SMILE: Semantically-guided Multi-attribute Image and Layout Editing

Andrés Romero[1] Luc Van Gool[1,2] Radu Timofte[1]
[1]Computer Vision Lab, ETH Zürich [2]KU Leuven
{roandres, vangool, timofter}@vision.ee.ethz.ch

Abstract

Attribute image manipulation has been a very active topic since the introduction of Generative Adversarial Networks (GANs). Exploring the disentangled attribute space within a transformation is a very challenging task due to the multiple and mutually-inclusive nature of the facial images, where different labels (eyeglasses, hats, hair, identity, etc.) can co-exist at the same time. Several works address this issue either by exploiting the modality of each domain/attribute using a conditional random vector noise, or extracting the modality from an exemplary image. However, existing methods cannot handle both random and reference transformations for multiple attributes, which limits the generality of the solutions. In this paper, we successfully exploit a multimodal representation that handles all attributes, be it guided by random noise or exemplar images, while only using the underlying domain information of the target domain. We present extensive qualitative and quantitative results for facial datasets and several different attributes that show the superiority of our method. Additionally, our method is capable of adding, removing or changing either fine-grained or coarse attributes by using an image as a reference or by exploring the style distribution space, and it can be easily extended to head-swapping and face-reenactment applications without being trained on videos.

1. Introduction

In this paper we tackle the problem of adding, removing, or manipulating facial attributes for either exemplar images or random manipulations, using a single model. For instance, given a person A, our system could aim at imposing the haircut of person B, eyeglasses of person C, hat of person D, earrings of person E, and randomly changing the background and the color of the hair. Particularly, the problem of manipulating multiple attributes has been coined 'multi-domain image-to-image (I2I) translation' [23, 7, 30].

Image-to-image translation methods have been traditionally categorized into two groups: latent and exemplar approaches. Latent approaches [46, 2, 6] require sampling from distribution in order to perform a cross-domain mapping, that is, to explore the underlying latent distribution and produce a plurality of representations given a single input. Conversely, exemplar-based approaches [37, 40, 10] require an additional image to condition the generation. There have been some efforts [14, 21, 8] trying to reconcile latent and exemplar approaches in a single and unified system. However, they consider domains with big gaps such as different kinds of animals, weather conditions, or male/female facial transformation.

Figure 1: SMILE learns a diverse manipulation for multiple attributes using a single generator. We do not use direct supervision as we simplify the problem using semantic segmentation manipulation as an intermediate stage. First, we manipulate eyeglasses, hair, earrings, and hat shapes from reference images on the semantic space (bottom right corner). Second, we impose the style of the reference attributes onto the manipulated semantic in order to synthesise the RGB output. Zoom in for better details.

Regarding facial manipulation, I2I translation approaches come with the additional constraint that some regions of the image (*e.g.*, background, clothes) or fixed characteristics of the face (*e.g.*, eyeglasses, hats) should remain unaltered during the transformation. Vanilla CycleGAN [45]-based approaches traditionally alter the general content and shift the colors of the input. To overcome this undesired property, latent generative approaches [23, 11, 38] have proposed attention mechanisms [29], performing architectural changes and introducing tailored loss func-

tions into the training framework, thus obtaining impressive results. Nevertheless, the transformations are mostly fine-grained and do not perform well for more global transformations such as a female to male, or short to long hair. Reference guided methods [44, 4, 39], on the other hand, either work on low-resolution scales or focus on local texture transformation. Recently, StarGANv2 [8] was proposed as a variant for multi-domain I2I translation. Nonetheless, it requires that the multiple domains are not activated at the same time and it does not perform well for fine-grained transformations.

In order to solve the aforementioned issues, we propose Semantically-guided Multi-attribute Image and Layout Editing (SMILE). With SMILE, we split the solution of this problem into two stages. We propose to introduce a segmentation space as an intermediate space between the high-level attribute semantic space (*e.g.*, eyeglasses, hair, hat) and the high-informative RGB image space, where instead of dealing with complex general and local transformations in the RGB space. First, we manipulate the semantic space according to the desired attributes, namely **semantic manipulation**. Second, we transform such semantic manipulations to photo-realistic RGB faces, namely **image synthesis**.

We enumerate our contributions as follows:

1. We propose a multi-attribute I2I transformation method for both fine-grained and more global attributes in the semantic space for both random and exemplar-guided synthesis.

2. We propose an extended version of StyleGAN2 [16] to deal with semantic masks and per-region-styles to perform random or exemplar-guided synthesis.

We depict diverse facial manipulations in Figure 1, and an overview of our system in Figure 2. Code source and pre-trained models will be released.

2. Related Work

Recently, Image-to-image (I2I) translation has become a very active topic thanks to the impressive advances in generative modeling methods, and in particular, Generative Adversarial Networks (GANs) [9]. Several novel and challenging problems have been successfully tackled with this technique, *e.g.*, multi-domain manipulation [7, 30], style transferring [13, 22], image inpainting [41, 26], image synthesis using semantic segmentation [27, 47, 20], image content manipulation [28], exploratory image super-resolution [24, 5].

2.1. Facial Attribute Manipulation

Since the face is one of the most common, yet interesting models, facial image editing has gained traction over the years [35, 34, 3]. Different works [7, 29, 19, 10] have included facial attribute information as conditions in GANs to manipulate eyeglasses, mouth expression, hair, and other attributes with remarkable results. Due to the mutually-inclusive representation of facial attributes, *i.e.*, different attributes can co-exist at the same time, multi-domain methods have received more attention as a unified and flexible way to deal with several domains. Nevertheless, modeling each attribute as a domain requires having a fully disentangled understanding of each attribute. Although multi-attribute latent manipulation has been widely studied [7, 11, 30, 42, 29, 19], multi-attribute exemplar attribute imposition has been less studied [10], and the combination has not been achieved yet.

