TEXT-DRIVEN IMAGE MANIPULATION VIA SEMANTIC-AWARE KNOWLEDGE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic-level facial attribute transfer is a special task to edit facial attribute, when reference images are viewed as conditions to control the image editing. In order to achieve better performance, semantic-level facial attribute transfer needs to fulfil two requirements: (1) specific attributes extracted from reference face should be precisely transferred to target face; (2) irrelevant information should be completely retained after transferring. Some existing methods locate and modify local support regions of facial images, which are not effective when editing global attributes; the other methods disentangle the latent code as different attribute-relevant parts, which may transfer redundant knowledge to target faces. In this paper, we first propose a novel text-driven directional latent mapping network with semantic direction consistency (SDC) constrain to explore the latent semantic space for effective attribute editing, leveraging the semantic-aware knowledge of Contrastive Language-Image Pre-training (CLIP) model as guidance. This latent space manipulation strategy is designed to disentangle the facial attribute, removing the redundant knowledge in the transfer process. And on this basis, a novel attribute transfer method, named semantic directional decomposition network (SDD-Net), is proposed to achieve semantic-level facial attribute transfer by latent semantic direction decomposition, improving the interpretability and editability of our method. Extensive experiments on CelebA-HQ dataset show that our method achieves impressive performance over the state-of-the-art methods.

1 INTRODUCTION

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have revolutionized a variety of fields due to its powerful ability to generate realistic and meaningful outputs. Recent works (Jahanian et al., 2019; He et al., 2019; Goetschalckx et al., 2019) have shown that deep generative models can capture real-world data distribution, and encode them into a semantically-rich latent space. Inspired by this, a lot of tasks draw their attention on the latent space manipulation, including image enhancement (Ledig et al., 2017; Yang et al., 2021b), editing (Shen et al., 2020; Härkönen et al., 2020; Patashnik et al., 2021), and discriminative tasks (Nitzan et al., 2021; Xu et al., 2021).

With the tremendous success of deep generative models, facial attribute editing (Yeh et al., 2017; Liu et al., 2019; He et al., 2020; Dorta et al., 2020), aiming to edit the specific attributes of the target facial image, has become topical. As a special case of facial attribute editing, facial attribute transfer (Xiao et al., 2018; Lin et al., 2018; Yin et al., 2019; Choi et al., 2018; 2020) uses knowledge from reference image as a condition to edit the corresponding attribute from the target image. In order to ensure that the manipulated facial image meets the requirements and interests, facial attribute transfer task tackles two challenges simultaneously: (1) **editing relevance:** the relevant attribute should be edited precisely according to the given condition; and (2) **keeping irrelevance:** the irrelevant part (e.g., identify information, background, or other attributes) should not be modified during attribute transfer. Due to strong entanglement of the attributes, meeting both requirements is an intractable task. For example, without fully disentanglement, transferring the "smile" attribute to the target facial image may cause that, another irrelevant but coupled attribute, e.g. "cheek color" attribute, would be changed during editing.

In view of the above issues, recently a variety of methods explore the attribute disentanglement in two ways. Some methods (He et al., 2020; Kwak et al., 2020) resort to the way of spatial attention

detection, which disentangle the attribute by searching specific support region spatially and only manipulate the image in such a confined area. Obviously, these methods totally ignore the facial details beyond the support region when the edited attribute is a global attribute, such as "smile" or "age". Meanwhile, other methods (Shen et al., 2020; Yang et al., 2021a; Patashnik et al., 2021) pay attention to the latent space factorization through pre-trained GAN. These methods employ the high-level semantic information as guidance to manipulate image in latent spaces, which are more suitable to handle both global and local attribute editing. However, owing to over-coupled semantic features, these methods are hard to manipulate specific attribute without powerful supervision.

To overcome the problems mentioned above, we explore the latent semantic space for disentangled attribute editing, and apply the discovered manipulation method to facial attribute transfer task. As for attribute editing task, in order to disentangle and edit the attribute specified by the text prompt, we design the directional latent mapping network, which leverages semantic direction consistency (SDC) loss to constrain the manipulation in the CLIP-space (Radford et al., 2021). The key idea of the SDC loss is employing the change direction of semantic feature to estimate the latent manipulation. Furthermore, in order to apply this effective editing method to facial attribute transfer task, we propose a novel semantic-level facial attribute transfer method driven by text prompt, named as semantic directional decomposition network (SDD-Net). The SDD-Net extracts and transfers the specific attribute without redundant information through attribute-manipulated semantic directional decomposition.

