Generative Modeling

More on GANs

Denis Derkach, Artem Ryzhikov, Sergei Popov

Laboratory for methods of big data analysis

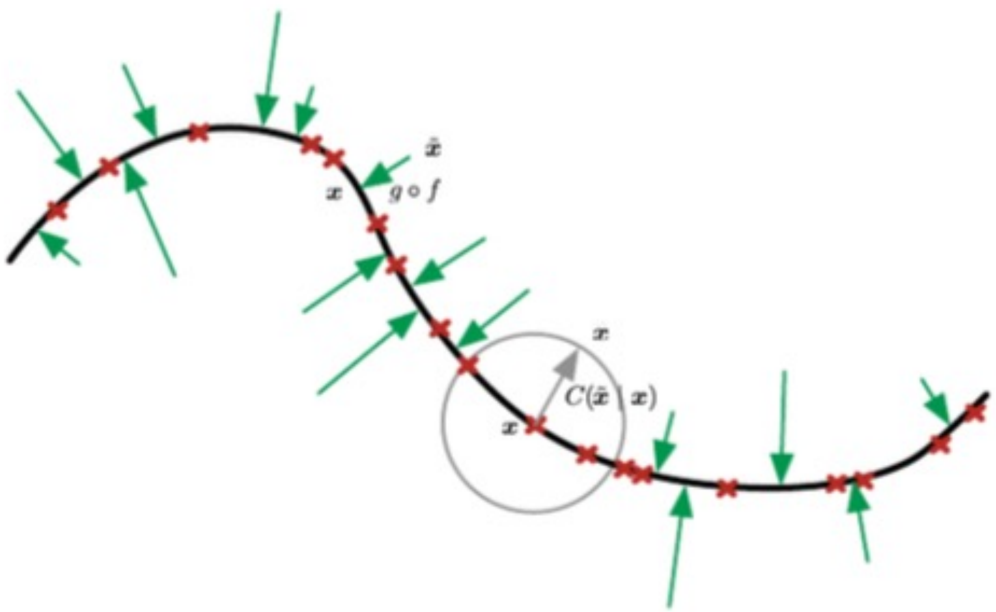
LAMBDA • HSE

Spring 2023

In this Lecture

- Special Discriminator Structures
  - Energy-based Generative Adversarial Network
  - Boundary Equilibrium Generative Adversarial Networks
  - Discriminator Rejection Sampling
- Additional tips for the infrastructure.

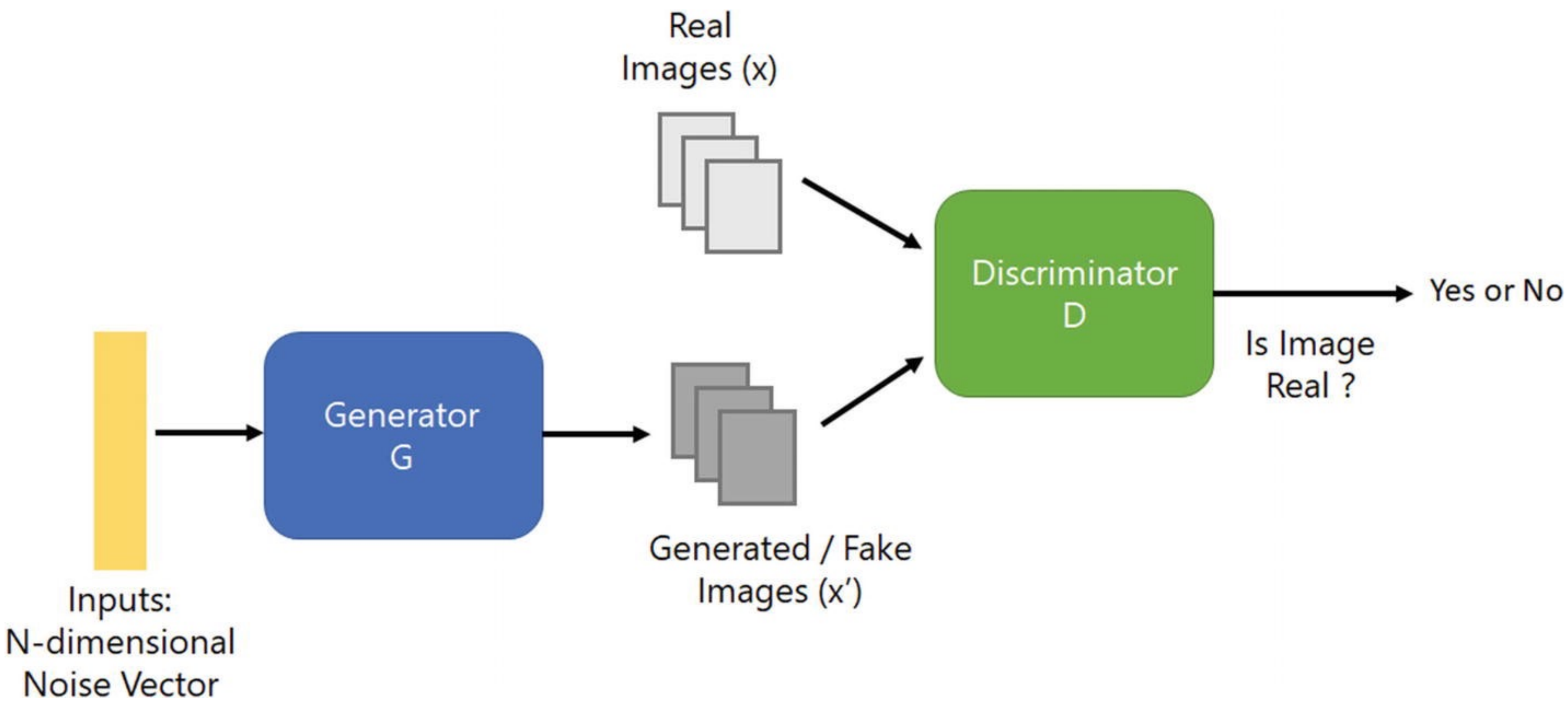
Reminder Contractive Denoising Autoencoders



