Diversity-aware Channel Pruning for StyleGAN Compression

Supplementary Material

1. Additional Results

Details about the datasets. We utilize three popular datasets in the research field of GAN: FFHQ, LSUN Church, LSUN Horse. FFHQ [4] contains $70,000$ human facial images with a resolution of 1024×1024, and we also test on its resized version with a resolution of 256×256. LSUN Church [10] comprises $126,227$ outdoor church images with a resolution of 256×256, and LSUN Horse [10] includes 2 million horse images with a resolution of 256×256. However, in alignment with the experimental setting of StyleGAN2, we only utilize a subset of 1 million images from the LSUN Horse dataset.

Image editing results of our compressed generator. We perform various real-world tasks, including style mixing, style interpolation, and latent editing using GANSpace [2] and StyleCLIP [8], as shown in Fig. 1.

For style mixing, we define coarse, middle, and fine layers as [0:2]th, [4:7]th, and [9:12]th layers, respectively, following baseline [9]. In this process, we inject each level of latent code from image B into image A. As in fine layer injection, our model successfully transplants the tone from image B while preserving the identity of image A.

For style interpolation, we perform linear interpolation between the inverted latent codes (w_A, w_B) to generate style-interpolated images; $w_{\text{interp}} = w_A \times (1 - \beta) + w_B \times \beta, \beta \in [0, 1]$. StyleKD shows a glasses artifact in the interpolated image despite neither image A nor B wearing the glasses. Another baseline, CAGAN shows low-quality inversion and interpolated images. In contrast, our model shows high-quality editing results that are similar to the teacher model.

For editing via GANSpace, we identify the important latent directions using PCA on 50,000 randomly sampled w. Then, we edit the inverted real-world images using the computed latent directions. As a result, we found 2th, 4th, and 9th latent direction captures the attributes “Turn left”, “Young”, and “Glasses”, and they are successfully applicable for editing in the compressed generator. Furthermore, we also confirmed that these directions are shared their semantic changes regardless of latent code w.

Furthermore, we validate the suitability of the proposed method in real-world applications, text-driven image editing. For experiments, we adopt the StyleCLIP [8] as editing method. As shown in Fig. 1, we observe that the compressed generator successfully works for both tasks, inverting and editing on the given real-world images and various input text prompts. These experimental results confirm the generative capability of our compression technique and provide strong evidence for the practical applicability of our compressed generator.

Table 1. **Actual speed comparison with teacher model.** Our compressed model significantly accelerates inference speed compared to the teacher model.

Inference Time (ms)	Teacher	Ours
FFHQ-256	12.90	5.48 (2.35x)
FFHQ-1024	45.23	12.03 (3.76x)

Actual speed gains. We measure the inference time per image for synthesizing 1,000 images (with a batch size of 4) on a single RTX 3070 GPU. The results are presented in the Tab. 1. In real-world scenarios, our pruning model ($p_r = 0.7$) achieves a substantial speedup of 2.35x faster in 256 resolution and an impressive 3.76x faster in 1024 resolution compared to the teacher model. We did not include the inference speeds of other baselines since the number of parameters and FLOPs of our model is identical to theirs. These results highlight the significant computational efficiency gained through our pruning approach, showing its practical properties in real-world applications.

Projection examples of the real-world dataset. We provide the projection results for real samples from HelenSet55 [7]. Note that, these real-world images are not included in our training dataset. Models trained on FFHQ-1024 datasets are used for this experiment. As shown in Fig. 2, we verify that our compressed models are able to capture sufficient information about the real samples to reconstruct them.

Additional comparison for sample diversity. Diversity refers to the generator’s capacity to produce varied images. To assess this, we begin by sampling 5,000 images from identical random latent vectors z for each model trained by FFHQ-256 dataset. Next, we select a reference sample and identify its nearest neighbors among the other sampled images, by measuring the L2 distance between these images. A smaller L2 distance implies the generator is producing similar images, whereas a larger value signifies diverse image generation. This process is repeated for all 5,000 samples, enabling us to provide not only the minimum distance but also the average and maximum distances for a comprehensive assessment of diversity.

Our model shows the largest distance between samples among compressed GANs, as shown in Tab. 2. It implies that our model generates more diverse samples from distinct latent vectors.

