MagGAN: High-Resolution Face Attribute Editing with Mask-Guided Generative Adversarial Network

Yi Wei[1], Zhe Gan[2], Wenbo Li[3], Siwei Lyu[4], Ming-Ching Chang[1], Lei Zhang[2], Jianfeng Gao[2], and Pengchuan Zhang[2]

[1] University at Albany, State University of New York, USA
[2] Microsoft Corporation, Redmond, USA
[3] Samsung Research America AI Center, USA
[4] University at Buffalo, State University of New York, USA

Abstract. We present *Mask-guided Generative Adversarial Network* (MagGAN) for high-resolution face attribute editing, in which semantic facial masks from a pre-trained face parser are used to guide the fine-grained image editing process. With the introduction of a mask-guided reconstruction loss, MagGAN learns to only edit the facial parts that are relevant to the desired attribute changes, while preserving the attribute-irrelevant regions (*e.g.*, hat, scarf for modification 'To Bald'). Further, a novel mask-guided conditioning strategy is introduced to incorporate the influence region of each attribute change into the generator. In addition, a multi-level patch-wise discriminator structure is proposed to scale our model for high-resolution (1024 × 1024) face editing. Experiments on the CelebA benchmark show that the proposed method significantly outperforms prior state-of-the-art approaches in terms of both image quality and editing performance.

1 Introduction

The demand of face editing is booming in the era of selfies. Both the research community, *e.g.*, [4,6,9,15,16,17,21,25,29,32,36,37,40,44], and the industry, *e.g.*, Adobe and Meitu, have extensively explored to improve the automation of face editing by leveraging user's specification of various facial attributes, *e.g.*, hair color and eye size, as the conditional input. Generative Adversarial Networks (GANs) [7] have made tremendous progress for this task. Prominent examples in this direction include AttGAN [9], StarGAN [6], and STGAN [25], all of which use an encoder-decoder architecture, and take both source image and target attributes (or, attributes to be changed) as input to generate a new image with the characteristic of target attributes.

Although promising results have been achieved, state-of-the-art methods still suffer from inaccurately localized editing, where regions irrelevant to the desired attribute change are often edited. For instance, STGAN [25] can make undesired editing by painting the scarf to white for "Pale Skin" (left) and the hat to

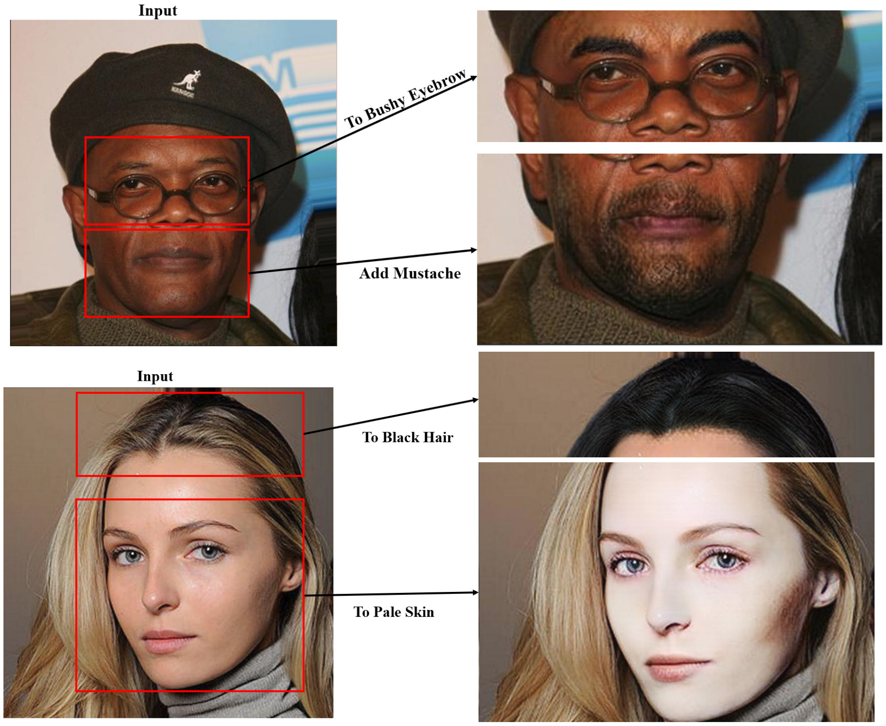

Fig. 1: Visual results of MagGAN on resolution 1024 × 1024. The specific subregions are cropped for better visualization

golden for "Blond Hair" (right) (see Figure 2). Solution to this problem requires notions of relevant regions that are editable *w.r.t.* the facial attribute edit types, while keeping the non-editable regions intact. To illustrate this concept of region-localized attribute editing, we refer to the facial regions that are editable when a specific attribute changes as *attribute-relevant* regions (such as the hair region for "To Blonde"). Regions that should not be edited (such as the hat and other non-hair regions for attribute "To Bald") are referred to as *attribute-irrelevant*. Ideal attribute editing generator will only edit attribute-relevant regions while keeping attribute-irrelevant regions intact, to minimize artifacts. The second issue of most existing methods is that they only work with images of low resolutions (128 × 128). How to edit facial attributes of high-resolution (1024 × 1024) images is less explored.

In order to address these challenges, we present the **Mask-guided Generative Adversarial Network** (MagGAN) for high-resolution face attribute editing. The proposed approach is built upon STGAN [25], which uses a difference attribute vector as conditional input, and a selective transfer unit for attribute editing. Based on this, a soft segmentation mask of common face parts from a pre-trained face parser is used to achieve fine-grained face editing. On one hand, the facial mask provides useful geometric constraints, which helps generate realistic face images. On the other hand, the mask also identifies each facial com-

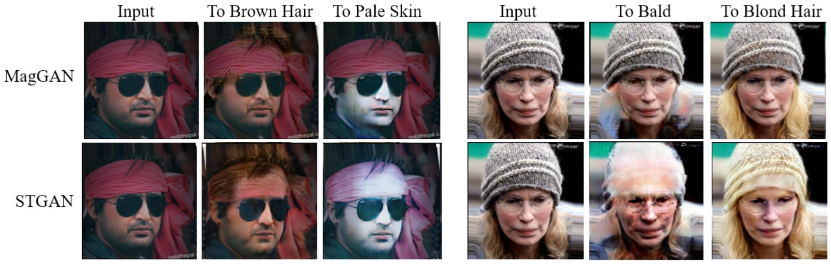

Fig. 2: MagGAN (1st row) can effectively apply accurate attribute editing while keeping attribute-irrelevant regions (*e.g.*, hat, scarf) intact. In comparison, the state-of-the-art STGAN [25] (2nd row) produces undesired modifications on these regions, *e.g.*, whitening the scarf while manipulating "Pale Skin"

ponent (*e.g.*, eyes, mouth, and hair), which is necessary for accurately localized editing. With the introduction of a mask-guided reconstruction loss, MagGAN can effectively focus on regions that are most related to the edited attributes, and keep the attribute-irrelevant regions intact, thus generating photo-realistic outputs.

