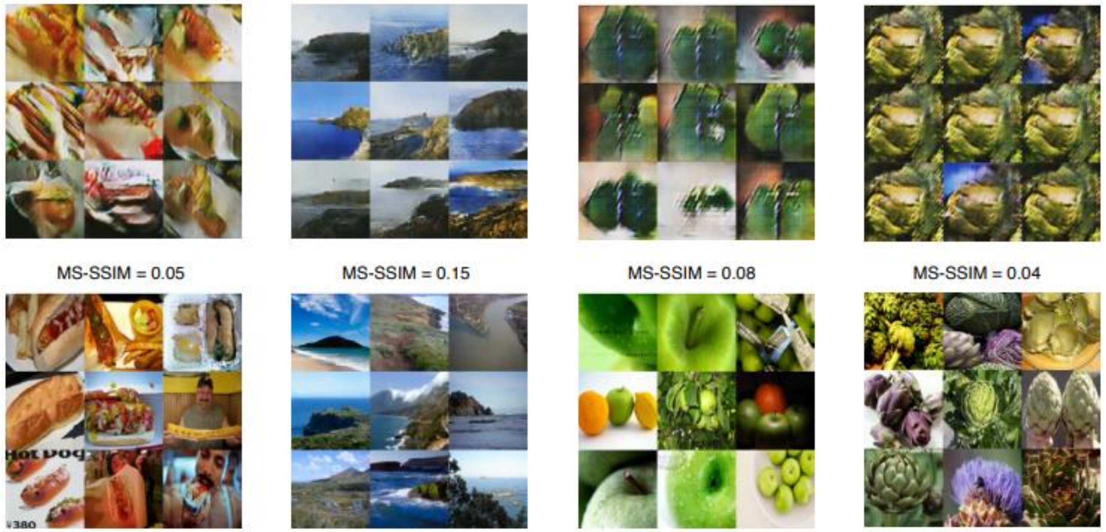
Label Noise Robust Generative Adversarial Networks

2018.12.3

Introduction

Success in Class Conditional GANs



It requires large amount accurate class-labeled data

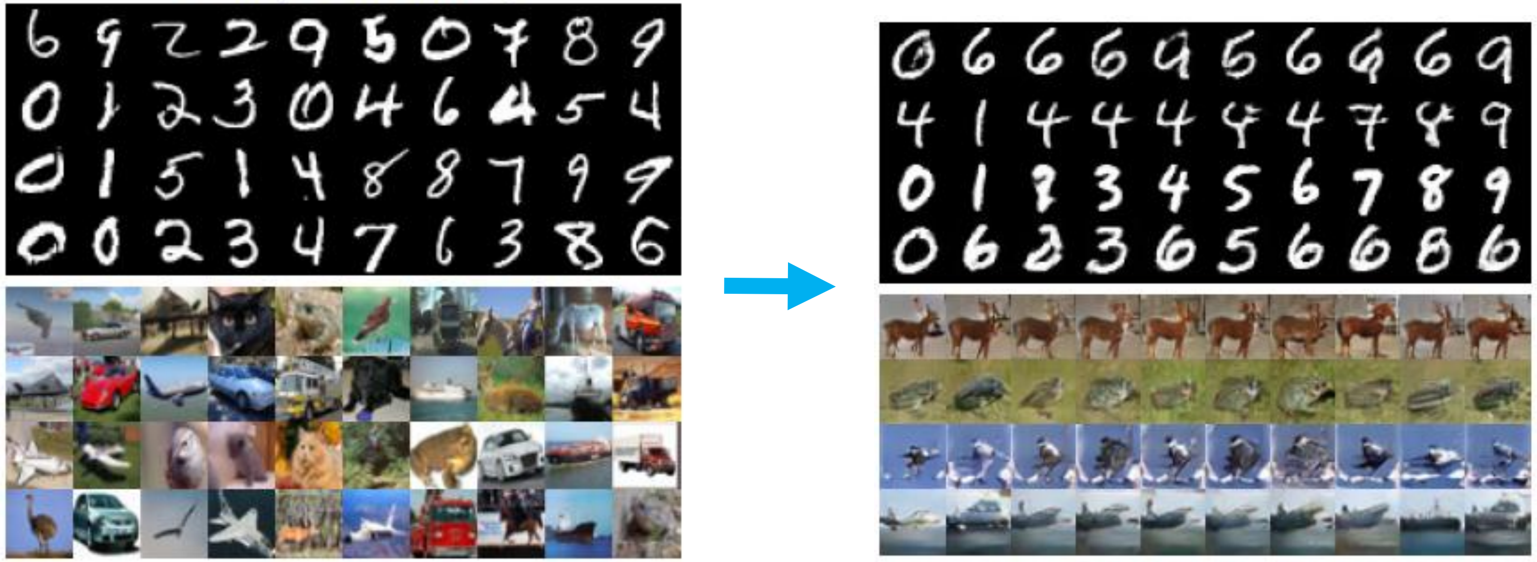
By incorporating class labels, it promote(to)

