Differentiable Augmentation for Data-Efficient GAN Training

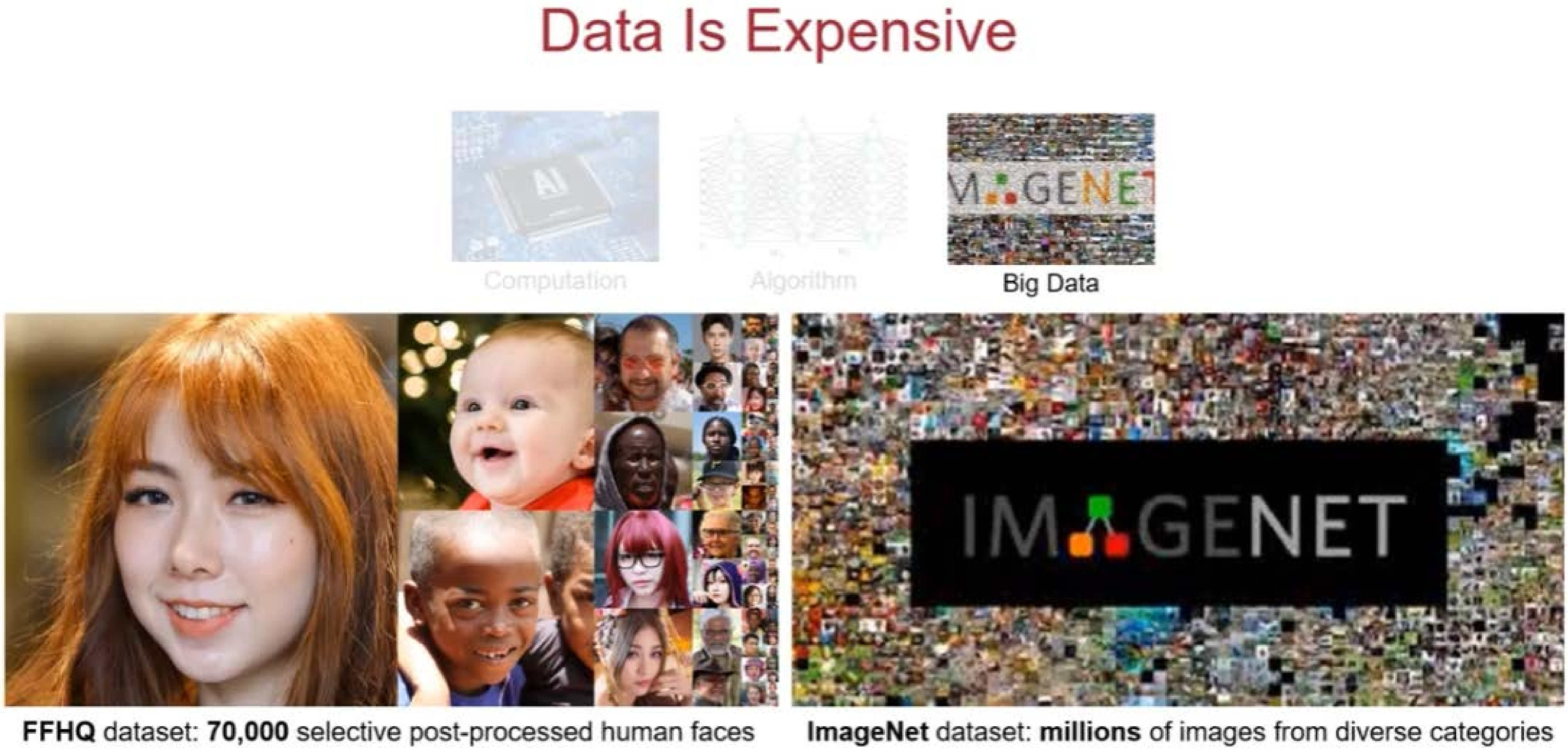
NeurIPS 2020

이한얼

DAVIAN

Data and Visual Analytics Lab

Intro



Data Is Expensive

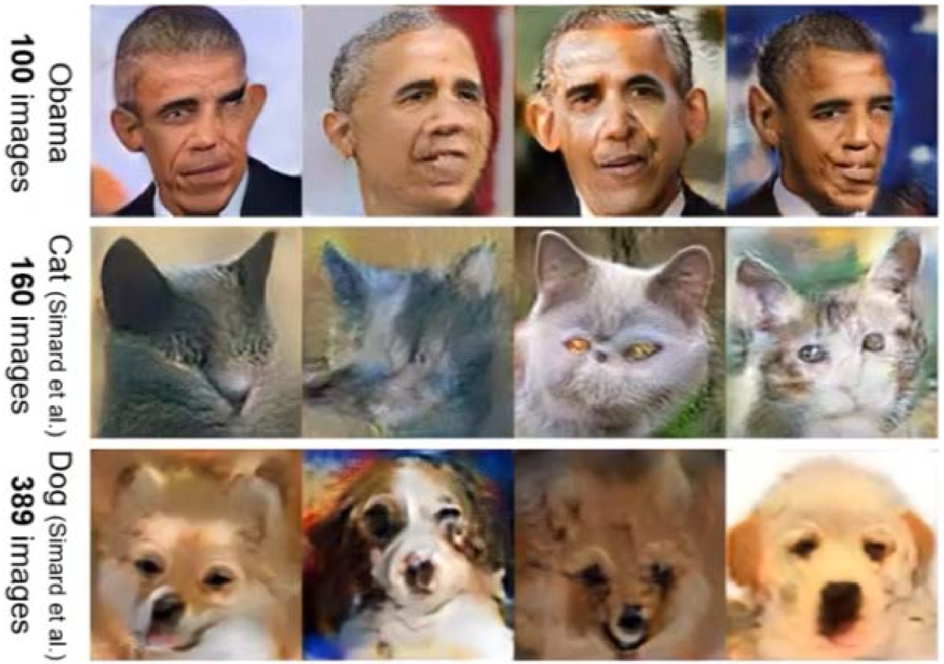
FFHQ dataset: **70,000** selective post-processed human faces

ImageNet dataset: **millions** of images from diverse categories

Months or even years to collect the data,
along with **prohibitive** annotation costs.

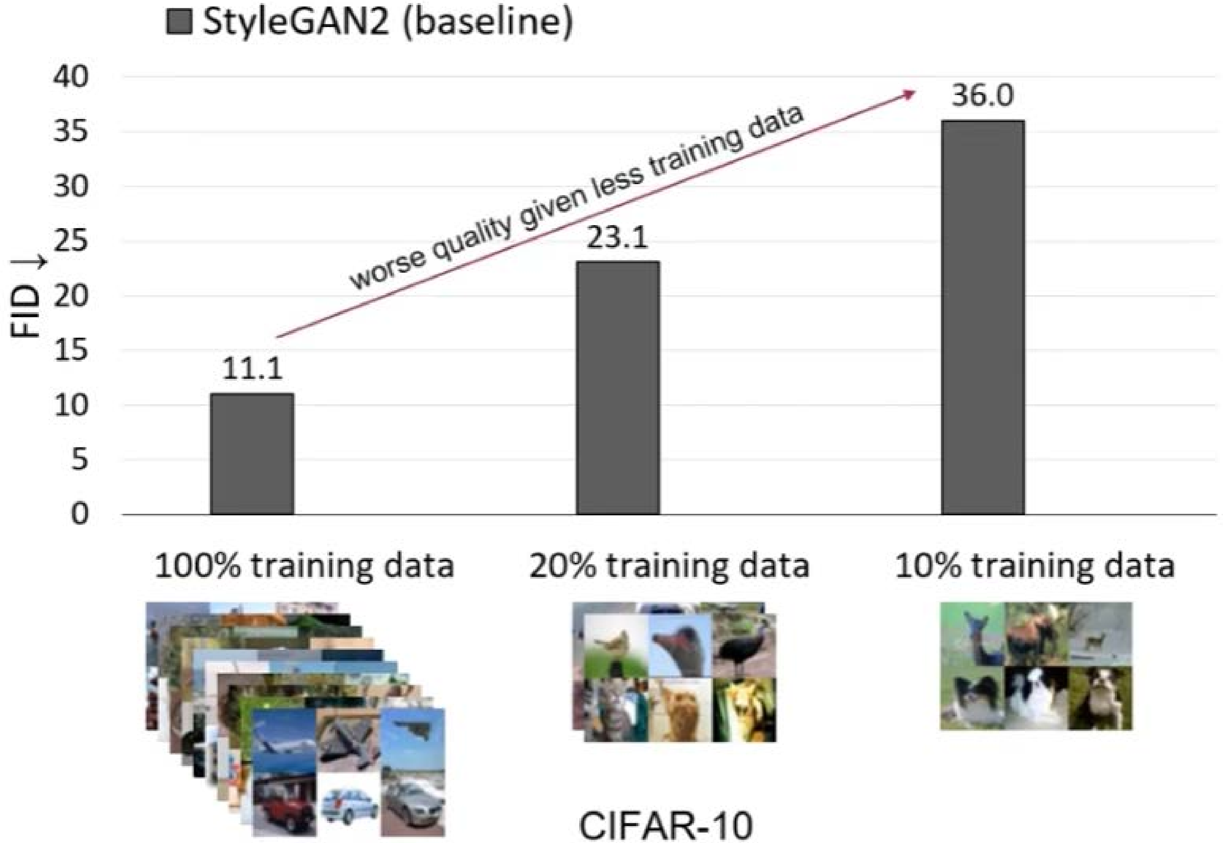
Intro

GANs Heavily Deteriorate Given Limited Data



Generated samples of StyleGAN2 (Karras et al.)
using only hundreds of images

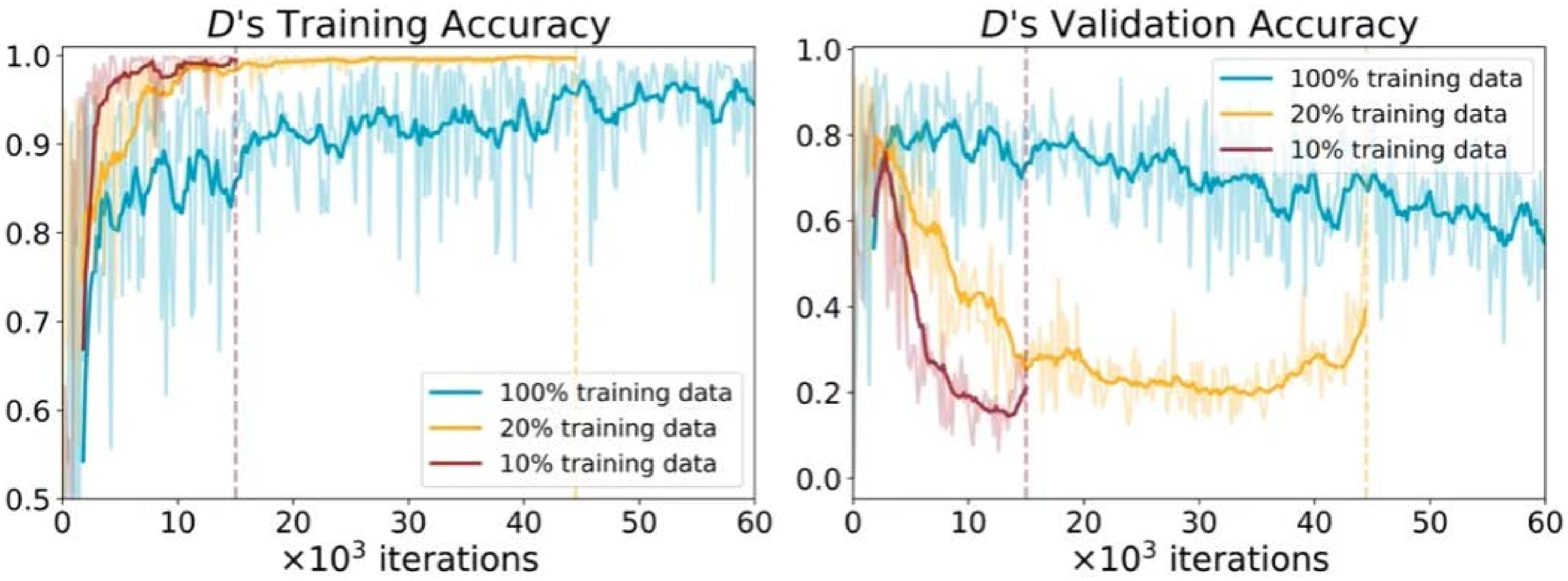
GANs Heavily Deteriorate Given Limited Data



