Portrait Eyeglasses and Shadow Removal by Leveraging 3D Synthetic Data

Junfeng Lyu Zhibo Wang Feng Xu
School of Software and BNRist, Tsinghua University

Abstract

*In portraits, eyeglasses may occlude facial regions and generate cast shadows on faces, which degrades the performance of many techniques like face verification and expression recognition. Portrait eyeglasses removal is critical in handling these problems. However, completely removing the eyeglasses is challenging because the lighting effects (*e.g.*, cast shadows) caused by them are often complex. In this paper, we propose a novel framework to remove eyeglasses as well as their cast shadows from face images. The method works in a detect-then-remove manner, in which eyeglasses and cast shadows are both detected and then removed from images. Due to the lack of paired data for supervised training, we present a new synthetic portrait dataset with both intermediate and final supervisions for both the detection and removal tasks. Furthermore, we apply a cross-domain technique to fill the gap between the synthetic and real data. To the best of our knowledge, the proposed technique is the first to remove eyeglasses and their cast shadows simultaneously. The code and synthetic dataset are available at* https://github.com/StoryMY/take-off-eyeglasses.

Figure 1. Our method allows to remove eyeglasses and their shadows simultaneously. It produces photo-realistic results under various illuminations, head poses and eyeglasses with different shapes and textures.

1. Introduction

A large portion of people wears eyeglasses in their daily lives. In their face photos, eyeglasses usually bring unwanted occlusions and cast shadows on faces, which lead to inaccuracy in many useful techniques like image-based face verification [42, 46], expression recognition [47], fatigue detection [13, 17, 40], *etc*. Besides, in photography, removing eyeglasses from portraits could be needed for aesthetic reasons, giving users a choice to edit their portraits. Therefore, it is beneficial to develop an automatic technique for portrait eyeglasses removal.

However, completely removing eyeglasses suffers some key challenges. First, to recover the occluded facial region and keep it consistent with the remaining regions is a difficult task as facial skin has rich details and complex reflectance. Second, only recovering the occluded region cannot ensure visually convincing results as eyeglasses also bring various lighting effects (*e.g.*, cast shadows, reflections and distortions) on face regions. Explicitly modeling these effects is extremely difficult as the physical rules to generate these effects are complicated. And it requires a delicate perception of the eyeglasses geometry, face geometry and lighting conditions, which are also difficult to obtain from a single portrait.

Recently, deep learning [30, 44] has shown its great potential in handling tasks related to face editing [14, 33], and has been successfully applied to portrait eyeglasses removal [22] with the help of the face datasets [27, 35] containing eyeglasses labels. However, these techniques only focus on the eyeglasses but not the corresponding lighting effects. ByeGlassesGAN [32] constructs paired data con-

taining some lighting effects for training. However, as it uses 2D methods to synthesize the data, the quality and realism are quite limited. Also, it does not take cast shadows into consideration.

In this paper, we propose a novel eyeglasses removal technique using a synthetic dataset which considers 3D shadows and uses a cross-domain training strategy to fill the gap between synthetic and real data. This method jointly removes eyeglasses and their cast shadows, generating more visually plausible results compared to the previous state-of-the-art methods. In order to facilitate learning the relation between eyeglasses and cast shadows, we introduce a novel mask-guided multi-step network architecture for eyeglasses removal. The proposed network first detects two masks for both eyeglasses and their cast shadows. Then, the estimated masks are used as guidance in the multi-step eyeglasses removal. We observe that the shadows to be removed are caused by the eyeglasses, and we use this fact to carefully construct our network where the eyeglasses and shadows are handled in well-designed orders in both the detection and removal tasks. In this way, the network can well take eyeglasses as an important prior when dealing with the shadows.

For training this network, we build a photo-realistic synthetic dataset using high-quality face scans collected by [52] and 3D eyeglasses models made by artists, with principled BSDF [37] to achieve high rendering quality. This dataset contains a large amount of data for supervised training, covering various identities, expressions, eyeglasses, and illuminations. Another benefit of using the synthetic dataset is that we can synthesize images that cannot be captured in real world, i.e., images with eyeglasses but no shadows and images with shadows but no eyeglasses. These images can be used as intermediate supervisions to train the proposed network.

Although the accurate 3D information and the high-end rendering technique improve the photo-realism of our synthetic data, the network still cannot generalize well to real images due to the gap between the synthetic and real domain. Inspired by [23] and [49], we develop a cross-domain segmentation module that leverages a real image dataset to build a uniform domain for both the real and synthetic images. This helps to prevent the proposed network from using domain-specific information to detect eyeglasses and their cast shadows.

In summary, our main contributions are listed as follows:

- We design a novel mask-guided multi-step network architecture which is the first attempt in the literature to remove both eyeglasses and their cast shadows from portraits and achieves high realism.
- We present a high-quality synthetic portrait dataset which provides both intermediate and final supervisions for training eyeglasses/shadows detection and removal networks.
- We introduce a cross-domain segmentation module to enhance the generalization capability on real face images.

