Text-guided Unsupervised Latent Transformation for Multi-attribute Image Manipulation

Xiwen Wei[1], Zhen Xu[1], Cheng Liu[2], Si Wu[1,3*], Zhiwen Yu[1], and Hau San Wong[4]
[1]School of Computer Science and Engineering, South China University of Technology
[2]Department of Computer Science, Shantou University
[3]Peng Cheng Laboratory
[4]Department of Computer Science, City University of Hong Kong
{202021044777, csxuzhen}@mail.scut.edu.cn, cliu@stu.edu.cn, {cswusi, zhwyu}@scut.edu.cn, cshswong@cityu.edu.hk

Abstract

Great progress has been made in StyleGAN-based image editing. To associate with preset attributes, most existing approaches focus on supervised learning for semantically meaningful latent space traversal directions, and each manipulation step is typically determined for an individual attribute. To address this limitation, we propose a Text-guided Unsupervised StyleGAN Latent Transformation (TUSLT) model, which adaptively infers a single transformation step in the latent space of StyleGAN to simultaneously manipulate multiple attributes on a given input image. Specifically, we adopt a two-stage architecture for a latent mapping network to break down the transformation process into two manageable steps. Our network first learns a diverse set of semantic directions tailored to an input image, and later nonlinearly fuses the ones associated with the target attributes to infer a residual vector. The resulting tightly interlinked two-stage architecture delivers the flexibility to handle diverse attribute combinations. By leveraging the cross-modal text-image representation of CLIP, we can perform pseudo annotations based on the semantic similarity between preset attribute text descriptions and training images, and further jointly train an auxiliary attribute classifier with the latent mapping network to provide semantic guidance. We perform extensive experiments to demonstrate that the adopted strategies contribute to the superior performance of TUSLT.

1. Introduction

Visual attributes represent semantically meaningful features inherent in images, and attribute manipulation has experienced great improvements, due to the advent of Generative Adversarial Network [13] (GAN)-based generative models, e.g. StyleGAN [21, 22] and StarGAN [7, 8]. Recent works [15, 37, 43] have discovered that the latent space of StyleGAN possesses semantic disentanglement properties, enabling a variety of image editing operations via latent transformations.

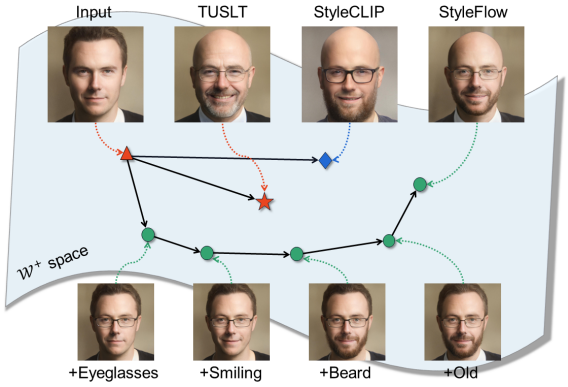

Figure 1. Visually comparing TUSLT with StyleFlow (supervised) and StyleCLIP (text-driven) in precisely manipulating multiple attributes and preserving irrelevant attributes.

StyleGAN-based methods for image attribute manipulation typically involve a large number of manual annotations or well-trained attribute classifiers. Furthermore, the discovered semantic latent directions are associated with individual attributes. The editing on a target attribute is carried out by moving the latent code of an input image along one of the directions. For K target attributes, these models require K transformation steps to handle the translation. As a result, they are not scalable to the increasing number of target attributes in multi-attribute transformation tasks. As shown in Figure 1, we test a state-of-the-art supervised

*Corresponding author.