Multi-attribute facial exemplar manipulation refers to extract some specific information from person A's face and impose it on person B's face, *e.g.*, make-up, eyeglasses, smile, hair. There are traditionally two different groups of methods doing this: makeup transferring [6, 37] and attribute manipulation [44, 4, 25, 39, 40, 10]. Make-up transferring methods focus on localized texture mappings, whereas attribute imposing methods are traditionally modeled as binary problems using the presence or absence of a selected feature. While the former allows for high-resolution transformations and require exemplar images, the latter normally operate at low resolution due to the intricate representations of multiple attributes in the RGB space, which traditionally operates at one model per domain.

Recently, StarGANv2 [8] has been introduced as an alternative for multi-domain facial attribute manipulation for both random sampling and reference guidance. Nonetheless, as we discuss in Section 4 the generalization capabilities of StarGANv2 are compromised when training with different and/or additional domains to Male/Female, and it also wanting when it comes to fine-grained transformations.

There is a common issue among the above-mentioned methods. When including several exemplar attribute manipulations that can co-exist at the same time, they struggle due to the inherent lack of exemplar supervision, *i.e.*, there is the only access to presence and absence of attributes in a high-level manner, which is the reason this problem is traditionally simplified by using one manipulation and one model at a time [44, 4, 25, 39, 40] or to operate at a low-resolution scale including several domains [10]. In contrast to previous works, we leverage the high-level semantic space to manipulate the high-informative RGB of attributes, which is a much easier space that allows us to manipulate either fine-grained or coarse attributes in a higher resolution space, so we can manipulate only the shape of the attribute, and transforming it into a specific style using a semantically guided RGB synthesis. Furthermore, our method combines the best of the two worlds by either using specific attribute imposition from exemplar images or exploring the latent space

using a fully disentangled representation.

2.2. Semantically-guided manipulation

Using semantic information for image synthesis is an emerging field, in which using semantic segmentation as input, aims at producing an RGB image that perfectly resembles the semantic regions in the input. Semantic manipulation allows finer control of the resulting image just by adjusting the input. To this end, inspired by pix2pixHD [36], SPADE [27] introduced a specialized and spatially driven normalization block in order to deal with the different masks in an up-sampling manner, producing impressive results for high-resolution synthesis. However, one critical issue about SPADE is the lack of control for each resulting semantic region. Recently, SEAN [47] and MaskGAN [20] modeled independent style representations for each semantic region, and in the same vein as SPADE, they introduce the semantic and style information combined in a W space distribution through adaptive normalization layers in the generator. It is worth mentioning that both SEAN [47] and MaskGAN [20] require an exemplar image to perform the generation, and this is particularly critical for attribute imposition as we would like to generate new content (*e.g.*, hat, eyeglasses) that can be hard to find in a dataset. Note that we refer to image manipulation as manipulating an existing image rather than hallucinating it as in vanilla GAN-based methods [27, 16].

SMILE is akin to both SEAN and MaskGAN, yet by leveraging StyleGAN2 [17] we extend the W latent distribution towards virtually any kind of style per semantic region. We accomplish this by replacing the normalization layers in the generator with semantically adaptive convolutions (SACs), and by using an alternative training scheme for both random generation (similar to StyleGAN2) and exemplar-guided generation (similar to SEAN).

3. Proposed Approach

Our main objective is to perform specific or global style-guided transformations using only the domain information as supervision. We argue that multi-domain exemplar style imposition (*e.g.*, wearing someone else's sunglasses or hair replacement) is a very challenging problem, which normally is simplified by assuming mutually exclusive domains [8], or one model per domain [4]. To this end, without simplifying this problem and inspired by recent developments in image synthesis [17], we develop our strategy in two stages: Semantic Manipulation ($\text{SMILE}_{\text{SEM}}$) and Region-wise Semantic Synthesis ($\text{SMILE}_{\text{SYN}}$).

We assume we have access to facial images, with their respective semantic segmentation, where each semantic region corresponds to a part of the face (*e.g.*, eyeglasses, eyes, mouth). First, we manipulate the semantic information, *i.e.*, the shape of each attribute, with high-level attribute information extracted either from a different semantic person or sampled from the distribution. In Figure 2 left, eyeglasses and hair shapes are extracted from the semantic reference. Second, we use the manipulated semantic information to transform it into RGB space, namely, style synthesis. For each semantic region, we extract the RGB information either from an exemplar image, from the original input, or sampled from the distribution. In Figure 2 right, eyeglasses, and hairstyles from the reference image are combined with the skin, mouth, and background from the original input.

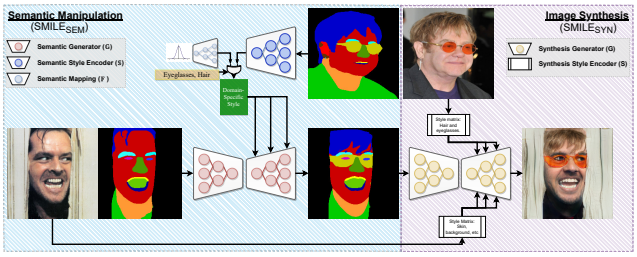


Figure 2: **Overview of SMILE.** We translate an image by either taking as input a random style or target attributes into the generator (we use a reference image in this example). We first manipulate the shape of the attribute by using a semantic segmentation map (left), and then we synthesize the style of each semantic region by using both input and reference styles to produce a photo-realistic merge of the two images (right). Our proposed approach SMILE is an ensemble $\text{SMILE}_{\text{SYN}} \circ \text{SMILE}_{\text{SEM}}$.