2. Related Works

Eyeglasses Removal. Early works [11, 38, 53, 54, 59] remove eyeglasses by statistical learning. The key assumption of these works is that the facial regions occluded by eyeglasses can be reconstructed from other faces without eyeglasses. However, these methods usually assume frontal faces and controlled environments, which limits their applications. Later works, *e.g.*, ERGAN [22] and ByeGlassesGAN [32], use deep neural networks in eyeglasses removal. ERGAN [22] proposes an unsupervised architecture for eyeglasses removal in the wild, while ByeGlassesGAN [32] manually constructs paired data and propose a multi-task framework for eyeglasses detection and removal. These methods can successfully remove eyeglasses in more general application scenarios. However, cast shadows caused by eyeglasses are often ignored in both methods as they do not explore the connections between eyeglasses and cast shadows. Unlike these methods, we found that by developing an architecture to learn this connection, the network can remove the eyeglasses and their cast shadows at the same time, generating more visually convincing results.

Face Attributes Manipulation. Facial image manipulation techniques [4, 14, 43] have been developed rapidly in recent years. Most of them jointly solve multi-label [9, 10, 19, 34, 55] or multi-style [3, 15, 24, 31, 64] issues. DFI [50] manipulates face attributes via interpolation of different feature vectors. AttGAN [19] manipulates facial images via attribute classification constraint and reconstruction learning. STGAN [34] incorporates difference attribute vector and selective transfer units (STUs) for arbitrary attribute editing. HiSD [33] proposes a hierarchical style disentanglement framework for image-to-image translation, which organizes the labels using a hierarchical tree structure and overcomes the disadvantages of previous joint methods [3, 41, 51, 56, 57, 60, 62]. Additionally, some works [12, 48] combine 3D Morphable Model (3DMM) with StyleGAN [28] to control facial images semantically. We found that manipulating external attributes (*e.g.*, hats or eyeglasses) is more difficult than manipulating internal attributes of faces as facial accessories often lead to occlusions or extra lighting effects (*e.g.*, cast shadows). Unlike the previous works, we focus on eyeglasses removal and aim to remove not only the eyeglasses but also their corresponding cast shadows.

Domain Adaptation for Segmentation. The majority of works in this task are usually designed for urban scenes. [21] combines both global and local alignment with a do-

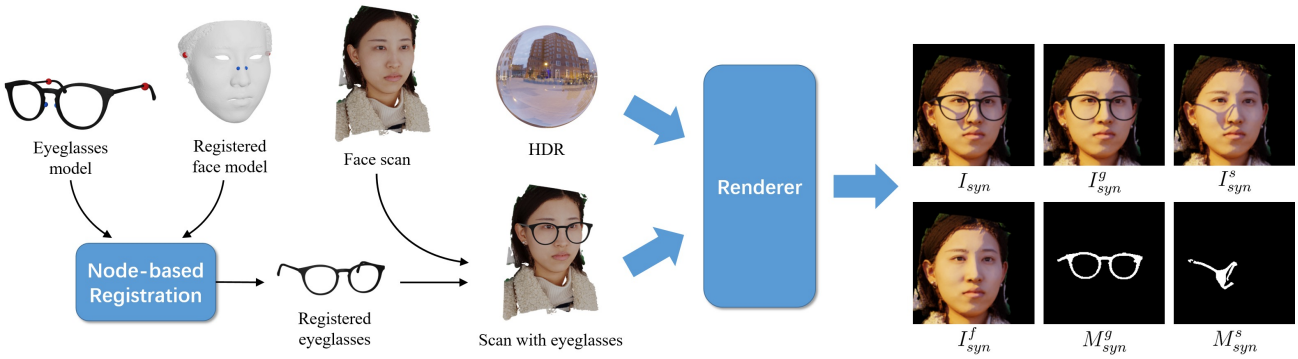

Figure 2. Illustration of portrait synthesis. We define two fix nodes (red), two floating nodes (blue) on the registered face model and their corresponding vertices on each eyeglasses model. With node-based registration, we compute a plausible pose to align the eyeglasses model with the face scan. Then, we combine them with a HDR lighting to render our synthetic data: I_{syn}, I^g_{syn}, I^s_{syn}, I^f_{syn}, M^g_{syn} and M^s_{syn}.

main adversarial training. [61] uses curriculum learning to address the domain adaptation. [8] proposes an unsupervised method to adapt segmenters across different cities. Other works [7,49] apply discriminators on the output space to align source and target segmentation, while [65] utilizes a conservative loss to naturally seek the domain-invariant representations. FDA [58] proposes a novel method that solves the domain adaptation via a simple Fourier Transform and its inverse. Based on the aforementioned methods, we additionally consider the relevance between the eyeglasses and cast shadows, and successfully bridge the gap between synthetic and real face images.

3. Portrait Synthesis with Eyeglasses

In order to build paired data for supervised training, we use 3D rendering to generate synthetic images. As shown in Fig. 2, we first make the face scan "wear" the 3D eyeglasses via node-based registration. Then, we render the scan with eyeglasses under a randomly chosen illumination. By setting the eyeglasses or their cast shadows to be visible or invisible, we can get four different types of rendered images. The masks of the eyeglasses and the cast shadows are also generated. Details are described as follows.

3.1. Data Preparation

For 3D face data, we directly use the dataset collected by [52]. This dataset contains the face scans of 438 subjects with 20 expressions, varying from male to female and young to old. In addition to raw scans, we also acquire the registered face models with the same topology. For 3D eyeglasses models, we ask professional artists to create 21 eyeglasses models, which contain various shapes and textures.