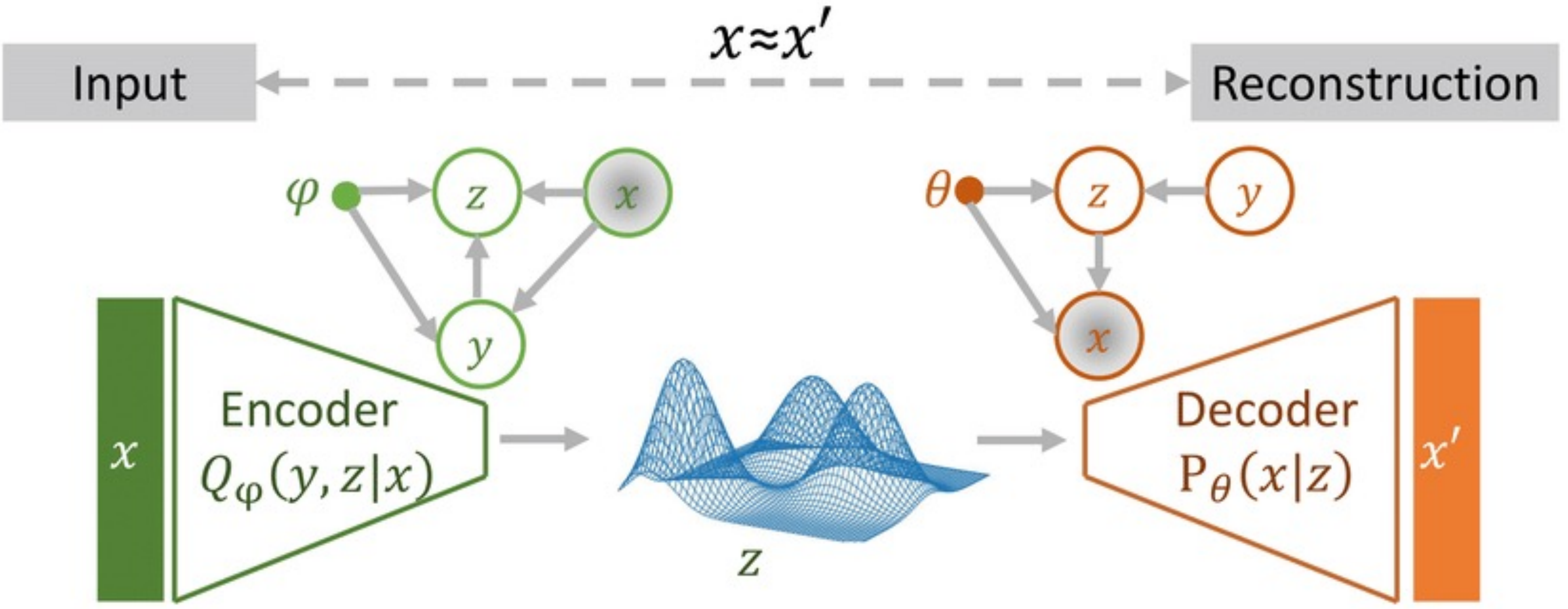
- The true signal is always situated on a manifold inside the R^D space.
- Denoising autoencoder is trained to map a corrupted data point $\tilde{x}$ back to the original data point x.

https://arxiv.org/abs/1305.6663

Vanilla GAN


Real
Images (x)
Discriminator
D
Yes or No
Is Image
Real ?
Generator
G
Generated / Fake
Images (x')
Inputs:
N-dimensional
Noise Vector

