Data InStance Prior (DISP) in Generative Adversarial Networks

Puneet Mangla[3], **Nupur Kumari**[2], **Mayank Singh**[2] & Balaji Krishnamurthy[1], Vineeth N Balasubramanian[3]

[1]Media and Data Science Research lab, Adobe [2]CMU [3]IIT Hyderabad, India

November 10, 2021

Contributions

Challange: In limited data regimes, GAN training typically diverges, and results in low quality and lack diversity.

Contributions

Challange: In limited data regimes, GAN training typically diverges, and results in low quality and lack diversity.

- We propose **D**ata **I**n**S**tance **P**rior (DISP) - novel transfer learning technique for GANs in low-data regime.

Contributions

Challange: In limited data regimes, GAN training typically diverges, and results in low quality and lack diversity.

- We propose **D**ata **I**n**S**tance **P**rior (DISP) - novel transfer learning technique for GANs in low-data regime.
- achieves SOTA in terms of image quality and diversity on few-shot, limited and large-scale image generation benchmarks.

Contributions

Challange: In limited data regimes, GAN training typically diverges, and results in low quality and lack diversity.

- We propose **D**ata **I**n**S**tance **P**rior (DISP) - novel transfer learning technique for GANs in low-data regime.
- achieves SOTA in terms of image quality and diversity on few-shot, limited and large-scale image generation benchmarks.

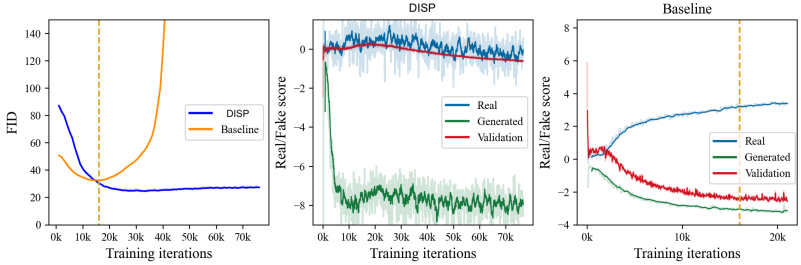


Figure: Comparison between DISP and Baseline when trained on 10% data of CIFAR-100.

Approach - Motivation

- Exploiting knowledge extracted from self-supervised/supervised networks, pre-trained on a rich and diverse source domains.

[1] Ke Li and Jitendra Malik. Implicit maximum likelihood estimation

Approach - Motivation

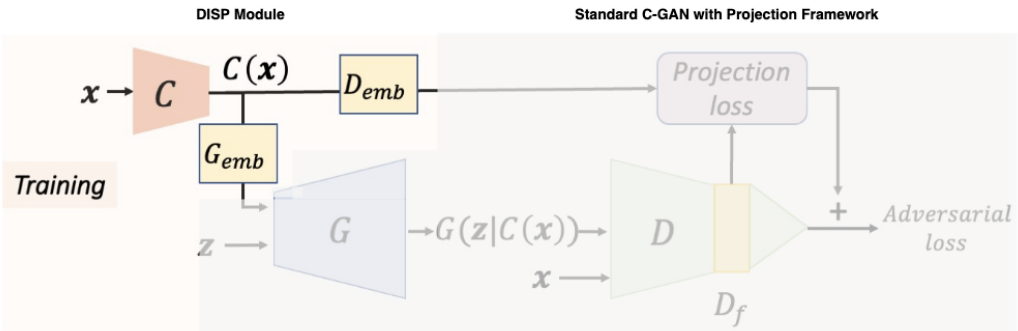
- Exploiting knowledge extracted from self-supervised/supervised networks, pre-trained on a rich and diverse source domains.

- Motivated from the reconstructive framework of IMLE [1] - propose a regularizer to prevent mode collapse and discriminator overfitting.

[1] Ke Li and Jitendra Malik. Implicit maximum likelihood estimation

Approach - Training

Given a pre-trained feature extractor $C : \mathbb{R}^p \to \mathbb{R}^d$, which is trained on a source domain using supervisory signals or self-supervision, we use its output $C(\mathbf{x})$ as the conditional information during GAN training.



Approach - Loss Function

$$
\begin{aligned}
L_D =& \mathbb{E}_{\mathbf{x}\sim q(x)}[\max(0, 1 - D(\mathbf{x}, C(\mathbf{x})))] \\
&+ \mathbb{E}_{\mathbf{x}\sim q(x), \mathbf{z}\sim p(\mathbf{z})}[\max(0, 1 + D(G(\mathbf{z}|C(\mathbf{x})), C(\mathbf{x})))] \\
L_G =& - \mathbb{E}_{\mathbf{x}\sim q(x), \mathbf{z}\sim p(\mathbf{z})}[D(G(\mathbf{z}|C(\mathbf{x})), C(\mathbf{x}))]
\end{aligned} \tag{1}
$$

$D(\mathbf{x}, \mathbf{y}) = D_{emb}(\mathbf{y}) \cdot D_f(\mathbf{x}) + D_l \circ D_f(\mathbf{x})$is the **c-GAN projection loss**[2]

[2]Takeru Miyato and Masanori Koyama. cgans with projection discriminator.