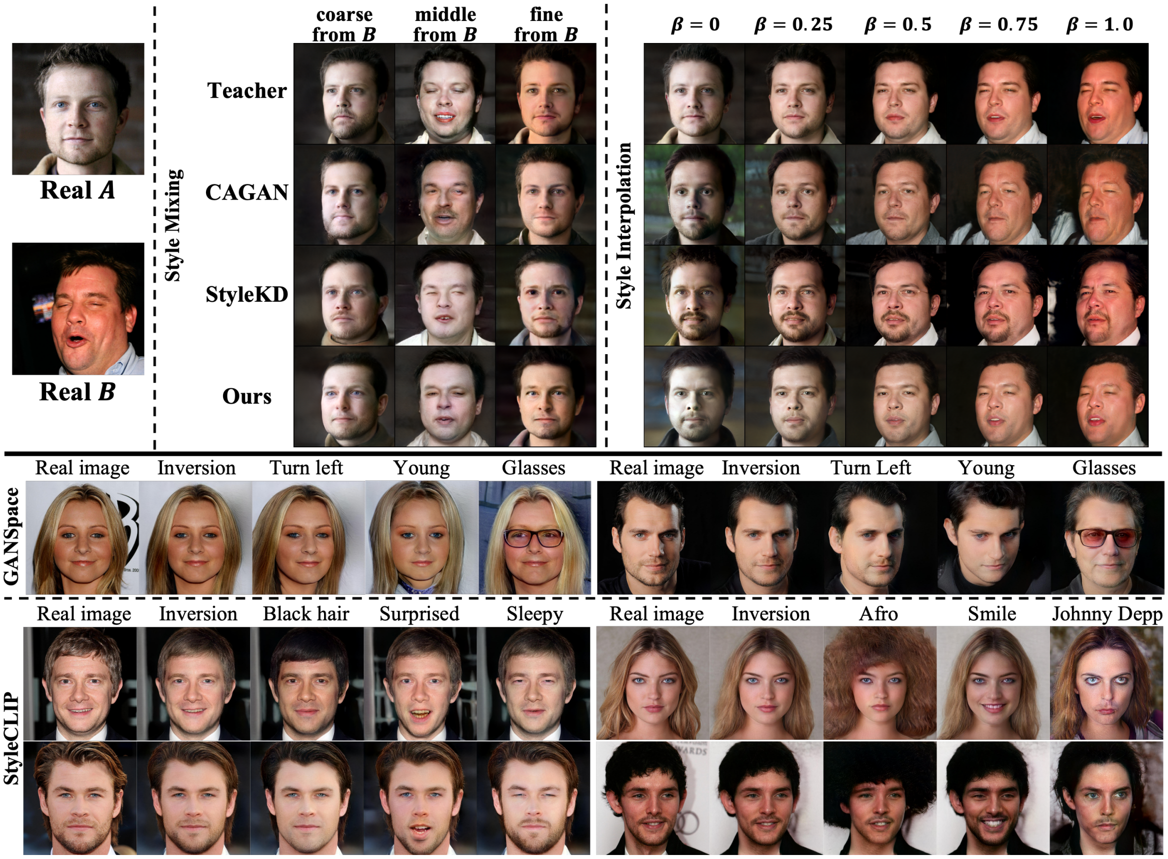


Figure 1. **Real image editing results by our model**. (Upper part) middle: style mixing, right: interpolation, and (Lower part) GANSpace and StyleCLIP examples. These results validate that our compressed GAN can be applied to real-world tasks.

Table 2. **Minimum and average L2 average distance of each generated image between the other generated images.** Please refer to the detailed calculations within the paragraph.

	Minimum	Average
Teacher	0.0449	0.1321
Ours	**0.0437**	**0.1313**
StyleKD	0.0427	0.1303
CAGAN	0.0424	0.1304

Additional generated samples from same noise input. We provide an additional visual comparison between the proposed method and baselines [7, 9]. Specifically, we synthesize the samples from the identical noise input for every methods on four datasets; FFHQ-256, FFHQ-1024, LSUN Church, and LSUN Horse. As a result, the generated samples from ours are more visually similar to samples from the teacher model compared to baselines, and this result is achieved consistently regardless of the type of dataset and its resolution, as shown in Fig. 3 and Fig. 4. Our method exhibits a high similarity to the teacher-generated images compared to the baselines. As a result, our model demonstrates an enhanced capability to preserve sample diversity of the teacher model.