3.2. Eyeglasses Alignment

In order to put eyeglasses on the plausible positions of the face, we manually label four anchor nodes ($\boldsymbol{A}_i, i \in \{1,2,3,4\}$) on each eyeglasses model and their corresponding vertices ($\boldsymbol{V}_i, i \in \{1,2,3,4\}$) on the template face model used for registration. Specifically, these four nodes consist of two fixed nodes on the face temples and two floating nodes on both sides of the nose as shown in Fig. 2. Then, we compute the rotation $\boldsymbol{R} \in SO(3)$, the translation $\boldsymbol{t} \in \mathbb{R}^3$ and the scaling $s \in \mathbb{R}$ by minimizing the distance between the nodes and their corresponding vertices using Singular Value Decomposition [39], expressed as

$$E(\boldsymbol{R}, \boldsymbol{t}, s; \boldsymbol{A}_i, \boldsymbol{V}_i) = \sum_{i=1}^{4} \|s \cdot \boldsymbol{R}\boldsymbol{A}_i + \boldsymbol{t} - \boldsymbol{V}_i\|_2^2. \quad (1)$$

According to our observation, people put their eyeglasses on different nose positions. To enrich the wearing styles of our synthetic data, we define various candidate pairs of floating nodes in the nose region of the face template and randomly choose one pair for eyeglasses alignment. Also, we randomly change the color of eyeglasses to enrich their textures.

3.3. Rendering Setting

The data variety and photo-realism are well considered in the portrait rendering. In detail, we first collect 367 HDR lightings from *Poly Heaven*[1] to increase the diversity of illuminations. During the rendering, the lighting variation is further augmented by setting a random rotation of the global scene. Besides, we render each face scan which randomly "wears" a pair of eyeglasses by a random head pose. For photo-realistic synthesis, we use the principled BSDF implementation in Blender to render our synthetic data with the rendering setting empirically adjusted by a professional artist.

For each rendering sample, we render four kinds of images with different visibility combinations of the eyeglasses and their cast shadows: I_{syn}, I^g_{syn}, I^s_{syn}, I^f_{syn}. The eyeglasses mask M^g_{syn} and the shadow mask M^s_{syn} are also synthesized in the rendering.

[1] https://polyhaven.com/

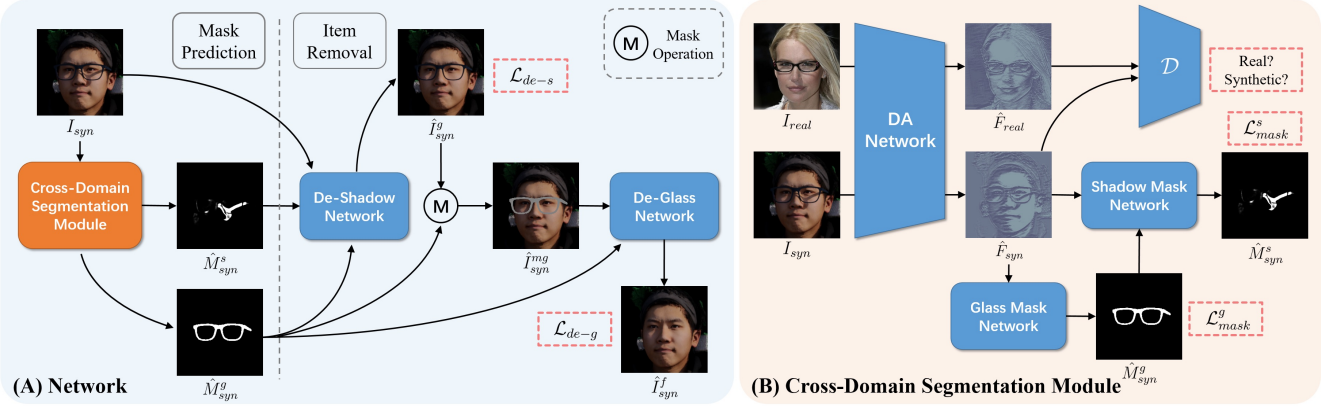

Figure 3. Illustration of the proposed network architecture. **(A)** Our network includes two stages: mask prediction stage and item removal stage. The mask prediction stage aims to estimate the eyeglasses mask and shadow mask via a cross-domain segmentation module. In the item removal stage, we successively employ a De-Shadow Network and a De-Glass Network to remove cast shadows and eyeglasses with the guidance of the two predicted masks. **(B)** In the cross-domain segmentation module, the Domain Adaptation (DA) Network normalizes the input images to uniform feature maps with the help of a discriminator. Then, the Glass Mask Network and Shadow Mask Network take the uniform feature maps to predict eyeglasses and shadows masks, respectively.

