

# SAT3D: Image-driven Semantic Attribute Transfer in 3D

## — Supplementary Material —

Anonymous Author(s)

Submission Id: 2290\*

### ABSTRACT

In this supplementary material, we provide more details of the descriptor group definition, visualization images and experimental results that could not be included in the main article due to the space limitation.

**Table 1: The descriptor group definition for semantic attributes of human faces. All the phrases are embedded in a template ‘a face with {}’ respectively before sent to CLIP.**

Attribute	Descriptor Groups
Hairstyle	[curly hair, straight hair, wavy hair, crew cut]
	[frizzy hair, smooth hair, crew cut]
	[tied-up hair, cropped hair, shoulder-length loose hair, chin-length loose hair]
	[pixie cut, crew cut, bowl cut, slicked-back, undercut, Bob cut, long hair]
	[thick hair, thin hair, baldness]
	[curtain bangs, choppy bangs, side-swept bangs, no bangs]
	[center-parting hair, side-parting hair, slicked-back hair, bangs, crew cut]
Hair color	[black hair, brown hair, gray hair, white hair, blond hair, auburn hair, ash green hair]
Eye region	[monolid eyes, double eyelid eyes]
	[deep-set eyes, protruding eyes]
	[blue eyes, brown eyes]
	[sparse eyebrows, full eyebrows]
Expression	[sharp and defined eyebrows, messy and undefined eyebrows]
	[slightly parted lips, tight lips, widely opened mouth, grinned, gentle smile, drooped mouth]
	[half-closed eyes, open eyes, closed eyes, squinting eyes]
	[wrinkled eyebrows, raised eyebrows, unfurled eyebrows]
Beard	[left gazing, right gazing, upward gazing, downward gazing]
	[no beard, beard]
	[bushy beard, sparse beard]
Eyeglasses	[stubble beard, goatee, anchor beard, full beard, mustache]
	[no glasses, glasses, sunglasses]
	[round glasses, oval glasses, square glasses, square glasses]
	[thin-rimmed glasses, thick-rimmed glasses, rimless glasses]
Eyeglasses	[red glasses, black glasses, golden glasses]

**Table 2: The descriptor group definition for semantic attributes of cars. All the phrases are embedded in a template ‘a model of {}’ respectively before sent to CLIP.**

Attribute	Descriptor Groups
Color	[red car, blue car, white car, black car, yellow car, green car, grey car, pink car]
Shape	[sports car, saloon, roadster, truck, van, classic car, SUV]

**Table 3: The descriptor group definition for semantic attributes of cats. All the phrases are embedded in a template ‘a cat with {}’ respectively before sent to CLIP.**

Attribute	Descriptor Groups
Fur	[black fur, white fur, fawn fur, brown fur, mottled fur, grey fur, red fur]
	[solid colored fur, calico fur, tortoiseshell fur, tabby fur]

## 1 DESCRIPTOR GROUPS

We provide the definition of descriptor groups that are used for attribute transfer in SAT3D on three domains. For facial attributes, the phrases presented in Table 1 are templated with ‘a face with {}’. Specifically, for Beard and Eyeglasses, the first descriptor group identifies the presence of attribute and the others describe attribute characteristics. When measuring the attribute similarity between two images, if the attribute is predicted to be present in both images by the first descriptor group, other descriptor groups are further utilized; otherwise, only the first descriptor group is used. The descriptor groups for cars and cats are listed in Table 2 and Table 3 respectively. The phrases are templated with ‘a model of {}’ and ‘a cat with {}’ respectively.

**Table 4: Quantitative comparisons of 3D-aware attribute transfer. AS measures the target attribute similarity between edited and reference images, while AP measures the irrelevant attribute similarity between edited and source images. The metric values are averaged over 5 views for each sample.**

Metrics	Method	Smiling	Beard	Eyeglasses
AS	Preim3D [4]	<b>0.0428</b>	0.0850	0.0657
	SAT3D (ours)	0.0432	<b>0.0816</b>	<b>0.0650</b>
AP	Preim3D [4]	<b>0.0264</b>	<b>0.0510</b>	0.0400
	SAT3D (ours)	0.0315	0.0545	<b>0.0398</b>