3.1. Semantic Manipulation ($\text{SMILE}_{\text{SEM}}$)

First, we build a simple yet powerful multi-domain I2I translation model using the semantic map. We rely on the semantic information as it is simpler and rich enough to spot and transform noticeable facial attributes like eyeglasses, hats, earrings, hair, bangs, and identity.

We build on top of StarGANv2 [8], and as depicted in Figure 2 (left), our system is composed of several networks: 1 semantic generator ($\mathbb{G}$), 1 semantic mapping network ($\mathbb{F}$) to map from a noise distribution to a shared latent space, and 1 semantic style encoder ($\mathbb{S}$) to map reference images to shared latent space. In order to perform unsupervised fine-grained and more global translations, we rely on several key assumptions as follows.

3.1.1 Model

We model the semantic manipulation as a problem of Image-to-Image translation, where the input and output are semantically segmented faces with access to binary attribute annotations, and the mapping function is $\mathbb{G}$.

Let $\mathcal{X}^r \in \mathbb{R}^{H \times W \times M}$ be the real image with M semantic channels, for instance a mask with a parsing of the face

where each channel represents different regions. Each image $\mathcal{X}^r$ has associated N binary attributes $y^r \in \mathbb{N}^{\{0,1\}} : \{y_0^r, \cdots y_i^r, \cdots y_{N-1}^r\}$, for instance wearing eyeglasses or not, wearing a hat or not. Importantly, as each attribute can have virtually countless appearances, this ground-truth information is unknown.

By using self-supervision we can extract each attribute style representation from the real image, that is, we assume that for each possible attribute y^r in a real image $\mathcal{X}^r$, there is one style associated $s^r \in \mathbb{R}^S : \{s_0^r, \cdots s_i^r, \cdots s_{N-1}^r\}$. For instance, for any given facial image x^i that has 4 binary labels $y^i = (1, 1, 0, 1)$, then each present attribute can have different shape, color, and appearance, namely style, so there is one style vectors per attribute ($\langle s_i \rangle$): $s^i = (\langle s_0 \rangle, \langle s_1 \rangle, \langle s_2 \rangle, \langle s_3 \rangle)$, that is, one latent representation for each label. Note that we assume that the absence of an attribute is also associated with a style distribution, and not as a deterministic zero vector. For instance, for the absence of black eyeglasses, there must be a style that hallucinates the region around the eyes.

Our purpose is to use the domain information as guidance for the style imposition. Particularly, for each domain we assume, a style distribution associated with the presence of it and a different style distribution associated with the absence of it.

Moreover, we can perform transformations ($\hat{X}$) in both directions: using an image as a reference by extracting the style from the style encoder ($\mathbb{S}$), and similarly, sampling from the style distribution and processing it through the mapping network ($\mathbb{F}$). Formally, we define these two transformations in Equation 1 and 2, respectively.

$$\hat{X}_{guided} = \mathbb{G}(X^r, \mathbb{S}(X^{ref})_{\hat{y}}) \tag{1}$$

$$\hat{X}_{random} = \mathbb{G}(X^r, \mathbb{F}(\mathcal{N}(0, I))_{\hat{y}}), \tag{2}$$

Where $\mathcal{N}(0, I)$ is a random vector sampled from the normal distribution. This is possible by assuming a shared latent space. Note that these transformations require the selection of the presence or absence of domains ($\hat{y}$) in each style mapping $\mathbb{S}$ and $\mathbb{F}$.

3.1.2 Training Framework

In this section, we explain in detail our method to work with either inclusive or exclusive domains, and also fine-grained or coarse transformations.

First, as each domain has two style distributions, we use the domain information in form of multi-task learning to inject the desired style representation into the generator. The resultant style is a weighted concatenation of all the attributes. Second, we replace the AdaIN and convolution layers with modulated convolutions [17], and we discuss this architectural change in Section 4. Third, we propose a novel training scheme critical for the success of the training stability.

During the forward pass, we first sample a noise vector ($\mathcal{N}(0, I)$) and randomly sample real domain labels ($\hat{y}$) to generate a mapping latent vector ($\hat{s}$), which is fed to the generator. The random style is defined in Equation 3:

$$\hat{s} = \mathbb{F}(\mathcal{N}(0, I))_{\hat{y}} \tag{3}$$

For the Discriminator, Mapping Network, and Style encoder, we use multi-task learning on the active domains and ignore the optimization for the zero-domain vectors.

Fake images are produced as $\hat{x} = \mathbb{G}(x, \hat{s})$. In contrast to StarGANv2, we only require one reconstruction step. We define the style reconstruction loss in Equation 4:

$$\mathcal{L}_{sty} = \min_{\mathbb{G},\mathbb{S}} \left[\|\hat{s} - \mathbb{S}(\mathbb{G}(x, \hat{s}))_{\hat{y}}\|_1 \right] \tag{4}$$

To further encourage diversity across the transformations, we follow the same pixel-wise style diversification as in StarGAN2. See Equation 5 for the style diversification loss.

$$\mathcal{L}_{sd} = \max_{\mathbb{G}} \left[\|\mathbb{G}(x, \hat{s})) - \mathbb{G}(x, \hat{s}'))\|_1 \right] \tag{5}$$