2. Further Analysis

Visual transition as strength of perturbation changes. We visually analyze the generated samples along with their perturbed counterparts generated from various strength parameters of the perturbations (α). As shown in Fig. 6, selecting $\alpha = 5, 10$ yields the perturbed samples with a sufficient magnitude of pixel-level changes in images. This observation aligns with our ablation study ("Ablation" section in main manuscript), which demonstrates that $\alpha = 5, 10$ is an adequate strength for perturbations as supported by $\text{FID}_{\text{early}}$ metric. Specifically, our ablation study denotes that the $\text{FID}_{\text{early}}$ metric values for $\alpha = 5, 10$ (12.08 and 12.09, respectively) are lower than $\alpha = 1$ (13.50). Therefore, when selecting the strength of perturbation changes, it is crucial to ensure that there are significant visual transitions in the perturbed samples to effectively capture the effects of latent perturbations.

More experiments in other network structure (Fast-

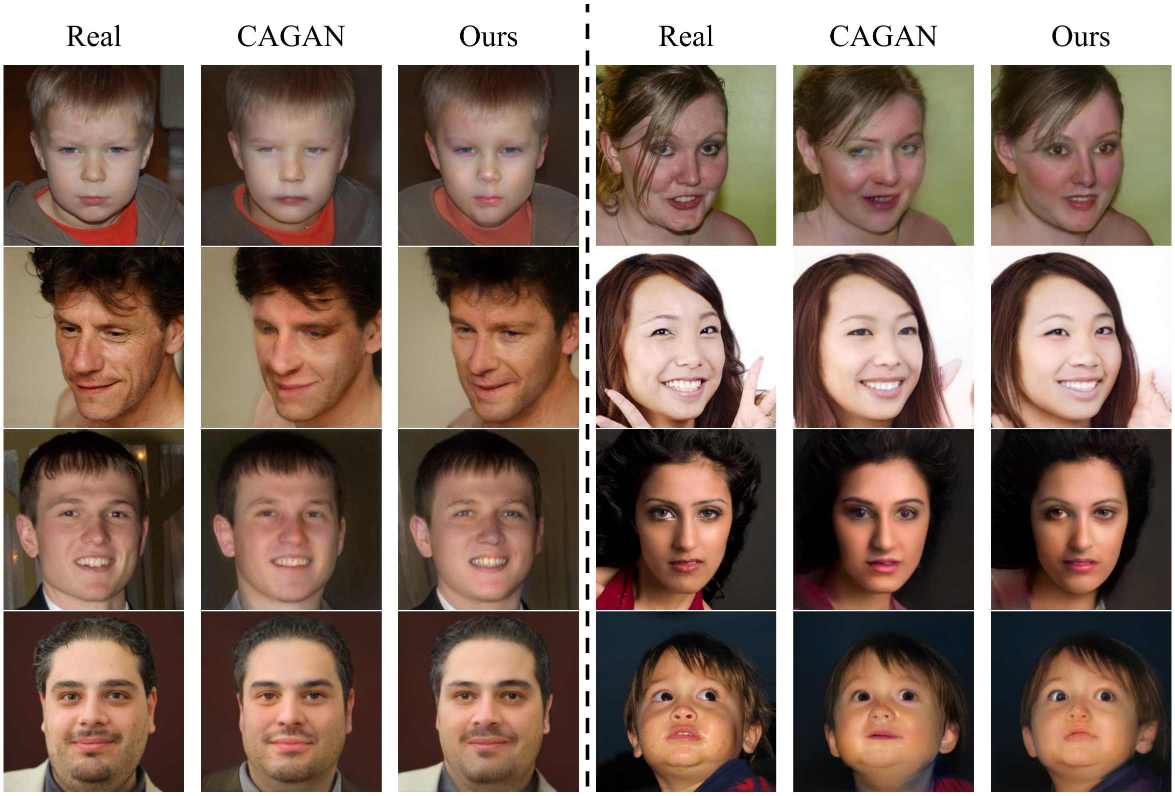


Figure 2. **Projection examples of the real-world dataset.** We visualize the real-world images from Helen-Set55 [7] and its projected results. We use the generators trained on FFHQ-1024 datasets. The visual similarity between the real image and the projected one shows that our compressed models have sufficient capability to express real samples. Note that, these real images are not included in our training dataset.

