Image-to-image Translation via Hierarchical Style Disentanglement

Xinyang Li[1], Shengchuan Zhang[1][*], Jie Hu[1], Liujuan Cao[1], Xiaopeng Hong[2], Xudong Mao[1],
Feiyue Huang[3], Yongjian Wu[3], Rongrong Ji[1]
[1]Xiamen University, [2]Xi'an Jiaotong University, [3]Tencent Youtu Lab

{imlixinyang, hujie.cpp, xudonmao}@gmail.com, {zsc_2016, caoliujuan, rrji}@xmu.edu.cn,
hongxiaopeng@mail.xjtu.edu.cn, {garyhuang,littlekenwu}@tencent.com

Abstract

Recently, image-to-image translation has made significant progress in achieving both multi-label (i.e., translation conditioned on different labels) and multi-style (i.e., generation with diverse styles) tasks. However, due to the unexplored independence and exclusiveness in the labels, existing endeavors are defeated by involving uncontrolled manipulations to the translation results. In this paper, we propose Hierarchical Style Disentanglement (HiSD) to address this issue. Specifically, we organize the labels into a hierarchical tree structure, in which independent tags, exclusive attributes, and disentangled styles are allocated from top to bottom. Correspondingly, a new translation process is designed to adapt the above structure, in which the styles are identified for controllable translations. Both qualitative and quantitative results on the CelebA-HQ dataset verify the ability of the proposed HiSD. The code has been released at https://github.com/imlixinyang/HiSD.

1. Introduction

Recently, deep learning based methods have achieved promising results in image-to-image translation area. Early works [56, 50, 26, 43] learn a deterministic mapping between two domains, which give rise to two emergent issues: translating the inputs conditioned on multiple labels, and generating diverse outputs with multiple styles. The former is termed the multi-label task, and the latter is termed the multi-style (or multi-modal) task. For the multi-label task, methods [5, 13, 25, 47] combine the labels into the translator. For the multi-style task, methods [18, 23, 1, 57] incorporate latent codes drawn from Gaussian noise into the translator. Recent unified solutions for these tasks can be classified into two categories. (i). Works [41, 45, 51, 24] learn the shared style by injecting the style code concatenated with the target labels into the generator. The shared
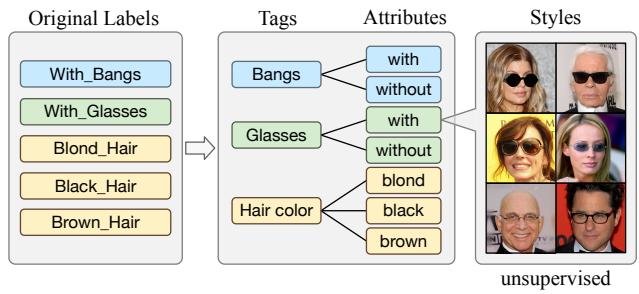
---

* Corresponding Author.

Figure 1: Hierarchical Style Disentanglement. The original labels are organized into independent tags and exclusive attributes. We aim to disentangle the styles to represent the clear manifestations in attributes, in an unsupervised way.

style code does not have an explicit effect on the source image without changed labels, which is shown in Figure 3(a). (ii). StarGANv2 [6] learns the mixed style by using the target label to index the mapped style code. It continues to use the hypothesis of StarGAN [5] that an image domain is the set of images sharing the same labels. The translations frequently involve unnecessary manipulations like changing facial identity and affecting background, as shown in Figure 3(b). In addition, they cannot independently learn the respective styles for bangs, glasses, and hair color. These uncontrollable translations severely limit their practical use.

We propose a novel framework, called Hierarchical Style Disentanglement, to solve the above limitations. We notice the general independence and exclusiveness among most label annotations. For example, in CelebA, original binary labels 'With_Bangs' and 'With_Glasses' are independent, while 'Blond_Hair' and 'Black_Hair' are exclusive. Accordingly, as shown in Figure 1, we organize the original labels into a hierarchical structure, including independent tags and exclusive attributes. The tags represent different accordance of attributes, and every image is relabeled to one of the attributes for each tag. For example, according to the tag 'Glasses', the attribute of images can be ei-
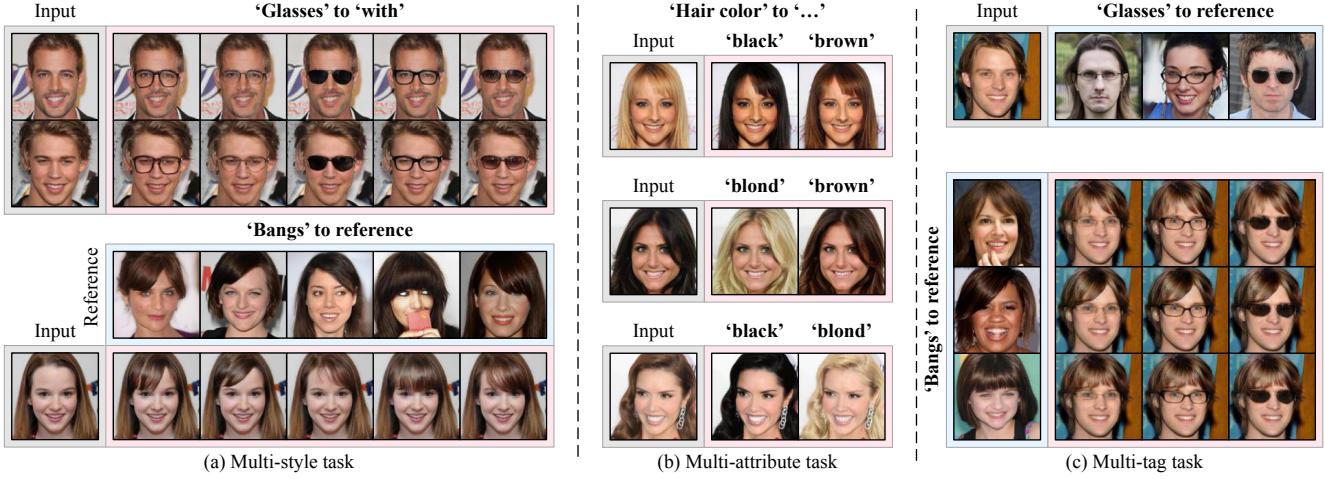
Figure 2: Selected results of our method on CelebA-HQ. **(a)**. The multi-style task, which aims to generate diverse tag-relevant styles. The styles in our framework can be either generated by random latent codes or extracted from reference images. **(b)**. The multi-attribute task, which aims to translate images into multiple possible attributes. **(c)**. The multi-tag task, which aims to manipulate multiple tags of images simultaneously and independently.
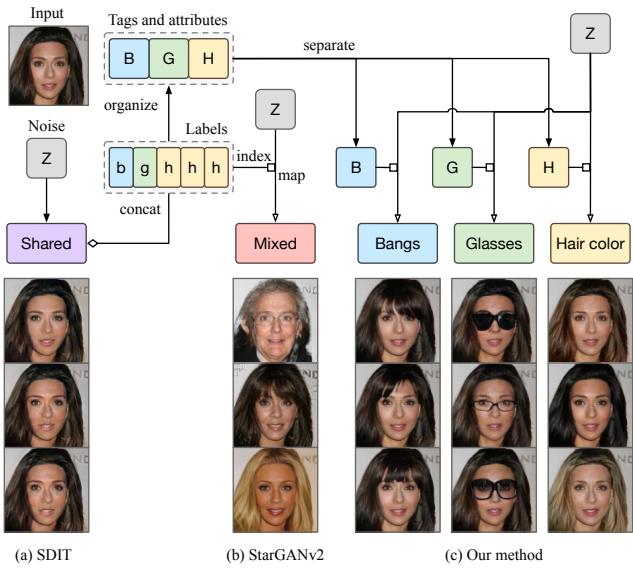

Figure 3: Comparison of different style codes. Our style codes are identified to the hierarchical structure.

ther 'with' or 'without'. Thus the multi-label issue is divided into two sub-tasks: multi-attribute task, which translates a tag to multiple possible attributes; and multi-tag task, which manipulates multiple tags simultaneously. However, the human-annotated attributes cannot represent the clear manifestations in images for tags. In this paper, we take the clear manifestations in images for tags as tag-relevant styles. The tag-relevant styles, which are identified to the tags and attributes, provide a more controllable manner for diverse translations. For example, the style for tag 'Glasses'

can disentangle different glasses, such as myopic glasses, sunglasses, and reading glasses in images, without supervised annotations. We introduce different modules to generate, extract, and efficiently manipulate the disentangled tag-relevant styles. In the cycle-translation path, we consistently optimize both generated and extracted styles to manipulate images realistically and accurately. Through cycle consistency and style consistency, the generated and extracted styles are guaranteed to include the detailed manifestations for tags. To guarantee the disentanglement, we introduce a local translator, which uses the attention mask to avoid the global manipulations; and a tag-irrelevant conditional discriminator, which uses redundant labels in the annotations to prevent that these implicit conditions are manipulated by the translations. In Figure 2, we show some selected results of our method on CelebA-HQ.

Our contributions include:

- We propose HiSD to address the issues in recent multi-label and multi-style image-to-image translation methods by organizing the labels into a hierarchical structure, where independent tags, exclusive attributes, and disentangled styles are allocated from top to bottom.

- To make the styles identified to the tags and attributes, we carefully redesign the modules, phases, and objectives. For unsupervised style disentanglement, we introduce two architectural improvements to avoid the global manipulations and implicit attributes to be manipulated during the translations.

- We conduct extensive experiments to prove the effectiveness of our model.