- models to selectively generate images condition on the class label.
- stabilize GAN training, which is typically unstable, and improve image quality.

Motivation

When train cGAN on noise labeled data…

Real (observable)



(b) Noisy labeled data
Flipped with probability 0.5

(c) cGAN trained with (b)
[Baseline]

The performance of models depends on labels accuracy

Challenge: Learn a clean label conditional distribution even when training labels are noisy.

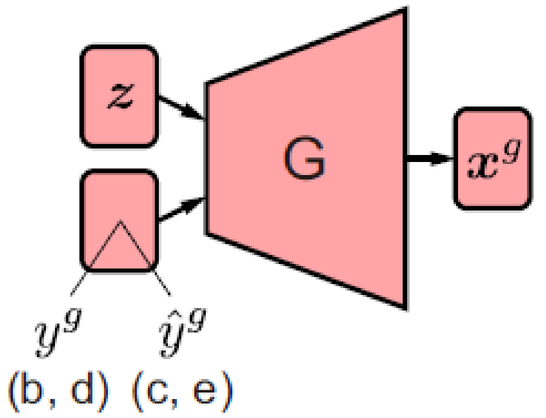
Background: cGAN

Conditional adversarial loss

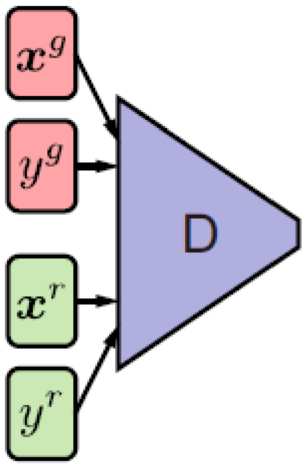
$$\mathcal{L}_{\text{cGAN}} = \mathbb{E}_{(\boldsymbol{x}^r, y^r) \sim p^r(\boldsymbol{x}, y)}[\log D(\boldsymbol{x}^r, y^r)] + \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z}), y^g \sim p(y)}[\log(1 - D(G(\boldsymbol{z}, y^g), y^g))]$$

Maximize the loss to optimize D

Minimize the loss to optimize G



(a) Conditional generator



(d) cGAN discriminator

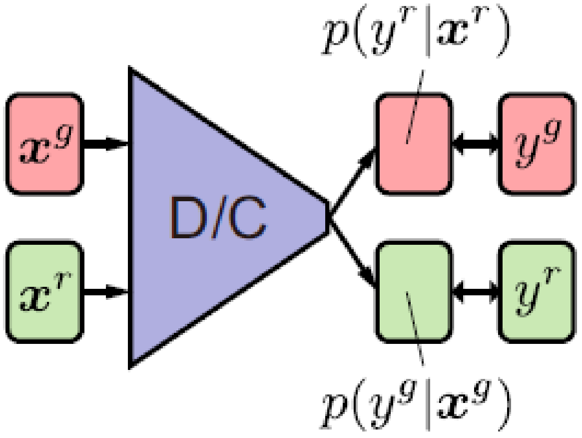
Background: ACGAN

Adversarial loss

$$\mathcal{L}_{\mathrm{GAN}} = \mathbb{E}_{\boldsymbol{x}^r \sim p^r(\boldsymbol{x})}[\log D(\boldsymbol{x}^r)] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}), y^g \sim p(y)}[\log(1 - D(G(\boldsymbol{z}, y^g)))]$$

Auxiliary classifier loss

$$\mathcal{L}^g_{\mathrm{AC}} = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}), y^g \sim p(y)}[-\log C(y = y^g | G(\boldsymbol{z}, y^g))]$$

