Ensembling Off-the-shelf Models for GAN Training

Nupur Kumari [1] Richard Zhang[2] Eli Shechtman[2] Jun-Yan Zhu[1]
[1]Carnegie Mellon University [2]Adobe

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× *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?

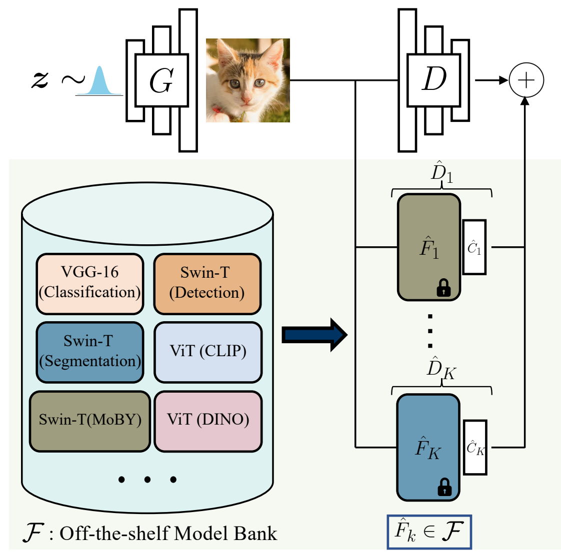

Figure 1. **Vision-aided GAN training**. The model bank $\mathcal{F}$ consists of widely used and state-of-the-art pretrained networks. We automatically select a subset $\{\hat{F}\}_{k=1}^{K}$ from $\mathcal{F}$, which can best distinguish between real and fake distribution. Our training procedure consists of creating an ensemble of the original discriminator D and discriminators $\hat{D}_k = \hat{C}_k \circ \hat{F}_k$ based on the feature space of selected off-the-shelf models. $\hat{C}_k$ is a shallow trainable network over the frozen pretrained features.

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.

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 given 1k training samples by 2-3$\times$ on the FID metric [35]. For LSUN CATS, we match 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. Logo-GAN [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 $\boldsymbol{x} \sim \mathbb{P}_{\mathcal{X}}$. The generator network G, maps latent vectors $\boldsymbol{z} \sim \mathbb{P}(\boldsymbol{z})$ (e.g., a normal distribution) to samples $G(\boldsymbol{z}) \sim \mathbb{P}_\theta$. The discriminator network D is trained adversarially to distinguish between the continuously changing generated distribution $\mathbb{P}_\theta$ and target real distribution $\mathbb{P}_{\mathcal{X}}$. GANs perform the minimax optimization $\min_G \max_D V(D, G)$, where

$$V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\mathcal{X}}}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \tag{1}$$

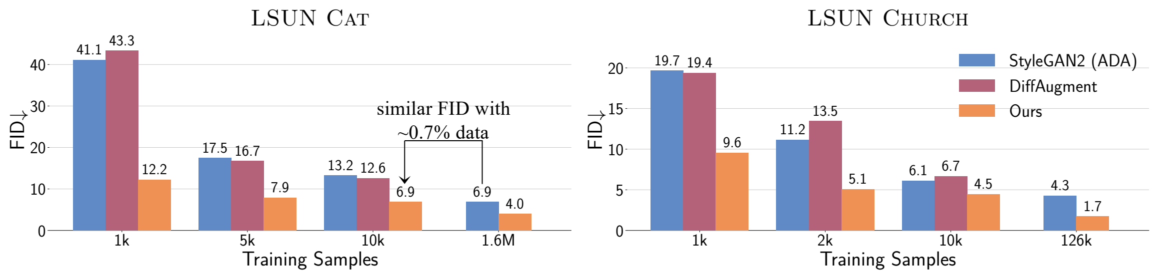

Figure 2. **Performance on LSUN Cat and LSUN Church**. We compare with the leading methods StyleGAN2-ADA [39] and DiffAugment [101] on different sizes of training samples and full-dataset. Our method outperforms them by a large margin, especially in limited sample setting. For LSUN Cat we achieve similar FID as StyleGAN2 [42] trained on full-dataset using only 0.7% of the dataset.