**Table 5: Quantitative comparisons of 2D attribute transfer.**

Metrics	Method	Smiling	Beard	Eyeglasses
AS	InterfaceGAN [8]	<b>0.0319</b>	0.0829	0.0437
	StyleCLIP [5]	0.0344	0.0940	0.0623
	SAT3D (ours)	0.0323	<b>0.0690</b>	<b>0.0419</b>
AP	InterfaceGAN [8]	0.0349	0.0463	0.0657
	StyleCLIP [5]	<b>0.0335</b>	<b>0.0428</b>	<b>0.0566</b>
	SAT3D (ours)	0.0391	<b>0.0428</b>	0.0575

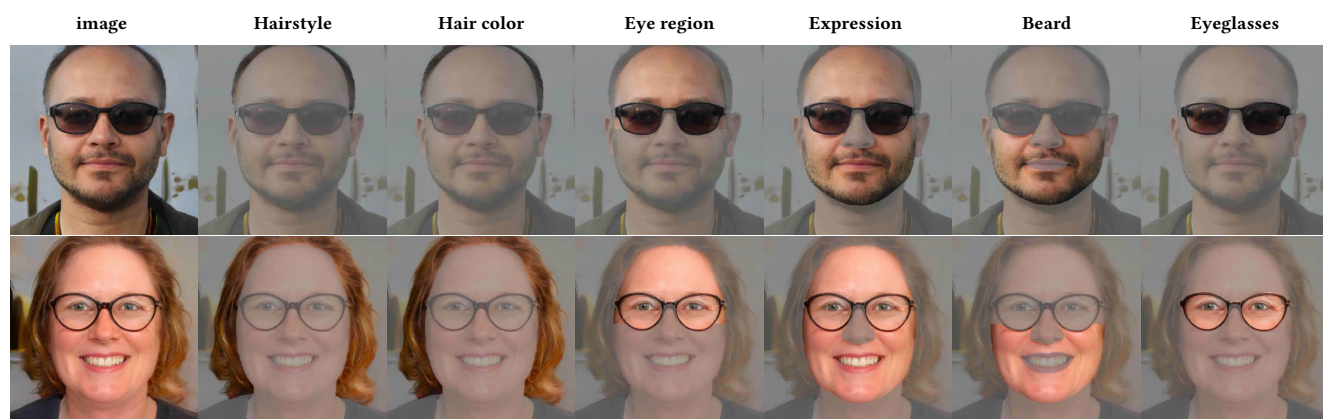


Figure 1: Region segmentation for the semantic attribute Fur of human faces.



Figure 2: Region segmentation for the semantic attribute Fur of cats.

