# Supplementary Materials

## 1 Experiment Details

**Data Preparation.** On FFHQ [2], we synthesize 500k images and use pre-trained CelebA classifiers provided by [2] to generate the pseudo labels on age, gender, eyeglasses and smile. On the other datasets, there are no existing attribute prediction models. Therefore, we synthesize 10k images and manually select 30 examples for each of the chosen key attributes presented in the paper, such as car colors, car rotations and cat colors.
**Model Implementation.** For each attribute, we train a classifier consisting of 2 fully-connected layers with hidden layer size 16 and ReLU as the activation function. We use the Adam optimizer with the default parameters, and use the model with the highest evaluation accuracy from the latest epoch.
**Interpolation Settings.** For each $W$ latent code corresponding to each test image, starting from its original logits of the target $l_k^0$, we generate a new image from the optimized latent code at each step, and obtain new logits $l_k^{'}$ for the target with the pretrained attribute classifier, until the amount of change in logits for the target $\Delta l_k = |l_k^{'} - l_k^0|$ is equal to $1.1 * |l_k^0|$, with tolerance of $0.05 * \sigma_k$, where $\sigma_k$ stands for the standard deviation of semantic $k$ logits in our image bank.

We use a fixed step size of 0.6 for our method and InterFaceGAN [4], and when the latent code overshoots, we go back to the previous position in the latent space and the step size is multiplied by 0.8. For GANSpace [1] and StyleSpace [5], we use the same bisection algorithm following [5], and the maximal/minimal manipulation strength are set to 10 and -10. For all methods, we set the maximal optimization step to 20.
**Accuracy and Attribute Dependency (AD) Calculation.** The full algorithm is described as follows.

- For the 3k images optimized by each method, we employ the pre-trained CelebA classifiers to score all 4 attributes. By thresholding the logits value, we obtain the labels for all attributes in the modified images. We then compare them to the original labels to compute the accuracy.
- For each set of images with target attribute $k \in A$ being manipulated, where $A$ stands for all attributes, we calculate the amount of change in the target attribute logits, normalized by the target population standard deviation with $x = \frac{\Delta l_k}{\sigma_k}$.
- We then group the images with their $x$ values in intervals with lengths of 0.2 (age and eyeglasses) or 0.1(gender and smile), starting from the smallest $x$ value, and ignore groups with less than 10 samples.
- Finally, for each group, we calculate mean-AD with $E(\frac{1}{|A|-1} \Sigma_{i \in A \setminus k} \frac{\Delta l_i}{\sigma l_i})$.

**Disentanglement Details.** We list the details of channel exclusion for our method in (Tab. 1).

Table 1: Channel exclusion details for manipulating images generated by StyleGAN2 [3] on FFHQ [2]. For each target attribute in the Target column, union of top channels for some/ all of the other attributes.

| Target | Gender | Smile | Eyeglasses | Age | Layers |
|---|---|---|---|---|---|
| Gender | x | 250 | 100 | 100 | 2-10 |
| Smile | x | x | x | x | 2-8 |
| Eyeglasses | 200 | 100 | x | 250 | 1-4 |
| Age | 200 | x | x | x | 4-10 |

## 2 Additional Ablation Studies

In this section, we present the effects of different training data sizes and the number of attributes excluded for disentanglement quantitatively.

**Number of Training Samples.** In (Tab. 2) we compare the accuracy of our method trained on different dataset sizes. The performance of our method can be further improved if trained on larger datasets.

Table 2: Attribute manipulation accuracy in a disentangled manner for our method trained on different numbers of samples for each attribute. A true positive should only have the target attribute changed.

| Accuracy | Gender | Smile | Eyeglasses | Age |
|---|---|---|---|---|
| Ours(1000) | 0.7204 | **0.9378** | **0.7197** | **0.6532** |
| Ours(100) | **0.7470** | 0.9253 | 0.6830 | 0.6089 |
| Ours(30) | 0.6937 | 0.9254 | 0.6626 | 0.6139 |

**Different Levels of Disentanglement** In section 4.3, we present a qualitative example of how excluding different attributes affects disentanglement for the gender attribute. In (Fig. 1) we evaluate it quantitatively and plot mean-AD for the 4 different choices. In particular, we only modify the gender attribute for the same 3k images used in our quantitative experiments, and compute the mean-AD with the same procedure.

## 3 Additional Qualitative Results

In this section, we present more results on the FFHQ, LSUN car and LSUN cat datasets in (Fig. 2, Fig. 4). We also show more real image editing results in (Fig. 3).