The key ingredient to stabilize our system relies on the reconstruction loss. As we are only learning $\mathbb{S}$ parameters using Equation 4, and we need to align the style encoder for both real and fake images, for the reconstruction loss we simply detach the weights from the graph. With this strategy, we force the two distributions $\mathbb{S}$ and $\mathbb{F}$ to be aligned. We found this trick to be crucial in the overall training framework. Therefore, the real style has the form of $\tilde{s} = \text{detach}(\mathbb{S}(x)_{y^r})$, and we define this loss in Equation 6.

$$\mathcal{L}_{rec} = \min_{\mathbb{G}} \left[\|x - \mathbb{G}(\mathbb{G}(x, \hat{s}), \tilde{s})\|_1 \right] \tag{6}$$

As it is common for adversarial approaches, we use the adversarial loss (L_{adv}) to produce photo-realistic images. We follow the same adversarial loss and regularizer as in StarGANv2. Our full loss function is defined in Equation 7.

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{sty}\mathcal{L}_{sty} + \lambda_{sd}\mathcal{L}_{sd}, \tag{7}$$

where λ represents the relative importance of each part within the system.

3.1.3 Experimental Setup

We build our system for 256×256 image size. For all our experiments we set $\lambda_{rec} = 1.0$, $\lambda_{sty} = 1.0$, and $\lambda_{ds} = 20.0$. We train our system during 200,000 iterations using a single GPU Titan Xp with a batch size of 6, and Adam Optimizer [18].

Please refer to the Supplementary Material for more details about the networks.

3.2. Improved Semantic Image Synthesis (SMILE$_{\text{SYN}}$)

In order to map from semantic regions to an RGB image, we use our semantic guided image synthesis method to perform the corresponding generation either by using an exemplar image or exploring the latent space.

Current methods [35, 34, 1] that use StyleGAN [16] for image manipulation, latent disentanglement, or image projection have to go through several steps: (i) train StyleGAN until convergence, (ii) study the latent space to produce meaningful yet visible disentangled representation which usually involves more training stages, and (iii) optimize the latent space for a reference image. We propose a method that only relies on the first step (i), and during inference, both the latent space manipulation and image reconstruction can be efficiently and effectively achieved.

3.2.1 Model

Recently, SEAN [47] and MaskGAN [20] have been proposed as strong alternatives for the generation of images using layout references by disentangling each style to each semantic region. However, the generation suffers from being tied to an exemplar image. In a similar direction, StyleGAN2 [17] is the current state-of-the-art for image generation. Inspired by Hong *et al*. [12] and SEAN, we replace StyleGAN2 modulated convolutions (ModConv) with improved semantic region-wise adaptive convolutions (SACs). Let w be the kernel weight, h the input features of the convolution, s the condition information, and σ_E the standard deviation also known as the demodulating factor, we define SACs in Equation 8.

$$\text{ModConv}_w(\mathbf{h}, s) = \frac{w * (s\mathbf{h})}{\sigma_E(w, s)} \Leftrightarrow s \in \mathbb{R}^{1\times C\times 1\times 1}$$
$$\text{SAC}_w(\mathbf{h}, \mathbf{s}) = \frac{w * (\mathbf{s} \odot \mathbf{h})}{\sigma_E(w, \mathbf{s})} \Leftrightarrow \mathbf{s} \in \mathbb{R}^{1\times C\times H\times W}, \quad (8)$$

where,

$$\mathbf{s} = \alpha_w SM + (1 - \alpha_w)M,$$

where SM is the per-region style matrix, which can be either extracted from an image or sampled from a Gaussian distribution, M is the required semantic mask, and α_w is a learned parameter that weights for the relative importance of each element at each layer of the network. This equation can also be seen as the SEAN [47] gamma factor. Please see [12] for further details on the mathematical development of Equation 8.

3.2.2 Training Framework

To couple this proposed scheme with the StyleGAN2 training framework, we propose an alternate scheme training. First, we update the generator ($\mathbb{G}$) and discriminator ($\mathbb{D}$) for a random generation as in StyleGAN2. Second, in addition to the generator and discriminator, we also update a style encoder network ($\mathbb{S}$) for exemplar-guided synthesis. For simplicity, we show the loss function for the generator during the reference synthesis in Equation 9. To this end, let x and m be the real image and its corresponding semantic map, respectively.

$$\mathcal{L}_{feat} = \min_{\mathbb{G},\mathbb{S}} \sum_{i=1}^{T-1} \frac{1}{N_i} \left[\| \mathbb{D}_k^{(i)}(x) = \mathbb{D}_k^{(i)}(\mathbb{G}(m, \mathbb{S}(x, m)) \|_1 \right]$$
$$\mathcal{L}_{\text{reference}} = \mathcal{L}_{adv} + \lambda_{feat}\mathcal{L}_{feat} \quad (9)$$

where T is the total number of layers in the discriminator, N is the number of elements in each layer, and λ_{feat} and λ_{pt} represents the importance of the feature matching loss [36], and it is set to 10. Note that the feature matching loss is only required for the reference update.

3.2.3 Experimental Setup

The generator uses semantic maps as the starting point for image synthesis. Instead of starting from a constant representation as in StyleGAN, and as the semantic segmentation information represents the high-level information of the data, we empirically found that starting from 8×8 yields better performance. Please refer to the Supplementary Material for this experiment and details about the networks.

Given current computational limitations to fully train StyleGAN2, we train our system during 300,000 iterations (roughly 3 weeks) using a single GPU Titan Xp with a batch size of 4 and image size of 256.