Another reason why existing methods cannot preserve the regions that should not be edited is about how the attribute change information is injected into the generator. Although most attribute changes lead to localized editing, the attribute change condition itself does not explicitly contain any spatial information. In order to better learn the alignment between attribute change and regions to edit, MagGAN further uses a novel mask-guided conditioning strategy that can adaptively learn *where to edit*.

To further scale our model for high-resolution (1024 × 1024) face editing (see Figure 1 for visual results), we propose to use a series of multi-level patch-wise discriminators. The coarsest-level discriminator sees the full downsampled image, and is responsible for judging the global consistency of generated images, while a finer-level discriminator only sees patches of the generated high-resolution image, and tries to classify whether these patches are real or not. Empirically, this leads to more stable model training for high-resolution face editing.

The main contributions of this paper are summarized as follows. (*i*) We propose MagGAN that can effectively leverage semantic facial mask information for fine-grained face attribute editing, via the introduction of a mask-guided reconstruction loss. (*ii*) A novel mask-guided conditioning strategy is further introduced to encourage the influenced region of each target attribute to be localized into the generator. (*iii*) A multi-level patch-wise discriminator structure scales up our model to deal with high-resolution face editing. (*iv*) State-of-the-art results are achieved on the CelebA benchmark, outperforming previous methods in terms of both visual quality and editing performance.

2 Related Work

The development of face editing techniques evolves along the automation of editing tools. In the early stage, researchers focused on developing attribute-

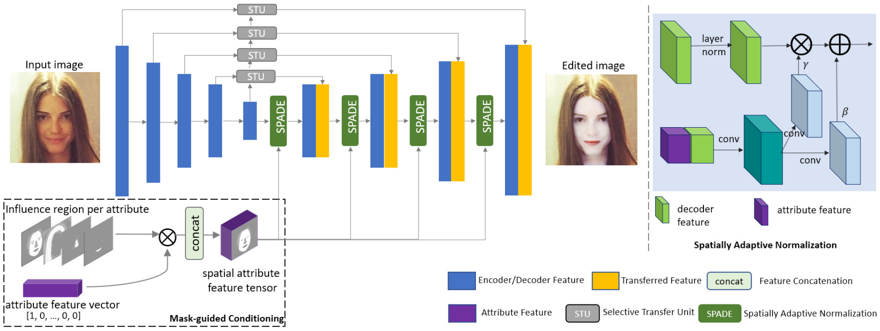

Fig. 3: Model architecture for the proposed Mask-guided GAN (MagGAN)

dedicated methods for face editing [3,22,26,33,34,43], *i.e.*, each model is dedicated to modifying a single attribute. However, such dedicated methods suffer from low automation level, *i.e.*, not being able to manipulate multiple attributes in one step. To this end, many works [6,9,15,16,17,21,25,29,32,36,37,40,44] started using attribute specifications, *i.e.*, semantically meaningful attribute vectors, as conditional input. Multiple attributes can be manipulated via changing the input attribute specifications. This work belongs to this category. Another line of works [5,30,35,38,46] improve the automation level of the face editing model by providing an exemplar image as the conditional input. Below, we briefly review recent attribute-specification based methods, and refer the readers to [45] for more details of methods that are not reviewed herein.

Many facial attributes are local properties (such as hair color, baldness, etc), and facial attribute editing should only change relevant regions and preserve regions not to be edited. StarGAN [6] and CycleGAN [29] introduced the cycle-consistency loss to conditional GAN so as to preserve attribute-irrelevant details and to stabilize training. AttGAN [9] and STGAN [25] found that the reconstruction loss of images not to be edited is at least as good as the cycle-consistency loss for preserving attribute-irrelevant regions. STGAN [25] proposed the selective transfer units to adaptively select and modify encoder features for enhanced attribute editing, achieving state-of-the-art performance on editing success rate. However, in this paper, we show that neither the cycle-consistency loss nor the reconstruction loss is sufficient to well preserve regions not to be edited (see Figure 2), and propose to utilize masks to solve this problem.

Semantic mask/segmentation provides geometry parsing information for image generation, see, *e.g.*, [12,31,23]. Semantic mask datasets and models are available for domains with important real applications, such as face editing [18,19] and fashion [24]. Recently, both [8] and [19] utilize mask information for facial image manipulation, where *a target/manipulated mask* is required in the manipulation process. In this paper, we focus on the setting of editing with attribute specifications, without requiring a target/manipulated mask. We only make use of a pre-trained face parser, instead of requiring users to provide the mask manually.

3 MagGAN

As illustrated in Figure 3, face editing is performed in MagGAN via an encoder-decoder architecture [9,6]. The design of Selective Transfer Units (STUs) in STGAN [25] is adopted to selectively transform encoder features according to the desired attribute change. Inspired by StyleGAN [14,31], the adaptive layer normalization [2,11] is used to inject conditions through the de-normalization process, instead of directly concatenating the conditions with the feature map. Our full encoder-decoder generator is denoted as:

$$\widehat{\mathbf{x}} = G(\mathbf{x}, \mathbf{att}_{\text{diff}}), \quad \mathbf{att}_{\text{diff}} = \mathbf{att}_t - \mathbf{att}_s, \tag{1}$$

where $\mathbf{x}$(or $\widehat{\mathbf{x}}$) $\in \mathbb{R}^{3\times H\times W}$ denote the input (or edited) image; $\mathbf{att}_s$(or $\mathbf{att}_t$) $\in \mathbb{R}^C$ are the source (or target) attributes. The generator takes the attribute difference $\mathbf{att}_{\text{diff}} \in \mathbb{R}^C$ as input, following [25].