Approach - Loss Function

$$
\begin{aligned}
L_D =& \mathbb{E}_{\mathbf{x}\sim q(x)}[\max(0, 1 - D(\mathbf{x}, C(\mathbf{x})))] \\
&+ \mathbb{E}_{\mathbf{x}\sim q(x), \mathbf{z}\sim p(\mathbf{z})}[\max(0, 1 + D(G(\mathbf{z}|C(\mathbf{x})), C(\mathbf{x})))] \qquad (1) \\
L_G =& - \mathbb{E}_{\mathbf{x}\sim q(x), \mathbf{z}\sim p(\mathbf{z})}[D(G(\mathbf{z}|C(\mathbf{x})), C(\mathbf{x}))]
\end{aligned}
$$

$$
D(\mathbf{x}, \mathbf{y}) = D_{emb}(\mathbf{y}) \cdot D_f(\mathbf{x}) + D_l \circ D_f(\mathbf{x}) \text{is the } \textbf{c-GAN projection loss}^2 \qquad (2)
$$

- Since $C(\mathbf{x})$ is extracted from a pre-trained network, above training objective leads to feature level knowledge distillation from C.

[2] Takeru Miyato and Masanori Koyama. cgans with projection discriminator.

Approach - Loss Function

$$
\begin{aligned}
L_D =& \mathbb{E}_{\mathbf{x}\sim q(x)}[\max(0, 1 - D(\mathbf{x}, C(\mathbf{x})))] \\
&+ \mathbb{E}_{\mathbf{x}\sim q(x), \mathbf{z}\sim p(\mathbf{z})}[\max(0, 1 + D(G(\mathbf{z}|C(\mathbf{x})), C(\mathbf{x})))] \qquad (1) \\
L_G =& - \mathbb{E}_{\mathbf{x}\sim q(x), \mathbf{z}\sim p(\mathbf{z})}[D(G(\mathbf{z}|C(\mathbf{x})), C(\mathbf{x}))]
\end{aligned}
$$

$$
D(\mathbf{x}, \mathbf{y}) = D_{emb}(\mathbf{y}) \cdot D_f(\mathbf{x}) + D_l \circ D_f(\mathbf{x}) \text{is the } \textbf{c-GAN projection loss}^2 \qquad (2)
$$

- Since $C(\mathbf{x})$ is extracted from a pre-trained network, above training objective leads to feature level knowledge distillation from C.
- It also acts as a regularizer on the discriminator to reduce overfitting.

[2] Takeru Miyato and Masanori Koyama. cgans with projection discriminator.

Approach - Loss Function

$$
\begin{aligned}
L_D =& \mathbb{E}_{\mathbf{x}\sim q(x)}[\max(0, 1 - D(\mathbf{x}, C(\mathbf{x})))] \\
&+ \mathbb{E}_{\mathbf{x}\sim q(x), \mathbf{z}\sim p(\mathbf{z})}[\max(0, 1 + D(G(\mathbf{z}|C(\mathbf{x})), C(\mathbf{x})))] \qquad (1) \\
L_G =& - \mathbb{E}_{\mathbf{x}\sim q(x), \mathbf{z}\sim p(\mathbf{z})}[D(G(\mathbf{z}|C(\mathbf{x})), C(\mathbf{x}))]
\end{aligned}
$$

$$
D(\mathbf{x}, \mathbf{y}) = D_{emb}(\mathbf{y}) \cdot D_f(\mathbf{x}) + D_l \circ D_f(\mathbf{x}) \text{is the } \textbf{c-GAN projection loss}^2 \qquad (2)
$$

- Since $C(\mathbf{x})$ is extracted from a pre-trained network, above training objective leads to feature level knowledge distillation from C.
- It also acts as a regularizer on the discriminator to reduce overfitting.
- Enforcing feature $D_f(G(\mathbf{z}|C(\mathbf{x})))$ to be similar to $D_{emb}(C(\mathbf{x}))$ promotes mode coverage of target data distribution.

[2] Takeru Miyato and Masanori Koyama. cgans with projection discriminator.

Approach - Inference

Let $\mathbb{D}_{prior} = \{C(\mathbf{x}_j)\}_{j=1}^{n}$. The generator requires access to $\mathbb{D}_{prior}$ for sample generation.