Table 3. Comparison on FastGAN ($p_r = 0.5$, 60K iters, 3 times)

FID ↓	Teacher	Scratch	StyleKD	Ours
Dog (256)	52.05 ± 0.2	56.34 ± 0.3	55.10 ± 0.7	**53.17 ± 0.3**

Table 4. Ablation study of the image difference metric (220K iters)

Dataset	L1	LPIPS	Dataset	L1	LPIPS
FFHQ-256	**7.05**	7.32	Horse-256	6.49	**6.18**

GAN). We conduct the other GAN architecture, FastGAN [6], distinct from the structure of StyleGAN. As shown in Tab. 3, proposed method shows superior performance even in a different network structure, validating the generalizability of the proposed method.

Utilization of LPIPS for capturing image difference. To investigate the effects of distance measure, we additionally conduct experiment that prunes channels with LPIPS as image difference. As reported in Tab. 4, we observe that two distance measures (L1, LPIPS) perform similarly. We hypothesize that semantic perturbation we used (i.e. PCA directions) already encourages the model to be aware of semantics, although guidance consists of pixel-level L1 distance.

Table 5. Quantitative comparisons with baselines ($p_r = 0.5$, 100K iters, FFHQ-256 dataset)

	Ours	StyleKD	CAGAN
FID ↓	**6.78**	8.79	11.68

More examples for pruned and not-pruned channels between the S^μ and S^σ scores. In Fig. 7, we further visualize channels in 5th layer following the experimental settings of Fig. 5 in main manuscript. The 214th channel exhibits low S^μ and high S^σ values with strong activation for 16th direction, associating with hair length.

Comparison with a different pruning ratio. We train ours and baseline methods up to 100K iters with $p_r = 0.5$. The proposed method demonstrates superior performance compared to the baselines, as shown in Tab. 5.

An overview of our method's implementations. Our pruning implementation follows the stages outlined below:

