

Describe What to Change: A Text-guided Unsupervised Image-to-Image Translation Approach

Yahui Liu

University of Trento, Italy
FBK, Italy
yahui.liu@unitn.it

Marco De Nadai

FBK, Italy
denadai@fbk.eu

Deng Cai

The Chinese University of
Hong Kong, China
thisisjcykcd@gmail.com

Huayang Li

Tencent AI Lab, China
alanili@tencent.com

Xavier

Alameda-Pineda
Inria Grenoble
Rhône-Alpes, France
xavier.alameda-
pineda@inria.fr

Nicu Sebe

University of Trento, Italy
Huawei Research, Ireland
niculae.sebe@unitn.it

Bruno Lepri

FBK, Italy
lepri@fbk.eu

ABSTRACT

Manipulating visual attributes of images through human-written text is a very challenging task. On the one hand, models have to learn the manipulation without the ground truth of the desired output. On the other hand, models have to deal with the inherent ambiguity of natural language. Previous research usually requires either the user to describe all the characteristics of the desired image or to use richly-annotated image captioning datasets. In this work, we propose a novel unsupervised approach, based on image-to-image translation, that alters the attributes of a given image through a command-like sentence such as “*change the hair color to black*”. Contrarily to state-of-the-art approaches, our model does not require a human-annotated dataset nor a textual description of all the attributes of the desired image, but only those that have to be modified. Our proposed model disentangles the image content from the visual attributes, and it learns to modify the latter using the textual description, before generating a new image from the content and the modified attribute representation. Because text might be inherently ambiguous (blond hair may refer to different shades of blond, e.g. golden, icy, sandy), our method generates multiple stochastic versions of the same translation. Experiments show that the proposed model achieves promising performances on two large-scale public datasets: CelebA and CUB. We believe our approach will pave the way to new avenues of research combining textual and speech commands with visual attributes.

CCS CONCEPTS

- Computing methodologies → Computer vision; Machine learning.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '20, October 12–16, 2020, Seattle, WA, USA

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7988-5/20/10...\$15.00
<https://doi.org/10.1145/3394171.3413505>

KEYWORDS

GANs, image translation, text-guided image manipulation, unsupervised learning

ACM Reference Format:

Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. 2020. Describe What to Change: A Text-guided Unsupervised Image-to-Image Translation Approach. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA*. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3394171.3413505>

1 INTRODUCTION

Editing image attributes on portable devices is an uncomfortable operation for most people, but especially for impaired users. On smartphones, changing the hair color on a picture requires to manually select the pixels to be altered and color them properly. Inspired by the emergence of voice assistants, which provide us a more convenient way to interact with machines through human language, we make pilot attempts to manipulate images through textual commands. Meeting this goal requires a deep integration of vision and language techniques, as previously achieved for other tasks, namely: image captioning [23, 52], text-to-image generation [25, 26, 38, 42, 53], text-based video editing [13] and drawing-based image editing [22, 39]. However, modifying parts of the image through natural language is still a challenging research problem.

Existing approaches, based on Generative Adversarial Networks (GANs) [1, 15], require either to describe all the characteristics of the desired image [42, 59] (e.g. “*young oval-faced lady with blond and black hair, smiling*”), or collecting a richly detailed dataset of supervised actions [11, 12] or captions [26, 38, 42, 43, 51]. For example, TAGAN [38] uses text input to change an image in order to have the attributes explained in the provided text description, but this method relies on detailed human-written captions. Similarly, El-Nouby *et al.* [11, 12] rely on annotated datasets to interact with the user and to generate semantically-consistent images from textual instructions. Two main issues have hindered the possibility to have an effective and unsupervised approach that follows human-written commands (e.g. “*make him smile*”). On the one hand, the literature on text-based image manipulation does not explicitly model image attributes, which would indicate if a specific part of the image is

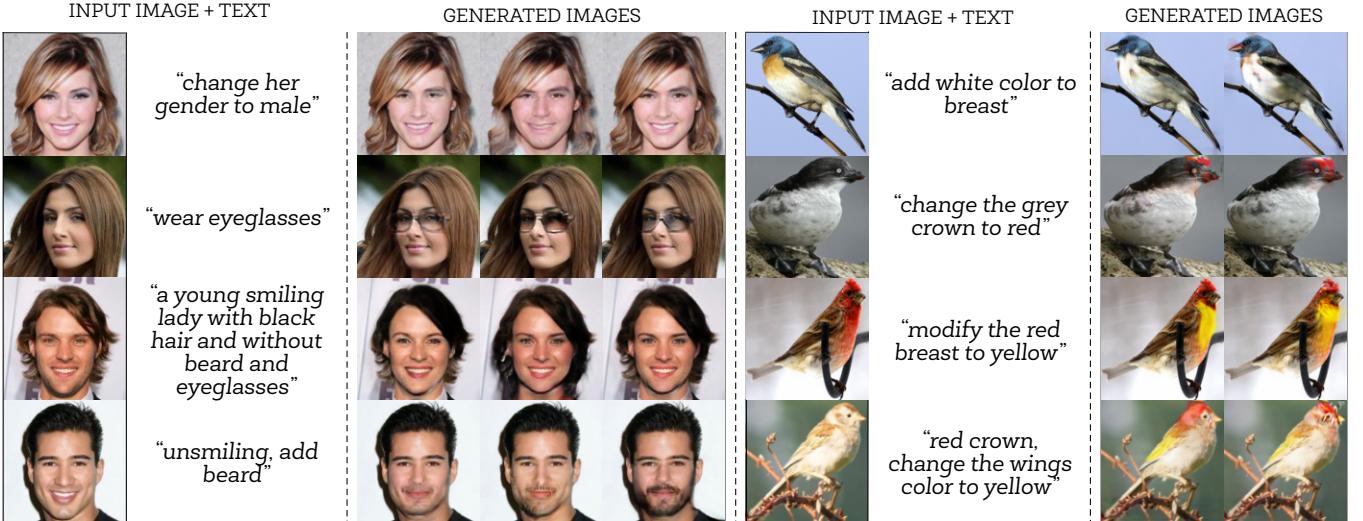


Figure 1: Our model allows to manipulate visual attributes through human-written text. To deal with the inherent ambiguity of textual commands, our model generates multiple versions of the same translation being as such multi-modal. Here, we see some examples of generated images from the CelebA [31] and CUB [48] datasets.

going to be changed or not. On the other hand, most GAN models assume a deterministic mapping, strongly limiting the capacity to handle the inherent ambiguity of human textual commands (e.g. “blond hair” might mean different shades of blond colors).

In this paper, we propose *Describe What to Change* (DWC), a novel method based on GANs to manipulate visual attributes from a human-written textual description of the required changes, thus named DWC-GAN. To explicitly model visual attributes and cope with the text ambiguity, our model translates images from one visual domain to other visual domains in a multi-modal (stochastic) fashion. Our contribution is three-fold:

- We propose the use of textual commands instead of image captions. As shown in Table 1, this leads to three advantages: (1) the commands can be more flexible than image captions and we can compose them for progressive manipulation; (2) the commands can be automatically generated, and (3) users are not required to know and mention all the modeled visual attributes in an input image, only the desired changes. DWC-GAN thus does not rely on human-annotated captioning text compared with state-of-the-art approaches [26, 38, 42, 43].
- Our model disentangles the content features and the attribute representation from the input image making the manipulation more controllable and robust. The attribute space follows a Gaussian Mixture Model (GMM), where the attributes of the source and the modified image follow the Gaussian distribution associated to their respective domains.
- To the best of our knowledge, we are the first to manipulate images through textual commands in a multi-modal fashion, allowing multiple versions of the same translation.

2 RELATED WORK

Our work is best placed in the literature of image-to-image translation and text-to-image generation. The former aims to transform

Input text	Features
Captioning text: This woman is old and smiling, with black hair, eyeglasses and no beard.	- Enumerates all the attributes; - No need to understand attributes from the input image; - Often trained with positive and negative annotation for each translation pair [26, 38].
Textual command: Change the hair to be black, increase her age.	- Focuses only on the differences; - Requires the extraction of attributes from the input image; - Trained without ground truth for each translation pair.

Table 1: Example of text describing a translation from an image of a young smiling woman with blond hair and eyeglasses to an older smiling woman with black hair and eyeglasses. Differently from captioning text, users are not required to know and mention all the modeled attributes.

an input image belonging to a visual domain (e.g. young people) to another domain (e.g. elderly people), while the latter tries to generate images from a textual description provided as input. These two fields have witnessed many improvements in quality and realism, thanks to the advent of GANs [15, 35] and, in particular, conditional GANs (cGANs) [35], which are conditioned by extra information (e.g. the hair colour in face generation). GANs aim to synthesize images through a min-max game between a discriminator, trying to discriminate between real and fake data, and a generator, seeking to generate data that resemble the real ones.