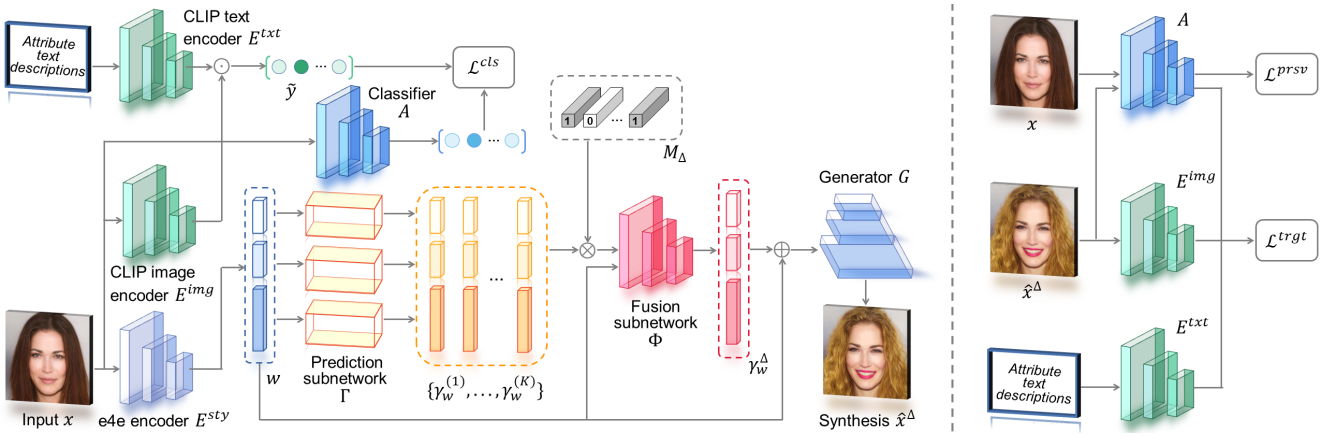

Figure 2. Overview of the proposed model, TUSLT, consisting of two learnable components: an auxiliary attribute classifier A trained on the CLIP-based labeled data, and a latent mapping network $\{\Gamma, \Phi\}$. Γ infers latent directions $\{\gamma_w^{(1)}, \ldots, \gamma_w^{(K)}\}$ for preset attributes, and Φ transforms the target-related directions as indicated by mask M_Δ into a residual vector γ_w^Δ, to which the initial latent code is added. Precise multi-attribute transfer is allowed by such a single transformation step, and the generator G synthesizes a new image reflecting the target attributes under the guidance of A and CLIP encoders.

model, StyleFlow [2], and find that multiple transformation steps lead to undesired deviation from the input image on irrelevant attributes. Compared to the state-of-the-art text-driven model, StyleCLIP [33], we can also achieve a better manipulation result by seeking a single latent transformation step for the task.

More specifically, we propose a Text-guided Unsupervised StyleGAN Latent Transformation (TUSLT) model that supports simultaneous manipulation on multiple attributes. As shown in Figure 2, the key is to jointly learn a mapping network to infer the latent transformation and an auxiliary attribute classifier to assess manipulation quality. We employ the Contrastive Language-Image Pre-training (CLIP) model [34] to generate pseudo-labeled data by measuring the semantic similarities between attribute text descriptions and training images. Compared to CLIP, the jointly trained classifier extracts domain-specific information to better characterize the differences among attributes. This benefits the mapping network to seek more suitable transformations, such that the synthesized images reflect target attributes. Further, we adopt a two-stage architecture for the mapping network: the earlier stage employs a prediction subnetwork to infer a set of semantic directions, and the latter stage operates on the resulting directions and nonlinearly fuses the target-related ones. The intermediate semantic directions are associated with preset attributes and tailored for the input image. This design allows us to deal with a wide range of attribute combinations in a single transformation step. We perform extensive experiments and provide both qualitative and quantitative results in diverse multi-attribute transformation tasks, showing the superiority of our model over the competing methods.

In summary, the main contributions of this work are given as follows: (a) The existing image editing methods focus on discovering semantic latent directions associated with individual visual attributes, and a sequential manipulation process is thus needed for multi-attribute manipulation. In contrast, the proposed model infers a single step of latent space walk to simultaneously manipulate multiple attributes. (b) Benefiting from the cross-modal text-image representation of CLIP, we jointly train a latent mapping network with an auxiliary attribute classifier, which leads to more precise attribute rendering without requiring additional manual annotations. (c) Due to the two-stage nature, our latent mapping network breaks down the challenging multi-attribute manipulation task into sub-tasks: inferring diverse semantic directions and integrating the target-related ones into a single transformation vector. This design gives our model interpretability and flexibility in dealing with a variety of attribute combinations.

2. Related Work

2.1. Generic Image-to-image Translation

As one of the earliest image translation models, pix2pix [18] learnt a cross-domain mapping via conditional GAN [31]. In addition to an adversarial training loss, the consistency regularization between each input image and the corresponding ground truth was imposed in the training process. To alleviate the problem of pairwise training data, many unpaired image translation models have been developed. UNIT [29] trained two generators to approximate the joint distribution of images from different domains. Cy-

cle consistency regularization was adopted in CycleGAN [48] and DiscoGAN [23]. To efficiently learn the mappings among multiple domains, StarGAN [7, 9] was a unified framework in which a single generator was trained to translate an input image into different domains, conditioned on domain information.

2.2. Attribute Manipulation

FaderNet [26] used an encoder-decoder architecture and learnt attribute-aware latent representations in an unsupervised manner. To ensure correct manipulation on the desired attributes, an attribute classification constraint was incorporated in AttGAN [16]. Instead of handling all target attributes, STGAN [28] contained a selective transfer module to only encode the changed attributes, based on the difference between source and reference images. In IOA-GAN [5], the relationship of preset attributes were leveraged via a Graph Convolutional Network (GCN) [25] to infer and inject an integrated embedding of attributes into a translation network.

The latent spaces of GANs demonstrate promising semantic organization [10, 12, 19]. With the progress of high-fidelity GAN inversion [14, 36, 40], there are many works on exploring the latent space of a pre-trained StyleGAN generator for image manipulation [3, 41]. SeFa [39] disentangled semantics from StyleGAN by decomposing the matrix of generator weights. In addition, LEFS [47] applied sparse representation learning to unsupervised semantic disentanglement. As a representative supervised method, InterFaceGAN [38] searched for semantic latent directions by solving a series of binary classification problems about preset attributes. Instead of latent directions, StyleSpace [43] explored style-associated channels with the help of a pre-trained classifier. Different from the above linear latent transformation methods, StyleFlow [2] employed a conditional flow model to learn non-linear paths for attribute manipulation.

