Adversarial Knowledge Distillation for a Compact Generator

Hideki Tsunashima[1,2], Hirokatsu Kataoka[2], Junji Yamato[3], Qiu Chen[3], Shigeo Morishima[4]

1) Waseda University, 2) AIST, 3) Kogakuin University, 4) Waseda Research Institute for Science and Engineering

Contents

- Background
- Adversarial Knowledge Distillation for Generative Models
- Experiments
  - Confirming the effectiveness of AKDG on a toy dataset
  - Compressing GANs on CIFAR-10
  - Compressing GANs on LSUN bedroom
  - Why does not AKDG work in case of the small batchsize?
- Conclusion and Future Work

GANs have the problem of large computation resources

Generative Adversarial Nets achieve the impressive progress in generative models. However, state-of-the-art GANs require large amount computation resources.

For instance, training StyleGAN [1] needs 7 days on 8 NVIDIA V100.
Moreover, the size of the pre-trained weights of StyleGAN is 310MB.

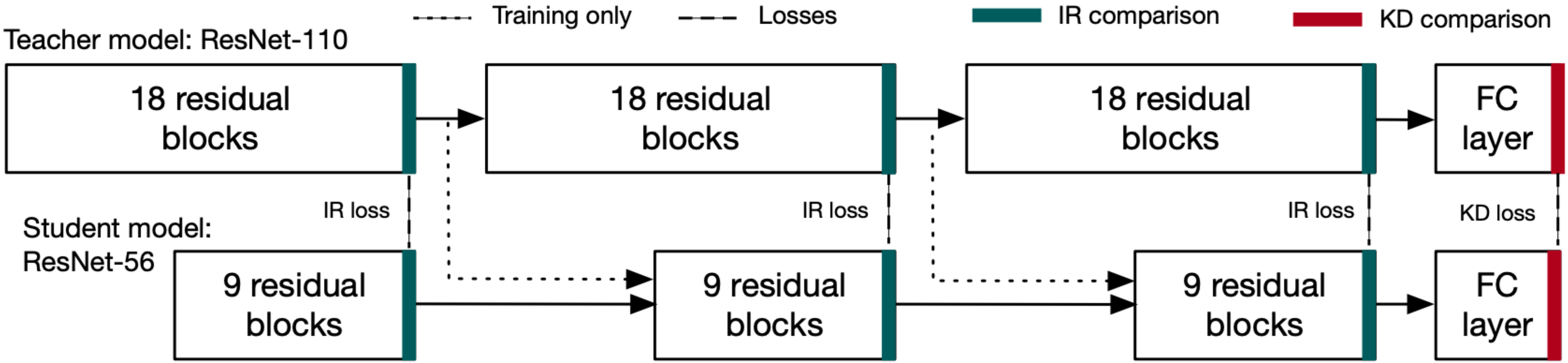
Impressive results but need large amount computation resources

The methods of compressing GANs

- The compression methods for unconditional GANs.

➢ LIT (Animesh et al., ICML2019) [2]

LIT reduces chunks of the residual blocks so that it reduces parameters of GANs. However, LIT can be only applied to a generator which has the ResNet architecture.