Image-to-image translation: Conditional GANs were first employed for image-to-image translation in pix2pix [21] for learning a mapping from one domain to another by minimizing the L_1 loss between the generated and the target image. However, pix2pix requires a large amount of paired data (e.g. colored-greyscale images),

which is often unrealistic to collect. Thus, the interest had shifted to unsupervised models that learn to translate images with unpaired data, for which the image content might vary (e.g. daylight-night images can have different people and cars). These models require additional constraints to better differentiate the domain-dependent parts of the image from the domain-independent ones. For example, Liu *et al.* [31] assume that the two domains share a common latent space. CycleGAN [57] requires an image translated from a domain A to a domain B to be translated back to A, and applies a consistency loss, which was later employed by many other works [7, 20, 30, 36]. MUNIT [20] assumes that images share a domain-invariant content latent space and a style latent space, which is specific to a domain. This choice allows for generating multiple samples of the translation by drawing from the style distribution. Most of these approaches are, however, limited to one-to-one domain translation, which requires training a large number of models in the case of multiple domains. StarGAN [7] proposed a unified approach for multi-domain translation by employing a binary domain vector as input, which specifies the target domain, and a domain classification loss that helps the translation. Recently, GMM-UNIT [29] proposed a unified approach for multi-domain and multi-modal translation by modeling attributes through a Gaussian mixture, in which each Gaussian component represents a domain. In this paper, inspired by GMM-UNIT, we build a translation system where both the original and the manipulated attributes follow a GMM.

Text-to-image generation: cGANs can be conditioned with complex information, such as human-written descriptions, and then can generate accurate images, as shown by Reed *et al.* [42, 43]. For example, StackGAN [53] uses two stages of GANs in order to generate high-resolution images. The first stage generates a low-resolution image with the primitive shape and colour of the desired object, while the second one further refines the image and synthesize 256x256 images. The same authors improved the model with multiple generators through StackGAN++ [54]. Qiao *et al.* [41] focused instead on semantic consistency, enforcing the generated images to have the same semantics with the input text description. In other words, the caption synthesized from the generated image should have the same meaning as the input text. Recently, Li *et al.* proposed StoryGAN [28], which generates a series of images that are contextually coherent with previously generated images and with the sequence of text descriptions provided by the user.

These approaches, however, require to accurately describe the picture to be generated, without allowing to start from an existing image and modify it through text.

Image manipulation: Modifying images through some user-defined conditions is a challenging task. Most of the previous approaches rely on conditional inpaiting [3, 22, 56], in which the network fills a user-selected area with pixel values coherent with the user preferences and the context. However, image editing does not necessarily require to select the exact pixels that have to be changed. Wang *et al.* [49] learned to change global image characteristics such as brightness and white balance from textual information. Zou *et al.* [59] proposed a network that colourizes sketches following the instructions provided by the input text specifications. El-Nouby *et al.* [11, 12] introduced a network that generates images from a dialogue with the user. A first image is generated from text, then objects are added according to the user preference provided as free

text. Chen *et al.* [5] learned with a recurrent model and an attention module to colourize greyscale images from a description of the colours or the scene (e.g. *“the afternoon light flooded the little room from the window”*). Recently, Cheng *et al.* [6] proposed a network, tailored for dialogue systems, that gets as inputs an image and a series of textual descriptions that are apt to modify the image. The network is encouraged to synthesize images that are similar to the input and the previously generated images. However, examples show that the content of the image might vary substantially. Nam *et al.* [38] and Li *et al.* [26] introduced instead TAGAN and ManiGAN that modify fine-grained attributes such as the colours of birds’ wings through text. The key idea of their methods is reconstructing the images with positive and negative sentences, where the positive/negative refers to captioning text matching/mismatching the corresponding image. The recurrent network is trained to learn the words that refer to the attributes in the images and it allows to change words to manipulate multiple attributes at once.

Current image manipulation methods based on text often rely on human-annotated captioning text [26, 38, 42, 43, 51], which describes the scene in the image. Moreover, they do not explicitly model the concept of attributes (i.e. domains), sometimes failing at balancing the trade-off between keeping the old content and changing it [38], which is a well-known issue in image-to-image translation. Last but not least, existing solutions are deterministic, thus limiting the diversity of the translated outputs, even if the provided text might have different meanings. For example, “*wear a beard*” might mean to have a goatee, short or long beard.

Therefore, we build on state-of-the-art approaches for image-to-image translation and on the image manipulation literature and propose to explicitly model the visual attributes of an image and learn how to translate them through automatically generated commands. We are, up to the best of our knowledge, the first proposing a multi-domain and multi-modal method that translates images from one domain to another one by means of the user descriptions of the attributes to be changed.

3 METHODOLOGY

Our model is designed to modify the attributes of an image \mathbf{x} through a human-written text \mathbf{t} , which describes the changes to be made in the image. In other words, we want to generate a new image $\hat{\mathbf{x}}$ that has the visual content of \mathbf{x} and all its attributes but those attributes that should be altered accordingly to the text \mathbf{t} . Different from previous works, \mathbf{t} is not a description of all the attributes of the generated image, but it is a direct command describing the modifications to be made (e.g. *change the hair color to black*). Thus, the proposed method should be able to model the attributes of the original image, the modifications of (some of) the attributes described in the text, and the attributes of the generated image. Moreover, it has to deal with the ambiguity of input text, which does not describe all the attributes of the target image, nor it is direct and clear as one-hot vector inputs.

The architecture, as shown in Figure 2, processes the input as follows. Given N attributes of a set of images, we first disentangle the visual content $\mathbf{c} \in \mathcal{C} \subset \mathbb{R}^C$ and the attributes $\mathbf{a} \in \mathcal{A} \subset \mathbb{R}^N$ of the input image \mathbf{x} through the encoders E_c and E_a , respectively. Then, E_t , which is a Recurrent Neural Network (RNN) module,

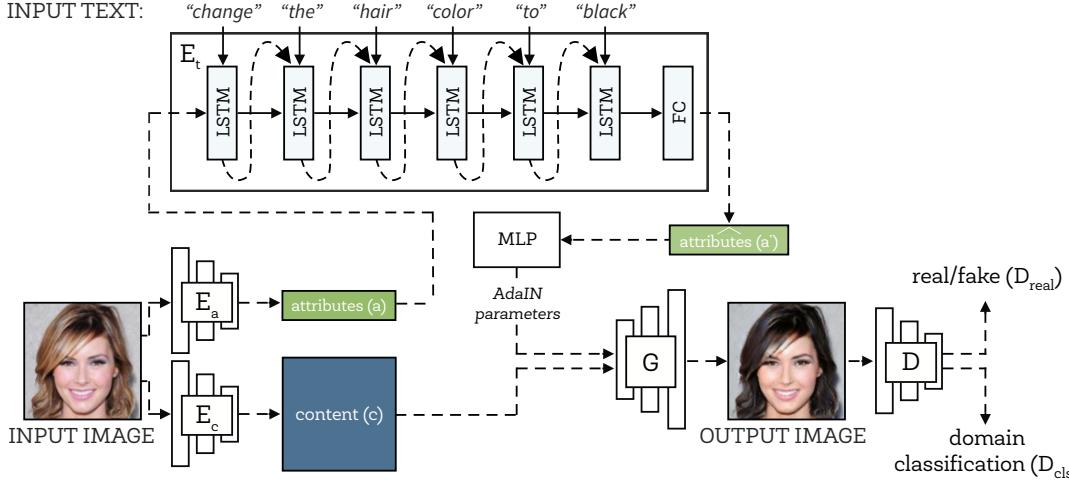


Figure 2: Architecture of our model. First, we disentangle the attributes and the content of the input image. Then, we modify the visual attributes of the original image using a text encoder. The generator uses a MLP to produce a set of AdaIN [19] parameters from the attribute representation. The content features is then processed by the parameterized generator to predict a realistic image with required attributes. Finally, the discriminator classifies whether the generated image is real or fake.

uses the input sentence t and the extracted attributes a to infer the desired (target) attributes $a' = E_t(t, a)$. The inferred target attributes are then used along with the extracted content c , by the generator G , to output the image $G(c, a')$. Finally, the discriminator D discerns between “real” or “fake” images (D_{real}) and recognizes the domain of the generated image (D_{cls}). Here, a domain is a combination of visual attributes. The model (content and attribute extractors, generator and discriminators) is learned in an end-to-end manner.

3.1 Assumptions

The basic assumption of DWC-GAN is that each image can be decomposed in a domain-invariant content space and a domain-specific attribute space [14, 20, 29]. We let the network model the high-dimension content features, while we represent the attributes of the image through Gaussian components in a mixture. This representation allows to model the combinations of attributes in a continuous space. We can then exploit this space to work with attribute combinations that have never (or little) been observed in the data. Formally and similarly to GMM-UNIT [29], we model the attributes with a K -component d -dimensional GMM: $p(a) = \sum_{k=1}^K \phi_k \mathcal{N}(a; \mu^k, \Sigma^k)$ where ϕ_k , μ^k and Σ^k denote respectively the weight, mean vector and covariance matrix of the k -th GMM component ($\phi_k \geq 0$, $\sum_{k=1}^K \phi_k = 1$, $\mu^k \in \mathbb{R}^d$ and $\Sigma^k \in \mathbb{R}^{d \times d}$ is symmetric and positive definite). Therefore, we model the attributes in a domain through a corresponding GMM component. Thus, for an image x from domain \mathcal{X}^k (i.e. $x \sim p_{\mathcal{X}^k}$), its latent attribute a is assumed to follow the k -th Gaussian component: $a^k \sim \mathcal{N}(\mu^k, \Sigma^k)$. We aim to manipulate a source image x from \mathcal{X}^k into domain \mathcal{X}^ℓ through a textual command $t^\ell \sim p_{\mathcal{T}^\ell}$, where $k, \ell \in \{1, \dots, K\}$.