2.3. Text-guided Image Editing

Cross-modal representation learning on visual and language data brings about substantive progress in text-guided image synthesis. An early attempt was to train a GAN-based model [31], conditioned on text embedding [35]. Further, a stacked structure [46] and an attention-based regularization approach [45] were designed to improve synthesis quality. In [30, 32], image content from visual attributes were disentangled and associated with text descriptions without using manually annotated data. ManiGAN [27] aimed to learn text and image cross-modality representations, such that semantic regions were associated with the corresponding text via an affine transformation. TediGAN [44] aligned and projected text and image representations into the latent space of StyleGAN. Further, StyleCLIP [33] integrated CLIP [34] with StyleGAN to leverage CLIP-based linguistic-visual semantic consistency regularization for better manipulation quality.

Among the aforementioned models, StyleFlow [2] and StyleCLIP [33] are the most relevant to our work. StyleFlow is a supervised method, and the learnt semantic directions are associated with individual attributes. For multi-attribute manipulation, StyleFlow needs a sequential transformation process, in which the target attributes are edited one at a time. In contrast, the proposed model learns to infer a single transformation step to simultaneously manipulate multiple attributes in an unsupervised manner. StyleCLIP trains a dedicated mapping network for each text description, and thus has limited scalability in handling a variety of attribute combinations. Unlike StyleCLIP, our model uses a two-stage mapping network and jointly trains an auxiliary attribute classifier without manual annotations. This structure benefits the scalability and semantic accuracy in changing multiple attributes.

3. Proposed Method

Our goal is to translate an input image into a new one reflecting the target attributes. Based on the linguistic-visual representation of CLIP, we can pseudo-label training images by measuring the semantic similarities between attribute text descriptions and training images. We believe that an auxiliary attribute classifier can capture the most discriminative information, which complements the role of CLIP encoders to a certain extent. Based on this, the constructed supervision is leveraged by jointly training a latent mapping network and the classifier. The former has a two-stage architecture to infer a diverse set of semantic latent directions followed by selectively integrating target-related ones, and the latter provides semantic guidance. As a result, multi-attribute manipulation can be precisely carried out in a single forward pass.

3.1. Auxiliary Attribute Classifier

CLIP consists of a text encoder E^{txt} and an image encoder E^{img}, and encodes both types of input into 512-D embedding vectors. Let $T = \{t^{(1)}, \dots, t^{(K)}\}$ denote a set of text prompts, and $t^{(i)}$ describes the i-th preset attribute. To identify the attributes reflected in images, we embed training images and T in the shared embedding space, and measure the semantic similarity as

$$\mathcal{S}^{(i)}(x) = \cos(E^{txt}(t^{(i)}), E^{img}(x)), \tag{1}$$

where $\cos(\cdot,\cdot)$ denotes the cosine distance between input vectors. $\mathcal{S}^{(i)}(x)$ should be larger when $t^{(i)}$ and x represent the same attribute. At this point, we pseudo-annotate training images, and the corresponding label $\tilde{y}$ is defined as

$$\tilde{y}^{(i)} = \begin{cases} 1, & \text{if } \mathcal{S}^{(i)}(x) > \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\tilde{y}^{(i)}$ indicates whether image x reflects attribute i, and τ is a threshold. Based on the constructed supervision, we train an auxiliary classifier A to capture the visual characteristics of preset attributes. For multi-attribute recognition, the evaluation term is defined as

$$\mathcal{L}^{cls} = \mathbf{E}_x \Big[\sum_{i=1}^{K} -\tilde{y}^{(i)} \log A^{(i)}(x) \Big], \tag{3}$$

where $A^{(i)}(\cdot)$ represents the predicted probability of an input reflecting attribute i. Different from E^{img} that is trained on generic data, A focuses more on the domain-specific data, and can be expected to learn the most discriminative features to identify preset attributes.

3.2. Latent Mapping Network

The latent space $\mathcal{W}$ of StyleGAN possesses semantic disentanglement properties, and the works [1, 36, 40] further extend to $\mathcal{W}+$ for better inversion quality. For a given input image x, we adopt the e4e model E^{sty} [40] to predict its latent code $w = E^{sty}(x)$ in the $\mathcal{W}+$ space.