2. Related Works

Generative Adversarial Networks. GANs [8] have gained remarkable results. After training, the generator is able to produce outputs which are similar to the real samples. Recently, many works optimize the training stability of GANs [30, 9, 34] or release the potential of GANs in different areas [39, 22, 40]. Specifically, We use multi-task GANs [32, 27] to make outputs satisfy the target attribute and conditional GANs [33] to ensure that the tag-irrelevant conditions remain satisfying during translations.

Image-to-image translation. Image-to-image translation has attracted increasing attentions since its widely practical use, such as colorization [52], super resolution [46], semantic synthesis [3] and domain adaption [15]. Our framework focuses on the broad concept of image-to-image translation to jointly solve multi-label [5, 13, 25, 47] and multi-style [18, 23, 1, 57] issues, and overcomes the disadvantages of previous joint frameworks [41, 45, 51, 24, 6, 48, 55, 49, 10].

Label-specific style. Label-specific style learns the clear manifestation of the specific binary label, which is a special case of our tag-relevant style. However, it ignores the exclusiveness of some labels. Early works [48, 55, 37, 4, 2] use a single or multi-valued style code while recent methods [49, 10] use the feature maps to represent the detailed manifestation. However, they suffer from bad visual quality or unaligned images because of the pose-variant feature maps and support only the reference-guided task.

3. Methods

The organized labels can be defined as a hierarchical structure. For a single image, its attribute for tag $i \in \{1, 2, ..., N\}$ can be defined as $j \in \{1, 2, ..., M_i\}$, where N is the number of tags and M_i is the number of attributes for tag i. However, the attribute j cannot represent the detailed manifestation for the tag i in images, which we called tag-relevant style $s_{i,j} \in \mathcal{S}_{i,j}$. Similarly, we denote the image which has the attribute j for tag i by $x_{i,j} \in \mathcal{X}_{i,j}$.

Our framework aims to generate, extract, and manipulate the tag-relevant style in the image. Each purpose corresponds to a specific module. There are two ways to get the style code for manipulation, as shown in Figure 5. For the latent-guided task, the style code is generated by the mapper module (**M**). Given a latent code $z \sim \mathcal{N}(0, 1)$ and an attribute j for tag i, **M** generates the style code $s_{i,j} = \mathbf{M}_{i,j}(z)$. For the reference-guided task, the style code is extracted by the extractor module (**F**). Given an image $x_{i,j}$ and a tag i, **F** learns to extract the style code $s_{i,j} = \mathbf{F}_i(x_{i,j})$. Then, the generated or extracted style is utilized to guide the manipulation. However, it is inefficient if we apply manipulation directly to the image when manipulating multiple tags. Instead, denote a source image by x,

we first convert it into its immediate feature by the encoder module (**E**), which is given by $e = \mathbf{E}(x)$. Then, to manipulate the feature, we introduce the translator module (**T**). Given a feature e and a tag-relevant style code $s_{i,j}$ for tag i, **T** learns to manipulate the specific tag of the feature by $\mathbf{T}_i(e, s_{i,j})$. In each translation, the feature can go through the translators multiple times. To get the translated image, we introduce the generator module (**G**) to convert the translated feature $\tilde{e}$ into the image. The translated image is given by $\tilde{x} = \mathbf{G}(\tilde{e})$. We also introduce the discriminator module (**D**) to determine whether an image, given the tag and attribute, is real or not. Notably, for modules which need the tag or attribute as input, we choose to use them to index the selection of specific layers of the modules rather than injecting them into the modules (*e.g.* using the individual module $\mathbf{M}_{i,j}(\cdot)$ rather than a single $\mathbf{M}(\cdot, i, j)$ with tag i and attribute j as inputs).

Formally, during the test, the source image x is firstly encoded to its immediate feature

$$e_0 = e = \mathbf{E}(x). \tag{1}$$

Second, to manipulate the source feature into target attributes $j_1, ..., j_l$ for multiple tags $i_1, ..., i_l$, respectively, where l is the number of manipulated tags, we input the feature into the specific translator one by one, for $k = 1, ..., l$,

$$e_k = \mathbf{T}_{i_k}(e_{k-1}, s_{i_k, j_k}), \tag{2}$$

where s_{i_k, j_k} can be either a latent-guided style $\mathbf{M}_{i_k, j_k}(z)$ or a reference-guided style $\mathbf{F}_{i_k}(x_{i_k, j_k})$. Finally, let $\tilde{e} = e_l$, the translated image is generated by

$$\tilde{x} = \mathbf{G}(\tilde{e}). \tag{3}$$

In particular, the test phases of the multi-style, multi-attribute, and multi-tag tasks are shown in Figure 6.

3.1. Training Phases

To independently optimize the modules for different tags and attributes, we randomly sample a tag i, a source attribute j, and a target attribute $\tilde{j}$ in each iteration. As shown in Figure 4, given a source image $x_{i,j} \in \mathcal{X}_{i,j}$, the training phases include:

Non-translation path. We get the first reconstruction image $x'_{i,j} = \mathbf{G}(\mathbf{E}(x_{i,j}))$ in this path.

Self-translation path. We get the second reconstruction image $x''_{i,j} = \mathbf{G}(\mathbf{T}(\mathbf{E}(x_{i,j}), s_{i,j}))$, where $s_{i,j} = \mathbf{F}_i(x_{i,j})$ is the extracted tag-relevant style code of the source image.

Cycle-translation path. In this path, we firstly generate the target tag-relevant style code $s_{i,\tilde{j}} = \mathbf{M}_{i,\tilde{j}}(z)$. Secondly, we render the generated style code $s_{i,\tilde{j}}$ into the feature of the source image $x_{i,j}$ and get the translated image $x_{i,\tilde{j}} = \mathbf{G}(\mathbf{T}(\mathbf{E}(x_{i,j}), s_{i,\tilde{j}}))$. Finally, the feature of the translated image $x_{i,\tilde{j}}$ and the original extracted style code $s_{i,j}$ are inputted into the translator and we get the third reconstruction image $x'''_{i,j} = \mathbf{G}(\mathbf{T}(\mathbf{E}(x_{i,\tilde{j}}), s_{i,j}))$.
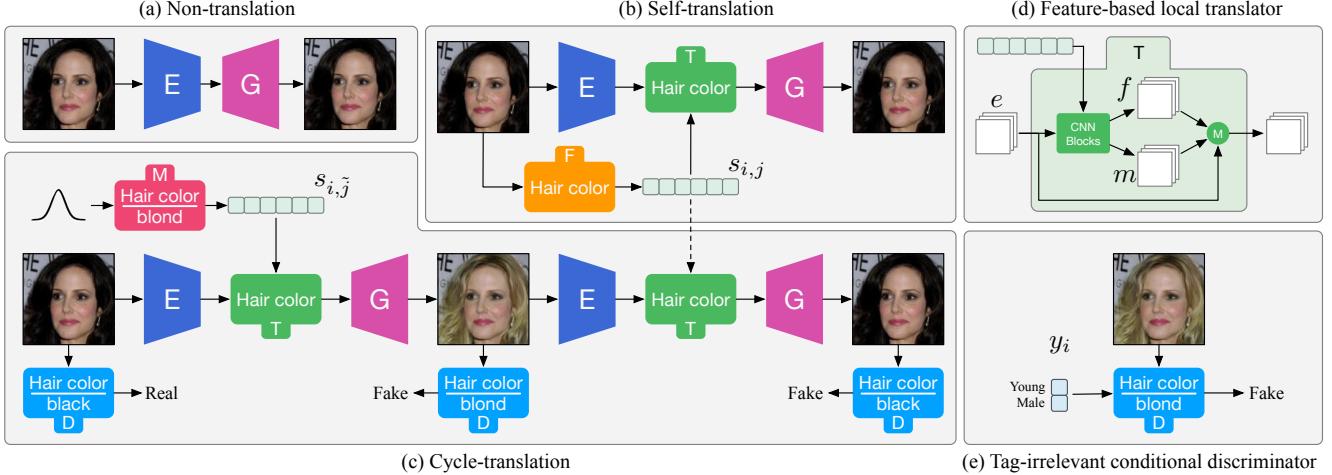
Figure 4: The training phases of our method include **(a)** a non-translation path, **(b)** a self-translation path, and **(c)** a cycle-translation path. We further avoid unnecessary manipulations by two essential architectural improvements: **(d)** a feature-based local translator and **(e)** a tag-irrelevant conditional discriminator.

3.2. Training Objectives

Adversarial objective. The adversarial objective of our method encourages realistic manipulations for both generated and extracted styles, which is defined as

$$
\begin{aligned}
\mathcal{L}_{adv} = & \; 2\mathbb{E}_{i,j,x}[\log(\mathbf{D}_{i,j}(x_{i,j}))] \\
& + \mathbb{E}_{i,j,x,\tilde{j},z}[\log(1 - \mathbf{D}_{i,\tilde{j}}(x_{i,\tilde{j}}))] \\
& + \mathbb{E}_{i,j,x,\tilde{j},z}[\log(1 - \mathbf{D}_{i,j}(\tilde{x}'''_{i,j}))],
\end{aligned}
\tag{4}
$$

where $x_{i,\tilde{j}}$ is the translated image using the style code generated by the mapper $\mathbf{M}$, and $\tilde{x}'''_{i,j}$ is the cycle-translated image using the style code extracted by the extractor $\mathbf{F}$. This objective not only encourages the mapper to accurately map the tag-specific attribute information into the generated tag-relevant style code, but also forces the extractor to extract the tag-specific attribute information from the image.

