

Mind the Gap in Distilling StyleGANs

Guodong Xu¹, Yuenan Hou², Ziwei Liu³, and Chen Change Loy³

¹ The Chinese University of Hong Kong

² Shanghai AI Laboratory

³ S-Lab, Nanyang Technological University
xg018@ie.cuhk.edu.hk, houyuenan@pjlab.org.cn,
{ziwei.liu, ccloy}@ntu.edu.sg

Abstract. StyleGAN family is one of the most popular Generative Adversarial Networks (GANs) for unconditional generation. Despite its impressive performance, its high demand on storage and computation impedes their deployment on resource-constrained devices. This paper provides a comprehensive study of distilling from the popular StyleGAN-like architecture. Our key insight is that the main challenge of StyleGAN distillation lies in the output discrepancy issue, where the teacher and student model yield different outputs given the same input latent code. Standard knowledge distillation losses typically fail under this heterogeneous distillation scenario. We conduct thorough analysis about the reasons and effects of this discrepancy issue, and identify that the mapping network plays a vital role in determining semantic information of generated images. Based on this finding, we propose a novel initialization strategy for the student model, which can ensure the output consistency to the maximum extent. To further enhance the semantic consistency between the teacher and student model, we present a latent-direction-based distillation loss that preserves the semantic relations in latent space. Extensive experiments demonstrate the effectiveness of our approach in distilling StyleGAN2 and StyleGAN3, outperforming existing GAN distillation methods by a large margin. Code is available at: <https://github.com/xuguodong03/StyleKD>

1 Introduction

GAN compression [32,22,23] has been actively studied to enable the practical deployment of powerful GAN models [16,18,19] on mobile applications and edge devices. Among these techniques, knowledge distillation (KD) [9] is a widely adopted training strategy for GAN compression. The objective of GAN distillation is to transfer the rich dark knowledge from the original model (teacher) to the compressed model (student) so as to mitigate the performance gap between these two models. There are two distillation strategies, i.e., pixel-level and distribution-level. The former minimizes the distance between generated images of two models, while the latter minimizes the distance between distributions. In this work, we focus on the first setting considering its prevalence in the GAN compression literature [6,32,22,23].

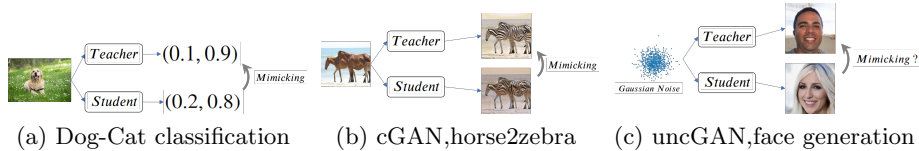


Fig. 1: Output discrepancy issue. For the classification task in (a), teacher and student naturally have similar output due to the label supervision. For the conditional GAN such as image-to-image translation in (b), teacher and student also have similar outputs because the input image imposes strong constraints on the output. However, for unconditional generation in (c), teacher and student may produce two images with totally different semantic features. In this condition, distillation is no longer meaningful and cannot bring gains to the student.

The majority of contemporary GAN distillation methods [6,32,22,23] focus on conditional GANs (cGANs), especially image-to-image translation [15,40], while the distillation of unconditional GANs (uncGANs) is relatively under-explored. Since there is a large difference between the learning dynamics of these two types of GANs, distillation methods tailored for cGANs cannot be directly applied to the unconditional setting.

We find that the main difficulty of uncGAN distillation lies in the *output discrepancy* between the teacher and student model. An example is shown in Fig. 1. In fact, the implicit prerequisite of KD is that teacher and student should have similar outputs for the same input, otherwise the mimicking supervision is no longer meaningful. This prerequisite is easier to be satisfied in most of cGANs, because the output space of cGANs can be narrowed down by the given conditional input, especially when the condition is strong [40,15]. Take the horse→zebra task as an example. An input horse image determines which region should be added with zebra stripes and which region is background that should not be changed. Two generated images in cGAN may differ in some low-level details such as the shape of zebra stripes, but would largely resemble in their structure. Unlike cGANs, as shown in our experiments, it is impossible for an uncGAN student with random initialization to learn similar mapping function to the teacher, even though we leverage distillation loss to enforce the agreement between the outputs of two models.

To study the aforementioned output discrepancy problem, we focus our attention on the StyleGAN family, e.g., StyleGAN2 [19] and StyleGAN3 [17], which is one of the most applied unconditional GANs in various downstream tasks [33,20,2]. We carefully examine each component of the StyleGAN-like student model through comparative experiments. We identify that the mapping network plays a crucial role in deciding the semantic information of the generated images. Based on this finding, we propose a simple yet effective initialization strategy for the student model, i.e., inheriting the weights from the teacher mapping network and keeping the remaining convolutional layers randomly initialized. Such initialization strategy can work well even in heterogeneous distillation

where the student architecture is obtained by neural architecture search (NAS) or manual design, and is totally different from the teacher model.

After resolving the output discrepancy problem, we further design an effective mimicking objective tailored for uncGAN distillation. As opposed to most of existing GAN distillation approaches that merely transfer the knowledge within a single image, we propose a novel latent-direction-based relational loss to fully exploit the rich relational knowledge between different images. Specifically, we exploit the good linear separability property of StyleGAN-like models in latent space and augment each latent code w by moving it along certain direction such that the resulting image only differs in a *single* semantic factor. Then, we compute the similarity matrix between original images and augmented images and take it as the dark knowledge to be mimicked by the student. The latent-direction-based augmentation disentangles various semantic factors and makes the learning of each factor easier, thus yielding better distillation performance.

Our **contributions** are summarized as follows: **1)** To the best of our knowledge, this is the first work that uncovers the *output discrepancy* issue in StyleGAN distillation. Through carefully designed comparative experiments, we identify that the mapping network is the determining factor to ensure output consistency. **2)** We propose a concise yet effective initialization strategy for the student to resolve the output discrepancy problem, demonstrating significant gains upon conventional uncGAN distillation. **3)** We further propose a latent-direction-based distillation loss to learn the rich relational knowledge between different images, and achieve state-of-the-art results in StyleGAN2/3 distillation, outperforming the existing state-of-the-art CAGAN [23] by a large margin.

2 Related Work

GAN Compression. We highlight a few recent methods among many GAN compression methods [28,5,6,32,22,10]. GAN Slimming [32] integrates model distillation, channel pruning and quantization into a unified framework. GAN Compression [22] searches a compact student architecture via NAS, and then forces the student to mimic the intermediate outputs and synthesized results of the teacher simultaneously. A common characteristic shared by these works is that they all focus on the cGANs such as pix2pixGAN [15] and CycleGAN [40].

Aguinaldo’s work [1] focuses on the uncGANs (DCGAN) distillation on low-resolution (32×32) datasets, where the easy setting makes it possible to solve the output discrepancy by adding L1 loss. Our work explores the distillation of StyleGAN-like models on high resolution (256/1024) images. In this case, output discrepancy issue becomes much more challenging. The more recent MobileStyleGAN [3] and Content-Aware GAN compression (CAGAN) [23] shift the attention to styleGANs. MobileStyleGAN compresses the model by mimicking the wavelet transformation of generated images. CAGAN estimates the contribution of each channel to the generated faces and eliminates channels with little contribution. Subsequently, the pruned model inherits the parameters from the original network for both mapping network and convolutional layers, and are

finetuned with adversarial loss and distillation loss afterwards. Though CAGAN involves the compression of uncGAN, it bypasses the issues of model heterogeneity between the teacher and student model by allowing the student to inherit the parameters. Such an requirement assumes the student to inherit the main structure of the teachers too despite pruning. As will be shown in the experiments, the performance of CAGAN greatly degrades in heterogeneous distillation. The proposed mimicking loss cannot guarantee the student to learn a similar mapping as the teacher. Moreover, we find that the content-aware pruning strategy in CAGAN is not an optimal solution for student initialization. With our proposed initialization strategy, the student model does not need to inherit any weights from convolutional layers of the teacher but achieves better results.