Our contributions are summarized as follows:

- We propose a novel method namely directional latent mapping network for facial attribute editing, which utilizes the semantic direction consistency regularization to ensure attribute disentanglement.
- To further take advantage of semantic direction constrain, we propose a text-driven semantic directional decomposition network (SDD-Net) for semantic-level attribute transfer, by transferring the knowledge from the reference image to the target image.
- Extensive experiments on CelebA-HQ (Karras et al., 2017) dataset show that our method achieves significant improvements over the state-of-art approaches.

2 Related Work

Latent Space Manipulation. Recent studies (Bau et al., 2020; Goetschalckx et al., 2019; Shen et al., 2020) have shown that numerous GAN models can encode rich crucial information in the intermediate latent space, such as $\mathcal{W}$, $\mathcal{W}+$ (Abdal et al., 2019), or *StyleSpace* $\mathcal{S}$ (Wu et al., 2021). By learning to modify the intermediate latent code, generative models can transfer attributes from one face to another face (Xiao et al., 2018; Choi et al., 2020). To find a latent code that allows for meaningful manipulation, some methods try to learn an effective encoder network, which inverts a real image into latent space: encoder4editing (e4e) (Tov et al., 2021) method presents an encoder that is specifically designed for balancing distortion-editability tradeoff and a distortion-perception tradeoff within the StyleGAN latent space; Stylespace (Wu et al., 2021) proposes a space of channel-wise style parameters that disentangle attributes by controlling attribute-related style channels; The pixel2style2pixel (pSp) (Richardson et al., 2021) method utilizes a novel encoder architecture that inverts a real image into $\mathcal{W}+$ space without optimization. Other methods mainly focus on finding such latent code modification approach as used to traverse latent space, result in the desired manipulation: AttGAN (He et al., 2019) applies attribute classification constraint to model the relation between the attributes and the latent representation; InterfaceGAN (Shen et al., 2020) decouples some entangled semantic features with subspace projection; GANspace (Härkönen et al., 2020) leverages principal component analysis to identify mainly direction in latent space; L2M-GAN (Yang et al., 2021a) imposes an orthogonality constraint to ensure disentanglement in latent space; StyleCLIP (Patashnik et al., 2021) leverages CLIP models to guide the attributes manipulation by latent semantic matching.

Different from the methods mentioned above, our method aligns the guidance direction, which is derived directly from text prompt, with the change direction of semantic features extracted by CLIP to guide the manipulation in the latent space. By constraining the consistency of these directions, our method can achieve disentangled attribute editing.

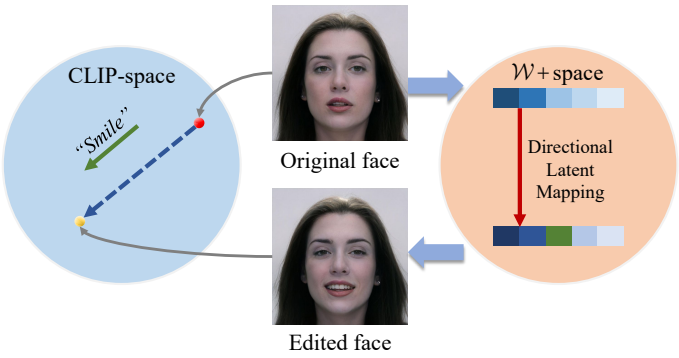

Figure 1: Overview of our idea for directional latent mapping network. By enforcing the change direction of semantic feature to align with desired direction in CLIP-space, our method can achieve impressive disentangled image manipulation.

Semantic-level Attribute Transfer. Semantic-level attribute transfer is a more challenging facial attribute editing task. Recent studies have attempted more detailed approaches for facial attribute transfer: StarGAN (Choi et al., 2018) applies cycle consistency to preserve identity, and uses classification loss to transfer between different domains. In addition, StarGANv2 (Choi et al., 2020) also could synthesize reference-guide images leveraging multiple domain translation. Some works are specialized for specific attributes: ExprGAN (Ding et al., 2018) proposes a model to learn the disentangled identity and expression representations explicitly for facial expression transfer. Besides, ERGAN (Hu et al., 2020) proposes a dual learning scheme to simultaneously learn two inverse manipulations for attribute transfer.

Different from these works, our proposed method focuses not only on one attribute, but also on various global or local attributes. Given an explicit text prompt, our method can automatically extract the specific attribute from the reference, and transfer the knowledge to the target image.

3 Methodology

In the following section, we first provide the preliminaries and problem formulation. Then, we introduce our text-driven directional latent mapping network for disentangled facial attribute editing. Finally, we describe the text-driven semantic directional decomposition network (SDD-Net) for semantic-aware facial attribute transfer.