VAE

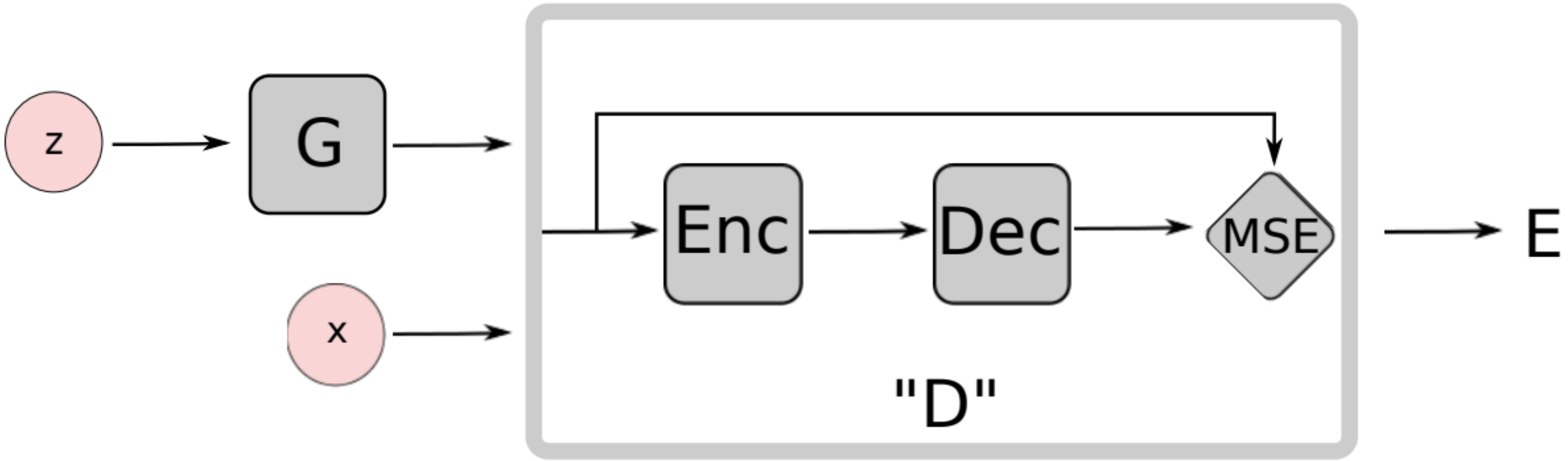


$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{Q_\phi(y,z|x)}\left[\log P_\theta(x,y,z) - \log Q_\phi(y,z|x)\right]$$

$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{Q_\phi(y,z|x)}\Bigg[\underbrace{\boxed{\log\frac{P(y)}{Q_\phi(y|x)}}}_{\text{Entropy}} + \underbrace{\boxed{\log\frac{P_\theta(z|y)}{Q_\phi(z|x,y)}}}_{\text{Regularization}} + \underbrace{\boxed{\log P_\theta(x|z)}}_{\text{Reconstruction}}\Bigg]$$

Energy-based GAN

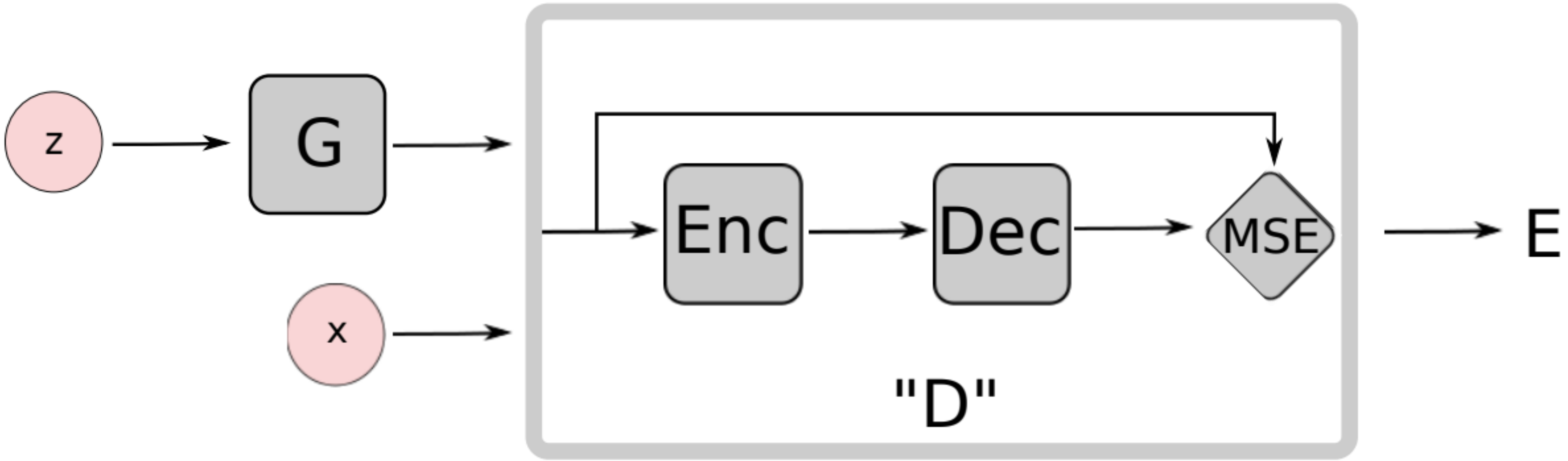
Autoencoder Discriminator



- Pros:
  - Takes into account the minibatch information properly.
  - Uses the full information about the manifold.

Junbo Zhao et al., ICLR 2017

Autoencoder Discriminator



- Use AE to extract latent features of the input image by an encoder and reconstruct it again with the decoder with MSE loss:

$$D(x) = ||Dec\big(Enc(x)\big) - x||$$

Junbo Zhao et al., ICLR 2017

EB-GAN training

For $[.]^+ = \max(0, .)$:

$$\mathcal{L}_D(x, z) = D(x) + [m - D(G(z))]^+;$$
$$\mathcal{L}_G(x, z) = D(G(z)),$$

- **autoencoder**: reconstruction cost $D(x)$ for real images is low;
- $D(x)$ is trained first several rounds;
- once $G(z)$ generates sufficiently good images $D(x)$ training resumes;
- repelling loss to address AE collapse problem:

$$f_{PT} = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \frac{S_i^T S_j}{||S_i|| ||S_j||}.$$

where $S \in \mathbb{R}^{s \times N}$ a batch of sample of size N representations taken from the encoder output layer.

GAN energy interpretation

For $[.]^+ = \max(0, .)$:

$$\mathcal{L}_D(x, z) = D(x) + [m - D(G(z))]^+;$$
$$\mathcal{L}_G(x, z) = D(G(z)),$$

- **autoencoder**: reconstruction cost $D(x)$ for real images is low;
- $D(x)$ does not have probability interpretation;
- one can use **energy interpretation** instead.

EB-GAN results

Figure 4: Generation from the grid search on MNIST. Left(a): Best GAN model; Middle(b): Best EBGAN model. Right(c): Best EBGAN-PT model.