Knowledge Distillation. KD [9] is originally proposed to achieve model compression [4] for image classification, whose target is to transfer the dark knowledge from one or multiple cumbersome networks (teacher) to a small compact network (student). Vanilla KD [9] proposes to match the outputs of two classifiers by minimizing the KL-divergence of the softened output logits. Besides the output logits, other intermediate outputs such as feature maps [26], attention maps [38,11], Gram matrices [36], pre-activations [8], relation [25,30] and self-supervision signals [29,35] can also serve as the dark knowledge. However, it should be careful when adapting KD from classification tasks to generation tasks. The output consistency prerequisite is naturally satisfied in image classification since the supervision of labels guarantees different models to converge to similar mappings. As discussed in Sec. 1, the consistency prerequisite does not naturally hold for uncGANs. Therefore, a special distillation technique tailored for uncGANs is required to cope with the output discrepancy problem.

StyleGAN Linear Property. As shown in StyleGAN [18], for a well-trained model, the w latent space consists of linear subspaces. It should be possible to find direction vectors that consistently correspond to individual factors of variation. Recently, some works [14,24,27,31] have been conducted to find these meaningful directions. Among them, SeFa [27] finds the latent directions by computing the eigenvalues of the transformation matrix in the ModConv [13] layer. We adopt it in our latent-direction-based loss due to its fast computation and high performance. A recent work StyleAlign [34] provides a thorough analysis about the property of StyleGAN latent space. It finds that the latent directions control similar semantic factors for two aligned models even they work on very different domains. This finding aligns with our observation that the mapping network plays a vital roles in determining the semantics of generated images.

3 Methodology

3.1 Preliminaries

StyleGAN. There are two modules in StyleGAN-like models [18,19,17], i.e., a mapping network $S(\cdot)$ that maps Gaussian noise z to the style vector w and a convolution backbone $C(\cdot)$ that takes w as input and generates images. The

style vector w is fed into the backbone $C(\cdot)$ through the modulated convolution (ModConv) layer [13,19]. StyleGAN allows the use of different w vectors in different ModConv layers. The image generation process can be formulated as:

$$G(z_1, z_2, \dots, z_L) = C(w_1, w_2, \dots, w_L) = C(S(z_1), S(z_2), \dots, S(z_L)), \quad (1)$$

where L is the number of ModConv layers in the backbone, and the i -th ModConv layer uses w_i that comes from z_i . We define the output consistency condition as:

$$G_s(z_1, z_2, \dots, z_L) = G_t(z_1, z_2, \dots, z_L), \quad (2)$$

where the s and t represent student and teacher, respectively. Equation 2 suggests that the generated images of two models should be the same if they use the same z at corresponding layers.

StyleGAN Compression. A typical StyleGAN compression approach [23] contains two steps, i.e., pruning and finetuning. In the pruning stage, unimportant / unnecessary channels will be removed according to some heuristics [12,21,7,23]. Note that pruning is only applied to the convolution backbone $C(\cdot)$ and the mapping network $S(\cdot)$ is kept *unchanged*. The pruned model will inherit the well-trained weights from the original model for both the mapping network and the convolution backbone [23]. In the finetuning stage, besides the normal adversarial loss, the pruned model is also required to mimic the original model’s output to compensate the performance degradation brought by channel reduction. A typical mimicking loss includes RGB loss and LPIPS loss [39]:

$$\mathcal{L}_{\text{rgb}} = \|G_s(z) - G_t(z)\|_1, \mathcal{L}_{\text{lpiips}} = \|F(G_s(z)) - F(G_t(z))\|_1, \quad (3)$$

where F is a well-trained frozen network that computes the perceptual distance between two images. L_{rgb} and L_{lpiips} require that the generated image of student should be close to that of teacher in RGB space and perceptual space, respectively. The final loss function in the finetuning stage is:

$$\mathcal{L} = \lambda_{\text{GAN}} \mathcal{L}_{\text{GAN}} + \lambda_{\text{rgb}} \mathcal{L}_{\text{rgb}} + \lambda_{\text{lpiips}} \mathcal{L}_{\text{lpiips}}, \quad (4)$$

where λ_* is the loss weight of each item.

3.2 Framework Overview of Unconditional GAN Distillation

Knowledge distillation is a common strategy that can bring improvements in classification tasks. However, in generation tasks, its prerequisite, namely the student and teacher having consistent outputs for the same input, is rarely mentioned. In the absence of this prerequisite, the influence of mimicking losses on the training of student remains largely unknown. Here, we hypothesize that RGB or LPIPS loss is not compatible with GAN loss when the output discrepancy occurs and distillation will also bring no benefit to the student. We examine this hypothesis both qualitatively and quantitatively.

Note that the three losses in Eq. 4 serve different roles. \mathcal{L}_{GAN} requires the student to generate realistic images while \mathcal{L}_{rgb} and $\mathcal{L}_{\text{lpiips}}$ encourage similarity

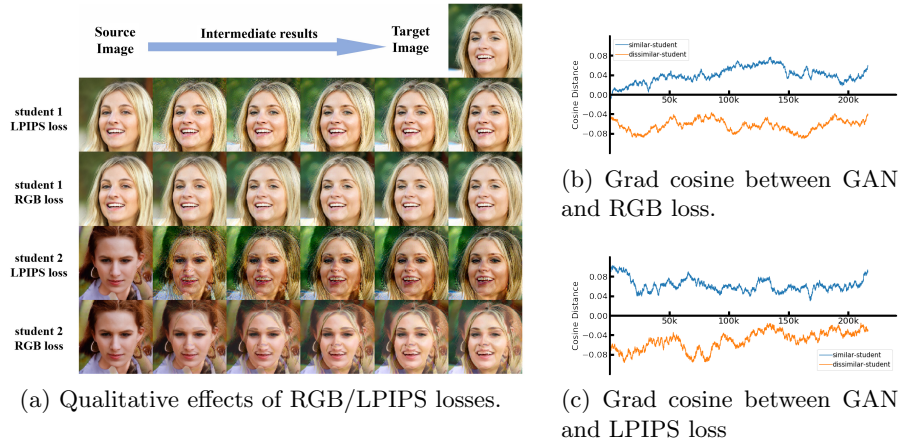


Fig. 2: (a) Student-1 has similar outputs with teacher and Student-2 has different outputs. The image in the top right corner is the teacher output. We demonstrate the intermediate results to show how RGB/LPIPS loss influences the image generation of student. (b)(c) Cosine distance between the gradient of GAN loss and RGB/LPIPS loss. The x -axis denotes training steps. For similar student, RGB/LPIPS loss is cooperating with GAN loss. For dissimilar student, RGB/LPIPS loss is competing with GAN loss.

between the generated images by student and those of the teacher. Intuitively, if the generated image of the student is totally different from that of teacher for the same input, \mathcal{L}_{rgb} and $\mathcal{L}_{\text{lpiips}}$ will result in images that are slightly closer to teacher but with much less realism. To examine this hypothesis, we remove the GAN loss in Eq. 4 and keep only RGB or LPIPS loss. We also cut off the gradient backward path between the student generator and generated images. In this condition, the gradient of RGB/LPIPS loss directly works on the images. The change of synthesized images reflects how RGB/LPIPS loss influences the generation process. We select two student models, i.e., student-1 that has similar output with teacher for the same input and student-2 that has totally different outputs from teacher. Two students have identical architectures. The mapping network of student-1 inherits from teacher and the mapping network of student-2 is randomly initialized. The effects of RGB/LPIPS loss are shown in Fig. 2a. We can find that the intermediate results are a mixup of source and target images to some extent. If the source image is in the neighbourhood of the target image (1st and 2nd rows), the intermediate results are still perceptually realistic. However, if the source image is totally different from the target image (3rd and 4th rows), the intermediate results are no longer realistic. Though RGB and LPIPS losses are reducing the distance between source and target images, they cannot guarantee a smooth and face-like interpolation in the dissimilar setting. And this unrealistic intermediate results naturally contradict with GAN loss.