It is challenging to directly infer the latent transformations for a variety of attribute combinations. To address this problem, we design a two-stage architecture for our latent mapping network. The first stage is based on a direction prediction subnetwork Γ that produces latent directions denoted by $\Gamma(w) = \{\gamma_w^{(1)}, \ldots, \gamma_w^{(K)}\}$, where $\gamma_w^{(i)}$ associates with the preset attribute i, conditioned on w. At the second stage, a fusion subnetwork Φ operates on the produced directions. We define a binary vector $\Delta \in \{0, 1\}^K$ to indicate target attributes. Φ learns to integrate the directions as indicated by Δ, and infer a residual vector γ_w^{Δ} defined as

$$\gamma_w^{\Delta} = \Phi(w, M_{\Delta} \otimes \Gamma(w)), \tag{4}$$

where the mask M_{Δ} is constructed by broadcasting Δ across an array having the same dimensionality as the latent code, and $\otimes$ denotes the Hadamard product. As a result, multi-attribute manipulation can be carried out by simply adding the initial latent code to the residual vector. The generator G of StyleGAN is employed to decode the resulting latent vector as

$$\hat{x}^{\Delta} = G(w + \alpha \gamma_w^{\Delta}), \tag{5}$$

where α controls the manipulation strength. Although our mapping network stacks two stages, each stage has access to the latent code of the input image.

To ensure that the intermediate directions are semantically meaningful, they are required to work directly on the manipulation process of individual attributes. In this case, Φ should not produce any changes to a single direction as input, and the corresponding invariance loss is formulated as follows:

$$\mathcal{L}^{sngl} = \mathbf{E}_x \Big[\sum_{i=1}^{K} \big\| \Phi(w, M_i \otimes \Gamma(w)) - \gamma_w^{(i)} \big\|_2 \Big], \tag{6}$$

where M_i denotes the mask for selecting $\gamma_w^{(i)}$ only. This design allows the mapping network to flexibly learn transformations with respect to different attribute combinations.

3.3. Semantic Evaluation

Target attribute identification. The synthesized image $\hat{x}^{\Delta}$ should properly reflect the target attributes as indicated by Δ. We leverage both classifier A and CLIP encoders $\{E^{txt}, E^{img}\}$ to impose a target attribute identification loss, and the formulation is expressed as follows:

$$\begin{aligned} \mathcal{L}^{trgt} = \mathbf{E}_{(x,\Delta)} \Big[& \sum_{i=1}^{K} -\Delta^{(i)} \log A^{(i)}(\hat{x}^{\Delta}) \\ & + \sum_{i=1}^{K} \Delta^{(i)} \big(1 - \cos(E^{txt}(t^{(i)}), E^{img}(\hat{x}^{\Delta}))\big) \Big]. \end{aligned} \tag{7}$$

In the above equation, the first term evaluates the predictions of A, and the second term measures the semantic consistency between the manipulation result and attribute text descriptions in the embedding space of CLIP. By minimizing $\mathcal{L}^{trgt}$, our mapping network seeks suitable transformations such that the target attributes can be well reflected in both views of A and CLIP. In this way, we significantly improve the semantic accuracy of the transformation.

Non-target attribute preservation. Minimizing the above losses does not guarantee that the synthesized images properly preserve the content of the input images while at the same time changing only the part related to the target attributes. To ensure that irrelevant attributes are unchanged before and after transformation, we further impose an attribute-aware consistency regularization on the mapping model. Specifically, we define the attribute-aware representation as

$$f_A^{(i)}(\hat{x}^{\Delta}) = \nu^{(i)} \otimes f_A(\hat{x}^{\Delta}), \tag{8}$$

where $f_A(\cdot)$ denotes the attribute classifier features, and $\nu^{(i)}$ represents the weight vector of the head associated with the i-th attribute. With the help of $\nu^{(i)}$, we can suppress the less informative features and highlight the useful ones due to the reason that the i-th head measures its emergence in images. By modulating the classifier features in this way, the resulting ones capture more specific information, and we thus measure the semantic consistency on non-target attributes as

$$\mathcal{L}^{prsv} = \mathbf{E}_{(x,\Delta)} \Big[\sum_{i=1}^{K} (1 - \Delta^{(i)}) \big\| f_A^{(i)}(\hat{x}^{\Delta}) - f_A^{(i)}(x) \big\|_2 \Big]. \tag{9}$$

Compared to the predictions of A, we find that the features $\{f_A^{(1)}, \ldots, f_A^{(K)}\}$ encode richer information on the attributes. Minimizing $\mathcal{L}^{prsv}$ enforces our model to preserve the non-target characteristics of the input image while manipulating the target attributes faithfully.