Approach - Inference

Let $\mathbb{D}_{prior} = \{C(\mathbf{x}_j)\}_{j=1}^{n}$. The generator requires access to $\mathbb{D}_{prior}$ for sample generation.

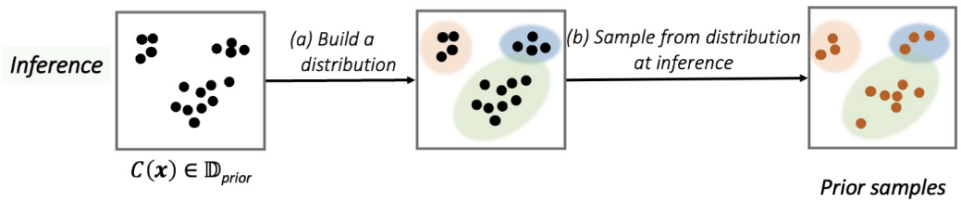
- **few-shot** and **limited** - size of $\mathbb{D}_{prior}$ is less, to create more variations, we generate images conditioned on prior samples from a mixup distribution.

Approach - Inference

Let $\mathbb{D}_{prior} = \{C(\mathbf{x}_j)\}_{j=1}^{n}$. The generator requires access to $\mathbb{D}_{prior}$ for sample generation.

- **few-shot** and **limited** - size of $\mathbb{D}_{prior}$ is less, to create more variations, we generate images conditioned on prior samples from a mixup distribution.
- **large-scale** - we learn a GMM on $\mathbb{D}_{prior}$. This enables memory efficient sampling of conditional prior.

$$G(\mathbf{z}|\mathcal{N}(\mu, \mathbf{\Sigma})) \text{ where } \mu, \mathbf{\Sigma} \sim \text{GMM}(G_{emb}(\mathbb{D}_{prior})) \tag{3}$$



Few-shot Image Generation

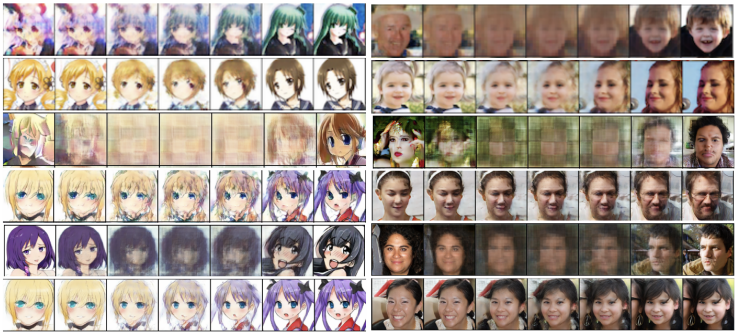
Few-shot image generation performance using 100 training images.

Method	Pre-training	SNGAN (128 x 128) Anime FID ↓	Anime P ↑	Anime R ↑	Faces FID ↓	Faces P ↑	Faces R ↑
From scratch	✗	120.38	0.61	0.00	140.66	0.31	0.00
+ DISP-Vgg16		**66.85**	**0.71**	**0.03**	**68.49**	**0.74**	**0.15**
TransferGAN [3]	✓	102.75	**0.70**	0.00	101.15	**0.85**	0.00
+ DISP-Vgg16		**86.96**	0.57	**0.02**	**75.21**	0.70	**0.10**
FreezeD [4]	✓	109.40	**0.67**	0.00	107.83	**0.83**	0.00
+ DISP-Vgg16		93.36	0.56	**0.03**	77.09	0.68	0.14
+ DISP-SimCLR		**89.39**	0.46	0.025	**70.40**	0.74	**0.22**
ADA [5]	✗	78.28	0.87	0.0	159.3	0.69	0.0
+ DISP-Vgg16		**60.8**	**0.90**	**0.003**	**79.5**	**0.85**	**0.004**
DiffAugment	✗	85.16	**0.95**	0.00	109.25	**0.84**	0.00
+ DISP-Vgg16		**48.67**	0.82	0.03	**62.44**	0.80	0.19
+ DISP-SimCLR		52.41	0.77	**0.04**	64.53	0.78	**0.22**

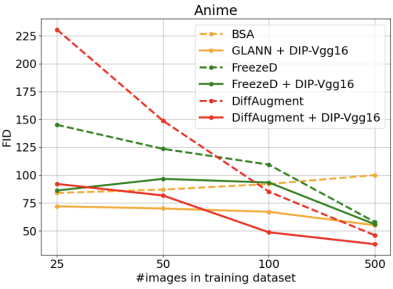
[3] Wang et al. Transferring gans: generating images from limited data

[4] Mo et al. Freeze the dis-criminator: a simple baseline for fine-tuning gans.