From quantitative perspective, we wish to prove that RGB/LPIPS loss is not compatible with GAN loss in the heterogeneous setting by gradient analysis.

In the training process, for each batch, we perform backward propagation for GAN loss, RGB loss and LPIPS loss, respectively, and obtain three gradients of these losses. We then compute the cosine distance between GAN gradient and RGB/LPIPS gradients. As shown in Fig. 2b and Fig. 2c, the cosine distance between GAN gradient and RGB/LPIPS gradients of dissimilar student is always negative, suggesting that RGB/LPIPS gradients are competing with GAN gradients. On the contrary, the cosine distance of similar student is positive, indicating that the distillation loss is driving the model in the same direction as the adversarial loss. Our analysis above suggests that distillation is not beneficial in heterogeneous setting. Having similar outputs for the same input z is the prerequisite for uncGAN distillation.

3.3 Effect of the Mapping Network

As we will show in the experiments, if the student is randomly initialized, it cannot learn consistent outputs as teacher even though we leverage RGB/LPIPS loss to force the agreement between the outputs of two models. We hypothesize that the mapping network $S(z)$ plays a key role in determining whether two models can have consistent outputs. If the gap between mapping networks of student and teacher is too large, it is hard for the student to learn outputs consistent with the teacher. This hypothesis comes from the following motivation.

Suppose the student has a different mapping network from the teacher and the consistency condition (Eq. 2) is still satisfied. Our goal is to derive a contradiction. For the convenience of the following discussion, we define:

$$G(z_1, z_2; k) = C(w_1, w_2; k) = C(w_1, \dots, w_1, w_2, w_1, \dots, w_1), \quad (5)$$

where all the ModConv layers use w_1 except that the k -th layer uses w_2 . The consistency condition of Eq. 2 requires that:

$$G_s(z_1, z_2; k) = G_t(z_1, z_2; k), 1 \leq k \leq L. \quad (6)$$

As shown in StyleGAN [18], for a well-trained model, it should be possible to find direction vectors that consistently correspond to individual factors of variation. An example is shown in appendix A3. Some individual semantic factors such as pose, glasses and hair color can be controlled by moving the style vector w of certain layer along a certain direction. Suppose the direction p at k -th layer controls the hair color of the generated face. The only difference between $C_t(w_0, w_1 + p; k)$ and $C_t(w_0, w_1; k)$ is that they are the same faces with different hair colors. The movement of w from w_1 to $w_1 + p$ corresponds to a consecutive change of hair color of the generated face. If we map the w back to the noise space:

$$z_1 = S_t^{-1}(w_1), \quad z_2 = S_t^{-1}(w_1 + p), \quad (7)$$

obviously, the line segment in w space corresponds to a curve in z space with two end points z_1 and z_2 due to the nonlinearity of $S_t(z)$. We denote this curve as $z_1 \widehat{z}_2$. Then $\{G_t(z_0, z; k) | z \in z_1 \widehat{z}_2\}$ represents a cluster of faces with different hair



Fig. 3: The mapping network $S(\cdot)$ determines whether the student can learn from the teacher’s output.

colors. According to the consistency constraint, $\{G_s(z_0, z; k) | z \in \widehat{z_1 z_2}\}$ should be the same cluster as $\{G_t(z_0, z; k) | z \in \widehat{z_1 z_2}\}$. We feed $z_0, z \in \widehat{z_1 z_2}$ into the student mapping network $S_s(\cdot)$:

$$w'_0 = S_s(z_0), \quad w'_1 w'_2 = S_s(\widehat{z_1 z_2}). \quad (8)$$

Since $S_s(\cdot)$ is different from and independent of $S_t(\cdot)$, the result $w'_1 w'_2$ is still a curve. Thus, the semantic factor of hair color in student model is controlled by a complex curve in w space, which contradicts the property of StyleGAN that various semantic factors are decoupled well in w space. Hence, having different mapping networks and consistency condition cannot hold simultaneously.

We further conduct experiments to examine our hypothesis. Specifically, we select four students according to whether the mapping network is from teacher or not and whether the convolution is from teacher or not. We use GAN loss, RGB loss and LPIPS loss to train these models. The mapping network and convolution are updated together. The results are shown in Fig. 3. The student that inherits weights from the teacher’s mapping network can learn a mapping that aligns well with the teacher’s output, no matter how the convolution $C(\cdot)$ is initialized. However, for the student whose mapping network is randomly initialized, there are no meaningful connections between student’s and teacher’s outputs. The analysis above clearly shows that the output consistency between student and teacher is determined by the mapping network.

3.4 Mapping Network Consistency in GAN Distillation

We have shown that the consistency between student and teacher outputs is the prerequisite of the distillation, and the mapping network determines whether two generators can have consistent outputs. Hence, to make distillation meaningful, it is necessary to impose extra constraints to guarantee the consistency between two mapping networks.

The simplest way is to keep the architecture of the mapping network unchanged and inherit teacher’s parameters directly. In fact, the parameters and FLOPs of the mapping network account for only 7.5% and 0.005% of the convolution backbone. Preserving the mapping network is thus feasible in practice.

If there is a strong demand on the compression of the mapping network, one can perform a two-stage training to ensure a small gap between student and teacher mapping networks. In the first stage, the student mapping network is forced to mimic outputs of the teacher mapping network:

$$\mathcal{L} = \mathbb{E}_{z \sim \mathcal{N}(0,1)} D(S_s(z), S_t(z)), \quad (9)$$

where $D(\cdot, \cdot)$ is a distance metric. Considering $S_t(\cdot)$ and $S_s(\cdot)$ are both shallow MLPs, the training cost of this stage is negligible (0.59% of the normal GAN training in the second stage). In the second stage, the mapping network and generator backbone are finetuned together using the loss in Eq. 4. We will explore the effects of compressing the mapping network in Sec. 4.1.

3.5 Latent-Direction-Based Relation Distillation

Under the premise that consistency condition is satisfied, we further propose to incorporate relation mimicking into GAN distillation. Conventional relation-based distillation [30] in classification tasks computes feature similarity matrices using the samples in a minibatch. Here, we tailor it to better cater to StyleGAN.