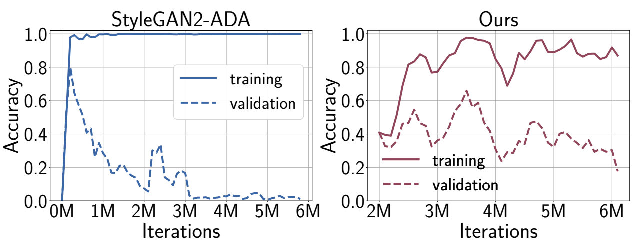

Figure 3. Training and validation accuracy w.r.t. training iterations for our DINO [11] based discriminator vs. baseline StyleGAN2-ADA discriminator on FFHQ 1k dataset. Our discriminator based on pretrained features has higher accuracy on validation real images and thus shows better generalization. In the above training, vision-aided adversarial loss is added at the 2M iteration.

Ideally, the discriminator should measure the gap between $\mathbb{P}_\mathcal{X}$ and $\mathbb{P}_\theta$ and guide the generator towards $\mathbb{P}_\mathcal{X}$. 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 $\mathbb{P}_\mathcal{X}$ and $\mathbb{P}_\theta$ 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:

$$\min_G \max_{D,\{C_n\}_{n=1}^N} V(D,G) + \overbrace{\sum_{n=1}^{N} V(D_n, G)}^{\text{vision-aided adversarial loss}}, \quad (2)$$
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 for 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 $K < N$:

$$\min_G \max_{D,\{\hat{C}_k\}_{k=1}^K} V(D,G) + \sum_{k=1}^{K} V(\hat{D}_k, G), \quad (3)$$

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

$$k = \arg\max_n \{\max_{C'_n} V(D'_n, G)\}, \quad (4)$$
where $D'_n = C'_n \circ F_n$.

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

Selected model set $\hat{\mathcal{F}} = \emptyset$
for $k = 1$ to K **do**
  Select best model $\hat{F}_k \in \mathcal{F}$ using Eqn. 4
  $\hat{\mathcal{F}} = \hat{\mathcal{F}} \cup \{\hat{F}_k\}$
  $\hat{D}_k = \hat{C}_k \circ \hat{F}_k$ $\triangleright$ $\hat{C}_k$ is a shallow trainable network
  $\mathcal{F} = \mathcal{F} \setminus \hat{F}_k$
  for $t = 1$ to T_k **do**
    Sample $\boldsymbol{x} \sim \{\boldsymbol{x}_i\}$
    Sample $\boldsymbol{z} \sim \mathbb{P}(\boldsymbol{z})$
    Update $D, \hat{D}_j \ \forall j = 1, \cdots, k$ using Eqn. 3
    Sample $\boldsymbol{z} \sim \mathbb{P}(\boldsymbol{z})$
    Update G using Eqn. 3
  end for
end for

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 $\{\boldsymbol{x}_i\}$ and generated images $\{G(\boldsymbol{z}_i)\}$ 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 $\mathbb{P}_{\mathcal{X}}$ or $\mathbb{P}_\theta$ 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) **K-fixed** model selection strategy chooses the K best off-the-shelf models at the start of training and trains until convergence and (2) **K-progressive** model selection strategy iteratively selects and adds the best, unused off-the-shelf model after a fixed number of iterations.

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.

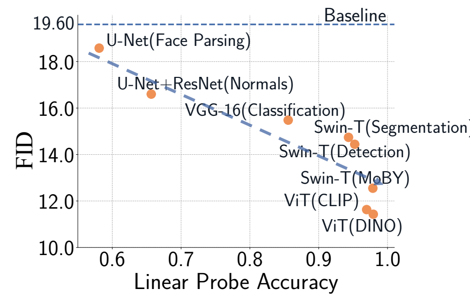

Figure 4. **Model selection using linear probing of pretrained features**. We show correlation of FID with the accuracy of a logistic linear model trained for real vs fake classification over the features of off-the-shelf models. Top dotted line is the FID of StyleGAN2-ADA generator used in model selection and from which we finetune with our proposed vision-aided adversarial loss. Similar analysis for LSUN CAT is shown in our arxiv paper.