4. Portrait Eyeglasses Removal Network

The architecture of the proposed network is illustrated in Fig. 3. Our network is designed based on the following considerations: *1)* ByeGlassesGAN [32] improves eyeglasses removal with a parallel segmentation task, which has demonstrated the importance of mask prediction in eyeglasses removal. Inspired by their method, we remove the eyeglasses in a more natural way by first explicitly detecting the eyeglasses in a *mask prediction stage* and then removing eyeglasses with the guidance of predicted masks in an *item removal stage*. *2)* We further enhance the eyeglasses removal performance by using a multi-step strategy in both above stages to treat the eyeglasses and their cast shadows in sequential order. Considering that the shadows to be removed are caused by eyeglasses, the eyeglasses should be a guidance in both shadow mask prediction and shadow removal. *3)* The proposed network is trained to remove eyeglasses using the synthetic dataset. To make it generalized to real images, we use a Domain Adaptation (DA) Network to convert the input images into uniform feature maps. The uniform feature maps eliminate the domain-specific information to confuse a discriminator but retain the structural information for eyeglasses and shadow mask prediction.

4.1. Mask Prediction Stage

Given an input portrait I with eyeglasses, our method estimates the eyeglasses mask $\hat{M}^g$ and shadows mask $\hat{M}^s$ in the mask prediction stage using a cross-domain segmentation module. This module is composed of a DA Network, a Glass Mask Network and a Shadow Mask Network.

In order to tackle the gap between synthetic and real domain, the DA Network is trained to transfer the input image I to a uniform domain, outputting the uniform feature map $\hat{F}$. Inspired by [23] and [49], we apply adversarial learning to find the uniform domain assisted with a discriminator $\mathcal{D}$. This discriminator $\mathcal{D}$ is trained to distinguish whether the feature map $\hat{F}$ is from a real image or a synthetic image while the DA Network aims to fool the discriminator. We utilize LSGAN [36, 63] for more stable training:

$$\mathcal{L}^{\mathcal{D}}_{adv} = (\mathcal{D}(\hat{F}_{syn}))^2 + (\mathcal{D}(\hat{F}_{real}) - 1)^2, \tag{2}$$

$$\mathcal{L}^{\mathcal{G}}_{adv} = (\mathcal{D}(\hat{F}_{syn}) - 1)^2, \tag{3}$$

where $\hat{F}_{real}$ and $\hat{F}_{syn}$ are the corresponding feature maps of real and synthetic data. Specifically, the DA Network consists of the first layer of a pre-trained VGG encoder [45] with fixed parameters, combined with six trainable ResNet blocks [18] with instance normalization.

We use a multi-step strategy to predict the eyeglasses mask $\hat{M}^g$ and the corresponding shadow mask $\hat{M}^s$ from the uniform domain feature $\hat{F}$. Instead of extracting these two masks together using a single network, we first estimate the eyeglasses mask $\hat{M}^g$ using a Glass Mask Network. Then, the previous outputs $\hat{F}$ and $\hat{M}^g$ are together fed into a Shadow Mask Network to predict the shadow mask $\hat{M}^s$, with the consideration that the eyeglasses masks could be a guidance in the shadow mask prediction. We learn the eyeglasses mask M^g and the cast shadow mask M^s in a supervised manner as follows,

$$\mathcal{L}^g_{mask} = L_{\mathcal{E}}(M^g_{syn}, \hat{M}^g_{syn}), \tag{4}$$

$$\mathcal{L}^s_{mask} = L_{\mathcal{E}}(M^s_{syn}, \hat{M}^s_{syn}), \tag{5}$$

$$L_{\mathcal{E}}(M, \hat{M}) = -M \log \hat{M} - (1 - M) \log(1 - \hat{M}), \tag{6}$$

where $L_{\mathcal{E}}$ is the widely used binary cross entropy (BCE) loss. Experiments in Sec. 5.2 demonstrate that with the

guidance of the estimated eyeglasses mask $\hat{M}^g$, the predicted shadow mask $\hat{M}^s$ will be more complete.

Overall, the training loss for the mask prediction stage is formulated as

$$\begin{aligned}\mathcal{L}_{predict} = \lambda_{adv}\mathcal{L}^{\mathcal{D}}_{adv} + \lambda_{adv}\mathcal{L}^{\mathcal{G}}_{adv} \\ + \lambda_{mask}\mathcal{L}^{g}_{mask} + \lambda_{mask}\mathcal{L}^{s}_{mask},\end{aligned} \tag{7}$$

where λ_{adv} and λ_{mask} are the weights for adversarial learning and mask prediction, respectively.

4.2. Item Removal Stage

This stage aims to remove eyeglasses and cast shadows, and we call it *item removal stage* for short. It takes the two predicted masks as clues to achieve clean eyeglasses and shadow removal. When removing these items, we also apply the multi-step strategy. However, different from the multi-step setup used in the mask prediction stage, in which our method first handles eyeglasses and then shadows, we deal with eyeglasses and shadows in an inverse order in this stage. This is because if we first remove the eyeglasses, the network will lose the abundant indications of shadow intensity and locations.

With an input image I, we first use a De-Shadow Network to remove the cast shadows of the eyeglasses. To help the network better locate the cast shadows to be removed, the estimated eyeglasses mask $\hat{M}^g$ and shadow mask $\hat{M}^s$ are also fed to the De-shadow Network. In order to learn the shadow-removed image I^g, we employ a L_1 regression loss, written as

$$\mathcal{L}_{de-s} = \|\hat{I}^g_{syn} - I^g_{syn}\|_1, \tag{8}$$

where $\hat{I}^g_{syn}$ indicates the output of our De-Shadow Network.