3.1 Avoid editing attribute-irrelevant regions

Although notable results have been achieved, existing work still suffers from inaccurately localized editing, where irrelevant regions unrelated to the desired attribute change are often made. For example, in Figure 2, STGAN [25] changes the scarf to white for "Pale Skin" (left), and changes the hat to golden for "Blond Hair" (right).

We leverage facial regions for effective facial attribute editing and modeling as a solution. We utilize a pre-trained face parser to provide soft facial region masks. Specifically, a modified BiseNet [39] trained on the CelebAMask-HQ dataset [20] [5] is used to generates 19-class region masks, including various facial components and accessories. For each attribute a_i, we define its *influence regions* represented by two probability masks $M_i^+, M_i^- \in [0,1]^{H\times W}$. If attribute a_i is strengthened during editing, the region characterized by M_i^+ is likely to be changed; if a_i is weakened, the region characterized by M_i^- is likely to be changed. For example, for "Pale Skin", both M_i^+ and M_i^- characterize the "skin" region; for "Bald", M_i^+ characterizes the "hair" region while M_i^- characterizes the region consisting of "background, skin, ears" and "ear rings". In this setup, we propose the following Mask-aware Reconstruction Error (MRE) to measure the *preserving quality* of the editing process (in preserving irrelevant regions that shall not be edited):

$$\text{MRE} = \frac{1}{HWC} \sum_{i=1}^{C} \left\| (1 - M_i^{\text{sgn}(\mathbf{att}_{\text{diff},i})})(G(\mathbf{x}, \mathbf{att}_{\text{diff},i}\mathbf{e}_i) - \mathbf{x}) \right\|_1, \tag{2}$$

where $\mathbf{att}_{\text{diff},i}$ is the i'th entry of $\mathbf{att}_{\text{diff}}$, and $\mathbf{e}_i$ is the vector with i'th entry 1 and all others 0, $M_i^{\text{sgn}(\mathbf{att}_{\text{diff},i})} \in \{M_i^+, M_i^-\}$. In the face editing experiments, since all attributes are binary and $\mathbf{att}_s \in \{0,1\}^C$, we take the attribute change vector $\mathbf{att}_{\text{diff}} := 1 - 2\mathbf{att}_s$. In this case, the image preservation error is computed when *only one* attribute is flipped each time, and MRE is the total error.

In § 4, we will report MRE for various previous methods and our models in Table 3. Existing approaches of both the cycle-consistency loss used in StarGAN [6] and the reconstruction loss in [9,25] are insufficient to preserve the regions that shall not be edited.

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

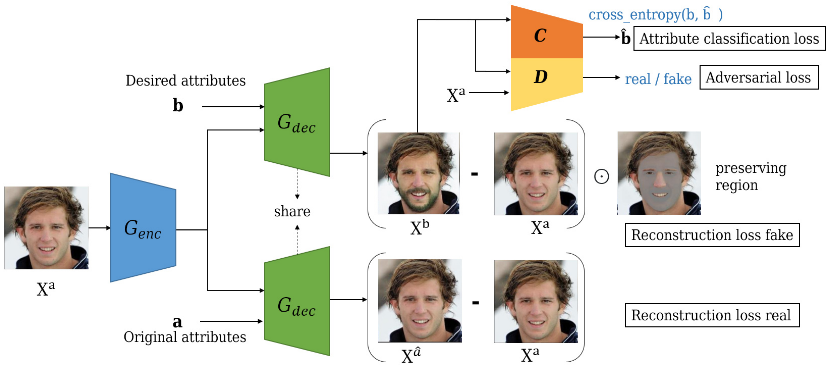

Fig. 4: MagGAN loss function design (§ 3.2). For better illustration, the preserving region is denoted by the non-grey region of human face

3.2 Loss functions for model training

We aim to optimize MagGAN regarding the following four aspects: (i) preservation accuracy for regions that should be preserved; (ii) reconstruction error of the original image; (iii) attribute editing success; and (iv) synthesized image quality. Therefore, we design four respective types of loss functions for MagGAN training, as illustrated in Figure 4.

Mask-guided reconstruction loss. Continue from the design of MRE (2), we propose the following mask-guided reconstruction loss:

$$L_G^{\mathrm{mre}} = \left\| M(\mathbf{att}_{\mathrm{diff}}, \mathbf{x}) \cdot (\mathbf{x} - G(\mathbf{x}, \mathbf{att}_{\mathrm{diff}})) \right\|_1 , \tag{3}$$

where $M(\mathbf{att}_{\mathrm{diff}}, \mathbf{x}) \in [0, 1]^{H \times W}$ is a probability mask of the regions to be preserved.

The preserved mask $M(\mathbf{att}_{\mathrm{diff}}, \mathbf{x})$ is computed from both the attribute difference $\mathbf{att}_{\mathrm{diff}}$ and the probability facial mask $\mathbf{M}$ of image $\mathbf{x}$. We first feed image $\mathbf{x}$ into a face parser, and obtain a probability map $\mathbf{M} \in [0, 1]^{19 \times H \times W}$ of the 19 facial parts, where $\sum_{i=1}^{19} \mathbf{M}_{i,h,w} = \mathbf{1}_{h,w}$. Since the semantic relationship between facial attributes and facial parts can be reasonably assumed to be constant, we explicitly define two binary relation matrices $\mathbf{AR}^+$ and $\mathbf{AR}^-$, the *attribute-part matrices* with dimension $C \times 19$, to characterize the relation between them. The i-th row of matrix $\mathbf{AR}^+$ or $\mathbf{AR}^-$ indicates which facial parts should be modified when the i-th attribute is strengthened, *i.e.*, $\mathbf{att}_{\mathrm{diff},i} > 0$, or weakened, *i.e.*, $\mathbf{att}_{\mathrm{diff},i} < 0$. Note that, if facial part has no explicit relationship with one attribute, the corresponding matrix entry of $\mathbf{AR}^+$,$\mathbf{AR}^-$ could be set to 0.