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

	Dataset	StyleGAN2	DiffAugment	ADA	Ours (w/ ADA)			Ours (w/ DiffAugment)		
					+1st D	+2nd D	+3rd D	+1st D	+2nd D	+3rd D
FFHQ	1k	62.16	27.20	19.57	11.43	**10.39**	10.58	**12.33**	13.39	12.76
	2k	42.62	16.63	16.06	10.17	8.73	**8.18**	10.01	**9.24**	10.99
	10k	16.07	8.15	8.38	6.90	6.39	**5.90**	6.94	**6.26**	6.43
LSUN Cat	1k	185.75	43.32	41.14	15.49	12.90	**12.19**	13.52	12.52	**11.01**
	2k	68.03	25.70	23.32	13.44	13.35	**11.51**	12.20	11.79	**11.33**
	10k	18.59	12.56	13.25	8.37	7.13	**6.86**	8.19	7.90	**7.79**
LSUN Church	1k	-	19.38	19.66	11.39	9.78	**9.56**	10.15	**9.87**	9.94
	2k	-	13.46	11.17	5.25	**5.06**	5.26	6.09	6.37	**5.56**
	10k	-	6.69	6.12	4.80	4.82	**4.47**	3.42	3.41	**3.25**

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 DiffAgument 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 MetFaces [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 $\hat{D}_k$ 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.

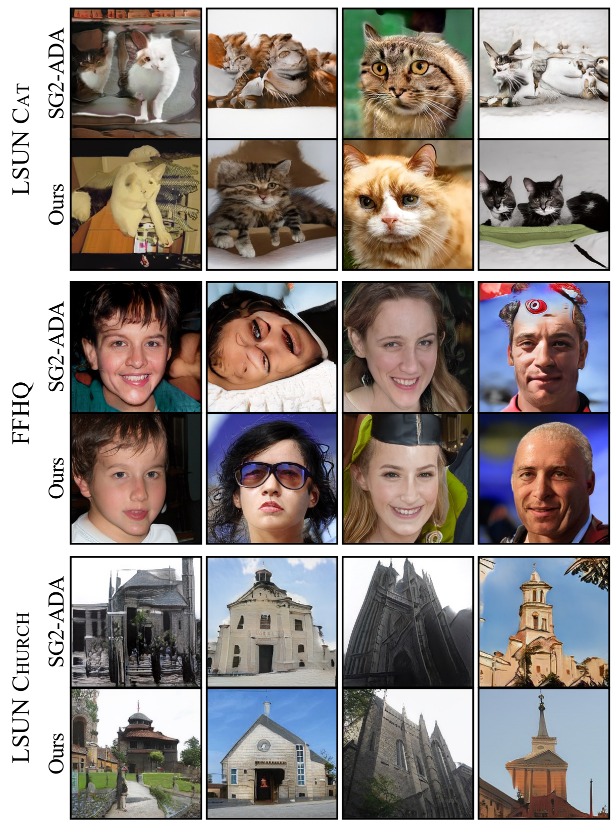

Figure 5. **LSUN Cat, FFHQ, and LSUN Church paired sample comparison in 1k training dataset setting**. For each dataset, the top row shows the baseline StyleGAN2-ADA samples, and the bottom row shows the samples by Our method for the same randomly sample latent code. We fine-tune the StyleGAN2-ADA model with our vision-aided adversarial loss. For the same latent code image quality improves with our method on average.

