Ensembling Off-the-shelf Models for GAN Training

Nupur Kumari 1  Richard Zhang2  Eli Shechtman2  Jun-Yan Zhu1
1Carnegie Mellon University  2Adobe

Abstract

The advent of large-scale training has produced a cornucopia of powerful visual recognition models. However, generative models, such as GANs, have traditionally been trained from scratch in an unsupervised manner. Can the collective “knowledge” from a large bank of pretrained vision models be leveraged to improve GAN training? If so, with so many models to choose from, which one(s) should be selected, and in what manner are they most effective?

We find that pretrained computer vision models can significantly improve performance when used in an ensemble of discriminators. Notably, the particular subset of selected models greatly affects performance. We propose an effective selection mechanism, by probing the linear separability between real and fake samples in pretrained model embeddings, choosing the most accurate model, and progressively adding it to the discriminator ensemble. Interestingly, our method can improve GAN training in both limited data and large-scale settings. Given only 10k training samples, our FID on LSUN CAT matches the StyleGAN2 trained on 1.6M images. On the full dataset, our method improves FID by 1.5 to 2\times on cat, church, and horse categories of LSUN.

1. Introduction

Image generation inherently requires being able to capture and model complex statistics in real-world visual phenomenon. Computer vision models, driven by the success of supervised and self-supervised learning techniques [15, 17, 33, 66, 78], have proven effective at capturing useful representations when trained on large-scale data [69, 92, 103]. What potential implications does this have on generative modeling? If one day, perfect computer vision systems could answer any question about any image, could this capability be leveraged to improve image synthesis models?

Surprisingly, despite the aforementioned connection between synthesis and analysis, state-of-the-art generative adversarial networks (GANs) [9, 39, 40, 101] are trained in an unsupervised manner without the aid of such pretrained networks. With a plethora of useful models easily available in the research ecosystem, this presents a missed opportunity to explore. Can the knowledge of pretrained visual representations actually benefit GAN training? If so, with so many models, tasks, and datasets to choose from, which models should be used, and in what manner are they most effective?

In this work, we study the use of a “bank” of pretrained deep feature extractors to aid in generative model training. Specifically, GANs are trained with a discriminator, aimed at continuously learning the relevant statistics differentiating real and generated samples, and a generator, which aims to reduce this gap. Naïvely using such strong, pretrained networks as a discriminator leads to the overfitting and overwhelming the generator, especially in limited data settings. We show that freezing the pretrained network (with a small, lightweight learned classifier on top as shown in Figure 1) provides stable training when used with the original, learned discriminator. In addition, ensembling multiple pretrained networks encourages the generator to match the real distribution in different, complementary feature spaces.

![Figure 1. Vision-aided GAN training. The model bank \( \mathcal{F} \) consists of widely used and state-of-the-art pretrained networks. We automatically select a subset \( \{ \hat{F}_k \}_{k=1}^K \) from \( \mathcal{F} \), which can best distinguish between real and fake distribution. Our training procedure consists of creating an ensemble of the original discriminator \( D \) and discriminators \( D_k = \hat{C}_k \circ \hat{F}_k \) based on the feature space of selected off-the-shelf models. \( \hat{C}_k \) is a shallow trainable network over the frozen pretrained features.](image-url)
To choose which networks work best, we propose to use an automatic model selection strategy, based on the linear separability of real and fake images in the feature space, and progressively add supervision from a set of available pretrained networks. In addition, we use label smoothing [72] and differentiable augmentation [39, 101] to stabilize the model training further and reduce overfitting.

We experiment on several datasets in both limited and large-scale sample setting to show the effectiveness of our method. We improve the state-of-the-art on FFHQ [41] and LSUN [92] datasets with 1k training samples by 2× and on the FID metric [35]. For LSUN CATs, our method improves the FID of StyleGAN2 trained on the full dataset (1.6M images) with only 10k samples, as shown in Figure 2. In the full-scale data setting, our method improves FID for LSUN CATs from 6.86 to 3.98, LSUN CHURCH from 4.28 to 1.72, and LSUN HORSE from 4.09 to 2.11. Finally, we visualize the internal representation of our learned models as well as training dynamics. Check out our code at https://github.com/nupurkmr9/vision-aided-gan. Full version of the paper is available at https://arxiv.org/abs/2112.09130.