To obtain M, we first gather all parts $\mathbf{AR}^* \in [0, 1]^{19}$ that are possibly influenced by attribute change $\mathbf{att}_{\mathrm{diff}}$, as,

$$\mathbf{AR}^* = \min\left\{1, \left(\mathbf{att}_{\mathrm{diff}}^{(+)}\right)^T \mathbf{AR}^+ + \left(\mathbf{att}_{\mathrm{diff}}^{(-)}\right)^T \mathbf{AR}^-\right\} , \tag{4}$$

where $\mathbf{att}_{\mathrm{diff}}^{(+)} = (\mathbf{att}_{\mathrm{diff}} > 0)$ and $\mathbf{att}_{\mathrm{diff}}^{(-)} = (\mathbf{att}_{\mathrm{diff}} < 0)$. Finally,

$$M_{h,w}(\mathbf{att}_{\mathrm{diff}}, \mathbf{x}) = \mathbf{1} - \sum_{i=1}^{C} \mathbf{M}_{i,h,w} * \mathbf{AR}_i^* . \tag{5}$$

The influence regions M_i^+ and M_i^- in (2) can also be computed this way, with $\mathbf{att}_{\text{diff}} = \mathbf{e}_i$ and $\mathbf{att}_{\text{diff}} = -\mathbf{e}_i$.

Reconstruction loss. Image reconstruction can be considered as a sub-task of image editing, because the generator should reconstruct the image when no edit is applied, $\mathbf{att}_{\text{diff}} = \mathbf{0}$. Therefore, the reconstruction loss is defined as

$$\mathcal{L}_G^{\text{rec}} = \|G(\mathbf{x}, \mathbf{0}) - x\|_1, \tag{6}$$

where the ℓ_1 norm is adopted to preserve the sharpness of the reconstructed image.

GAN loss for enhancing image quality. The synthesized image quality is enhanced by the generative adversarial networks, where we use an unconditional image discriminator D_{adv} to differentiate real images from edited images. In particular, a Wasserstein GAN (WGAN) [1] is utilized:

$$\mathcal{L}_{D_{\text{adv}}} = \mathbb{E}_{\widehat{\mathbf{x}}}[D_{\text{adv}}(\widehat{\mathbf{x}})] - \mathbb{E}_{\mathbf{x}}[D_{\text{adv}}(\mathbf{x})] + \lambda\, \mathbb{E}_{\mathbf{x}_{\text{int}}}[(\|\nabla_{\mathbf{x}_{\text{int}}} D_{\text{adv}}(\mathbf{x}_{\text{int}})\|_2 - 1)^2], \tag{7}$$

where $\widehat{\mathbf{x}}$ is the generated image and $\mathbf{x}_{\text{int}}$ is sampled along lines between the latent space of pairs of real and generated image.

The generator G, instead, tries to fool the discriminator by synthesizing more realistic images:

$$\mathcal{L}_G^{\text{gan}} = -\,\mathbb{E}_{\mathbf{x},\mathbf{att}_{\text{diff}}}[D_{\text{adv}}(G(\mathbf{x}, \mathbf{att}_{\text{diff}}))]. \tag{8}$$

Attribute classification loss. To ensure that the edited image indeed has the target attribute $\mathbf{att}_t$, an attribute classifier D_{att} is trained on the ground-truth image attribute pairs $(\mathbf{x}, \mathbf{att}_s)$ with the standard cross-entropy loss:

$$\mathcal{L}_{D_{\text{att}}} = \mathbb{E}_{\mathbf{x}}[KL(D_{\text{att}}(\mathbf{x}), \mathbf{att}_s)]\,. \tag{9}$$

The generator is trying to generate images that maximize its probability to be classified with the target attribute $\mathbf{att}_t$:

$$\mathcal{L}_G^{\text{cls}} = -\,\mathbb{E}_{\mathbf{x},\mathbf{att}_{\text{diff}}}[KL(D_{\text{att}}(G(\mathbf{x}, \mathbf{att}_{\text{diff}})), \mathbf{att}_t)]\,. \tag{10}$$

In summary, the loss to train the MagGAN generator G is

$$\mathcal{L}_G = L_G^{\text{gan}} + \lambda_1 \mathcal{L}_G^{\text{rec}} + \lambda_2 \mathcal{L}_G^{\text{cls}} + \lambda_3 L_G^{\text{mre}}. \tag{11}$$

In experiments, we always take $\lambda_1 = 100$ and $\lambda_2 = 10$. We vary λ_3 to examine the effect of our proposed mask-guided reconstruction loss.

3.3 Mask-guided conditioning in the generator

Another reason why the previous methods cannot preserve the regions that shall not be edited is about how the attribute change information is injected into the generator. Although most attribute changes should lead to localized editing, the attribute change condition $\mathbf{att}_{\text{diff}} \in \mathbb{R}^C$ does not explicitly contain any spatial information. In STGAN [25] (and other previous works for face attribute editing), this condition is replicated to have the same spatial size of some hidden feature tensor, and then concatenated to it in the generator. For example, in the SPADE block in Figure 3 (Right), $\mathbf{att}_{\text{diff}}$ is replicated spatially to be $\mathbf{Att}_{\text{diff}} \in \mathbb{R}^{C\times H\times W}$ (the purple block)[6], and then concatenated to the decoder feature (the green

[6] We use $\mathbf{att} \in \mathbb{R}^C$ to denote attributes without spatial dimension and $\mathbf{Att} \in \mathbb{R}^{C\times H\times W}$ for attributes with spatial dimensions.

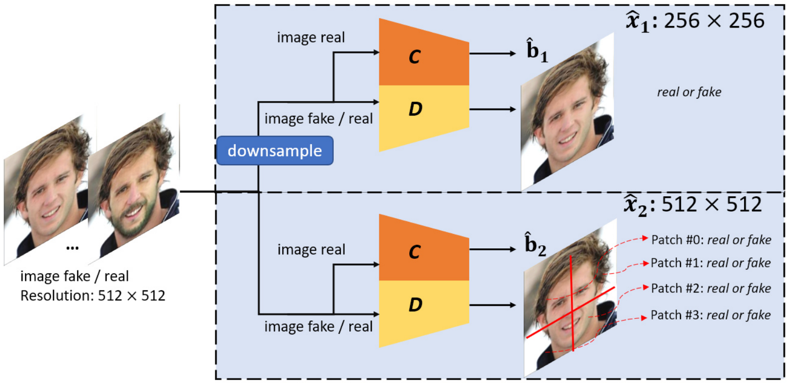

Fig. 5: Illustration of multi-level patch-wise discriminators

block). It is hoped that the generator will learn by itself the localized property of attribute editing from this concatenated tensor. However, in practice, this is insufficient, even with the mask-guided reconstruction loss (3).