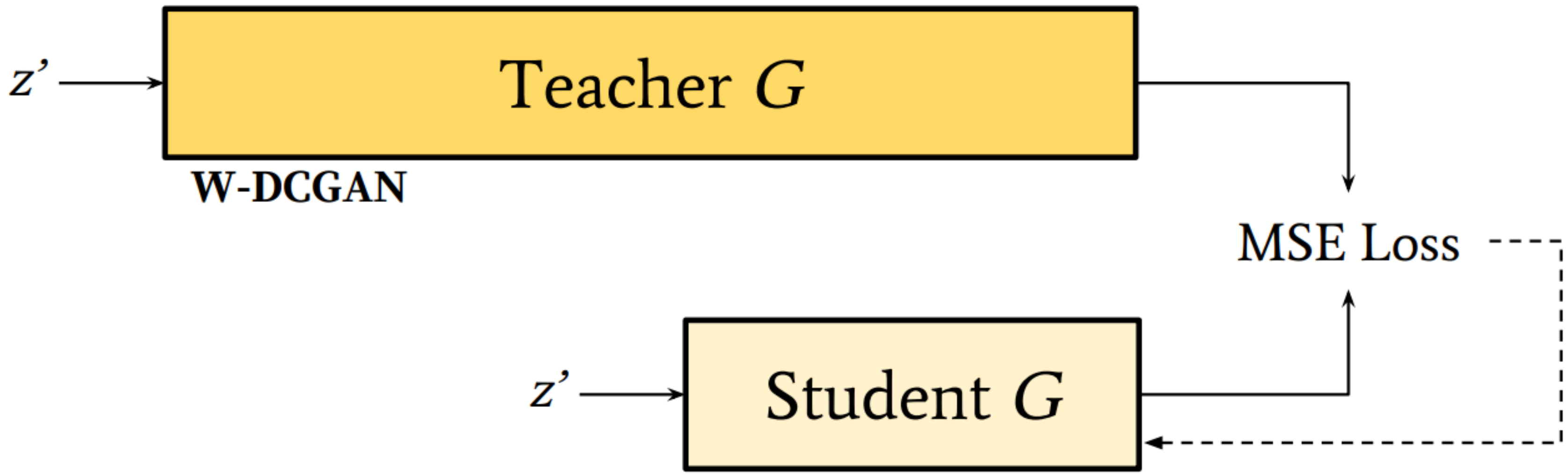
The methods of compressing GANs

- The compression methods for unconditional GANs

➢ MSE-method (Aguinaldo et al., arXiv preprint 2019) [3]

MSE-method takes the Mean Squared Error (MSE) between the generated images of a teacher and student.

However, MSE-method aggravates quality of the generated images from vanilla GANs



The methods of compressing GANs

- The compression methods for conditional GANs