3.2 Multi-modal image generation

In order to have a multi-modal (non-deterministic) translation, we enforce both the sampled and extracted attributes to follow the

image domain distribution, that is $\forall \ell, k$:

$$G(E_c(x), a) \sim p_{\mathcal{X}^k}, \quad \forall a \sim \mathcal{N}(\mu^k, \Sigma^k), x \sim p_{\mathcal{X}^\ell}.$$

We ensure these two properties with the following losses.

Reconstruction losses. These losses, introduced at first in [20, 57], force the output to be consistent with the original content and attributes. Specifically, the *self-reconstruction loss* ensures that the original image is recovered if its attribute and content is used in the generator. If $c_x = E_c(x)$ and $a_x = E_a(x)$, then for all k :

$$\mathcal{L}_{rec_s} = \mathbb{E}_{x \sim p_{\mathcal{X}^k}} [\|G(c_x, a_x) - x\|_1].$$

The *content and attribute reconstruction losses* [20] are used to constrain and learn the content and attribute extractors:

$$\mathcal{L}_{rec_c} = \mathbb{E}_{x \sim p_{\mathcal{X}^k}, a \sim \mathcal{N}(\mu^k, \Sigma^k)} [\|E_c(G(c_x, a)) - c_x\|_1],$$

$$\mathcal{L}_{rec_a} = \mathbb{E}_{x \sim p_{\mathcal{X}^k}} [\|E_a(G(c_x, a)) - a_x\|_1].$$

The *cycle reconstruction loss* [57] enforces consistency when translating an image into a new domain and then back to the original one:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x \sim p_{\mathcal{X}^k}, a \sim \mathcal{N}(\mu^\ell, \Sigma^\ell)} [\|G(E_c(G(c_x, a)), a_x) - x\|_1],$$

where the use of the \mathcal{L}_1 loss inside the expectation is motivated by previous works [21] showing that \mathcal{L}_1 produces sharper results than \mathcal{L}_2 . To encourage the generator to produce diverse images, we explicitly regularize G with the *diversity sensitive loss* [8, 32]:

$$\mathcal{L}_{ds} = \mathbb{E}_{x \sim p_{\mathcal{X}^k}, a_1, a_2 \sim \mathcal{N}(\mu^\ell, \Sigma^\ell)} [\|G(c_x, a_1) - G(c_x, a_2)\|_1]$$

Domain losses. Similarly to StarGAN [7], for any given input image x , we would like the method to classify it as its original domain, and to be able to generate an image in any domain from its content. Therefore, we need two different losses, one directly applied to the original images, and a second one applied to the generated images:

$$\mathcal{L}_{cls}^D = \mathbb{E}_{x \sim p_{\mathcal{X}^k}, d_{\mathcal{X}^k}} [-\log D_{cls}(d_{\mathcal{X}^k} | x)] \quad \text{and}$$

$$\mathcal{L}_{cls}^G = \mathbb{E}_{x \sim p_{\mathcal{X}^k}, d_{\mathcal{X}^\ell}, a \sim \mathcal{N}(\mu^\ell, \Sigma^\ell)} [-\log D_{cls}(d_{\mathcal{X}^\ell} | G(c_x, a))],$$

where $d_{\mathbf{x}^\ell}$ is the label of ℓ -th domain. Importantly, the discriminator D is trained using the first loss, while the generator G is trained using the second loss.

Adversarial losses. These terms enforce the generated images to be indistinguishable from the real images by following the formulation of LSGAN [33]:

$$\begin{aligned}\mathcal{L}_{GAN}^D &= \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{X}^\ell}} [D_{\text{real}}(\mathbf{x})^2] + \\ &\quad \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{X}^\ell}, \mathbf{a} \sim \mathcal{N}(\boldsymbol{\mu}^\ell, \boldsymbol{\Sigma}^\ell)} [(1 - D_{\text{real}}(G(\mathbf{c}_x, \mathbf{a})))^2] \\ \mathcal{L}_{GAN}^G &= \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{X}^\ell}, \mathbf{a} \sim \mathcal{N}(\boldsymbol{\mu}^\ell, \boldsymbol{\Sigma}^\ell)} [D_{\text{real}}(G(\mathbf{c}_x, \mathbf{a}))^2]\end{aligned}$$

3.3 Attribute manipulation losses

As mentioned above, the underlying statistical assumption for the attributes is to follow a GMM – one component per domain. Very importantly, both the attribute representations extracted from image by E_a and the attribute representations obtained from E_t should follow the assumption and correspond to the correct component. We do that by imposing that the Kullback–Leibler (KL) divergence of the extracted and manipulated attributes correspond to the Gaussian component of the original and targetted domain respectively. Recalling that $\mathbf{a}_x = E_a(\mathbf{x})$, we write:

$$\begin{aligned}\mathcal{L}_{KL} &= \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{X}^\ell}, \mathbf{t} \sim p_{\mathbf{T}^\ell}} [\mathcal{D}_{KL}(\mathbf{a}_x \| \mathcal{N}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)) + \\ &\quad \mathcal{D}_{KL}(E_t(\mathbf{t}, \mathbf{a}_x) \| \mathcal{N}(\boldsymbol{\mu}^\ell, \boldsymbol{\Sigma}^\ell))]\end{aligned}$$

where $\mathcal{D}_{KL}(p \| q) = - \int p(t) \log \frac{p(t)}{q(t)} dt$ is the KL.

Intuitively, the second KL divergence enforces the images generated from the manipulated attributes to follow the distribution of the target domain:

$$G(\mathbf{c}_x, E_t(\mathbf{t}, \mathbf{a}_x)) \sim p_{\mathbf{X}^\ell}, \quad \forall \mathbf{x} \sim p_{\mathbf{X}^\ell}, \mathbf{t} \sim p_{\mathbf{T}^\ell}$$

Finally, the full objective function of our network is:

$$\begin{aligned}\mathcal{L}_D &= \mathcal{L}_{GAN}^D + \lambda_{cls} \mathcal{L}_{cls}^D \\ \mathcal{L}_G &= \mathcal{L}_{GAN}^G + \lambda_{rec_s} \mathcal{L}_{rec_s} + \mathcal{L}_{rec_c} + \mathcal{L}_{rec_a} + \\ &\quad \lambda_{cyc} \mathcal{L}_{cyc} - \mathcal{L}_{ds} + \lambda_{KL} \mathcal{L}_{KL} + \mathcal{L}_{cls}^G\end{aligned}$$

where $\{\lambda_{rec_s}, \lambda_{cyc}, \lambda_{KL}\}$ are hyper-parameters of weights for corresponding loss terms. The value of most of these parameters come from the literature. We refer to Supplementary Material for the details. Note that these losses are required to constrain down the difficult problem of image to image translation, disentangle content and style, and achieve stochastic results by sampling from the GMM distribution of attributes. Without these losses, the problem would be much less constrained and difficult.

3.4 Domain Sampling

We propose a strategy to obtain diverse manipulated results during testing by following the assumption introduced in Section 3.1. After extracting the target attributes \mathbf{a}' from the original image \mathbf{x} and input text \mathbf{t} , we assign it to a closest component $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ of the GMM by following: $k^* = \arg \max_{k \in \{1, \dots, K\}} \{\Phi(\mathbf{a}'; \boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)\}$, where Φ is the Gaussian probability density function. Then, we can randomly sample several \mathbf{a}^* from the component $\mathcal{N}(\boldsymbol{\mu}^{k^*}, \boldsymbol{\Sigma}^{k^*})$. With such different sampled \mathbf{a}^* 's and the trained non-linear generator G , our model achieves multi-modality in any domain, which is different

from state-of-the-art approaches [38, 54] that use random noise sampled from a standard normal distribution $\mathcal{N}(0, 1)$ to generate multiple diverse results independently of the domain.

Sampling all the target attributes might hurt translation performance. Thus, we propose an additional constraint that the ideal attribute representation is a mixture of domain sampling \mathbf{a}^* and extracted representation $E_a(\mathbf{x})$, where \mathbf{a}^* copies the attribute representation from $E_a(\mathbf{x})$ when a target attribute is the same as the attribute in the input image.