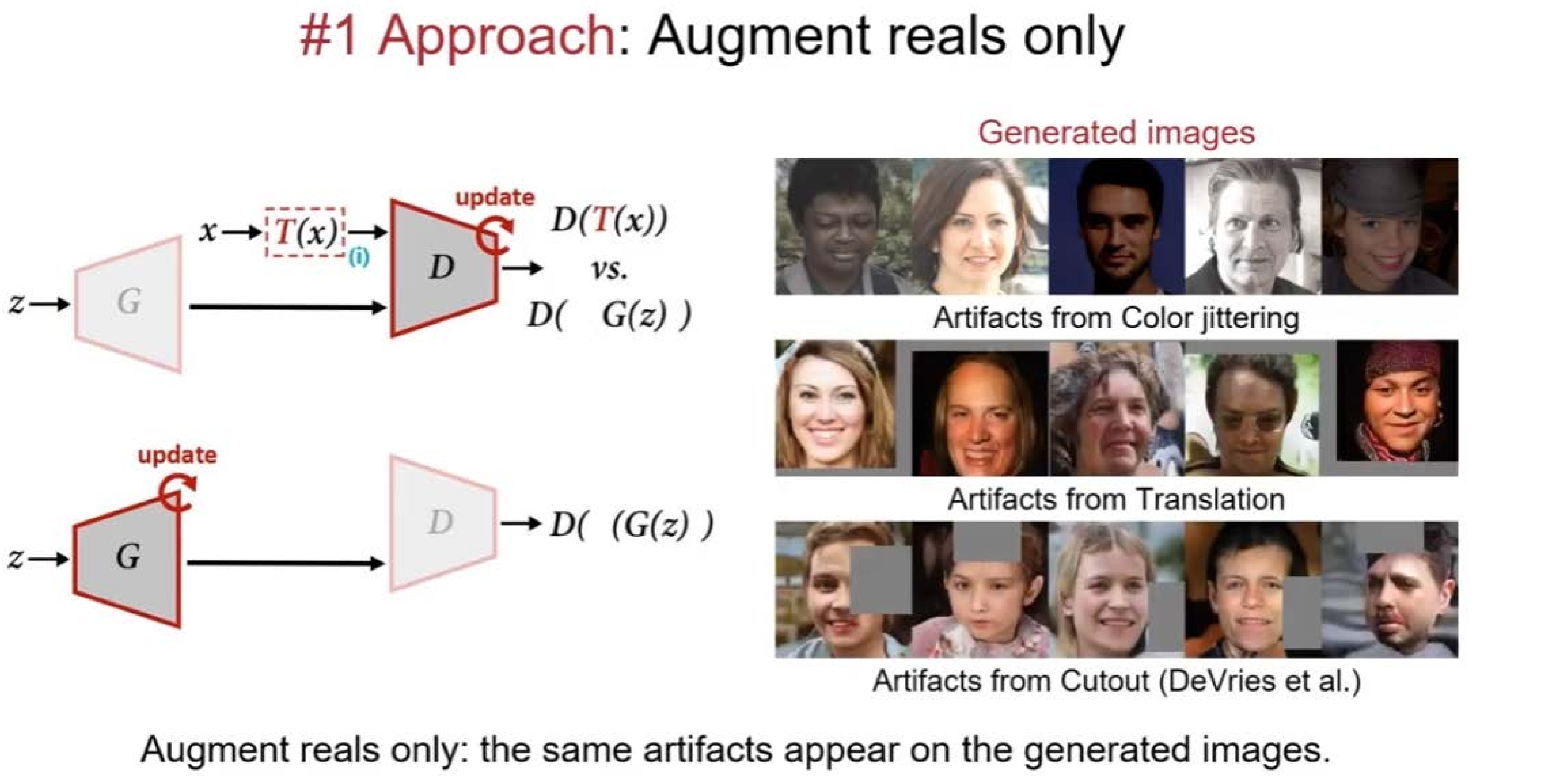
CIFAR-10

Intro

Discriminator Overfitting



Intro



#1 Approach: Augment reals only

Generated images

Artifacts from Color jittering

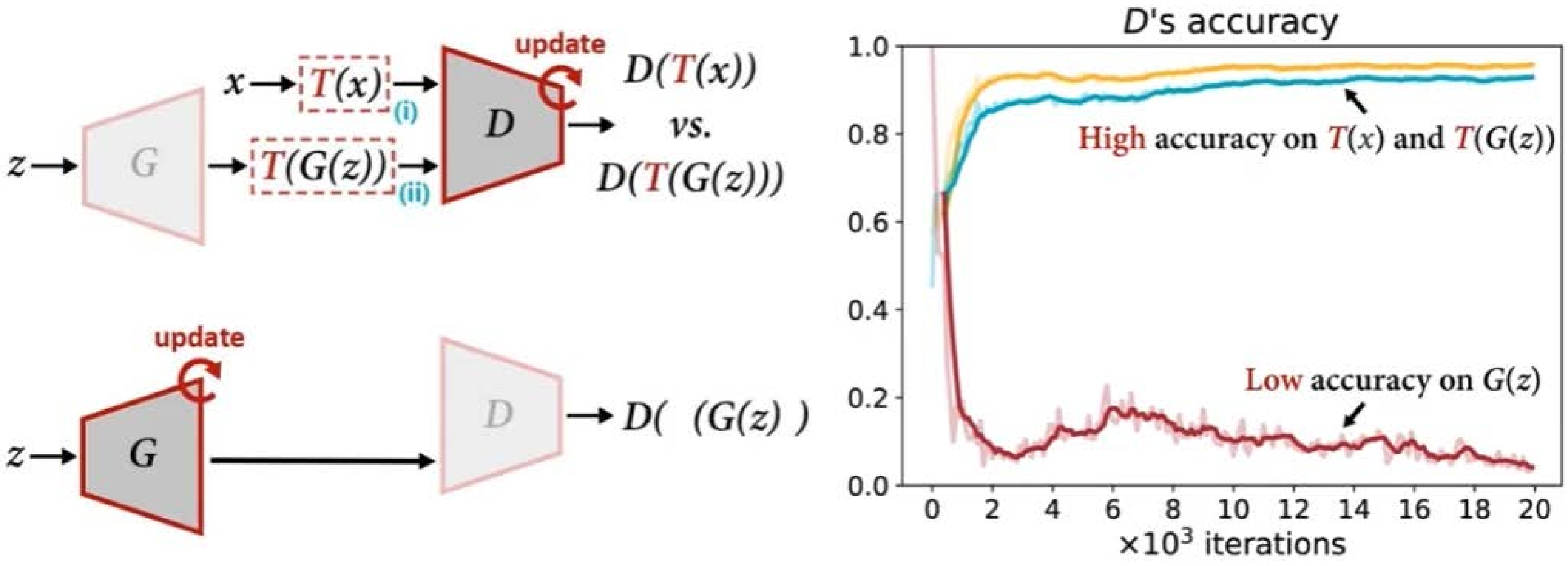
Artifacts from Translation

Artifacts from Cutout (DeVries et al.)

Augment reals only: the same artifacts appear on the generated images.

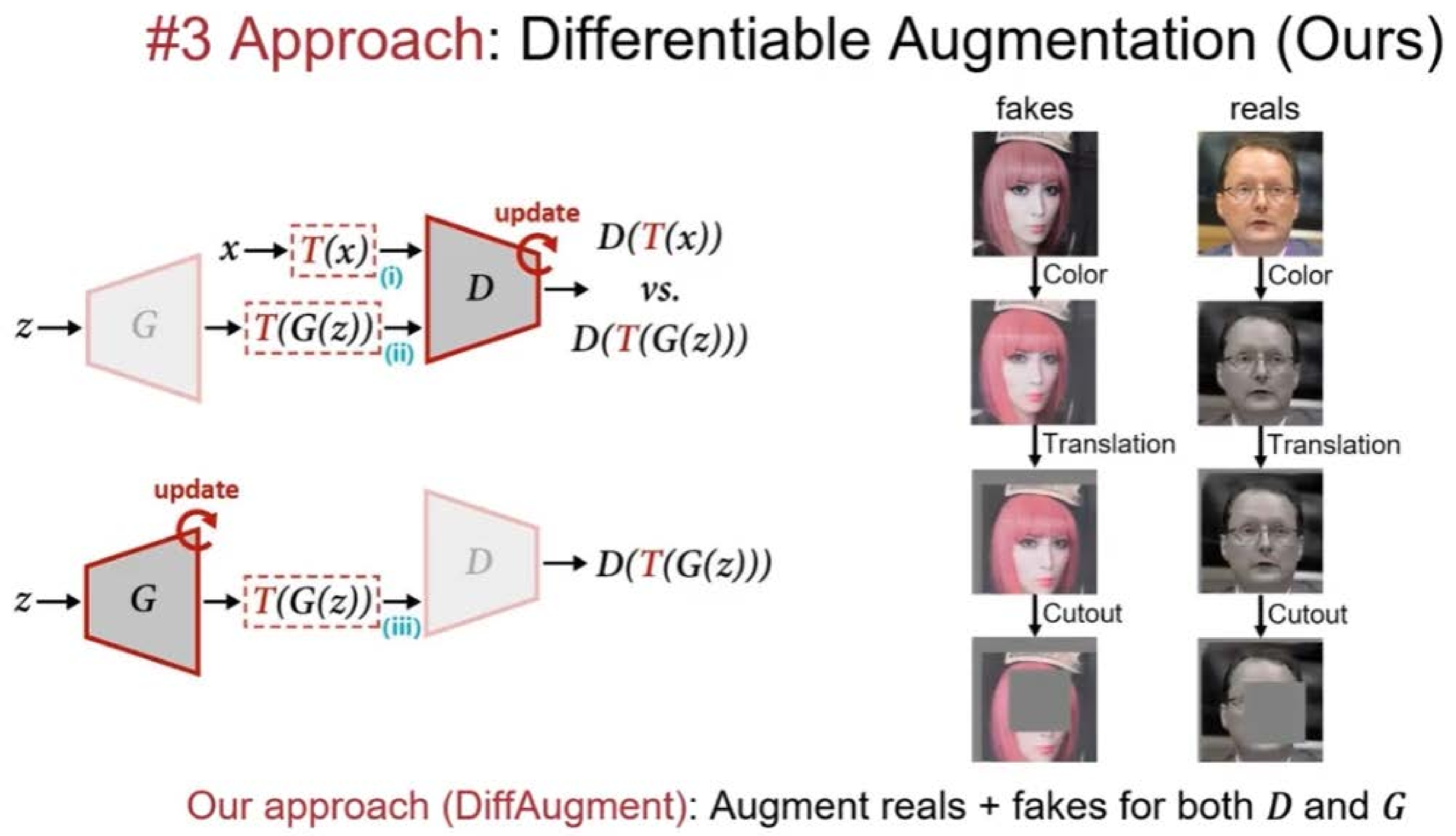
Intro

#2 Approach: Augment reals & fakes for *D* only



Augment ***D*** only: the unbalanced optimization cripples training.