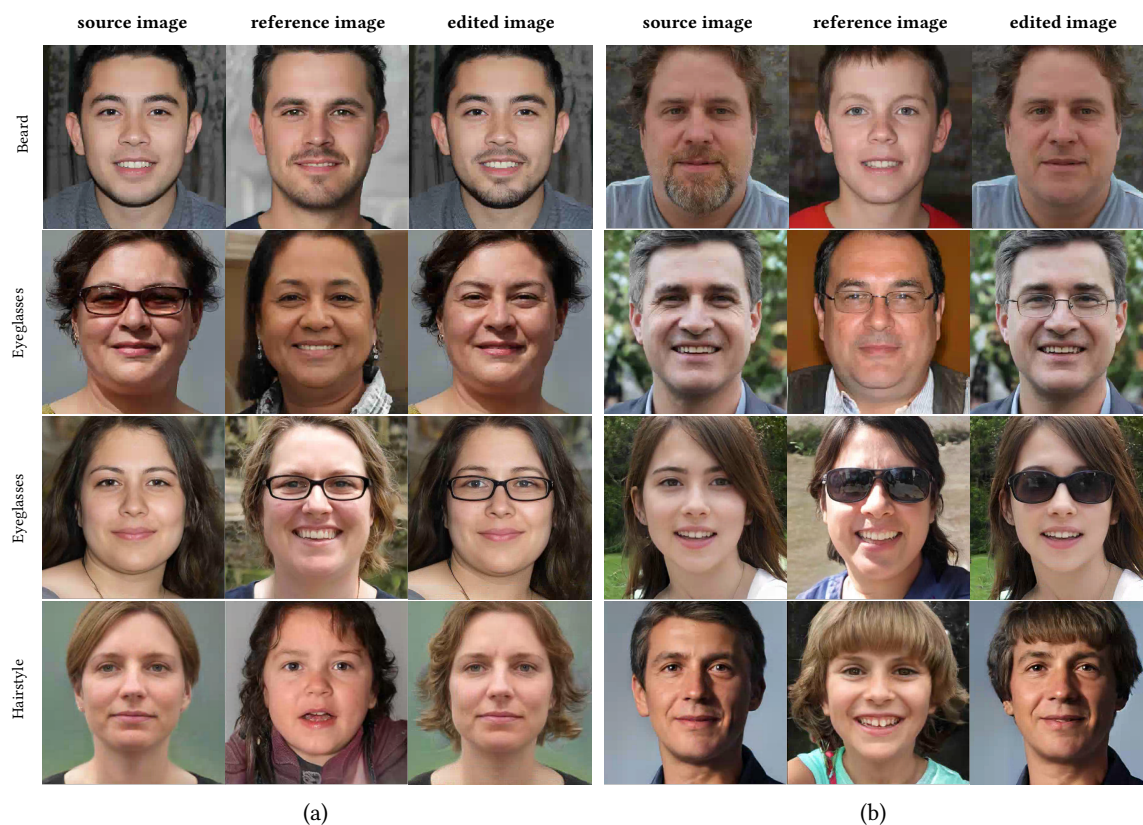


Figure 3: Visualization of attribute transfer on the EG3D generator pre-trained on FFHQ dataset.