3.4. Model Optimization

By integrating the above aspects of classifier training, mapping regularization and evaluation, the optimization formulation of the learnable components is expressed as follows:

$$\begin{aligned} &\min_{A} \mathcal{L}^{cls}, \\ &\min_{\Gamma} \mathcal{L}^{trgt} + \mathcal{L}^{prsv} + \lambda \mathcal{L}^{loc} \\ &\min_{\Phi} \mathcal{L}^{trgt} + \mathcal{L}^{prsv} + \lambda \mathcal{L}^{loc} + \mu \mathcal{L}^{sngl}, \end{aligned} \tag{10}$$

where the local searching term $\mathcal{L}^{loc}$ is defined as

$$\mathcal{L}^{loc} = \mathbf{E}_{(x,\Delta)} \big[\max(\|\gamma_w^{\Delta}\|_2 - \epsilon, 0) \big], \tag{11}$$

and ϵ is a margin. Minimizing $\mathcal{L}^{loc}$ prevents the modified latent code from deviating too far from the initial one. In addition, λ and μ in Eq.(10) are the weighting factors that control the relative importance of the regularization terms, compared to the semantic evaluation terms, respectively. We jointly train $\{A, \Gamma, \Phi\}$ from scratch to flexibly transform input images by randomly specifying the target attributes. Note that we do not perform a dedicated optimization for individual attribute (combination). In the test stage, both Γ and Φ are fixed, and attribute manipulation is performed in a single forward pass for any given data.

4. Experiments

In this section, we first describe the test datasets and evaluation setups. Next, we investigate the effect of the main components of TUSLT. We further compare our model against multiple leading methods by performing both user studies and quantitative evaluation. Note that all the experimental results are obtained by applying the model on unseen images during the training phase.

4.1. Experimental Setup

Datasets and preset attributes CelebA-HQ [20] is a widely used benchmark for facial image editing, and contains 30k high-resolution face images of celebrities. We follow [33] to set 38 attributes on hair style, hair color, expression, gender, age and others. We also use the animal and anime face datasets, AFHQ-cats/dogs [9] and Danbooru AnimeFace [4], with large intra-domain differences to evaluate the proposed model in manipulating 4 and 6 preset attributes, respectively.

Figure 3. Single-attribute transformation results of StyleCLIP and TUSLT.

Architecture and hyperparameters. The proposed model consists of three pre-trained networks (CLIP text and image encoders $\{E^{txt}, E^{img}\}$ and an e4e encoder E^{sty}) and two learnable components (an attribute classifier A and a latent mapping network $\{\Gamma, \Phi\}$). We adopt a ResNet-50 [6] for A, and the network architectures of Γ and Φ consist of 4 fully connected layers. The threshold τ in Eq.(1), the coefficient α in Eq.(5), the weighting factors λ and μ in Eq.(10) and the margin ϵ in Eq.(11) are set to 0.7, 0.1, 0.8, 1 and 0.01, respectively. The proposed model is trained using the Adam optimizer [24] with a learning rate of 0.5. There are 50k training iterations, and each batch contains 2 images. We empirically find that our model can converge to a good solution without heavy tuning.

Evaluation protocols. For all competing methods, we use the open source codes or pre-trained models. We assess the diversity and manipulation quality of synthesized images using the Fréchet Inception Distance (FID) [17]. On the other hand, we measure the correctness of attribute manipulation by an independent attribute classifier [28], and report the Target Attribute Recognition Rate (TARR). To quantitatively assess the model performance in irrelevant attribute preservation, we use the metric of IDentity Distance (IDD) before and after transformation in the feature space of a well-trained face recognition network [11].

4.2. Analysis of Main Components

Semantically meaningful directions. We first perform an experiment to demonstrate the effectiveness of the prediction subnetwork Γ in inferring diverse semantic directions associated with preset attributes. For individual attributes, the manipulation is performed along the directions produced by Γ. In Figure 3, we visually compare with the main competing method, StyleCLIP, and observe that TUSLT is able to render more precise semantics about the target attributes.

Precise manipulation on multiple attributes. To assess the fusion subnetwork Φ, we can feed the directions

Figure 4. Ablative experiment results. The combinations of target attributes are (*upper*) 'Bald +Beard +Happy' and (*bottom*) 'Happy +Makeup +Bushy-eyebrows'.

Figure 5. Comparison between TUSLT and variants in terms of (*left*) TARR and (*right*) IDD.

$\gamma_{CLIP} = \{\gamma^{(1)}_{CLIP}, \ldots, \gamma^{(K)}_{CLIP}\}$ learnt by StyleCLIP, and the resulting model is referred to as 'γ_{CLIP}+Φ'. On the other hand, we build another variant 'TUSLT w/o A' by disabling the attribute classifier. In addition to visualizing the synthesis results in Figure 4, we report the TARR and IDD scores of the variants in Figure 5. 'γ_{CLIP}+Φ' achieves comparable performance with StyleCLIP in terms of TARR, while the IDD value of the former is significantly lower than that of the latter. This confirms the effectiveness of Φ in identity preservation. The results of 'TUSLT w/o A' are unsatisfactory due to the lack of domain-specific knowledge learnt from the target data. In contrast, benefiting from A, our full model exhibits more precise manipulation ability.

Irrelevant attribute preservation. It is also an important aspect to preserve the visual characteristics of the input image apart from the target attributes before and after transformation. We construct the attribute-aware features by Eq.(8) and impose the corresponding regularization term $\mathcal{L}^{prsv}$ defined in Eq.(9). We remove $\mathcal{L}^{prsv}$ from the overall training loss to build a variant 'TUSLT w/o $\mathcal{L}^{prsv}$'. In Figures 4-5, we can observe that removing $\mathcal{L}^{prsv}$ leads to more significant changes on the target attributes (a higher TARR score), while at the same time increasing the identity distance between input and transformed images. The results suggest that our strategy leads to a significant improvement in the maintenance of irrelevant attributes.