3.3. Datasets

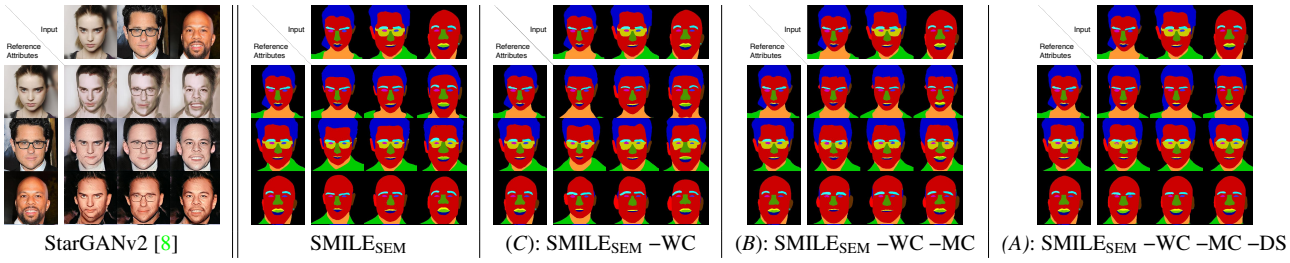
Semantic Manipulation We validate our semantic manipulation method in CelebA-HQ [15] that consists of multiple facial attribute labels. Since we are tackling semantic manipulation, we selected 6 visible attributes that were not related to facial texture: eyeglasses, hat, amount of hair, bangs, earrings, and identity[1]. For the semantic segmentation labels, we use the ones provided by CelebA-Mask [20].

Semantic Image Synthesis Semantic Image Synthesis only requires generating photo-realistic images using a semantic segmentation as input, and as this scheme does not need having access to facial attribute labels, we validate this part of the system using the FFHQ [16] dataset.

3.4. Evaluation Framework

For our entire system, we study independent performances under two circumstances: generation by latent space and generation by exemplar images. Since our proposed solution splits into two stages, we evaluate the semantic manipulation and image synthesis independently.

[1] Identity refers to the Male/Female label in the CelebA-HQ dataset.



StarGANv2 [8] | SMILE$_{\text{SEM}}$ | (*C*): SMILE$_{\text{SEM}}$ −WC | (*B*): SMILE$_{\text{SEM}}$ −WC −MC | (*A*): SMILE$_{\text{SEM}}$ −WC −MC −DS

Figure 3: Qualitative results of the ablation experiments for semantic manipulation. We use reference images to perform attribute transformation, *i.e.*, for this visualization we transfer all the reference attributes (identity, eyeglasses, baldness) to the input. Please zoom in for a detailed assessment. WC, MC, and DS stand for Weighted Classes, Modulated Convolutions, and Detaching Style, respectively, *i.e.* −WC means we substract the weighted classes experiment from the pipeline.

	CelebA-HQ [15] — **Latent Synthesis**							
	Pose			Attributes		Reconstruction	Perceptual	
	Roll↓	Pitch↓	Yaw↓	AP↑	F1↑	mIoU↑	FID↓	Diversity↑
StarGANv2 [8]	2.952 ± 0.856	16.900 ± 6.264	29.331 ± 8.134	0.795 ± 0.092	0.797 ± 0.079	0.964 ± 0.012	81.945 ± 24.276	0.018 ± 0.008
SMILE$_{\text{SEM}}$	2.589 ± 0.684	15.082 ± 4.097	11.286 ± 1.983	0.960 ± 0.031	0.946 ± 0.032	0.989 ± 0.002	43.151 ± 15.527	0.399 ± 0.020
(*C*): SMILE$_{\text{SEM}}$ −WC	2.683 ± 0.792	18.628 ± 6.243	10.553 ± 2.560	0.965 ± 0.028	0.953 ± 0.027	0.986 ± 0.002	48.123 ± 14.759	0.390 ± 0.013
(*B*): SMILE$_{\text{SEM}}$ −WC −MC	2.732 ± 0.681	18.172 ± 4.500	17.626 ± 7.250	0.940 ± 0.039	0.928 ± 0.038	0.987 ± 0.003	46.797 ± 14.204	0.395 ± 0.013
(*A*): SMILE$_{\text{SEM}}$ −WC −MC −DS	2.359 ± 0.678	13.520 ± 4.476	15.424 ± 6.432	0.889 ± 0.062	0.884 ± 0.051	0.994 ± 0.001	61.015 ± 22.235	0.382 ± 0.014
	CelebA-HQ [15] — **Reference Synthesis**							
	Pose			Attributes		Reconstruction	Perceptual	
	Roll↓	Pitch↓	Yaw↓	AP↑	F1↑	mIoU↑	FID↓	Diversity↑
StarGANv2 [8]	2.472 ± 0.726	14.691 ± 3.987	31.071 ± 15.769	0.811 ± 0.086	0.806 ± 0.077	0.971 ± 0.012	72.910 ± 18.961	0.214 ± 0.051
SMILE$_{\text{SEM}}$	1.948 ± 0.450	13.225 ± 3.428	9.439 ± 1.826	0.942 ± 0.030	0.928 ± 0.031	0.989 ± 0.002	50.257 ± 24.735	0.129 ± 0.083
(*C*): SMILE$_{\text{SEM}}$ −WC	2.182 ± 0.652	17.142 ± 6.113	9.117 ± 1.280	0.943 ± 0.031	0.930 ± 0.029	0.986 ± 0.002	52.327 ± 23.352	0.111 ± 0.064
(*B*): SMILE$_{\text{SEM}}$ −WC −MC	2.277 ± 0.595	16.362 ± 4.304	14.952 ± 5.364	0.919 ± 0.047	0.909 ± 0.043	0.987 ± 0.003	53.298 ± 23.361	0.132 ± 0.057
(*A*): SMILE$_{\text{SEM}}$ −WC −MC −DS	2.011 ± 0.698	10.811 ± 4.247	13.765 ± 7.567	0.899 ± 0.063	0.887 ± 0.060	0.994 ± 0.001	65.863 ± 26.084	0.136 ± 0.058