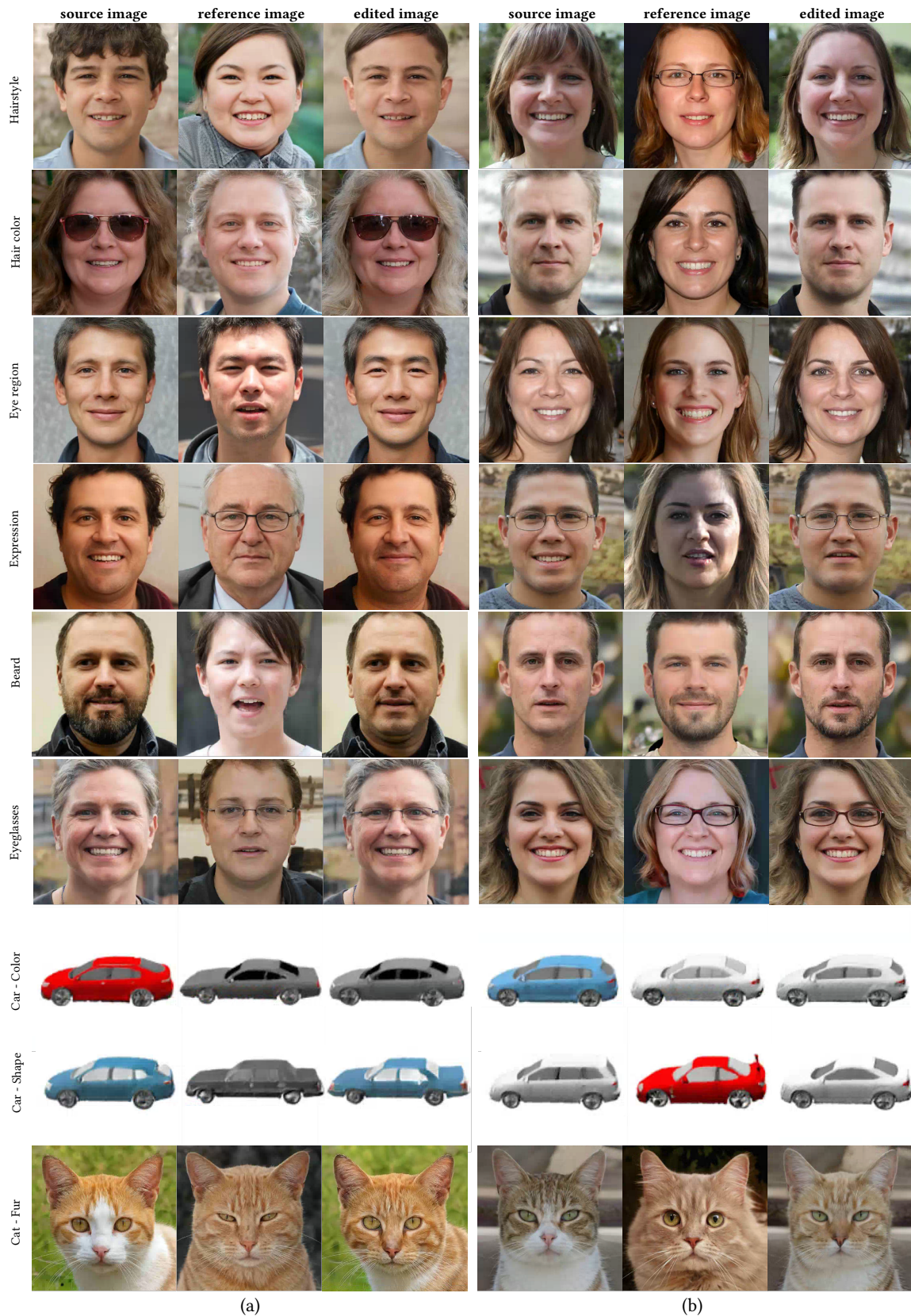
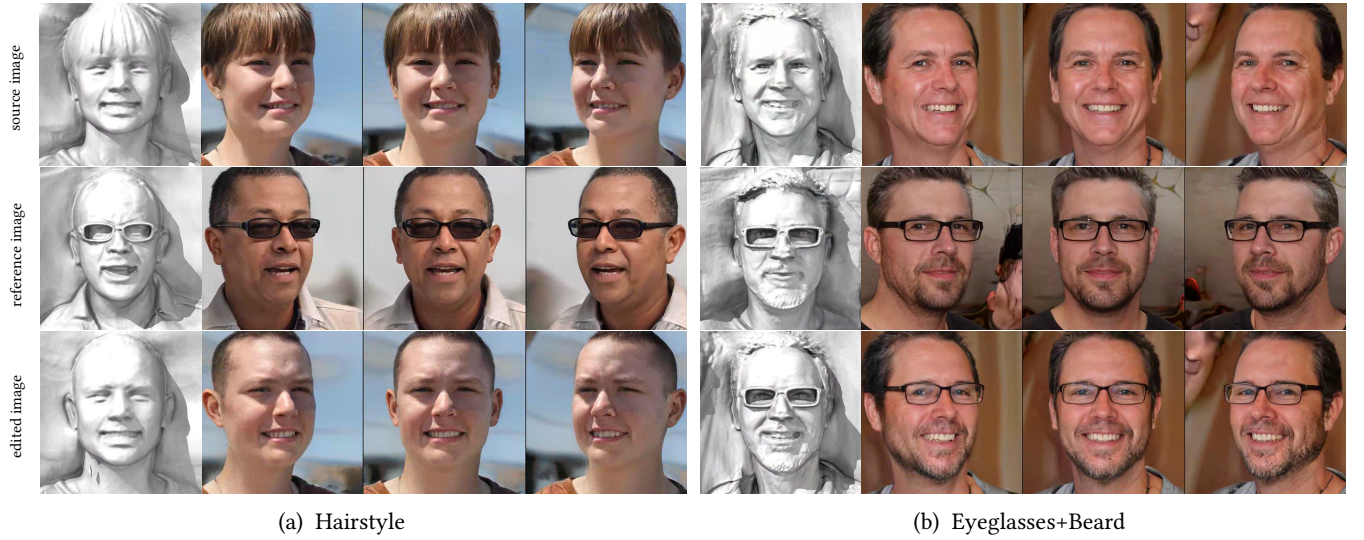
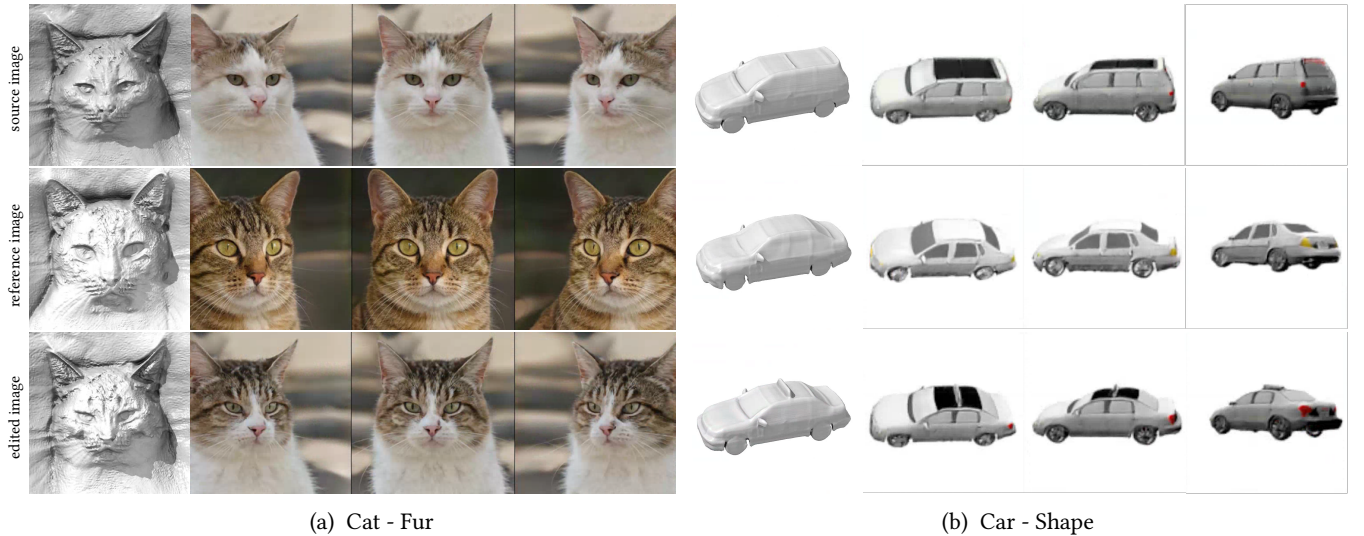