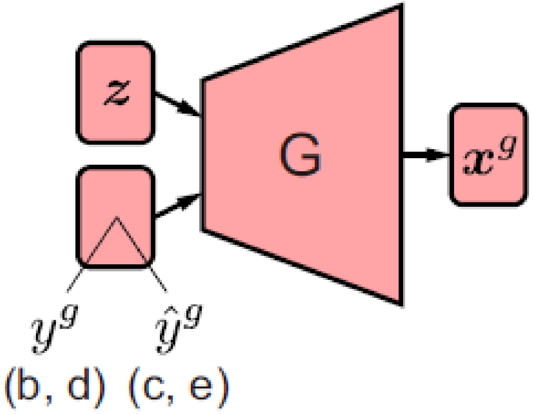


(b) AC-GAN discriminator

Full objective

$$\mathcal{L}_{D/C} = -\mathcal{L}_{\mathrm{GAN}} + \lambda^r_{\mathrm{AC}}\mathcal{L}^r_{\mathrm{AC}},$$
$$\mathcal{L}_G = \mathcal{L}_{\mathrm{GAN}} + \lambda^g_{\mathrm{AC}}\mathcal{L}^g_{\mathrm{AC}},$$

D/C and G are optimized by minimizing $\mathcal{L}_{D/C}$ and $\mathcal{L}_G$



(a) Conditional generator

Problem and Idea

Problem statement

Assume class dependent noise in which each clean label $\hat{y} = i$ is corrupted to a noisy label $\tilde{y} = j$ with probability $p(\tilde{y} = j|\hat{y} = i)$, independently of x.

Noise transition model

Noise transition matrix: $T_{i,j} = p(\tilde{y} = j|\hat{y} = i)$

where $T = (T_{i,j}) \in [0,1]^{c \times c}$ $(\sum_i T_{i,j} = 1)$

Incorporate the noise transition model into ACGAN and cGAN.

Model: rAC-GAN

The auxiliary loss is reformulated as:

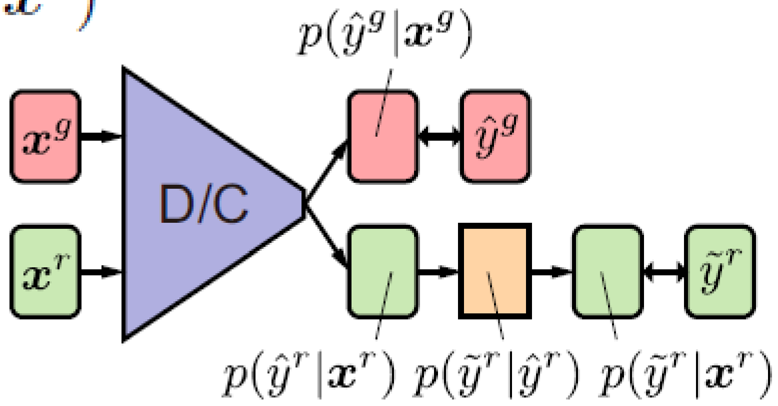
noise label classifier

$$\mathcal{L}^{r}_{\mathrm{rAC}} = \mathbb{E}_{(\boldsymbol{x}^r,\tilde{y}^r)\sim p^r(\boldsymbol{x},\tilde{y})}[-\log \tilde{C}(\tilde{y} = \tilde{y}^r|\boldsymbol{x}^r)] \rightarrow \hat{C}(\hat{y} = \hat{y}^r|\boldsymbol{x}^r)$$

$$= \mathbb{E}_{(\boldsymbol{x}^r,\tilde{y}^r)\sim p^r(\boldsymbol{x},\tilde{y})}[-\log \sum_{i=1}^{c} p(\tilde{y} = \tilde{y}^r|\hat{y} = i)\hat{C}(\hat{y} = i|\boldsymbol{x}^r)]$$

clean label classifier

$$= \mathbb{E}_{(\boldsymbol{x}^r,\tilde{y}^r)\sim p^r(\boldsymbol{x},\tilde{y})}[-\log \sum_{i=1}^{c} T_{i,\tilde{y}^r}\hat{C}(\hat{y} = i|\boldsymbol{x}^r)], \quad (6)$$



(c) rAC-GAN discriminator

Note: this formulation(called the forward correction) is often used in label-noise robust classification model.