Boundary equilibrium GAN

Wasserstein Distance lower bound

- **Wasserstein distance:**

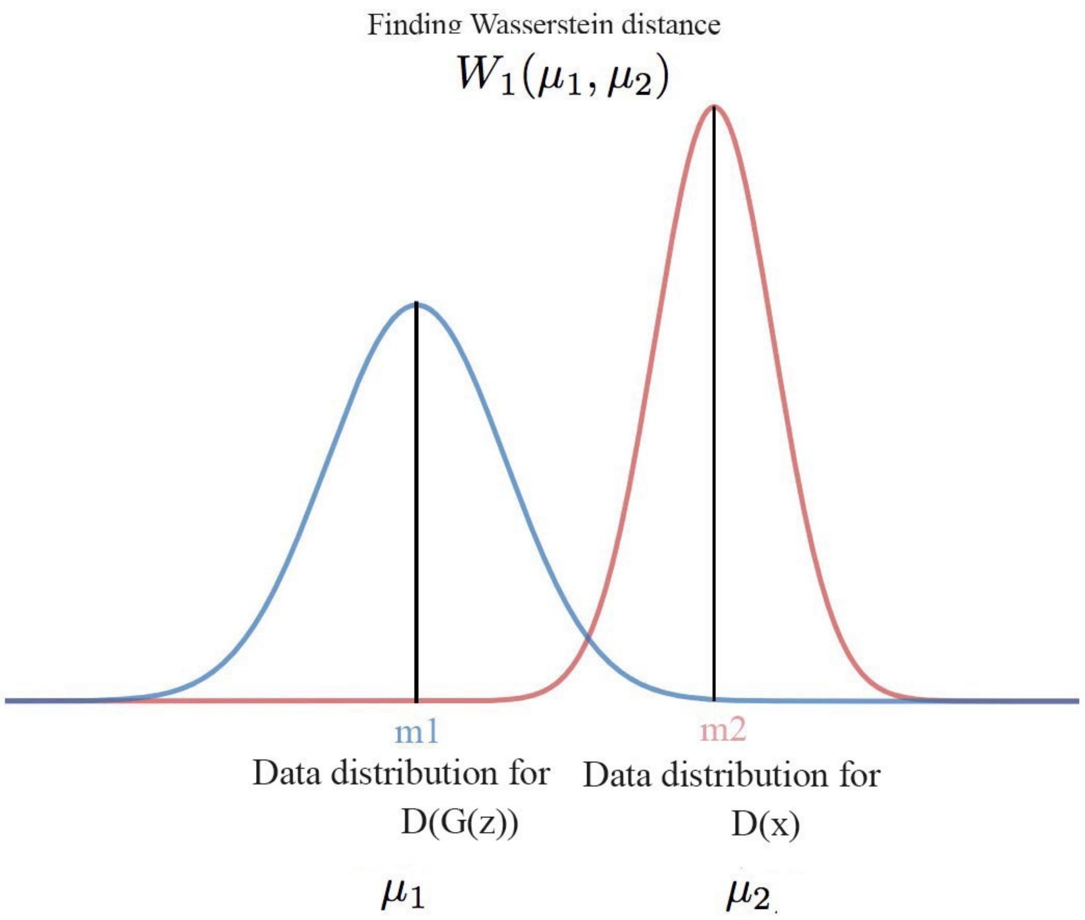
$$W(\mu 1, \mu 2) = \inf_{\gamma \in \Pi} \mathbb{E}_{(x,y)\sim\gamma} ||x - y||$$

- **Jensen's inequality**:

$$W(\mu 1, \mu 2) \geq \inf_{\gamma \in \Pi} |\mathbb{E}_{(x,y)\sim\gamma} |x - y|| =$$

$$= |m_1 - m_2|,$$

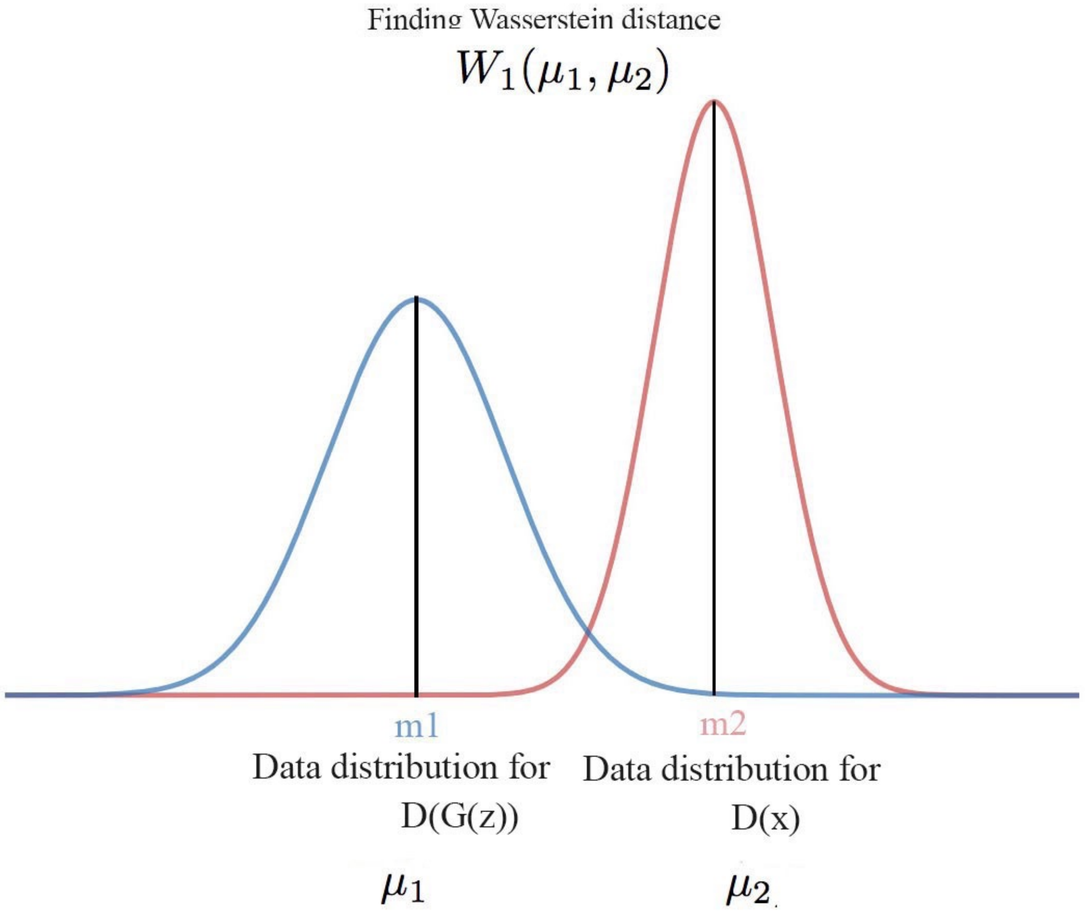
where m_i are the mean on μ_i.