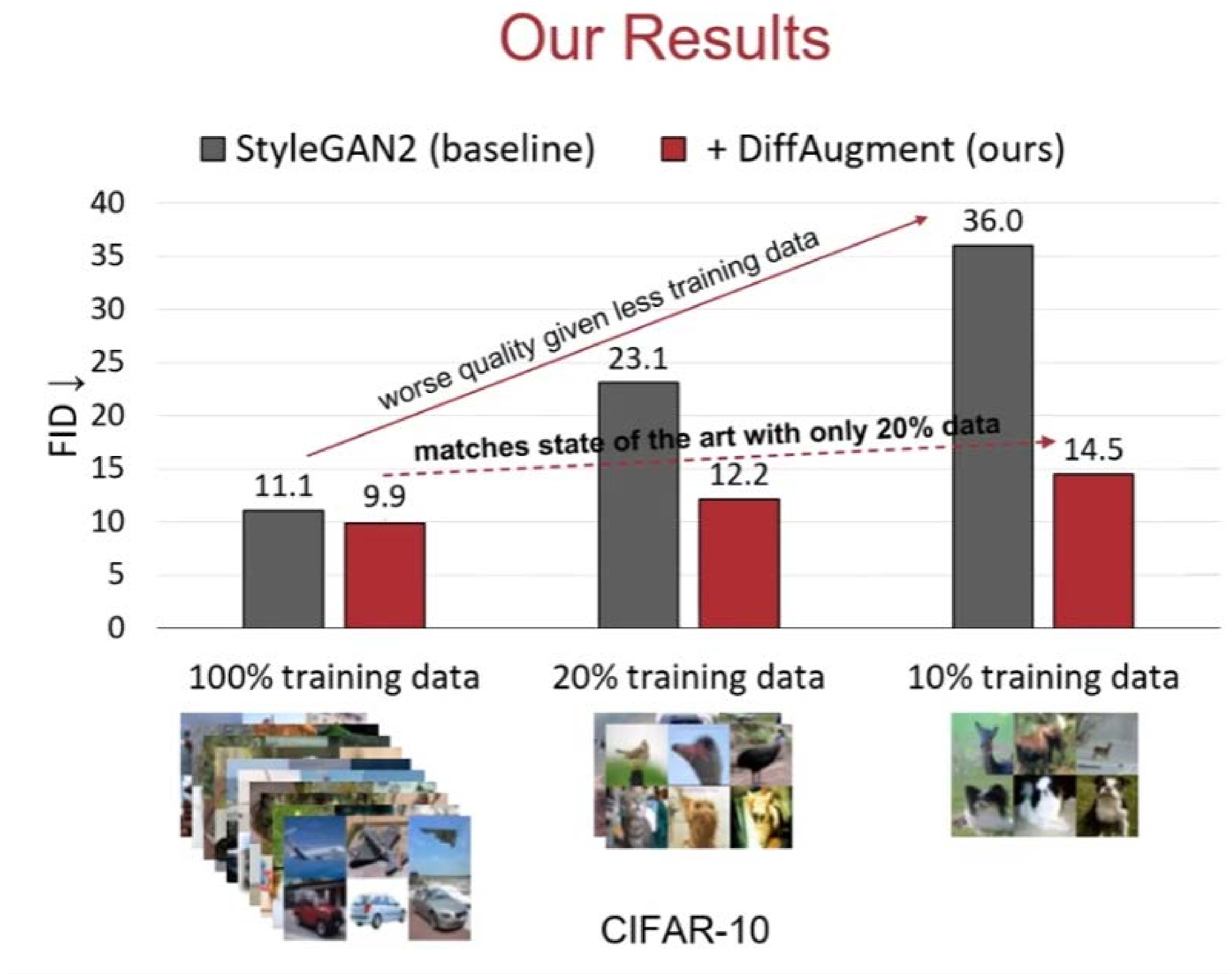
Intro



#3 Approach: Differentiable Augmentation (Ours)

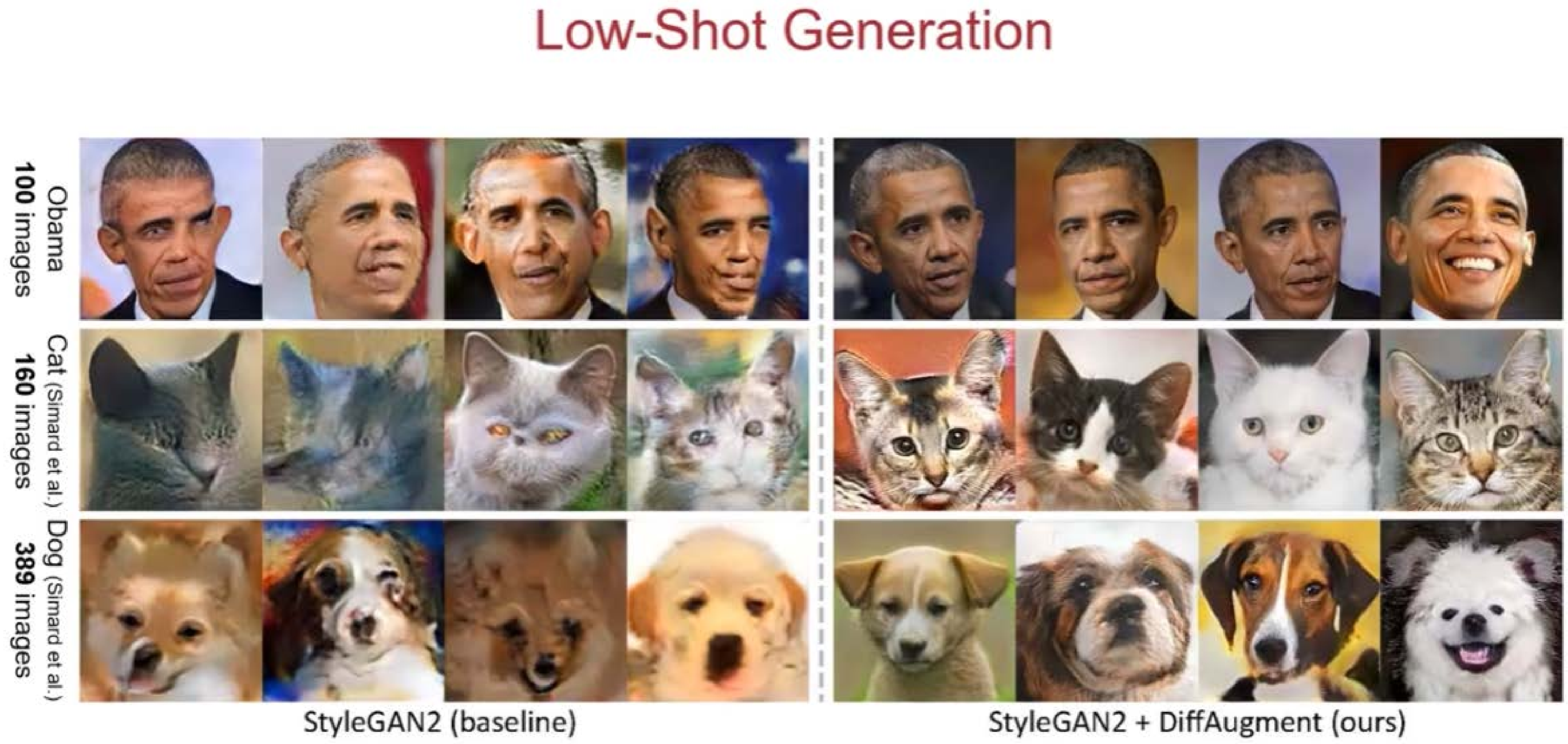
Our approach (DiffAugment): Augment reals + fakes for both D and G

Intro



Our Results

Intro



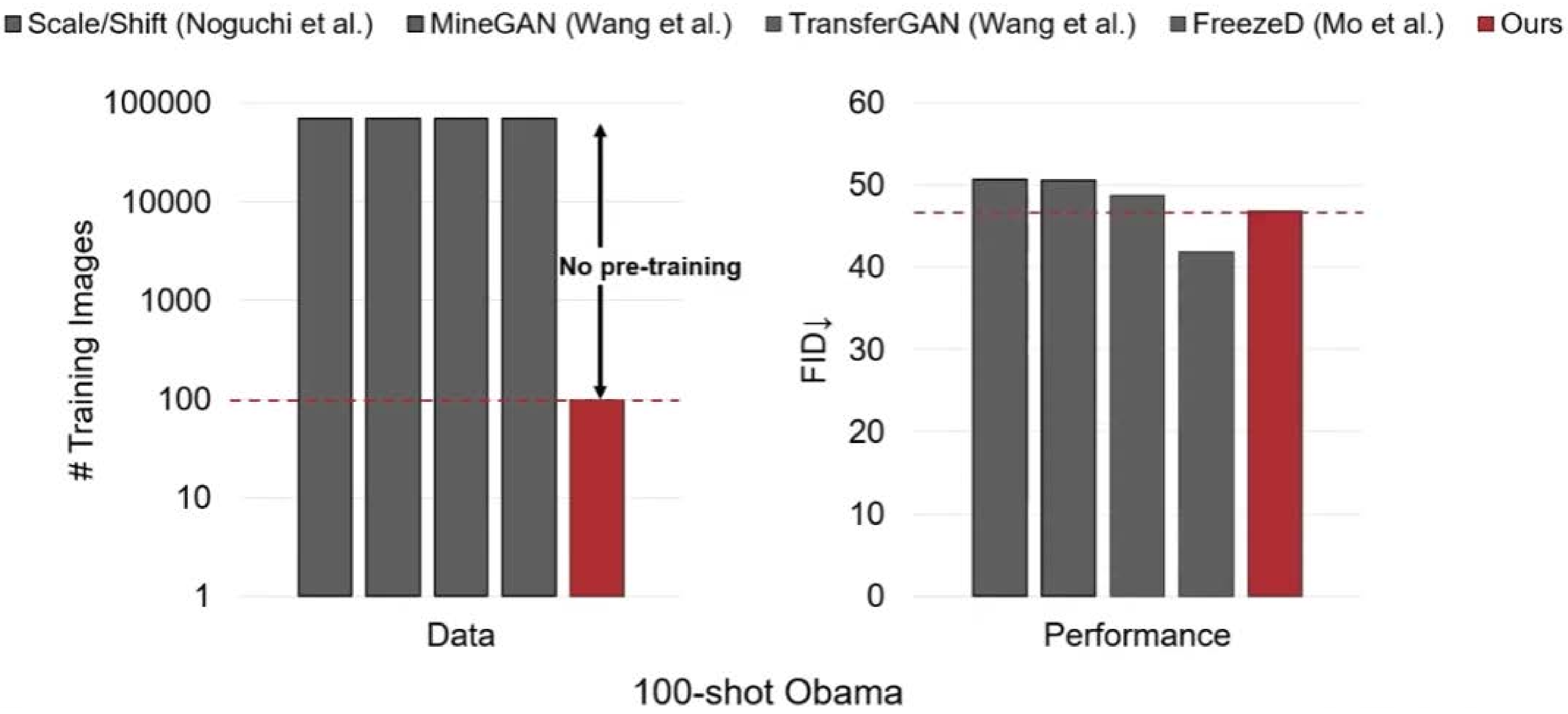
Low-Shot Generation

StyleGAN2 (baseline)

StyleGAN2 + DiffAugment (ours)

Intro

Fine-Tuning vs. Ours



Contribution

1) Because the discriminator is memorizing the exact training set. Propose Differentiable Augmentation, a method that improves data efficiency of GANs by imposing various types of differentiable augmentation on both real and fake samples.
2) Achieve a state-of-the-art FID of 6.80 with an IS of 100.8 on ImageNet 128 X 128, 2-4 X reductions of FID given 1,000 images on FFHQ and LSUN.

- 일반적인 GANs 모델의 학습에서, training data 전체 숫자의 10%, 20% 비율로 더 적은 수의 데이터(이미지)를 사용하면서도, 성능 감소는 최소화하도록 하는 새로운 관점의 augmentation 제안
- real image(x)뿐만 아니라 fake image (G(z))에도 augmentation(DiffAugment)을 적용하며, Discriminator Update과정 뿐만 아니라 Generator Update과정에서도 DiffAugment를 적용하여 성능 향상하고 training 안정화
- 여러 가지 모델(BigGAN, CR-BigGAN, StyleGAN2)별, 데이터 셋(CIFAR-10, CIFAR-100, FFHQ, LSUN)별로 성능 평가

Link : https://arxiv.org/pdf/2006.10738.pdf

Backgrounds

- **Regularization for GANs**
  - GAN training often requires additional regularization such as the instance noise, Jensen-Shannon regularization, gradient penalties, spectral normalization, adversarial defense regularization, and consistency regularization to penalize sudden changes in the discriminator's output within a local region of the input.