We propose to inject this inductive bias that the influence region of each attribute change is localized into the generator directly, by making use of masks. We view the i-th channel of $\mathbf{Att}_{\text{diff}}$, denoted as $\mathbf{Att}_{\text{diff}}^{(i)} \in \mathbb{R}^{H \times W}$, as the condition to edit attribute a_i. In previous work, $\mathbf{Att}_{\text{diff}}^{(i)} = \mathbf{att}_{\text{diff},i}\mathbf{1}$ that is uniform across the spatial dimension. Specifically, we propose:

$$\mathbf{Att}_{\text{diff}}^{(i)} = \mathbf{att}_{\text{diff},i} M_i^{\text{sgn}(\mathbf{att}_{\text{diff},i})}, \tag{12}$$

where M_i^+ and M_i^- are the influence regions of attribute a_i defined in (2). We illustrate this mask-guided conditioning process in Figure 3 (bottom-left). Finally, we simply replace the original replicated tensor with the mask-guided attribute condition tensor, and obtain *a generator with mask-guided conditioning*. Note that this mask-guided conditioning technique is generally applicable to both generators with and without SPADE.

The blending trick is another simple approach to preserve the attribute-irrelevant regions. More specifically, with the probability mask of attribute-irrelevant regions $M(\mathbf{att}_{\text{diff}}, \mathbf{x})$ defined in (3), we simply add a linear layer at the end of the generator:

$$\widehat{x} = M(\mathbf{att}_{\text{diff}}, \mathbf{x}) * x + (1 - M(\mathbf{att}_{\text{diff}}, \mathbf{x})) * G(\mathbf{x}, \mathbf{att}_{\text{diff}}). \tag{13}$$

This blending trick improves our MagGAN performance in terms of MRE, but visually it introduces sharp transitions at the boundary of regions to be preserved. Therefore, we do not include this trick in our final MagGAN. More discussions are in Supplementary.

3.4 Multi-level patch-wise discriminators for high-resolution face editing

We describe our approach to scale up image editing in high resolutions. First of all, we empirically found that a single "shallow" discriminator cannot learn some global concepts, such as Male/Female, leading to low editing success. On the other hand, a single "deep" discriminator makes the adversarial training very unstable, leading to low image quality.

Inspired by PatchGAN [12] and several multi-level generation works [41,42,13], we propose to use a series of multi-level patch-wise "shallow" discriminators, as illustrated in Figure 5, for high-resolution face editing. The architecture of the discriminators are exactly the same without sharing weights. The coarsest-level discriminator (D_1) see the full downsampled image, and is responsible for global consistency in the image generation. The attribute classifier C_1 associated with it is effective in attribute classification, as in the low-resolution image editing case. The finer-level discriminators (D_2, etc.) see patches of the generated high-resolution image instead of the full one, and determine whether these patches are real or not. To maintain an unified architecture for discriminators across different levels, we still associate the finer-level discriminator with a classifier (C_2), which takes the average pooled feature as input for classification. The total loss for all PatchGAN discriminators are defined as:

$$\mathcal{L}_D = \frac{1}{P}\sum_{i=1}^{P}\left(\mathcal{L}_{D^i_{\text{att}}} + \mathcal{L}_{D^i_{\text{adv}}}\right), \tag{14}$$

where D^i_{att}, D^i_{adv} denote the attribute classifier and image discriminator of the ith PatchGAN discriminator, P is the number of total discriminators. In practice, we found these finer-level discriminators improve the editing performance.

Note that our generator only generates high-resolution images, which can be directly downsampled to lower resolutions and fed to coarse-level discriminators. On the contrary, generators in previous works [41,42,13] generate a high-resolution image in a multi-stage manner for the sake of training stability. They generate low-resolution images as intermediate outputs, which are fed to coarse-level discriminators. Our approach is simple in comparison, and we did not observe any training stability issue.

4 Experiments

Dataset and pre-processing. We use CelebA dataset [27] for evaluation. CelebA contains over 200K facial images with 40 binary attribute labels for each image. To apply CelebA to high-resolution face editing, we process the original web images by cropping, aligning and resizing into 1024 × 1024. When loading images for editing, they are re-scaled to match the target resolution. The images are divided into the training set, validation set and test set. Following the repository of STGAN[7], we take 637 images from the validation set to assess the training process. We use the rest of the validation set and the training set to train our model. The test set (nearly 20K) is used for evaluation. We consider 13 distinctive attributes including: *Bald*, *Bangs*, *Black Hair*, *Blond Hair*, *Brown Hair*, *Bushy Eyebrows*, *Eyeglasses*, *Male*, *Mouth Slightly Open*, *Mustache*, *No Beard*,

[7] STGAN: https://github.com/csmliu/STGAN

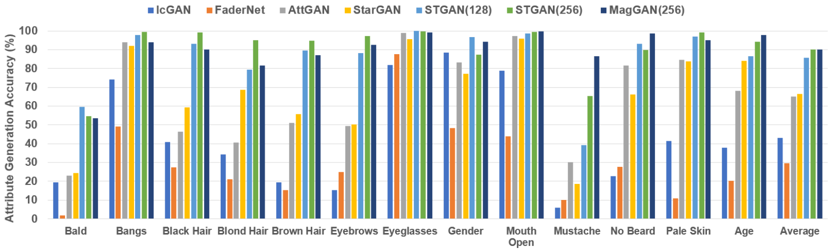

Fig. 6: Facial attribute editing accuracy of IcGAN [32], FaderNet [16], StarGAN [6], AttGAN [9], STGAN [25], STGAN(256) and our model MagGAN(256) (from left to right in rainbow colors in order). The last two models naming with "(256)" are the ones with image resolution 256 that are resized into 128 for evaluation

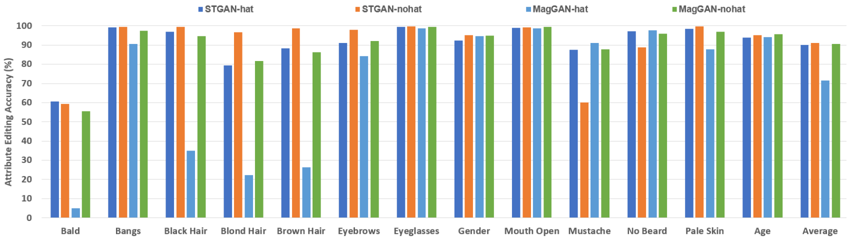

Fig. 7: Facial attribute editing accuracy of STGAN and MagGAN on hat samples and non hat samples of resolution 256 × 256

Pale Skin and *Young*. Since most images in CelebA have lower resolution than 1024×1024, our "high-resolution" MagGAN models are not exactly trained with true high-resolution images. However, our results show the ability of MagGAN scale up to 1024×1024 resolution.