2. Related Work

Improving GAN training. Since the introduction of GANs [30], significant advances have been induced by architectural changes [40, 41, 67], training schemes [38, 96], as well as objective functions [4, 5, 21, 24, 54, 55]. The learning objectives often aim to minimize different types of divergences between real and fake distribution. The discriminators are typically trained from scratch and do not use pretrained networks. Especially for the limited data setting, the discriminator is prone to overfit the training set [39, 90, 101].

Use of pretrained models in image synthesis. Pretrained models have been widely used as perceptual loss functions [23, 27, 37] to measure the distance between an output image and a target image in deep feature space. The loss has proven effective for conditional image synthesis tasks such as super-resolution [47], image-to-image translation [14, 62, 86], and neural style transfer [27]. Zhang et al. [98] show that deep features can indeed match the human perception of image similarity better than classic metrics. Sungatullina et al. [80] propose a perceptual discriminator to combine perceptual loss and adversarial loss for unpaired image-to-image translation. This idea was recently used by a concurrent work on CG2real [68]. Another recent work [26] proposes the use of pretrained objects detectors to detect regions in the image and trains object-specific discriminators. Our work is inspired by the idea of perceptual discriminators [80] but differs in three ways. First, we focus on a different application of unconditional GAN training rather than image-to-image translation. Second, instead of using a single VGG model, we ensemble a diverse set of feature representations that complement each other. Finally, we propose an automatic model selection method to find models useful for a given domain. A concurrent work [74] propose to reduce overfitting of perceptual discriminators [80] using random projection and achieve better and faster GAN training.

Loosely related to our work, other works have used pretrained models for clustering, encoding, and nearest neighbor search during their model training. LogoGAN [71] uses deep features to get synthetic clustering labels for conditional GAN training. InclusiveGAN [93] improves the recall of generated samples by enforcing each real image to be close to a generated image in deep feature space. Shocher et al. [76] uses an encoder-decoder based generative model with pretrained encoder for image-to-image translation tasks. Pretrained features have also been used to condition the generator in GANs [13, 53]. Different from the above work, our method empowers the discriminator with pretrained models and requires no changes to the backbone generator.

Use of pretrained models in image editing. Pretrained models have also been used in image editing once the generative model has been trained. Notable examples include image projection with a perceptual distance [1, 105], text-driven image editing with CLIP [64], finding editable directions using attribute classifier models [75], and extracting semantic editing regions with pretrained segmentation networks [106]. In our work, we focus on using the rich knowledge of computer vision models to improve model training.

Transfer learning. Large-scale supervised and self-supervised models learn useful feature representations [11, 15, 34, 44, 66, 89] that can transfer well to unseen tasks, datasets, and domains [22, 36, 43, 61, 70, 84, 91, 94, 95]. In generative modeling, recent works propose transferring the weights of pretrained generators and discriminators from a source domain (e.g., faces) to a new domain (e.g., portraits of one person) [31, 50, 56, 59, 60, 87, 88, 99]. Together with differentiable data augmentation techniques [39, 83, 101, 102], they have shown faster convergence speed and better sampling quality for limited-data settings. Different from them, we transfer the knowledge of learned feature representations of computer vision models. This enables us to leverage the knowledge from a diverse set of sources at scale.

3. Method

Generative Adversarial Networks (GANs) aim to approximate the distribution of real samples from a finite training set \( x \sim P_X \). The generator network \( G \), maps latent vectors \( z \sim P(z) \) (e.g., a normal distribution) to samples \( G(z) \sim P_G \). The discriminator network \( D \) is trained adversarially to distinguish between the continuously changing generated distribution \( P_G \) and target real distribution \( P_X \). GANs perform the minimax optimization \( \min_{G} \max_{D} V(D, G) \), where

\[
V(D, G) = E_{x \sim P_X} \left[ \log D(x) \right] + E_{z \sim P(z)} \left[ \log(1 - D(G(z))) \right].
\]
The discriminator should measure the gap between $P_X$ and $P_\theta$ and guide the generator towards $P_X$. However, in practice, large capacity discriminators can easily overfit on a given training set, especially in the limited-data regime [39, 101]. Unfortunately, as shown in Figure 3, even when we adopt the latest differentiable data augmentation [39] to reduce overfitting, the discriminator still tends to overfit, failing to perform well on a validation set. In addition, the discriminator can potentially focus on artifacts that are indiscernible to humans but obvious for machines [85].

