

# Disentangled Face Attribute Editing via Instance-Aware Latent Space Search

Yuxuan Han<sup>1</sup>, Jiaolong Yang<sup>2†</sup>, Ying Fu<sup>1\*</sup>

<sup>1</sup>Beijing Institute of Technology

<sup>2</sup>Mircosoft Research Asia

{hanyuxuan, fuying}@bit.edu.cn, jiaoyan@microsoft.com

## Abstract

Recent works have shown that a rich set of semantic directions exist in the latent space of Generative Adversarial Networks (GANs), which enables various facial attribute editing applications. However, existing methods may suffer poor attribute variation disentanglement, leading to unwanted change of other attributes when altering the desired one. The semantic directions used by existing methods are at attribute level, which are difficult to model complex attribute correlations, especially in the presence of attribute distribution bias in GAN’s training set. In this paper, we propose a novel framework (**IALS**) that performs **I**nstance-**A**ware **L**atent-**S**pace **S**earch to find semantic directions for disentangled attribute editing. The instance information is injected by leveraging the supervision from a set of attribute classifiers evaluated on the input images. We further propose a Disentanglement-Transformation (*DT*) metric to quantify the attribute transformation and disentanglement efficacy and find the optimal control factor between attribute-level and instance-specific directions based on it. Experimental results on both GAN-generated and real-world images collectively show that our method outperforms state-of-the-art methods proposed recently by a wide margin. Code is available at <https://github.com/yxuhan/IALS>.

## 1 Introduction

The task of face attribute editing aims to alter a given face image towards a given attribute, such as age, gender, expression, and eyeglasses. A successful editing should not only output high-quality results with accurate target attribute, but also well preserve all other image content characterized by the complementary attributes. Face attribute editing has attracted much attention in recent years and numerous algorithms have been proposed [Shen and Liu, 2017; Choi *et al.*, 2018; Zhang *et al.*, 2018; Bahng *et al.*, 2020; Awiszus *et al.*, 2019; Gu *et al.*, 2019]. Notwithstanding the

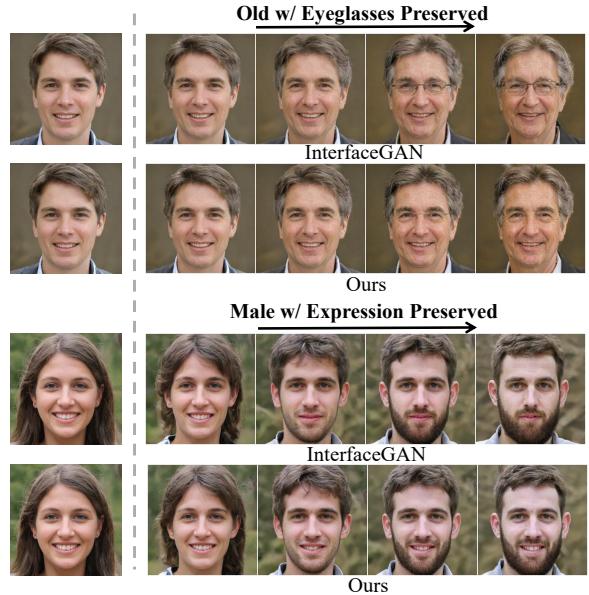


Figure 1: Our goal is to change the primal attribute of a given face (e.g., age and gender here) while preserving other condition attributes (e.g., eyeglasses and expression). Despite the latent space directions of the primal and condition attributes in the InterfaceGAN method have been orthogonalized, changing primal attribute still leads to unwanted condition attribute change. Our IALS method can produce satisfactory disentangled editing results.

promising results demonstrated by these methods, this task is still quite challenging due to the high output dimension and need for precisely disentangling factors of variation corresponding to different face attributes.

Face image synthesis has achieved tremendous success recently with Generative Adversarial Networks (GANs) [Goodfellow *et al.*, 2014]. State-of-the-art GAN models [Karras *et al.*, 2019; Karras *et al.*, 2020] can generate high-fidelity face images from the learned latent space. This motivates several works [Shen *et al.*, 2020; Voynov and Babenko, 2020; Shen and Zhou, 2020; Härkönen *et al.*, 2020] to edit face attribute by reusing the knowledge learned by GAN models. Specifically, for each attribute, they search a corresponding direction in the GAN latent space, such that moving a latent code along this direction can lead to the desired change of

\*Corresponding Author: fuying@bit.edu.cn.

†Work of JY was done in September 2020.

this attribute in the generated images. Although the target attributes can be changed effectively by these methods, disentangled editing is still problematic. As illustrated in Figure 1, eyeglasses may appear when age is changed from young to old, despite the latent space direction of age have been made orthogonal to eyeglasses via projection [Shen *et al.*, 2020]. There are at least two possible reasons for this issue: *i*) the distribution bias of GAN’s training set (*e.g.*, elder people tend to wear eyeglasses in FFHQ [Karras *et al.*, 2019]), and *ii*) the attribute-level directions cannot handle complex attribute distributions and are not effective for attribute variation disentanglement.

In this paper, we propose a novel framework to search the semantic directions in GAN latent space for disentangled face attribute editing. Instead of naively using fixed, *attribute-level* directions, we opt for dynamically searching *instance-aware* directions, where instance refers to the input image to be edited. The intuition behind is that by leveraging the instance information, the complementary attributes of the instance can be explicitly and effectively preserved, leading to disentangled editing results. To render the directions instance-aware, we consider the instance-specific direction obtained by back-propagating the gradient of off-the-shelf attribute CNN classifiers on the input image, and introduce a control factor to balance the attribute-level and instance-specific direction components. We propose a Disentanglement-Transformation (*DT*) metric to quantitatively evaluate the editing results and select the control factor that leads to the highest *DT* metric.

We test our method on both GAN-generated images and real ones, the latter of which are achieved by GAN inversion [Abdal *et al.*, 2019] and re-generation. Experiments show that our method can achieve high-quality disentangled face editing results, outperforming state-of-the-art methods on both GAN-generate and real-world images by a wide margin. In summary, our contributions include:

- We propose a novel face attribute editing framework that searches for instance-aware semantic direction in GAN latent space, which explicitly promotes attribute variation disentanglement;
- We propose a Disentanglement-Transformation (*DT*) metric to quantitatively evaluate the editing efficacy and optimize our algorithm by leveraging this metric;
- We achieve high-quality results on both GAN-generated and real images that significantly outperform existing methods.

## 2 Related Work

Face attribute editing aims to manipulate the interested face attribute while preserving the rest. To achieve this goal, previous methods often leverage the conditional GAN model [Mirza and Osindero, 2014; Odena *et al.*, 2017]. These methods usually design the loss function [Shen and Liu, 2017; Bahng *et al.*, 2020; Zhu *et al.*, 2017] or network architecture [Choi *et al.*, 2018; Zhang *et al.*, 2018; Liu *et al.*, 2019; Lin *et al.*, 2019; He *et al.*, 2019] manually to improve the output quality. Recently, 3D prior (*e.g.* 3DMM [Blanz and

Vetter, 1999]) is also introduced to encode the synthetic image [Deng *et al.*, 2020; Tewari *et al.*, 2020]. These methods can generate high-quality results, but the diversity of controllable attributes is limited by the 3D priors (*e.g.*, it cannot well model and edit gender information). Other recent methods leverage a high-quality synthetic face image dataset for disentangled representation training [Kowalski *et al.*, 2020].

Another set of works [Shen *et al.*, 2020; Voynov and Babenko, 2020; Shen and Zhou, 2020; Plumerault *et al.*, 2020; Härkönen *et al.*, 2020; Goetschalckx *et al.*, 2019] propose to edit face attribute by moving the latent code along a specific semantic direction in the latent space of well-trained unconditional GAN model [Karras *et al.*, 2018; Karras *et al.*, 2019; Karras *et al.*, 2020; Goodfellow *et al.*, 2014]. Shen *et al.* [Shen *et al.*, 2020] learn an SVM boundary to separate the latent space into the opposite semantic label and output the normal vector of the SVM boundary as the semantic direction. Härkönen *et al.* [Härkönen *et al.*, 2020] sample a collection of latent codes and perform PCA on them to find principle semantic directions. However, these methods search the semantic directions on the attribute level, which cannot handle complex attribute correlations. Our method dynamically searches instance-aware semantic direction, which is effective for attribute variation disentanglement. It can edit various face attributes and obtain high-quality results.

## 3 Method

This section introduces the novel face attribute editing framework proposed by this paper. We begin by introducing face attribute editing with GANs and briefly revisiting prior methods using attribute-level latent directions, after which we present our instance-aware direction search algorithm.