1. Prepare the teacher model (f, g), along with the perturbation vector d.

2. Sample a latent vector w and its perturbed counterpart $(w + \alpha d)$.

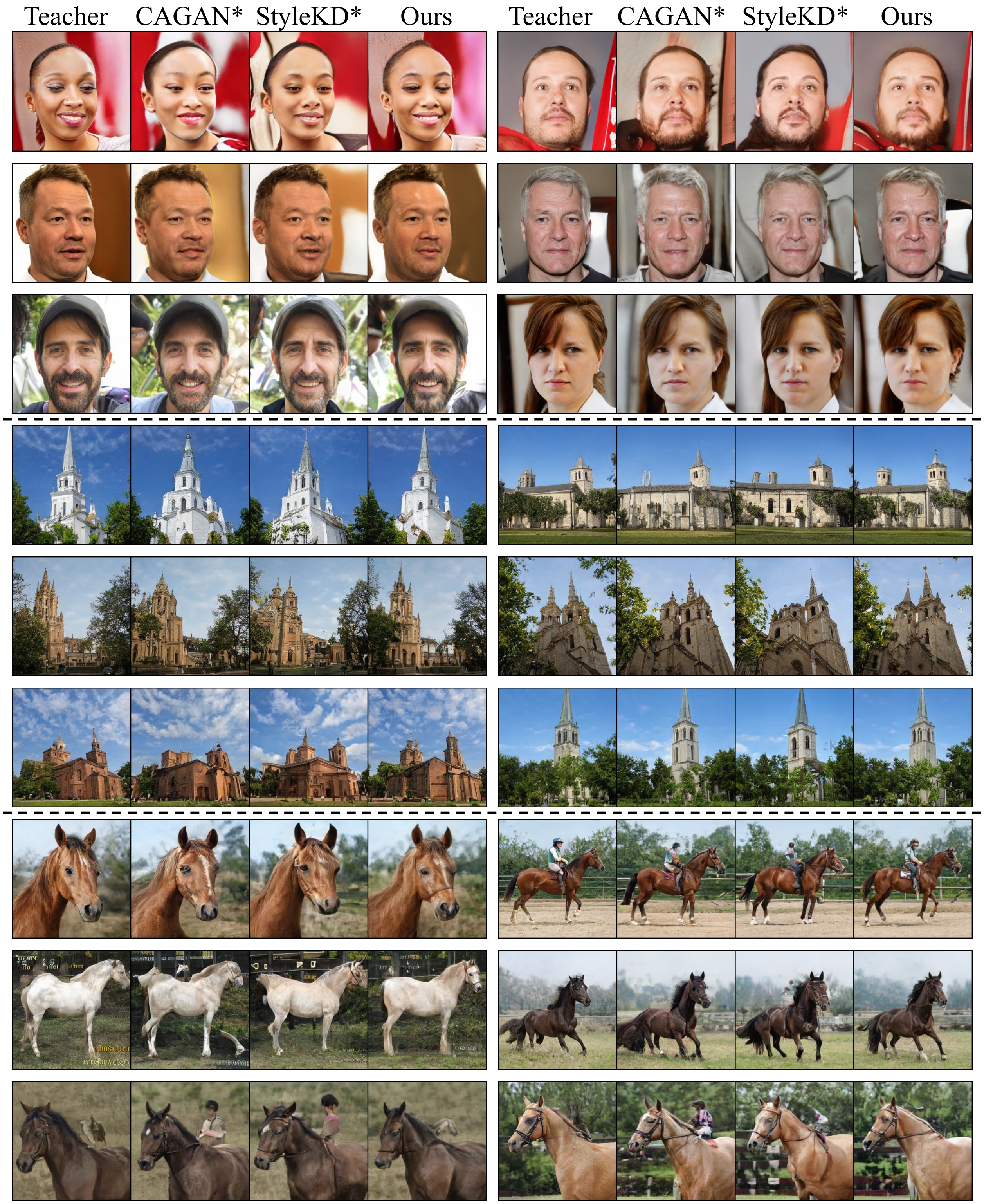


Figure 3. **Qualitative comparison with baselines on various datasets.** For comparison, we visualize the generated samples from ours and baselines [7, 9] in FFHQ-256, LSUN Church, and LSUN Horse datasets. Each half of the row corresponds to samples generated from the same noise vector z.