To address the above issues, we propose ensembling a diverse set of deep feature representations as our discriminator. This new source of supervision can benefit us in two manners. First, training a shallow classifier over pretrained features has higher accuracy on validation real images and thus shows better generalization. In the above training, vision-aided adversarial loss is added at the 2M iteration.

![Figure 2. Performance on LSUN Cat and LSUN Church.](image)

Ideally, the discriminator should measure the gap between $P_X$ and $P_\theta$ and guide the generator towards $P_X$. However, in practice, large capacity discriminators can easily overfit on a given training set, especially in the limited-data regime [39, 101]. Unfortunately, as shown in Figure 3, even when we adopt the latest differentiable data augmentation [39] to reduce overfitting, the discriminator still tends to overfit, failing to perform well on a validation set. In addition, the discriminator can potentially focus on artifacts that are indiscernible to humans but obvious for machines [85].

To address the above issues, we propose ensembling a diverse set of deep feature representations as our discriminator. This new source of supervision can benefit us in two manners. First, training a shallow classifier over pretrained features is a common way to adapt deep networks to a small-scale dataset, while reducing overfitting [16, 29]. As shown in Figure 3, our method reduces the discriminator overfitting significantly. Second, recent studies [6, 95] have shown that deep networks can capture meaningful visual concepts from low-level visual cues (edges and textures), to high-level concepts (objects and object parts). A discriminator built on these features may better match human perception [98].

### 3.1. Formulation

Given a set of pretrained feature extractors $\mathcal{F} = \{F_n\}_{n=1}^N$, which learns to tackle different vision tasks, we train corresponding discriminators $\{D_n\}_{n=1}^N$. We add small classifier heads $\{C_n\}_{n=1}^N$ to measure the gap between $P_X$ and $P_\theta$ in the pretrained models’ feature spaces. During discriminator training, the feature extractor $F_n$ is frozen, and only the classifier head is updated. The generator $G$ is updated with the gradients from $D$ and the discriminators $\{D_n\}$ based on pretrained feature extractors. In this manner, we propose to leverage pretrained models in an adversarial fashion for GAN training, which we refer to as Vision-aided Adversarial training:

$$\min_D \max_G V(D, G) + \sum_{n=1}^N V(D_n, G) , \quad (2)$$

where $D_n = C_n \circ F_n$. Here, $C_n$ is a small trainable head over the pretrained features. The above training objective involves the sum of discriminator losses based on all available pretrained models $\{F_n\}$. Solving this at each training iteration would be computationally and memory-intensive. Using all pretrained models would force a significant reduction in batch size to fit all models into memory, potentially hurting performance [9]. To bypass the computational bottleneck, we automatically select a small subset of $K$ models, where $K < N$:

$$\min_D \max_{\mathcal{F}, \{C_n\}_{n=1}^K} V(D, G) + \sum_{k=1}^K V(\hat{D}_k, G), \quad (3)$$

where $\hat{D}_k = \hat{C}_k \circ \hat{F}_k$ denotes the discriminator corresponding to $k$th selected model, and $k \in \{1, \ldots, K\}$.

### 3.2. Model Selection

We choose the models whose off-the-shelf feature spaces best distinguish samples from real and fake distributions. Given the pretrained model’s features of real and fake images, the strongest adversary from the set of models is $\hat{F}_k$, where

$$k = \arg \max_{\mathcal{F}} \max_{C_n} \max_{D_n} V(D_n', G),$$

where $D_n' = C_n \circ F_n$. 

10653
Algorithm 1 GAN training with Vision-aided Adversarial loss.

Input: \( G, D \) trained with standard GAN loss for baseline number of iterations. Off-the-shelf model bank \( \mathcal{F} = \{ F_n \}_{n=1}^N \). Training data \( \{ x_i \} \).

