large pose estimation," *Procedia Computer Science*, vol. 131, pp. 226–233, 2018.
[3] D. Marelli, S. Bianco, and G. Ciocca, "Faithful fit, markerless, 3d eyeglasses virtual try-on," in *International Conference on Pattern Recognition*, 2021, pp. 460–471.
[4] ——, "A web application for glasses virtual try-on in 3d space," in *2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT)*, 2019, pp. 299–303.
[5] ——, "Designing an ai-based virtual try-on web application," *Sensors*, vol. 22, no. 10, p. 3832, 2022.
[6] X. Li, S. Zhang, J. Hu, L. Cao, X. Hong, X. Mao, F. Huang, Y. Wu, and R. Ji, "Image-to-image translation via hierarchical style disentanglement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8639–8648.
[7] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9243–9252.
[8] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9841–9850, 2020.
[9] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1532–1540.
[10] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.
[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[13] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2085–2094.

[14] Z. Xu, T. Lin, H. Tang, F. Li, D. He, N. Sebe, R. Timofte, L. Van Gool, and E. Ding, "Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 229–18 238.

[15] Y. Yu, F. Zhan, R. Wu, J. Zhang, S. Lu, M. Cui, X. Xie, X.-S. Hua, and C. Miao, "Towards counterfactual image manipulation via clip," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3637–3645.

[16] R. Paiss, H. Chefer, and L. Wolf, "No token left behind: Explainability-aided image classification and generation," in *European Conference on Computer Vision*, 2022, pp. 334–350.

[17] X. Hou, L. Shen, O. Patashnik, D. Cohen-Or, and H. Huang, "Feat: Face editing with attention," *arXiv preprint arXiv:2202.02713*, 2022.

[18] Y. Lyu, T. Lin, F. Li, D. He, J. Dong, and T. Tan, "Deltaedit: Exploring text-free training for text-driven image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6894–6903.

[19] X. Liu, C. Gong, L. Wu, S. Zhang, H. Su, and Q. Liu, "Fusedream: Training-free text-to-image generation with improved clip+gan space optimization," *arXiv preprint arXiv:2112.01573*, 2021.

[20] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–13, 2022.

[21] T. Zhao, T. Zhang, M. Zhu, H. Shen, K. Lee, X. Lu, and J. Yin, "Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations," *arXiv preprint arXiv:2207.00221*, 2022.

[22] T. Wei, D. Chen, W. Zhou, J. Liao, Z. Tan, L. Yuan, W. Zhang, and N. Yu, "Hairclip: Design your hair by text and reference image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 072–18 081.

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *6th International Conference on Learning Representations*, 2018.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[26] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *2012 Fourth International Workshop on Quality of Multimedia Experience*, 2012, pp. 37–38.

[27] B. Hu, Z. Zheng, P. Liu, W. Yang, and M. Ren, "Unsupervised eyeglasses removal in the wild," *IEEE Transactions on Cybernetics*, vol. 51, no. 9, pp. 4373–4385, 2020.

[28] Y.-H. Lee and S.-H. Lai, "Byeglassesgan: Identity preserving eyeglasses removal for face images," in *European Conference on Computer Vision*, 2020, pp. 243–258.

[29] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[30] R. Plesh, P. Peer, and V. Struc, "Glassesgan: Eyewear personalization using synthetic appearance discovery and targeted subspace modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 847–16 857.

[31] J. Sun, Q. Deng, Q. Li, M. Sun, M. Ren, and Z. Sun, "Anyface: Free-style text-to-face synthesis and manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 687–18 696.

[32] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Tedigan: Text-guided diverse face image generation and manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2256–2265.

[33] M. Zhu, J. Li, N. Wang, and X. Gao, "Learning deep patch representation for probabilistic graphical model-based face sketch synthesis," *International Journal of Computer Vision*, vol. 129, pp. 1820–1836, 2021.

[34] ——, "Knowledge distillation for face photo–sketch synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 893–906, 2020.

[35] ——, "A deep collaborative framework for face photo–sketch synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3096–3108, 2019.

[36] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5104–5113.

[37] Y. Wang, L. Qi, Y.-C. Chen, X. Zhang, and J. Jia, "Image synthesis via semantic composition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 749–13 758.

[38] J. Huang, L. Jing, Z. Tan, and S. Kwong, "Multi-density sketch-to-image translation network," *IEEE Transactions on Multimedia*, vol. 24, pp. 4002–4015, 2022.