- **Data augmentation**
  - Transformation such as cropping, flipping, scaling, color jittering, and region making(Cutout) are commonly-used augmentation for vision models.
  - Different from the classifier training where the label is invariant to transformations of the input, the goal of generative models is to learn the data.

Original

Augmented

Base

Binary Mask

RGB Mask

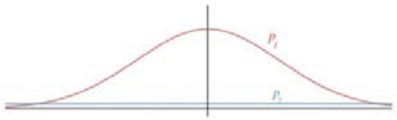
Backgrounds

- **Frechet inception distance (FID)**
  - FID is an evaluation metric for measuring the model performance by calculating the Wasserstein-2 distance of feature space representations between the generated outputs and the real images.
  - A low FID score means that the generated image is close to the real image distribution.

$$FID(x,g) = \left|\left|\mu_x - \mu_g\right|\right|_2^2 + Tr\left(\Sigma_x + \Sigma_g - 2\left(\Sigma_x\Sigma_g\right)^{\frac{1}{2}}\right)$$

- **Inceptions Score (IS)**
  - GAN의 성능평가에 생성된 영상의 품질, 생성된영상의 다양성 두 가지 기준을 사용: 생성된 영상의 품질과 다양성
  - KL-divergence는 두 데이터의 확률 분포가 얼마나 다른지 수치적으로 표현이 가능

$$InceptS \;=\; \exp\left(E_x KL\left(p(y|x)\,||\,p(y)\right)\right) = \exp\left(E_x E_{p(y|x)}\left[\log\left(\frac{p(y|x)}{p(y)}\right)\right]\right)$$



Methods

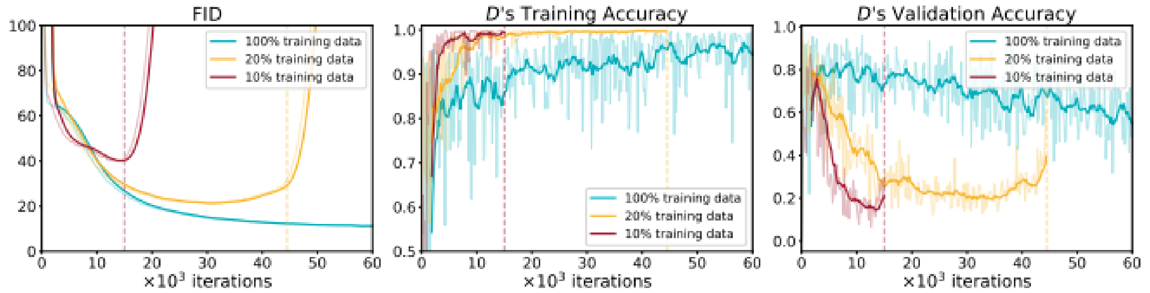


Figure 1: **BigGAN heavily deteriorates given a limited amount of data.** *left*: With 10% of CIFAR-10 data, FID increases shortly after the training starts, and the model then collapses (red curve). *middle*: the training accuracy of the discriminator D quickly saturates. *right*: the validation accuracy of D dramatically falls, indicating that D has memorized the exact training set and fails to generalize.

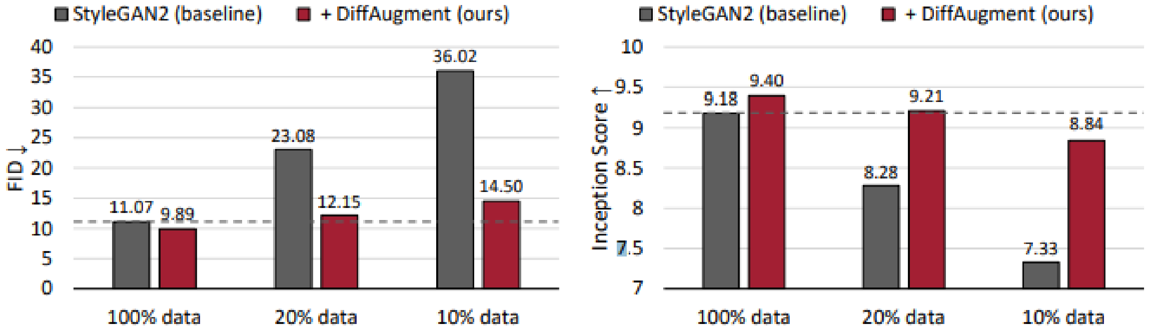


Figure 2: **Unconditional generation results on CIFAR-10.** StyleGAN2's performance drastically degrades given less training data. With DiffAugment, we are able to roughly match its FID and outperform its Inception Score (IS) using only 20% training data. FID and IS are measured using 10k samples; the validation set is used as the reference distribution for FID calculation.

Typical Loss Function of GAN

$$L_D = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[f_D(-D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_D(D(G(\boldsymbol{z})))],$$
$$L_G = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_G(-D(G(\boldsymbol{z})))].$$

Different Function Added on DiffAugment

$$f_D(x) = f_G(x) = \log\left(1 + e^x\right)$$
$$f_D(x) = \max(0, 1 + x) \text{ and } f_G(x) = x$$

Loss Function of DiffAugment

$$L_D = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[f_D(-D(T(\boldsymbol{x})))] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_D(D(G(\boldsymbol{z})))],$$
$$L_G = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_G(-D(G(\boldsymbol{z})))].$$

Network architecture : same (StyleGAN2, BigGAN, CR-BigGAN)

Augmentation : + DiffAugment

Methods

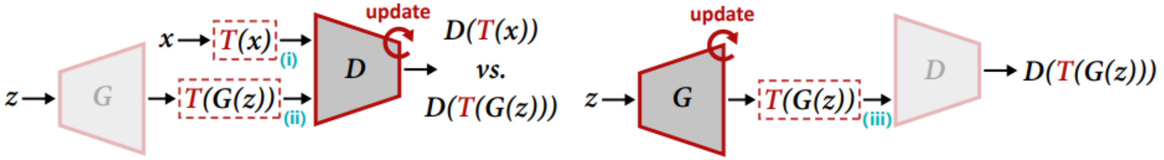
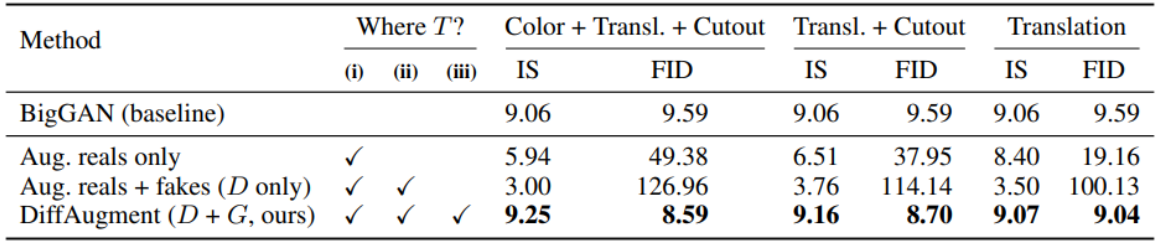


Figure 4: **Overview of DiffAugment** for updating D (left) and G (right). DiffAugment applies the augmentation T to both the real samples $\boldsymbol{x}$ and the generated output $G(\boldsymbol{z})$. When we update G, gradients need to be back-propagated through T, which requires T to be differentiable w.r.t. the input.



Method	Where T?			Color + Transl. + Cutout		Transl. + Cutout		Translation	
	(i)	(ii)	(iii)	IS	FID	IS	FID	IS	FID
BigGAN (baseline)				9.06	9.59	9.06	9.59	9.06	9.59
Aug. reals only	✓			5.94	49.38	6.51	37.95	8.40	19.16
Aug. reals + fakes (D only)	✓	✓		3.00	126.96	3.76	114.14	3.50	100.13
DiffAugment ($D + G$, ours)	✓	✓	✓	**9.25**	**8.59**	**9.16**	**8.70**	**9.07**	**9.04**

Table 1: **DiffAugment vs. vanilla augmentation strategies** on CIFAR-10 with 100% training data. "Augment reals only" applies augmentation T to (i) only (see Figure 4) and corresponds to Equations (3)-(4); "Augment D only" applies T to both reals (i) and fakes (ii), but not G (iii), and corresponds to Equations (5)-(6); "DiffAugment" applies T to reals (i), fakes (ii), and G (iii). (iii) requires T to be differentiable since gradients should be back-propagated through T to G. DiffAugment corresponds to Equations (7)-(8). IS and FID are measured using 10k samples; the validation set is the reference distribution. We select the snapshot with the best FID for each method. Results are averaged over 5 evaluation runs; all standard deviations are less than 1% relatively.

(i) Augment reals only

$$L_D = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[f_D(-D(T(\boldsymbol{x})))] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_D(D(G(\boldsymbol{z})))],$$
$$L_G = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_G(-D(G(\boldsymbol{z})))].$$

(ii) Augment reals & fakes for D only

$$L_D = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[f_D(-D(T(\boldsymbol{x})))] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_D(D(T(G(\boldsymbol{z}))))]$$
$$L_G = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_G(-D(G(\boldsymbol{z})))].$$

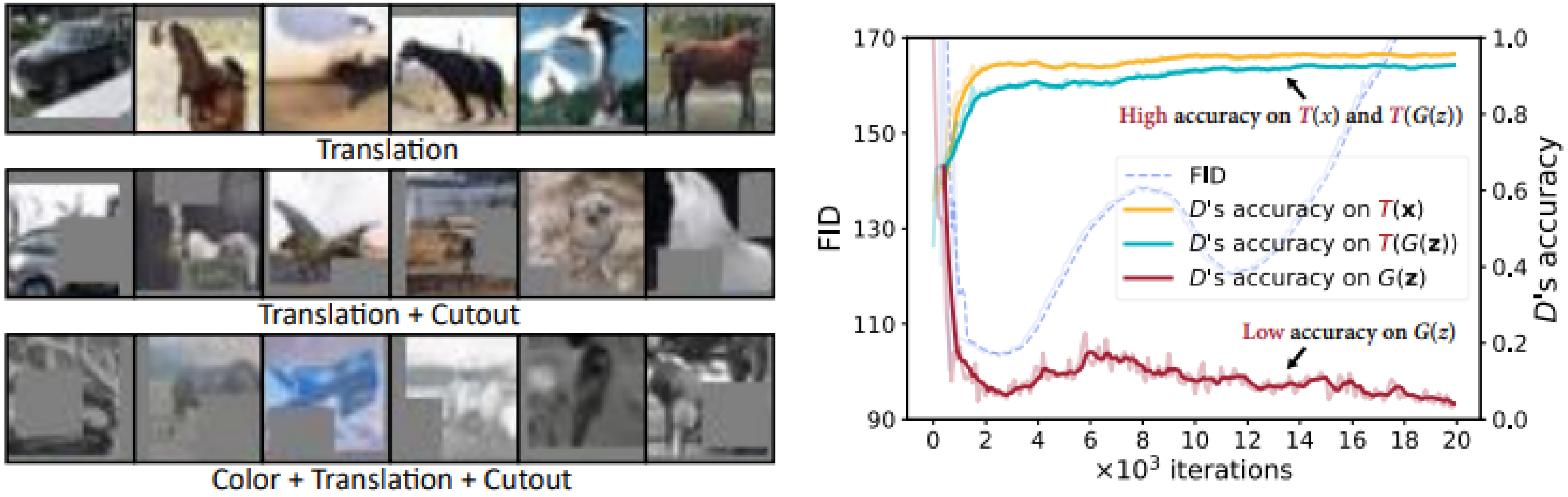
(iii) Differentiable Augmentation (Ours)

$$L_D = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}(\boldsymbol{x})[f_D(-D(T(\boldsymbol{x})))] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_D(D(T(G(\boldsymbol{z}))))],$$
$$L_G = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_G(-D(T(G(\boldsymbol{z}))))].$$

Network architecture : same (StyleGAN2, BigGAN, CR-BigGAN)

Augmentation : + DiffAugment

Methods



(a) "Augment reals only": the same augmentation artifacts appear on the generated images.

(b) "Augment D only": the unbalanced optimization between G and D cripples training.