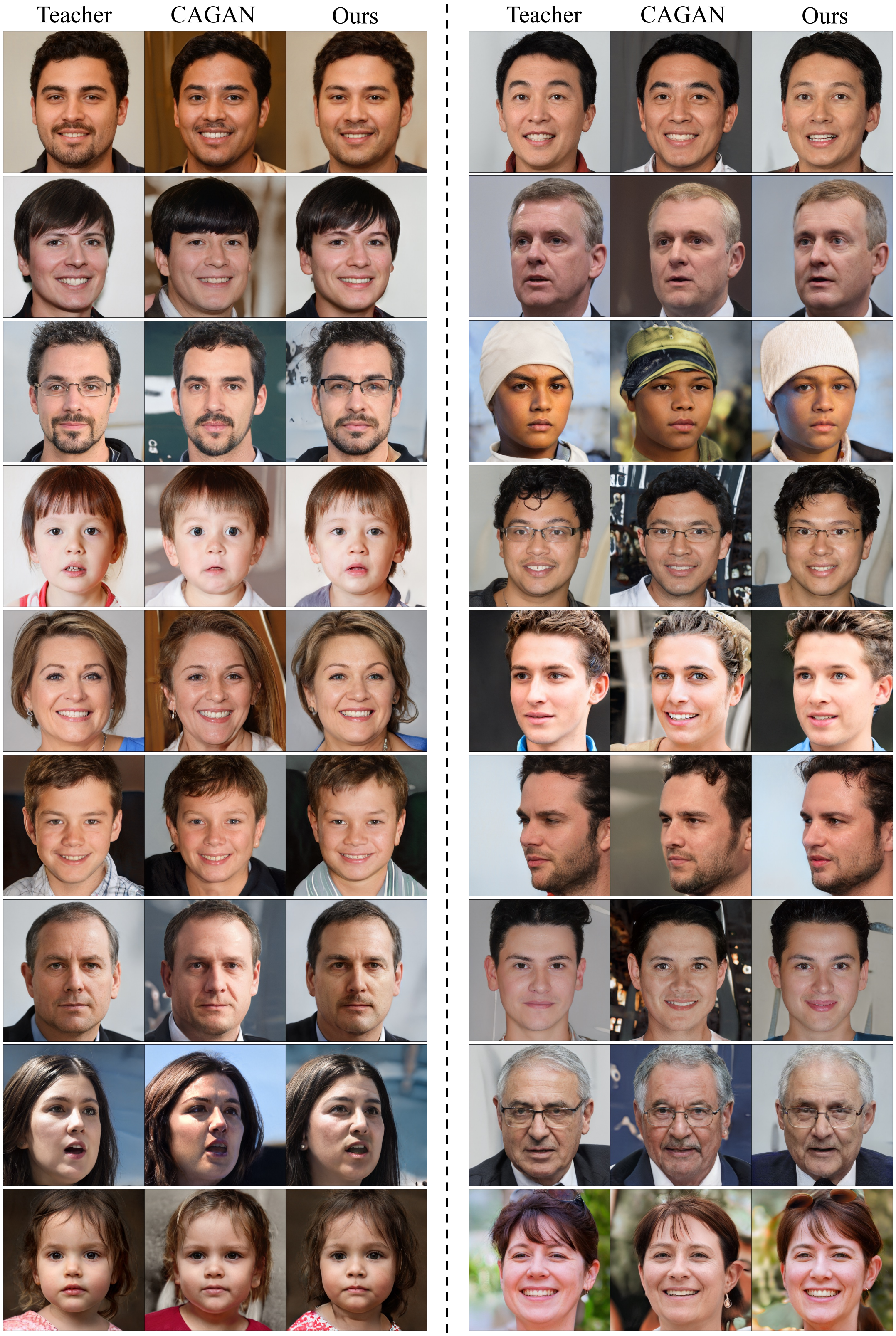


Figure 4. **Qualitative comparison with baselines on the high-resolution dataset.** For comparison, we visualize generated samples from ours and baselines [7, 9] in FFHQ-1024 dataset. Each half of the row corresponds to samples generated from the same noise vector z.