Hyperparameters: \( K \): maximum number of pretrained models to use. \( \{ T_k : k = 1 \cdots K \} \): training intervals before adding next pretrained model.

1: Selected model set \( \hat{\mathcal{F}} = \emptyset \)
2: for \( k = 1 \) to \( K \) do
3: \( \hat{\mathcal{F}} = \hat{\mathcal{F}} \cup \{ \hat{F}_k \} \)
4: \( \hat{D}_k = \hat{C}_k \circ \hat{F}_k \) \( \triangleright \hat{C}_k \) is a shallow trainable network
5: \( \mathcal{F} = \mathcal{F} \setminus \hat{F}_k \)
6: for \( t = 1 \) to \( T_k \) do
7: \( \text{Sample } x \sim \{ x_i \} \)
8: \( \text{Sample } z \sim \tilde{F}(z) \)
9: Update \( D, D_j \) \( \forall j = 1, \cdots, k \) using Eqn. 3
10: Update \( G \) using Eqn. 3
11: end for
12: end for
13: Output: \( G \) with best training set FID

Here \( F_n \) is frozen, and \( C'_n \) is a linear trainable head over the pretrained features. In the case of limited real samples available and for computational efficiency, we use linear probing to measure the separability of real and fake images in the feature space of \( F_n \).

We split the union of real training samples \( \{ x_i \} \) and generated images \( \{ G(z_i) \} \) into training and validation sets. For each pretrained model \( F_n \), we train a logistic linear discriminator head to classify whether a sample comes from \( \mathbb{P}_X \) or \( \mathbb{P}_0 \) and measure \( V(D'_n, G) \) on the validation split. The above term measures the negative binary cross-entropy loss and returns the model with the lowest error. A low validation error correlates with higher accuracy of the linear probe, indicating that the features are useful for distinguishing real from generated samples and using these features will provide more useful feedback to the generator. We empirically validate this on GAN training with 1k training samples of FFHQ and LSUN Cat datasets. Figure 4 shows that the GANs trained with the pretrained model \( F_n \) with higher linear probe accuracy in general achieve better FID metrics.

To incorporate feedback from multiple off-the-shelf models, we explore two variants of model selection and ensembling strategies—(1) \textbf{K-fixed} model selection strategy chooses the \( K \) best off-the-shelf models at the start of training and trains until convergence and (2) \textbf{K-progressive} model selection strategy iteratively selects and adds the best, unused off-the-shelf model after a fixed number of iterations.

\textbf{K-progressive model selection.} We find including multiple models in a progressive manner has lower computational complexity compared to the K-fixed strategy. This also helps in the selection of pretrained models, which captures different aspects of the data distribution. For example, the first two models selected through the progressive strategy are usually a pair of self-supervised and supervised models. For these reasons, we primarily perform all of our experiments using the progressive strategy. We also show a comparison between the two strategies in our arxiv version.

\textbf{Discussion.} The idea of linear separability as a metric has been previously used for evaluating GAN via classifier two-sample tests [52, 100]. We adopt this in our work to evaluate the usefulness of available off-the-shelf discriminators, rather than evaluating generators. “Linear probing” is also a common technique for measuring the effectiveness of intermediate features spaces in both self-supervised [15, 32, 97] and supervised [3] contexts, and model selection has been explored in previous works to predict expert models for transfer learning [25, 58, 65]. We explore this in context of generative modeling and propose a progressive addition of next best model to create an ensemble [12] of discriminators.