Table 1: Quantitative contribution of each component of our system for Latent Synthesis manipulation (upper part) and Exemplar Image manipulation (lower part). ↓ and ↑ mean that lower is better and higher is better, respectively. Note that Diversity computes the LPIPS perceptual *dissimilarity* across different styles for a single input, therefore higher is better. WC, MC, and DS stand for Weighted Classes, Modulated Convolutions, and Detaching Style, respectively.

Semantic Manipulation There are two main aspects we consider for the proper evaluation of our system: the transformation mapping must resemble the pose of the input, and the output semantic map must contain the target attributes. To this end, we use an off-the-shelf pose estimator [31] (HopeNet) and use the training set of CelebA-HQ to train an attribute classifier using MobilenetV2 [33]. For the entire CelebA-HQ test set, we manipulate each image using a specific attribute and keeping the others unaltered (for instance only male ↔ female), and we perform 10 transformations per image. We then extract the average Yaw, Pitch and Roll using HopeNet, and the average attribute scores using MobilenetV2, that is, the Root-Mean Squared Error (RMSE), Average Precision (AP), and F1 score between the test set and generated images per attribute, for the entire set of attributes. Additionally, following the same protocol, we also report the FID between the training set and generated images for each attribute. In addition to FID, we compute the perceptual Diversity metric across each image in the test set and 10 different transformations. For both perceptual metrics, we follow the same validation protocol as in StarGANv2. Furthermore, we also report the mean Intersection over Union (mIoU) over the input image and the reconstructed cycle image. (Equation 6).

Semantic Image Synthesis As it is common for image synthesis, we report the Fréchet Inception Distance [32] (FID), and the Perceptual Similarity Score (LPIPS) [43] as a measure of dissimilarity across transformations (Diversity). We strictly follow the same evaluation framework proposed in StyleGAN2 for FID. Since we have to use real semantic annotations for the evaluation protocol, we use 10,000 samples (the entire test set of FFHQ) to compute the FID score. For Diversity, we generate 10 different samples from a single semantic input, and compute the LPIPS score across each pair, for all possible pairs. In our case, the LPIPS score is associated with diversity rather than similarity.

4. Discussion

In this section, we discuss in detail the aspects that strengthen our method. We depict in Figure 3 and 4, and quantitatively evidence in Table 1 and 2 each part of our system. Note that the numbers reported in Table 1 are the average scores for the 8 different attribute manipulations. Please refer to the Supplementary Material for a detailed table for each attribute manipulation.

As our framework is an ensemble of two approaches, *i.e.*, (SMILE$_{\text{SEM}}$) and (SMILE$_{\text{SYN}}$), we do not compare directly with purely RGB methods. Conversely, we validate

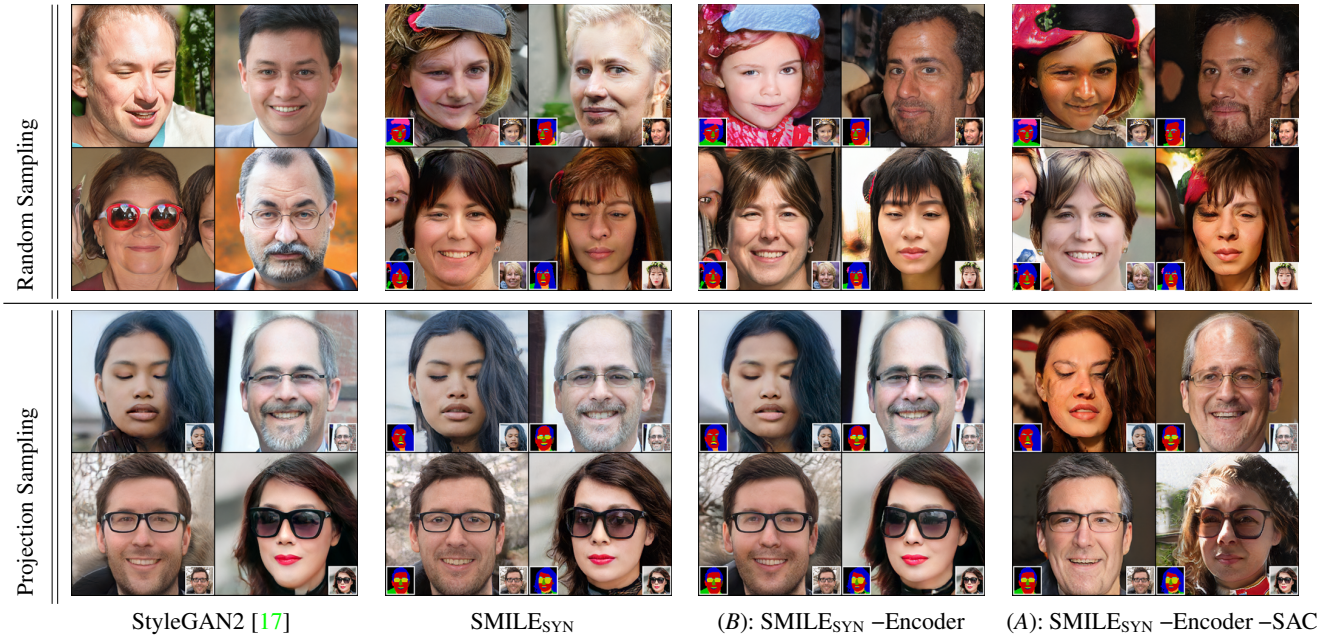


Figure 4: Ablation experiments for image synthesis. We present our full system and without our two major contributions for SMILE$_{SYN}$: SAC layers and Reference Style Encoder. We show StyleGAN2 for reference. Upper and bottom parts of the figure shows qualitative results for random sampling and projection reconstruction, respectively. We show the semantic maps and reference images in the bottom corners of each image. Zoom in for better details.