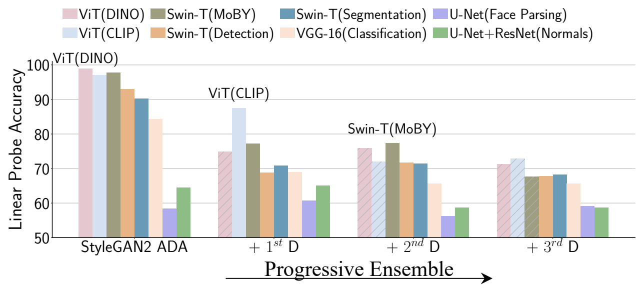

Figure 6. **Linear probe accuracy of off-the-shelf models during our K-progressive ensemble training** on FFHQ 1k. For the StyleGAN2-ADA, ViT (DINO) model has the highest accuracy and is selected first, then ViT (CLIP) and then Swin-T (MoBY). As we train with vision-aided discriminators, linear probe 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 judgment regarding the improved sample quality. We compare StyleGAN2-ADA and our method trained on 1k samples of LSUN CAT, LSUN CHURCH, and FFHQ datasets. Since we fine-tune StyleGAN2-ADA with our method, the same latent code corresponds to similar images for the two models, as also shown in Figure 5. For randomly sampled latent codes, we show the two images generated by our method

Dataset	StyleGAN2 (F)		Ours (w/ ADA)		ADM
	FID ↓	PPL ↓	FID ↓	PPL ↓	FID ↓
FFHQ-1024	**2.98**	144.62	3.01	**127.58**	-
LSUN CAT-256	6.86	437.13	**3.98**	**420.15**	5.57*
LSUN CHURCH-256	4.28	**343.02**	**1.72**	388.94	-
LSUN HORSE-256	4.09	337.98	**2.11**	**307.12**	2.57*

Table 2. **Results on full-dataset setting**. we improve the FID metric on LSUN categories by a significant margin. On the FFHQ dataset we improve the PPL metric. $*$ means directly reported from the ADM paper [20].

Figure 7. **Qualitative comparison of our method with StyleGAN2-ADA on AFHQ.** *Left:* randomly generated samples for both methods. *Right:* For both our model and StyleGAN2-ADA, we independently generate 5k samples and find the worst-case samples compared to real image distribution. We first fit a Gaussian model using the Inception [81] feature space of real images. We then calculate the log-likelihood of each sample given this Gaussian prior and show the images with minimum log-likelihood (maximum Mahalanobis distance).

Dataset	Transfer	StyleGAN2			StyleGAN2-ADA			Ours (w/ ADA)		
		FID ↓	KID ↓	Recall ↑	FID ↓	KID ↓	Recall ↑	FID ↓	KID ↓	Recall ↑
AFHQ Dog	✗	22.35	10.05	0.20	7.60	1.29	0.47	**4.73**	**0.39**	**0.60**
	✓	9.28	3.13	0.42	7.52	1.22	0.43	**4.81**	**0.37**	**0.61**
AFHQ Cat	✗	5.16	1.72	0.26	3.29	0.72	0.41	**2.53**	**0.47**	**0.52**
	✓	3.48	1.07	0.47	3.02	**0.38**	0.45	**2.69**	0.62	**0.50**
AFHQ Wild	✗	3.62	0.84	0.15	3.00	0.44	0.14	**2.36**	**0.38**	**0.29**
	✓	**2.11**	**0.17**	0.35	2.72	0.17	0.29	2.18	0.28	**0.38**
MetFaces	✓	57.26	2.50	**0.34**	17.56	1.55	0.22	**15.44**	**1.03**	0.30

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 $\times 10^3$ units following [39].

Method	Bridge		AnimalFace Cat		AnimalFace Dog	
	FID ↓	KID↓	FID ↓	KID ↓	FID↓	KID↓
DiffAugment	54.50	15.68	43.87	7.56	60.50	20.13
ADA	-	-	38.01	5.61	52.59	14.32
Ours +1st D	44.18	9.27	30.62	1.15	34.23	2.01
Ours +2nd D	**33.89**	**2.35**	28.01	0.37	33.03	**1.37**
Ours +3rd D	34.35	2.96	**27.35**	**0.34**	**32.56**	1.67

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 $\times 10^3$ units.

Model Selection	FFHQ 1k			LSUN Cat 1k		
	+1st D	+2nd D	+3rd D	+1st D	+2nd D	+3rd D
Best	11.43	**10.39**	10.58	15.49	12.90	**12.19**
Random	15.48	12.54	11.92	19.02	15.12	14.28
Worst	15.48	15.45	13.88	19.02	17.53	17.66

Table 5. **FID↓ metric for models trained with different model selection strategies in K-progressive vision-aided training.** 1^{st} *Row:* model selection with best linear probe accuracy. 2^{nd} *Row:* randomly selecting from the bank of off-the-shelf models. 3^{rd} *Row:* model selection with least linear probe accuracy.