3.1 Preliminaries

StyleGAN. The StyleGAN (Karras et al., 2019; 2020) generator consists of two main components: mapping network and synthesis network. The former translates latent code z to latent code w, which is in semantic-rich latent space $\mathcal{W}$, and the latter utilizes latent code w to synthesize final images through different layers. Due to the rich information inside, the latent code w can control the multi-granularity semantic features of the synthetic image, which is utilized for effective facial attribute transfer in semantic-level. In this paper, we leverage the pre-trained synthesis network as our image generator.

StyleCLIP. The StyleCLIP (Patashnik et al., 2021) method first combines StyleGAN and CLIP as a strong tool for text-driven image editing. The key idea of StyleCLIP is to leverage the CLIP as latent manipulation guidance. By mapping multi-modal inputs to the CLIP-space, StyleCLIP could ensure that the synthetic image matches with the text prompt in semantic-level. The objective is given by:

$$\mathcal{L}_{styleclip} = D_{\text{CLIP}}(G(w), text), \tag{1}$$

where D_{CLIP} is the cosine distance metric in CLIP-space, and G is the pre-trained StyleGAN generator.

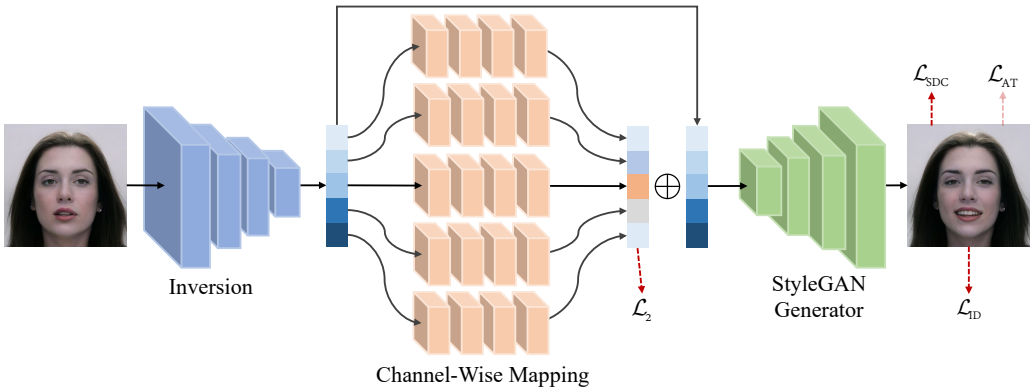

Figure 2: The architecture of our directional latent mapping network. Given the text prompt, we manipulate the latent code inverted by the facial images. Then we use g group channel-wise mapping networks to manipulate the different parts of latent code w respectively. Multiple loss functions are used to constrain the synthetic image to fulfil the requirements.

3.2 Problem Formulation

Our method leverages two latent space: $\mathcal{W}+$ space and CLIP-space, to transfer semantic-level facial attributes. For better expression, let $\mathcal{X}$ and $\mathcal{Y}$ denote the set of images and semantic features in CLIP-space, respectively. The image $x \in \mathcal{X}$ is inverted into corresponding latent code $w \in \mathcal{W}+$. t denotes the input text prompt, with corresponding semantic feature y_t. Meanwhile, the parameters of pre-trained StyleGANv2 generator G are frozen during training.

Given the text prompt t, the goal of facial attribute transfer model is to train a latent mapping network, which translates $w \backslash y$ to $\hat{w} \backslash \hat{y}$, and synthesizes the edited image $\hat{x}$ that meets the specific requirements. Our basic editing model can be formally definded as $\hat{x} = G(M(x, t))$. Hence given the reference image $x_{ref} \in \mathcal{X}$, our semantic-aware attribute transfer model can be formally definded as $\hat{x} = G(M(x, x_{ref}, t))$.

3.3 Directional Latent Mapping

Only focusing on matching $\hat{y}$ with y_t may cause irrelevant attribute changed. Therefore, we enforce the mapping network to focus on the change direction of semantic features in CLIP-space, and illustrate this idea in Figure 1. The details of the proposed directional latent mapping network is described as follows:

Architecture. It has been shown that the different layers of synthesis network, which corresponds to different parts of the latent code, control different granularity of semantic feature. In order to better exploit this property, we design the directional latent mapping network, and depict it in Figure 2. To a great extent, the degree of attributes disentanglement depends on the disentanglement of latent features. Therefore, we split the layers into g groups, instead three (coarse, medium, and fine), with g fully connected mapping networks, one for each group. The divided part of the latent code can be denoted as $w = [w_1, w_2, ..., w_g]$, so the mapping network is defined by:

$$M(w) = [M_1(w_1), M_2(w_2), ..., M_g(w_g)], \tag{2}$$

where M_i is the i-th mapping network, and $[\cdot, \cdot]$ means concat operation. Then we use skip-connection operation to obtain the final manipulated latent code $\hat{w} = w + \alpha M(w)$, and feed it to the pre-trained StyleGANv2 generator G to get the final manipulated facial image $\hat{x} = G(\hat{w})$.