MagGAN exploits the information of facial masks, which are obtained using a pre-trained face parser with 19 classes (as mentioned in § 3.1). Instead of taking a multi-label hard mask, we take the probability of each class as soft masks with smooth boundaries, which leads to improved generation quality. All the facial masks are stored in resolution 256×256. The two attribute-part relation matrices $\mathrm{AR}^+, \mathrm{AR}^- \in [0,1]^{13\times 19}$ described in § 3.2 characterize the relation between each edit attribute and corresponding facial component changes. Detailed definitions are in Supplementary.

Quantitative evaluation. The performance of attribute editing are measured in three aspects, *i.e.*, (*i*) mask-aware reconstruction error (MRE), (*ii*) attribute editing accuracy and (*iii*) image quality.

Table 1 shows that MagGAN decreases the MRE significantly, indicating better preserving of regions that should be intact. This improvement is also obvious in the editing results in Figure 8. Table 1 also reports the PSNR/SSIM score of the reconstructed image by keeping target attribute vector the same as the source one (detailed definition in Supplementary). MagGAN also improves PSNR/SSIM significantly.

We also report the attribute editing accuracy by employing the pre-trained attribute classification model from [25]. We follow the evaluation protocol used

Table 1: Comparison of quantitative results with SOTA

Methods	MRE ↓	FID ↓	Avg Acc	PSNR	SSIM
AttGAN(128)	0.0713	10.23	64.9%	24.07	0.841
STGAN(128)	0.0627	7.75	85.8%	31.67	0.948
STGAN(256)	0.0530	1.21	90.4%	37.61	0.959
MagGAN(256)	0.0163	**1.10**	90.0%	40.25	0.984
MagGAN(512)	0.0141	1.20	89.1%	41.42	0.987
MagGAN(1024)	**0.0130**	1.31	**91.0%**	**42.94**	**0.994**

in [9,25]. For each test image, reverse one of its 13 attributes at a time (1 → 0 or 0 → 1), and generate an image after each reversion; so there are 13 edited images for each input image. The widely used evaluation metric is *attribute editing accuracy*, which measures the successful manipulation rate for the reversed attribute each time, but ignores the attribute preservation error. Figure 6 reports the facial attribute manipulation accuracy of previous works IcGAN [32], FaderNet [16], AttGAN [9], StarGAN [6], STGAN [25] and our proposed MagGAN. To build the strongest baseline, we also train our own STGAN model at resolution 256 × 256, optimizing all possible parameters; see details of the hyperparameter tuning in Supplementary.

High editing accuracy *v.s.*attribute-irrelevant region preserving. As shown in Table 1, MagGAN at resolution 256 outperforms all the previous reported numbers except STGAN(256) on average accuracy. In Figure 6, compared with STGAN(256), MagGAN(256) is better in "Mustache", "No beard", "Gender", "Age" and worse in "Bald", "Bangs", "Black Hair", "Blonde Hair", "Brown Hair". We conjecture that STGAN(256) achieves this high accuracy by editing hat or scarf when they appear in the image; like coloring the hat to golden to get an editing success of "Blonde Hair". To verify this assumption, we separate the testing set into two groups – samples with hat, samples without hat by measuring the area ratio of hat in the face masks (we select threshold 0.1 to decide if the sample contains a hat). The attribute editing accuracy is evaluated on the two subsets respectively. Results in Figure 7 show that the editing accuracy of MagGAN decreases a lot on hat subset on several hat-related attributes, *e.g.*, "Bald", "Black Hair", but on par with STGAN on non hat subsets. In this sense, MagGAN editing success is even higher than our strongest baseline STGAN(256) since it can preserve the attribute irrelevant regions, making editing more real.

To measure the image quality, we report FID (Fréchet Inception Distance) score [10]. The FID score measures the distance between the Inception-v3[8] activation distributions of original images and the edited images. Table 1 shows that the FID score improves significantly from resolution 128 to 256, but then get stalled and insensitive to image quality for 256 and higher resolutions. This is because the input size for Inception-v3 model is 299, and thus resolution increase from 128 to 256 is significant. However, all high resolution generations are first downsampled to evaluate the FID score. After all, MagGAN at all

[8] We pretrained an Inception-V3 model that achieves 92.69% average attribute classification accuracy on all 40 attributes of CelebA dataset.

Table 2: Results of user study for ranking methods on two subsets considering hat wearing

Winner method	w/ hat	w/o hat	Overall
MagGAN	**59.2%**	**52.1%**	**55.7%**
STGAN	37.7%	45.3%	41.5%
Tie	3.1 %	2.6%	2.8 %

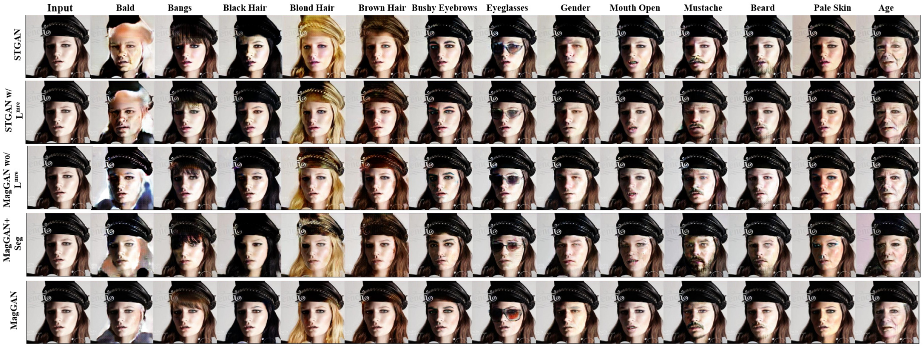

Fig. 8: Visual results of MagGAN variants on resolution 256 × 256. Each column represents edited images through one attribute reversing editing

resolutions achieves the comparable result with the best FID score. Finally, due to smaller batches in training for high resolutions, FID scores of MagGAN(512) and MagGAN(1024) are slightly lower than those of MagGAN(256).