➢ GAN Compressions (Li et al., CVPR2020) [4]

➢ Distilling portable Generative Adversarial Networks for Image Translation (Chen et al., AAAI2020) [5]

➢ AGD (Fu et al., ICML2020) [6]

➢ Learning Efficient GANs via Differentiable Masks and co-Attention Distillation (Li et al., arXiv preprint 2020) [7]

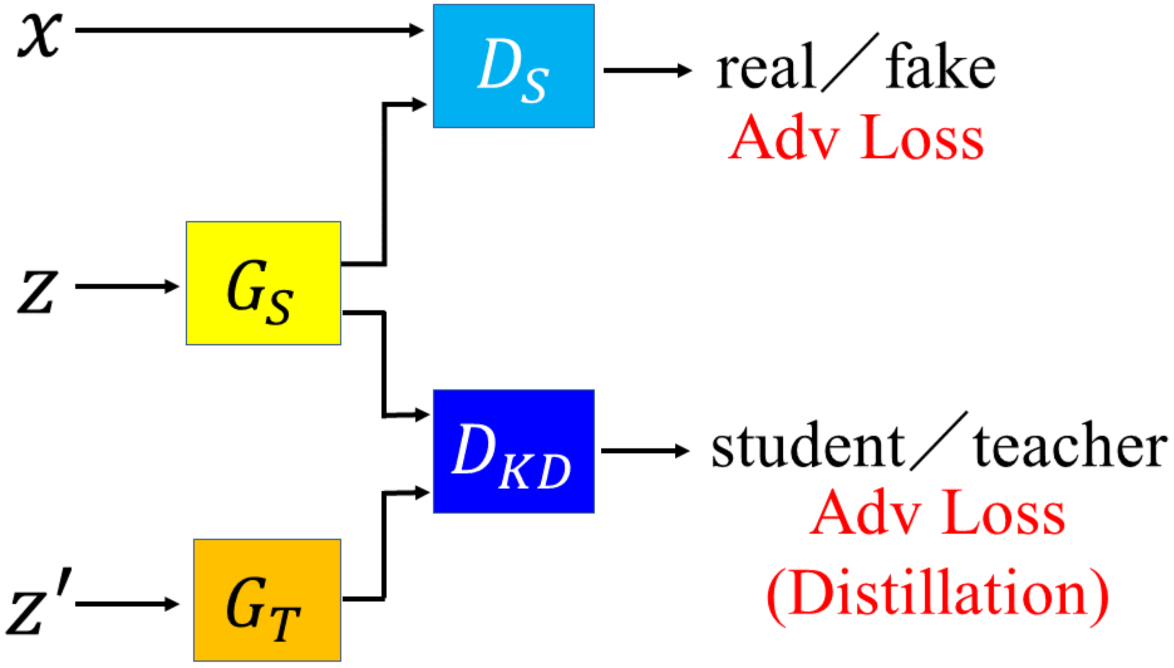
These methods are specialized to conditional GANs.
Therefore, these are out of scope in our paper.