Figure 4: Supplementary attribute transfer results of SAT3D on EG3D generator.





**Figure 5: Multi-view visualization of attribute transfer on pre-trained facial EG3D generator. The source and reference images are sampled from the style space. For attribute transfer with real images, both the source and reference images require only one arbitrary viewpoint, which can be inverted into the latent space by pivotal tuning inversion (PTI) [7] or pseudo-multi-view optimized HFGI3D [9].**



**Figure 6: Multi-view visualization of attribute transfer results on cats and cars.**

## 2 EXPERIMENTAL RESULTS

**Quantitative results.** Quantitative evaluations are conducted on 3D-aware and 2D semantic attribute transfer respectively. For evaluation, we define descriptor groups for the 40 attribute categories of classifiers in [2] and utilize the zero-shot prediction capability of CLIP [6] to provide quantitative measurement, which provides more comprehensive description than binary classifiers. The attribute transfer and preservation losses defined in main article are applied on the 40 attributes as Attribute Similarity (AS) and Attribute Preserving (AP) metrics, i.e.,  $AS = \mathcal{L}_{ref}$ ,  $AP = \mathcal{L}_{src}$ .

We perform evaluation on three common attributes "Smiling", "Beard" and "Eyeglasses". For each attribute, we generate attribute editing results on the testing set of CelebAMask-HQ [3] dataset for all methods. Specifically, "Smiling" editing is applied on 1737 images that are classified as not smiling, "Beard" editing is applied on 925 images of men, and "Eyeglasses" editing is applied on all of the 2824 images. With a reference image selected for each attribute, we calculate the AS and AP metrics to measure the target attribute similarity of each edited-reference image pair and the irrelevant attribute similarity of each edited-source image pair respectively.





Figure 7: Influence of editing intensity  $\delta$ .