### 3.3. Training Algorithm

As shown in Algorithm 1, our final algorithm consists of first training a GAN with standard adversarial loss [30, 42]. Given this baseline generator, we search for the best off-the-shelf models using linear probing and introduce our proposed loss objective during training. In the K-progressive strategy, we add the next discriminator based on off-the-shelf model’s features after training for a fixed number of iterations proportional to the number of available real training samples. The new discriminator is added to the snapshot with the best training set FID in the previous stage. During training, we perform data augmentation through horizontal flipping and use differentiable augmentation techniques [39, 101] and one-sided label smoothing [72] as a regularization. We also observe that only using off-the-shelf models as the discriminator leads to divergence. Thus, the benefit is brought by ensembling the original discriminator and the newly added off-the-shelf models. We show results with the use of three
<table>
<thead>
<tr>
<th>Dataset</th>
<th>StyleGAN2</th>
<th>DiffAugment</th>
<th>ADA</th>
<th>Ours (w/ ADA)</th>
<th>Ours (w/ DiffAugment)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+1st D</td>
<td>+2nd D</td>
</tr>
<tr>
<td>FFHQ</td>
<td>1k</td>
<td>62.16</td>
<td>27.20</td>
<td>19.57</td>
<td>11.43</td>
</tr>
<tr>
<td></td>
<td>2k</td>
<td>42.62</td>
<td>16.63</td>
<td>16.06</td>
<td>10.17</td>
</tr>
<tr>
<td></td>
<td>10k</td>
<td>16.07</td>
<td>8.15</td>
<td>8.38</td>
<td>6.90</td>
</tr>
<tr>
<td>LSUN CAT</td>
<td>1k</td>
<td>185.75</td>
<td>43.32</td>
<td>41.14</td>
<td>15.49</td>
</tr>
<tr>
<td></td>
<td>2k</td>
<td>68.03</td>
<td>25.70</td>
<td>23.32</td>
<td>13.44</td>
</tr>
<tr>
<td></td>
<td>10k</td>
<td>18.59</td>
<td>12.56</td>
<td>13.25</td>
<td>8.37</td>
</tr>
<tr>
<td></td>
<td>2k</td>
<td>-</td>
<td>13.46</td>
<td>11.17</td>
<td>5.25</td>
</tr>
<tr>
<td></td>
<td>10k</td>
<td>-</td>
<td>6.69</td>
<td>6.12</td>
<td>4.80</td>
</tr>
</tbody>
</table>

Table 1. **FFHQ and LSUN results** with varying training samples from 1k to 10k. FID↓ is measured with complete dataset as reference distribution. We select the best snapshot according to training set FID, and report mean of 3 FID evaluations. In Ours (w/ ADA) we finetune the StyleGAN2-ADA model, and in Ours (w/ DiffAugment) we finetune the model trained with DiffAugment while using the corresponding policy for augmentation. Our method works with both ADA and DiffAugment strategy for augmenting images input to the discriminators.

pretrained models and observe minimal benefit with the progressive addition of next model if the linear probe accuracy is low and worse than the models already in the selected set.

4. Experiments

Here we conduct extensive experiments on multiple datasets of different resolutions with the StyleGAN2 architecture. We show results on FFHQ [41], LSUN CAT, and LSUN CHURCH datasets [92] while varying training sample size from 1k to 10k, as well as with the full dataset. For real-world limited sample datasets, we perform experiments on the cat, dog, and wild categories of AFHQ [18] dataset at 512 resolution and MEFACES [39] at 1024 resolution.

**Baseline and metrics.** We compare with state-of-the-art methods for limited dataset GAN training, StyleGAN2-ADA [39] and DiffAugment [101]. We compute the commonly used Fréchet Inception Distance (FID) metric [35] using the clean-fid library [63] to evaluate models. We also report KID [8], precision and recall [46] metrics for the experiments in our arxiv version.

**Off-the-shelf models.** We include eight large-scale self-supervised and supervised networks. Specifically, we perform experiments with CLIP [66], VGG-16 [78] trained for ImageNet [19] classification, and self-supervised models, DINO [11] and MoBY [89]. We also include face parsing [48] and face normals prediction networks [2]. Finally, we have Swin-Transformer [51] based segmentation model trained on ADE-20K [104] and object detection model trained on MS-COCO [49]. Full details of all models are given in the arxiv version.

**Vision-aided discriminator’s architecture.** For discriminator $D_k$ based on pretrained model features, we extract spatial features from the last layer and use a small Conv-LeakyReLU-Linear-LeakyReLU-Linear architecture for binary classification. In the case of big transformer networks, such as CLIP and DINO, we explore a multi-scale architecture that works better. For all experiments, we use three pretrained models selected by the model selection strategy during training. Details about the architecture, model training, memory requirements, and hyperparameters are provided in the arxiv version.
Linear probe accuracy of off-the-shelf models during our K-progressive ensemble training on FFHQ 1k. For the StyleGAN2-ADA, ViT (DINO) model has the highest accuracy and is selected first, then ViT (CLIP) and then Swin-T (MoBY). As we train with vision-aided discriminators, linear probe accuracy decreases for most of the pretrained models. Similar trend for all our experiments are shown in the arxiv version.