Qualitative evaluation. Apart from the quantitative evaluation, we visualize some facial attribute editing results at resolution 256 × 256 in Figure 8, and compare our proposed model with the state-of-the-art method, *i.e.*, STGAN [25] (as it is the strongest baseline) and other variations.

User Study. We conduct user study on Amazon Turk to compare the generation quality of STGAN and MagGAN. To verify that MagGAN performs better on editing attribute relevant regions, we randomly choose 100 input samples from test set, 50 samples with hat or scarf and 50 samples without (since STGAN usually fails on person wearing hat). For each sample, 5 attribute editing tasks are performed by STGAN and MagGAN (500 comparison pairs in total). All 5 tasks are randomly chosen from 13 attributes, for subjects with hat, we increase the chance to select hair related attributes. The users are instructed to choose the best result which changes the attribute more successfully considering image quality and identity preservation. To avoid human bias, each sample pair is evaluated by 3 volunteers. The results are shown in Table 2, MagGAN outperforms STGAN on both hat samples and without hat samples.

5 Ablation Study

We conduct three groups of ablation comparisons in image resolution of 256 × 256, to verify the effectiveness of the proposed modules individually: (*i*) mask guided

Table 3: Comparison of variants of MagGAN on 256 × 256

Methods	MRE ↓	FID ↓	Avg Acc	PSNR	SSIM
(i) STGAN	0.0530	1.21	90.4%	37.61	0.959
(ii) STGAN+cycle	0.0530	1.31	87.3%	36.14	0.970
(iii) STGAN w/ L^{mre}	0.0289	1.33	**95.6%**	38.48	**0.984**
(iv) MagGAN w/o L^{mre}	0.0397	1.22	89.6%	39.35	0.980
(v) MagGAN w/o SP	**0.0161**	1.23	89.9%	**40.40**	0.982
(vi) MagGAN	0.0163	**1.10**	90.0%	40.25	**0.984**
(vii) MagGAN+Seg	0.0612	2.39	90.3%	40.10	0.983

reconstruction loss, (*ii*) spatially modified attribute feature, and (*iii*) usage of SPADE normalization.

We consider seven variants, *i.e.*, (*i*) STGAN: STGAN at resolution 256 × 256, (*ii*) STGAN+cycle: STGAN with cycle-consistency loss instead of its original reconstruction loss, (*iii*) STGAN w/ L^{mre}: STGAN plus mask guided reconstruction loss, (*iv*) MagGAN w/o L^{mre}: MagGAN trained without mask guided reconstruction loss, (*v*) MagGAN w/o SP: MagGAN without using SPADE, (*vi*) MagGAN: our proposed model with the usage of mask-guided reconstruction loss and make-guided attribute conditioning. (*vii*) MagGAN+Seg: Instead of using a pre-trained face parser, build a face segmentation branch (adopting FCN[28] architecture) into generator as sub-task, making the whole model fully trainable.

Mask-guided reconstruction loss. We compare three reconstruction loss: (i) STGAN with only the reconstruction loss computed by reconstructed images, (ii) cycle-consistency loss which is applied in StarGAN [6], (iii) two parts of reconstruction loss (computed on reconstructed images and synthesized images respectively) proposed in § 3.2. Row 1-3 of Table 3 report the quantitative results of STGAN applying each type of reconstruction loss respectively. We observe that adding mask guided reconstruction loss to generator training can effectively reduce Mask-aware Reconstruction Error (MRE). In Figure 8, the synthesized image of STGAN w/ L^{mre} on attribute "Bald" and "Blonde Hair" also proves this assumption. But since the spatial information of mask is not directly injected into generator, STGAN w/ L^{mre} still cannot preserve the attribute-irrelevant regions well.

Mask-guided attribute conditioning. Utilizing mask-guided attribute conditioning instead of the spatially uniformed attribute conditioning provides generator with more spatial information of the interest regions. From Table 3, (i) *v.s.*(iv), (iii) *v.s.*(vi) illustrate that the MRE score decreases obviously when mask-guided attribute conditioning is applied in generator. It implies that generator effectively takes the regions of interest and edits on these local regions. Taking advantage of both mask-guided reconstruction loss and attribute conditioning strategy, MagGAN achieves the best MRE and FID. And the visual results in Figure 8 also show that MagGAN makes accurate editing on hair related attributes ('Bold', 'Blonde Hair', *etc*), by preserving the region of hat while only remove or paint the hair. MagGAN w/o SP and MagGAN perform nearly the same as (v) *v.s.*(vi), which demonstrates the denormalization method does not affect much on performance. Finally, the quantitative results and visual

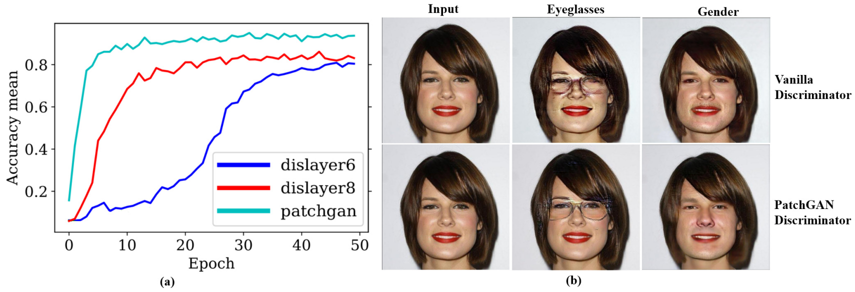

Fig. 9: Comparison of training with vanilla single discriminator and multi-level PatchGAN discriminators on resolution 1024×1024: (a) attribute editing accuracy and (b) visual results

results of (vii) MagGAN+Seg are bad, which indicates the incorporating mask segmentation branch as part of the generator is not a good choice. Since the mask-guided reconstruction loss and attribute conditioning requires accurate masks, training segmentation branch with generator from scratch makes the model hard to train and undermines the editing accuracy.