[39] X. Hou, X. Zhang, Y. Li, and L. Shen, "Textface: Text-to-style mapping based face generation and manipulation," *IEEE Transactions on Multimedia*, vol. 25, pp. 3409–3419, 2023.

[40] L. Liang and X. Zhang, "Adaptive label propagation for facial appearance transfer," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3068–3082, 2019.

[41] S. Liu, R. Bao, D. Zhu, S. Huang, Q. Yan, L. Lin, and C. Dong, "Fine-grained face editing via personalized spatial-aware affine modulation," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[42] Y. Liu, Y. Chen, L. Bao, N. Sebe, B. Lepri, and M. De Nadai, "Isf-gan: An implicit style function for high-resolution image-to-image translation," *IEEE Transactions on Multimedia*, vol. 25, pp. 3343–3353, 2023.

[43] W. Zheng, Q. Li, X. Guo, P. Wan, and Z. Wang, "Bridging clip and stylegan through latent alignment for image editing," *arXiv preprint arXiv:2210.04506*, 2022.

[44] H. Wang, G. Lin, A. G. del Molino, A. Wang, Z. Yuan, C. Miao, and J. Feng, "Maniclip: Multi-attribute face manipulation from text," *arXiv preprint arXiv:2210.00445*, 2022.

[45] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426–2435.

[46] P. Chandramouli and K. V. Gandikota, "Ldedit: Towards generalized text guided image manipulation via latent diffusion models," in *33rd British Machine Vision Conference*, 2022.

[47] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 208–18 218.

[48] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[49] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *9th International Conference on Learning Representations*, 2021.

[50] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 674–10 685.

[51] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[52] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[53] T. Smith and J. Guild, "The cie colorimetric standards and their use," *Transactions of the optical society*, vol. 33, no. 3, p. 73, 1931.

[54] K. Singh, S. Schaub-Meyer, and S. Roth, "s^2-flow: Joint semantic and style editing of facial images," in *33rd British Machine Vision Conference*, 2022, p. 821.

[55] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local

nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6626–6637, 2017.

[57] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross-attention control," in *The Eleventh International Conference on Learning Representations*, 2023.

[58] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, and K. Kersting, "Sega: Instructing diffusion using semantic dimensions," *arXiv preprint arXiv:2301.12247*, 2023.

[59] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[60] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

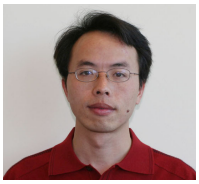
[61] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6038–6047.

[62] L. Tsaban and A. Passos, "Ledits: Real image editing with ddpm inversion and semantic guidance," *arXiv preprint arXiv:2307.00522*, 2023.

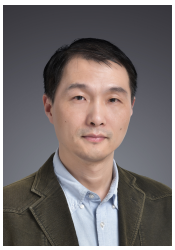
[63] I. Huberman-Spiegelglas, V. Kulikov, and T. Michaeli, "An edit friendly ddpm noise space: Inversion and manipulations," *arXiv preprint arXiv:2304.06140*, 2023.

Jiacheng Wang received the B.E. degree in electronic and information engineering, in 2020 from the Huazhong University of Science and Technology, Wuhan, China, where he is currently working toward the M.S. degree with the School of Electronic Information and Communications. His research interests include image generation and manipulation, computer vision, and machine learning.

Dr. Ping Liu (Senior Member, IEEE) serves as a Research Scientist at the Center for Frontier AI Research (CFAR) under A*STAR in Singapore. Prior to his tenure at CFAR, he was affiliated with the Center for Artificial Intelligence at the University of Technology Sydney. Dr. Liu earned his Bachelor's degree from Wuhan University of Technology, his Master's from Huazhong University of Science and Technology, both in Wuhan, China, and his Ph.D. from the Department of Computer Science and Engineering at the University of South Carolina, Columbia, SC, USA. He specializes in computer vision, machine learning, and deep learning methodologies.



Dr. Jingen Liu (Senior Member, IEEE) is an AI Researcher at Disney Streaming Advanced Research. Before joining Disney, he was a researcher at JD.com Silicon Valley Lab from 2018 to 2022, SRI International from 2011 to 2018, and a research fellow at the University of Michigan, Ann Arbor, from 2010 to 2011. He received his Ph.D. degree in 2009 from UCF. His research is focused on computer vision, multimedia, and machine learning.



Dr. Wei Xu (Member, IEEE) received the Ph.D.degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008. He is currently an Associate Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology. His research interests include automatic singing, multimedia, and machine learning.