3.5 Unsupervised Attention

Attention mechanisms have proven successful in preserving the visual information that should not be modified in an image-to-image translation [29, 30, 34, 40]. For example, GMM-UNIT [29] claims that models are inclined to manipulate the intensity and details of pixels that are not related to the desired attribute transformation. Therefore, we follow the state-of-the-art approaches and add a convolutional layer followed by a sigmoid layer at the end of G to learn a single channel attention mask M in an unsupervised manner. Then, the final prediction $\tilde{\mathbf{x}}$ is obtained from a convex combination of the input and the initial translation: $\tilde{\mathbf{x}} = \hat{\mathbf{x}} \cdot M + \mathbf{x} \cdot (1 - M)$.

4 EXPERIMENTAL SETUP

4.1 Datasets

CelebA. The CelebFaces Attributes (CelebA) dataset [31] contains 202,599 face images of celebrities where each face is annotated with 40 binary attributes. This dataset is composed of some attributes that are mutually exclusive (e.g. either male or female) and those that are mutually inclusive (e.g. people could have both blond and black hair). To be consistent with previous papers [7, 29], we select a subset of attributes, namely *black hair*, *blond hair*, *brown hair*, *male/female*, *young/old*, *smile*, *eyeglasses*, *beard*. As we model these attributes as different GMM components, our model allows the generation of images with up to 2^8 different combinations of the attributes. Note that in the dataset only 100 combinations exist. As preprocessing, we center crop the initial 178×218 images to 178×178 and resize them to 128×128 . We randomly select 2,000 images for testing and use all remaining images for training.

CUB Birds. The CUB dataset [48] contains 11,788 images of 200 bird species, and provides hundreds of annotated attributes for each bird (e.g. color, shape). We select some color attributes *{grey, brown, red, yellow, white, buff}* of three body parts *{crown, wing, breast}* of the birds, yielding a total of $K = 18$ attributes used to represent around one thousand domains appearing in the dataset. Differently from the CelebA dataset, the selected colors could appear in all selected body parts. Hence, the colors and body parts follow a hierarchical structure, requiring the model to modify the color of a specific body part. We crop all images by using the annotated bounding boxes of the birds and resize images to 128×128 .

Additional details can be found in the Supplementary Material.

4.2 Automatic Text Description

The text encoder E_t has to be trained with a dataset describing the changes to be made in the visual attributes of the input image. Therefore, we here propose a protocol to generate the corpus automatically for the datasets provided multiple annotated

attributes or labels. Given an image \mathbf{x} , we can collect a set of attributes $\{p_1, \dots, p_N\}$ and obtain the corresponding real-valued vector $\mathbf{p} \in \{0, 1\}^N$ for generating the description text. We denote \mathbf{p}' as the randomly assigned real-valued vector of a target image. After observing the human-used commands of manipulating images, we propose an automatic description text generation method through three different strategies:

Step-by-step actions. We can compare the differences between \mathbf{p} and \mathbf{p}' element-by-element. For example, if the hair colors in the source image and target image are “brown” and “black”, respectively. We can describe the action like “*change the brown hair to black*”, “*change the hair color to black*”, “*black hair*”. Similarly, we can describe all actions that change the image one by one. For the same attributes in \mathbf{p} and \mathbf{p}' , we can provide empty description or the no-change commands (e.g. “*do not change the hair color*”, “*keep his hair color unchanged*”). Then, the sequence of step-by-step actions are disrupted randomly. A full example for this strategy is “*make the face older, translate the face to be smiling, remove the eyeglasses, do nothing on the gender, change the hair color to black, wear a beard*”.

Overall description. In practice, it’s feasible to describe the attributes in the target images to manipulate images, which is similar to the text-to-image task [26, 38, 42, 43]. Here, we completely describe the appeared attributes in \mathbf{p}' . For example, “*a smiling young man with black hair, wearing a beard and without eyeglasses*”. In addition, the untouched attributes are not regarded as existing attributes in the target image by default.

Mixed description. To increase the generalization, we mix the two previous strategies by randomly mentioning some attributes in \mathbf{p} and specifying the exact change of attributes in \mathbf{p}' . For example, “*translate the young miss to be an old smiling man with black hair and wearing beards and eyeglasses*”. In this case, the attributes of both the input image and input text are mentioned, which is more complex for the language model.

During training, these three strategies are randomly selected for each input image, and are used for both the CelebA and the CUB Birds datasets. More details are provided in the supplementary material. This dataset is released along with the paper to allow future work and better comparisons with our results.

4.3 Metrics

We quantitatively evaluate our model through the image quality, diversity and the accuracy of generated images. We evaluate the image quality through the Fréchet Inception Distance (FID) [17] and the Inception Score (IS) [44], the diversity is measured by the Learned Perceptual Image Patch Similarity (LPIPS) [55], the mean accuracy (Mean Acc) of multi-label classification, while we also measure the accuracy through a user study.

FID and IS. We randomly select 10000 source images and 10000 target images. The models transfer all source images with the same attributes of target images through textual commands, as shown in Section 4.2. Hence, the FID is estimated using 10000 generated images vs the selected target 10000 images. The IS is evaluated using Inception-v3 [46] and 10000 generated images for FID.

LPIPS. The LPIPS distance is defined as the L_2 distance between the features extracted by a deep learning model of two images. This distance has been demonstrated to match well the human perceptual

similarity [55]. Thus, following [20, 58], we randomly select 100 input images and translate them to different domains. For each domain translation, we generate 10 images for each input image and evaluate the average LPIPS distance between the 10 generated images. Finally, we get the average of all distances. Higher LPIPS distance indicates better diversity among the generated images. We show in the results both the mean and standard deviation.

Accuracy and Realism. Through these two metrics we evaluate how humans perceive generated images. *Accuracy* measures whether the attributes of the generated images are coherent with the target attributes (0: incorrect; 1: correct). *Realism* quantifies how realistic a generated image is perceived (0: bad, unrecognizable; 1: neutral, recognizable but with obvious defects; 2: good, recognizable but with slight defects; 3: perfect, recognizable and without any defect). We test Accuracy and Realism by randomly sampling 50 images with the same translation condition and collect more than 20 surveys from different people with various background.

4.4 Baseline Models

StackGAN++ [54] is a text-to-image model that takes text as input and encodes the text into embedding for the decoding network. Here, we train/test this model by using the embedding of the manipulation text as input.

TAGAN [38] and **ManiGAN** [26] are two image-conditioned text-to-image models that take the original image and detailed description text of the target image as input. We follow the original method and train/test the model on our automatic description text.

StarGAN [7] is a unified multi-domain translation model that takes the original image and target attribute vector as input. Here, we add the same RNN module (i.e. the text understanding module) used in our framework to StarGAN to extract attributes from the input text. After that, we feed the embedding of the input text and original image as input. We call this modified version as StarGAN*.

For all the baseline models, we use the same pre-trained word embeddings in 300 dimensions for the input description text. The embeddings are trained by the skip-gram model described in [4] on Wikipedia using fastText¹.



Figure 3: Qualitative evaluation for different textual input on CelebA dataset. Our model generates high-quality images that are consistent with the textual commands.

¹<https://github.com/facebookresearch/fastText>

Method	IS↑	FID↓	LPIPS↑	Accuracy↑	Realism↑	Params ↓	Training Speed (s/iter) ↓
StackGAN++ [54]	1.444	285.48	.292±.053	.000	.00	50.30M	.082±.003
TAGAN [38]	1.178	421.84	.024±.012	.000	.00	44.20M	.804±.001
StarGAN* [7]	2.541	50.66	.000±.000	.256	1.17	54.10M	.130±.001
DWC-GAN (proposed)	3.069	32.14	.152±.003	.885	2.25	32.75M	.172±.010

Table 2: Quantitative evaluation on the CelebA dataset. There is no captioning text on CelebA.

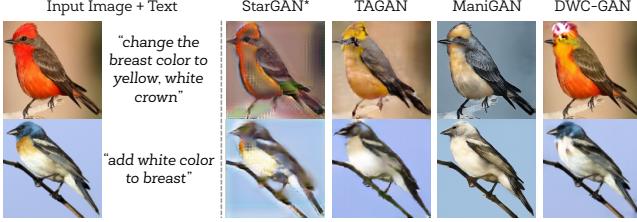


Figure 4: Qualitative comparisons for different textual input on CUB datasets. For reference, we show also the results of StarGAN* [7], TAGAN [38] and ManiGAN [26].

5 RESULTS

We begin by quantitatively comparing our model with state of the art. Table 2 shows that our model generates better quality images than all competing methods, using around half the parameters and achieving comparable training speed. Specifically, DWC-GAN outperforms our most similar competitor, TAGAN, attaining higher IS (3.069 vs 1.178) and lower FID (32.14 vs 421.84) with less required parameters (32.75M vs 44.20M). Surprisingly, the diversity of generated images is lower than StackGAN++. However, the qualitative results show that StackGAN++ generates diverse but very noisy images, in which the attributes are not easily understandable (see the Supplementary Material for details). Contrarily, Figure 3 shows that DWC-GAN generates high-quality images that are also consistent with the manipulation described by the text. Our approach outperforms state of the art in all the metrics in the CUB dataset as well, as shown in both Figure 3 and Table 3. Additional results can be seen in Figure 1 and in the Supplementary Material.