Specifically, for a given teacher model, we compute its meaningful latent directions (LD) that control a single semantic factor and store them in a dictionary $\{d_1, d_2, \dots, d_m\}$. Note that the latent direction is related to a specific layer. For example, if d_i is computed in k -layer, then only $C_t(w, w + d_i; k)$ has single semantic factor difference with $C_t(w)$. $C_t(w, w + d_i; j)_{j \neq k}$ does not have this property. In the training stage, we feed a batch of noise $\{z_i\}_{i=1:N}$ into the mapping network and obtain $\{w_i\}_{i=1:N}$. For each w_i we randomly sample a latent direction d from the dictionary. Thus, $C_t(w_i)$ and $C_t(w_i, w_i + \alpha d; k)$ (k is the layer related to d) are two images with single semantic factor difference with α controls the moving distance. We denote the intermediate features of $C_t(w_i)$ and $C_t(w_i, w_i + \alpha d; k)$ as f_i and f'_i , respectively. Then the similarity matrix M between original view and augmentation view can be computed as $A_{i,j} = f_i \cdot f'_j$. We then convert the similarity into probability via the softmax operation and minimize the distance using KL-divergence loss:

$$M_{i,j} = \frac{\exp(A_{i,j})}{\sum_{k=1}^N \exp(A_{i,k})}, \quad \mathcal{L}_{\text{LD}} = - \sum_{i,j} M_{i,j}^t \log M_{i,j}^s. \quad (10)$$

The final learning objective is the combination of Eq. 4 and \mathcal{L}_{LD} .

4 Experiments

We conduct experiments mainly on StyleGAN2/3 since they are the most powerful unconditional GANs so far. We use the FFHQ [18] and LSUN church [37]

Table 1: Effect of initialization. Surprisingly, we find that inheriting only mapping network is the best solution.

mapping network Initialization	Convolution Initialization	Mimicking Loss	Student FID
random	random	No Mimic	10.92
		RGB	10.78
		RGB + LPIPS	11.27
random	inherit	RGB	10.81
		RGB+LPIPS	10.88
inherit	inherit	No Mimic	10.54
		RGB	9.41
		RGB + LPIPS	8.61
		RGB + LPIPS + LD	8.45
inherit	random	RGB	9.42
		RGB + LPIPS	8.23
		RGB + LPIPS + LD	7.94

datasets. We adopt Fréchet Inception Distance (FID), Perceptual Path Length (PPL) [18] and PSNR [23] between real and projected images as evaluation metrics. More qualitative results are shown in the appendix A.7.

For the ablation study in Sec. 4.1, we train the models on resolution 256×256 and use a smaller batch size of 8 to save the computation cost. For the comparison with state-of-the-art methods in Sec. 4.2, we train the models on both resolutions of 256×256 and 1024×1024 . We also use a batch size of 16 that is the same as CAGAN [23] to ensure a fair comparison.

4.1 Ablation Study

The Initialization of the Student Model. Previous works usually treat StyleGAN2 as an integral module and initialize the mapping network and convolution backbone in the same way (from scratch or inherits teacher parameters). Based on our analysis in Sec. 3.3 that the mapping network plays a key role in determining the semantics of generated images, here we separate the mapping network $S(z)$ from the convolution backbone $C(w)$ and test three initialization strategies: 1) both $S(z)$ and $C(w)$ are randomly initialized, 2) both $S(z)$ and $C(w)$ are initialized with teacher weights, 3) only $S(z)$ inherits teacher weights and $C(w)$ is randomly initialized.

The results are shown in Table 1. For the setting where $S(z)$ and $C(w)$ are both randomly initialized, RGB loss can only bring marginal improvement. RGB+LPIPS even performs worse than No-Mimic, indicating that distillation cannot work well when output discrepancy occurs. If $S(z)$ and $C(w)$ both inherit teacher weights, the mimicking loss can achieve 1-2 FID improvement. To explore the effect of the mapping network, we also try inheriting only $S(z)$ and

Table 2: How to deal with the mapping network in StyleGAN2 distillation. The mapping network is comprised of MLPs. The numbers inside and outside the “[]” is the number of channels in each layer and the number of layers, respectively. The mapping network of the teacher is [512]*8. The FLOPs saving is computed with regard to the total FLOPs (the mapping network and convolution layers).

Setting	mapping network Architecture	FLOPs Saving	$D(\cdot, \cdot)$	$\ S_s(z) - S_t(z)\ _1$	Student FID
Random Initialization	[512]*8	0%	N/A	1.027	11.78
Two-Stage	[512]*8	0%	L1	0.156	9.69
	[512]*8	0%	L2	0.260	10.80
	[512]*5	0.0019%	L1	0.197	10.38
	[390]*7+[512]	0.0019%	L1	0.210	10.55
	[256]*7+[512]	0.0034%	L1	0.245	10.86
Inheriting	[512]*8	0%	N/A	0	8.30

surprisingly find that this initialization obtains the best result. And loading $C(w)$ hampers the performance of distillation. This result contradicts with the conclusion in CAGAN. It shows that the general pruning strategy, i.e., determining which channels should be removed, is not important. Randomly initialization of convolution layers is the optimal solution if the mapping network is kept.

The Effects of Mapping Network Compression. We conduct experiments to investigate how to deal with the mapping network in StyleGAN-like models compression. Specifically, we consider three settings: 1) student has the same mapping network architecture as teacher but with random initialization, 2) student mapping network has a different architecture and uses the two-stage training strategy, 3) student has the same architecture and inherits weights from the teacher. For all the settings, the convolution backbones are randomly initialized. For the two-stage setting, we also explore how the architecture of the mapping network and mimicking loss in Eq. 9 affect the final performance. To emphasize the importance of the mapping network, we also list the average L1 distance between $S_s(z)$ and $S_t(z)$ before entering the normal GAN training stage.

The results are shown in Table. 2. ‘Random Initialization’ obtains the worst FID because the output discrepancy makes the distillation ineffective. The ‘Two-Stage’ strategy improves the results by narrowing the gap between $S_s(z)$ and $S_t(z)$. From several two-stage settings, we can find that L1 is a better mimicking loss than L2 and reducing the number of layers is better than reducing the number of channels in each layer. It is also worth noting that there is a strong positive correlation between $|S_s(z) - S_t(z)|$ and FID, indicating that the gap between $S_s(z)$ and $S_t(z)$ determines the output consistency and further determines the influence of distillation. Though the two-stage strategy brings performance gains, there is still a large gap between it and the ‘Inheriting’ variant. Thus, we conclude that the modification to the mapping network will greatly harm the final performance and the two-stage strategy can only mitigate the degradation

Table 3: Comparison with SOTA methods. “↓” (“↑”) denotes the lower (higher) the better. “†” denotes that the numbers come from CAGAN [23]. **Bold** font denotes the results that outperform CAGAN. “heter” denotes the heterogeneous setting where the student is not a subnet of the teacher. Since StyleGAN3 removes the PPL loss in the training stage, we also do not measure the PPL for StyleGAN3. The PSNR (proposed by CAGAN) is a special-designed metric to measure the face projection ability. Thus, we do not measure it for the LSUN church dataset. We compute PSNR using our own implementation and leave the result of GAN slim blank due to the lack of the corresponding checkpoint.