Theoretical Analysis

Theorem 1. *When T is nonsingular,*

$$\underset{\hat{C}}{\operatorname{argmin}}\, \mathbb{E}_{(\boldsymbol{x}^r,\tilde{y}^r)\sim p^r(\boldsymbol{x},\tilde{y})}[-\log \sum_{i=1}^{c} T_{i,\tilde{y}^r}\hat{C}(\hat{y}=i|\boldsymbol{x}^r)]$$
$$= \underset{\hat{C}}{\operatorname{argmin}}\, \mathbb{E}_{(\boldsymbol{x}^r,\hat{y}^r)\sim p^r(\boldsymbol{x},\hat{y})}[-\log \hat{C}(\hat{y}=\hat{y}^r|\boldsymbol{x}^r)]. \qquad (8)$$

clean label classifier

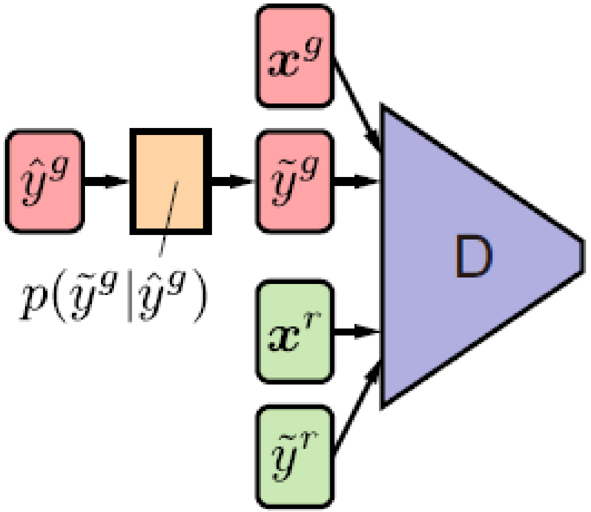
Minimizing $\mathcal{L}_{rAC}^{r}$ → Clean label classifier $\hat{C}$

Hence, G is optimized using $\hat{C}$ rather than using $\tilde{C}$

$$\mathcal{L}_{\mathrm{rAC}}^{g} = \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}),\hat{y}^g\sim p(\hat{y})}[-\log \hat{C}(\hat{y}=\hat{y}^g|G(\boldsymbol{z},\hat{y}^g))]$$

Model: cGAN

$$\mathcal{L}_{\text{cGAN}} = \mathbb{E}_{(\boldsymbol{x}^r, y^r)\sim p^r(\boldsymbol{x}, y)}[\log D(\boldsymbol{x}^r, y^r)] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}), y^g\sim p(y)}[\log(1 - D(G(\boldsymbol{z}, y^g), y^g))]$$



(e) rcGAN discriminator

When y^r is noisy($\tilde{y}^r$ is given), cGAN learns the noisy label conditional generator $G(z, \tilde{y}^g)$ → $G(z, \hat{y}^g)$ clean

Sample $\tilde{y}^g$ from $\tilde{y}^g \sim p(\tilde{y}|\hat{y}^r)$ and refine the above equation as:

$$\mathcal{L}_{\text{rcGAN}} = \mathbb{E}_{(\boldsymbol{x}^r, \tilde{y}^r)\sim p^r(\boldsymbol{x}, \tilde{y})}[\log D(\boldsymbol{x}^r, \tilde{y}^r)] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}), \hat{y}^g\sim p^g(\hat{y}), \tilde{y}^g\sim p(\tilde{y}|\hat{y}^g)}[\log(1 - D(\underline{G(\boldsymbol{z}, \hat{y}^g)}, \tilde{y}^g))],$$

clean

Theoretical Analysis

In an optimal condition, the following theorem hold.

Theorem 2. *When there is a unique probability distribution $p(\hat{y})$ that induces the distribution $p(\tilde{y})$, G is optimal if and only if $p^g(\boldsymbol{x}, \hat{y}) = p^r(\boldsymbol{x}, \hat{y})$.*

↓

This supports the idea that, in an optimal condition, rcGAN learns $G(z, \hat{y})$ such that $p^g(x, \hat{y}) = p^r(x, \hat{y})$.