Reconstruction objective. All final outputs of the non-translation, self-translation, and cycle-translation paths are the reconstruction images of the source image. Thus, we apply a reconstruction objective to make the reconstruction images equal to the source image, which is

$$
\begin{aligned}
\mathcal{L}_{rec} = & \; \mathbb{E}_{i,j,x}[\|x'_{i,j} - x_{i,j}\|_1] \\
& + \mathbb{E}_{i,j,x}[\|x''_{i,j} - x_{i,j}\|_1] \\
& + \mathbb{E}_{i,j,x,\tilde{j},z}[\|x'''_{i,j} - x_{i,j}\|_1].
\end{aligned}
\tag{5}
$$

Specifically, the first two terms encourage the consistency between the features whether through the translator $\mathbf{T}$ or not. They are significant for our framework to manipulate multiple tags during test, which is not directly involved during training. The importance of these two terms has been proved in ModularGAN [54].

The final term utilizes the cycle consistency to encourage the extracted tag-relevant style to be accurate, which needs models to extract the detailed manifestation of the source image and render it to manipulate the translated image, so that the cycle-translated image can be equal to the source one (*e.g.*, the model needs to extract the specific style of glasses so that it can translate the non-glasses translated image back into the source image).

Style objective. The extracted style code of the translated image is supposed to be equal to the generated style code. So we introduce the style objective

$$
\mathcal{L}_{sty} = \mathbb{E}_{i,j,x,\tilde{j},z}[\|\mathbf{F}_i(x_{i,\tilde{j}}) - s_{i,\tilde{j}}\|_1],
\tag{6}
$$

which encourages the consistency between the generated and extracted styles [18, 23, 45, 24, 51, 6]. On the one hand, it encourages the mapper $\mathbf{M}$ to generate the accurate tag-relevant style code, which can be equally extracted by the extractor $\mathbf{F}$ as well. On the other hand, it also encourages the translator $\mathbf{T}$ to fully utilize the style code and forces both generated and extracted styles to be camera-ready.

Full objective. Finally, the optimization of the full objective function can be written as

$$
\min_{\mathbf{E},\mathbf{G},\mathbf{T},\mathbf{F},\mathbf{M}} \max_{\mathbf{D}} \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{sty}\mathcal{L}_{sty},
\tag{7}
$$

where λ_{rec} and λ_{sty} are hyper-parameters that control the relative importance of reconstruction and style objectives compared to the adversarial objective, respectively. The full objective guarantees the style codes to catch the clear manifestations of different tags with unnecessary manipulations. We introduce two architectural improvements to avoid unnecessary manipulations and make the style codes further disentangled without extra objectives [37, 2, 16].

4

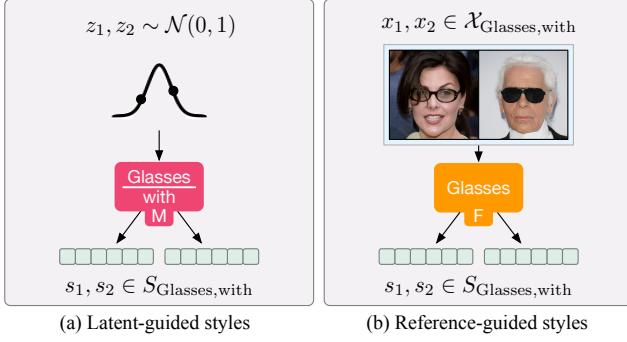(a) Latent-guided styles   (b) Reference-guided styles

Figure 5: During test, the tag-relevant styles can be either **(a)** latent-guided (*i.e.* generated by random latent codes) or **(b)** reference-guided (*i.e.* extracted from reference images).

3.3. Feature-based Local Translator

The translator in our framework affects the features rather than the images [31, 38, 7, 35]. To utilize the tag locality, we introduce a feature-based local translator. Denote the original feature by e, the translator outputs two folds of features m and f with the same size (*i.e.* height, width, and channel) of e. Then the translated feature is given by

$$\sigma(m) \cdot e + (1 - \sigma(m)) \cdot f, \qquad (8)$$

where $\sigma(\cdot)$ is the sigmoid function and $\sigma(m)$ is an attention mask. The attention mask in our translator is both spatial-wise and channel-wise. This design can avoid global manipulations like background and illumination during translations, with negligible additional calculation and no regularization objective.

3.4. Tag-irrelevant Conditional Discriminator

For different attributes, imbalanced phenomenons of implicit conditions are widespread in the real-world datasets. In CelebA-HQ, there are 83.3% male and 65.7% aged in images with attribute 'with' for tag 'Glasses', while the percentages decrease to 36.0% and 20.0% correspondingly in images with attribute 'without' for tag 'Glasses'. The discriminator will force the translations to manipulate these implicit conditions. We address this problem by injecting the tag-irrelevant conditions (*e.g.* labels 'Male' and 'Young') into the discriminator. Denote the tag-irrelevant conditions of the original image $x_{i,j}$ for tag i by y_i, we replace Equation 4 by

$$\begin{aligned}
\mathcal{L}'_{adv} = {} & 2\mathbb{E}_{i,j,x}[\log(\mathbf{D}_{i,j}(x_{i,j}, y_i))] \\
& + \mathbb{E}_{i,j,x,\tilde{j},z}[\log(1 - \mathbf{D}_{i,\tilde{j}}(x_{i,\tilde{j}}, y_i))] \\
& + \mathbb{E}_{i,j,x,\tilde{j},z}[\log(1 - \mathbf{D}_{i,j}(x'''_{i,j}, y_i))].
\end{aligned} \qquad (9)$$

Consequently, the discriminator will notice the imbalanced phenomenons and encourage the translations to not manipulate the tag-irrelevant implicit conditions.

(a) Multi-style task

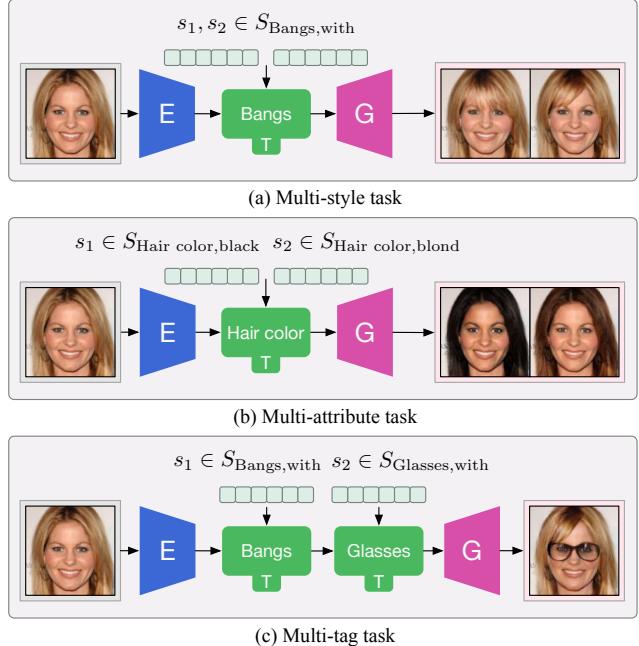(b) Multi-attribute task

(c) Multi-tag task

Figure 6: Test phases for **(a)** the multi-style task, **(b)** the multi-attribute task and **(c)** the multi-tag task. Note that the multi-tag task is not involved during training but achieved by the re-designed training phases and objectives indirectly.

4. Experiments

In this section, we conduct a set of experiments to prove the effectiveness of our method. All experiments are conducted using unseen images during training.

Dataset. We choose CelebA-HQ [19] as our dataset, which contains 30,000 facial images with label annotations such as hair color, gender, and presence of glasses. CelebA-HQ is more challenging than the original CelebA [28]. The original labels 'With_Bangs', 'With_Glasses', 'Blond_Hair', 'Black_Hair', and 'Brown_Hair' in CelebA-HQ are organized into three tags 'Hair color' (with attributes 'blond', 'black' and 'brown'), 'Bangs' (with attributes 'with' and 'without'), 'Glasses' (with attributes 'with' and 'without') for experiments. For tag-irrelevant conditions, we choose two major labels, 'Male' and 'Young' to avoid the manipulation of gender and age. We split the 30000 images in CelebA-HQ into 3000 images as the test set and 27000 as the training set.

Baselines. We use SDIT [45] (with shared style), Star-GANv2 [6] (with mixed style), and ELEGANT [49] (with label-specific style) as our baselines. All models of baselines are trained using the implementations provided by their authors. We train three independent models for Star-GANv2 for each tag to avoid the exponential number of label domains. For other implementation details, please refer to our supplemental material.

Figure 7: Qualitative results of the multi-style task, including **(a)** the latent-guided task and **(b)** the reference-guided task. Please zoom in for more details.

4.1. Multi-style Task

In this section, we evaluate our method on translating a tag of images into another attribute with diverse outputs. We respectively manipulate tag 'Bangs' to attribute 'with' and tag 'Glasses' to attribute 'with' as our examples from two perspectives: latent-guided task and reference-guided task.

Latent-guided task. Figure 7(a) provides a qualitative comparison of the competing methods. We cannot provide

the results of ELEGANT because it only has the reference-guided capacity. The visual quality and diversity of SDIT are limited. StarGANv2 manipulates tag-irrelevant details (*e.g.* changing the hair color when manipulating tag 'Bangs' and translating the young female into an aged male when manipulating tag 'Glasses'). Our method manipulates accurate tag-relevant styles for both tags, with high visual quality, satisfying diversity of the manipulated attributes, and maintaining tag-irrelevant details.