Training Objective. The edited desired attribute is determined by the textual prompt t. Without paired training data, the mapping network could not correctly manipulate the latent code. In order to obtain extra powerful supervision, we use CLIP model to effectively extract semantic features (attributes) y of the corresponding images and text:

$$\begin{aligned} y_i; \hat{y}_i &= E_I(x; \hat{x}), \\ y_t &= E_T(t), \end{aligned} \tag{3}$$

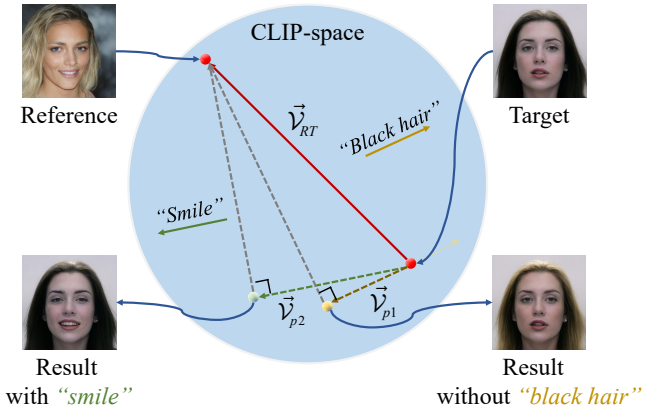

Figure 3: The illustration of attribute transfer loss $\mathcal{L}_{\text{AT}}$ in our Semantic Directional Decomposition Network (SDD-Net). We embed both the text prompt and reference\target face into CLIP-space. When the projected vector $\vec{\mathcal{V}}_p$ has the same direction with desired direction, the same attribute (e.g., smile) is transferred to the target face. Reversely, the opposite attribute (e.g., black hair) will be transferred when the direction is inverse.

where E denotes the pre-trained multi-modal feature extractor integrated in CLIP, y_i and $\hat{y}_i$ denotes the extracted semantic features of x and $\hat{x}$, respectively. y_t is the desired semantic feature extracted from textual prompt t.

In the latent CLIP-space, simply optimizing the matching degree between $\hat{y}_i$ and y_t may cause irrelevant attribute to be changed. Therefore, rather than optimizing the matching degree to the utmost extent, we enforce the change direction between y and $\hat{y}_i$ to align with the y_t, and propose the semantic direction consistency (SDC) loss. The SDC loss is given by:

$$\begin{aligned} \vec{I} &= \hat{y}_i - y_i, \\ \vec{T} &= y_t, \\ \mathcal{L}_{\text{SDC}} &= 1 - S(\vec{I}, \vec{T}), \end{aligned} \tag{4}$$

where $S(\cdot, \cdot)$ is the similarity measurement. In this paper, we use the effective cosine similarity as the measurement.

The above objective can guide the mapping network to manipulate the latent code along the direction of textual prompt. In order to preserve the irrelevant parts, we use the following identity loss:

$$\mathcal{L}_{\text{ID}} = 1 - \langle R(G(w)), R(G(\hat{w})) \rangle \,, \tag{5}$$

where R is a pre-trained ArcFace (Deng et al., 2019) network for face recognition, and $\langle \cdot, \cdot \rangle$ computes the cosine similarity between its two arguments. Meanwhile, we use L_2 distance in $\mathcal{W}+$ space to control the degree of manipulation. Then, the whole training objective for directional latent mapping network is denoted as:

$$\arg\min_{w \in \mathcal{W}+} \lambda_{\text{SDC}} \mathcal{L}_{\text{SDC}}(w, t) + \lambda_{L2} ||M(w)||_2 + \lambda_{\text{ID}} \mathcal{L}_{\text{ID}}(w). \tag{6}$$

3.4 Semantic Directional Decomposition

Inspired by the above method, we found that a semantic attribute translation can be represented as a directional vector in CLIP-space. Based on this, we leverage the same architecture of directional mapping network, and propose a facial attribute transfer method via semantic-aware knowledge transfer, named as semantic directional decomposition network (SDD-Net).