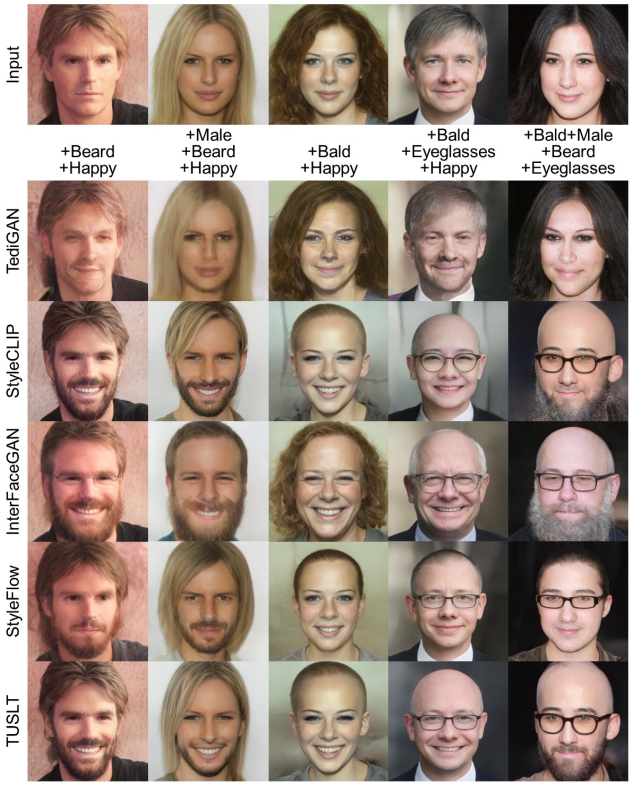

Figure 6. Multi-attribute manipulation results of TUSLT and competing methods. Note that we examine the common attributes which all the methods are able to manipulate.

4.3. Human Evaluation

We perform user studies to evaluate TUSLT and a number of main leading methods in multi-attribute manipulation tasks. For InterFaceGAN and StyleFlow, we perfor-

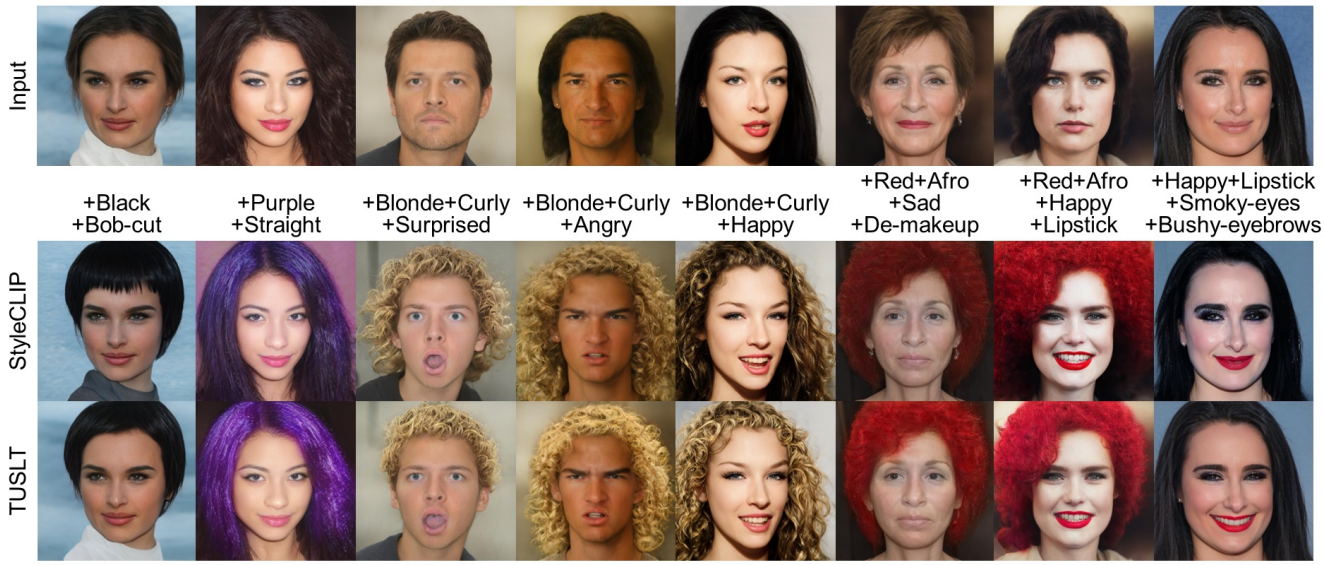

Figure 7. Diverse image synthesis results of TUSLT and StyleCLIP.

Table 1. The result of ranking TUSLT and competing models in multi-attribute transfer tasks. MA, IAP and VR denote Manipulation Accuracy, Irrelevant Attribute Preservation and Visual Realism, respectively.