Differently from state of the art, DWC-GAN explicitly models the attributes of images in a smooth and continuous space through a GMM. This allows to translate images in multiple domains, but also to have multi-modal results. As Figure 1 shows, our model generates different realizations of the same manipulation. For example, in “add beards” the model synthesizes various versions of the subjects’ beards, as the command is open to more than one interpretation. State-of-the-art models such as TAGAN and StarGAN* do instead generate results with no, or low, diversity (see the Supplementary Material for additional details).

Our model uses the source attributes and the human-written sentence to generate the desired attributes. The results of StackGAN++ show that it is not feasible to only encode textual commands to generate meaningful target images, while StarGAN* shows that taking into account only the text attributes as an additional condition is not enough. In particular, StarGAN* exhibits a visible mode collapse in this setting. These results show that applying and modifying the existing baselines is not feasible to solve our task.

We trained our model with automatically generated text, thus without relying on human-annotated datasets as state of the art does [38, 42, 43]. We observe that while TAGAN does not perform well with our corpus, we have consistent results with different strategies of automatic text generation, and even with a text strategy that is similar to their dataset (the Overall strategy). This result proves that annotating the dataset with automatically generated textual commands is a feasible strategy to train a model that manipulates images through text.

Human evaluation We evaluated StarGAN* and DWC-GAN by a user study where more than 20 people were asked to judge the Accuracy and Realism of generated results. On average, 82.0% of the images generated by DWC-GAN were judged as correct, while StarGAN* generated only 25.6% of correct results. Regarding Realism, DWC-GAN had an average score of 2.25, while StarGAN* had 1.17. Specifically, in 50.50% of the people said DWC-GAN were perfect, recognizable and without defects while only 2.82% of the people said the same for StarGAN*. StarGAN* results are indeed often judged “neutral, recognizable but with obvious defects” (52.91% of times). StackGAN++ and TAGAN do not achieve accuracy and realism higher than zero because they do not generate recognizable images. We refer to the Supplementary Material for additional details.

These results show that DWC-GAN generates correct images, which can be mistaken with the real images at higher rates than competitors.

Method	IS↑	FID↓	LPIPS↑	Params ↓
StackGAN++ [54]	1.029	278.60	.028±.009	50.30M
TAGAN [38]	4.451	50.51	.060±.024	44.20M
ManiGAN [26]	4.136	11.74	.001±.000	163.34M
StarGAN* [7]	4.343	109.89	.000±.000	54.10M
DWC-GAN (proposed)	4.599	2.96	.081±.001	33.53M

Table 3: Quantitative evaluation on the CUB dataset.

Progressive Manipulation A few state of the art approaches for text-to-image generation and image manipulation are expressly designed for conversational systems [6, 11, 12, 28, 37]. In Figure 15 we show that we can repeatedly apply our method to generated images to have a manipulation in multiple steps. This result highlights the consistency of our generated results, which can be used as input of our model, and that DWC-GAN might be applied in an interactive environment.

Domain Interpolation Our method generates interpolated images from two sentences. Figure 6 shows that by manipulating an image in two different ways, it is possible to generate the intermediate steps that go from one textual manipulation to the other.

User Study on Textual Commands Our training corpus is automatically generated, thus it might suffer limited generalizability.



Figure 5: An example of progressive manipulation. Our method can be used in an interactive environment.

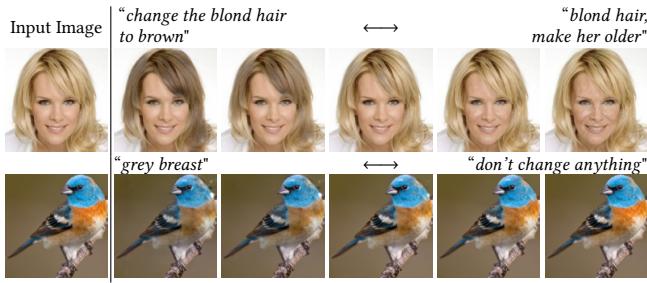


Figure 6: Domain interpolation given an input image.

We conduct an experiment to verify these possible issues by asking 15 different people to describe in natural language how to edit an image A to make it similar to another image B. We asked people to do this experiment for 15 randomly chosen images. The resulting collected textual commands and their corresponding images are fed into our model to evaluate the results. We found that our model can be generalized to real human-written descriptions well. Details are presented in the Supplementary Material.

Attention Visualization We visualize the unsupervised attention mask in Figure 7. It shows the localization area of different manipulations, which indeed helps to focus on the attribute-related pixels and preserves the background pixels.

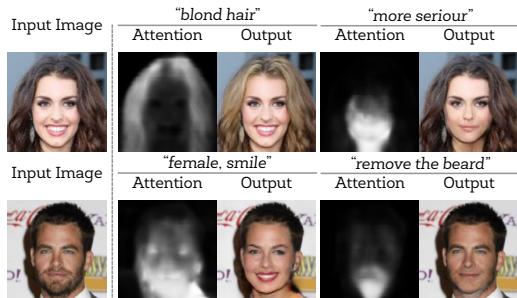


Figure 7: Unsupervised learned attention in DWC-GAN.

Ablation Study We here investigate the effect of removing some components from our architecture. Results are showed in Table 4. First, we remove the *Cycle consistency*, which is widely employed by the literature of image-to-image translation [20, 57, 58]. We observe that both the image quality and diversity significantly performs worse without the cycle consistency, which indicates that the Cycle consistency loss indeed constrains the network on the translation. Then, when we remove the *Attribute reconstruction*, which is applied to learn the attribute extractor E_a . Here, we use an attribute

classifier to measure the F1 score. It shows that the constraint on the attribute latent variable slightly decreases the classification performance (see the Supplementary Material for details on the attribute classifier). However, both image quality and diversity are slightly improved. We also evaluate the contribution of pre-trained word embeddings by using our model without them. We observe that the use of pre-trained word embeddings improves the quality of generated images and converge the training faster. We observe that diversity sensitive loss indeed increases the diversities during the sampling, while slightly reduces the performances on FID and IS. In addition, \mathcal{L}_{cyc} , \mathcal{L}_{rec_a} and using pre-trained embeddings can also have effects on the diversities. We investigate whether the contribution of our model relies only on the use of the GMM to model the attributes. Thus, we constrain down our model to a StarGAN*-like model by setting the covariance matrix of the GMM to zero ($\forall k \Sigma^k = 0$). As expected, we observe that our model behaves similarly but with no diversity (LPIPS: $.000 \pm .000$). Moreover, this proves that our model outperforms StarGAN* even in this case (see Table 2 for a comparison).

Model	IS↑	FID↓	LPIPS↑	F1 (%)↑
DWC-GAN w/o \mathcal{L}_{cyc}	2.589	97.94	.042±.001	96.32
DWC-GAN w/o embedd.	2.782	73.90	.033±.001	98.06
DWC-GAN w/o \mathcal{L}_{rec_a}	2.961	32.35	.072±.001	98.49
DWC-GAN $\forall k \Sigma^k = 0$	2.872	33.17	.000±.000	98.64
DWC-GAN w/o \mathcal{L}_{ds}	3.148	31.29	.061±.001	94.69
DWC-GAN	3.069	32.14	.152±.003	94.80

Table 4: Ablation study performance on the CelebA dataset.

6 CONCLUSION

In this paper, we presented a novel method to manipulate visual attributes through text. Our model builds upon recent literature of image to image translation, using unpaired images and modeling visual attributes through a GMM. Users interact with the model by feeding an image and a text that describes the desired alteration to the original image. The model extracts from the input image the visual attributes, understand the text and the desired transformation, and sample multiple images having the content of the input image and the style of the desired output. By sampling multiple synthesized images, our model deals with the inherent ambiguity of natural language and shows the user multiple transformations (e.g. multiple shadows and styles of blonde hair) from which she/he might select the best output.

To the best of our knowledge, we are the first at exploiting automatically generated sentences and unpaired images to manipulate image attributes with textual commands. We showed that our method can model multiple attributes, and that can be applied in a wide range of situations such as interpolation and progressive manipulation. We foresee that our work will stimulate further research on attribute-based image manipulation and generation and will stimulate software integration in virtual assistants such as Alexa or image editors such as Photoshop.

REFERENCES

- [1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *ICML*. 214–223.

- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [3] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2019. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 59.
- [4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.
- [5] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. 2018. Language-based image editing with recurrent attentive models. In *CVPR*.
- [6] Yu Cheng, Zhe Gan, Yitong Li, Jingjing Liu, and Jianfeng Gao. 2018. Sequential attention gan for interactive image editing via dialogue. In *AAAI*.
- [7] Yunjeon Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*. 8789–8797.
- [8] Yunjeon Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *CVPR*.
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [10] Carl Doersch. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).
- [11] Alaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. 2018. Keep Drawing It: Iterative language-based image generation and editing. In *NIPS*.
- [12] Alaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. 2019. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *ICCV*. 10304–10312.
- [13] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.
- [14] Abel Gonzalez-Garcia, Joost Van De Weijer, and Yoshua Bengio. 2018. Image-to-image translation for cross-domain disentanglement. In *Advances in neural information processing systems*. 1287–1298.
- [15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*.
- [18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [19] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*. 1501–1510.
- [20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *ECCV*. 172–189.
- [21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*. 1125–1134.
- [22] Younghoo Jo and Jongyoul Park. 2019. SC-FEGAN: Face Editing Generative Adversarial Network with User’s Sketch and Color. In *ICCV*.
- [23] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One Model To Learn Them All. *ArXiV* (2017).
- [24] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- [25] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2019. Controllable Text-to-Image Generation. *arXiv preprint arXiv:1909.07083* (2019).
- [26] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2019. ManiGAN: Text-Guided Image Manipulation. *arXiv preprint arXiv:1912.06203* (2019).
- [27] Xuhong Li, Yves GRANDVALET, and Franck DAVOINE. 2018. Explicit Induction Bias for Transfer Learning with Convolutional Networks. In *ICLR*.
- [28] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuxin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. StoryGAN: A Sequential Conditional GAN for Story Visualization. In *CVPR*. 6329–6338.
- [29] Yahui Liu, Marco De Nadai, Gloria Zen, Nicu Sebe, Bruno Lepri, and Xavier Alameda-Pineda. 2020. GMM-UNIT: Unsupervised Multi-Domain and Multi-Modal Image-to-Image Translation via Attribute Gaussian Mixture Modeling. *arXiv preprint arXiv:2003.06788* (2020).
- [30] Yahui Liu, Marco De Nadai, Gloria Zen, Nicu Sebe, and Bruno Lepri. 2019. Gesture-to-gesture translation in the wild via category-independent conditional maps. *ACM MM* (2019).
- [31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. 3730–3738.
- [32] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. 2019. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*. 1429–1437.
- [33] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *ICCV*.
- [34] Youssef Alami Mejjaati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. 2018. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*. 3693–3703.
- [35] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
- [36] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. 2019. Instance-aware Image-to-Image Translation. In *ICLR*.
- [37] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2019. Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods. *arXiv preprint arXiv:1907.09358* (2019).
- [38] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. 2018. Text-adaptive generative adversarial networks: manipulating images with natural language. In *NIPS*.
- [39] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *CVPR*.
- [40] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfelix, and Francesc Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 818–833.
- [41] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. MirrorGAN: Learning Text-to-image Generation by Redescription. In *CVPR*. 1505–1514.
- [42] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. In *ICML*.
- [43] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016. Learning what and where to draw. In *NIPS*. 217–225.
- [44] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *NIPS*. 2234–2242.
- [45] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.
- [46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.
- [47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*. 6924–6932.
- [48] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. *Technical Report* (2011).
- [49] Hai Wang, Jason D Williams, and SingBing Kang. 2018. Learning to Globally Edit Images with Textual Description. *arXiv preprint arXiv:1810.05786* (2018).
- [50] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* (2015).
- [51] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*. 1316–1324.
- [52] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *ICCV*. 4894–4902.
- [53] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *CVPR*. 5907–5915.
- [54] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI* 41, 8 (2018), 1947–1962.
- [55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. 586–595.
- [56] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2019. Pluralistic Image Completion. In *CVPR*. 1438–1447.
- [57] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.
- [58] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward multimodal image-to-image translation. In *NIPS*. 465–476.
- [59] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. 2019. Language-based Colorization of Scene Sketches. *ACM Transactions on Graphics* 38 (2019), 16.

A IMPLEMENTATION DETAILS

We built our model upon the recent state-of-the-art approaches, namely MUNIT [20], BicycleGAN [58], StarGAN [7] and TAGAN [38]. We apply Instance Normalization (IN) [47] to the content encoder E_c , while we use Adaptive Instance Normalization (AdaIN) [19] and Layer Normalization (LN) [2] for the decoder G . We use a 2 layers LSTM [18] with Dropout [45] rate 0.1 for the text encoder E_t . For the discriminator network, we use Leaky ReLU [50] with a negative slope of 0.2. The details of the architecture are shown in Table 5. For the test of the two datasets, we use 2 scales for the discriminator, in which the resolutions of input images are 128×128 and 64×64, respectively.

We use the Adam optimizer [24] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and an initial learning rate of 0.0001. The learning rate is decreased by half every 2e5 iterations. We set maximal iterations 1e6 for CelebA and 3e5 for CUB dataset. In all experiments, we use a batch size of 1. And we set the loss weights to $\lambda_{rec_s} = 10$, $\lambda_{cyc} = 10$, and $\lambda_{KL} = 0.1$. Random mirroring is applied during training.

In the experiments, we use a simplified version of the GMM, which satisfies the following properties:

- The mean vectors are placed on the vertices of Z -dimensional regular simplex, so that the mean vectors are equidistant (and disentangled). Here $Z = d * K$ from the main paper.
- The covariance matrices are diagonal, with the same on all the components. In other words, each Gaussian component is *spherical*, formally: $\Sigma_k = \sigma_k \mathbf{I}$, where \mathbf{I} is the identity matrix.

Changing the parameter σ_k to a number $\lambda\sigma_k$ does not change the model as in normal VAEs [10]. However, when λ is bigger than the distance between means in the simplex, this poses issues on the assumption of independent and disentangled latent factors.

Regarding the baseline models, we modify the baseline models based on their corresponding released codes, including StackGAN++¹, StarGAN², TAGAN³ and ManiGAN⁴. To train StarGAN on this task, we add a module E_t same as ours and learn to predict target attributes guided by binary cross-entropy loss. For fairness, all models take use the same corpus and images during the training. To test the training speed in Table 1 of the main paper we set the batch size 1 and use a single GTX TITAN Xp with 12GB Memory for all models.

We release the source code of our model and the StarGAN* implementation at <https://xxx>⁵

B ATTRIBUTE CLASSIFIER

We evaluate the consistency of attributes labels through a model that is trained to classify faces on the four tested combinations of attributes. We finetune an ResNet-50 [16] model trained on ImageNet [9] by using the L^2 -SP regularization [27] on all the layers, with $\alpha = 0.01$ and $\beta = 0.01$. The list of training examples is the same as the main model, thus keeping the hold-out dataset as unseen. From the training list we carved out 10% of the data for the validation and 20% for testing.

¹<https://github.com/hanzhanggit/StackGAN-v2>

²<https://github.com/yunjey/stargan>

³<https://github.com/woozu/tagan>

⁴<https://github.com/mrlibw/ManiGAN>

⁵Temporary attached in the submission as supplementary file to keep the anonymity of the authors. It will be stored in a permanent link upon acceptance.

We use the Adam optimizer [24] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and an initial learning rate of 0.001. The learning rate is divided by 10 whenever the training loss does not decrease for two epochs. The training is early stopped when the validation loss does not decrease after 10 epochs.

We apply a central crop and random rotation until at most 10 degrees data augmentation techniques during training. The images are then resized to 299 × 299 pixels. The accuracy is 0.91 and 0.91 in the training and test set respectively.

We use this classifier to evaluate the accuracy of our transformations through the Micro-F1 (the F1 score that weights equally all test samples) score between the desired attributes and the generated ones.

C AUTOMATIC TEXT PROTOCOL

We propose an automatic text protocol to collect the description text, which takes use of the attribute annotations only. Hence, we launch our experiments based on two multi-label recognition or classification [31, 48] datasets. In CelebA [31] dataset, we select 8 attributes from annotated attributes {“Black_Hair”, “Blond_Hair”, “Brown_Hair”, “Male”, “Smiling”, “Young”, “Eyeglasses”, “No_Beard”}, which including both the style (e.g. hair color) and content (e.g. gender, age, beard and eyeglasses). In CUB [48] dataset, we select three body parts of the birds {“wing”, “crown”, “breast”} and each part can be consist of a set of colors {“grey”, “brown”, “red”, “yellow”, “white”, “buff”}.

To automatically collect manipulation text, we provide some common templates and phrases to collect the text, such as “change ... to ...”, “with ... hair”, “make the face/bird ...”, “a man/woman/bird with/has ...”. We randomly permute the description sequence of the attributes and replace some words with synonymy (e.g. “change”, “translate”, “modify” and “make”) to improve the generalization. Similarly, we randomly select the synonymy named entities during generate text (e.g., “man”, “male”, “boy”, “gentleman” and “sir” for the attribute *Male*). As shown in Figure 9, it indicates that our model generates consistent results with various textual commands for the same manipulation. We also test the sensitivity of the strategy, as the choice of the strategy might influence the results of the model. Table 6 shows that our model is robust to the choice of the textual strategy.

D USER STUDY

We ask different people with various backgrounds to describe how to change the “image A” to be similar to “image B”, as shown in Figure 10 and Figure 11. In Figure 8 we show the result of the human evaluation on the realism.

E DOMAIN INTERPOLATION

We show more results in Figure 13 and Figure 14 on various face images. Figure 16 show examples of domain interpolation.

F DOMAIN SAMPLING

As shown in Figure 12, we apply the domain sampling to obtain diverse results that are consistent with the same textual commands.