Given the textual prompt t, we want to transfer the attribute, extracted from the reference image x_{ref}, to the target image x_{tar}. The y_{ref} and y_{tar} are their corresponding semantic features respectively. And the semantic feature of textual prompt t is y_t.

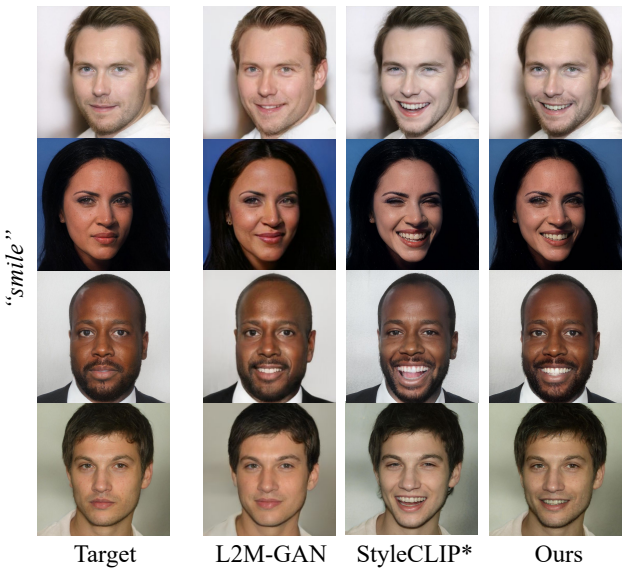

Figure 4: Qualitative results for facial attribute editing on attribute "smile". The left column are the target images sampled from CelebA-HQ dataset. The other column from left to right are the editing results of L2M-GAN (Yang et al., 2021a), StyleCLIP* (Patashnik et al., 2021), and our directional latent mapping network.

Due to the above direction-based manipulation, we can treat the facial attribute transfer problem as the attribute semantic feature projection problem. In the CLIP-space, a facial image can be treated as the composition of different semantic features. Therefore, we assume that different facial images can be converted to each other in semantic-level. In other words, one semantic feature can be translated to another semantic feature.

Under such an assumption, we first define the RT (reference-target) vector as: $\vec{\mathcal{V}_{RT}} = y_{ref} - y_{tar}$, which is the variance of two semantic features, and also is the direction of translation. We set the y_t as the projection direction. In order to extract knowledge from the reference, we project RT vector $\vec{\mathcal{V}_{RT}}$ onto desired direction $\vec{\mathcal{V}_t} = y_t$:

$$\vec{\mathcal{V}_p} = \vec{\mathcal{V}_{RT}} \cdot \left(\frac{\vec{\mathcal{V}_t}}{|\mathcal{V}_t|} \right)^2, \tag{7}$$

where $\vec{\mathcal{V}_p}$ is the projected vector. Then we add it to the target semantic feature y_{tar} to get the final goal:

$$y_{goal} = y_{tar} + \beta \vec{\mathcal{V}_p}, \tag{8}$$

where β is the hyperparameter.

Training Objective. We use same architecture of directional latent mapping network to obtain the manipulated image $\hat{x} = G(M(\hat{w}))$, then match the semantic feature $\hat{y} = E_I(\hat{x})$ with y_{goal} in CLIP-space. As illustrated in Figure 3, the attribute transfer loss $\mathcal{L}_{\text{AT}}$ is given by:

$$\mathcal{L}_{\text{AT}} = \text{MSE}(\hat{y}, y_{goal}), \tag{9}$$

where $\text{MSE}(\cdot, \cdot)$ is the mean squared error.

Meanwhile, we use the same $L2$ loss and identity loss to preserve the irrelevant parts. The whole training objective of SDD-Net is:

$$\arg\min_{w \in \mathcal{W}+} \lambda_{\text{AT}} \mathcal{L}_{\text{AT}} + \lambda_{\text{L2}} \mathcal{L}_2 + \lambda_{\text{ID}} \mathcal{L}_{\text{ID}}. \tag{10}$$