Method	MA	IAP	VR
TUSLT (Ours)	**1.25**	**1.31**	**1.44**
StyleCLIP [33]	2.51	2.44	2.81
StyleFlow [2]	2.63	3.25	2.56
InterFaceGAN [38]	3.69	3.31	3.38
TediGAN [44]	4.94	4.69	4.81

Table 2. Quantitative comparison of TUSLT and competing methods.

Method	FID (↓)	IDD (↓)	PSNR (↑)	SSIM (↑)	TARR (↑)
TUSLT (Ours)	**56.91**	**0.45**	**24.92**	**0.75**	**87.85**
StyleCLIP [33]	63.93	0.48	19.24	0.71	85.74
StyleFlow [2]	61.53	0.46	21.84	0.73	85.10
InterFaceGAN [38]	69.31	0.51	17.92	0.68	81.80
TediGAN [44]	58.74	0.49	20.61	0.69	18.93

m multi-step translations to manipulate multiple target attributes. There are 18 attribute combinations, and 10 questions for each one. For each question, given an input image and the target attributes, the options are the images synthesized by the competing methods, and the order is randomly shuffled. A total of 50 validated workers are instructed to rank the synthesized images in terms of manipulation accuracy, irrelevant attribute preservation and visual realism. Table 1 summarizes the average ranking values of the competing methods. TUSLT is able to produce the best transformation results in most cases. We also provide the representative synthesis results of the competing methods in Figure 6. We observe that TediGAN fails in most of the manipulation tasks. In contrast, TUSLT produces images with more precise manipulation results, compared to the main competing models, StyleFlow and StyleCLIP. We can also find that TUSLT is the only model which is able to successfully manipulate all four attributes in the last example. In Figure 7, we visually compare with StyleCLIP in more tasks, and find that TUSLT yields more natural and plausible images.

4.4. Quantitative Evaluation

We further perform a quantitative comparison between the proposed approach and competing methods. Note that this experiment involves the common attribute combinations which all of the competing models are able to manipulate. In Table 2, we report the average quantitative comparison results of our TUSLT and the competing methods in terms of FID, TARR and IDD. In addition, we adopt the metrics of Peak Signal-to-Noise Ratio (PSNR) and Structure SIMilarity (SSIM) to measure low to mid-level similarity between the input and synthesized images. The manipulations are performed on 5 attribute combinations. TUSLT outperforms the competing models by a large margin in both manipulation quality and irrelevant attribute preservation. In particular, our model achieves the highest PSNR/TARR score of 24.92/87.85, which is higher than that of the second best method (StyleFlow 21.84/85.10) by about 3 points. We further perform a comparison with a CLIP-based image manipulation method, HairCLIP, which focuses on editing hair color and style. Some representative synthesized images are shown in Figure 8, and the corresponding quantitative assessment on 6 attribute combina-

Figure 8. Visual comparison between TUSLT and HairCLIP in hair attribute manipulation.

Table 3. Quantitative comparison of TUSLT and HairCLIP.

Method	FID (↓)	PSNR (↑)	s_{CLIP} (↑)
TUSLT (Ours)	**50.10**	**27.14**	**0.32**
HairCLIP [42]	50.83	26.31	0.30

Figure 9. Synthesis results of TUSLT on AFHQ.

tions is reported in Table 3. To measure manipulation accuracy, we compute the cosine similarity s_{CLIP} between the attribute text descriptions and synthesized images in the CLIP feature space. HairCLIP performs less satisfactorily in rendering hair color as described in the text prompts.

Discussion. One can find that the results lead to the same conclusions as our user studies. This is mainly due to the joint training of an auxiliary attribute classifier with CLIP-based supervision, and thus precise semantics of preset attributes can be captured. Another reason is the adopted two-stage architecture for latent transformation. Fusing semantically meaningful latent directions simplify the task of multi-attribute manipulation. This allows the proposed model to flexibly transform the input image to reflect a variety of attribute combinations.

4.5. Results on AFHQ and AnimeFace

We also show the ability of the proposed model to manipulate multiple attributes on AFHQ-cats/dogs and AnimeFace. For AFHQ, we follow the setting of StyleCLIP to specify preset attribute text descriptions and train TUSLT

Figure 10. Synthesis results of TUSLT on AnimeFace.

to capture the corresponding semantics. For AnimeFace, we use the same setting as the experiments on CelebA-HQ. The results shown in Figures 9-10 confirm again that our model is scalable to multi-attribute transformations on animal and anime facial images and lead to significant visual modification.