Figure 5: **Understanding why vanilla augmentation strategies fail:** (a) "Augment reals only" mimics the same data distortion as introduced by the augmentations, *e.g.*, the translation padding, the Cutout square, and the color artifacts; (b) "Augment D only" diverges because of the unbalanced optimization — D perfectly classifies the augmented images (both $T(\boldsymbol{x})$ and $T(G(\boldsymbol{z}))$ but barely recognizes $G(\boldsymbol{z})$ (i.e., fake images without augmentation) from which G receives gradients.

Methods

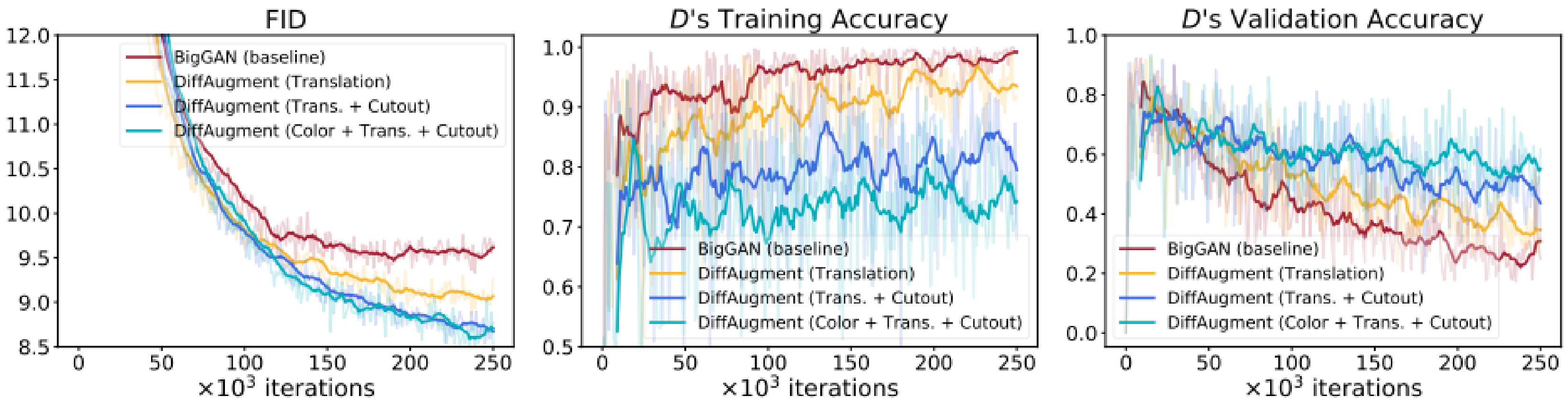
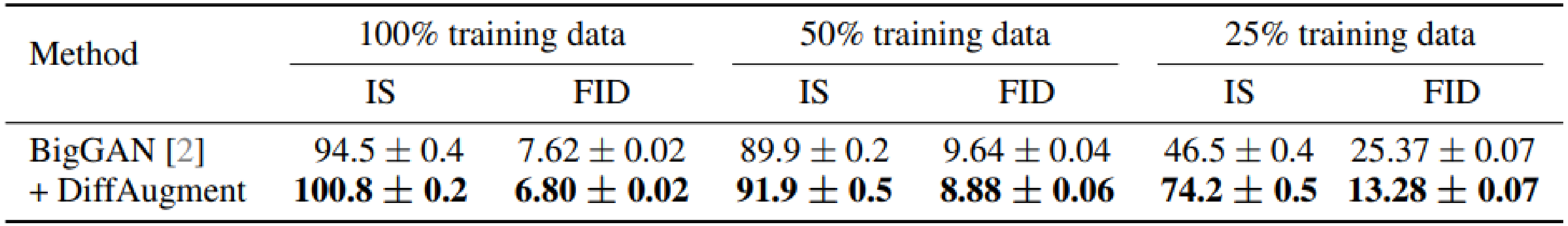


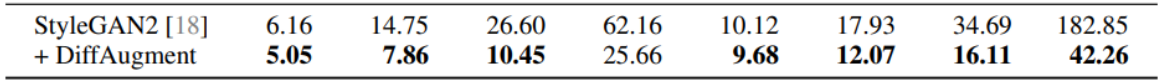
Figure 6: **Analysis of different types of DiffAugment** on CIFAR-10 with 100% training data. A stronger DiffAugment can dramatically reduce the gap between the discriminator's training accuracy (middle) and validation accuracy (right), leading to a better convergence (left).



Method	100% training data		50% training data		25% training data	
	IS	FID	IS	FID	IS	FID
BigGAN [2]	94.5 ± 0.4	7.62 ± 0.02	89.9 ± 0.2	9.64 ± 0.04	46.5 ± 0.4	25.37 ± 0.07
+ DiffAugment	**100.8 ± 0.2**	**6.80 ± 0.02**	**91.9 ± 0.5**	**8.88 ± 0.06**	**74.2 ± 0.5**	**13.28 ± 0.07**

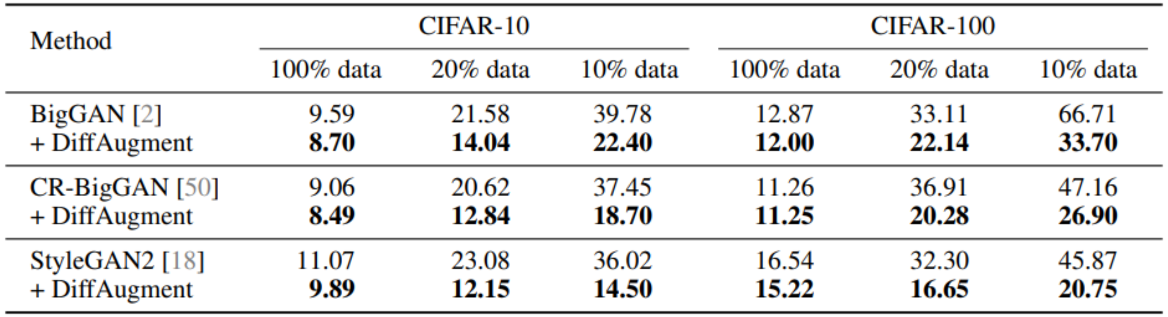
Table 2: **ImageNet** 128×128 results without the truncation trick [2]. IS and FID are measured using 50k samples; the validation set is used as the reference distribution for FID. We select the snapshot with the best FID for each method. We report means and standard deviations over 3 evaluation runs.

Experiments



StyleGAN2 [18]	6.16	14.75	26.60	62.16	10.12	17.93	34.69	182.85
+ DiffAugment	**5.05**	**7.86**	**10.45**	25.66	**9.68**	**12.07**	**16.11**	**42.26**

Table 3: **FFHQ and LSUN-Cat** results with 1k, 5k, 10k, and 30k training samples. With the fixed *Color* + *Translation* + *Cutout* DiffAugment, our method improves the StyleGAN2 baseline and is on par with a concurrent work ADA [16]. FID is measured using 50k generated samples; the full training set is used as the reference distribution. We select the snapshot with the best FID for each method. Results are averaged over 5 evaluation runs; all standard deviations are less than 1% relatively.



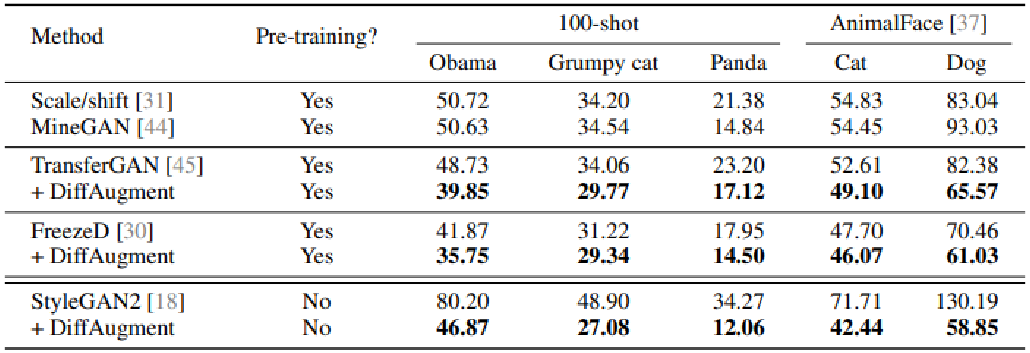
Method	CIFAR-10			CIFAR-100		
	100% data	20% data	10% data	100% data	20% data	10% data
BigGAN [2]	9.59	21.58	39.78	12.87	33.11	66.71
+ DiffAugment	**8.70**	**14.04**	**22.40**	**12.00**	**22.14**	**33.70**
CR-BigGAN [50]	9.06	20.62	37.45	11.26	36.91	47.16
+ DiffAugment	**8.49**	**12.84**	**18.70**	**11.25**	**20.28**	**26.90**
StyleGAN2 [18]	11.07	23.08	36.02	16.54	32.30	45.87
+ DiffAugment	**9.89**	**12.15**	**14.50**	**15.22**	**16.65**	**20.75**

Table 4: **CIFAR-10 and CIFAR-100** results. We select the snapshot with the best FID for each method. Results are averaged over 5 evaluation runs; all standard deviations are less than 1% relatively. We use 10k samples and the validation set as the reference distribution for FID calculation, as done in prior work [50]. Concurrent works [14, 16] use a different protocol: 50k samples and the training set as the reference distribution. If we adopt this evaluation protocol, our BigGAN + DiffAugment achieves an FID of 4.61, CR-BigGAN + DiffAugment achieves an FID of 4.30, and StyleGAN2 + DiffAugment achieves an FID of 5.79.

For DiffAugment, we adopt Translation + Cutout for the BigGAN models,

Color + Cutout for StyleGAN2 with 100% data,

and Color + Translation + Cutout for StyleGAN2 with 10% or 20% data



Method	Pre-training?	100-shot			AnimalFace [37]	
		Obama	Grumpy cat	Panda	Cat	Dog
Scale/shift [31]	Yes	50.72	34.20	21.38	54.83	83.04
MineGAN [44]	Yes	50.63	34.54	14.84	54.45	93.03
TransferGAN [45]	Yes	48.73	34.06	23.20	52.61	82.38
+ DiffAugment	Yes	**39.85**	**29.77**	**17.12**	**49.10**	**65.57**
FreezeD [30]	Yes	41.87	31.22	17.95	47.70	70.46
+ DiffAugment	Yes	**35.75**	**29.34**	**14.50**	**46.07**	**61.03**
StyleGAN2 [18]	No	80.20	48.90	34.27	71.71	130.19
+ DiffAugment	No	**46.87**	**27.08**	**12.06**	**42.44**	**58.85**

Table 5: **Low-shot generation** results. With only **100** (Obama, Grumpy cat, Panda), **160** (Cat), or **389** (Dog) training images, our method is on par with the transfer learning algorithms that are pre-trained with **70,000** images. FID is measured using 5k generated samples; the training set is the reference distribution. We select the snapshot with the best FID for each method.