Multi-level PatchGAN discriminator for high resolution editing. We apply PatchGAN discriminator to supervise training of high resolution image generation. We are able to scale the generated image resolution up to 1024 × 1024. In Figure 9, we compare the 1024 version of training with a single discriminator and with our proposed multi-level PatchGAN discriminators. Under this setting, PatchGAN has 3 discriminators working on resolution 256 × 256, 512 × 512 and 1024, respectively. In Figure 9 (a), when applying single vanilla discriminator, the generator converges slower than using PatchGAN discriminator and early stops at low editing accuracy. In Figure 9 (b), editing effects on "Eyeglasses", "Gender" from PatchGAN are more obvious than original discriminator. We assume PatchGAN discriminators provide more supervise signal on global and local regions, thus helping generator learns more discriminative features for each attribute. See more visual results in supplementary.

6 Conclusion

In this paper, we propose MagGAN for high-resolution face image editing. The key novelty of our work lies in the use of facial masks for achieving more accurate local editing. Specifically, the mask information is used to construct a mask-guided reconstruction loss and mask-guided conditioning in the generator. MagGAN is further scaled up for high-resolution face editing with the help of PatchGAN discriminators. To our knowledge, it is the first time face attribute editing is able to be applied on resolution 1024 × 1024.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017) 7
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016) 5
3. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: NIPS. pp. 2172–2180 (2016) 4
4. Chen, Y.C., Shen, X., Lin, Z., Lu, X., Pao, I.M., Jia, J.: Semantic component decomposition for face attribute manipulation. In: CVPR (2019) 1
5. Chen, Y.C., Xu, X., Tian, Z., Jia, J.: Homomorphic latent space interpolation for unpaired image-to-image translation. In: CVPR (2019) 4
6. Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR. pp. 8789–8797 (2018) 1, 4, 5, 10, 11, 13
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014) 1
8. Gu, S., Bao, J., Yang, H., Chen, D., Wen, F., Yuan, L.: Mask-guided portrait editing with conditional gans. In: CVPR (2019) 4
9. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing **28**(11), 5464–5478 (2019) 1, 4, 5, 10, 11
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint arXiv:1706.08500 (2017) 11
11. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017) 5
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017) 4, 9
13. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018) 9
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR **abs/1812.04948** (2018) 5
15. Klys, J., Snell, J., Zemel, R.S.: Learning latent subspaces in variational autoencoders. In: NeurIPS. pp. 6445–6455 (2018) 1, 4
16. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: Manipulating images by sliding attributes. In: NIPS. pp. 5969–5978 (2017) 1, 4, 10, 11
17. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: ICML. pp. 1558–1566 (2016) 1, 4
18. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: ECCV (2012) 4
19. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: towards diverse and interactive facial image manipulation. arXiv preprint arXiv:1907.11922 (2019) 4
20. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: towards diverse and interactive facial image manipulation. arXiv preprint arXiv:1907.11922 (2019) 5
21. Li, H., Dong, W., Hu, B.: Facial image attributes transformation via conditional recycle generative adversarial networks. J. Comput. Sci. Technol. **33**(3), 511–521 (2018) 1, 4

22. Li, M., Zuo, W., Zhang, D.: Deep identity-aware transfer of facial attributes. CoRR **abs/1610.05586** (2016) 4
23. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. In: CVPR (2019) 4
24. Liang, X., Xu, C., Shen, X., Yang, J., Liu, S., Tang, J., Lin, L., Yan, S.: Human parsing with contextualized convolutional neural network. In: ICCV (2015) 4
25. Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., Wen, S.: STGAN: A unified selective transfer network for arbitrary image attribute editing. CVPR pp. 3673–3682 (2019) 1, 2, 3, 4, 5, 7, 10, 11, 12
26. Liu, M., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NIPS. pp. 700–708 (2017) 4
27. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015) 9
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015) 13
29. Lu, Y., Tai, Y., Tang, C.: Attribute-guided face generation using conditional cyclegan. In: ECCV. pp. 293–308 (2018) 1, 4
30. Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., Gool, L.V.: Exemplar guided unsupervised image-to-image translation. CoRR **abs/1805.11145** (2018) 4
31. Park, T., Liu, M., Wang, T., Zhu, J.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019) 4, 5
32. Perarnau, G., van de Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional gans for image editing. CoRR **abs/1611.06355** (2016) 1, 4, 10, 11
33. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: CVPR. pp. 1225–1233 (2017) 4
34. Wang, Y., Wang, S., Qi, G., Tang, J., Li, B.: Weakly supervised facial attribute manipulation via deep adversarial network. In: WACV. pp. 112–121 (2018) 4
35. Xiao, T., Hong, J., Ma, J.: ELEGANT: exchanging latent encodings with GAN for transferring multiple face attributes. In: ECCV. pp. 172–187 (2018) 4
36. Xie, D., Yang, M., Deng, C., Liu, W., Tao, D.: Fully-featured attribute transfer. CoRR **abs/1902.06258** (2019) 1, 4
37. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: ECCV. pp. 776–791 (2016) 1, 4
38. Yin, W., Liu, Z., Loy, C.C.: Instance-level facial attributes transfer with geometry-aware flow. CoRR **abs/1811.12670** (2018) 4
39. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: ECCV (2018) 5
40. Zhang, G., Kan, M., Shan, S., Chen, X.: Generative adversarial network with spatial attention for face attribute editing. In: ECCV. pp. 422–437 (2018) 1, 4
41. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017) 9
42. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. TPAMI (2018) 9
43. Zhang, J., Shu, Y., Xu, S., Cao, G., Zhong, F., Liu, M., Qin, X.: Sparsely grouped multi-task generative adversarial networks for facial attribute manipulation. In: ACM MM. pp. 392–401 (2018) 4
44. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: CVPR. pp. 4352–4360 (2017) 1, 4

45. Zheng, X., Guo, Y., Huang, H., Li, Y., He, R.: A survey to deep facial attribute analysis. CoRR **abs/1812.10265** (2018) 4
46. Zhou, S., Xiao, T., Yang, Y., Feng, D., He, Q., He, W.: Genegan: Learning object transfiguration and object subspace from unpaired data. In: BMVC (2017) 4