Contents

- Adversarial Knowledge Distillation for Generative Models

Adversarial Knowledge Distillation for Generative models

We propose Adversarial Knowledge Distillation for Generative models (AKDG)
The key of AKDG is the additional discriminator which judges a teacher or student.



AKDG (ours)

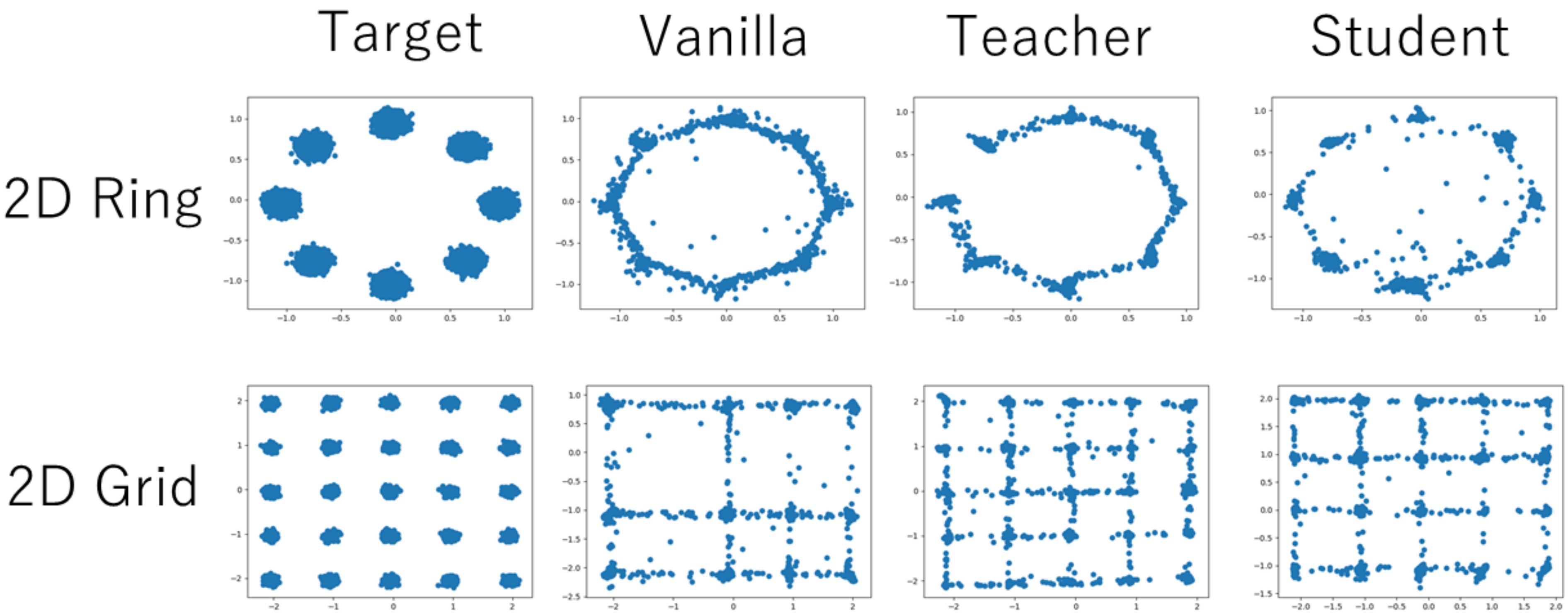
Contents

- Experiments
  - Confirming the effectiveness of AKDG on a toy dataset
  - Compressing GANs on CIFAR-10
  - Compressing GANs on LSUN bedroom
  - Why does not AKDG work in case of the small batchsize?
- Conclusion and Future Work

Confirming the effectiveness of AKDG on a toy dataset

The connection between modes is sparse in the student trained with AKDG on 2D Ring.
The student does not miss the modes in comparison with the vanilla model on 2D Grid.

We confirm AKDG appropriately work.



Compressing GANs on CIFAR-10

IS: Inception Score, FID: Fréchet Inception Distance

Quality and diversity of the images generated by the student trained with AKDG is the best.

Moreover, AKDG is robust in terms of the difference of architecture.

Methods	# layers	# params	# mult-add	IS ↑	FID ↓
Teacher (SNGAN [4])	8	4.1M	793M	8.49(±0.13)	16.50
DCGAN vanilla [40]	5	2.6M(37.6% ↓)	229M(71.1% ↓)	6.86(±0.07)	31.58
DCGAN-MSE [21]	5	2.6M(37.6% ↓)	229M(71.1% ↓)	6.45(±0.07)	40.06
DCGAN-AKDG (ours)	5	2.6M(37.6% ↓)	229M(71.1% ↓)	**7.05**(±0.07)	**29.52**
MLP vanilla [40]	4	2.2M(46.3% ↓)	2.2M(99.7% ↓)	4.66(±0.02)	61.30
MLP-MSE [21]	4	2.2M(46.3% ↓)	2.2M(99.7% ↓)	3.97(±0.05)	108.92
MLP-AKDG (ours)	4	2.2M(46.3% ↓)	2.2M(99.7% ↓)	**5.14**(±0.05)	**53.73**

Compressing GANs on CIFAR-10

IS: Inception Score, FID: Fréchet Inception Distance

Quality and diversity of the images generated by the student trained with AKDG is the best.

Moreover, AKDG is robust in terms of the difference of architecture.

Teacher (Miyato et al.)

Vanilla (Mescheder et al.)

AKDG (ours)