Method	Realism(FID)			Diversity(User Study%)			Disentanglement(FID)		
	L	R	G	L	R	G	L	R	G
SDIT [45]	33.73	33.12	00.61	09.94	08.25	01.69	80.25	79.72	00.53
StarGANv2 [6]	26.04	25.49	00.55	29.18	26.63	02.55	90.08	78.03	12.05
ELEGANT [49]	-	22.96	-	-	38.80	-	-	75.03	-
Ours	**21.37**	**21.49**	**00.12**	**50.00**	**48.97**	**01.03**	**71.85**	**71.48**	**00.37**

Table 1: Quantitative results of the multi-style task. (**L**: latent-guided; **R**: reference-guided; **G**: gap between L and R.)

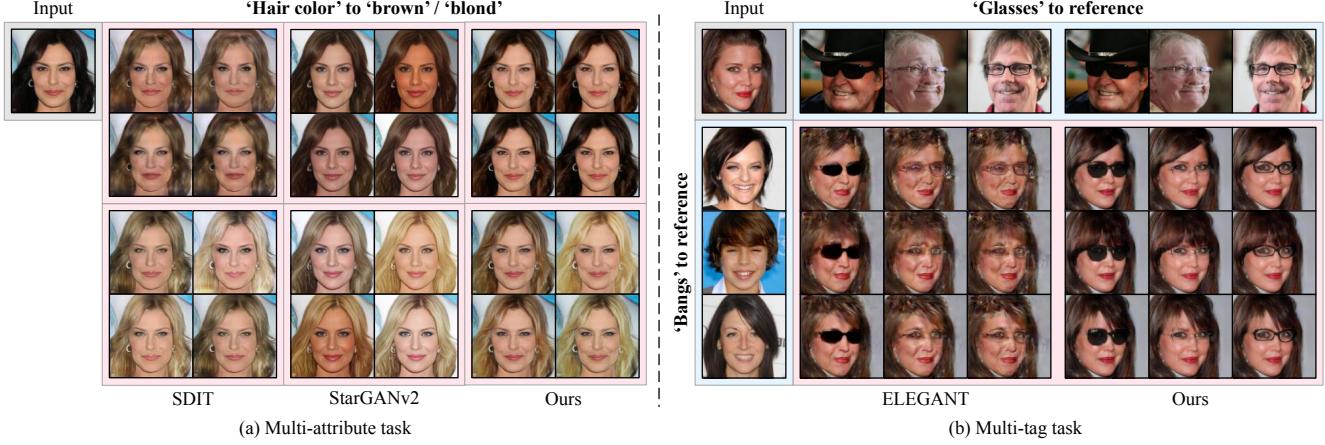(a) Multi-attribute task   (b) Multi-tag task

Figure 8: Qualitative results of **(a)** the multi-attribute task and **(b)** the multi-tag task.

Reference-guided task. Figure 7(b) provides a qualitative comparison of the competing methods. In this task, the styles are extracted from the reference images. SDIT cannot effectively extract the styles. The expression of attributes comes from the change of the labels. StarGANv2 extracts and manipulates many distinctive styles (*e.g.* hair color, background, and facial identity). ELEGANT succeeds in extracting and manipulating the accurate styles, but it generates visible artifacts in the last row of the left results. Another limitation is that the transfer of unaligned images is ineffective, because it uses pose-variant feature maps to represent the styles, while we use pose-invariant global style codes. Our method transfers the styles for both tags accurately.

We also perform three quantitative comparisons between our method and baselines, including:

Realism. To qualitatively evaluate realism, we calculate the Frechét inception distance (FID) [14]. For each test image without bangs, we translate it into images with bangs using 5 style codes, which are generated by randomly sampled latent codes for the latent-guided task or extracted from random samples from images with bangs for the reference-guided task. We then calculate the average FID between the translated images and real images with bangs. Table 1 shows the quantitative comparison of the competing methods. Our method outperforms all baselines in terms of realism.

Diversity. To qualitatively evaluate diversity, the widely used metric is the learned perceptual image patch similarity (LPIPS) [53]. However, it encourages the models to manipulate the image as much as possible, which is not desired in practical use. Hence, we choose to use user study for quantitative comparison of diversity. For each task, the percentage is given by asking users which diversity is more preferred, compared to the latent-guided results of our method. The results in Table 1 show that the baselines cannot obtain a vote bigger than our method (*i.e.* 50%), which means that users prefer the diversity of our method more.

Disentanglement. To qualitatively evaluate the disentanglement of the tag-relevant styles, we translate the young male images without bangs into 5 images with bangs and use the translated images and real young male images with bangs to calculate the average FID distance. This metric challenges the methods to only transfer the tag-relevant styles and maintain the tag-irrelevant details. If the model transfers tag-irrelevant details, such as changing the gender of the images, the FID between the translated images (some female images) and the real images (all male images) will increase. The results show that the disentanglement of our method is superior to the baselines.

Moreover, we also report the capacity gap, which is the L1-norm between the values of the latent-guided task and reference-guided task. The results show that our method gets the most balanced performance.
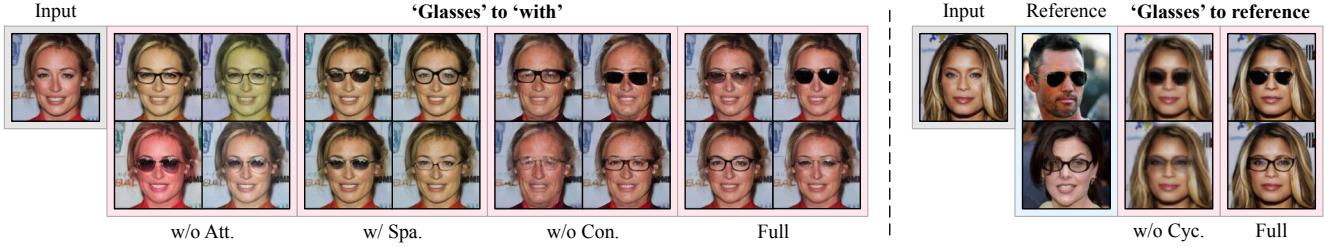
Figure 9: Qualitative results of the ablation study. Results to the left of the dotted line are latent-guided while results to the right of the dotted line are reference-guided.

4.2. Multi-label Task

Multi-label task in previous methods is divided into two sub-tasks in our method: multi-attribute task and multi-tag task. The former translates images into multiple possible attributes for one tag only. The latter manipulates multiple tags of images.

Multi-attribute task.. We translate the tag 'Hair color' of images from 'black' into two attributes 'brown' and 'blond' for example. As shown in Figure 8(a), SDIT gets worse visual quality than other methods, and StarGANv2 manipulates tag-irrelevant details (*e.g.* skin color and illumination). Our method accurately manipulates the disentangled tag-relevant styles, which satisfy both attributes for 'Hair color'. The visual quality and diversity of our method are also desirable.

Multi-tag task. An ideal method is supposed to independently manipulate each tag, which cannot be performed by SDIT and StarGAN due to their shared or mixed style code. We manipulate two tags, 'Bangs' and 'Glasses', simultaneously. As shown in Figure 8(b), simultaneous translations of ELEGANT for multiple tags lead to interference (*e.g.* changing the style of bangs while changing the style of glasses). Furthermore, the results of our method demonstrate the disentanglement of our tag-relevant style codes.

4.3. Ablation Study

In this experiment, we evaluate several ablations of our method with different settings to investigate the effect of each additional component in our framework. The qualitative and quantitative results are shown in Figure 9 and Table 2, respectively. The first ablation removes the feature-based local translator (w/o Att.). The result shows that it cannot focus on the tag-relevant style (*e.g.* changing the color and illumination of the images). The second ablation uses the translator with only spatial-wise attention (w/ Spa.). Although it can obtain tag-relevant diversity, the color distribution of the input image changes. The third ablation removes the tag-irrelevant conditions for the discriminator (w/o Con.). As expected, it is observed to change the gender and age of the input, but it achieves the best

Setting	Realism(FID)			Disentanglement(FID)		
	L	R	G	L	R	G
w/o Att.	22.30	21.83	00.47	<u>73.91</u>	<u>71.50</u>	02.41
w/ Spa.	23.94	23.94	**00.00**	75.17	75.17	**00.00**
w/o Con.	**20.63**	**20.66**	<u>00.03</u>	74.91	73.61	01.30
w/o Cyc.	22.31	28.39	06.08	74.18	80.34	06.16
Full	<u>21.37</u>	<u>21.49</u>	00.12	**71.85**	**71.48**	<u>00.37</u>

Table 2: Quantitative results of the ablation study. (**Bold**: best; <u>Underline</u>: runner-up.)

performance in terms of the realism due to the imbalanced phenomenon in the test images. The last ablation removes the adversarial objective of the cycle-translation output (w/o Cyc.). The results of this ablation in the latent-guided task are similar to the full model, but in the reference-guided task, it is observed to output fuzzy results. The realism and disentanglement metrics of it in the reference-guided task are the worst among all ablations. We suspect that the optimization with only the reconstruction objective of cycle-translation output is unclear for the framework, especially for small objects like glasses. Our full model gets the best performance in qualitative results and sacrifices a bit of realism for disentanglement in quantitative results, which proves the significance of each component in our framework.

5. Conclusion

In this paper, we propose the Hierarchical Style Disentanglement for image-to-image translation, with scalability and controllable diversity. The key idea is to organize the labels into a hierarchical tree structure, which consists of independent tags, exclusive attributes, and unsupervised but disentangled styles. Extensive qualitative and quantitative experiments prove the effectiveness of our method. Moreover, it is easy to adapt our method to joint training of multiple datasets [5], continuous learning of new tags [54], semi-supervised learning of semi-labeled dataset [24], and few-shot learning of tags with many attributes [27]. We believe that our method is more suitable for practical use than the state-of-the-art image-to-image translation methods.