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 $53.8\% \pm 1.3$. For the LSUN Church dataset, our method is preferred over StyleGAN2-ADA with $60.5\% \pm 1.7$, and for the LSUN Cat dataset $63.5\% \pm 1.6$. 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,

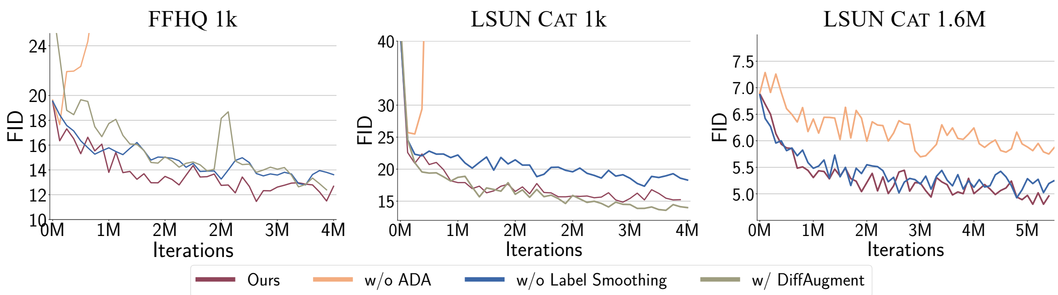

Figure 8. **Ablation of augmentation and label smoothing** on FFHQ and LSUN CAT with 1k training samples and LSUN CAT full-dataset setting. We show the plot of FID w.r.t training iterations when ADA [39] augmentation and label smoothing [72] are individually removed from our training. Without differentiable 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].

Method	FFHQ 1k	LSUN CAT 1k	LSUN CAT 1.6M
StyleGAN2-ADA	19.57	41.14	6.86
Ours (w/ ViT (CLIP))	**11.63**	**15.49**	**4.61**
Ours w/ fine-tune ViT (CLIP)	✗	✗	✗
Ours w/ ViT random weights	19.10	33.77	6.35
Ours w/ multi-discriminator	17.59	37.01	✗
Longer StyleGAN2-ADA	19.07	39.36	6.52

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.

101, 102] which is one of the important factors that enable the effective use of pretrained features. Label smoothing [72] further improves the training dynamics, especially in a limited sample setting. We ablate each of these component and show its contribution in Figure 8 on FFHQ and LSUN CAT dataset in 1k sample setting, and LSUN CAT full-dataset setting. Figure 8 shows that replacing ADA [39] augmentation strategy with DiffAugment [101] in our method also performs comparably. Moreover, in the limited sample setting, without data augmentation, model collapses very early in training, and FID diverges. The role of label smoothing is more prominent in limited data setting e.g. LSUN CAT 1k.

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 in our vision-aided adversarial loss; (2) Randomly initializing the feature extractor network ViT (CLIP); (3) Training with two discriminators, where the 2$^{\text{nd}}$ 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

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 2

[2] Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. Cross-modal deep face normals with deactivable skip connections. In *CVPR*, 2020. 5

[3] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 4

[4] Isabela Albuquerque, João Monteiro, Thang Doan, Breandan Considine, Tiago Falk, and Ioannis Mitliagkas. Multi-objective training of generative adversarial networks with multiple discriminators. In *ICML*, 2019. 2

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *ICML*, 2017. 2

[6] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020. 3

[7] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR*, 2019. 6

[8] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018. 5

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 1, 3, 8

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 7

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3, 5

[12] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *ICML*, 2004. 4

[13] Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero. Instance-conditioned gan. In *NeurIPS*, 2021. 2

[14] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 2

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 4

[16] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019. 3

[17] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[18] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 5

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[20] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 6

[21] Thang Doan, Joao Monteiro, Isabela Albuquerque, Bogdan Mazoure, Audrey Durand, Joelle Pineau, and R Devon Hjelm. On-line adaptative curriculum learning for gans. In *AAAI*, 2019. 2