Compressing GANs on LSUN bedroom

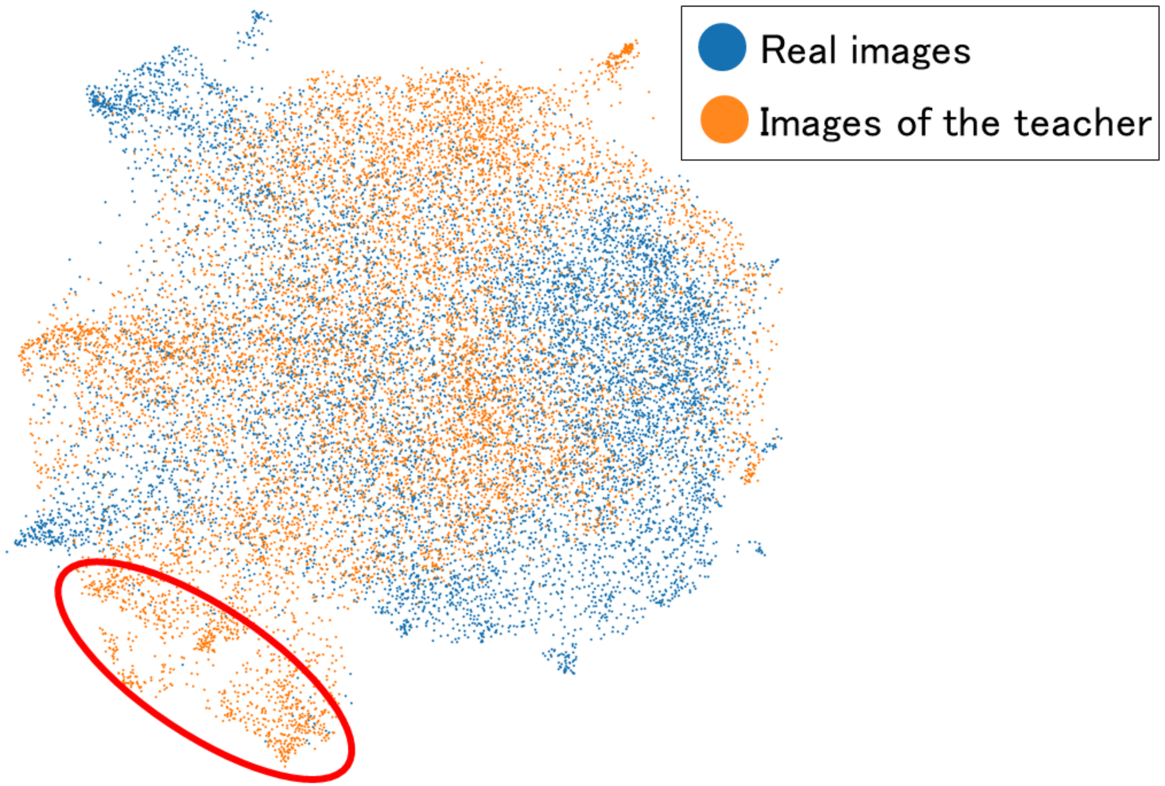
Quality of the images generated by the student with AKDG is the best.

Methods	# layers	# params	# mult-add	FID ↓
Teacher (PGGAN [2])	15	18.3M	8.9B	21.22
DCGAN vanilla (batch64) [40]	8	3.7M(79.8% ↓)	785M(91.2% ↓)	44.72
DCGAN vanilla (batch2048) [40]	8	3.7M(79.8% ↓)	785M(91.2% ↓)	37.89
DCGAN-MSE [21]	8	3.7M(79.8% ↓)	785M(91.2% ↓)	98.96
DCGAN-AKDG (ours)	8	3.7M(79.8% ↓)	785M(91.2% ↓)	**27.86**

Teacher (Karras et al.) Vanilla (Mescheder et al.) **AKDG (ours)**

Why does not AKDG work in case of the small batchsize?

We visualize the real images and the images of the teacher using UMAP.
Red circle indicates images of the teacher which does not overlap the real images.
This distribution gap incurs the collapse of training.



Conclusion and Future Work

- Conclusion

We propose the novel compression method for unconditional generative models. In all experiments, our AKDG outperforms the conventional compression method.

- Future Work

1. We will fix the problem of large batchsize training.
2. We will incorporate other network compression methods, such as pruning, quantization and tensor factorization.

Reference

[1] Karras et al., “A Style-Based Generator Architecture for Generative Adversarial Networks”, CVPR2019.
[2] Koratana et al., “LIT: Learned Intermediate Representation Training for Model Compression”, ICML2019.
[3] Aguinaldo et al., “Compressing GANs using Knowledge Distillation”, arXiv preprint, 2019.
[4] Li et al., “GAN Compression: Efficient Architectures for Interactive Conditional GANs”, CVPR2020.
[5] Chen et al., “Distilling portable Generative Adversarial Networks for Image Translation”, AAAI2020.
[6] Fu et al., “AutoGAN-Distiller: Searching to Compress Generative Adversarial Networks”, ICML2020.
[7] Li et al., “Learning Efficient GANs via Differentiable Masks and co-Attention Distillation”, arXiv preprint, 2020.