### 3.1 Semantic Direction for Attribute Editing

Given a pretrained generator  $G$  from a state-of-the-art GAN model, *e.g.*, StyleGAN [Karras *et al.*, 2019], which maps a latent vector  $z$  to a face image, attribute editing can be achieved by moving  $z$  along a certain direction in the latent space. For real images, one can first embed them into the latent space to obtain the latent vector  $z$  and then modify it. The key is to find suitable semantic directions for attribute editing.

Let  $\mathcal{A}$  denote a collection of face attributes, *e.g.*, age, gender, expression, and eyeglasses. For each attribute  $X \in \mathcal{A}$ , existing methods seek for a direction  $d_X$  corresponding to  $X$ . For example, the InterfaceGAN method [Shen *et al.*, 2020] first generates a large corpus of images  $G(z)$  by randomly sampling  $z$ . Then it labels the attributes of these images using a set of CNN binary classifiers  $H(\cdot)$ . Finally, it trains a SVM to separate each attribute label in the GAN latent space using these samples and outputs the normal vector of the SVM boundary as the semantic direction  $d_X$  for each attribute. To achieve disentangled editing, it proposes to edit the primal attribute  $A$  while preserving one condition attribute  $B$  (or more) via direction orthogonalization:

$$d_{A|B} = d_A - \langle d_A, d_B \rangle d_B, \quad (1)$$

where  $\langle \cdot, \cdot \rangle$  denotes inner product. Directions with more than one condition attribute can be obtained similarly as discussed in [Shen *et al.*, 2020].

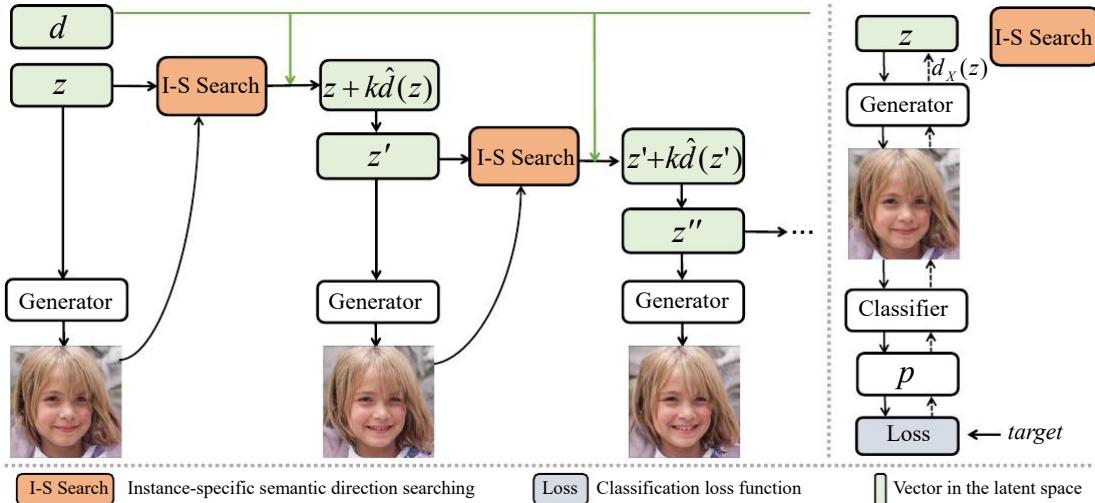


Figure 2: The overview of our face attribute editing framework. The left part shows our instance-aware semantic direction searching method in an incremental update scheme. Here,  $d$  and  $\hat{d}(z)$  stand for the attribute-level and instance-aware semantic directions, respectively. The right part illustrates the instance-specific semantic search process.

It can be seen that the directions so-obtained are at attribute level. They are fixed for each attribute and are instance-agnostic. We instead propose to incorporate instance information into direction search for better editing performance.

### 3.2 Instance-Aware Semantic Direction Search

Our instance-aware semantic direction consists of two parts: *instance-specific* and *attribute-level* directions. Next we first introduce the two components and then describe how to combine them.

#### Instance-Specific Semantic Direction

The generator  $G$  maps a latent code  $z$  to an image and an attribute classifier  $H$  maps an image  $x \in \mathcal{X}$  to an attribute label. We can bridge the GAN latent space and attribute space via compositing  $H$  and  $G$ , *i.e.*  $H(G(\cdot))$ . For attribute  $X$ , we can search for the instance-specific semantic direction for instance  $z$ , denoted by  $d_X(z)$ , via minimizing the following loss:

$$\arg \min_{d_X(z)} L(H(G(z + d_X(z))), y), \quad (2)$$

where  $y$  is the target attribute label and  $L(\cdot, \cdot)$  is the classification loss with binary cross entropy function

$$L(x, y) = -y \log x - (1 - y) \log(1 - x). \quad (3)$$

We simply use gradient descent to search for  $d_X(z)$  in Eq. (2), where  $d_X(z)$  is updated by using an incremental direction updating scheme. To streamline the presentation, more details of incremental update is deferred to a later section. We further normalize  $d_X(z)$  as the final instance-specific semantic direction for  $z$ :

$$\begin{aligned} d_X(z) &= \frac{-\nabla_z L(H(G(z)), y)}{\|\nabla_z L(H(G(z)), y)\|_2} \\ &= (2y - 1) \frac{\nabla_z H(G(z))}{\|\nabla_z H(G(z))\|_2}. \end{aligned} \quad (4)$$

Eq. (4) shows that opposite directions can be obtained with  $y = 0$  and  $1$ , respectively.

#### Attribute-Level Semantic Direction

Attribute-level directions aggregate the information across all training instances thus have higher resistance to noise. Similar to previous method, we also leverage attribute-level directions for attribute editing and use the one computed by InterfaceGAN as the default option.

Here we propose another way to compute the attribute-level direction. The intuition is to bridge the gap between the local instance-level and the global attribute-level information via sampling and averaging. Specifically, we can first randomly sample a set of GAN latent codes  $z$  and generate face images by them accordingly. Then, we compute the instance-specific direction for each sample and each attribute according to Eq. (4). Finally, for each attribute, we can average the instance-specific directions from all samples as the attribute-level direction  $d_X$ . We find that these  $d_X$  lead to similar attribute editing quality compared to the counterparts computed by InterfaceGAN, but this approach is simpler and easier to implement and eliminates the need for training SVMs.

#### Instance-Aware Semantic Direction

Our instance-aware semantic direction is constructed by injecting instance-specific information into the direction search process. Specifically, we formulate it as the combination of the aforementioned attribute-level and instance-specific directions:

$$\hat{d}_X(z) = \lambda d_X + (1 - \lambda) d_X(z), \quad (5)$$

where  $\lambda \in [0, 1]$  is the control factor to balance these two components. For conditional attribute editing, we rewrite Eq. (1) with our instance-aware semantic direction as

$$\hat{d}_{A|B}(z) = \hat{d}_A(z) - \langle \hat{d}_A(z), \hat{d}_B(z) \rangle \hat{d}_B(z), \quad (6)$$

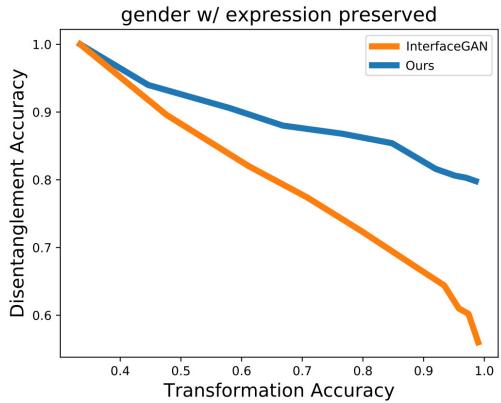


Figure 3: An example  $DT$  curve with gender as the primal attribute and expression as the condition attribute.

where  $A$  and  $B$  are the primal and conditional attributes, respectively. We set two different values for  $\lambda_1$  and  $\lambda_2$  for these two attributes respectively, considering we might require a different quantity of instance information when editing or disentangling face attributes. Note that Eq. (6) degenerates to Eq. (1) when  $\lambda_1 = \lambda_2 = 1$ . In practice, we first solve for the  $(\lambda_1, \lambda_2)$  pair which can produce best results on a sample set, after which we fix them for the facial attribute editing task. Next, we discuss how to solve the  $(\lambda_1, \lambda_2)$  pair.

As mentioned previously, a good attribute editing should *i*) transform the primal attribute  $A$  to target label accurately, and *ii*) preserve the condition attribute  $B$  as much as possible. According to these criteria, we propose to use a Disentanglement-Transformation ( $DT$ ) curve to evaluate the editing results for a pair of primal and condition attributes on a set of samples. In a  $DT$  curve, the  $x$ -axis represents the transformation accuracy  $p$ , *i.e.*, the ratio of samples for which the primal attribute has been transformed into target label correctly in the editing results. The  $y$ -axis represents the disentanglement accuracy  $q$ , *i.e.*, the ratio of samples for which the condition attribute on the edited images is consistent with their original ones.