### 4.1. FFHQ and LSUN datasets

Table 1 shows the results of our method when the training sample is varied from 1k to 10k for FFHQ, LSUN Cat, and LSUN CHURCH datasets. The considerable gain in FID for all settings shows the effectiveness of our method in the limited data scenario. To qualitatively analyze the difference between our method and StyleGAN2-ADA, we show randomly generated samples from both models given the same latent code in Figure 5. Our method improves the quality of the worst samples, especially for FFHQ and LSUN Cat. Figure 6 shows the accuracy of linear probe over the pretrained models’s features as we progressively add the next discriminator. To analyze the overfitting behavior of discriminators, we also evaluate its training and validation accuracy across iterations. Compared to the baseline StyleGAN2-ADA discriminator, our vision-aided discriminator shows better generalization on the validation set specifically for limited-data regime as shown in Figure 3.

**Full-dataset training.** In the full-dataset setting, we fine-tune the trained StyleGAN2 (config-F) [42] generator with our method. Table 2 shows the comparison of StyleGAN2 and ADM [20] with our method trained using three vision-aided discriminators. We report both FID and Perceptual Path Length (PPL) [41] (W space) metric. On LSUN Cat, our method improves FID from 6.86 to 3.98, on LSUN CHURCH from 4.28 to 1.72, and on LSUN HORSE from 4.09 to 2.11. For FFHQ dataset, our method improves the PPL metric from 144.62 to 127.58 and has similar performance on FID metric. Perceptual path length has been shown to correlate with image quality and indicates a smooth mapping in generator latent space [42]. We also compare the generator trained with our method to StyleGAN2 using GAN dissection [7] in the arxiv version.

**Human preference study.** As suggested by [45] we perform a human preference study on Amazon Mechanical Turk (AMT) to verify that our results agree with the human judg-
which was not used during our training and example with the best FID and show an average of three evaluations. KID is shown in these results correlate with the improved FID metric. We
Table 4. Low-shot generation results on 100-shot Bridge dataset [101], AnimalFace cat and dog [77] categories. Our method significantly improves FID and KID compared to leading methods for few-shot GAN training. KID is shown in $\times 10^5$ units.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bridge</th>
<th>AnimalFace Cat</th>
<th>AnimalFace Dog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FID ↓</td>
<td>KID ↓</td>
<td>FID ↓</td>
</tr>
<tr>
<td>DiffAugment</td>
<td>54.50</td>
<td>43.87</td>
<td>15.68</td>
</tr>
<tr>
<td>ADA</td>
<td>-</td>
<td>-</td>
<td>38.01</td>
</tr>
<tr>
<td>Our +1st D</td>
<td>44.18</td>
<td>9.27</td>
<td>30.62</td>
</tr>
<tr>
<td>+2nd D</td>
<td>33.89</td>
<td>2.35</td>
<td>28.01</td>
</tr>
<tr>
<td>+3rd D</td>
<td>34.35</td>
<td>2.96</td>
<td>27.35</td>
</tr>
</tbody>
</table>

Table 3. Results on AFHQ and MetFaces. Our method, in general, results in lower FID and higher Recall. In transfer setup we fine-tune from a FFHQ trained model of similar resolution with $D$ updated according to FreezeD technique [56] similar to [39]. We select the snapshot with the best FID and show an average of three evaluations. KID is shown in $\times 10^5$ units following [39].