D. Berthelot et al., 1703.10717

Wasserstein Discriminator

- We have $D(x)$ as AE:
  $D(x) = ||Dec\big(Enc(x)\big) - x||$
  $\mathcal{L}_D = W(\mu_1 \mu_2) \geq |m_1 - m_2|.$
- We can use $D(x)$ in minibatch instead of mean:
  $$\mathcal{L}_D = D(x) - D\big(G(z)\big).$$
- We thus optimize W between losses.
- No need for K-Lipshitz, since no Kantorovich-Rubinstein duality is used.



Equilibrium term

- we need to maintain balance between G and D:

$$\mathbb{E}(D(x)) = \mathbb{E}(D(G(z)))$$

- we thus can use a parameter to balance the impact:

$$\gamma = \frac{\mathbb{E}(D(x))}{\mathbb{E}(D(G(z))}.$$

- γ can be chosen to sharpen the image.

BEGAN formulation

- We thus can write full optimization for BEGAN

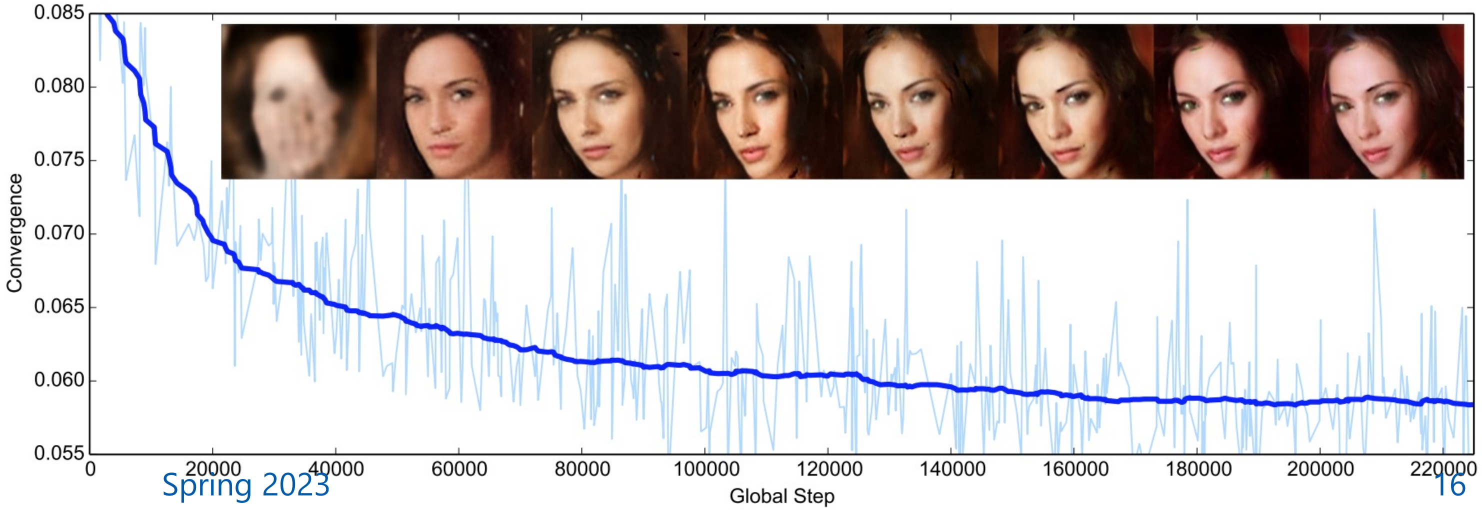
$$\mathcal{L}_D = D(x) - k_t D(G(z));$$

$$\mathcal{L}_G = D(G(z));$$

$$k_{t+1} = k_t + \lambda_k(\gamma D(x) - D(G(z))).$$

- Dropping γ leads to mode collapse.
- To monitor the convergence:

$$M_{global} = D(x) + (\gamma D(x) - D(G(z)))$$



BEGAN results

(c) Our results (128x128 with 128 filters)