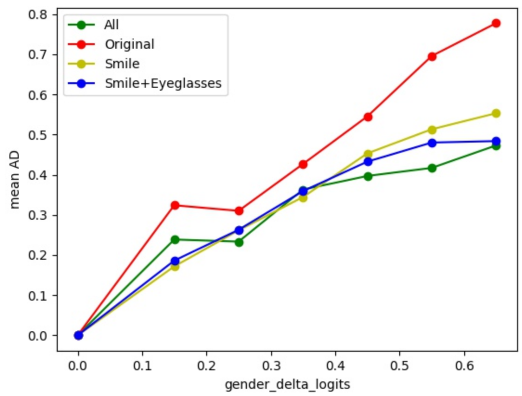



Fig. 1: Mean-AD for editing gender with top channels for predicting different attributes excluded. When all 3 attributes (age, smile, eyeglasses) are excluded (green), we achieve the lowest mean-AD as change in the amount of logits increases, i.e., the best disentanglement.

## 4 Effects of Per-Layer Editing

Multiple previous works [1,2] have shown that different style layers in StyleGAN [2] control different levels of semantics, e.g., early layers control attributes that affect the image globally, such as the orientation of an object, whereas later layers control more fine-grained details, such as the tone of the image. In (Fig. 5) we inspect the effects of interpolating the style vector in different layers. We notice that the major meaningful changes happen when interpolating the top & middle layers, whereas the bottom layers only change irrelevant details like the overall color tone.

## 5 Limitation

Despite the overall success of our method in achieving versatile controls, we notice some failure cases with unrealistic results and identity changes as shown in (Fig. 6). For LSUN cars rotation, the process sometimes lacks 3D understanding, as the gradient tends to point in the direction that a minimal amount of change is required to achieve the target. For FFHQ, control for the identity attribute is challenging as it has high correlations with multiple attributes, such as age. Nevertheless, as the control for each labelled attribute learned by our method covers some part of the identity attribute, in general conditioning on multiple attributes should preserve the identity properly, such as adjusting the smile direction by excluding the union of top 100 channels for the rest.


Fig. 2: Additional StyleGAN2 [3] FFHQ [2] editing results in a disentangled manner.

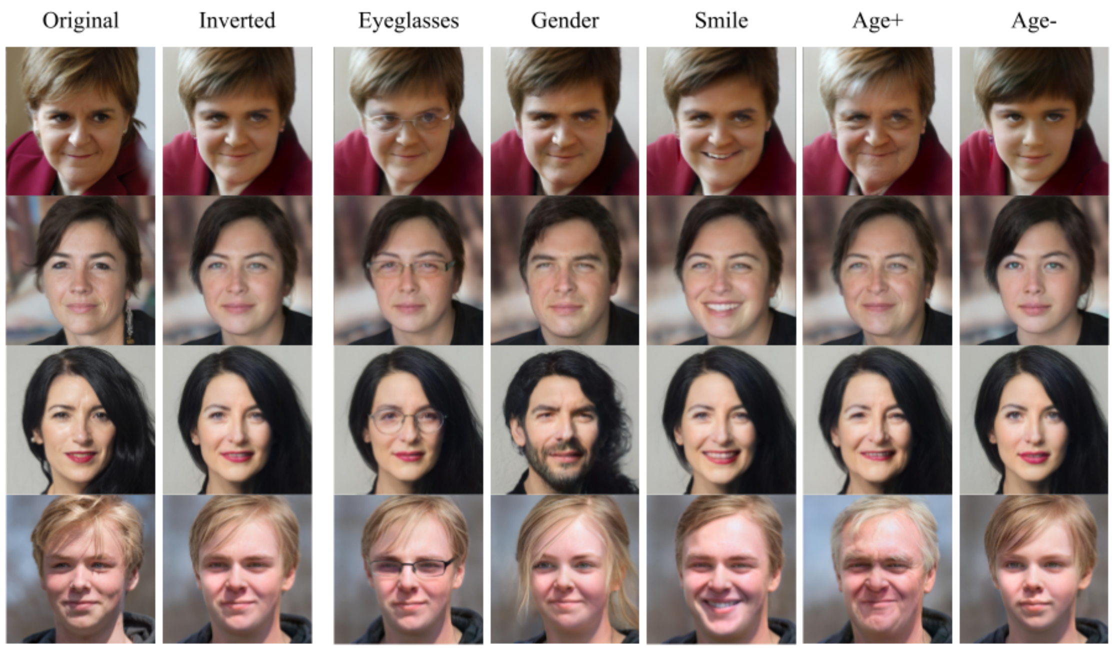
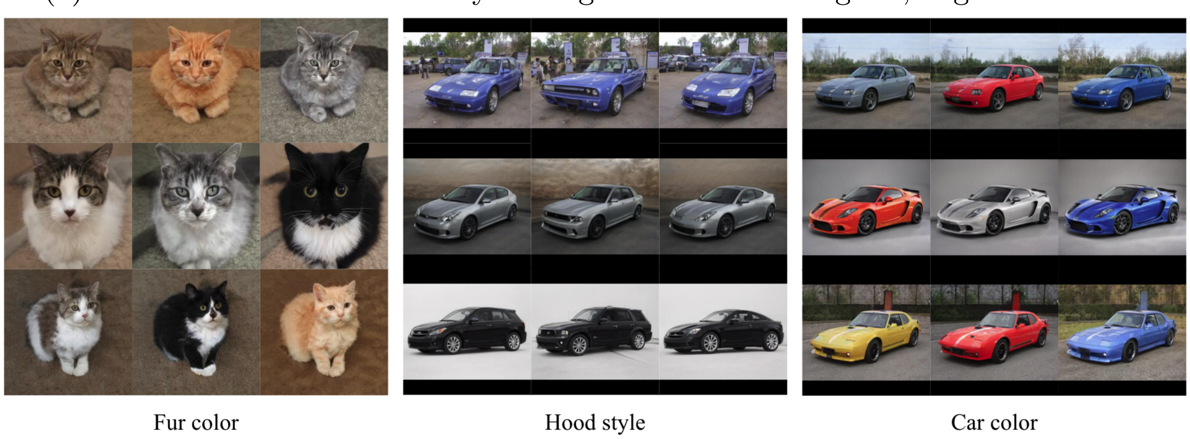