Specifically, we randomly sample a set of latent codes in the GAN latent space and obtain the corresponding images generated by  $G$ . Then we edit these images by changing the latent code  $z$  to  $z + k \cdot n \cdot d_{A|B}(z)$ , where  $k \in \mathbb{R}$  is the step size and  $n \in \mathbb{N}$  is the number of steps. The transformation and disentanglement accuracy  $(p_n, q_n)$  are evaluated by CNN classifiers for the attributes. By continuously changing  $n$  and evaluating  $(p_n, q_n)$ , we obtain a  $DT$  curve as illustrated in Figure 3 (in practice, the  $DT$  curve of our method is obtained via varying the incremental updating steps; see next section).

With  $DT$  curves generated, we can further compute the AUC (Area Under Curve) of  $DT$ , and we choose the  $(\lambda_1, \lambda_2)$  pair that maximizes the average AUC for all possible primal-condition attribute pairs via:

$$\arg \max_{\lambda_1, \lambda_2} \frac{1}{N} \sum_{A, B \in \mathcal{A}} \int_0^1 q(A, B, \lambda_1, \lambda_2, p) dp, \quad (7)$$

where  $N$  is the normalization factor. We numerically estimate the continuous integral in Eq. (7) using quadrature.

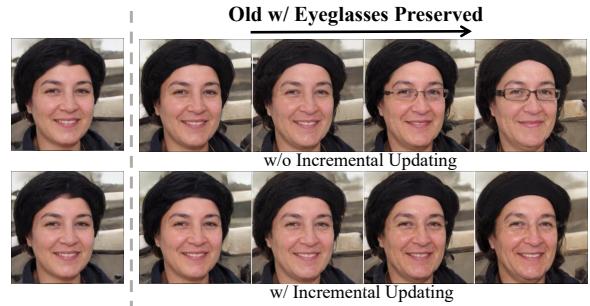


Figure 4: Ablation study of the incremental updating scheme.

### 3.3 Incremental Direction Updating

Here, we present an incremental instance-aware direction search scheme for our editing method. We alternate between searching for the semantic direction and updating the latent code (and the output image). Given the current latent code  $z$  and generated image  $G(z)$ , we search for the new direction  $\hat{d}(z)$  and update the latent code as  $z' = z + k \cdot \hat{d}(z)$ . The whole process is illustrated in Figure 2.

## 4 Experiment

In this section, we evaluate our IALS method and compare it against the state-of-the-arts. Due to space limitation, more experimental results and discussions are included in the *suppl. material*.

### 4.1 Settings and Implementation Details

We consider five face attributes in our experiment, *i.e.*, expression, age, gender, eyeglasses, and pose, as in [Shen *et al.*, 2020]. We will focus on the former four attributes in our experiments as we find pose is well disentangled from other attributes in GAN latent space (similar observations are found in [Shen *et al.*, 2020; Deng *et al.*, 2020]). We test our framework on the  $\mathcal{W}$  space of StyleGAN generator trained on the FFHQ [Karras *et al.*, 2018] and CelebA-HQ [Karras *et al.*, 2019] datasets. The attribute classifiers  $H(\cdot)$  are ResNet-18 [He *et al.*, 2016] networks trained on the CelebA dataset [Liu *et al.*, 2015]. The  $DT$  metrics are computed by another set of attribute classifiers with ResNet-50 structure. We empirically set the step size of incremental updating in our method to 0.1 in the following experiments.

### 4.2 Ablation Study

In this part, we investigate the effect of direction control factor and incremental updating scheme in our IALS framework.

#### Control Factor Pairs

To study the behavior of different combination weights for attribute-level and instance-specific directions, and solve for control factors  $\lambda_1$  and  $\lambda_2$ , we evaluate the  $DT$  metric with  $\lambda_1$  and  $\lambda_2$  evenly sampled in  $[0, 1]$  with a step size of 0.25 (hence 25  $(\lambda_1, \lambda_2)$  pairs in total). For each  $(\lambda_1, \lambda_2)$  pair, we set  $n_{max} = 20$  (*i.e.* sample 20 points on the  $DT$  curve) and  $k = 0.1$  and adopt the trapezoidal quadrature formula to approximate the integral in AUC computation defined in Eq. (7).

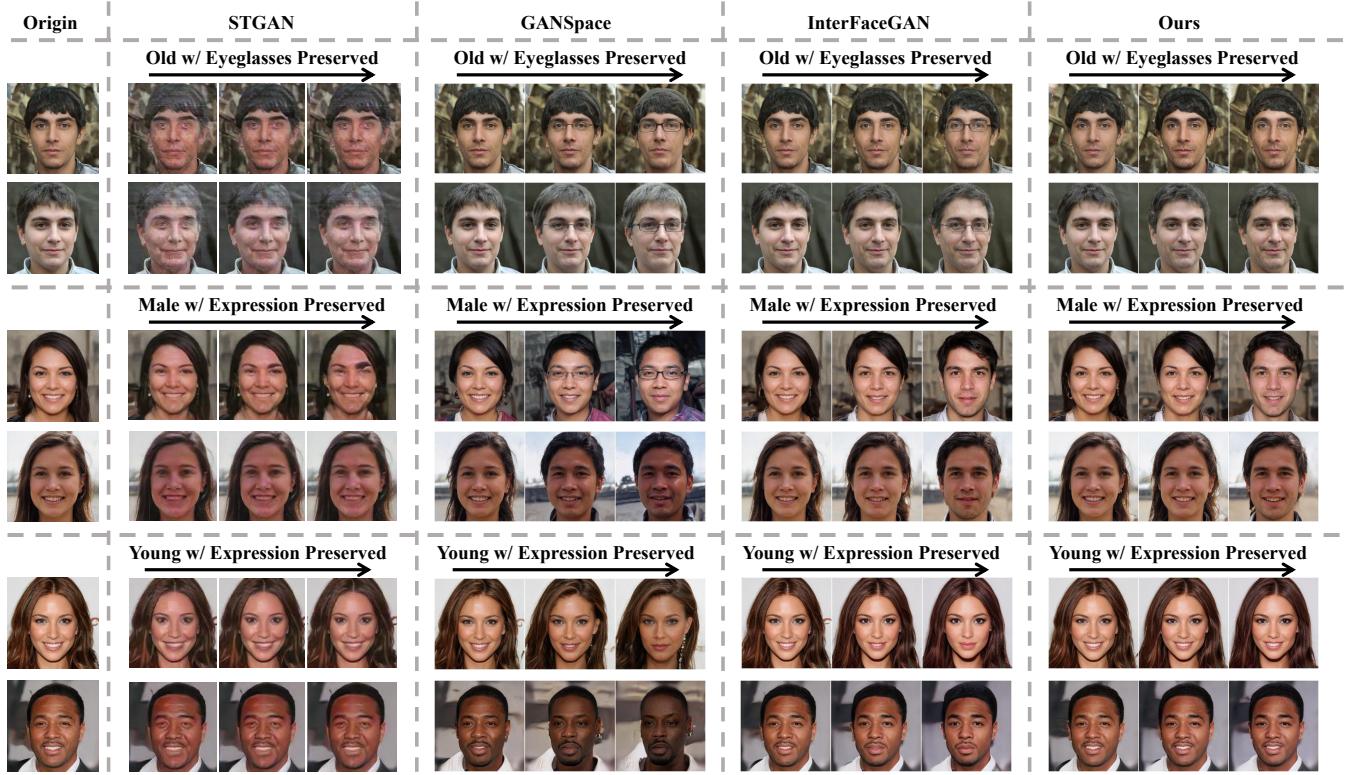


Figure 5: Qualitative comparison of face attribute editing results between our method and other competitors on GAN-generated images. By our method, not only the primal attributes are successfully changed but also the condition attributes are preserved much better than STGAN, GANSpace and InterfaceGAN.

$\lambda_1 \setminus \lambda_2$	0.0	0.25	0.5	0.75	1.0
0.0	0.8393	0.8454	0.8445	0.8444	0.8302
0.25	0.8717	0.8675	0.8684	0.8706	0.8463
0.5	0.8882	0.8889	0.8886	0.8922	0.8660
0.75	<b>0.9025</b>	0.8992	0.9010	0.9003	0.8742
1.0	0.8999	0.9003	0.9006	0.9013	0.8662

Table 1: The average AUC (Area Under Curve) of all  $DT$  curves with different  $(\lambda_1, \lambda_2)$  choices.

Table 1 shows the average AUC of all  $DT$  curves with different primal-condition attribute combinations. We have the following three observations: *i*) using attribute-level direction alone (*i.e.*,  $\lambda_1 = \lambda_2 = 1$ ) is clearly non-optimal; *ii*) the attribute-level information is more important to edit primal attribute ( $\lambda_1 \geq 0.75$ ); and *iii*) adding instance information to condition attribute direction search significantly improves the editing results ( $\lambda_2 \leq 0.75$ ).