Model	Dataset	Reso.	Methods	RAM	FLOPs	FID (↓)	PPL (↓)	PSNR (↑)
StyleGAN2	FFHQ	256	Teacher	30.0M	45.1B	4.5	0.162	34.26
			Baseline	5.6M	4.1B	9.79	0.156	33.17
			GAN slim	-	5.0B	12.4†	0.313†	-
			CAGAN	5.6M	4.1B	7.9†	0.143†	33.34
			Ours	5.6M	4.1B	7.25	0.135	33.49
			CAGAN-heter	3.4M	2.7B	13.75	0.158	33.19
		Ours-heter	3.4M	2.7B	9.96	0.141	33.54	
		1024	Teacher	49.1M	74.3B	2.7	0.162	33.52
			GAN slim	-	23.9B	10.1†	0.211†	-
			CAGAN	9.2M	7.0B	7.6†	0.157†	32.63
	Ours		9.2M	7.0B	7.19	0.128	32.70	
	LSUN Church	256	Teacher	30.0M	45.1B	4.92	0.168	N/A
CAGAN			5.6M	4.1B	8.57	0.146	N/A	
Ours			5.6M	4.1B	7.96	0.136	N/A	
StyleGAN3	FFHQ	256	Teacher	30.0M	45.1B	4.41	N/A	34.30
			CAGAN	5.6M	4.1B	7.75	N/A	33.39
			Ours	5.6M	4.1B	7.14	N/A	33.58

to a certain degree. Considering that the scale of the original $S_t(z)$ is negligible compared to the convolution backbone, the best practice in StyleGAN2 compression is to preserve the mapping network architecture and inherit the weights from the teacher mapping network.

4.2 Comparison with State-of-the-Art Methods

Quantitative Results. We compare our method with the GAN Slimming [32] and CAGAN [23] methods. Since our method does not focus on the pruning, we directly adopt the student architecture used in CAGAN, i.e., a network that is the same as teacher but with fewer channels. We also compare with CAGAN in the heterogeneous setting where the student is not a subnet of the teacher. Specifically, we modify the kernel size of the second convolution layer in each residual block from 3 to 1, thus inheriting teacher convolution parameters is infeasible. Since CAGAN did not notice the output discrepancy issue and always initialize the mapping network and convolution backbone in the same way, we assume it does not inherit weights from teacher in the heterogeneous setting.



Fig. 4: StyleGAN2 synthesized results on FFHQ 256×256 .

The results are shown in Table 3. For the distillation of StyleGAN2 on FFHQ dataset, our method outperforms CAGAN on FID by 0.65 and 0.41 on resolution 256×256 and 1024×1024 , respectively, showing that our method can generate more realistic images. Note that these improvements are not marginal considering the images generated by CAGAN are already of high quality. For the PPL metric that measures the smoothness of latent space, we outperform CAGAN by 6% (relative improvement) on resolution 256×256 . The gap is even larger (18.5%) on resolution 1024×1024 . For PSNR that is related to the image projection ability, our method also surpasses CAGAN, demonstrating that our method can model the face distribution in real world better. Our superiority is much more significant in the heterogeneous setting, showing that our method can be applied in a more general situation where the student is not necessary to be a subnet of the teacher. On LSUN Church dataset, our method still achieves better results than CAGAN on both FID and PPL, showing that our method not only handles those well-aligned settings, but also works well in complex outdoor scenes. On StyleGAN3, our method also brings more gains, indicating that the proposed method has good generalization ability in various StyleGAN-like models.

Qualitative Results. We show StyleGAN2 generation results of FFHQ on resolution 256×256 in Fig. 4. For Two-Stage, we compress the original 8-layer mapping network into 5 layers. The images of each row are generated using the same input noise z . Note that all the students are trained with mimicking loss. Random $S_s(z)$ cannot make the student model generate images consistent with the teacher due to the different mapping networks. The Two-Stage method mitigates output discrepancy issue by directly mimicking the mapping network, but there still exist semantic differences from the teacher. Compared to CAGAN, our generated images have fewer artifacts and are more similar to the teacher in various semantic features such as the face color, haircut and expression.

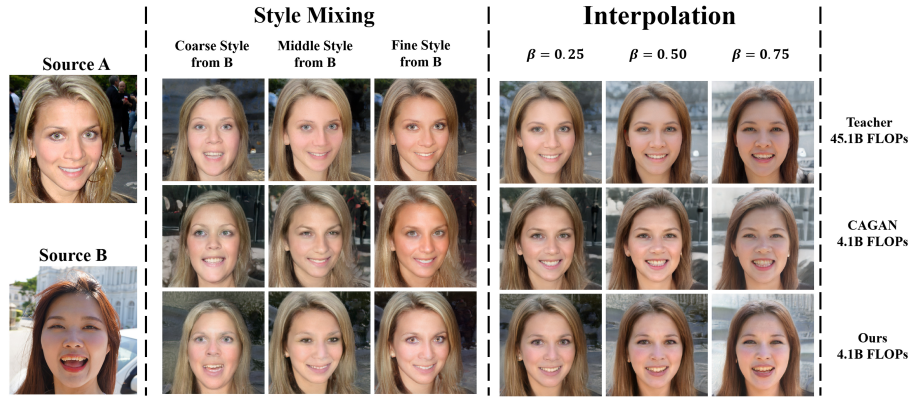


Fig. 5: In coarse style mixing, our result corresponds better with source B on the mouth and face shape. In fine style mixing, our result corresponds better with source B on skin color. CAGAN also generates artifacts on hair in middle and fine style cases.

Image Editing. We demonstrate an image editing case in Fig. 5. Specifically, we apply style mixing and interpolation to the image. The implementation details and more results are shown in the appendix A2.

5 Conclusion

In this paper, we uncover the output discrepancy issue in uncGAN distillation. Through comparative experiments, we find that the mapping network is the key to the output discrepancy and propose a novel initialization strategy of student, which can help resolve the output discrepancy issue. The proposed latent-direction-based distillation loss further improves the distillation efficacy and we achieve state-of-the-art results in StyleGAN2/3 distillation, outperforming the rival method by a large margin on image realism, latent space smoothness and image projection fidelity.

Limitations. The computation and memory footprint of our method are larger than previous methods because it needs to compute the similarity between the original batch and transformed batch. Besides, we only consider the output discrepancy issue in unconditional GANs. In fact, this problem also exists in conditional setting when the condition is not strong enough (e.g., the conditional input is the class label). How to analyze the output discrepancy issues of uncGANs and cGANs in a more general form is also a direction worth exploring.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also supported by Singapore MOE AcRF Tier 2 (MOE-T2EP20120-0001).

References

1. Agualdo, A., Chiang, P.Y., Gain, A., Patil, A., Pearson, K., Feizi, S.: Compressing gans using knowledge distillation. arxiv:1902.00159 (2019)
2. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: The IEEE International Conference on Computer Vision (ICCV) (October 2021)
3. Belousov, S.: Mobilestylegan: A lightweight convolutional neural network for high-fidelity image synthesis. arxiv:2104.04767 (2021)
4. Buciluundefined, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006)
5. Chang, T.Y., Lu, C.J.: Tinygan: Distilling biggan for conditional image generation. In: The Asian Conference on Computer Vision (ACCV) (2020)
6. Chen, H., Wang, Y., Shu, H., Wen, C., Xu, C., Shi, B., Xu, C., Xu, C.: Distilling portable generative adversarial networks for image translation. arXiv:2003.03519 (2020)
7. He, Y., Kang, G., Dong, X., Fu, Y., Yang, Y.: Soft filter pruning for accelerating deep convolutional neural networks. arXiv:1808.06866 (2018)
8. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015)
10. Hou, L., Yuan, Z., Huang, L., Shen, H., Cheng, X., Wang, C.: Slimmable generative adversarial networks. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2021)
11. Hou, Y., Ma, Z., Liu, C., Loy, C.C.: Learning lightweight lane detection cnns by self attention distillation. In: The IEEE International Conference on Computer Vision (ICCV). pp. 1013–1021 (2019)
12. Hu, H., Peng, R., Tai, Y.W., Tang, C.K.: Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. arXiv:1607.03250 (2016)
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
14. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
16. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196 (2018)
17. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

19. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
20. Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W.T., Isola, P., Globerson, A., Irani, M., Mosseri, I.: Explaining in style: Training a gan to explain a classifier in stylespace. =arXiv:2104.13369 (2021)
21. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. arXiv:1608.08710 (2017)
22. Li, M., Lin, J., Ding, Y., Liu, Z., Zhu, J.Y., Han, S.: Gan compression: Efficient architectures for interactive conditional gans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Liu, Y., Shu, Z., Li, Y., Lin, Z., Perazzi, F., Kung, S.Y.: Content-aware gan compression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
24. Peebles, W., Peebles, J., Zhu, J.Y., Efros, A.A., Torralba, A.: The hessian penalty: A weak prior for unsupervised disentanglement. In: The European Conference on Computer Vision (ECCV) (2020)
25. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
26. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (ICLR) (2015)
27. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
28. Shu, H., Wang, Y., Jia, X., Han, K., Chen, H., Xu, C., Tian, Q., Xu, C.: Co-evolutionary compression for unpaired image translation. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
29. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (ICLR) (2020)
30. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
31. Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the gan latent space. In: International Conference on Machine Learning (ICML) (2020)
32. Wang, H., Gui, S., Yang, H., Liu, J., Wang, Z.: Gan slimming: All-in-one gan compression by a unified optimization framework. In: The European Conference on Computer Vision (ECCV) (2020)
33. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
34. Wu, Z., Nitzan, Y., Shechtman, E., Lischinski, D.: Stylealign: Analysis and applications of aligned stylegan models. arxiv:2110.11323 (2021)
35. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: The European Conference on Computer Vision (ECCV) (2020)
36. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
37. 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:1506.03365 (2016)

38. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (ICLR) (2017)
39. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
40. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: The IEEE International Conference on Computer Vision (ICCV) (2017)

A Appendix

A.1 Implementation Details

Training Hyperparameters. For the mapping network mimicking of the first stage, we use Adam as the optimizer with a initial learning rate of 0.05. We train for 50k steps and the batch size is set as 4096. For the normal GAN training of the second stage, we use Adam optimizer with a initial learning rate of 0.002 and 450k iterations. For the α that controls the offset along latent direction, we sample it from a Gaussian distribution $\mathcal{N}(0, 5)$. We set λ_{GAN} , λ_{rgb} , λ_{lips} and λ_{LD} to be 1, 3, 3 and 30, respectively. The features that are used to compute LD loss come from the outputs of 64/128/256 resolution blocks.

Evaluation Metrics. Fréchet Inception Distance (FID) is a commonly used metric to evaluate the realism of generated images. The generated images and real images are fed into a inception network and then a Fréchet distance is computed between their corresponding feature maps. We use the implementation of FID in CAGAN [23]. Specifically, we use 50K real images and 50K generated images to compute statistics, respectively. Perceptual Path Length (PPL) is proposed in StyleGAN [18] to measure the smoothness of latent space. We adopt the PPL implementation in CAGAN [23] for a fair comparison. PSNR and LPIPS are used by CAGAN to evaluate the image projection ability. A given real image is first mapped back to the latent space through optimizer such as L-BFGS. The projected image is obtained by feeding this resulting latent code to the generator. Then, the PSNR and LPIPS distance are computed between the projected image and the original image again. A smaller value indicates that the generator can model the distribution in real world better. We compute these two metrics using our own implementation.

A.2 Distillation without GAN Loss

In Section. 3.3 of the main paper, we highlight that the mapping network decides whether a student can learn similar output to that of the teacher. To further examine this hypothesis, we train the student in a fully supervised manner. Specifically, we remove the GAN loss and treat the z and $G_t(z)$ as input/label pairs to train the student network. The result is shown in Fig. A1. It shows that the student cannot learn any meaningful content in the distillation process without a suitable mapping network. It yields the same face-like output for all the input noise.

A.3 Latent-Direction-Based Distillation Loss

The proposed latent-direction-based loss is essentially a relation loss. We are interested in whether the benefit brought by \mathcal{L}_{LD} comes from relation mimicking or from the latent-direction-based augmentation. Specifically, we consider three variants: 1) Single View, namely the similarity is computed inside the normal

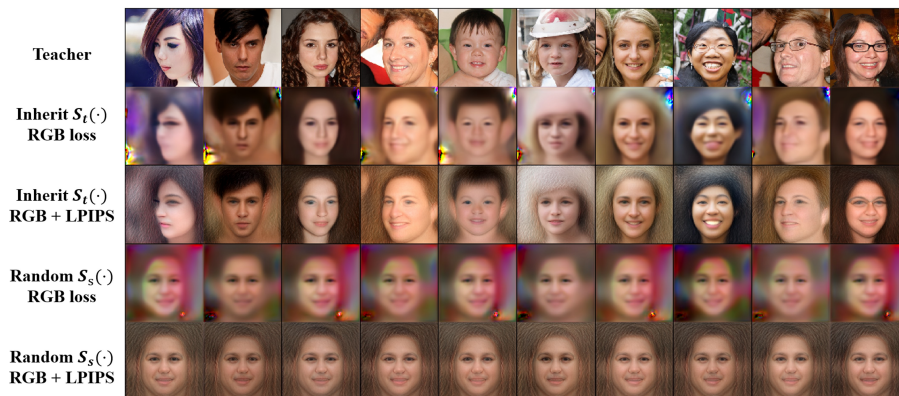


Fig. A1: Distillation without GAN loss.

Table A1: Ablation study about relation mimicking. Single View brings marginal improvement. Random Offset even has negative effect. Our LD loss consistently improves the performances of both RGB and RGB+LPIPS.

Mimicking Loss	\mathcal{L}_{LD}	FID
RGB	N/A	9.41
RGB + Random Offset	KL	9.80
RGB + Single View	KL	9.47
RGB + LD	L2	9.16
RGB + LD	KL	9.05
RGB + LPIPS	N/A	8.61
RGB + LPIPS + LD	L2	8.64
RGB + LPIPS + LD	KL	8.26

samples rather than between normal samples and augmented samples, 2) Random Offset, namely we move w along a random direction to get f'_i instead of along the latent direction, 3) Our latent-direction-based method (abbreviated as LD).

A.4 Image Editing

We demonstrate the superiority of our method on image editing, including style mixing and interpolation. Given two real face images I_A, I_B , we first project them back to the latent space and get w_A, w_B . Both w_A and w_B are of shape $L \times D$, where L is the number of convolution layers and D is the dimension of latent code. For style mixing, we replace the i -th vector in w_A with that from w_B . We set $i \in [1, 3]$, $i \in [5, 8]$ and $i \in [10, 13]$ for coarse, middle and fine style mixing, respectively. For interpolation, we linearly combine the latent code with β controls the weight: $w = \beta \cdot w_A + (1 - \beta) \cdot w_B$, and then feed w into generator to get the interpolation results. We edit the images on resolution 256×256 .