Few-shot Image Generation



(a) (b)



(c)

Figure: **(a) and (b)**: Sample interpolations between two generated images for models trained in few-shot setting : Scratch *(Row 1)*, Scratch + DISP-Vgg16 *(Row 2)*, FreezeD *(Row 3)*, FreezeD + DISP-Vgg16 *(Row 4)*, DiffAugment *(Row 5)*, DiffAugment + DISP-Vgg16 *(Row 6)*. **(c)**: FID (lower is better) performance graph of few-shot image generation by varying the training samples from 25 to 500 images of Anime dataset for different approaches on SNGAN model.

Limited Image Generation

Comparison of FID on Unconditional CIFAR-10 and CIFAR-100 image generation while varying the amount of training data.

Method	CIFAR-10			CIFAR-100		
	100% data	20% data	10% data	100% data	20% data	10% data
BigGAN	17.22	31.25	42.59	20.37	33.25	42.43
+ DISP	9.70	16.24	27.86	12.89	21.70	31.48
+ DiffAugment [6]	10.39	15.12	18.56	13.33	19.78	23.80
+ DiffAugment & DISP	**9.52**	**14.24**	**18.50**	**12.70**	**16.91**	**20.47**
StyleGAN2* [7]	11.07	23.08	36.02	16.54	32.30	45.87
+ DiffAugment*	9.89	12.15	14.5	15.22	16.65	20.75
+ DiffAugment & DISP	**9.50**	**10.92**	**12.03**	**14.45**	**15.52**	**17.33**

[6] Zhao et al.Differentiable augmentation for data-efficient gan training

[7] Karras et al. Analyzing and improving the image quality of stylegan.

Limited Image Generation

(a) Places (2.5k) (b) FFHQ (2K) (c) CUB (6K)

Figure: Samples of generated image in limited data training setting : FreezeD *(Row 1)*, FreezeD + DISP-Vgg16 *(Row 2)*, DiffAugment *(Row 3)* and DiffAugment + DISP-Vgg16 *(Row 4)*.

Large Scale Image Generation

Comparison of DISP with Baseline [8], SSGAN [9] and Self-Cond GAN [10] in large-scale image generation setting on FID, Precision and Recall metrics.

Method	**CIFAR-10**			**CIFAR-100**			**FFHQ**			**LSUN-Bedroom**			**ImageNet32x32**		
	FID ↓	P ↑	R ↑	FID ↓	P ↑	R ↑	FID ↓	P ↑	R ↑	FID ↓	P ↑	R ↑	FID ↓	P ↑	R ↑
Baseline	19.73	0.64	**0.70**	24.66	0.61	**0.67**	21.67	0.77	0.47	9.89	0.58	0.42	16.19	0.60	**0.67**
SSGAN	15.65	0.67	0.68	21.02	0.61	0.65	-	-	-	7.68	0.59	0.50	17.18	0.61	0.65
Self-Cond GAN	16.72	0.71	0.64	21.8	0.64	0.60	-	-	-	-	-	-	15.56	**0.66**	0.63
DISP-Vgg16	**11.24**	**0.74**	0.64	**15.71**	**0.70**	0.62	**15.83**	0.76	**0.55**	4.99	**0.66**	**0.54**	**12.11**	0.64	0.62
DISP-SimCLR	14.42	0.68	0.65	20.08	0.67	0.62	16.62	**0.77**	0.53	**4.92**	0.62	0.53	14.99	0.60	0.63

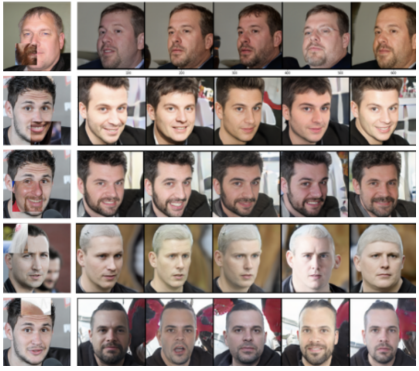
[8] Brock et al. Largescale gan training for high fidelity natural image synthesis.

[9] Chen et al. Self-supervised gans via auxiliary rotationloss.

[10] Liu et al. Diverse image generation via self-conditioned gans.

Semantic Diffusion

Exploit $C(\mathbf{x})$, to get some control over the high-level semantics (e.g. hair, gender, glasses, etc in case of faces) of generated image.



(a) Custom Editing - First column shows human-edited version where certain portion of image is substituted with another to achieve desired semantics. Rest columns correspond to images generated when Vgg16 features of human-edited version is provided as prior to DISP module.

(b) Sketch-to-Image - First column shows sketch describing desired high-level semantics. Rest columns correspond to images generated when Vgg16 features of the sketch version is provided as prior in DISP module.

The End