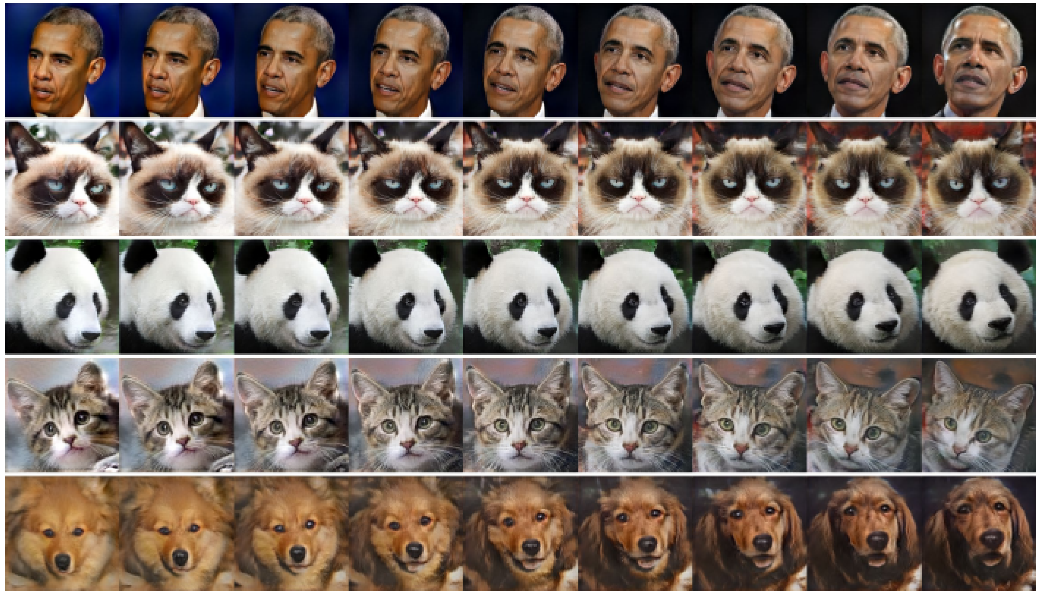
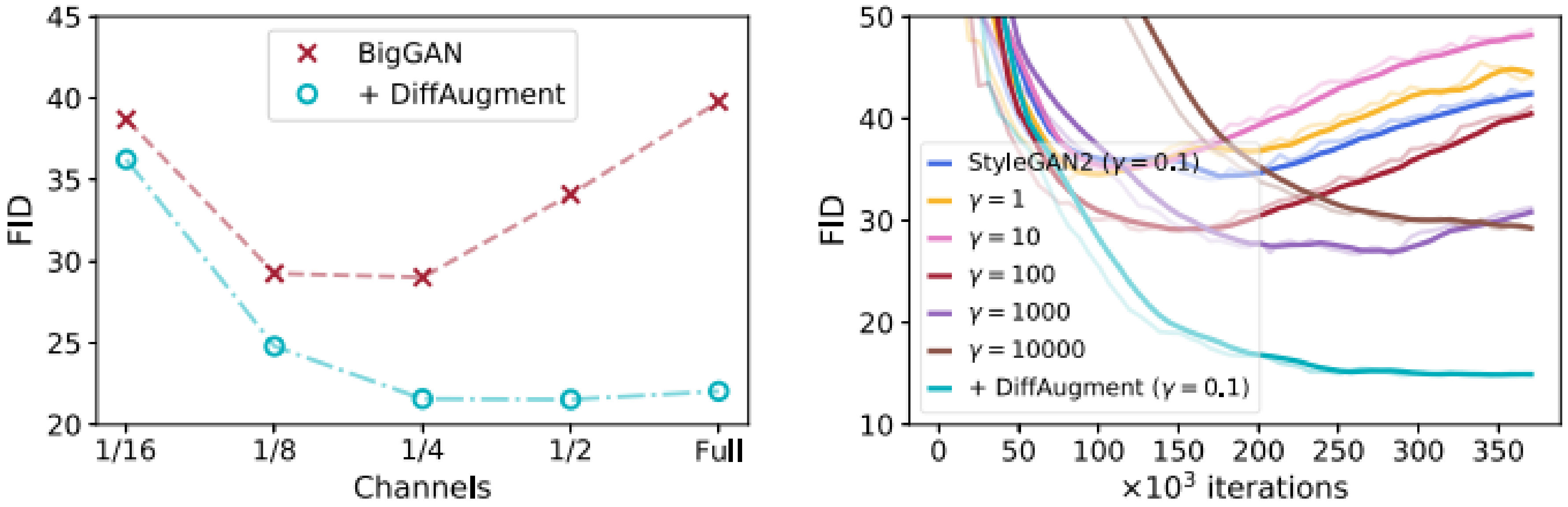


Figure 7: **Style space interpolation** of our method for low-shot generation without pre-training. The smooth interpolation results suggest little overfitting of our method even given small datasets.

Experiments



(a) Impact of model size.

(b) Impact of R_1 regularization γ.

Figure 8: **Analysis of smaller models or stronger regularization** on CIFAR-10 with 10% training data. (a) Smaller models reduce overfitting for the BigGAN baseline, while our method dominates its performance at all model capacities. (b) Over a wide sweep of the R_1 regularization γ for the baseline StyleGAN2, its best FID (26.87) is still much worse than ours (14.50).

Experiments

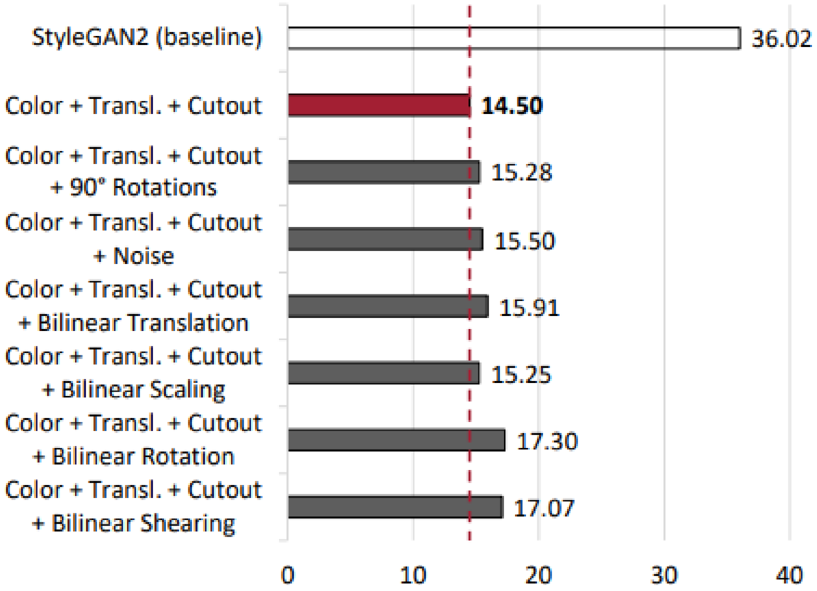


Figure 9: Various types of DiffAugment consistently outperform the baseline. We report StyleGAN2's FID on CIFAR-10 with 10% training data.

(i) Augment reals only

$$L_D = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[f_D(-D(T(\boldsymbol{x})))] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_D(D(G(\boldsymbol{z})))],$$

$$L_G = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_G(-D(G(\boldsymbol{z})))].$$

(i i) Augment reals & fakes for D only

$$L_D = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[f_D(-D(T(\boldsymbol{x})))] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_D(D(T(G(\boldsymbol{z}))))],$$

$$L_G = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_G(-D(G(\boldsymbol{z})))].$$

(iii) Differentiable Augmentation (Ours)

$$L_D = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}(\boldsymbol{x})[f_D(-D(T(\boldsymbol{x})))] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_D(D(T(G(\boldsymbol{z}))))], \tag{7}$$

$$L_G = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[f_G(-D(T(G(\boldsymbol{z}))))]. \tag{8}$$

Results

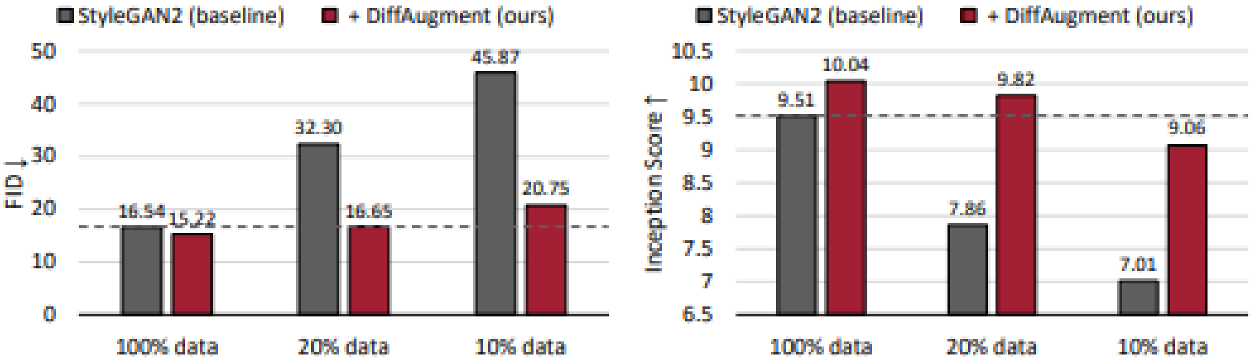
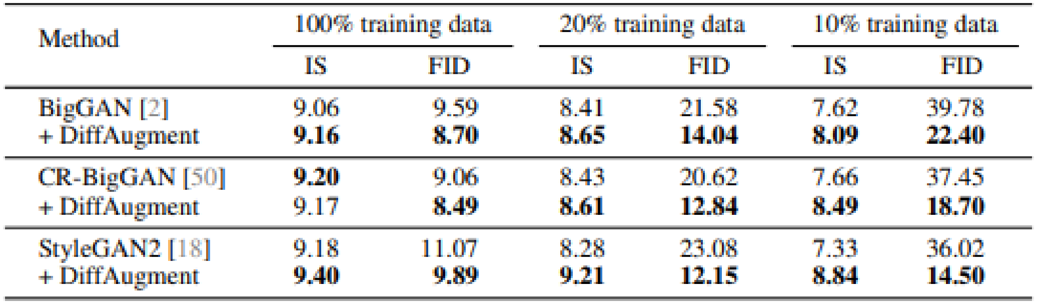


Figure 10: Unconditional generation results on **CIFAR-100**. We are able to roughly match StyleGAN2's FID and outperform its IS using only 20% training data.



Method	100% training data		20% training data		10% training data	
	IS	FID	IS	FID	IS	FID
BigGAN [2]	9.06	9.59	8.41	21.58	7.62	39.78
+ DiffAugment	**9.16**	**8.70**	**8.65**	**14.04**	**8.09**	**22.40**
CR-BigGAN [50]	**9.20**	9.06	8.43	20.62	7.66	37.45
+ DiffAugment	9.17	**8.49**	**8.61**	**12.84**	**8.49**	**18.70**
StyleGAN2 [18]	9.18	11.07	8.28	23.08	7.33	36.02
+ DiffAugment	**9.40**	**9.89**	**9.21**	**12.15**	**8.84**	**14.50**

Table 6: **CIFAR-10 results.** IS and FID are measured using 10k samples; the validation set is the reference distribution for FID calculation. We select the snapshot with the best FID for each method. Results are averaged over 5 evaluation runs; all standard deviations are less than 1% relatively.

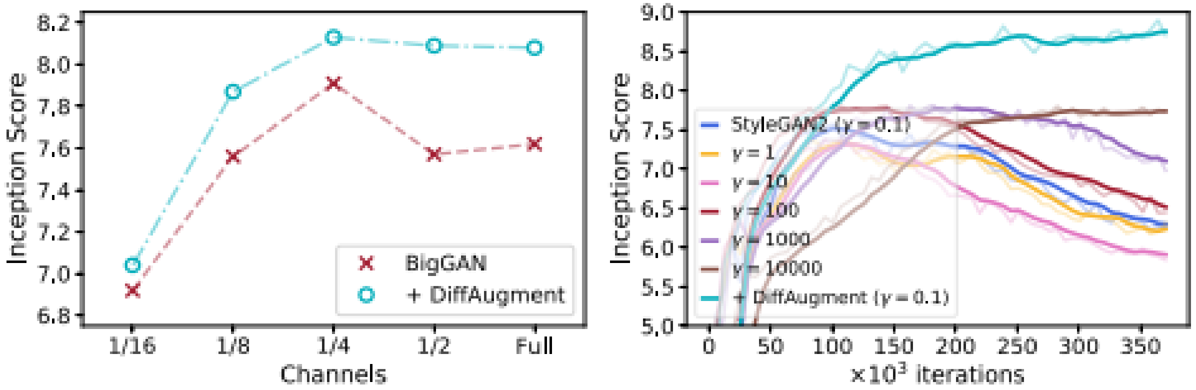
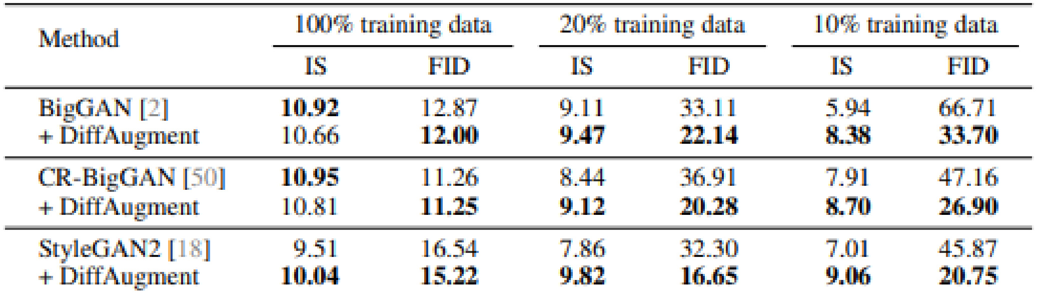


Figure 11: **Analysis of smaller models or stronger regularization** on CIFAR-10 with **10%** training data. *left*: Smaller models reduce overfitting for the BigGAN baseline, while our method outperforms it at all the model capacities. *right*: Over a wide sweep of the R_1 regularization γ for the baseline StyleGAN2, its best IS (7.75) is still **12%** worse than ours (8.84).