6. Acknowledge

G. Module Architecture

The architectural details of HiSD are shown in Figure 10. The mapper module (**M**) consists of an MLP. The tag i and attribute j are used to index before the first layer and middle layer, respectively. The extractor module (**F**) consists of five downsampling blocks, of which inherit pre-activation residual units (ResBlock) [12]. The tag i is used to index before the last layer. The encoder module (**E**) consists of two downsampling blocks while the generator module (**G**) consists of two upsampling blocks. We use the Instance Normalization (IN) [44] in these two shared modules. The translator module (**T**) consists of eight immediate blocks with Adaptive Instance Normalization (AdaIN) [17] and decreased channel dimension. The tag-relevant style code is injected into all AdaIN layers, providing scaling and shifting vectors through a linear layer. The tag i is used to index before the first layer. The discriminator module (**D**) uses the same architecture as **F** but uses both the tag i and the attribute j to index before the last layer. For all residual units, We use the Leaky ReLU (LReLU) [29] as the activation function. The Average Pooling and Nearest Neighbor Upsampling are used to resample the feature maps.

H. Implementation Details

The batch size is 8 and the model is trained for 200K iterations. The images of CelebA-HQ are resized to 128×128. The training time is around 40 hours on a single GTX 1080Ti GPU with our implementation in PyTorch [36]. To stabilize the training, we adopt the hinge version of adversarial loss [34] with R1-regularization [32] using $\gamma = 1$. We use the Adam [21] optimizer with $\beta_1 = 0$ and $\beta_2 = 0.99$. The learning rate is 0.0001 except for the mapper, of which the learning rate is 0.000001 [20]. We use the historical average version [19] of the intermediate modules for test where the update weight is 0.001. We initialize the weights of all modules using He initialization [11] and set all biases to zero, except for the biases associated with the scaling vectors of AdaIN that are set to one. For hyper-parameters, we easily set $\lambda_{rec} = 1$ and $\lambda_{sty} = 1$ in all experiments.

I. Comparison without Cherry-picking

We provide some additional qualitative results without cherry-picking of baselines and our method for the latent-guided multi-style task in Figures 11, 12, and 13. The results are completely random without manual selection. Besides the limitations we mentioned in the paper, the baselines (*i.e.* SDIT and StarGANv2) are observed to suffer from mode-collapse. Furthermore, the satisfying diversity of our generated results demonstrates the effectiveness of HiSD.

J. Interpolation of Tag-relevant Styles

We show the interpolation results by interpolating between the extracted tag-relevant style codes from two different reference images (with the same attribute or not) in Figures 14, 15, and 16. The interpolation is smooth, which implies that the space of each tag-relevant style is continuous. The continuous tag-relevant style space allows the translations to manipulate images with a novel tag-relevant style which is not seen by the framework during training.

K. Visualization of Tag-relevant Styles

We further explore the style space learned by the extractor by using t-SNE to visualize the extracted tag-relevant styles in a two-dimensional space. As shown in Figure 17, for all tags, images with the same tag-specific attribute are grouped together in the style space. Notably, the attribute is not inputted into the extractor in our method. For each tag, there is a main direction and various secondary directions between different attributes. InterFaceGAN [42] takes advantage of the main direction to manipulate attributes with unsupervised GANs [20, 19]. However, it cannot guarantee the disentanglement, especially for unnecessary global and identity manipulations. The images at the edge of the style space are always the most obvious examples for a specific attribute, while the images in the middle space are always confusing ones. More interestingly, for tag 'Hair color', the tag-relevant styles of images with attribute 'brown' are extracted to be a middle state between 'blond' and 'black'.
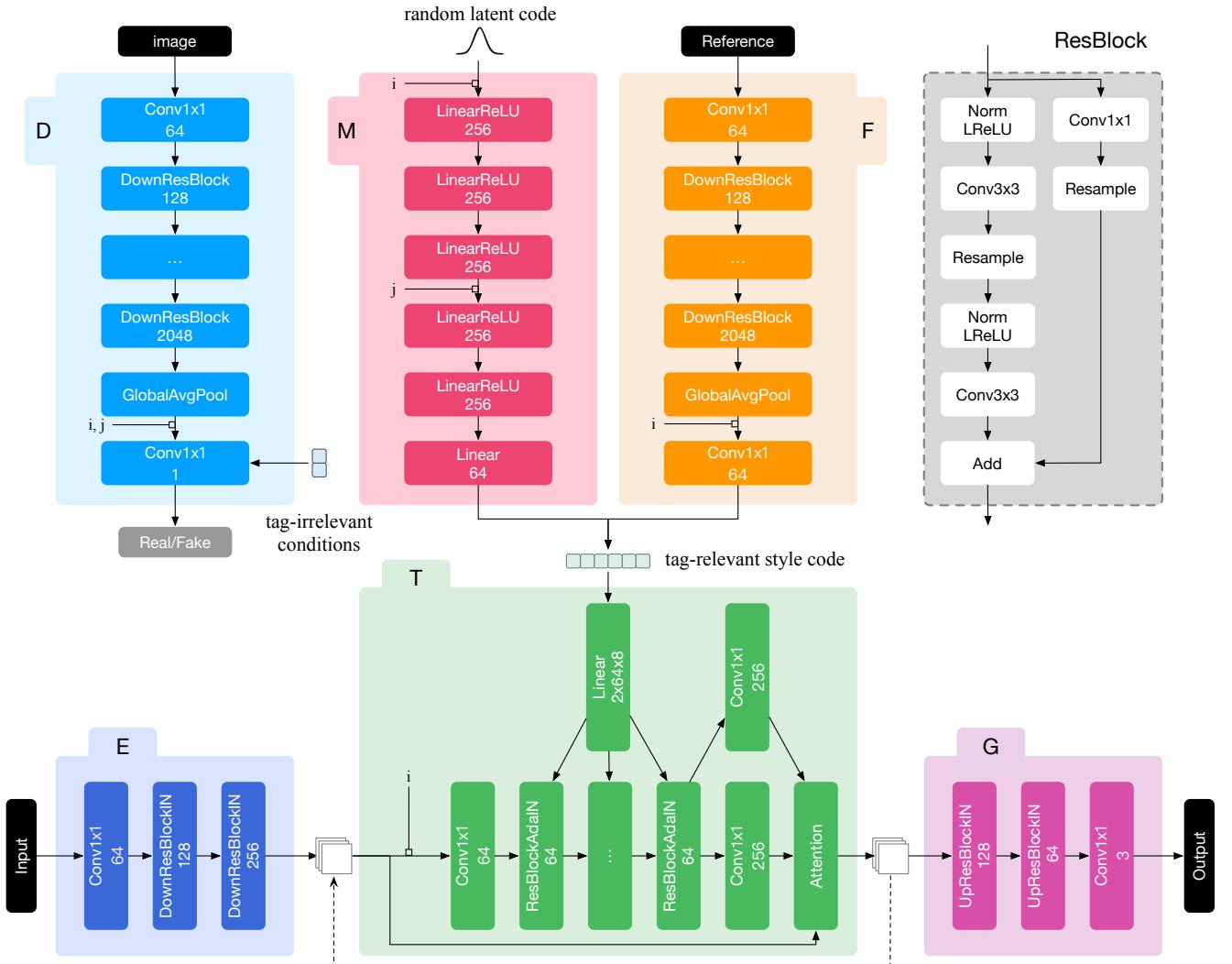
Figure 10: Architectural details of HiSD. To manipulate the input image, we first encode the image into its feature by **E**. Then, the feature is manipulated by a single or multiple T. The manipulation is guided by the tag-relevant style code which can be either generated by **M** or extracted by **F**. Finally, the output image is generated by **G**. **D** is used to determine whether a image, given tag and attribute, is real or not. The details of ResBlocks (*i.e.* DownResBlock, DownResBlockIN, ResBlockAdaIN, UpResBlockIN) are shown at the upper right corner.