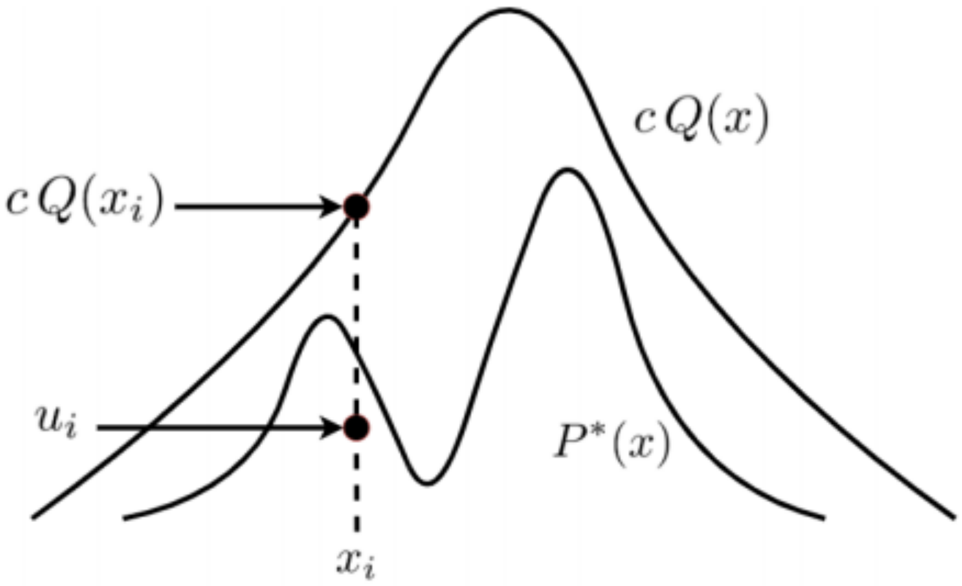
(d) Mirror interpolations (our results 128x128 with 128 filters)

Wrap-up

- We can change the architecture of discriminator.
- This might lead to change of the optimization idea.
- If we use autoencoder as discriminator we have access to the energy instead of probability.
- We can optimize Wasserstein distance not only for datasets, but also for results of function.

Rejection Sampling

Rejection Sampling



Input: $P^*(X), Q(X), c$
Output: $\mathcal{S} = \{x_i\}_{i=1}^n \sim P^*(X)$
$\mathcal{S} \leftarrow \varnothing$
for *sample index* i *from* 1 *to* n **do**
  $x_i \sim Q(X)$
  $u_i \sim U(0, c\,Q(x_i))$
  if $u_i < P^*(x_i)$ **then**
    Accept x_i: $\mathcal{S} \leftarrow \mathcal{S} \cup \{x_i\}$
  else
    Reject x_i: $i \leftarrow i - 1$

B. Ghojogh et al., 2011.00901

Ideal Discriminator

- Ideal discriminator:

$$D^*(x) = \frac{p(x)}{p(x) + q_\theta(x)}.$$

- Remember f-GAN idea of last layer:

$$D^*(x) = \frac{1}{1 + e^{-\widetilde{D}^*(x)}} = \frac{p(x)}{p(x) + q_\theta(x)}.$$

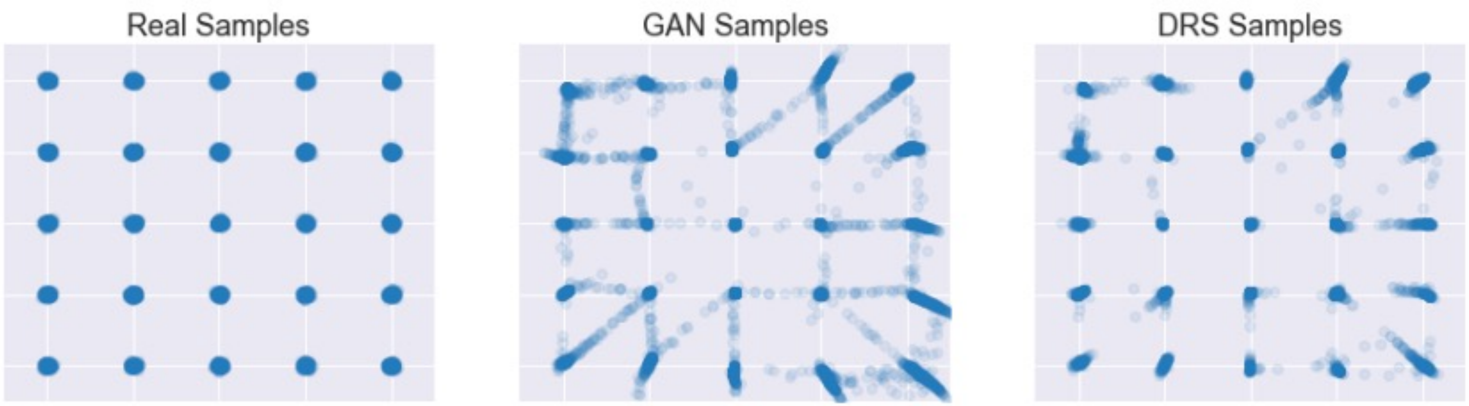
- Thus:

$$\frac{p(x)}{q_\theta(x)} = e^{\widetilde{D}^*(x)}.$$

- This defines constant for rejection sampling.

S. Azadi et al., NeurIPS 2019

Discriminator Rejection Sampling



ImageNet	IS	FID
Without DRS	52.34 ± 0.45	18.21 ± 0.14
With DRS	61.44 ± 0.09	17.14 ± 0.09

Results suggest that the quality of sampling is improved

Your GAN is secretly an energy based model

- Previous results can be revisited (with acceptance probability):

$$\frac{p(x)}{q_\theta(x)} = e^{\widetilde{D}^*(x)}.$$

- And applied to the latent space. This will create a rule for new latent space distribution:

$$p_t(z) = p_0(z) r(z) / C.$$

- Which can be rewritten as:

$$p_t(z) \; = \; e^{-E(z)} \, / Z_0, \text{ with tractable } E(z):$$

$$E(z) = \; -\log p_0(z) - \widetilde{D} \; \left(G(z)\right).$$

- This can be used to define MCMC in latent space and later obtain $x \sim G(z)$.