We also evaluate the ability of DWC-GAN to synthesize new images with attributes that are extremely scarce or non present in

Part	Input → Output Shape	Layer Information
E_c	$(h, w, 3) \rightarrow (h, w, 64)$	CONV-(N64, K7x7, S1, P3), IN, ReLU
	$(h, w, 64) \rightarrow (\frac{h}{2}, \frac{w}{2}, 128)$	CONV-(N128, K4x4, S2, P1), IN, ReLU
	$(\frac{h}{2}, \frac{w}{2}, 128) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	CONV-(N256, K4x4, S2, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
E_a	$(h, w, 3) \rightarrow (h, w, 64)$	CONV-(N64, K7x7, S1, P3), ReLU
	$(h, w, 64) \rightarrow (\frac{h}{2}, \frac{w}{2}, 128)$	CONV-(N128, K4x4, S2, P1), ReLU
	$(\frac{h}{2}, \frac{w}{2}, 128) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	CONV-(N256, K4x4, S2, P1), ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{8}, \frac{w}{8}, 256)$	CONV-(N256, K4x4, S2, P1), ReLU
	$(\frac{h}{8}, \frac{w}{8}, 256) \rightarrow (\frac{h}{16}, \frac{w}{16}, 256)$	CONV-(N256, K4x4, S2, P1), ReLU
	$(\frac{h}{16}, \frac{w}{16}, 256) \rightarrow (\frac{h}{32}, \frac{w}{32}, 256)$	CONV-(N256, K4x4, S2, P1), ReLU
	$(\frac{h}{32}, \frac{w}{32}, 256) \rightarrow (1, 1, 256)$	GAP
E_t	$(256,) \rightarrow (Z,)$	FC-(NZ)
	$(256,) \rightarrow (Z,)$	FC-(NZ)
E_t	$(300(+Z), T) \rightarrow (2400,)$	LSTM(W300, H300, L2, BiD, Dropout)
	$(2400,) \rightarrow (Z,)$	FC-(NZ)
G	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), AdaIN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), AdaIN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), AdaIN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), AdaIN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{2}, \frac{w}{2}, 128)$	UPCONV-(N128, K5x5, S1, P2), LN, ReLU
	$(\frac{h}{2}, \frac{w}{2}, 128) \rightarrow (h, w, 64)$	UPCONV-(N64, K5x5, S1, P2), LN, ReLU
	$(h, w, 64) \rightarrow (h, w, 3)$	CONV-(N3, K7x7, S1, P3), Tanh
D	$(h, w, 3) \rightarrow (\frac{h}{2}, \frac{w}{2}, 64)$	CONV-(N64, K4x4, S2, P1), Leaky ReLU
	$(\frac{h}{2}, \frac{w}{2}, 64) \rightarrow (\frac{h}{4}, \frac{w}{4}, 128)$	CONV-(N128, K4x4, S2, P1), Leaky ReLU
	$(\frac{h}{4}, \frac{w}{4}, 128) \rightarrow (\frac{h}{8}, \frac{w}{8}, 256)$	CONV-(N256, K4x4, S2, P1), Leaky ReLU
	$(\frac{h}{8}, \frac{w}{8}, 256) \rightarrow (\frac{h}{16}, \frac{w}{16}, 512)$	CONV-(N512, K4x4, S2, P1), Leaky ReLU
	$(\frac{h}{16}, \frac{w}{16}, 512) \rightarrow (\frac{h}{32}, \frac{w}{32}, 512)$	CONV-(N512, K4x4, S2, P1), Leaky ReLU
	$(\frac{h}{32}, \frac{w}{32}, 512) \rightarrow (\frac{h}{32}, \frac{w}{32}, 1)$	CONV-(N1, K1x1, S1, P0)
	$(\frac{h}{32}, \frac{w}{32}, 512) \rightarrow (1, 1, n)$	CONV-(Nn, K $\frac{h}{8} \times \frac{w}{8}$, S1, P0)

Table 5: Network architecture. We use the following notation: Z : the dimension of attribute vector, n : the number of the domains, N : the number of output channels, K : kernel size, S : stride size, P : padding size, CONV: a convolutional layer, GAP: a global average pooling layer, UPCONV: a $2 \times$ bilinear upsampling layer followed by a convolutional layer, FC: fully connected layer, T : the number of the words in the text sentence, W : the dimension of word embedding, H : hidden size of the LSTM layer, BiD: using bidirectional encoding, L : the number of LSTM layers.

Strategy Name	Metrics		
	IS↑	FID↓	F1↑
Step-by-step	3.070	32.25	94.14
Overall	3.081	32.07	94.78
Mixed	3.078	32.16	94.80
Random	3.069	32.14	94.80

the training dataset, as shown in Figure 17. To do so, we select three combinations of attributes consisting of less than two images in the CelebA dataset: [1, 1, 0, 0, 1, 1, 0, 1] ("Black_Hair, Blond_Hair, Female, Smiling, Young, No_Eyeglasses, No_Beard") and [1, 1, 0, 0, 1, 0, 0, 1] ("Black_Hair, Blond_Hair, Female, Smiling, Old, No_Eyeglasses, No_Beard"). We show that the model can effectively generate configurations

Table 6: Quantitative comparison on the different automatic text strategies on the CelebA dataset.

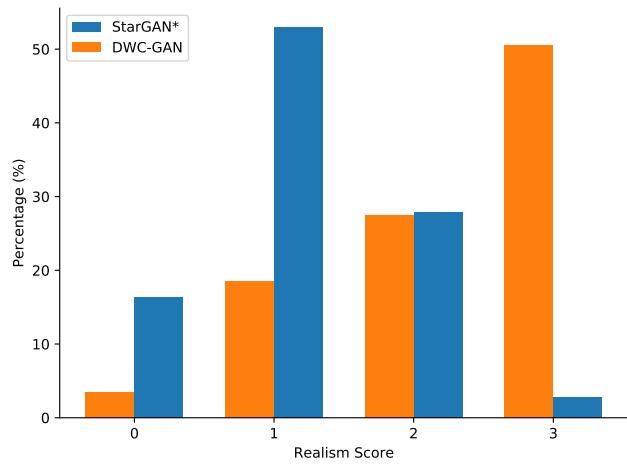


Figure 8: Distributions on the Realism human evaluation.

in previously unseen domains if we describe them through the textual condition. These generated images show that our model can effectively do few- or zero-shot generation.

Input Image	Text Strategies	Input Text	Output Image
	Step-by-step	<p>"put on smile, no beard, translate her hair colour from black into blond, change the sir to be a girl"</p> <p>change her face smile, remove the moustache, change the gentleman to be a female, translate hair color blond</p>	
	Overall	<p>"it is a smiling young miss without sunglasses , with blond hair and no beards</p> <p>this is an smile young girl with blond hair, without whiskers and without eyeglasses</p>	
	Mixed	<p>make the man with black hair to a smiling young lady take off beards, with blond hair and no eyeglasses</p> <p>modify the young smile male with black hair to a smile young female take off glasses and take off whiskers and with blond hair</p>	

Figure 9: Examples on automatic text strategies. It shows a consistent performance with various textual commands for the same manipulation.

Image A	Image B	Human-written Commands	Output
		<p>"change his hair to be black and the gender to female"</p> <p>"female, smile, black hair"</p>	   
		<p>"could you modify the gender and change the hair color to black"</p>	   
		<p>"change the gender from male to female and put a smile on the face"</p>	  
		<p>"change hair color to black, change the gender, smile"</p> <p>"change hair color to brown, unsmiling"</p>	   
		<p>"brown hair color, no smile"</p>	  
		<p>"can you change the hair color to brown"</p>	
		<p>"change the blond hair to brown hair"</p>	
		<p>"make the face with brown hair"</p>	

Figure 10: Examples on asking users to describe how to change image A to make it similar to image B.

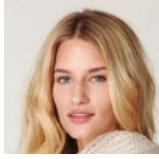
Image A	Image B	Human-written Commands	Output
		"change the hair to be brown, make him a girl, remove the beard, keep him smile"	    
		"change the gender to female, remove the beards"	
		"change hair color to brown, female, remove beard"	
		"change the gender to female with long brown hair, no beard"	
		"change the gender and remove the beard"	
		"change hair color to brown, smile, older"	  
		"change hair color to brown . change age to old"	
		"can you make it older and change the hair color to brown"	
		"change the hair color to brown . make the age older around 50 . make her smile"	
		"change the hair color to brown, make her smile and older"	

Figure 11: Examples on asking users to describe how to change image A to make it similar to image B.



Figure 12: Examples of domain sampling on CelebA dataset.



Figure 13: Examples of translations on CelebA dataset.