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Transfer</th>
<th>StyleGAN2</th>
<th>StyleGAN2-ADA</th>
<th>Ours (w/ ADA)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FID ↓</td>
<td>KID ↓</td>
<td>Recall ↑</td>
<td>FID ↓</td>
</tr>
<tr>
<td>AFHQ DOG</td>
<td>✗</td>
<td>22.35</td>
<td>10.05</td>
<td>0.20</td>
</tr>
<tr>
<td>✓</td>
<td>9.28</td>
<td>3.13</td>
<td>0.42</td>
<td>7.52</td>
</tr>
<tr>
<td>AFHQ CAT</td>
<td>✗</td>
<td>5.16</td>
<td>1.72</td>
<td>0.26</td>
</tr>
<tr>
<td>✓</td>
<td>3.48</td>
<td>1.07</td>
<td>0.47</td>
<td>3.02</td>
</tr>
<tr>
<td>AFHQ WILD</td>
<td>✗</td>
<td>3.62</td>
<td>0.84</td>
<td>0.15</td>
</tr>
<tr>
<td>✓</td>
<td>2.11</td>
<td>0.17</td>
<td>0.35</td>
<td>2.72</td>
</tr>
</tbody>
</table>

| MetFaces | ✓ | 57.26 | 2.50 | 0.34 | 17.56 | 1.55 | 0.22 | 15.44 | 1.03 | 0.30 |

and StyleGAN2-ADA for six seconds to the test subject and ask to select the more realistic image. We perform this study for 50 test subjects per dataset, and each subject is shown a total of 55 images. On the FFHQ dataset, human preference for our method is 53.8% ± 1.3. For the LSUN CHURCH dataset, our method is preferred over StyleGAN2-ADA with 60.5% ± 1.7, and for the LSUN CAT dataset 63.5% ± 1.6. These results correlate with the improved FID metric. We also show FID evaluation using features of SwAV [10] model which was not used during our training [45,57] and example images from our study in the arxiv version.

### 4.2. AFHQ and MetFaces

To further evaluate our method on real-world limited sample datasets, we perform experiments on MetFaces (1336 images) and AFHQ dog, cat, wild categories with $\sim 5k$ images per category. We compare with StyleGAN2-ADA under two settings, (1) Fine-tuning StyleGAN2-ADA model with our loss (2) Fine-tuning from a StyleGAN2 model trained on FFHQ dataset of same resolution (transfer setup) using FreezeD [56]. The second setting evaluates the transfer learning capability when fine-tuned from a generator trained on a different domain. Table 3 shows the comparison of our method with StyleGAN2 and StyleGAN2-ADA on multiple metrics. We outperform or perform on-par compared to the existing methods in general. Figure 7 shows the qualitative comparison between our method and StyleGAN2-ADA.

### 4.3. Low-shot Generation

To test our method to the limit of low-shot samples, we evaluate our method when only 100-400 samples are available. We finetune StyleGAN2 model with our method on AnimalFace cat (169 images) and dog (389 images) [77], and 100-shot Bridge-of-Sighs [101] datasets. For differentiable augmentation, we use ADA except for the 100-shot dataset where we find that DiffAugment [101] works better than ADA [39], and therefore employ that. Our method leads to considerable improvement over existing methods on FID and KID metrics as shown in Table 4. We show latent space interpolations and nearest neighbour test in our arxiv paper.

### 4.4. Ablation Study

#### Our model selection vs. random selection.

We showed earlier in Figure 4 that FID correlates with model selection ranking in vision-aided GAN training with a single pretrained model. To show the effectiveness of model selection in K-progressive strategy, we compare it with (1) random selection of models during progressive addition and (2) selection of models with least linear probe accuracy. The results are shown in Table 5. We observe that using any of the pretrained models from the model bank already provides benefit in FID, but with our model selection, it can be improved further. More details regarding selected off-the-shelf models are provided in the arxiv version.

#### Role of data augmentation and label smoothing.

Here, we investigate the role of differentiable augmentation [39,83,
Figure 8. Ablation of augmentation and label smoothing on FFHQ and LSUN Cat with 1k training samples and LSUN Cat full-dataset setting. We show the plot of FID w.r.t training iterations when ADA [39] augmentation and label smoothing [72] are individually removed from our training. Without differentiable augmentation, our model training quickly collapses in limited sample setting. Even for full-dataset, using differentiable augmentation for vision-aided discriminator results in better FID. Label smoothing has a reasonable effect in case of LSUN Cat 1k and is marginally helpful for FFHQ 1k. We also change the augmentation technique to DiffAugment [101] for both original and vision-aided discriminator and observe that it performs comparable to ADA [39].