**Discussion** Table 1 shows a performance plateau for a range of  $\lambda$ 's ( $\lambda_1 \geq 0.75$ ,  $\lambda_2 \leq 0.75$ ), indicating that our method is insensitive to parameter selection within a reasonable range. It also shows that adding moderate instance-level information for the condition attribute would significantly better than using attribute-level information alone ( $\lambda_2 = 1$ ), demonstrating the efficacy of our framework. In the following we simply use  $(\lambda_1, \lambda_2) = (0.75, 0)$  for our editing method.

### Incremental Updating

We further study the efficacy of our incremental updating scheme for instance-aware direction search. Figure 4 shows a typical result when changing age from young to old while preserving the eyeglasses attribute of the original face image. It can be seen that the results are clearly inferior if we do not use incremental updating (*i.e.*, we keep using the instance-aware direction obtained in the first iteration).

### 4.3 Comparison with the State-of-the-arts

We compare our IALS method with several state-of-the-art face attribute editing method proposed recently, including InterfaceGAN [Shen *et al.*, 2020], GANSpace [Härkönen *et al.*, 2020], and STGAN [Liu *et al.*, 2019].

Both InterfaceGAN and GANSpace are methods based on GAN latent space search. InterfaceGAN is a supervised semantic direction search method as mentioned in the previous section, while GANSpace is an unsupervised one which performs PCA on the sampled latent codes to find principle semantic directions in the latent space. We assign the directions found by GANSpace to interpretable meanings following [Shen and Zhou, 2020]. The STGAN method is based on image generation using conditional GAN. It uses an encoder-decoder architecture with a well-designed selective transfer unit for attribute editing. For these methods, we use the code or pretrained models released by the authors for evaluation and comparison.

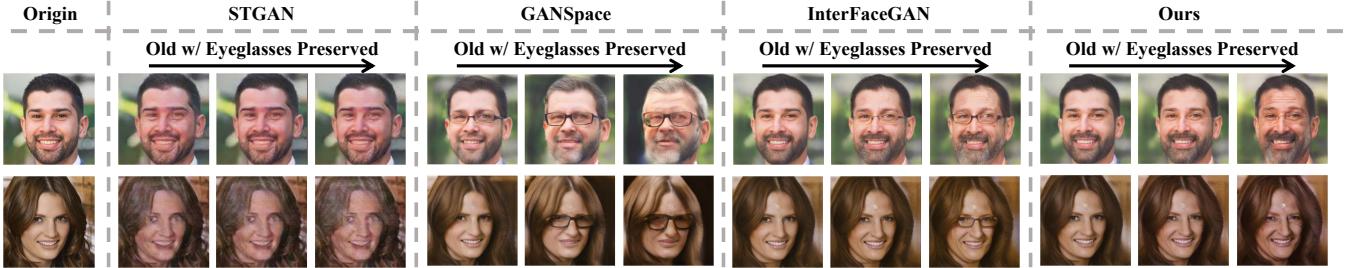


Figure 6: Qualitative comparison of face attribute editing results between our method and other competitors on real images from FFHQ. Our method yields higher fidelity results than STGAN and better attribute variation disentanglement than GANSpace and InterfaceGAN. Photos by Flickr users Elio Yañez (Creative Commons BY 2.0 License) and sexinhose (Public Domain Mark 1.0 License).

## Qualitative Results

we compare our method with the competitors on synthetic images and real images respectively. We adopt the conditional manipulation setup [Shen *et al.*, 2020] with one primal attribute to be changed and one condition attribute to be preserved for the semantic direction based methods, and directly send the source image to the conditional-GAN based method.

Some typical qualitative results on GAN-generated images are provided in Figure 5. It can be seen that as the degree of attribute change increases, STGAN generated blurry images while the semantic direction based methods outputted high-fidelity results. Furthermore, our method can well preserve the condition attribute of the original image when editing the primal attribute while the other two semantic direction based competitors often fail. InterfaceGAN and GANSpace only use attribute-level direction while ignoring the instance information when facial editing. By contrast, our method combines instance-specific and attribute-level information which results in much better disentangled facial editing results. The results also demonstrate that our method outperforms state-of-the-art methods on GAN-generated images.

Figure 6 shows two typical examples to illustrate the attribute editing efficacy of different methods on real images<sup>1</sup> where the goal is to edit age while preserving the eyeglasses attribute. We find that STGAN outputted blurry results again. On the other hand, the condition attributes (*i.e.* eyeglasses) are changed by GANSpace and InterfaceGAN. In contrast, our method obtains superior results in terms of both image quality and attribute variation disentanglement. It produces high fidelity results with satisfactory attribute modification and preservation.

## Quantitative Results

To further evaluate the performance of our IALS method, we conducted a user study. we recruited 100 people and presented them with 720 groups of data, with each group consisting of 5 images - the original face image and the facial editing results of our method and the other 3 competitors. Each person was randomly assigned 18 groups of data and asked to choose all of the results they are satisfied with according to three criteria: the result looks natural, the primal attribute is

<sup>1</sup>For InterfaceGAN, GANSpace and our method, we firstly embed the given images into the  $\mathcal{W}^+$  latent space of StyleGAN using [Abdal *et al.*, 2019] and then edit them.

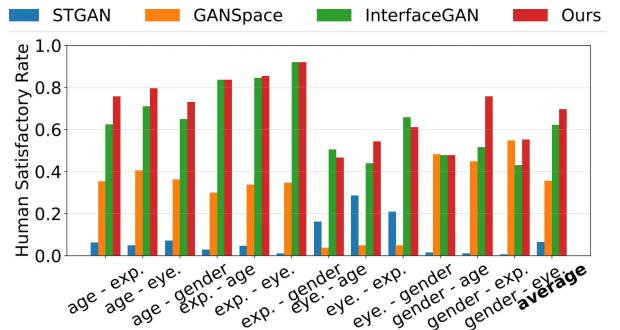


Figure 7: Human satisfactory rate comparison between our method and other competitors on different primal-condition attribute pairs (exp. and eye. denote expression and eyeglasses respectively)

well changed, and condition attribute is well preserved. Note that we did not ask them to select the best result, as optimal facial editing results are often not unique.

The results in Fig. 7 show that our method obtained the highest average satisfactory rate (69.66% for ours vs. 62.06% for InterfaceGAN, 35.44% for GANSpace and 6.48% for STGAN). It significantly outperforms the other methods on some challenging attribute pairs such as gender as the primal attribute and expression as the condition attribute (76.44% for ours vs. 51.92% for InterfaceGAN).

## 5 Conclusion

We have proposed a novel IALS framework to achieve disentangled face attribute editing with GAN latent space search. The key ingredient is the integration of attribute-level and instance-specific directions, which leads to accurate target attribute control and, more importantly, highly disentangled attribute editing. To better integrate the attribute-level and instance-specific directions, we introduce a Disentanglement-Transformation (*DT*) metric to find suitable control factor. The experiments collectively show that our approach obtains significantly better results than previous methods.

## Acknowledgements

This research was supported by the National Natural Science Foundation of China under Grants No. 61827901 and No. 62088101.

# Supplementary Material

## 6 More Implementation Details

We consider five attributes in our implementation, *i.e.*, expression, age, gender, eyeglasses and pose, as in [Shen *et al.*, 2020]. Note that our method can be extended to edit any other attributes in the presence of corresponding CNN classifier. To search instance-specific directions, we train a set of binary CNN classifiers  $H(\cdot)$  on the CelebA dataset [Liu *et al.*, 2015] with  $256 \times 256$  image resolution. Unlike other attributes, pose is not annotated in CelebA. We therefore construct the training set for it by ourselves. Specifically, we first use the landmarks provided by CelebA annotations to estimate the Euler angles of the face images, and then label images with yaw angle greater than  $30^\circ$  as a positive sample and less than  $-30^\circ$  as a negative sample, respectively. In the training phase, we first center-crop the original  $178 \times 218$  images and then resize them to  $256 \times 256$  before feeding them to  $H(\cdot)$  (we adopt the ResNet-18 [He *et al.*, 2016] architecture), similar to other attributes. Table 2 shows the precision and recall of our trained classifiers for all attributes.

For instance-specific direction searching, we resize the images generated by  $G(\cdot)$  (*i.e.*, the StyleGAN [Karras *et al.*, 2019] generator) from  $1024 \times 1024$  to  $256 \times 256$  before feeding them to  $H(\cdot)$ . In the instance-aware direction constructing process, we first normalize the instance-specific direction  $d_X(z)$  and the attribute-level direction  $d_X$  to unit length, and then combine them by the control factor  $\lambda$ . In addition, we normalize  $\hat{d}_X(z)$  as the final instance-aware semantic direction for  $z$  on attribute  $X$ .