T. Che et al., NeurIPS 2020

Energy-based sampling: results

Figure 4. Top-5 nearest neighbor images (right columns) of generated samples (left column).

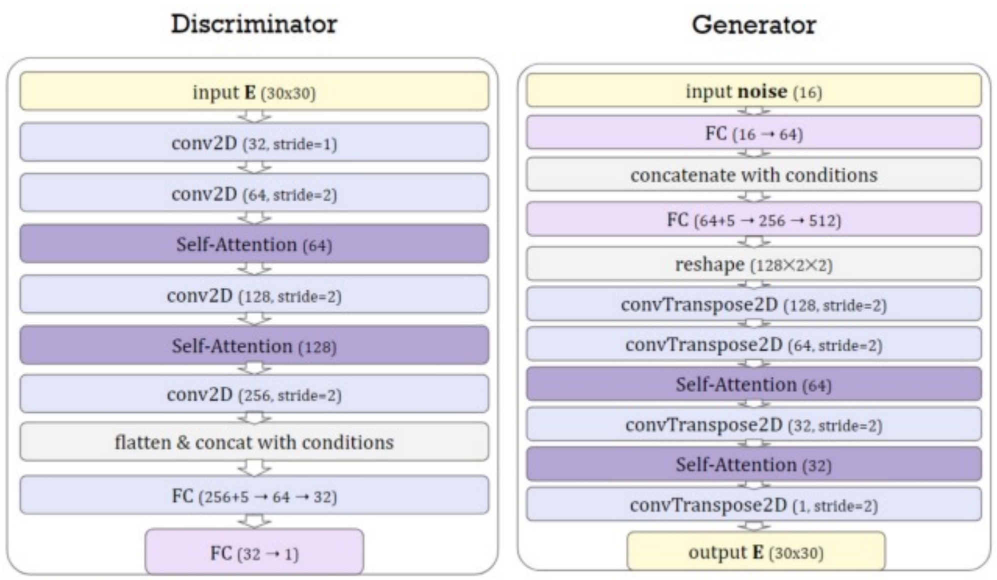
Discussion

- GAN's discriminator can enable better modeling of the data distribution with Discriminator Driven Latent Sampling.
- The major advantage of DDLS is that it allows MCMC sampling in the latent space.

Implementing GANs

Motivation

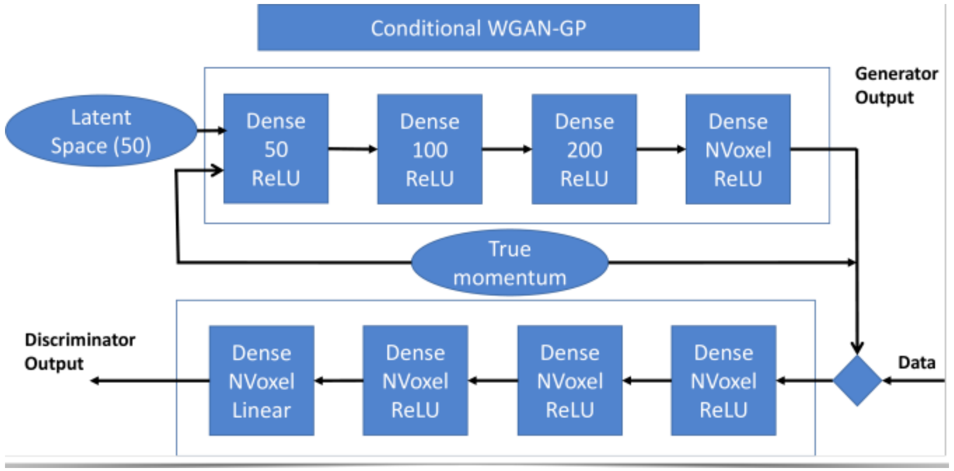
- GANs, being the most popular generative models, often aim for production.
- The sampling speed thus becomes important.
- Easiest way: implement model description.



EPJ Web of Conferences 251, 03043 (2021)