Input

'Bangs' to 'with'

'Glasses' to 'with'
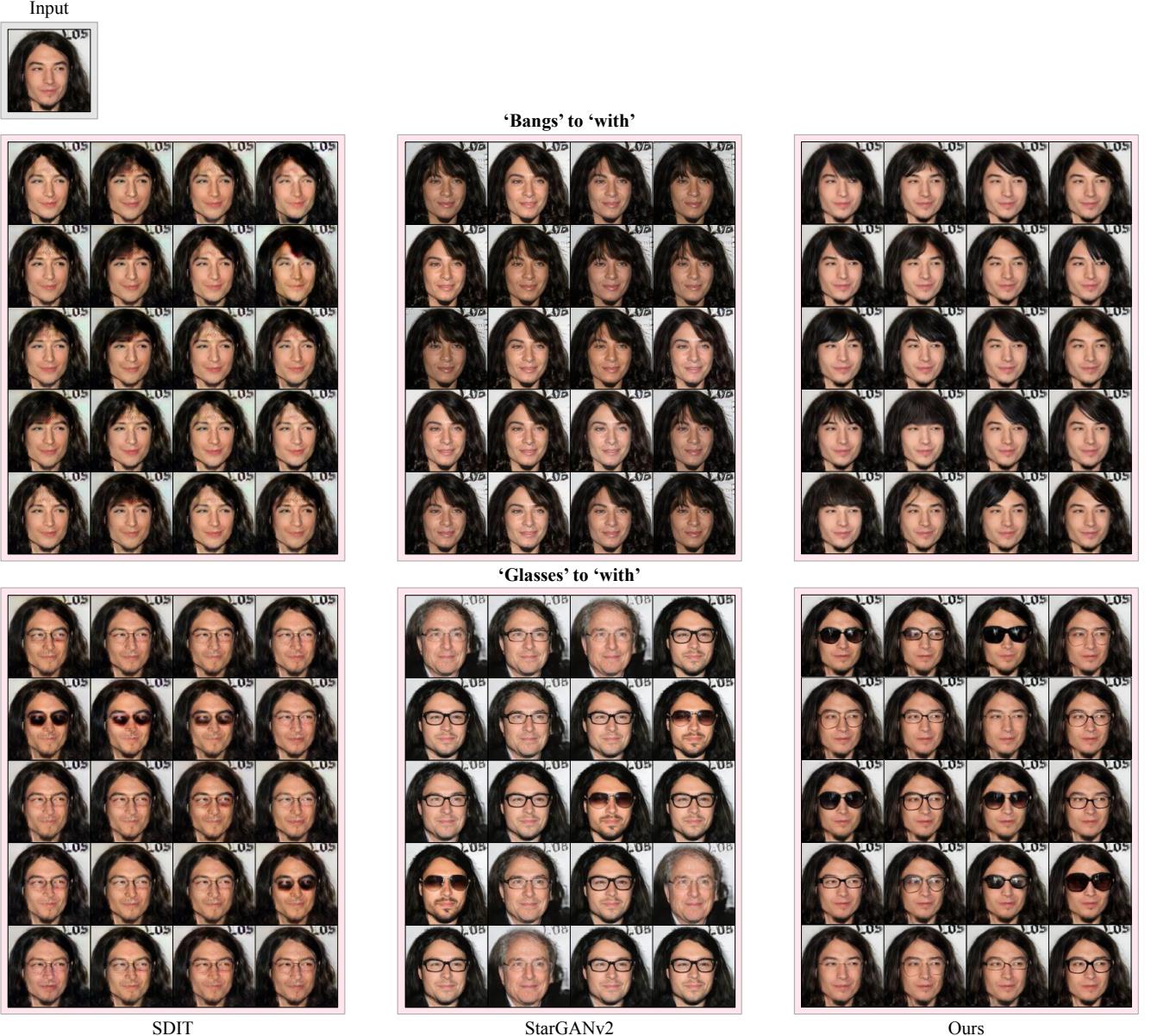

| SDIT | StarGANv2 | Ours |

Figure 11: Additional qualitative results without cherry-picking of the baselines and our method for the latent-guided multi-style task. We respectively manipulate the input image to attribute 'with' for tag 'Bangs' and 'Glasses' by using 20 random latent codes, which are drawn from Gaussian distribution, to generate diverse outputs.

Input

'Bangs' to 'with'

'Glasses' to 'with'

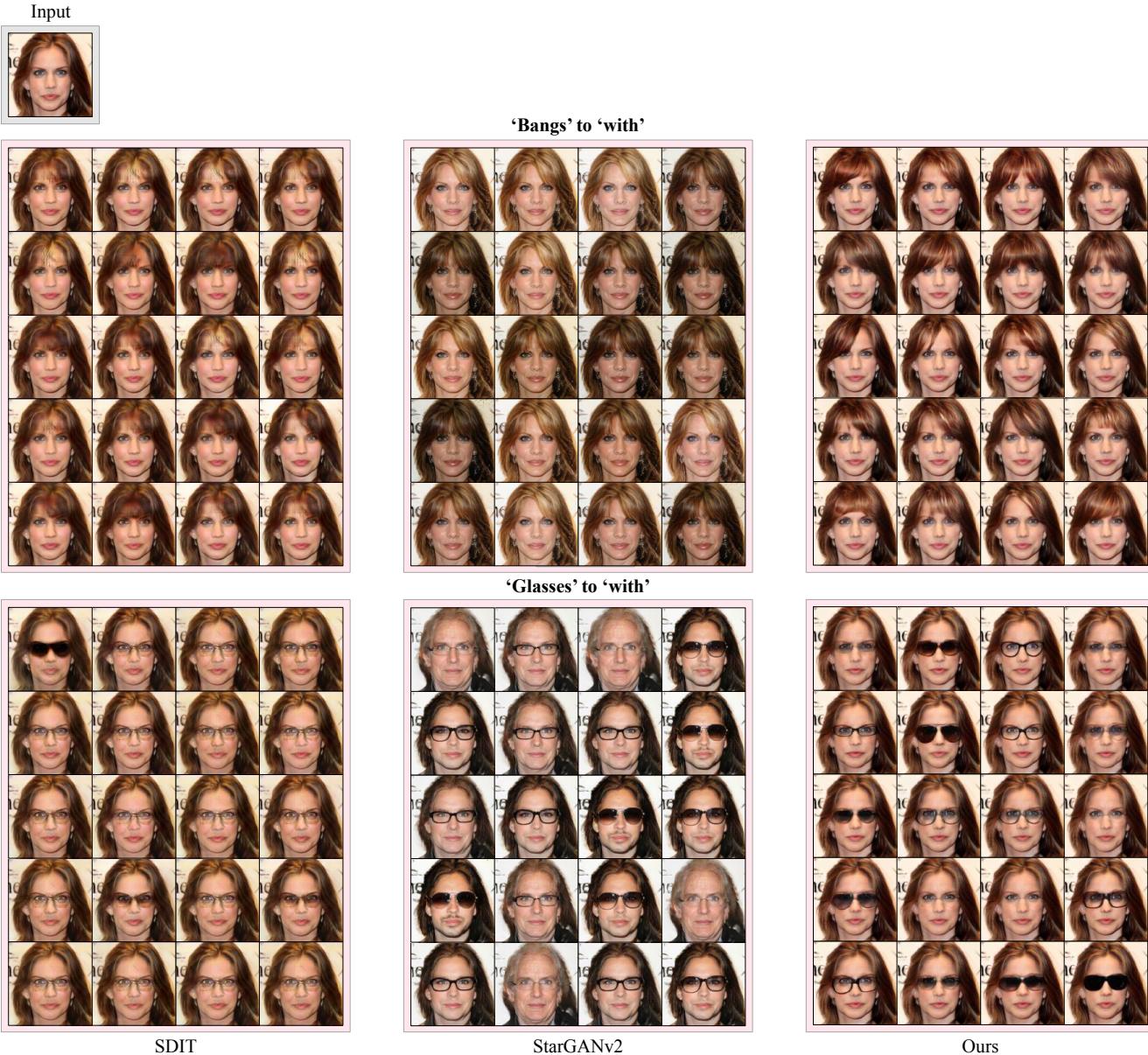| SDIT | StarGANv2 | Ours |

Figure 12: More additional qualitative results without cherry-picking for the latent-guided multi-style task.

Input

'Bangs' to 'with'

'Glasses' to 'with'

SDIT   StarGANv2   Ours

Figure 13: More additional qualitative results without cherry-picking for the latent-guided multi-style task.
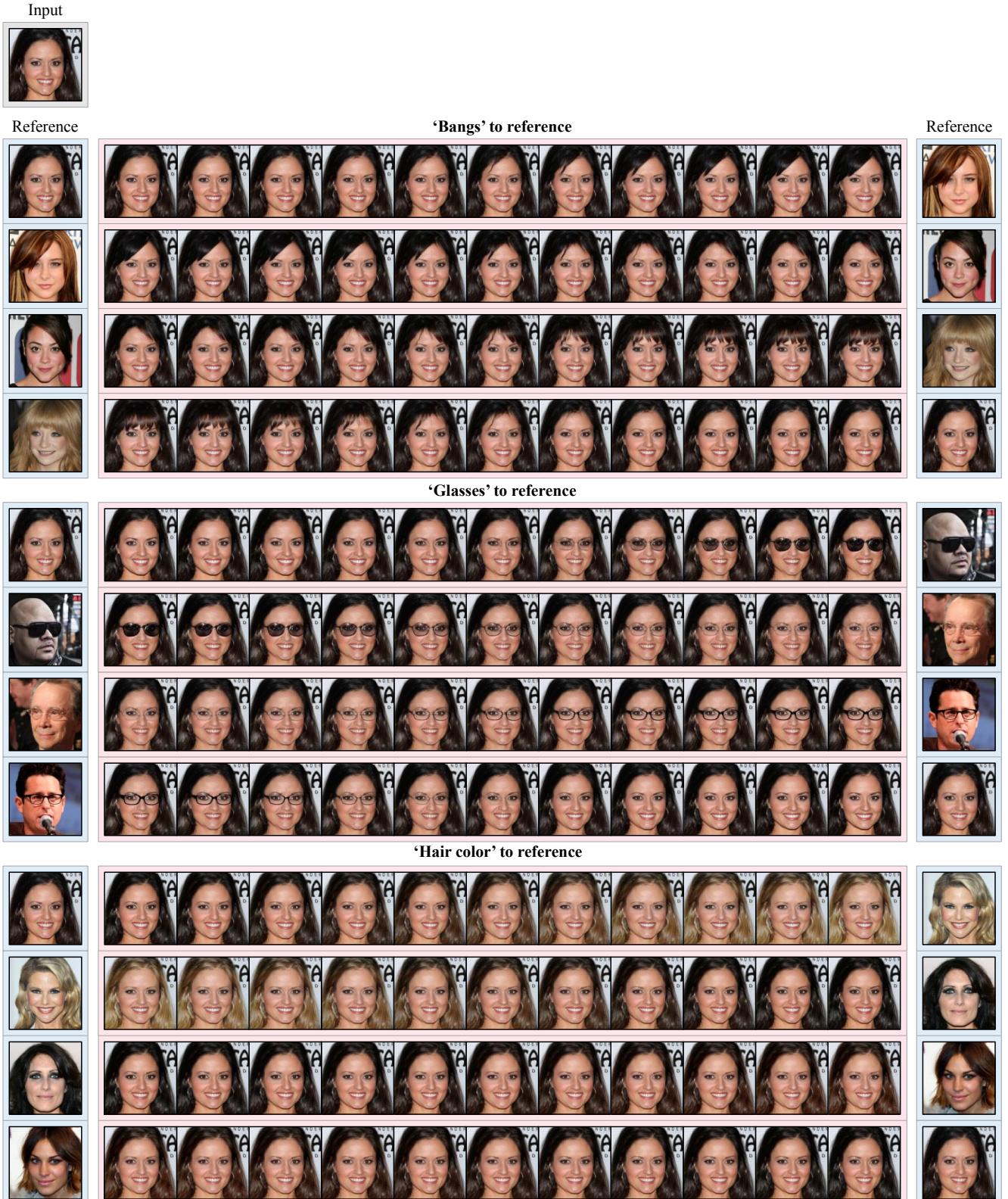
Input