Fig. 3: Additional real image editing results.


(a) Additional LSUN car binary editing results. Left: Original, Right: Modified.




(b) Additional LSUN car & LSUN cat multiclass editing results. Left: Original, Middle & Right: Modified

Fig. 4: Additional StyleGAN2 [3] LSUN [6] editing results.

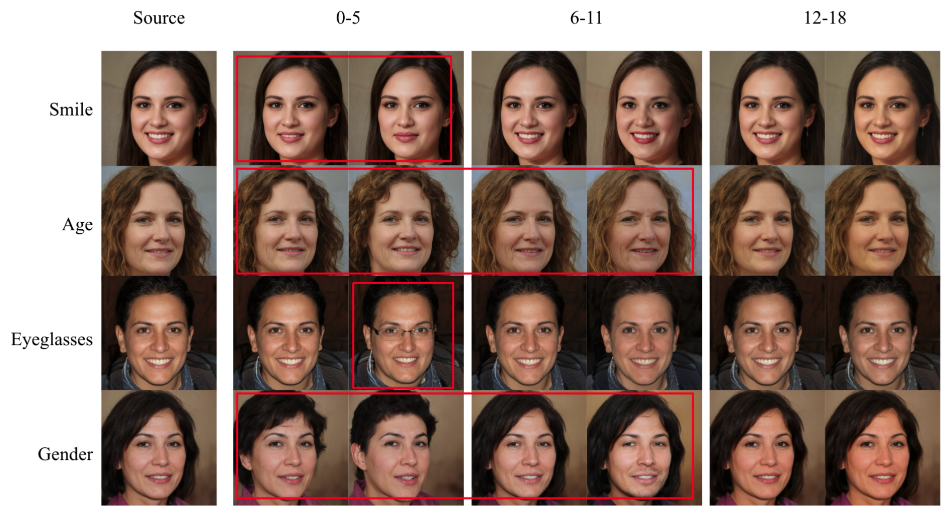



Fig. 5: Editing results for modifying the top(0-5)/middle(6-11)/bottom(12-18) style layers in StyleGAN2 [3] using our control directions. Meaningful changes are indicated in the red rectangles.

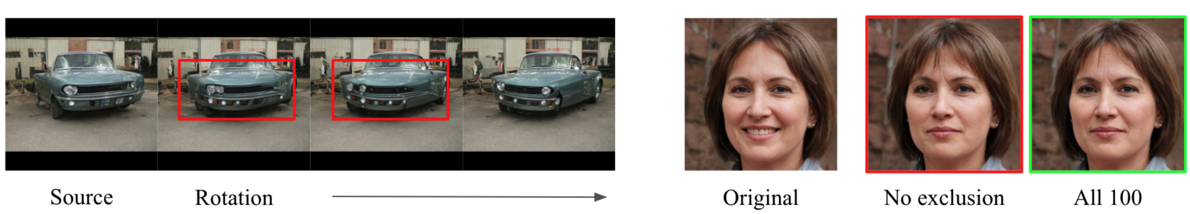



Fig. 6: Left: Failure case for rotation. Right: Identity change when removing smile is less obvious in the disentangled direction (green).

# References

1. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. Advances in Neural Information Processing Systems **33**, 9841–9850 (2020) 1, 3
2. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 1, 2, 3, 4
3. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020) 2, 4, 5, 6
4. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE transactions on pattern analysis and machine intelligence (2020) 1
5. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12863–12872 (2021) 1
6. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015) 5