[22] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014. 2

[23] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *NeurIPS*, 2016. 2

[24] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. In *ICLR*, 2017. 2

[25] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *CVPR*, 2019. 4

[26] Raghudeep Gadde, Qianli Feng, and Aleix M Martinez. Detail me more: Improving gan's photo-realism of complex scenes. In *ICCV*, 2021. 2

[27] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[28] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *ICCV*, 2019. 8

[29] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 3

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 4

[31] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao. Lofgan: Fusing local representations for few-shot image generation. In *ICCV*, 2021. 2

[32] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 4

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two

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

[36] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 2

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

[38] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2

[39] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 1, 2, 3, 4, 5, 7, 8

[40] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 1, 2

[41] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 5, 6

[42] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3, 4, 6

[43] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, 2019. 2

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 2

[45] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022. 6, 7

[46] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019. 5

[47] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[48] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 5

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[50] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *ICLR*, 2021. 2

[51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[52] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *ICLR*, 2017. 4

[53] Puneet Mangla, Nupur Kumari, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Data instance prior (disp) in generative adversarial networks. *arXiv preprint arXiv:2012.04256*, 2020. 2

[54] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *CVPR*, 2017. 2

[55] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018. 2

[56] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In *CVPRW*, 2020. 2, 7

[57] Stanislav Morozov, Andrey Voynov, and Artem Babenko. On self-supervised image representations for gan evaluation. In *ICLR*, 2021. 7

[58] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Deep ensembles for low-data transfer learning. *arXiv preprint arXiv:2010.06866*, 2020. 4

[59] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *ICCV*, 2019. 2

[60] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *CVPR*, 2021. 2

[61] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 2

[62] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2

[63] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021. 5

[64] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 2

[65] Joan Puigcerver, Carlos Riquelme, Basil Mustafa, Cedric Renggli, André Susano Pinto, Sylvain Gelly, Daniel Keysers, and Neil Houlsby. Scalable transfer learning with expert models. In *ICLR*, 2021. 4

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 5

[67] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2

[68] Stephan R Richter, Hassan Abu AlHaija, and Vladlen Koltun. Enhancing photorealism enhancement. *arXiv preprint arXiv:2105.04619*, 2021. 2

[69] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large

scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1

[70] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2

[71] Alexander Sage, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Logo synthesis and manipulation with clustered generative adversarial networks. In *CVPR*, 2018. 2

[72] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 2, 4, 8

[73] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 8

[74] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *NeurIPS*, 2021. 2

[75] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE TPAMI*, 2020. 2

[76] Assaf Shocher, Yossi Gandelsman, Inbar Mosseri, Michal Yarom, Michal Irani, William T Freeman, and Tali Dekel. Semantic pyramid for image generation. In *CVPR*, 2020. 2

[77] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. *PAMI*, 34(7):1354–1367, 2011. 7

[78] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1, 5

[79] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 8

[80] Diana Sungatullina, Egor Zakharov, Dmitry Ulyanov, and Victor Lempitsky. Image manipulation with perceptual discriminators. In *ECCV*, 2018. 2

[81] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 6

[82] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 8

[83] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. Towards good practices for data augmentation in gan training. *arXiv preprint arXiv:2006.05338*, 2, 2020. 2, 7

[84] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[85] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020. 3

[86] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2

[87] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *CVPR*, 2020. 2

[88] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *ECCV*, 2018. 2

[89] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021. 2, 5

[90] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. In *NeurIPS*, 2021. 2

[91] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014. 2

[92] 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, 2, 5

[93] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative models. In *ECCV*, 2020. 2

[94] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 2

[95] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 2, 3

[96] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 2

[97] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 4

[98] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 3

[99] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *ICML*, 2020. 2

[100] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *ICLR*, 2021. 4

[101] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 1, 2, 3, 4, 5, 7, 8

[102] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020. 2, 7

[103] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 40(6):1452–1464, 2017. 1

[104] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 5

[105] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 2

[106] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Barbershop: Gan-based image compositing using segmentation masks, 2021. 2