After removing the cast shadows, we use a De-Glass Network to further remove the eyeglasses in the next step. The large variety of eyeglasses textures in real world will lower the performance of eyeglasses removal. To enhance the robustness of our method, we adopt a mask operation to set the pixel values of the eyeglasses regions to $\mathbf{0}$. This operation eliminates the texture of eyeglasses from $\hat{I}^g$, forcing the De-Glass Network to remove the eyeglasses only according to the structure instead of textures. Finally, the De-Glass Network takes the masked shadow-removed result $\hat{I}^{mg}$ and the estimated eyeglasses mask $\hat{M}^g$ as input and learns the eyeglasses-removed image I^f via the following constraint:

$$\mathcal{L}_{de-g} = \|\hat{I}^f_{syn} - I^f_{syn}\|_1. \tag{9}$$

where $\hat{I}^f_{syn}$ represents the output of our De-Glass Network.

To sum up, the training loss for the item removal stage is formulated as

$$\mathcal{L}_{remove} = \lambda_{de-s}\mathcal{L}_{de-s} + \lambda_{de-g}\mathcal{L}_{de-g}, \tag{10}$$

where λ_{de-s} and λ_{de-g} are the weights for shadow and eyeglasses removal, respectively.

5. Experiments

In this section, we first describe the datasets and our implementation details. Then, we compare our method with the state-of-the-art eyeglasses removal and image-to-image translation methods qualitatively and quantitatively. Finally, we evaluate the key contributions of the proposed method via ablation study. Note that besides the results in Fig. 1, we will show more various results in our supplementary material.

Dataset. We use our synthetic dataset described in Sec. 3 and CelebA [35] to train the proposed network. For synthetic dataset, we randomly sample 73 identities of the 438 identities. Each identity contains 20 face scans with different expressions. We combine the face scans randomly with 5 eyeglasses and 4 HDR lightings, finally generating 29,200 training samples. CelebA is a real-world portrait dataset that contains 202,599 face images of 10,177 identities and is annotated with 5 landmarks and 40 binary attributes for each image. Using the attributes labels, we split 13,193 images with eyeglasses and 189,406 images without eyeglasses from it. Additionally, we adopt FFHQ [28] and MeGlass [16] for testing. FFHQ contains 70,000 high-quality portraits and it also covers accessories like eyeglasses. Using face parsing [1], we roughly split 11,778 images with eyeglasses from it. MeGlass is a dataset containing 1,710 identities and each identity has images with and without eyeglasses. This dataset is essential for identity preservation validation in Sec. 5.1.2. We refer to [28] to align all the images to a size of 256×256 using facial landmarks.

Implementation Details. Our method is implemented with PyTorch. We use Adam optimizer [29] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is 0.0001 and the batch size is 8. For the weights in the objective functions in Eq. (7) and Eq. (10), we set $\lambda_{adv} = 0.1$, $\lambda_{mask} = 1$, $\lambda_{de-s} = 1$, and $\lambda_{de-g} = 1$. Apart from the DA Network and the discriminator, all the other networks utilize the architecture in [26]. In practice, we first train the cross-domain segmentation module for 30 epochs and fix it when training networks in the item removal stage, which needs 80 epochs. The total training process costs about two days on a single GTX 1080 GPU.

5.1. Comparison with State-of-the-art Methods

We compare our method with state-of-the-art eyeglasses removal methods: ERGAN [22] and ByeGlassesGAN [32], as well as image-to-image translation methods including CycleGAN [63], StarGAN [9], ELEGANT [57], pix2pix [25] and HiSD [33]. To ensure fair comparisons, all these methods and our method are not trained on the testing dataset. Specifically, to compare with ERGAN and HiSD,

Figure 4. Qualitative cross-dataset results of different methods on FFHQ dataset (top) and MeGlass (bottom).

we directly use their released models which are trained on CelebA and CelebA-HQ [27], respectively. For CycleGAN, StarGAN and ELEGANT, we train them on the task of eyeglasses removal using their codes and the CelebA dataset. As pix2pix needs paired data, we train it on our synthetic data using the released code. As we cannot reach the authors of ByeGlassesGAN [32] to conduct a comparison experiment, we just show qualitative comparison using the images posted in their paper. Note that the purpose of the comparisons is not to purely compare different methods in the same setting but to demonstrate which solution better solves the problem.

5.1.1 Qualitative Comparison

We first compare the visual quality of our method with prior works on various images from FFHQ and MeGlass, covering different ages, genders, head poses, illuminations, eyeglasses shapes and textures. As shown in Fig. 4, our method achieves the best quality compared to the previous works. ELEGANT fails to remove the frames of the eyeglasses on all the test images. ERGAN can remove eyeglasses, but the eyeglasses regions are always blurred. CycleGAN, StarGAN and pix2pix preserve the high-frequency details in the whole eyeglasses regions, but they cannot completely remove eyeglasses for some samples. HiSD seems competitive to ours on some easy samples, but it fails to remove sharp cast shadows (1st row) as well as eyeglasses with unusual shape (5th row) and texture (2nd row). Benefiting from the mask-guided learning and our synthetic data, our method can remove various eyeglasses and the corresponding cast shadows. In addition, it generates photo-realistic contents in the regions occluded by eyeglasses or shadows, and retains the consistency with the global illumination and the skin texture of the surrounding regions. For ByeGlassesGAN [32], we only perform the comparison using the images posted in their paper. Results are shown in Fig. 5 and we can see that our method outperforms theirs in the shadow removal.