Reference   'Bangs' to reference   Reference

'Glasses' to reference

'Hair color' to reference

Figure 14: Interpolation results between two extracted tag-relevant styles for different tags. We use the linear interpolation between the style codes extracted from two different reference images to observe the continuous manipulations of the output images.

Input

Reference 'Bangs' to reference Reference

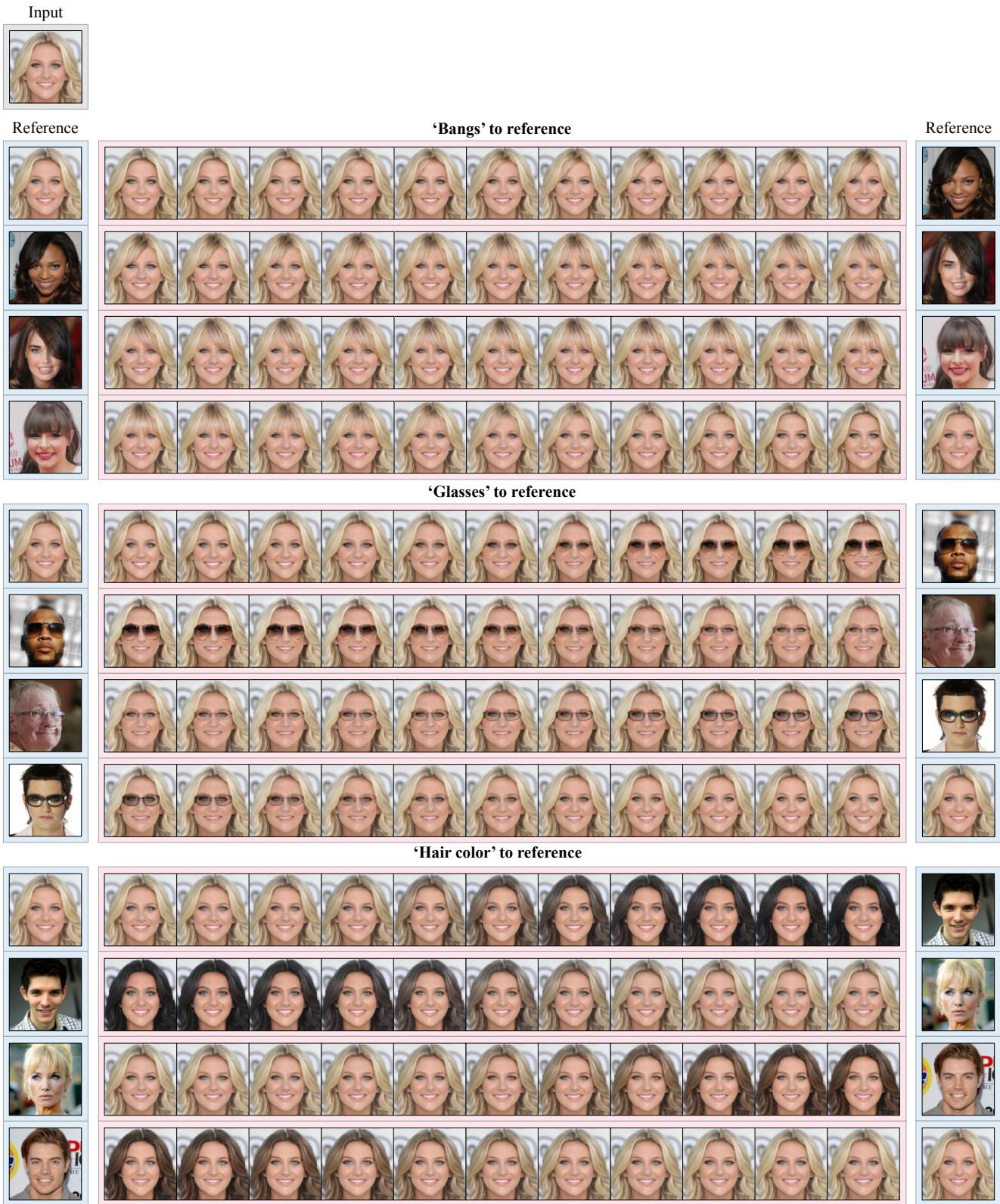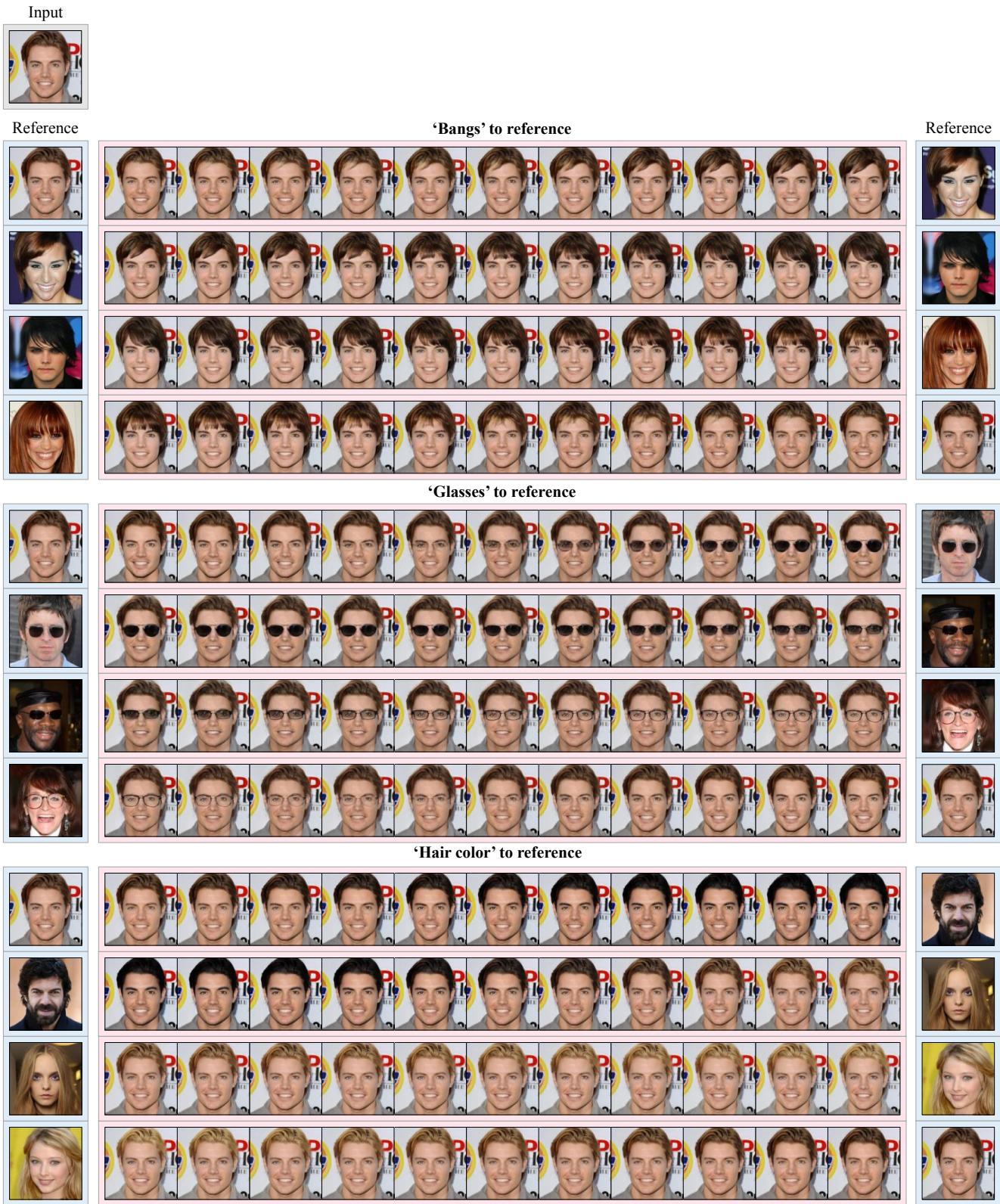'Glasses' to reference

'Hair color' to reference

Figure 15: More interpolation results between two extracted tag-relevant styles for different tags.