A.5 StyleGAN2 Linear Separability

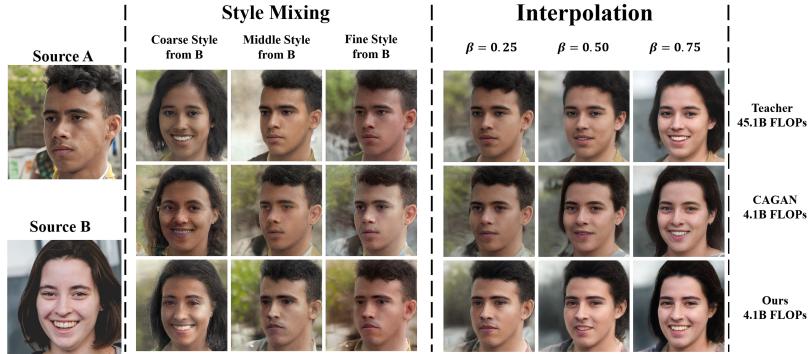
A well-trained StyleGAN2 model is linear separable in the latent space. An example is shown in Fig. A3. The results are shown in Fig. A2. For style mixing, CAGAN always has artifacts in face shape (coarse style) and skin color (middle shape). In contrast, the synthesized results of our method are more realistic and correspond better with two source images. In the coarse style case, our result corresponds well on face shape and facial components with source B. In the fine style case, our result corresponds well on lighting and skin color with source B. For interpolation, we also observe a smoother change than CAGAN, showing that our method learns a better structure in the latent space.

A.6 Image Projection

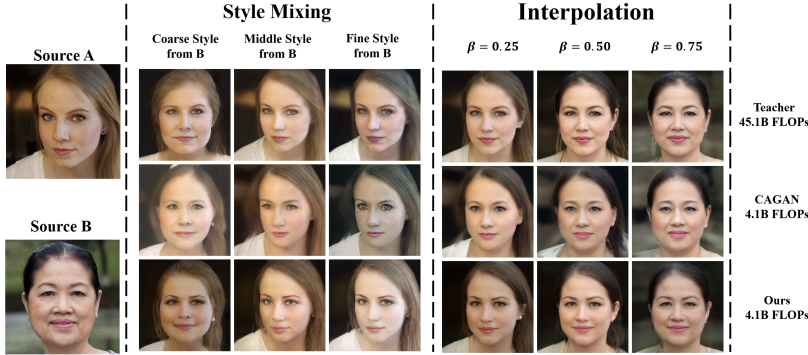
We show image projection results of our method in Fig. A4. All the real images come from Helen Set55 [23] and are not seen in the training stage. Our model reconstructs them with high quality.

A.7 Generation Results

We show more generation results of FFHQ and LSUN church datasets in Fig. A5 and Fig. A6, respectively.



(a) In coarse style mixing, CAGAN generates glasses, which does not appear in both source images. CAGAN also produces blurry images in middle style case. In contrast, our style mixing results are more realistic and more similar to teacher.



(b) CAGAN generates lighting artifacts in coarse case and skin color artifacts in fine case, while our results are more realistic. In interpolation of CAGAN, the earrings disappear in $\beta = 0.25$ but appear again in $\beta = 0.50$. In contrast, our results are much smoother.

Fig. A2: Image editing results.



Fig. A3: StyleGAN2 shows good factorization in the w space. It is possible to control a single semantic factor such as pose, lighting condition, glasses and hair color by moving the style vector w of a certain layer along a specific direction.

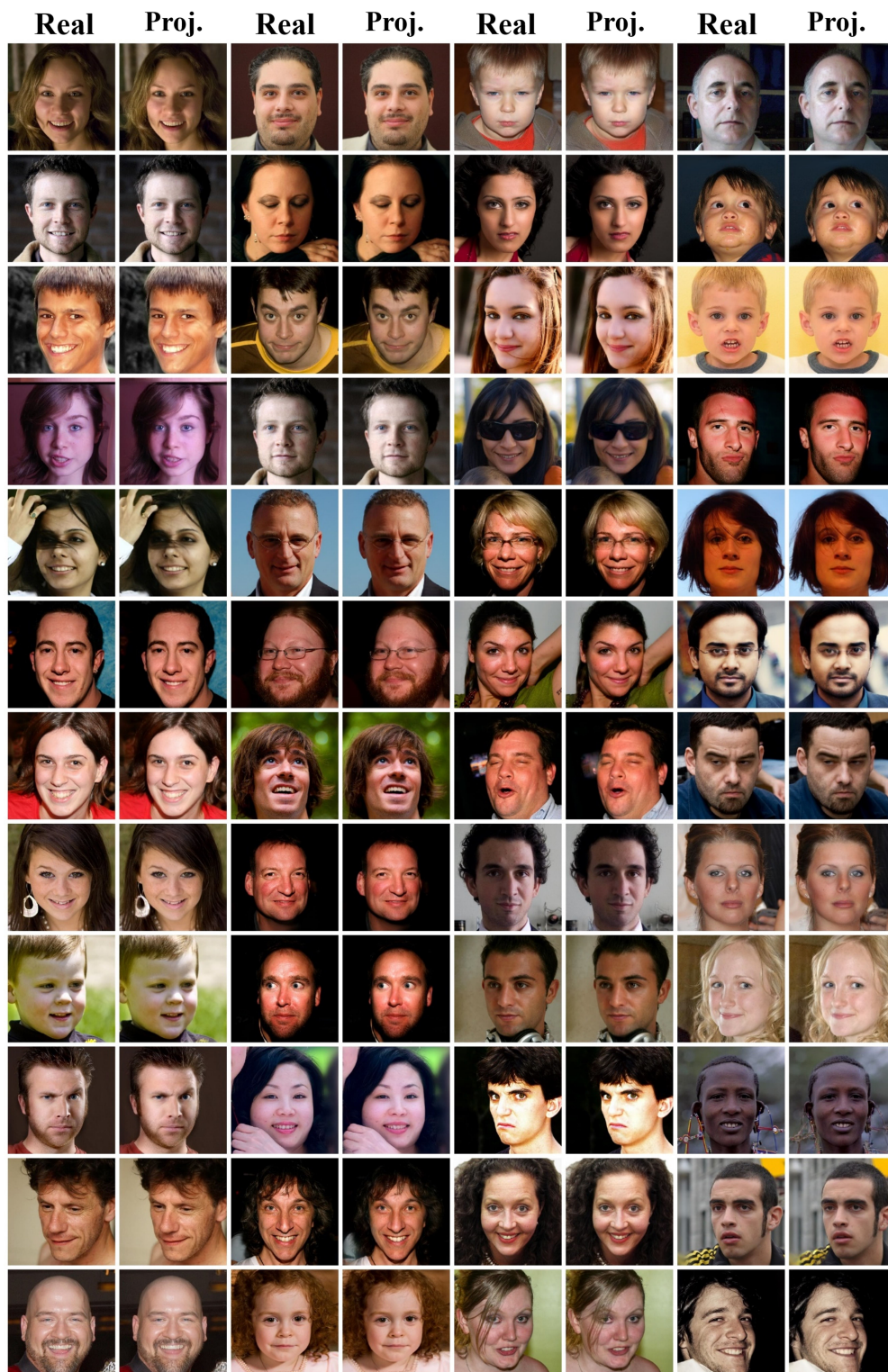


Fig. A4: Image projection results. In each pair, the left image is from real world (not from training set) and the right image is the projected result by our model. Our method can model the real face distribution well.