Theoretical Analysis

The optimal discriminator D for fixed G is
$$D(\boldsymbol{x}, \tilde{y}) = \frac{p^r(\boldsymbol{x}, \tilde{y})}{p^r(\boldsymbol{x}, \tilde{y}) + p^g(\boldsymbol{x}, \tilde{y})}$$

Then G is optimal if and only if

$$p^g(\boldsymbol{x}, \tilde{y}) = p^r(\boldsymbol{x}, \tilde{y})$$

Label corruption occurs with $p(\tilde{y}|\hat{y})$, independently of x

$$p(\boldsymbol{x}, \tilde{y}) = p(\tilde{y}|\boldsymbol{x})p(\boldsymbol{x}) = \sum_{\hat{y}} p(\tilde{y}|\hat{y})p(\hat{y}|\boldsymbol{x})p(\boldsymbol{x})$$
$$= \sum_{\hat{y}} p(\tilde{y}|\hat{y})p(\boldsymbol{x}, \hat{y}).$$

$$\sum_{\hat{y}} p(\tilde{y}|\hat{y})p^g(\boldsymbol{x}, \hat{y}) = \sum_{\hat{y}} p(\tilde{y}|\hat{y})p^r(\boldsymbol{x}, \hat{y})$$

When there is a unique probability distribution $p(\hat{y})$ that induces the distribution $p(\tilde{y})$, $p^g(x, \hat{y}) = p^r(x, \hat{y})$

Experiment

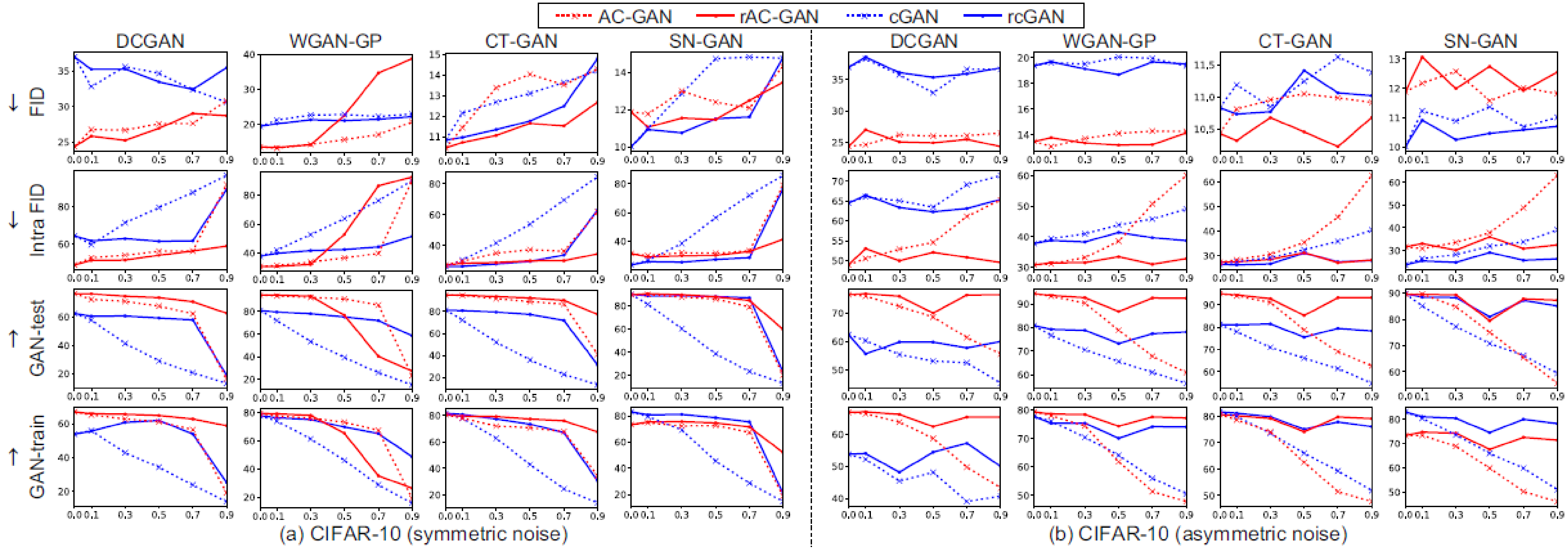
Dataset: CIFAR10 and CIFAR100

Asymmetric(class-dependent) noise: Ground truth labels are flipped with probability μ by mimicking real mistakes between similar classes.