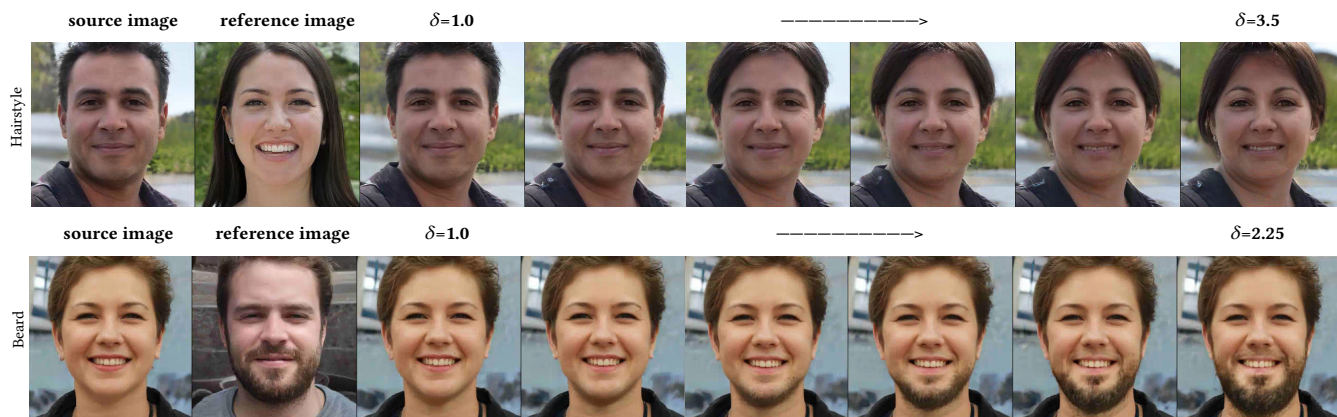


Figure 8: Typical failure cases on attribute transfer.

For 3D-aware evaluation, the metric values are averaged over 5 views.

As displayed in Table 4 and Table 5, with comparable AP performance, the generated "Beard" and "Eyeglasses" of our SAT3D are more similar to the reference image with lower AS, demonstrating the customizability of SAT3D. The "Beard" and "Eyeglasses" attributes have multifaceted characteristics, e.g., the style and sparseness of "Beard", and the shape and rims of "Eyeglasses". However, for "Smiling", the metrics can only measure the magnitude of the smile. Because of the relatively homogeneous characteristic of "Smiling", the editing results of all methods have similar features, and the advantages of our approach are suppressed.

Notably, although our evaluation metrics have more descriptive dimensions relative to traditional binary classifiers, they still cannot fully characterize the visual features of attributes, which are not suitable for the novel image-driven semantic attribute transfer task. Structural similarity evaluation is more reasonable, but there is no such quantitative evaluation method available yet. Therefore, our experimental results are mainly presented in images, which can be more directly to indicate the effectiveness of our proposed method. **Qualitative results.** For background loss, we perform image segmentation using BiSeNet [10] for human faces and DeepLabv3 [1] for cats. The segmented regions for semantic attributes are exemplified in Figure 1 and Figure 2. Supplementary attribute transfer results of SAT3D are displayed in Figure 3 and Figure 4. Besides, in Figure 5 and Figure 6, we further provide multi-view visualization for different domains.

**Influence of editing intensity.** We provide more examples in Figure 7 to visualize the influence of editing intensity  $\delta$  on attribute transfer. The proper value of  $\delta$  varies for different source-reference image pairs, roughly within the range [1.0, 2.25]. In practice, we can generate multiple edited images of intensities within this range and select the optimal one.

### 3 LIMITATION

For completeness, we provide some typical failure cases in Figure 8. The limitations mainly originate from two issues: the latent space of the pre-trained generator is insufficiently disentangled on current attributes; the disparity between the source and reference image is overly significant, resulting in large distance in latent space, making it difficult to fully migrate target attributes without disturbing others or sacrificing the quality of the generated image. Taking Hairstyle as example, which is the most complicated attribute, the key obstacles in transferring are the variations across gender and hair length. For Beard, the major challenge is to find potential latent codes for female without distortion.

### REFERENCES

- [1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).
- [2] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.
- [3] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5549–5558.

- [4] Jianhui Li, Jianmin Li, Haoji Zhang, Shilong Liu, Zhengyi Wang, Zihao Xiao, Kaiwen Zheng, and Jun Zhu. 2023. Preim3d: 3d consistent precise image attribute editing from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8549–8558.
- [5] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
- [7] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics* 42, 1 (2022), 1–13.
- [8] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 2004–2018.
- [9] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. 2023. High-fidelity 3D GAN Inversion by Pseudo-multi-view Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 321–331.
- [10] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision*. 325–341.