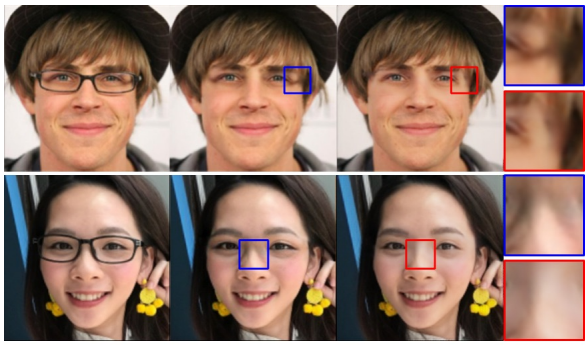

Figure 5. Qualitative comparison with ByeGlassesGAN [32] using images from their paper.

5.1.2 Quantitative Results

For quantitative comparisons, we first use Fréchet Inception Distance (FID) [20] to evaluate the realism of the generated images. Then, we apply face recognition technique to evaluate the ability of identity preservation. Finally, we adopt a user study to further evaluate the visual quality of eyeglasses removal.

Realism. First, we process the images with eyegalsses in FFHQ by different methods. Then, we compute the FID between the eyeglasses-removed results and the images without eyeglasses in FFHQ. The results (Tab. 1, 1st col) show that our method is competitive with HiSD and outperforms the other methods. This indicates that images generated by ours and HiSD are probably close to the real images without eyeglasses. Note that realism is a subjective measurement that can not be fully represented by FID. For further evaluation, we adopt a **user study** later.

Identity Preservation. To evaluate the identity preservation ability, we use some metrics commonly used in face recognition [5, 6], including the True Accept Rate at False Accept Rate (TAR@FAR) and Rank-1. To compute these metrics, we first collect 1,227 image triplets from MeGlass dataset. Each triplet contains three images of the same identity: two without eyeglasses and one with eyeglasses. Then, we input the image with eyeglasses into different methods to acquire corresponding eyeglasses-removed results. Finally, we select the first eyeglasses-free image in the triplet as the gallery and all the other images as probes to compute the metrics based on a pre-trained face recognition network [2]. As shown in Tab. 1, the second eyeglasses-free image in the triplet (*noglass*) achieves high face recognition accuracy as it is a real image containing full identity information. However, the accuracy will degrade when taking the images with eyeglasses as the probe (*glass*), indicating the negative effects of eyeglasses in face recognition. ERGAN, CycleGAN, ELEGANT and pix2pix lead to the further degradation of face recognition after eyeglasses removal while StarGAN and HiSD enhance the metrics. Our method exhibits the most significant increase, which stands for the best ability of eyeglasses removal and identity preservation.

	FID↓	MOS↑	TAR@FAR↑ $1e^{-2}$	TAR@FAR↑ $1e^{-3}$	Rank-1↑
glass	-	-	0.6025	0.3349	0.3716
ERGAN [22]	38.61	1.10	0.2839	0.1005	0.1439
CycleGAN [63]	38.10	2.21	0.5856	0.3186	0.3431
ELEGANT [57]	43.13	1.12	0.3531	0.1507	0.1862
StarGAN [9]	40.93	1.51	0.6435	0.3773	0.4107
HiSD [33]	**26.74**	3.17	0.6329	0.3757	0.3903
pix2pix [25]	41.42	1.52	0.5687	0.3015	0.3422
Ours	26.89	**4.43**	**0.6702**	**0.4315**	**0.4621**
noglass	-	-	0.8295	0.6430	0.6625

Table 1. Quantitative results of different methods.

User Study. A user study is conducted to further evaluate the visual quality of eyeglasses removal. In detail, we combine the results of different methods together with the input image to construct a "question". Participants are asked to give their opinions based on the visual quality, scoring different results from 1 to 5 (1 for the worst, 5 for the best). In total, we invite 40 participants and each participant is asked to answer 20 randomly sampled "questions". As shown in Tab. 1, our method has the highest Mean Opinion Score (MOS), indicating the superiority of our technique.

5.2. Ablation Study

In this subsection, we evaluate the performance of our key contributions in the mask prediction stage and the item removal stage.