GAN configurations: DCGAN, WGAN-GP, CT-GAN and SNGAN

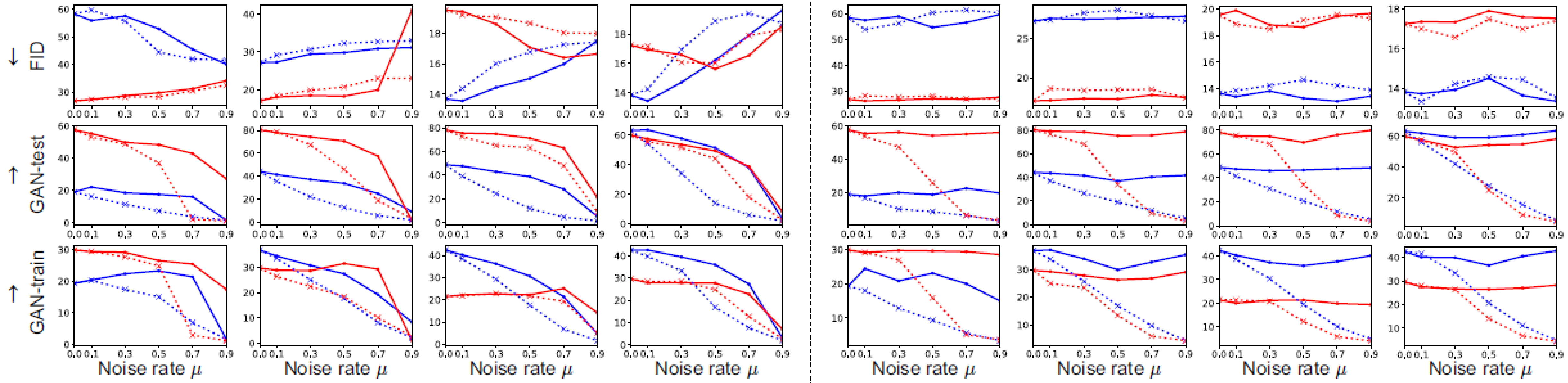
Evaluation metrics: FID, gan-Test and gan-Train

Experiment



AC-GAN
rAC-GAN
cGAN
rcGAN
DCGAN
WGAN-GP
CT-GAN
SN-GAN
↓ FID
↓ Intra FID
↑ GAN-test
↑ GAN-train
(a) CIFAR-10 (symmetric noise)
(b) CIFAR-10 (asymmetric noise)