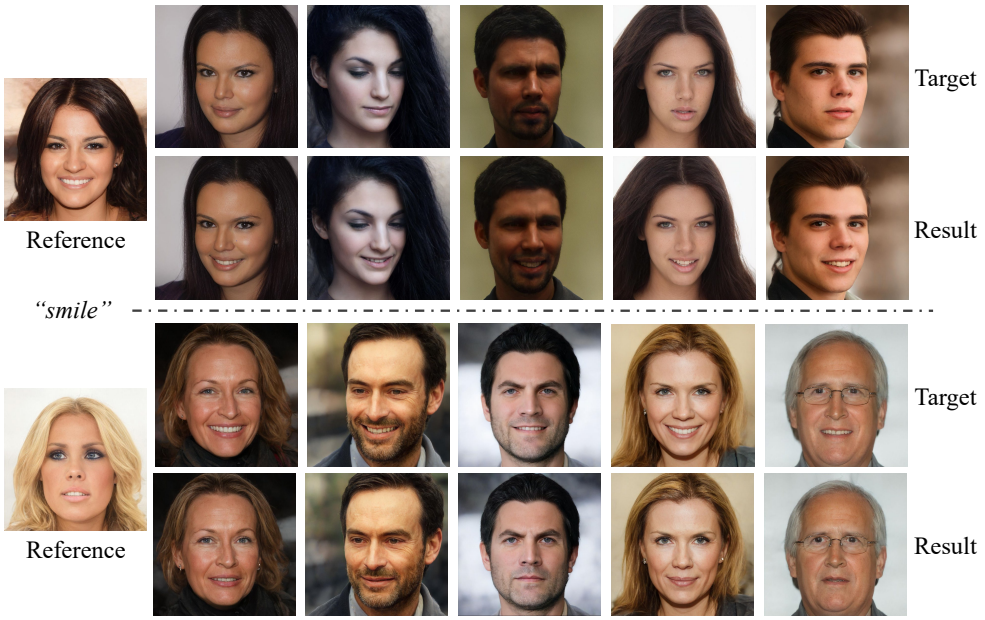

Figure 5: The qualitative results of our SDD-Net on single attribute transfer. Reference images in the left column. Target images in the first row of each parts, the rest is our manipulated images. Given the text prompt, our SDD-Net transfer the specific attribute ("smile") to the target images (top part) when the reference have the specific attribute. By contrast, the reverse attribute ("unsmiling") will be transferred (bottom part) when the specific attribute is not contained in the reference. Notice that, we only input one text prompt. So the SDD-Net could determine the forward or reverse transferring direction, according to the reference.

4 Experiments

In this section, we first introduce the involved dataset and the implementation details. Then we will present the comparison results with several state-of-the-art facial attribute editing and facial attribute transfer methods to prove the effectiveness of our proposed method. Finally, the ablation studies will be presented to prove the effectiveness of our method.

4.1 Dataset

In order to achieve text-driven facial attribute editing and transfer, we choose the widely-used CelebA-HQ (Karras et al., 2017) dataset, which consists of 30,000 high quality facial images picked from the original CelebA (Liu et al., 2015) dataset. The size of each high quality image is 1024×1024. In the original dataset, each image has 40 attributes annotations inherited from the original CelebA. However in this work, we remove these annotations, and leverage CLIP model as powerful supervision. We also use the standard training, validation and test splits inherited from CelebA dataset.

4.2 Implementation Details

In this subsection we provide the implementation details of our proposed networks. All images taken from the CelebA-HQ are inverted by e4e (Tov et al., 2021). We set the batch size and the number of total iterations to 5 and 50k respectively, during training. Our image editing is performed on StyleGANv2 pre-trained on FFHQ (Karras et al., 2019) dataset. We keep the StyleGANv2 generator fixed during training. The CLIP model is pre-trained on 400 million image-text pairs. We use the Vision Transformer (ViT) (Dosovitskiy et al., 2020) and a normal Transformer (Vaswani et al., 2017), which are integrated in CLIP model, as our image encoder and text encoder, respectively. Our directional latent mapping module is initialized using the pre-trained StyleGANv2, and trained

Figure 6: Experiments on reference-guided image synthesis on CelebA-HQ. Reference and target images both in the first row. The second and third raw are the synthetic results of StarGANv2 (Choi et al., 2020) and our SDD-Net respectively.

using Adam with the learning rate 5e-3. We set $g = 9$, $\alpha = 0.1$, and $\beta = 130$. For facial attribute editing, the hyperparameters are empirically set as $\lambda_{SDC} = 1$, $\lambda_{L2} = 0.4$, and $\lambda_{\text{ID}} = 0.02$. For facial attribute transfer, the hyperparameters are empirically set as $\lambda_{\text{AT}} = 1$, $\lambda_{L2} = 0.2$, and $\lambda_{\text{ID}} = 0.02$. Our methods are trained on PyTorch with a single TITAN RTX GPU. We optimize training objective through gradient descent, by back-propagating the gradient through the pre-trained and fixed StyleGAN generator G and CLIP multi-modal encoder E.