5. Conclusion

We propose a text-guided unsupervised multi-attribute manipulation model to edit images in a single latent transformation step. Benefiting from the cross-modal image and text representation of CLIP, we can jointly train an auxiliary attribute classifier and a latent mapping network for precise attribute manipulation. We suggest two main reasons why the proposed model is able to successfully manipulate multiple attributes on diverse input images, compared to the leading methods. First, the use of a two-stage architecture enables our model to take care of all preset individual attributes, and thus provide the flexibility to handle diverse attribute combinations. Second, in contrast to CLIP encoders, the classifier learns domain-specific features to identify preset attributes, and offers semantic guidance not only for manipulating target attributes, but for preserving irrelevant attributes of the input image. This work significantly increases the scalability of StyleGAN-based image attribute manipulation without causing any manual annotation cost.

Acknowledgments

This work was supported in part by the China Scholarship Council, in part by the National Natural Science Foundation of China (Project No. 62072189), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11206622), and in part by the Natural Science Foundation of Guangdong Province (Project No. 2022A1515011160).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: how to embed images into the StyleGAN

latent space? In *Proc. International Conference on Computer Vision*, 2019. 4

[2] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. StyleFlow: attribute-conditioned exploration of StyleGAN-Generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics*, 40(3):1–21, 2021. 2, 3, 7

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: age transformation using a style-based regression model. *ACM Transactions on Graphics*, 40(4):1–12, 2021. 3

[4] Gwern Branwen Aaron Gokaslan Anonymous, the Danbooru community. Danbooru2018: A large-scale crowdsourced and tagged anime illustration dataset. https://www.gwern.net/Danbooru2018, January 2019. 5

[5] Binod Bhattarai and Tae-Kyun Kim. Inducing optimal attribute representations for conditional GANs. In *Proc. European Conference on Computer Vision*, 2020. 3

[6] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019. 5

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 1, 3

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: diverse image synthesis for multiple domains. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: diverse image synthesis for multiple domains. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 3, 5

[10] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: uncovering the local semantics of GANs. In *Proc. IEEE conference on Computer Vision and Pattern Recognition*, 2020. 3

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: additive angular margin loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5

[12] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Philip Isola. GANalyze: toward visual definitions of congnitive image properties. In *Proc. International Conference on Computer Vision*, 2019. 3

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Neural Information Processing Systems*, 2014. 1

[14] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code GAN prior. In *Proceedings of IEEE conference on computer vision and pattern recognition*, pages 3012–3021, 2020. 3

[15] Erik Harkonen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: discovering interpretable GAN controls. In *Proc. Neural Information Processing Systems*, 2020. 1

[16] 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. 3

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, and Bernhard Nessler. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. Neural Information Processing Systems*, 2017. 5

[18] Philip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[19] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *arXiv:1907.07171*, 2019. 3

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. International Conference on Learning Representation*, 2018. 5

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[23] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proc. International Conference on Machine Learning*, 2017. 3

[24] Diederik P. Kingma and Jimmy Lei Ba. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representation*, 2015. 5

[25] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proc. International Conference on Learning Representation*, 2017. 3

[26] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: manipulating images by sliding attributes. In *Proc. Neural Information Processing Systems*, 2017. 3

[27] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. ManiGAN: text-guided image manipulation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[28] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3, 5

[29] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proc. Neural Information Processing Systems*, 2017. 2

[30] Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. Describe

what to change: a text-guided unsupervised image-to-image translation approach. In *Proc. ACM International Conference on Multimedia*, 2020. 3

[31] M. Mirza and S. Osindero. Conditional generative adversarial nets. In *arXiv:1411.1784*, 2014. 2, 3

[32] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In *Proc. Neural Information Processing Systems*, 2018. 3

[33] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: text-driven manipulation of StyleGAN imagery. In *Proc. International Conference on Computer Vision*, 2021. 2, 3, 5, 7

[34] Alec Redford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning*, 2021. 2, 3

[35] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proc. International Conference on Machine Learning*, 2016. 3

[36] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Deniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3, 4

[37] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[38] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2004–2018, 2020. 3, 7

[39] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[40] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGan image manipulation. *ACM Transactions on Graphics*, 40(4):1–14, 2021. 3, 4

[41] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Cross-domain and disentangled face manipulation with 3D guidance. *IEEE Transactions on Visualization and Computer Graphics (Early Access)*, pages 1–15, 2021. 3

[42] Tianyi Wei, Dongdong Chen, Wenbo Zhou, and Jing Liao. HairCLIP: design your hiar by text and reference image. In *Proc. IEEE conference on Computer Vision and Pattern Recognition*, 2022. 8

[43] Zongze Wu, Dani Lischinski, and Eli Shechtman. StyleSpace analysis: disentangled controls for StyleGAN image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 3

[44] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. TediGAN: text-guided diverse face image generation and manipulation. In *arXiv:2012.03308*, 2020. 3, 7

[45] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In *Proc. IEEE conference on Computer Vision and Pattern Recognition*, 2018. 3

[46] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. International Conference on Computer Vision*, 2017. 3

[47] Yutong Zheng, Yu-Kai Huang, Ran Tao, Zhiqiang Shen, and Marios Savvides. Unsupervised disentanglement of linear-encoded facial semantics. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[48] Jun-Yan Zhu, Taesung Park, Philip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. International Conference on Computer Vision*, 2017. 3