Experiment

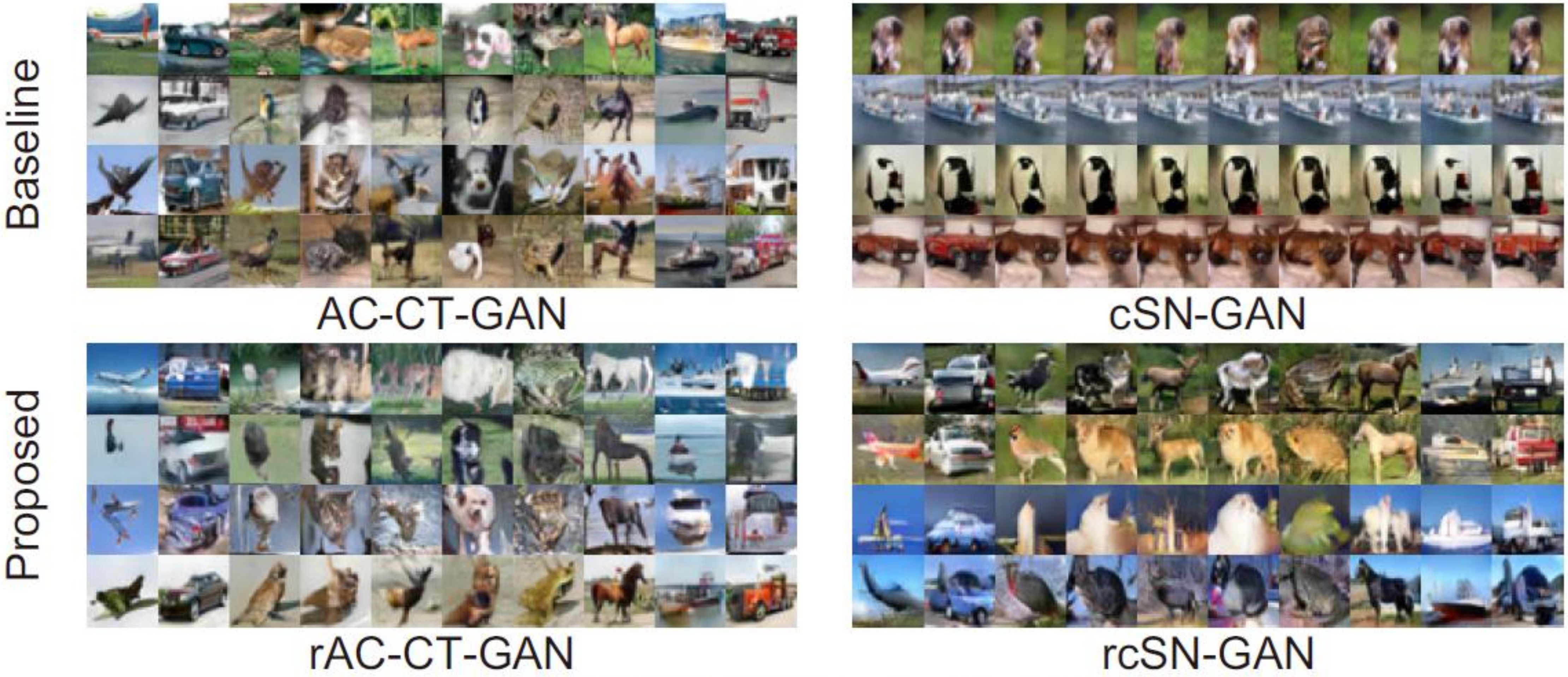


(c) CIFAR-100 (symmetric noise)

(d) CIFAR-100 (asymmetric noise)

Model	Metric	CIFAR-10 (symmetric noise)					CIFAR-10 (asymmetric noise)					CIFAR-100 (symmetric noise)					CIFAR-100 (asymmetric noise)				
		0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
rAC-CT-GAN with T'	FID ↓	10.9	11.4	11.3	11.5	13.0	10.8	10.2	10.2	10.4	11.0	19.7	19.3	17.7	17.3	18.5	19.4	19.3	19.7	18.8	19.0
	Intra FID ↓	28.7	**31.0**	**30.1**	**31.7**	**38.9**	28.5	**27.4**	**31.2**	**35.0**	**36.8**	–	–	–	–	–	–	–	–	–	–
	GAN-test ↑	95.3	93.2	**92.0**	87.7	**70.4**	94.9	92.9	**85.2**	**78.5**	**76.6**	**76.6**	67.1	**68.1**	*1.0*	*2.5*	74.1	68.9	*28.7*	7.2	2.2
	GAN-train ↑	78.7	**75.9**	**76.9**	**73.7**	**63.4**	79.8	**79.5**	**74.0**	**69.1**	**67.3**	21.2	21.4	23.3	*1.0*	2.3	19.1	19.9	10.7	5.5	3.9
rcSN-GAN with T'	FID ↓	10.7	11.9	12.4	12.1	15.0	10.8	10.8	11.0	10.9	11.3	14.3	16.6	17.5	20.0	19.8	13.8	14.1	14.7	14.7	13.9
	Intra FID ↓	25.5	**29.4**	**29.4**	**29.7**	87.4	25.7	26.0	**28.7**	32.6	**33.9**	–	–	–	–	–	–	–	–	–	–
	GAN-test ↑	**85.3**	**79.0**	**84.8**	**82.8**	15.9	86.6	**87.2**	**84.0**	74.9	**71.2**	53.4	36.6	**37.7**	*1.0*	1.7	**65.0**	**63.0**	**32.4**	*7.8*	3.8
	GAN-train ↑	80.7	**78.1**	**77.4**	**75.6**	15.0	80.5	**79.0**	**75.7**	**69.3**	**65.7**	40.1	32.8	**31.3**	*1.0*	1.8	41.7	**39.3**	20.1	*6.1*	3.9

Experiment



(a) CIFAR-10 (symmetric noise)

Thanks