Mask Prediction. We first conduct ablation studies for the mask prediction stage. The first ablation removes the DA Network with two new segmentation networks trained on synthetic data only and tested on real data directly (*w/o DA*). Results in Fig. 6 show that without domain adaptation, the estimated eyeglasses masks are sometimes incomplete and thus lead the shadow mask prediction to produce even worse results. The second ablation removes the multi-step strategy (in mask prediction) by using a single network to estimate the masks of eyeglasses and shadows together (*w/o multi-step*). With the help of the DA network, the eyeglasses masks are properly estimated. However, as the eyeglasses masks can not help the shadow masks estimation in the single-step setting, the estimated shadow masks still have noticeable artifacts. To further evaluate our assumption that the eyeglasses mask can guide the task of shadow mask prediction as shadows are caused by eyeglasses, we further conduct another ablation setting where the shadow mask is first predicted and then used as guidance in the eyeglasses mask prediction (*SM-guided GM*). Its results show that this multi-step setup will lead to worse shadow mask estimation. This further indicates the correctness of our as-

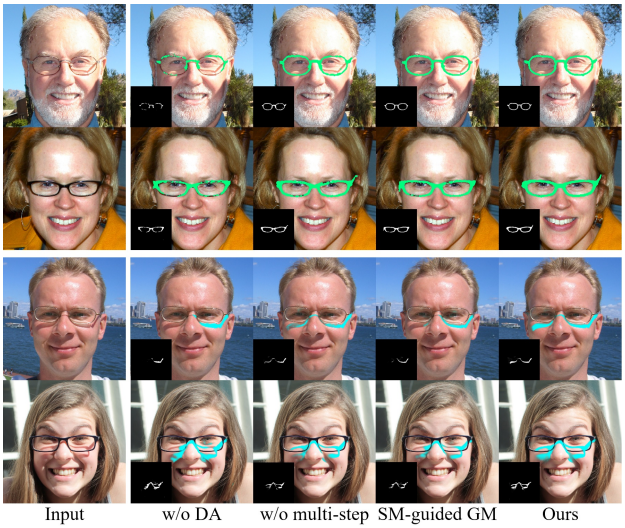

Figure 6. Visualization of eyeglasses masks (green) and shadow masks (blue) of different ablations in the mask prediction stage.

	FID↓	TAR@FAR↑ $1e^{-2}$	TAR@FAR↑ $1e^{-3}$	Rank-1↑
w/o DA	27.45	0.6463	0.3977	0.4392
w/o multi-step	27.18	0.6683	0.4262	0.4458
SM-guided GM	27.30	0.6641	0.4201	0.4523
GM-guided SM (ours)	**26.89**	**0.6702**	**0.4315**	**0.4621**
w/o SM	33.89	0.6586	0.3989	0.4327
w/o GM	42.80	0.6567	0.3846	0.4221
w/o multi-step	28.66	0.6675	0.4197	0.4498
De-Glass First	29.58	0.6590	0.4115	0.4417
De-Shadow First (ours)	**26.89**	**0.6702**	**0.4315**	**0.4621**

Table 2. Quantitative comparison of different ablations in the mask prediction stage (top) and item removal stage (bottom).

sumption, and the order of the two tasks is important due to the causality between eyeglasses and shadows.

Item Removal. Here, we evaluate the effect of mask guidance and the multi-step strategy (in item removal) by comparing different ablation settings. We first train two ablation settings without using the shadow mask or the eyeglasses mask (*w/o SM* and *w/o GM*), respectively. We also remove eyeglasses and shadows using one network to construct the third ablation setting (*w/o multi-step*). Similar to the mask prediction stage, we also invert the order of De-Shadow and De-Glass Network to get the fourth setting (*De-Glass First*). Qualitative results in Fig. 7 obviously show that *w/o SM* is weak at shadow removal while *w/o GM* fails to remove the complete eyeglasses. Besides, *w/o multi-step* and *De-Glass First* also have noticeable degradation compared to the proposed method. Quantitative results in Tab. 2 also manifest the advantages of the proposed method.

6. Limitations

Extensive experiments have shown that the proposed method achieves promising performance on real-world images across age, gender, head pose, illumination and eyeglasses. However, it currently does not perform well on images with extreme head pose or eyeglasses with colored lenses as shown in Fig. 8. A large head pose often results in extreme lens distortion, which is expensive to simulate in the portrait synthesis. Eyeglasses with colored lenses, *e.g.* sunglasses, are still difficult to remove due to the complete occlusions of eyes. A possible solution is to add more samples of these cases into the training dataset, which will be included in our future work.

Figure 7. Qualitative results of different ablations in the item removal stage.

Figure 8. Limitations. Extreme head pose with effects of lenses (left) and colored lenses (right). These cases are difficult for most of the existing methods. Here, we only show comparisons to the most competitive method (HiSD).

7. Conclusion

In this paper, we introduce a novel eyeglasses removal technique that first detects and then removes the eyeglasses using the mask-guided multi-step network architecture. To our best knowledge, the proposed method is the first attempt to remove the eyeglasses and their cast shadows simultaneously from a single portrait. Besides, we build a high-quality synthetic portrait dataset, which provides intermediate and final supervisions. In order to fill the gap between the synthetic and real domain, we apply the cross-domain segmentation module to predict the masks of eyeglasses and their cast shadows from a uniform domain for removal guidance. Both qualitative and quantitative experiments demonstrate that our method better preserves the original identity and achieves high realism on real portraits.

Acknowledgements. We thank SenseTime Group Limited for providing computing resources. This work was supported by Beijing Natural Science Foundation (JQ19015), the NSFC (No.61727808, 62021002), the National Key R&D Program of China 2018YFA0704000. This work was supported by the THUIBCS and BLBCI. Feng Xu is the corresponding author.