Method	100% training data		20% training data		10% training data	
	IS	FID	IS	FID	IS	FID
BigGAN [2]	**10.92**	12.87	9.11	33.11	5.94	66.71
+ DiffAugment	10.66	**12.00**	**9.47**	**22.14**	**8.38**	**33.70**
CR-BigGAN [50]	**10.95**	11.26	8.44	36.91	7.91	47.16
+ DiffAugment	10.81	**11.25**	**9.12**	**20.28**	**8.70**	**26.90**
StyleGAN2 [18]	9.51	16.54	7.86	32.30	7.01	45.87
+ DiffAugment	**10.04**	**15.22**	**9.82**	**16.65**	**9.06**	**20.75**

Table 7: **CIFAR-100 results.** IS and FID are measured using 10k samples; the validation set is the reference distribution for FID calculation. We select the snapshot with the best FID for each method. Results are averaged over 5 evaluation runs; all standard deviations are less than 1% relatively.

Results

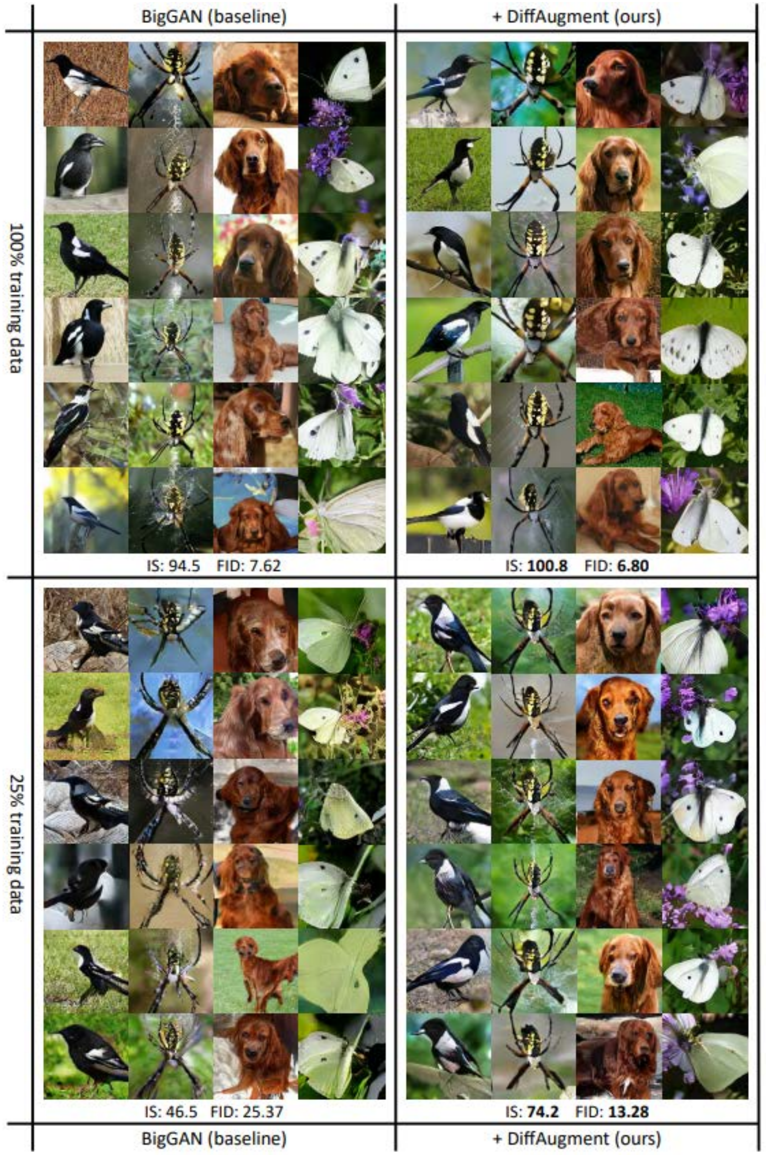


Figure 12: Qualitative comparison on **ImageNet** 128×128 without the truncation trick [2].

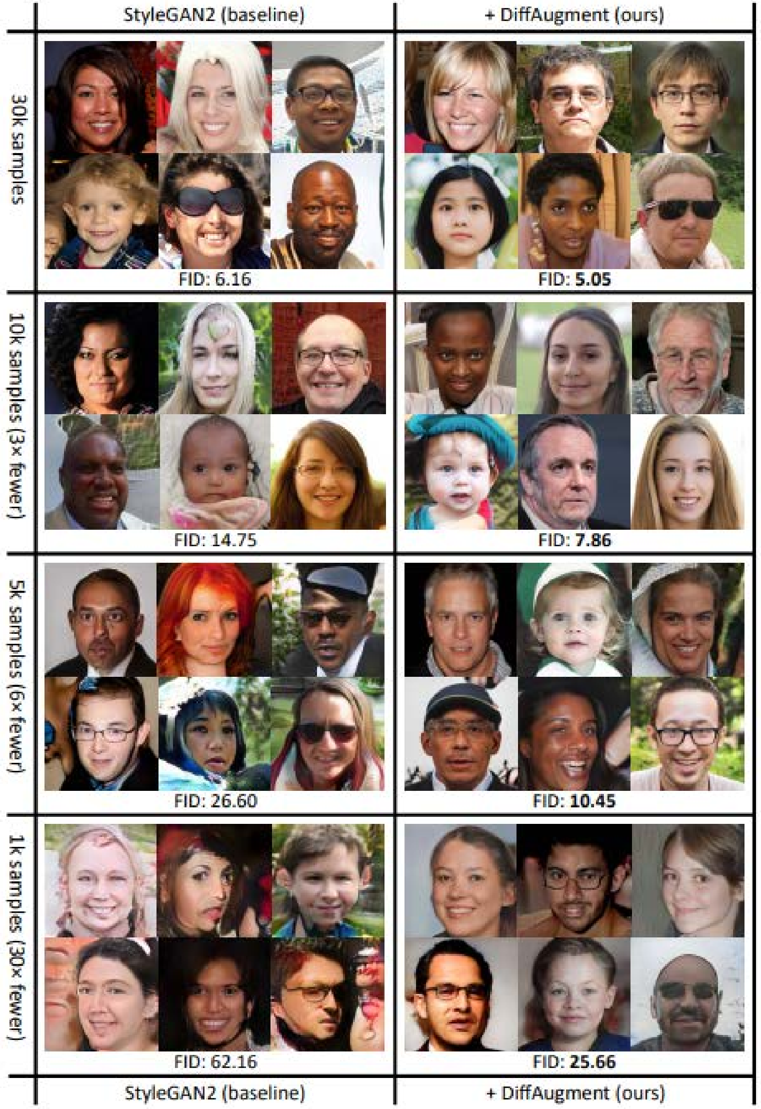


Figure 13: Qualitative comparison on **FFHQ** at 256×256 resolution with 1k, 5k, 10k, and 30k training images. Our method consistently outperforms the StyleGAN2 baselines [18] under different data percentage settings.

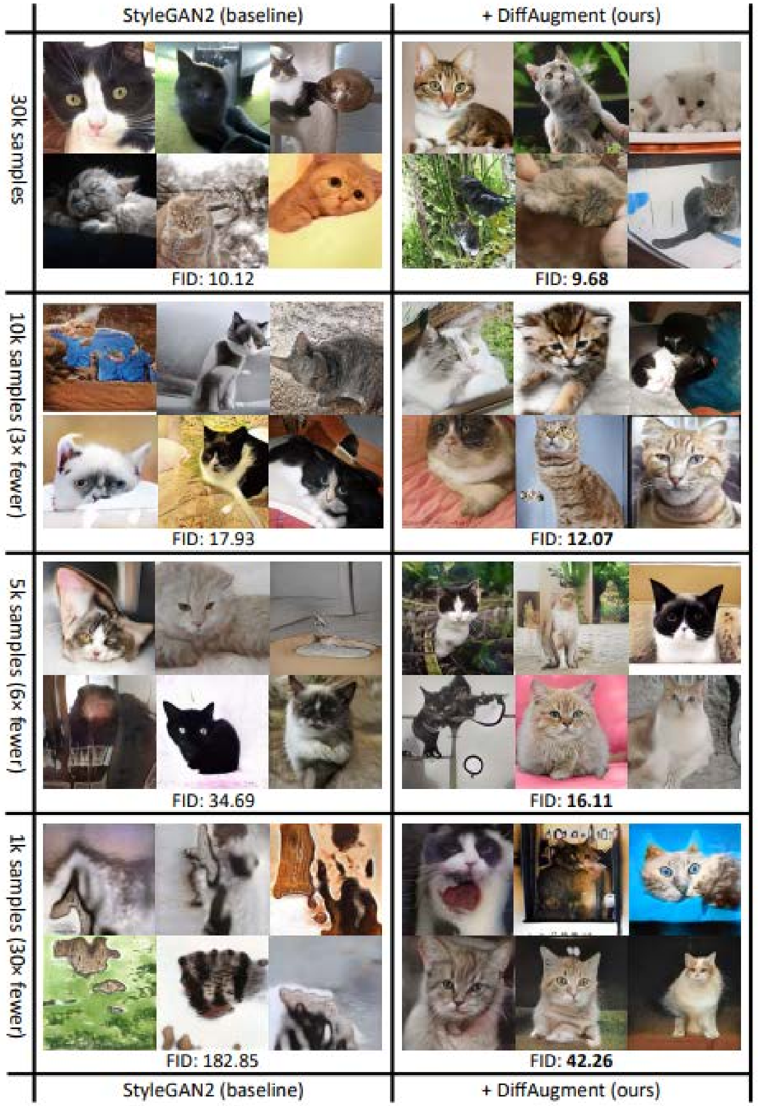


Figure 14: Qualitative comparison on **LSUN-cat** at 256×256 resolution with 1k, 5k, 10k, and 30k training images. Our method consistently outperforms the StyleGAN2 baselines [18] under different data percentage settings.

Results

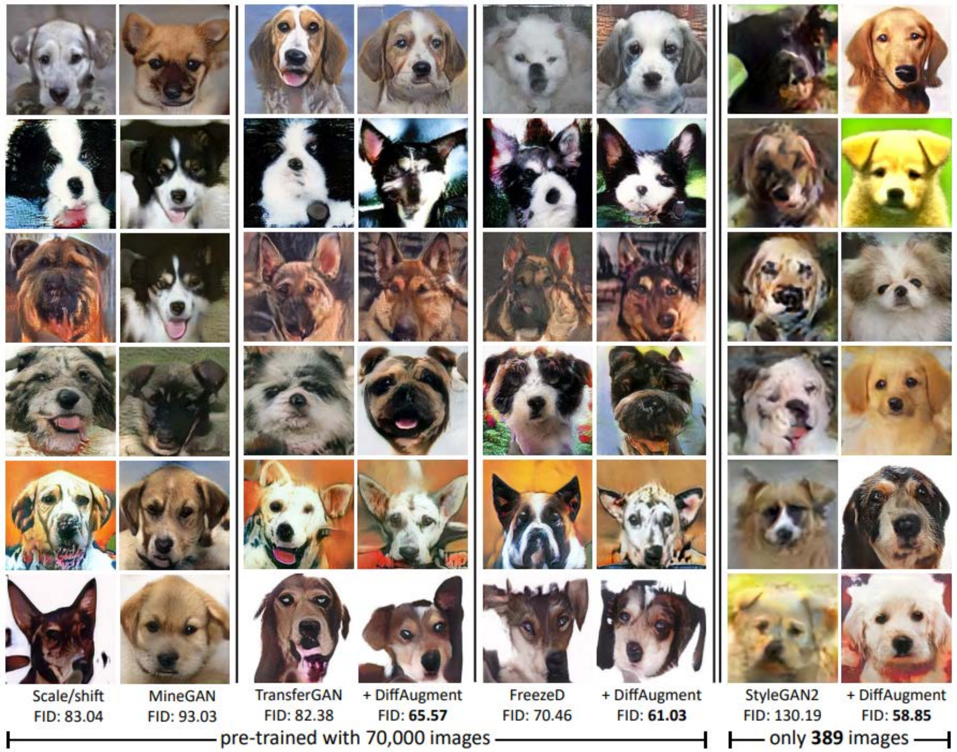


Figure 19: Qualitative comparison on the **AnimalFace-dog** [37] dataset.