The pseudocode of our method is provided in Algorithm 1.

---

**Algorithm 1** Instance-Aware Semantic Direction Search for Face Attribute Editing

**Input:** latent vector  $z$ , StyleGAN generator  $G$ , CNN classifier  $H$ , attribute-level direction  $d_X$ , target label  $y$ , control factor  $\lambda$ , step size  $k$ , number of step  $N$ .

**Output:** a list  $\mathcal{I}$  containing  $N$  images with continuous attribute variation towards the target label.

```
1: Let  $i \leftarrow 0$ ,  $\mathcal{I} \leftarrow \emptyset$ .
2: while  $i < N$  do
3:    $d_X(z) = \frac{\partial L(H(G(z)), y)}{\partial z}$  (Eq. 4 in the main paper)
4:    $d_X(z) \leftarrow \text{Normalize}(d_X(z))$ 
5:    $d_X \leftarrow \text{Normalize}(d_X)$ 
6:    $\hat{d}_X(z) = \lambda d_X + (1 - \lambda)d_X(z)$  (Eq. 5)
7:    $\hat{d}_X(z) = \text{Normalize}(\hat{d}_X(z))$ 
8:    $z \leftarrow z + k \cdot \hat{d}_X(z)$ 
9:    $\mathcal{I} \leftarrow \text{AppendList}(\mathcal{I}, G(z))$ 
10:   $i \leftarrow i + 1$ 
11: end while
12: return  $\mathcal{I}$ 
```

---

## 7 More Results and Comparisons

### 7.1 More Results of Our Method

We show more real-image editing results of our method on different attributes (including pose) in Figs. 8 and 9. We can see that our method can obtain high-quality disentangled results when editing various face attributes. It implies the effectiveness of our method.

### 7.2 More Results on Multiple Condition Attributes

In Fig 10, we show the facial editing results of our method with a different number of condition attributes. It can be observed that our method is able to handle complex disentanglement between face attributes by adding more condition attributes into the conditional manipulation operation using the strategy presented in [Shen *et al.*, 2020].

### 7.3 More Ablation Study on Control Factors.

In Figs. 11 and 12, we show the qualitative ablation results of varying  $\lambda$  in the primal and condition directions, respectively. We can see that *i*) our method would fail to change the desired primal attribute if we abandon the attribute-level information, and *ii*) our method can effectively obtain disentangled results when adding instance information, which is consistent with the conclusion in the main paper.

### 7.4 More Comparison with the State-of-the-Arts

Here, we provide more qualitative results compared to the state-of-the-art methods, including the semantic direction based methods (InterfaceGAN [Shen *et al.*, 2020] and GANSpace [Härkönen *et al.*, 2020]), and the conditional-GAN based method [Liu *et al.*, 2019] on GAN-generated images. We adopt the conditional manipulation setup [Shen *et al.*, 2020] for the semantic direction based methods, and directly send the source image to the conditional-GAN based method, as in the main paper. Besides, the results in Figs. 13, 14, and 15 show that our method performs better than the other competitors.

### 7.5 More Comparison with InterfaceGAN on attribute-level direction

In the main paper, we point out that the attribute-level direction computed by our method leads to similar editing quality with the counterpart computed by InterfaceGAN [Shen *et al.*, 2020]. Here, we show qualitative comparison in Fig. 16 and demonstrate their similar efficacy. Note that our process is easier to implement, as we eliminate the need for training SVMs. Our main contribution is to use the instance-aware semantic direction for face attribute editing.

## 8 More Details of the Incremental Updating Scheme

In the main paper, we propose an incremental instance-aware direction search scheme, instead of keep using the instance aware direction  $\hat{d}_X(z)$  of the original latent code  $z$  when editing. We treat the latent codes formed in the editing process (*e.g.*  $z + k\hat{d}_X(z)$ , where  $k$  is the step size) as different instances and use their instance-aware direction to edit them.

	Smiling	Gender	Age	Eyeglasses	Pose
Precision	0.9407	0.9613	0.9609	0.9951	0.9884
Recall	0.9094	0.9855	0.8593	0.9264	0.9906

Table 2: Performance of the CNN attribute classifiers (ResNet-18) trained on CelebA. These classifiers are used to edit attributes in our method.

For example, we denote  $z' = z + k\hat{d}_X(z)$  and adopt  $\hat{d}_X(z')$  to edit  $z'$  instead of directly using  $\hat{d}_X(z)$ . Note that the incremental updating scheme makes no difference if we use the attribute-level direction alone in the editing process.

## 9 Discussion about Condition Attribute Selection

In practice, one can freely select one or multiple condition attributes for primal attribute editing, depending on the targeted editing goals. In general, to achieve accurate and disentangled editing for a primal attribute, selection of condition attribute is closely related to the data distribution of GAN’s training set. For example, elder people tend to wear eyeglasses in FFHQ [Karras *et al.*, 2019], so age and eyeglass attributes are likely to be entangled with each other in the latent space. One can also follow this strategy for condition attribute selection.

## 10 Limitations

We have demonstrated that our method is effective for disentangled face attribute editing. It still has some limitations of the semantic direction based facial editing algorithm. For example, our method cannot model non-binary attributes (*e.g.* hair color and hairstyle), as it needs a binary classification boundary for facial editing. In addition, we observed that the image editing quality on embedded latent code obtained by GAN inversion approaches [Abdal *et al.*, 2019; Abdal *et al.*, 2020] is slightly lower than random latent codes from the GAN generator. A recent work from [Zhu *et al.*, 2020] proposed to eliminate the above performance gap by adding a regularization term to encourage the embedded latent code to be semantically meaningful, which we will explore in our future work.

## References

- [Abdal *et al.*, 2019] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, pages 4431–4440, 2019.
- [Abdal *et al.*, 2020] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, pages 8296–8305, 2020.
- [Awiszus *et al.*, 2019] Maren Awiszus, Hanno Ackermann, and Bodo Rosenhahn. Learning disentangled representations via independent subspaces. In *ICCVW*, 2019.
- [Bahng *et al.*, 2020] Hyojin Bahng, Sunghyo Chung, Seungjoo Yoo, and Jaegul Choo. Exploring unlabeled faces for novel attribute discovery. In *CVPR*, pages 5821–5830, 2020.
- [Blanz and Vetter, 1999] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999.
- [Choi *et al.*, 2018] 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*, pages 8789–8797, 2018.
- [Deng *et al.*, 2020] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, pages 5154–5163, 2020.
- [Goetschalckx *et al.*, 2019] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, pages 5744–5753, 2019.
- [Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.
- [Gu *et al.*, 2019] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *CVPR*, pages 3436–3445, 2019.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [He *et al.*, 2019] 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.
- [Härkönen *et al.*, 2020] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NIPS*, 2020.
- [Karras *et al.*, 2018] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.
- [Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CVPR*, pages 4401–4410, 2019.
- [Karras *et al.*, 2020] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020.
- [Kowalski *et al.*, 2020] Marek Kowalski, Stephan J. Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson,

- and Jamie Shotton. Config: Controllable neural face image generation. In *ECCV*, 2020.
- [Lin *et al.*, 2019] Yu-Jing Lin, Po-Wei Wu, Che-Han Chang, Edward Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *ICCV*, pages 5914–5922, 2019.
- [Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.
- [Liu *et al.*, 2019] 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*, pages 3673–3682, 2019.
- [Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [Odena *et al.*, 2017] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, pages 2642–2651, 2017.
- [Plumerault *et al.*, 2020] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. In *ICLR*, 2020.
- [Shen and Liu, 2017] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *CVPR*, pages 1225–1233, 2017.
- [Shen and Zhou, 2020] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020.
- [Shen *et al.*, 2020] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, pages 9243–9252, 2020.
- [Tewari *et al.*, 2020] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Perez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*, pages 6142–6151, 2020.
- [Voynov and Babenko, 2020] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*, 2020.
- [Zhang *et al.*, 2018] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *ECCV*, pages 422–437, 2018.
- [Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017.
- [Zhu *et al.*, 2020] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. *ECCV*, 2020.



Figure 8: More qualitative results of our method on real image editing. The images are taken from FFHQ.



Figure 9: Qualitative results of our method on real image editing. The images are taken from FFHQ.