References

[1] https://github.com/zllrunning/face-parsing.PyTorch. 5

[2] https://github.com/timesler/facenet-pytorch. 7

[3] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2018. 2

[4] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6713–6722, 2018. 2

[5] Lacey Best-Rowden, Hu Han, Charles Otto, Brendan F Klare, and Anil K Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security*, 9(12):2144–2157, 2014. 7

[6] Jui-Shan Chan, Gee-Sern Jison Hsu, Hung-Cheng Shie, and Yan-Xiang Chen. Face recognition by facial attribute assisted network. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3825–3829, 2017. 7

[7] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2019. 3

[8] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017. 3

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 5, 7

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[11] M. De Smet, R. Fransens, and L. Van Gool. A generalized em approach for 3d model based face recognition under occlusions. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1423–1430, 2006. 2

[12] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020. 2

[13] Kartik Dwivedi, Kumar Biswaranjan, and Amit Sethi. Drowsy driver detection using representation learning. In *2014 IEEE international advance computing conference (IACC)*, pages 995–999. IEEE, 2014. 1

[14] Yue Gao, Fangyun Wei, Jianmin Bao, Shuyang Gu, Dong Chen, Fang Wen, and Zhouhui Lian. High-fidelity and arbitrary face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16115–16124, 2021. 1, 2

[15] Jingtao Guo, Zhenzhen Qian, Zuowei Zhou, and Yi Liu. Mulgan: Facial attribute editing by exemplar. *arXiv preprint arXiv:1912.12396*, 2019. 2

[16] Jianzhu Guo, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Face synthesis for eyeglass-robust face recognition. *arXiv preprint arXiv:1806.01196*, 2018. 5

[17] Mehdi Hajinoroozi, Zijing Mao, and Yufei Huang. Prediction of driver's drowsy and alert states from eeg signals with deep learning. In *2015 IEEE 6th international workshop on computational advances in multi-sensor adaptive processing (CAMSAP)*, pages 493–496. IEEE, 2015. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[19] 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*, 28(11):5464–5478, Nov 2019. 2

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[21] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[22] Bingwen Hu, Zhedong Zheng, Ping Liu, Wankou Yang, and Mingwu Ren. Unsupervised eyeglasses removal in the wild. *IEEE Transactions on Cybernetics*, 2020. 1, 2, 5, 7

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2, 4

[24] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017. 5, 7

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 5

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1, 6

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2, 5

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[31] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[32] Yu-Hui Lee and Shang-Hong Lai. Byeglassesgan: Identity preserving eyeglasses removal for face images. In *European Conference on Computer Vision*, pages 243–258. Springer, 2020. 1, 2, 4, 5, 6, 7

[33] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8639–8648, June 2021. 1, 2, 5, 7

[34] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 1, 5

[36] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 4

[37] Stephen McAuley, Stephen Hill, Naty Hoffman, Yoshiharu Gotanda, Brian Smits, Brent Burley, and Adam Martinez. Practical physically-based shading in film and game production. In *ACM SIGGRAPH 2012 Courses*, pages 1–7. 2012. 2

[38] Jeong-Seon Park, You Hwa Oh, Sang Chul Ahn, and Seong-Whan Lee. Glasses removal from facial image using recursive error compensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):805–811, 2005. 2

[39] Long Quan and Zhongdan Lan. Linear n-point camera pose determination. *IEEE Transactions on pattern analysis and machine intelligence*, 21(8):774–780, 1999. 3

[40] Akshay Rangesh, Bowen Zhang, and Mohan M. Trivedi. Driver gaze estimation in the real world: Overcoming the eyeglass challenge. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1054–1059, 2020. 1

[41] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1

[43] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4030–4038, 2017. 2

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[46] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 1

[47] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013. 1

[48] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2

[49] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 2, 3, 4

[50] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017. 2

[51] Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, and Luis Herranz. Sdit: Scalable and diverse cross-domain image translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1267–1276, 2019. 2

[52] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Trans. Graph.*, 39(6), Nov. 2020. 2, 3

[53] Wai Keung Wong and Haitao Zhao. Eyeglasses removal of thermal image based on visible information. *Information Fusion*, 14(2):163–176, 2013. 2

[54] Chenyu Wu, Ce Liu, Heung-Yueng Shum, Ying-Qing Xy, and Zhengyou Zhang. Automatic eyeglasses removal from face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):322–336, 2004. 2

[55] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y. Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. *2019 IEEE/CVF*

International Conference on Computer Vision (ICCV), pages 5913–5921, 2019. 2

[56] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Dna-gan: Learning disentangled representations from multi-attribute images. *arXiv preprint arXiv:1711.05415*, 2017. 2

[57] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–187, September 2018. 2, 5, 7

[58] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 3

[59] Dong Yi and Stan Z Li. Learning sparse feature for eyeglasses problem in face recognition. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 430–435. IEEE, 2011. 2

[60] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *Advances in Neural Information Processing Systems*, 2019. 2

[61] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 2020–2030, 2017. 3

[62] Shuchang Zhou, Taihong Xiao, Yi Yang, Dieqiao Feng, Qinyao He, and Weiran He. Genegan: Learning object transfiguration and attribute subspace from unpaired data. *arXiv preprint arXiv:1705.04932*, 2017. 2

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 4, 5, 7

[64] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017. 2

[65] Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 568–583, 2018. 3