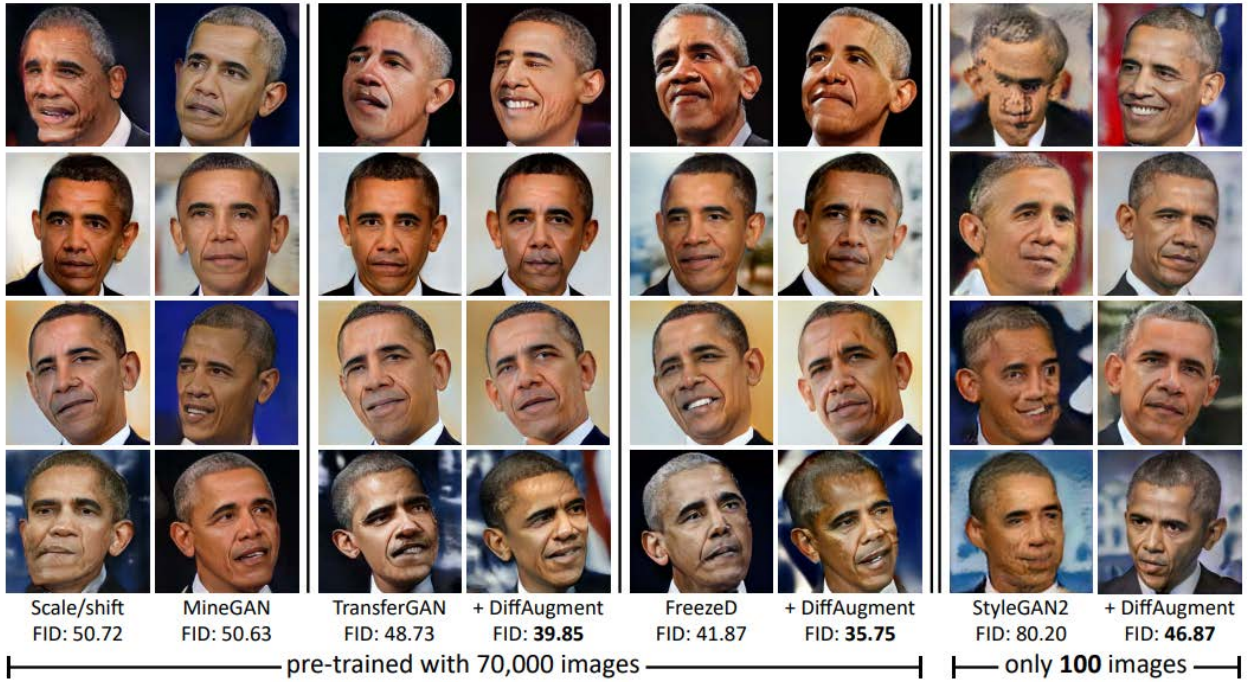
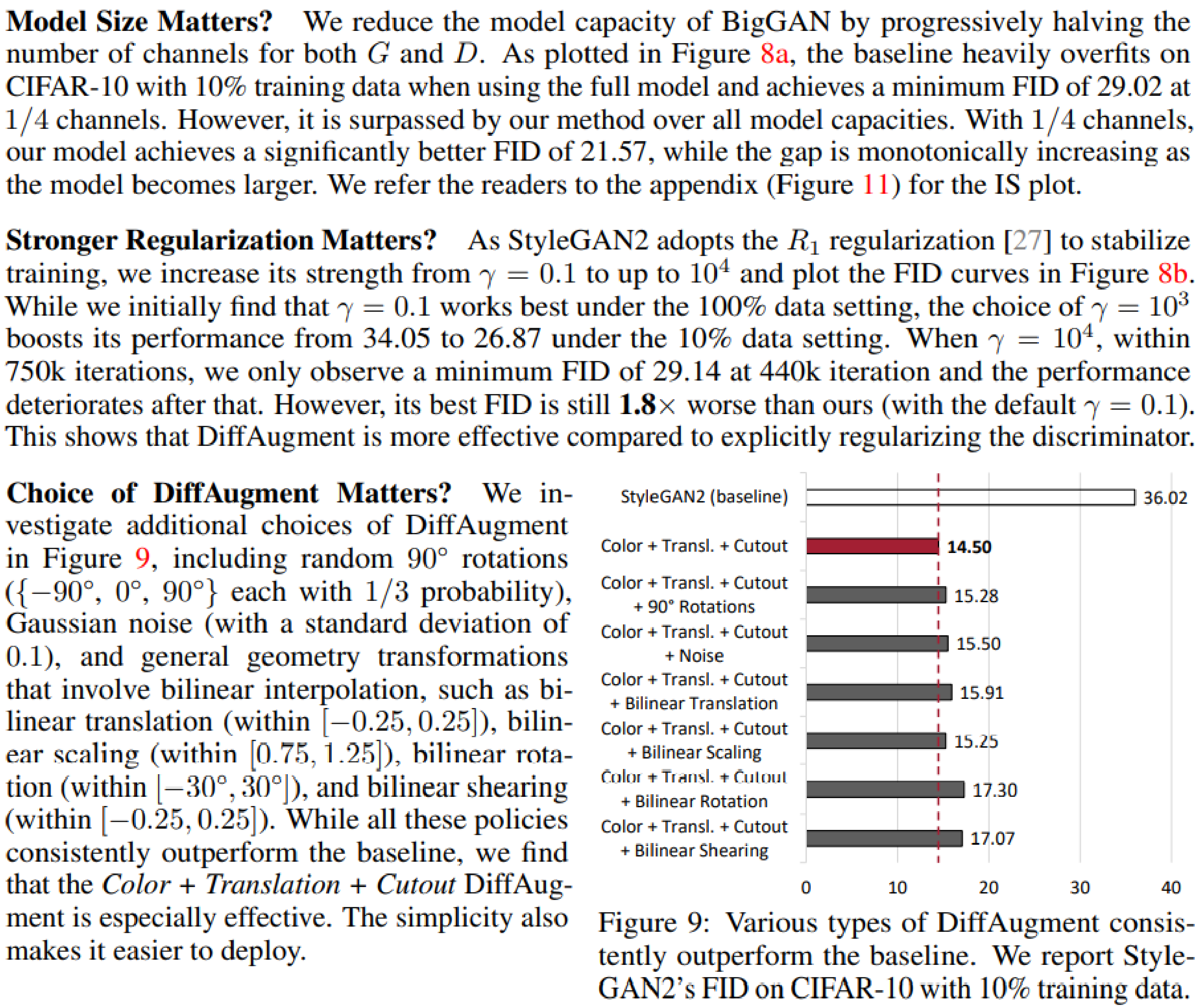


Figure 20: Qualitative comparison on the **100-shot Obama** dataset.

Results



Model Size Matters? We reduce the model capacity of BigGAN by progressively halving the number of channels for both G and D. As plotted in Figure 8a, the baseline heavily overfits on CIFAR-10 with 10% training data when using the full model and achieves a minimum FID of 29.02 at $1/4$ channels. However, it is surpassed by our method over all model capacities. With $1/4$ channels, our model achieves a significantly better FID of 21.57, while the gap is monotonically increasing as the model becomes larger. We refer the readers to the appendix (Figure 11) for the IS plot.

Stronger Regularization Matters? As StyleGAN2 adopts the R_1 regularization [27] to stabilize training, we increase its strength from $\gamma = 0.1$ to up to 10^4 and plot the FID curves in Figure 8b. While we initially find that $\gamma = 0.1$ works best under the 100% data setting, the choice of $\gamma = 10^3$ boosts its performance from 34.05 to 26.87 under the 10% data setting. When $\gamma = 10^4$, within 750k iterations, we only observe a minimum FID of 29.14 at 440k iteration and the performance deteriorates after that. However, its best FID is still **1.8**× worse than ours (with the default $\gamma = 0.1$). This shows that DiffAugment is more effective compared to explicitly regularizing the discriminator.

Choice of DiffAugment Matters? We investigate additional choices of DiffAugment in Figure 9, including random 90° rotations ($\{-90°, 0°, 90°\}$ each with $1/3$ probability), Gaussian noise (with a standard deviation of 0.1), and general geometry transformations that involve bilinear interpolation, such as bilinear translation (within $[-0.25, 0.25]$), bilinear scaling (within $[0.75, 1.25]$), bilinear rotation (within $[-30°, 30°]$), and bilinear shearing (within $[-0.25, 0.25]$). While all these policies consistently outperform the baseline, we find that the *Color* + *Translation* + *Cutout* DiffAugment is especially effective. The simplicity also makes it easier to deploy.

StyleGAN2 (baseline)
36.02
Color + Transl. + Cutout
14.50
Color + Transl. + Cutout + 90° Rotations
15.28
Color + Transl. + Cutout + Noise
15.50
Color + Transl. + Cutout + Bilinear Translation
15.91
Color + Transl. + Cutout + Bilinear Scaling
15.25
Color + Transl. + Cutout + Bilinear Rotation
17.30
Color + Transl. + Cutout + Bilinear Shearing
17.07
0
10
20
30
40

Figure 9: Various types of DiffAugment consistently outperform the baseline. We report StyleGAN2's FID on CIFAR-10 with 10% training data.

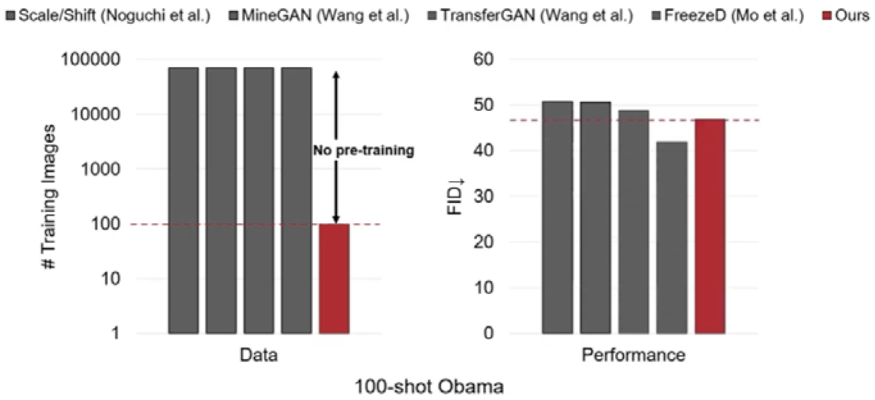
Conclusion

- DiffAugment reveals valuable observations that augmenting both real and fake samples effectively prevents the discriminator from over-fitting, and that the augmentation must be differentiable to enable both generator and discriminator training.
- Extensive experiments consistently demonstrate its benefits with different network architectures (StyleGAN2 and BigGAN), supervision settings, and objective functions, across multiple datasets (ImageNet, CIFAR, FFHQ, LSUN, and 100-shot datasets).

- We can get better performance on our model

  than 3 of 4 Fine-Tuning models.

Fine-Tuning vs. Ours



Thank you!

Reference

- Video Presentation Link : https://www.youtube.com/watch?v=SsqcjS6SVM0