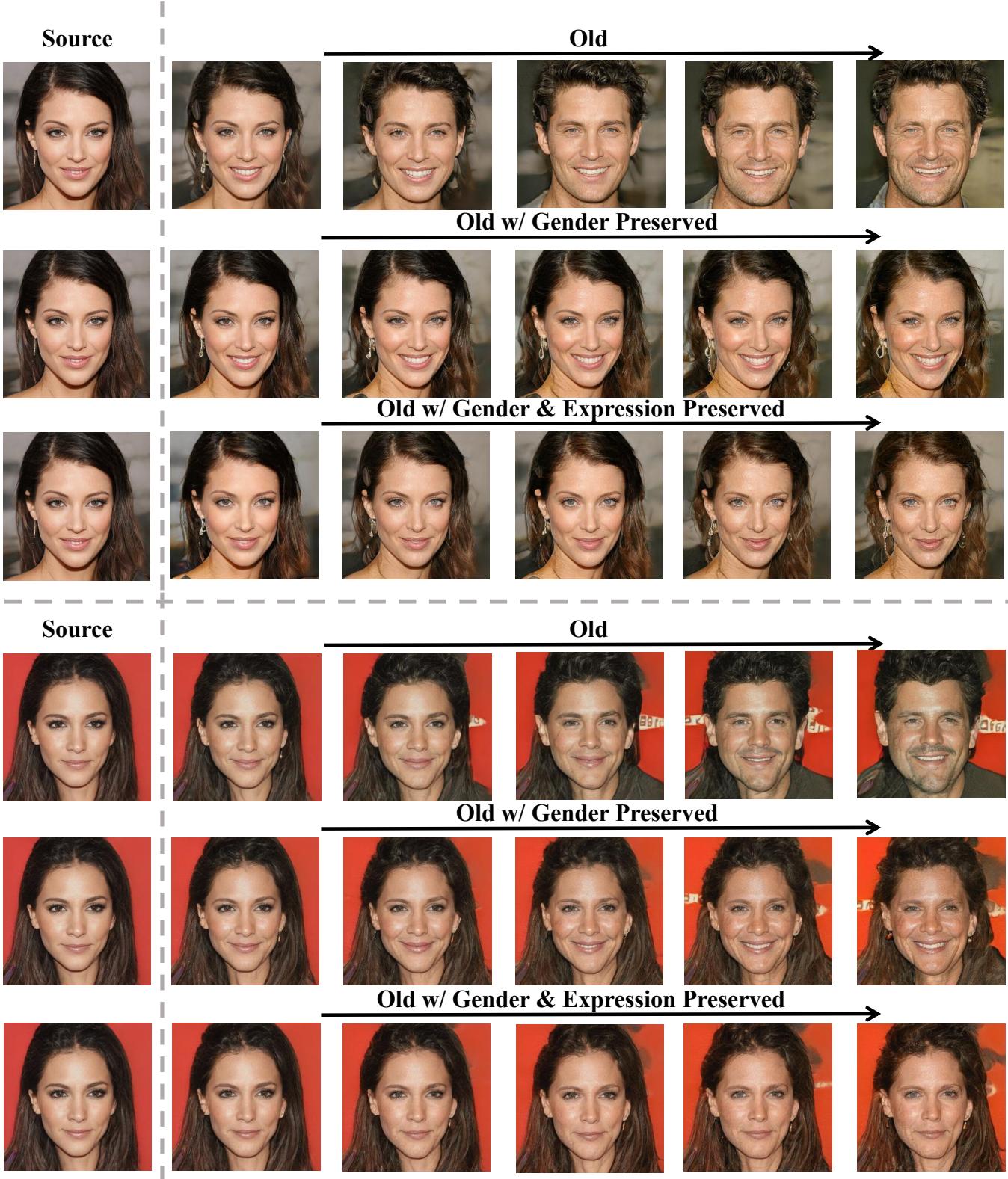


Figure 10: Qualitative results of our method on multiple condition attributes.

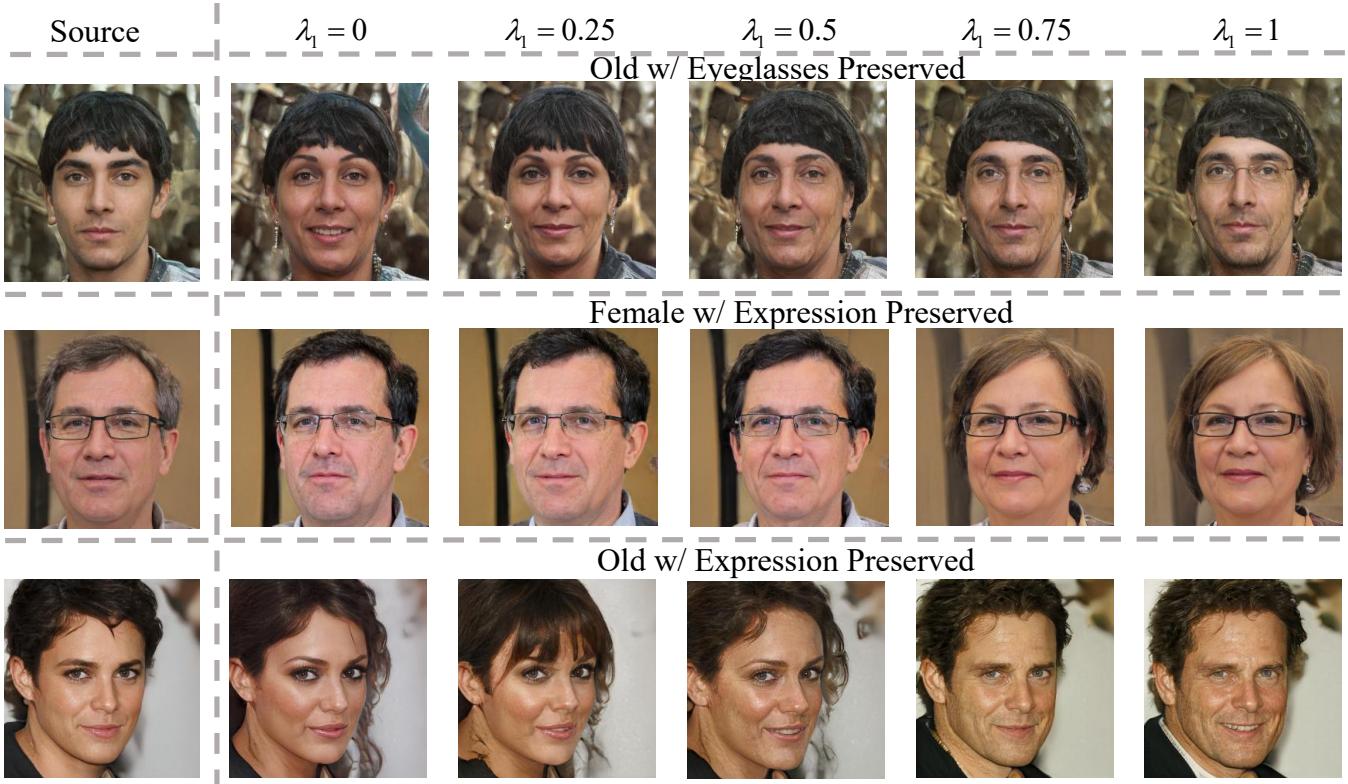


Figure 11: Qualitative ablation study on  $\lambda_1$  (we choose  $\lambda_1 = 0.75$  in this paper based on the *DT* metric).