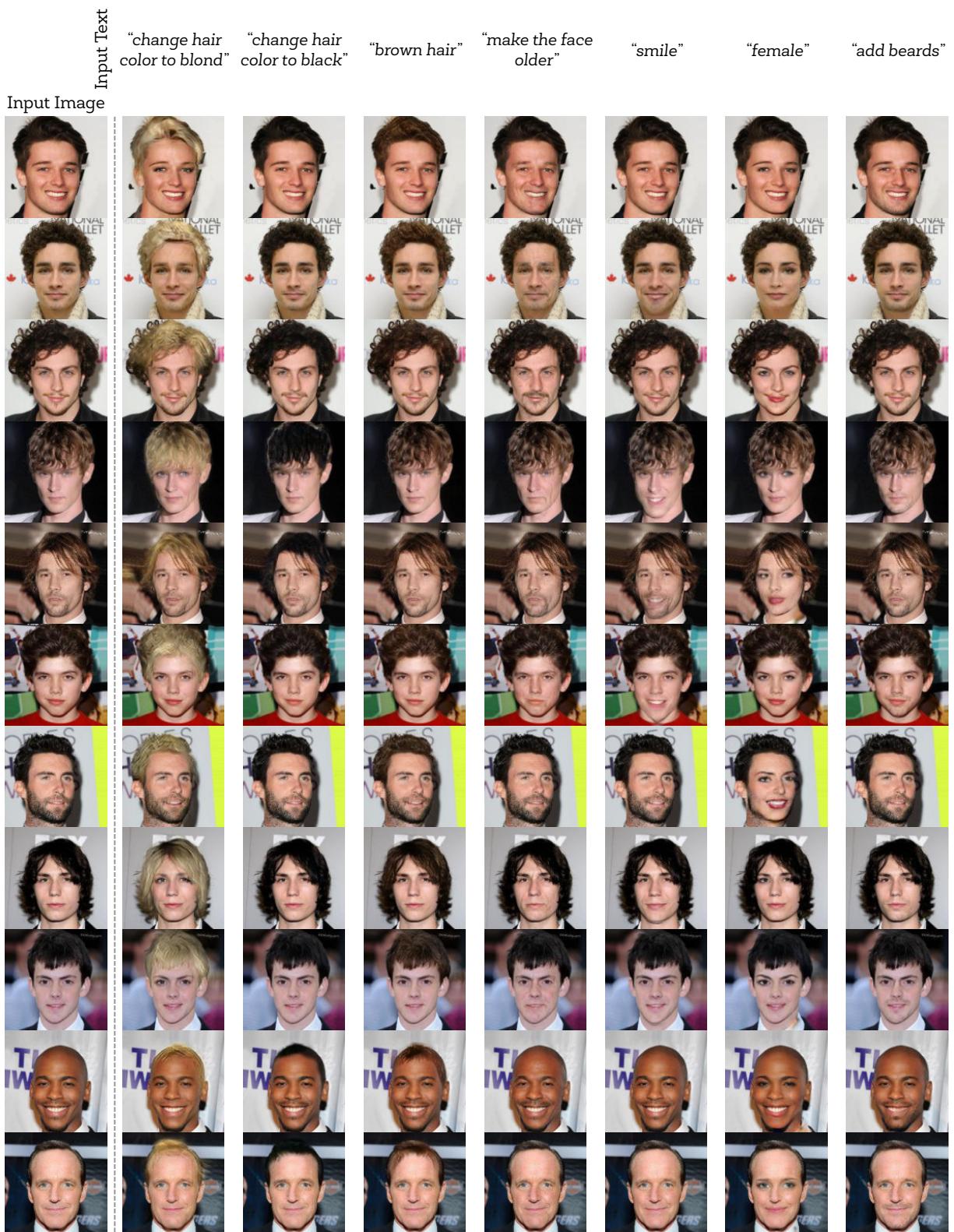


Figure 14: Examples of translations on CelebA dataset.

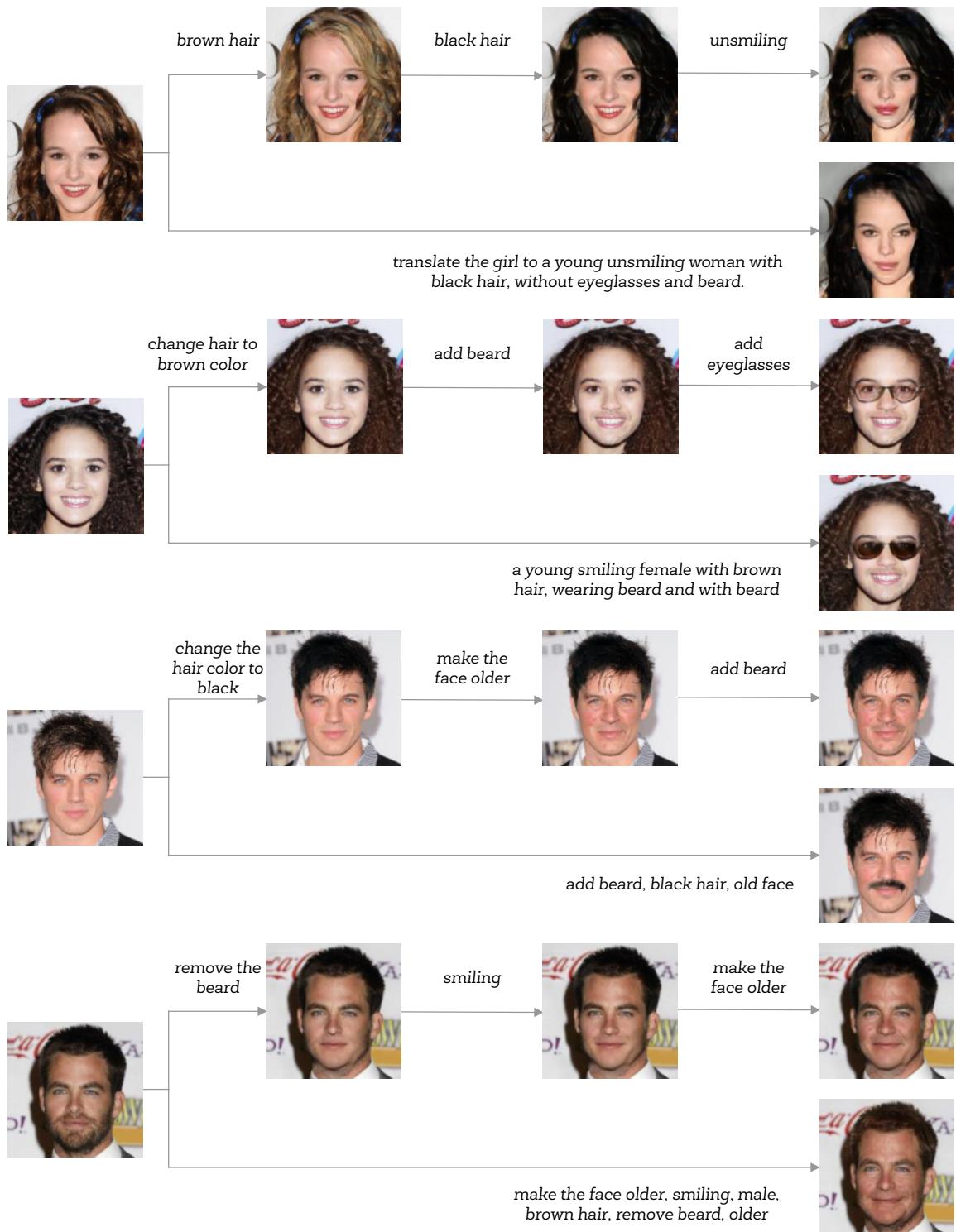


Figure 15: Examples of progressive manipulation.

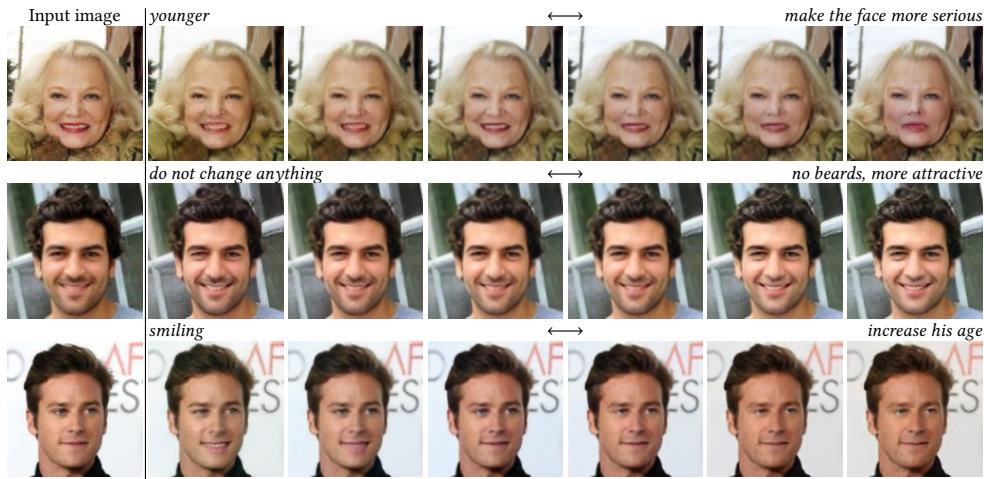


Figure 16: Examples of domain interpolation given an input image.



Figure 17: Generated images in previously unseen combinations of attributes.