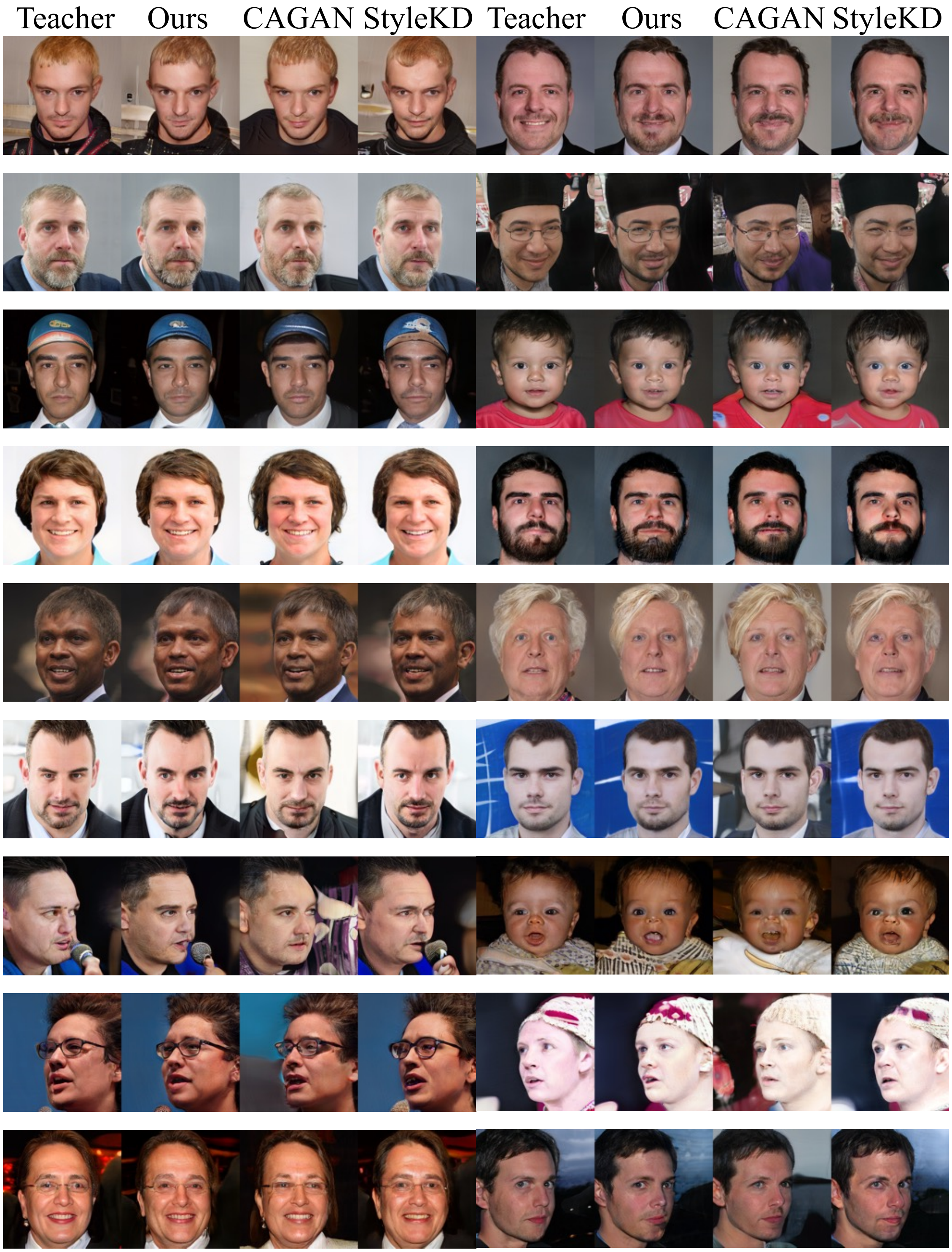


Figure 5. **Qualitative comparison with baselines on StyleGAN3.** For comparison, we visualize generated samples from ours and baselines [7, 9] on StyleGAN3-T [5] in the FFHQ-256 dataset. Each half of the row corresponds to samples generated from the same noise vector z.

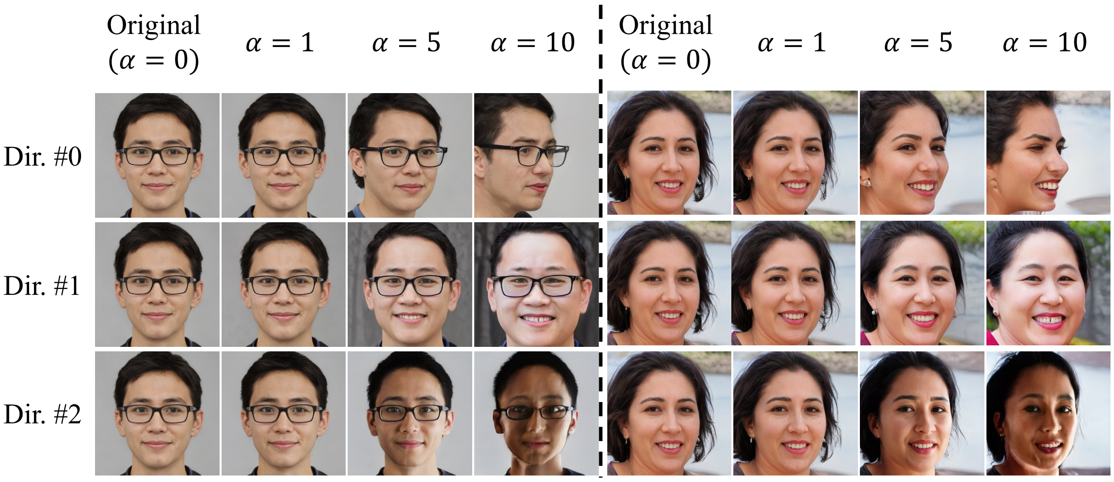
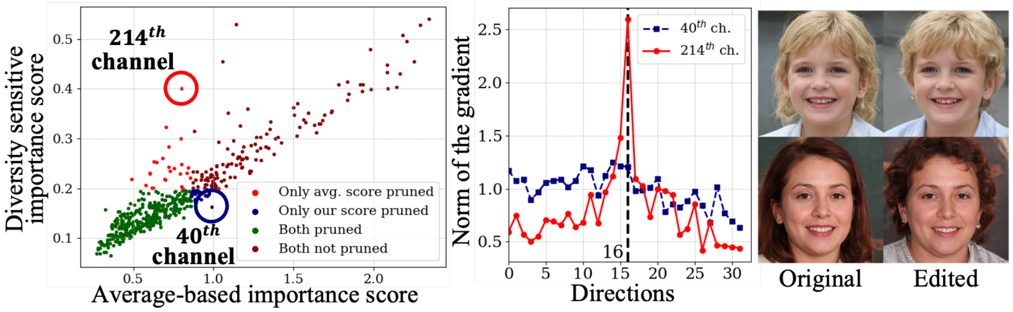