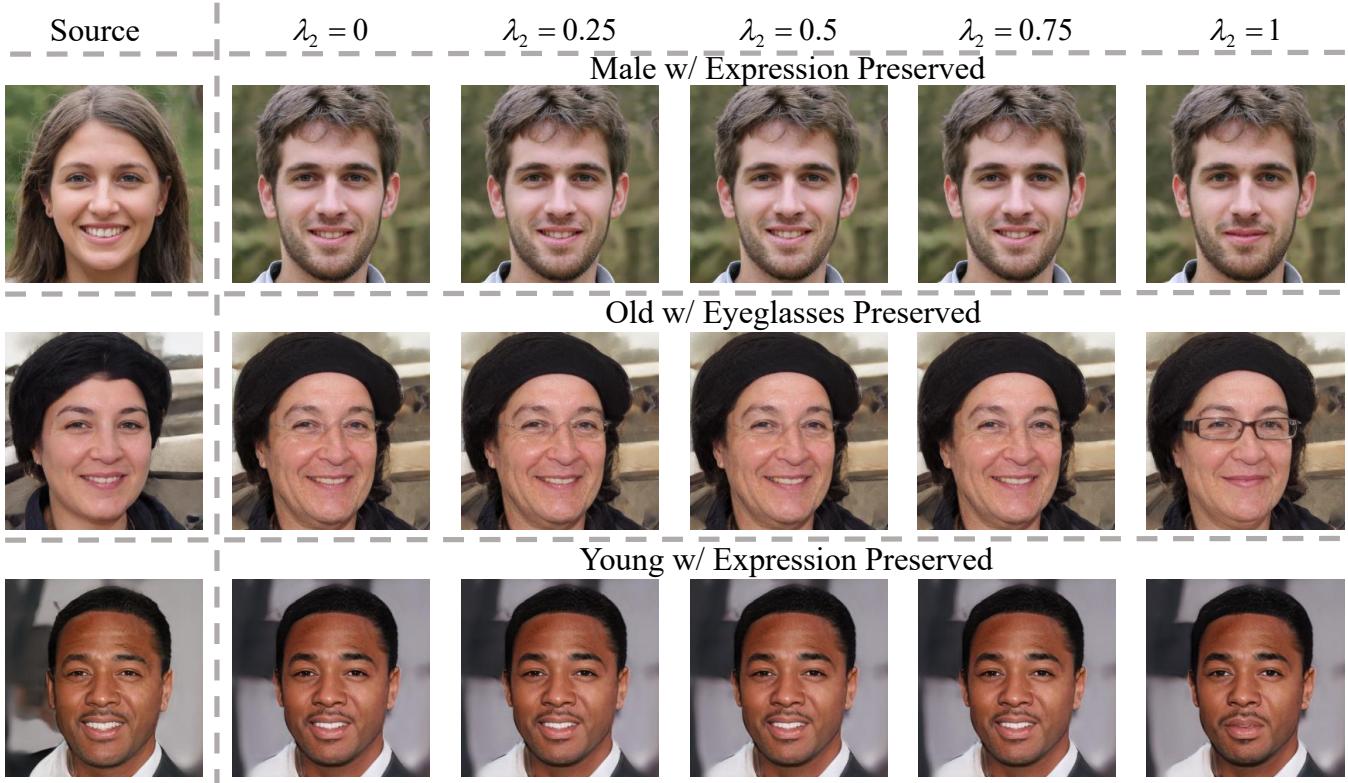


Figure 12: Qualitative ablation study on  $\lambda_2$  (we choose  $\lambda_2 = 0$  in this paper based on the *DT* metric).

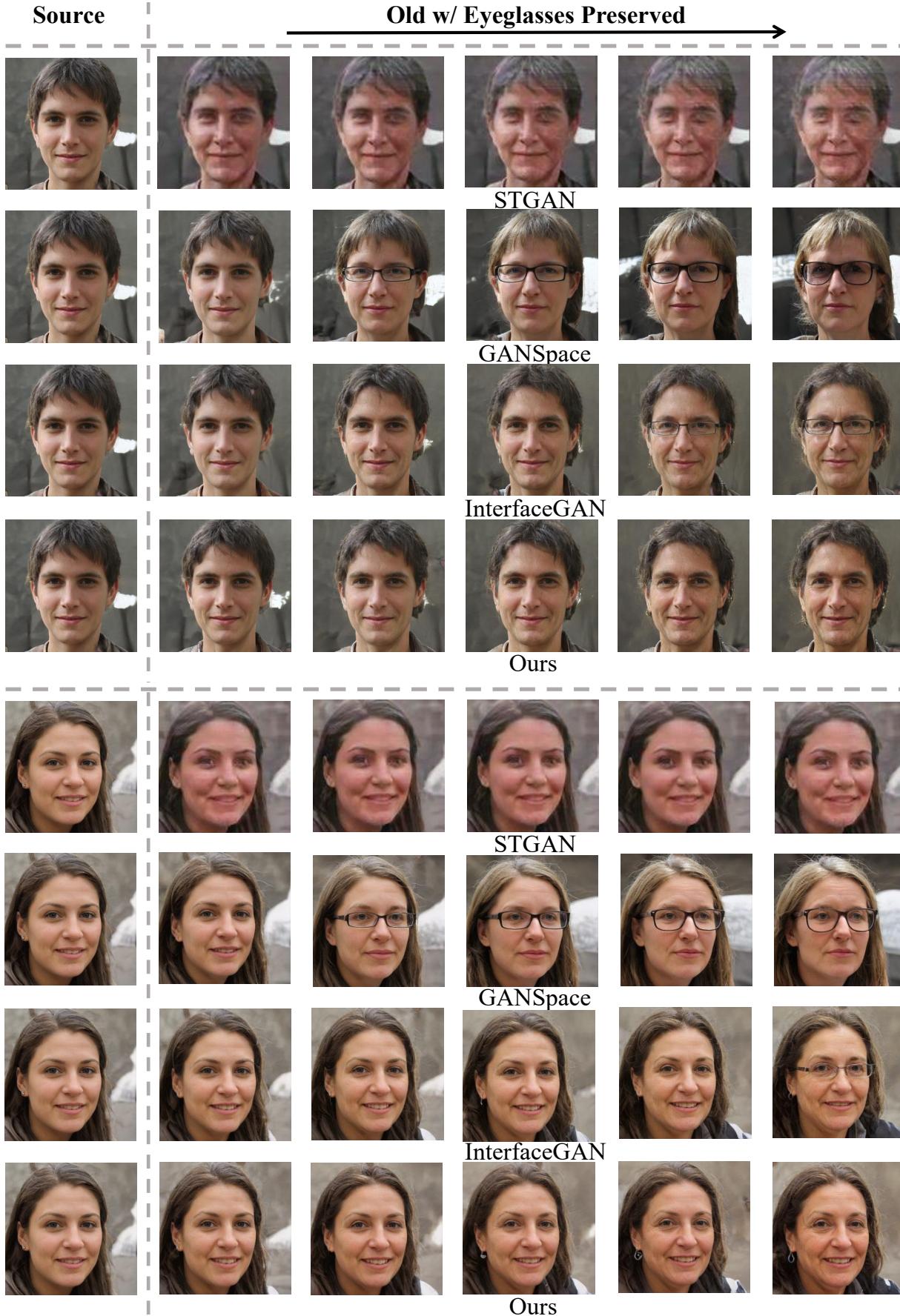


Figure 13: More qualitative comparison with the state-of-the-art on GAN-generated images.

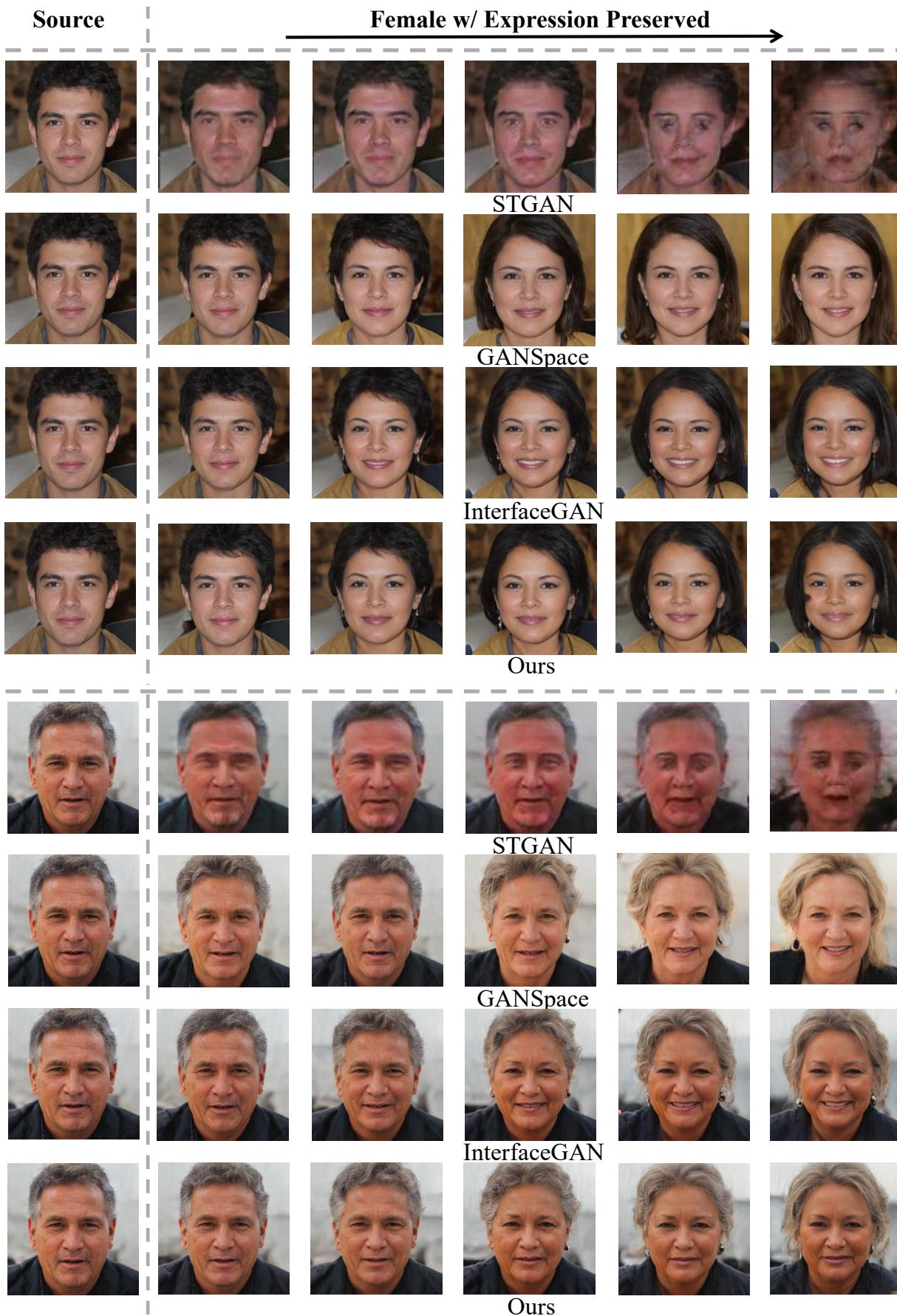


Figure 14: More qualitative comparison with the state-of-the-art on GAN-generated images.