4.3 Baseline Methods

Facial Attribute Editing Methods. We first compare our directional latent mapping network with the state-of-the-art facial attribute editing methods (i.e., L2M-GAN (Yang et al., 2021a), and StyleCLIP (Patashnik et al., 2021)) for the specific attribute: "smile". Due to the requirements of high-level semantic-aware knowledge, the smile attribute has become one of the most challenging global attributes. For L2M-GAN, we set the attribute domains as two ("smile" and "sad"), leveraging the domain label to guide the manipulation. For StyleCLIP, we use the latent mapper network as the baseline model, which is marked as StyleCLIP*.

Facial Attribute Transfer Methods. The StarGANv2 (Choi et al., 2020) learns to transform a source image reflecting the style of a given reference image. We compare our SDD-Net with StarGANv2 method on reference-guided image synthesis, which is a challenging facial attribute transfer task. Reference-guided image synthesis requires that various high-level attributes in the reference images, such as hairstyle, makeup, beard and expression, should be transferrd to the target images, while the irrelevant information such as pose and identity should be preserved. For fair comparison, we resize the images to 256×256 in reference-guided image synthesis.

4.4 Results and Analysis

Facial Attribute Editing. The qualitative results are shown in Figure 4. After careful comparison, we have following observations: (1) The L2M-GAN performs well at disentangling attributes during editing. However, due to the strong constrain of orthogonal loss, the manipulated attribute is not obvious. (2) Although the StyleCLIP* method can make the best use of semantic knowledge of CLIP-space, the irrelevant attributes are changed in facial attribute editing without correct guidance direction. (3) Our directional latent mapping network manipulates the attribute correctly and naturally, which demonstrates that the proposed semantic directional consistency (SDC) loss could enforce the editing model to change specific attributes while preserving irrelevant parts. It also proves that there exists latent directions corresponding to different semantic properties in latent space. By manipulating latent code along such direction or its opposite direction, we can add or remove attributes.

Facial Attribute Transfer. Inspired by the above direction-based latent space manipulation for facial attribute transfer, we leverage this peculiarity for facial attribute transfer task.

We first execute our SDD-Net on the "smile" attribute for single attribute transfer, and input the corresponding text prompt to the model. The qualitative results are shown in Figure 5. We observe that when reference has the consistent attribute appointed by the prompt, the SDD-Net could correctly transfer the specific attribute to the target face with irrelevant information preserved, as shown in the top part. On the contrary, when the reference has the opposite attribute, our SDD-Net also could transfer opposite attribute to the target without extra guidance. By fully leveraging the knowledge of CLIP-space, our SDD-Net could find the semantic-aware latent direction in the latent space. The experiments show the excellent performance of our SDD-Net.

For the reference-guide image synthesis, we set the projected vector $\vec{\mathcal{V}_p}$ is equal to the RT vector $\vec{\mathcal{V}_{RT}}$. Figure 6 provides qualitative comparison of the results. We observe that StarGANv2 mothed synthesizes images with a same style code. However, the results show that StarGANv2 model manipulates images in a limited space. As a result, the attributes of the synthetic faces tend to be exactly alike. In addition, the expression attribute such as "smile" is ignored during transferring. Compared to StarGANv2, our SDD-Net transfers the multiple meaningful attributes to each of the target faces in semantic-level. Meanwhile, our method manipulates the image along the latent direction leveraging the knowledge of CLIP in latent space, which allows our method to synthesize realistic facial images, rather than simple style transfer.

4.5 Ablation Study

We conduct a qualitative ablation study, as shown in Figure 7, and show the significance of identity loss. We observe that, when hyperparameter $\lambda_{\text{ID}} = 0.1$, the ID loss hinders the attribute transfer. Then we experiment with $\lambda_{\text{ID}} = 0$, the attribute could be correctly transferred, but the identity information is changed. To trade-off, we set the $\lambda_{\text{ID}} = 0.02$.

Figure 7: The ablation study of identity loss. Under each column we specify (λ_{ID}) identity loss. Obviously the ID loss is significant for facial attribute transfer.

5 Conclusions

In this paper, we first propose directional latent mapping network for text-driven facial attribute editing. By leveraging semantic direction consistency (SDC) loss, the directional latent mapping network could correctly edit relevant attribute while preserving irrelevant attributes. And on this basis, we propose semantic directional decomposition network (SDD-Net) for text-driven facial attribute transfer, which correctly transfers the semantic-aware attributes of reference image to the target image. Experiments show that our method achieves impressive performance on CelebA-HQ dataset.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4432–4441, 2019.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020.

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*, pp. 8789–8797, 2018.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Garoe Dorta, Sara Vicente, Neill DF Campbell, and Ivor JA Simpson. The gan that warped: Semantic attribute editing with unpaired data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5356–5365, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5744–5753, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020.