<table>
<thead>
<tr>
<th>Method</th>
<th>FFHQ 1k</th>
<th>LSUN Cat 1k</th>
<th>LSUN Cat 1.6M</th>
</tr>
</thead>
<tbody>
<tr>
<td>StyleGAN2-ADA</td>
<td>19.57</td>
<td>41.14</td>
<td>6.86</td>
</tr>
<tr>
<td>Ours (w/ ViT (CLIP))</td>
<td>11.63</td>
<td>15.49</td>
<td>6.35</td>
</tr>
<tr>
<td>Ours w/ fine-tune ViT (CLIP)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ours w/ ViT random weights</td>
<td>17.59</td>
<td>37.01</td>
<td>6.52</td>
</tr>
<tr>
<td>Ours w/ multi-discriminator</td>
<td>19.07</td>
<td>39.36</td>
<td>6.52</td>
</tr>
</tbody>
</table>

Table 6. Additional ablation studies evaluated on FID metric. Having two discriminators during training (frozen with random weights or trainable) or standard adversarial training for more iterations leads to only marginal benefits in FID. Thus the improvement in our method is through an ensemble of original and vision-aided discriminators. ✓ means FID increased to twice the baseline, and therefore, we stop the training run.

<table>
<thead>
<tr>
<th>Method</th>
<th>FFHQ 1k</th>
<th>LSUN Cat 1k</th>
<th>LSUN Cat 1.6M</th>
</tr>
</thead>
<tbody>
<tr>
<td>StyleGAN2-ADA</td>
<td>19.57</td>
<td>41.14</td>
<td>6.86</td>
</tr>
<tr>
<td>Ours (w/ ViT (CLIP))</td>
<td>11.63</td>
<td>15.49</td>
<td>6.35</td>
</tr>
<tr>
<td>Ours w/ fine-tune ViT (CLIP)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ours w/ ViT random weights</td>
<td>17.59</td>
<td>37.01</td>
<td>6.52</td>
</tr>
<tr>
<td>Ours w/ multi-discriminator</td>
<td>19.07</td>
<td>39.36</td>
<td>6.52</td>
</tr>
</tbody>
</table>

Additional ablation study. Here we further analyze the importance of our design choice. All the experiments are done on LSUN Cat and FFHQ. We compare our method with the following settings: (1) Fine-tuning ViT (CLIP) network as well as in our vision-aided adversarial loss; (2) Randomly initializing the feature extractor network ViT (CLIP); (3) Training with two discriminators, where the 2nd discriminator is of same architecture as StyleGAN2 original discriminator; (4) Training the StyleGAN2-ADA model longer for the same number of iterations as ours with standard adversarial loss. The results are as shown in Table 6. We observe that the baseline methods provide marginal improvement, whereas our method offers significant improvement over StyleGAN2-ADA, as measured by FID. We show more ablation experiments and results with BigGAN [9] architecture in our arxiv version.

5. Limitations and Discussion

In this work, we propose to use available off-the-shelf models to help in the unconditional GAN training. Our method significantly improves the quality of generated images, especially in the limited-data setting. While the use of multiple pretrained models as discriminators improves the generator, it has a few limitations. First, this increases memory requirement for training. Exploring the use of efficient computer vision models [73, 82] will potentially make our method more accessible. Second, our model selection strategy is not ideal in the low-shot settings when only a dozen samples are available. We observe increased variance in the linear probe accuracy with sample size ~ 100 which can lead to ineffective model selection. We plan to adopt few-shot learning [28, 79] methods for these settings in future.

Nonetheless, as more and more self-supervised and supervised computer vision models are readily available, they should be used to good advantage for generative modeling. This paper serves as a small step towards improving generative modeling by transferring the knowledge from large-scale representation learning.

Acknowledgment. We thank Muyang Li, Sheng-Yu Wang, Chonghyuk (Andrew) Song for proofreading the draft. We are also grateful to Alexei A. Efros, Sheng-Yu Wang, Taesung Park, and William Peebles for helpful comments and discussion. The work is partly supported by Adobe Inc., Kwai Inc, Sony Corporation, and Naver Corporation.
References


time-scale update rule converge to a local nash equilibrium. In NeurIPS, 2017. 2, 5


[37] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV, 2016. 2