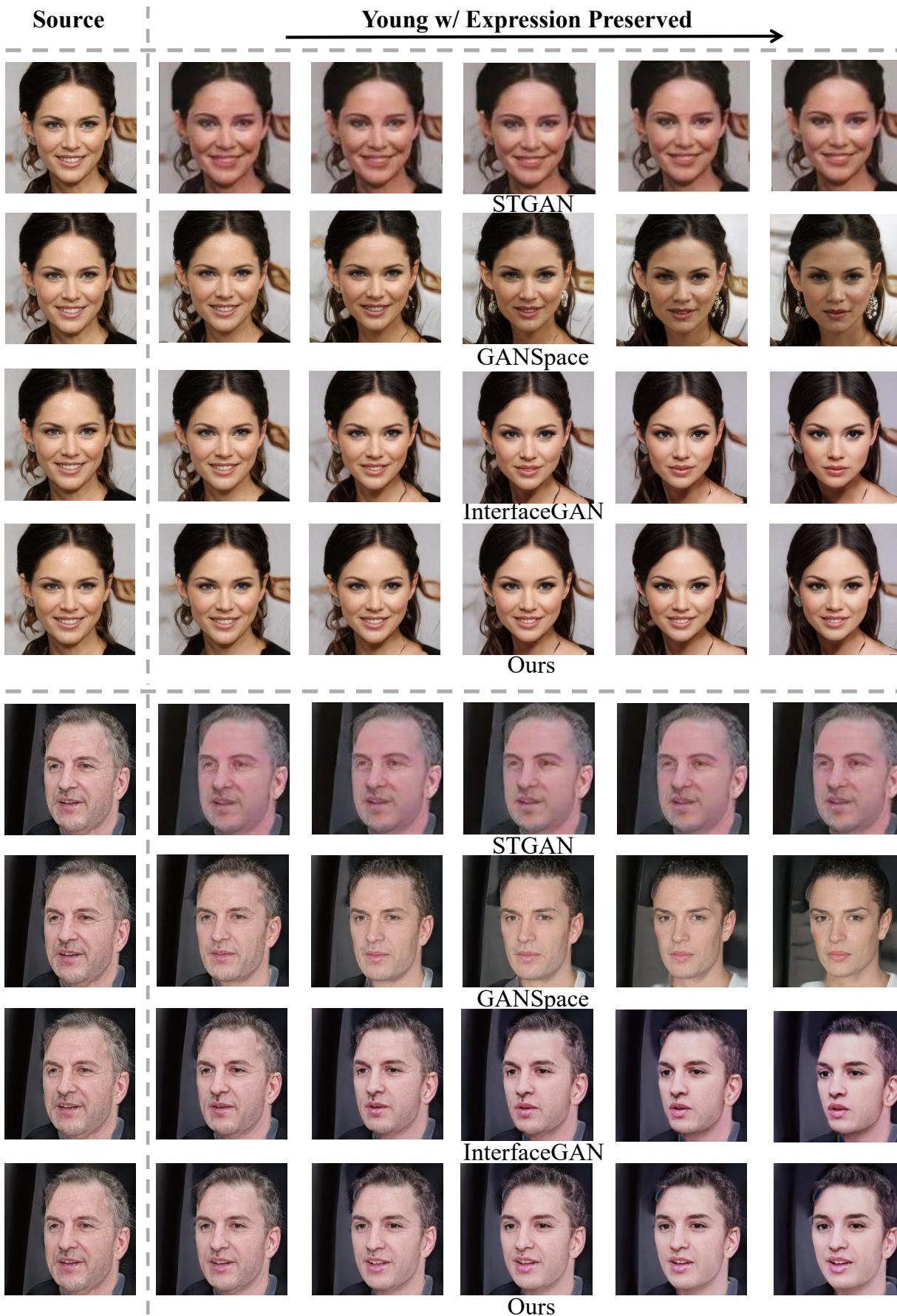


Figure 15: More qualitative comparison with the state-of-the-art on GAN-generated images.

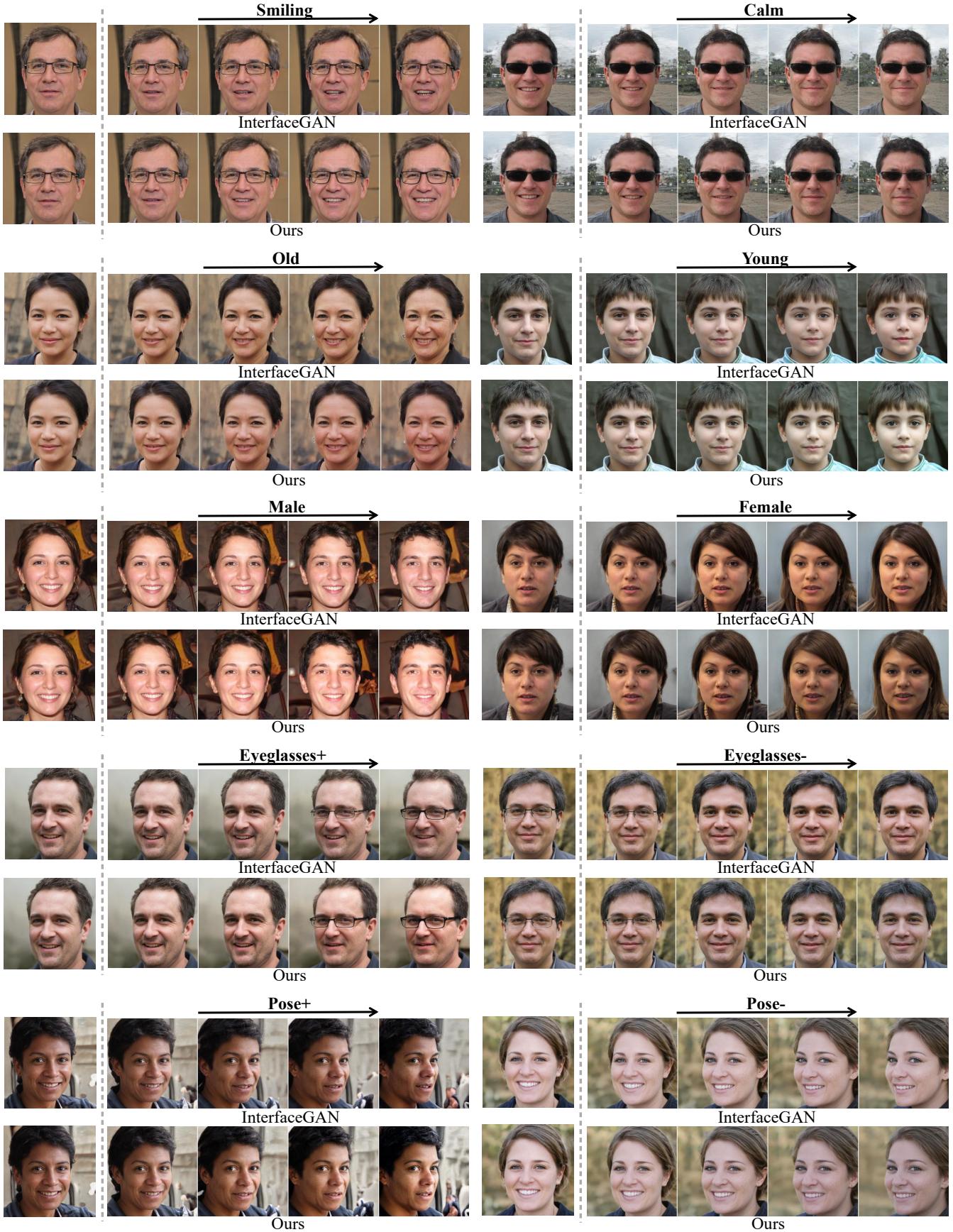


Figure 16: Qualitative comparison of face image editing using the attribute-level directions computed by our method and InterfaceGAN.