FFHQ [16]										
Experiment	Latent Synthesis			Reference Synthesis					Training [days]	# Params [millions]
	FID↓	Diversity↑	Runtime [s/img]	LPIPS↓	PSNR↑	SSIM↑	RMSE↓	Runtime [s/img]		
StyleGAN2 [17]	15.15[2]	-	0.03	0.14 ± 0.02	20.13 ± 1.14	0.66 ± 0.03	0.11 ± 0.02	120	2.5	30.0
SMILE$_{SYN}$	16.99	0.43 ± 0.03	0.13	0.21 ± 0.06	18.19 ± 2.84	0.54 ± 0.10	0.13 ± 0.03	0.13	17.5	42.2
(*B*): SMILE$_{SYN}$ −Encoder	13.08	0.42 ± 0.04	0.13	0.18 ± 0.06	17.86 ± 2.97	0.60 ± 0.09	0.13 ± 0.05	210	9.4	39.9
(*A*): SMILE$_{SYN}$ −Encoder −SAC	24.12	0.08 ± 0.03	0.60	0.42 ± 0.07	10.38 ± 2.07	0.34 ± 0.08	0.31 ± 0.07	180	3.8	36.1
SPADE [27]	-	-	-	0.40 ± 0.02	12.33 ± 0.69	0.40 ± 0.03	0.25 ± 0.02	0.56	1	92.5
SEAN [47]	-	-	-	0.24 ± 0.02	16.68 ± 0.82	0.52 ± 0.03	0.15 ± 0.01	0.28	4	266.9

Table 2: Image synthesis quantitative evaluation under different configurations, and in comparison with recent works.

our framework as independent stages and perform an extensive comparison with strong baselines, and state-of-the-art methods in the image synthesis task.

4.1. Semantic Manipulation (SMILE$_{SEM}$)

We use the state-of-the-art method [8] in multi-domain image manipulation as backbone for our method. As StarGANv2 was proposed for mutually exclusive domains, we first extend it to deal with co-existing domains. We apply the concatenation of all the target styles as an input in addition to the RGB image. As Figure 3 shows and Table 1 indicate, StarGANv2 does not generalize well to different domains. Interestingly, we found that StarGANv2 struggles when trained when using domains different from Male/Female. We hypothesize that it is due to the fact that the style encoder extracts general characteristics of the entire image and thanks to the lack of supervision it cannot focus on fine-grained styles. To circumvent this problem, we instead use the high-level semantic information as input and perform manipulations in this space (A), as the shape of each attribute is much easier to handle in the semantic space. This change leads to cleaner and sharper transformations that better approximate the desired domain.

Moreover, we found that (B) disabling the gradients of the style encoder during the reconstruction pass, is sufficient for the overall training framework and reduces the training time by half. Our rationale is as follows: as we are injecting a random style through the mapping function, and the style encoder learns to reconstruct it, then we can assume that after enough iterations the style encoder extracts the corresponding style from the real image.

Simplifying our system for the semantic space brings the problem of losing the texture information. We found that using Adaptive Instance Normalization layers (AdaIN [13]) deforms the input's image pose, in particular Yaw, dur-

[2]We report the StyleGAN2 FID for a model trained with batch size 4 and 300,000 iterations. It is possible that this result does not match with the one reported in the original paper.

ing the transformation, and yet it still minimizes the proposed formulation in Equation 7. Therefore, we noticed that (C) conditionally modulating the weights of the convolutions [17] alleviates this issue and the output resembles the pose of the input.

Furthermore, As we assumed equal dimensionality contribution per domain, transformations produced in stage (C) are not diverse enough across domains. By closely inspecting the resultant images we found that the identity domain is the least diverse in spite of being the most abstract and the one that models the biggest part of the image (for instance facial structure, clothes, hairstyle, etc). Finally, for our proposed approach (SMILE$_{\text{SEM}}$), we weighted the identity domain to have more representation in the final latent vector with respect to the others, so this domain has a bigger impact than other domains (*e.g.*, eyeglasses, bangs, etc) in the reconstruction style loss, and it can consequently produce more diverse transformations. This change of dimensionality for the identity domain is inspired by an observation from the semantic space. Most of the selected domains have one specific corresponding channel in the semantic space that facilitates the style encoding, yet it is not the case for the identity.

Additionally, to further assess a quantitative disentangled level of the manipulations, we studied how each independent transformation affects the unaltered attributes. We accomplished this feat by computing Precision and Recall curves over each manipulation. Please refer to the Supplementary Material for the generated curves.

4.2. Semantically Image Synthesis (SMILE$_{\text{SYN}}$)

StyleGAN2 [17] is the state-of-the-art method in image synthesis. Using this method to perform attribute manipulation or modifying an existing image is very challenging, and usually involves different post-processing techniques. In order to modify StyleGAN2 backbone to be able to perform both disentangled representations in the semantic space and modify existing images, we introduce different subtle but critical changes to the architecture to build SMILE$_{\text{SYN}}$. See Figure 4 and Table 2 for qualitative and quantitative ablative comparison, respectively.