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.

Zhenliang He, Meina Kan, Jichao Zhang, and Shiguang Shan. Pa-gan: Progressive attention generative adversarial network for facial attribute editing. *arXiv preprint arXiv:2007.05892*, 2020.

Bingwen Hu, Zhedong Zheng, Ping Liu, Wankou Yang, and Mingwu Ren. Unsupervised eyeglasses removal in the wild. *IEEE Transactions on Cybernetics*, 2020.

Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

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

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Jeong-gi Kwak, David K Han, and Hanseok Ko. Cafe-gan: Arbitrary face attribute editing with complementary attention feature. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 524–540. Springer, 2020.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9455–9464, 2018.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3673–3682, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Yotam Nitzan, Rinon Gal, Ofir Brenner, and Daniel Cohen-Or. Large: Latent-based regression through gan semantics. *arXiv preprint arXiv:2107.11186*, 2021.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2287–2296, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12863–12872, 2021.

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)*, pp. 168–184, 2018.

Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4432–4442, 2021.

Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2951–2960, 2021a.

Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 672–681, 2021b.

Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5485–5493, 2017.

Weidong Yin, Ziwei Liu, and Chen Change Loy. Instance-level facial attributes transfer with geometry-aware flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9111–9118, 2019.

A MORE RESULTS

A.1 RESULTS ON VARIOUS ATTRIBUTES

As shown in Figure 8, we employ our SDD-Net on other challenging single attributes such as "black hair" and "bangs", which needs the method to understand the global semantic information. For "black hair" attribute, the SDD-Net projects the RT vector to get the final direction of projected vector. Owing to that the hair color is the opposite of given text prompt, the SDD-Net manipulates the target face in a reverse direction. For "bangs" attribute, our SDD-Net could correctly transfer the fine-grained semantic feature to the target, which leads to refine attribute transfer.

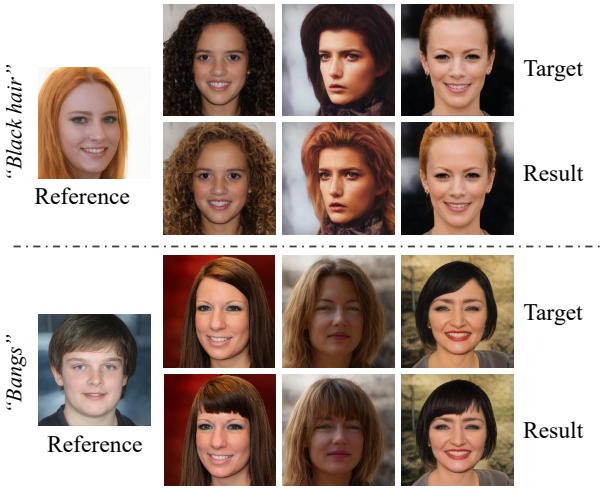

Figure 8: The experiment results on various single attributes such as "black hair" and "bangs".

A.2 RESULTS ON IRRELEVANT ATTRIBUTES.

To further verify the effectiveness of our SDD-Net, we design this experiment to transfer irrelevant attributes while preserving the relevant attribute, as illustrated in Figure 9. We enforce the semantic feature of the manipulated images to change to y_{goal} along the projected vector $\vec{\mathcal{V}_p}$ in CLIP-space. Here we use the vertical vector $\vec{\mathcal{V}_v} = \vec{\mathcal{V}_{RT}} - \vec{\mathcal{V}_p}$, to transfer all of attributes except "smile".

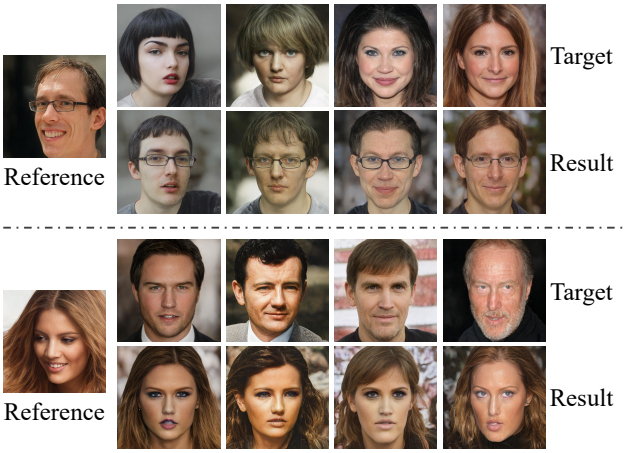

Figure 9: The experiment results on irrelevant attributes transfer. Note that every attributes of the reference, except "smile", are transferred to the target.