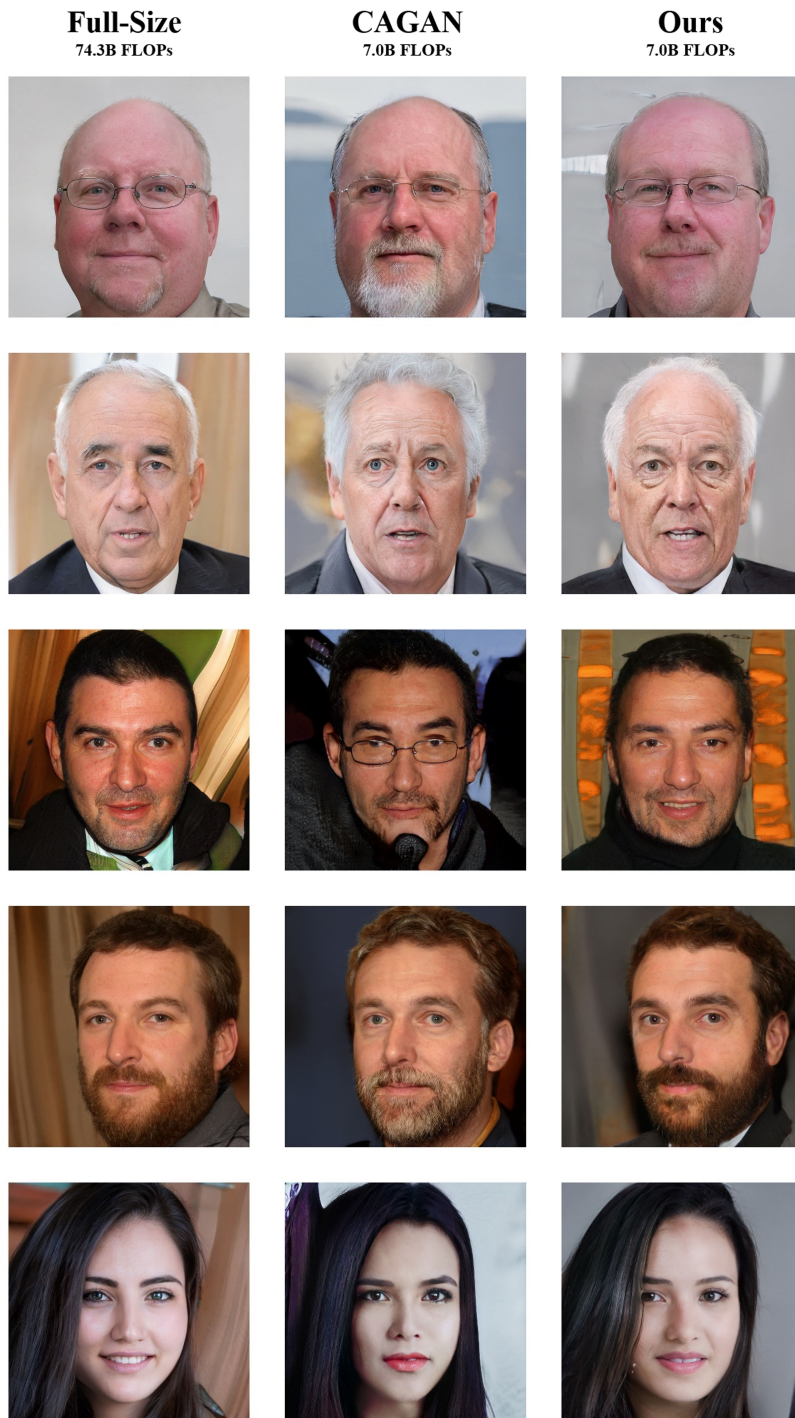


Fig. A5: Generation results on resolution 1024×1024 . The synthesized images of Our method are of better quality than CAGAN. In several semantic factors such as beard, haircut and glasses, our results are more similar to the full-size model even though we do not inherit convolution weights.

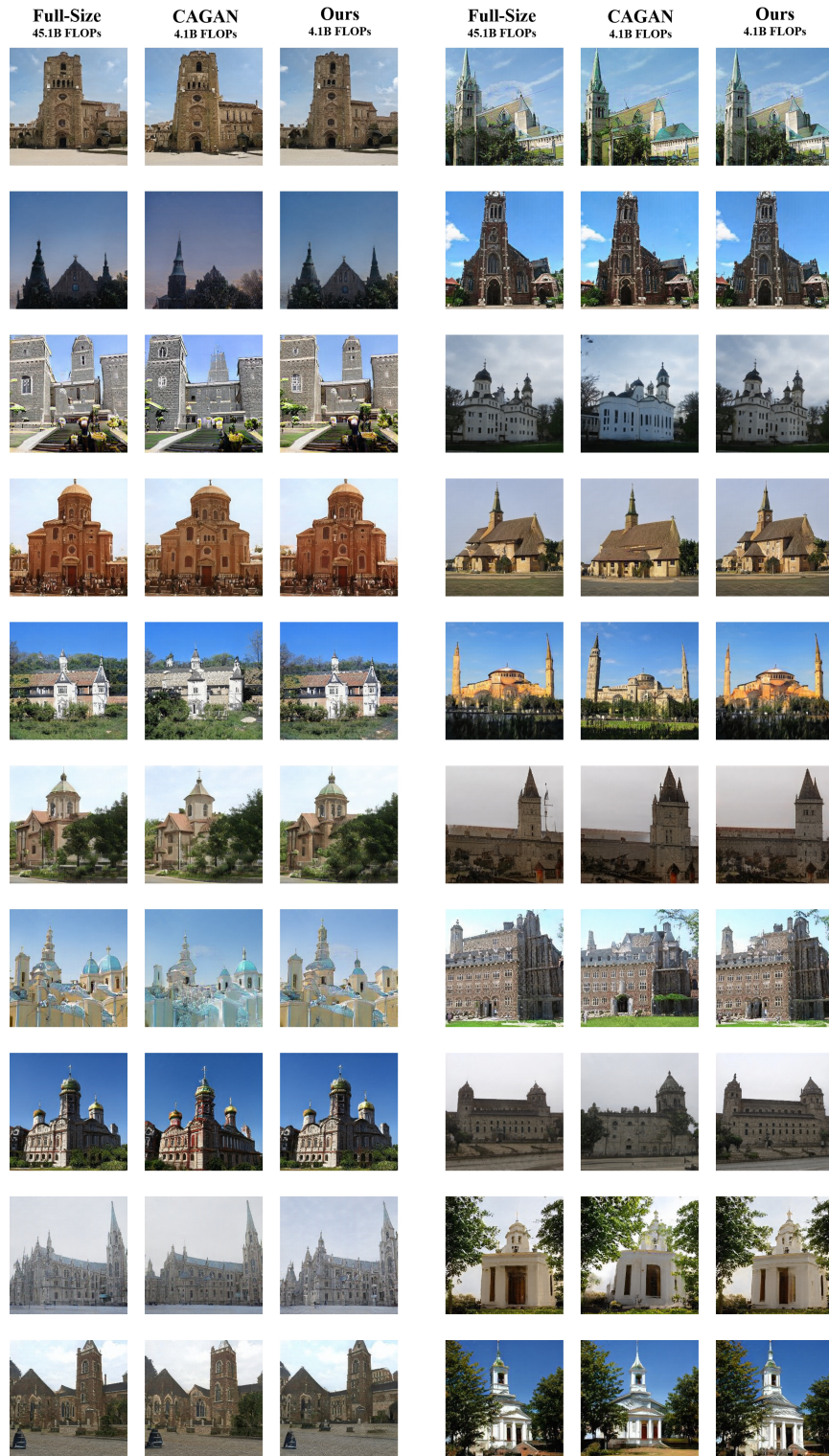


Fig. A6: Generation results on LSUN church on resolution 256×256 .