Input

Reference   **'Bangs' to reference**   Reference

'Glasses' to reference
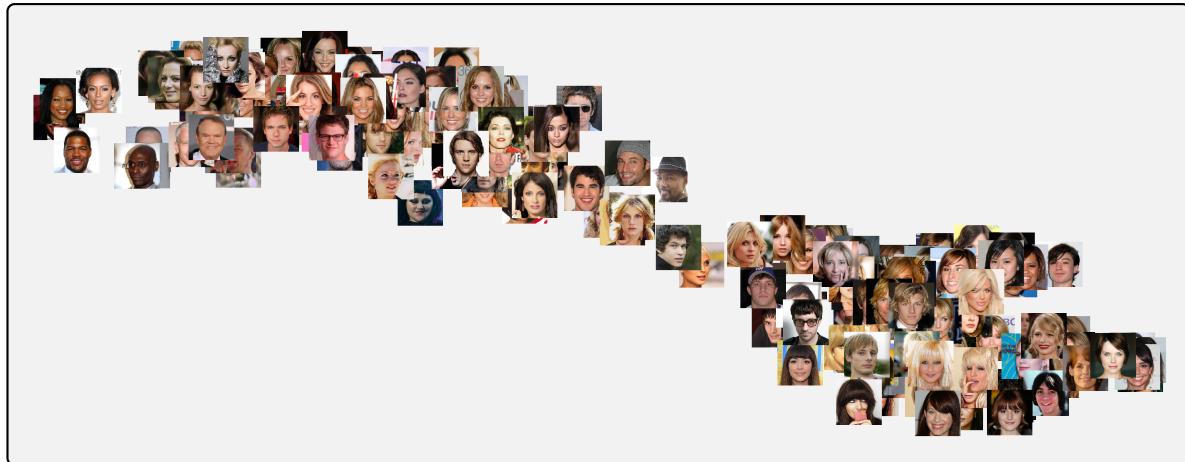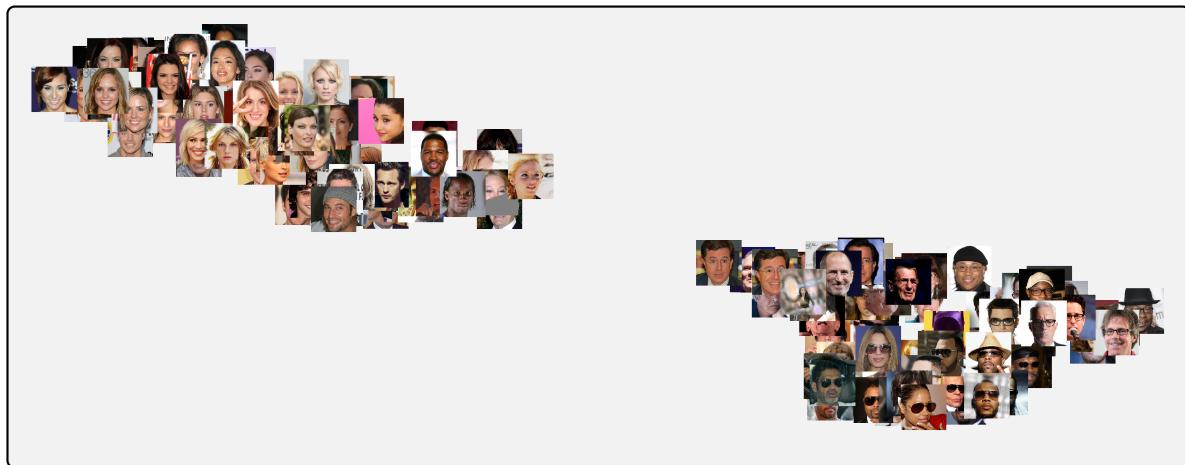

'Hair color' to reference

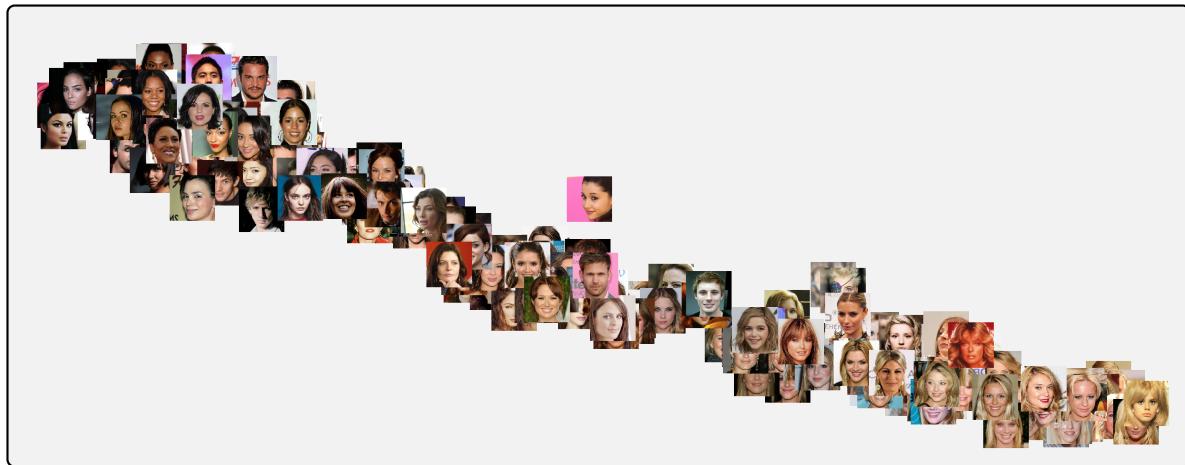Figure 16: More interpolation results between two extracted tag-relevant styles for different tags.

(a) 'Bangs'

(b) 'Glasses'

(c) 'Hair color'

Figure 17: 2-D representation of the extracted tag-relevant styles from 180 images using t-SNE for different tags. Please zoom-in for details.

References

[1] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented Cycle-GAN: Learning Many-to-Many Mappings from Unpaired Data. In *ICML*, 2018. 1, 3

[2] Sagie Benaim, Michael Khaitov, Tomer Galanti, and Lior Wolf. Domain intersection and domain difference. In *ICCV*, 2019. 3, 4

[3] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 3

[4] Ying Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *CVPR*, 2019. 3

[5] 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 *CVPR*, 2018. 1, 3, 8

[6] 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 *CVPR*, 2018. 1, 3, 4, 5, 7

[7] Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Chinnam. Spa-gan: Spatial attention gan for image-to-image translation. *TMM*, 2020. 5

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017. 3

[10] Jingtao Guo, Zhenzhen Qian, Zuowei Zhou, and Yi Liu. MulGAN: Facial Attribute Editing by Exemplar. *arXiv preprint arXiv:1912.12396*, 2019. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *CVPR*, 2015. 9

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 9

[13] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. AttGAN: Facial Attribute Editing by Only Changing What You Want. *TIP*, 2019. 1, 3

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 7

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 3

[16] Jie Hu, Rongrong Ji, ShengChuan Zhang, Xiaoshuai Sun, Qixiang Ye, Chia-Wen Lin, and Qi Tian. Information competing process for learning diversified representations. In *NeurIPS*, 2019. 4

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 9

[18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 1, 3, 4

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 5, 9

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 9

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 9

[22] 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 *CVPR*, 2017. 3

[23] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 1, 3, 4

[24] Xinyang Li, Jie Hu, Shengchuan Zhang, Xiaopeng Hong, Qixiang Ye, Chenglin Wu, and Rongrong Ji. Attribute guided unpaired image-to-image translation with semi-supervised learning. *arXiv preprint arXiv:1904.12428*, 2019. 1, 3, 4, 8

[25] 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 *CVPR*, 2019. 1, 3

[26] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 1

[27] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019. 3, 8

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[29] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 9

[30] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 3

[31] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *NeurIPS*, 2018. 5

[32] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018. 3, 9

[33] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[34] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 3, 9

18

[35] Ron Mokady, Sagie Benaim, Lior Wolf, and Amit Bermano. Mask based unsupervised content transfer. 2019. 5

[36] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 9

[37] Ori Press, Tomer Galanti, Sagie Benaim, and Lior Wolf. Emerging disentanglement in auto-encoder based unsupervised image content transfer. In *ICLR*, 2018. 3, 4

[38] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, 2018. 5

[39] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 3

[40] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. *ICML*, 2017. 3

[41] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation. In *ICCVW*, 2019. 1, 3

[42] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *TPAMI*, 2020. 9

[43] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. 2019. 1

[44] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 9

[45] Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, and Luis Herranz. Sdit: Scalable and diverse cross-domain image translation. In *ACMMM*, 2019. 1, 3, 4, 5, 7

[46] Bingzhe Wu, Haodong Duan, Zhichao Liu, and Guangyu Sun. Srpgan: Perceptual generative adversarial network for single image super resolution. In *CVPR*, 2017. 3

[47] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *ICCV*, 2019. 1, 3

[48] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Dna-gan: Learning disentangled representations from multi-attribute images. *arXiv preprint arXiv:1711.05415*, 2017. 3

[49] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *ECCV*, 2018. 3, 5, 7

[50] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 1

[51] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *NeurIPS*, 2019. 1, 3, 4

[52] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 3

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[54] Bo Zhao, Bo Chang, Zequn Jie, and Leonid Sigal. Modular generative adversarial networks. In *ECCV*, 2018. 4, 8

[55] Shuchang Zhou, Taihong Xiao, Yi Yang, Dieqiao Feng, Qinyao He, and Weiran He. Genegan: Learning object transfiguration and attribute subspace from unpaired data. *arXiv preprint arXiv:1705.04932*, 2017. 3

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1

[57] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 1, 3