We first (A) replace the StyleGAN2 style condition by semantic information in a SPADE [27], which implies there is not diversity in the generation as it is a deterministic mapping. Next (B), to introduce both semantic information and per-style region into our framework, we replace all the Modulated Convolutions by SAC layers (Equation 8). We use the full style per-region matrix in conjunction with the segmentation mask to generate diverse images controlled by random noise in each region. Only with SAC layers our method outperforms the state-of-the-art for face generation. Finally, in order to generate a fast reference synthesis, we incorporate a style encoder. SMILE$_{\text{SYN}}$ is trained in an alternate fashion by generating random and projected images using either the latent space or the reference style encoder, respectively. As a result of the alignment between the mapping random style and style encoder, we encountered a trade-off in the performance. This trade-off is expected due to the random nature of the latent sampling, *i.e.*, as the mapping network can receive very different styles for skin, neck, and ears, this is not the case for the reference synthesis. This trade-off is strongly evidenced in the runtime that each method requires to reconstruct a reference image. SMILE$_{\text{SYN}}$ compares favorably with StyleGAN2 yet does not require a post-optimization process for image projection. Furthermore, as our ablation study lies in the StyleGAN domain, we also compare our method with fully I2I translation methods. We show comparison with respect to SPADE [27] and SEAN [47], which are reference methods for semantic image synthesis. SMILE$_{\text{SYN}}$ performs better, faster, and requires fewer parameters.

Figure 5: Our model learns a diverse manipulation for multiple attributes using a single generator and keeping the identity of the input. The style imposition is as follows: eyeglasses, clothes and hair; hair and earrings; and hair, respectively.

As our final goal is to manipulate attributes in the RGB space, *i.e.*, SMILE = SMILE$_{\text{SYN}}$ $\circ$ SMILE$_{\text{SEM}}$, see Figure 1 and 5 for a complete visualization of multi-attribute semantic manipulation and image synthesis.

5. Conclusions

We introduced SMILE, a method for multi-attribute image-to-image translation using random sampling or image guiding reference. We show that using a semantic segmentation space as an intermediate step is a much easier manipulation task, which can be further transformed into RGB by more sophisticated semantically driven image synthesis schemes. We extensively show that our method outperforms previous state-of-the-art baselines StarGANv2 [8] and StyleGAN2 [17] for both image manipulation and image synthesis.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4432–4441, 2019. 5

[2] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 195–204, 2018. 1

[3] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[4] Sagie Benaim, Michael Khaitov, Tomer Galanti, and Lior Wolf. Domain intersection and domain difference. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3445–3453, 2019. 2, 3

[5] Marcel Christoph Bühler, Andrés Romero, and Radu Timofte. DeepSEE: deep disentangled semantic explorative extreme super-resolution. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2

[6] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40–48, 2018. 1, 2

[7] 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 (CVPR)*, pages 8789–8797, 2018. 1, 2

[8] 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 (CVPR)*, 2020. 1, 2, 3, 6, 7, 8

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 2

[10] Jingtao Guo, Zhenzhen Qian, Zuowei Zhou, and Yi Liu. Mulgan: Facial attribute editing by exemplar. *arXiv preprint arXiv:1912.12396*, 2019. 1, 2

[11] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 1, 2

[12] Sarah Jane Hong, Martin Arjovsky, Ian Thompson, and Darryl Barnhardt. Low distortion block-resampling with spatially stochastic networks. *arXiv preprint arXiv:2006.05394*, 2020. 5

[13] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 2, 7

[14] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 1

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 5, 6

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 2, 3, 5, 7

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 3, 4, 5, 7, 8

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[19] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5967–5976, 2017. 2

[20] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5549–5558, 2020. 2, 3, 5

[21] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Kumar Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, pages 1–16, 2020. 1

[22] Minxuan Lin, Fan Tang, Weiming Dong, Xiao Li, Chongyang Ma, and Changsheng Xu. Distribution aligned multimodal and multi-domain image stylization. *arXiv preprint arXiv:2006.01431*, 2020. 2

[23] 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 *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3673–3682, 2019. 1

[24] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2437–2445, 2020. 2

[25] Ron Mokady, Sagie Benaim, Lior Wolf, and Amit Bermano. Mask based unsupervised content transfer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 2

[26] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. SESAME: semantic editing of

scenes by adding, manipulating or erasing objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 7, 8

[28] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[29] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018. 1, 2

[30] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. SMIT: stochastic multi-label image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 0–0, 2019. 1, 2

[31] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2074–2083, 2018. 6

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2234–2242, 2016. 6

[33] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. 6

[34] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[35] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[36] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018. 3, 5

[37] Yaxing Wang, Abel Gonzalez-Garcia, Luis Herranz, and Joost van de Weijer. Controlling biases and diversity in diverse image-to-image translation. *Computer Vision and Image Understanding*, 202:103082, 2019. 1, 2

[38] 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. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5914–5922, 2019. 1

[39] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Dna-gan: Learning disentangled representations from multi-attribute images. In *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, 2018. 2

[40] 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. 1, 2

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4471–4480, 2019. 2

[42] 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 (NeurIPS)*, pages 2994–3004, 2019. 2

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 6

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

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 1

[46] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 465–476, 2017. 1

[47] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5104–5113, 2020. 2, 3, 5, 7, 8