Figure 6. **Visual transition as strength of perturbation changes.** We first obtain three directional vectors (Dir.) using GANSpace [2]. Next, we generate the samples and their perturbed counterparts with the different strength parameters for the perturbations (α). When $\alpha = 1$, perturbed samples only show minor changes to identify the image difference that latent changes lead to. Conversely, when $\alpha = 5, 10$, the perturbed samples show sufficient pixel-level differences to detect the effect of latent variations. Similarly, our ablation study (Sec. 4.4 in main manuscript) validates that $\alpha = 5, 10$ are the proper strength of perturbations by achieving lower FID$_{\text{early}}$ metric. Therefore, when selecting the strength of perturbation changes, it is crucial to ensure sufficient pixel-level differences in the perturbed samples to effectively capture the effects of latent variations.



(a) Channel scatter plot (S^{μ}, S^{σ}) (b) Grad. norm of each dir. (c) Edited 16^{th} dir.

Figure 7. **Additional examples for pruned channels** (a) We provide additional scatter plot as same as Fig. 5 in main manuscript. (b) The 214th channel exhibits high sensitivity to the 16th direction from PCA. (c) The 16th direction corresponds to an hair length related perturbation. The S^{μ} score prunes the 214th channel, while the S^{σ} score preserves this channel, which demonstrates high sensitivity to the hair length variations.

3. Generate two images $g(w)$ and $g(w + \alpha d)$, and perform backpropagation from the loss $\mathcal{L}_{\text{diff}} = |g(w) - g(w + \alpha d)|$.

4. Accumulate the gradients $\mathbf{G}_{\text{perturb}}$.

5. For 1,000 iterations, repeat steps 2 to 4.

6. Calculate the diversity-sensitive importance score S^{σ} and prune channels based on this score.

Broader Impacts. In today's social media landscape, the generation of fake images of celebrities or sports stars using generative models is a major concern. Our proposed compression method not only address the computational challenges but also brings attention to the potential misuse of such techniques. To mitigate the negative impact of fake images, detection models [1, 3] offers a solution to minimize the harm caused by these fake images. It is crucial to also consider the ethical implications of such technologies and promote responsible use to prevent malicious exploitation.

Limitations. The proposed pruning method aims to preserve the sample diversity of teacher network as much as possible. Hence, the samples that can potentially disturb the effective transfer of knowledge of teacher (e.g. samples with degenerated quality) also can be preserved in the student network. This may hinder the further improvement in the generation performance of student network.

References

[1] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5012–5019, 2021. 7

[2] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 1, 7

[3] Young-Jin Heo, Woon-Ha Yeo, and Byung-Gyu Kim. Deepfake detection algorithm based on improved vision transformer. *Applied Intelligence*, 53(7):7512–7527, 2023. 7

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[5] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 6

[6] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2020. 3

[7] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Federico Perazzi, and Sun-Yuan Kung. Content-aware gan compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12156–12166, 2021. 1, 2, 3, 4, 5, 6

[8] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1

[9] Guodong Xu, Yuenan Hou, Ziwei Liu, and Chen Change Loy. Mind the gap in distilling stylegans. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 423–439. Springer, 2022. 1, 2, 4, 5, 6

[10] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1