vs

100X



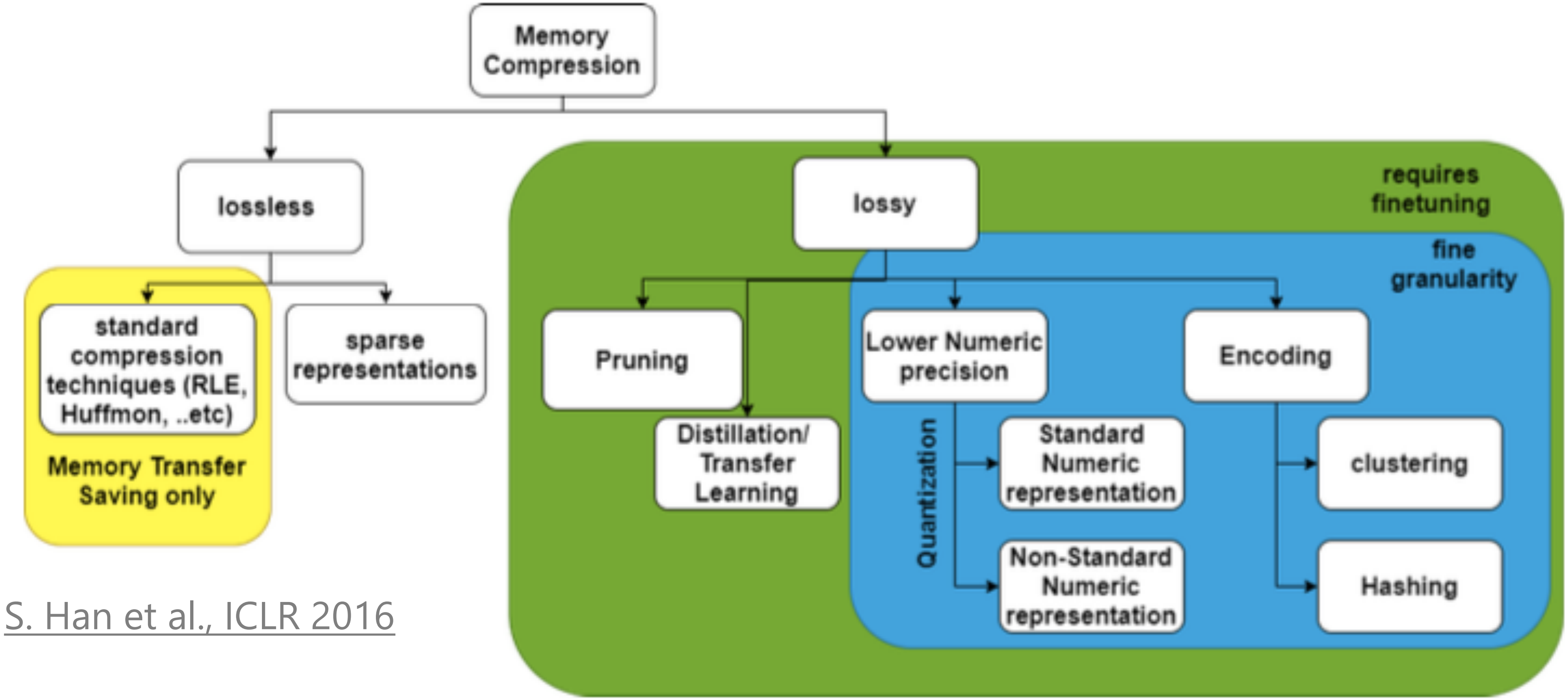
Rogachev et al., ACAT 2021
Ahmed et al., CERN seminar’2022

Acceleration techniques

- Memory compression – minimize memory requirements.
- Computation optimization – decrease number of mathematical operations.
- Dataflow optimization – maximize data reuse and minimize ineffectual operations.

D. Tantawy et al., A survey on GAN acceleration using memory compression techniques, Journal of Engineering and Applied Sciences, 2021

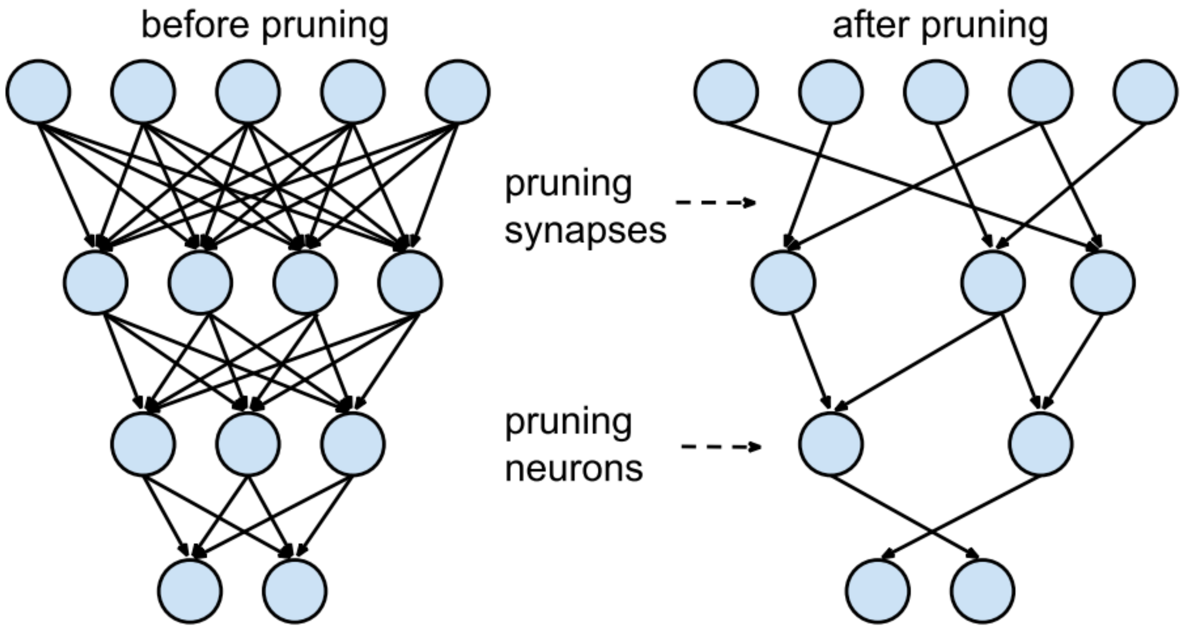
Memory Compression



S. Han et al., ICLR 2016

Pruning Decisions

- Criteria
  - random;
  - threshold based;
  - evolutionary.
- Granularity
  - structured;
  - unstructured.
- Application
  - Before/after training.
- D vs. G



Pruning for GANs

- Out-of-the-box techniques fail:
  - weak evaluation metrics;
  - unstable training;
  - high-dimensional input/output spaces.

C. Yu et al., Self-Supervised GAN compression, NeurIPS'20

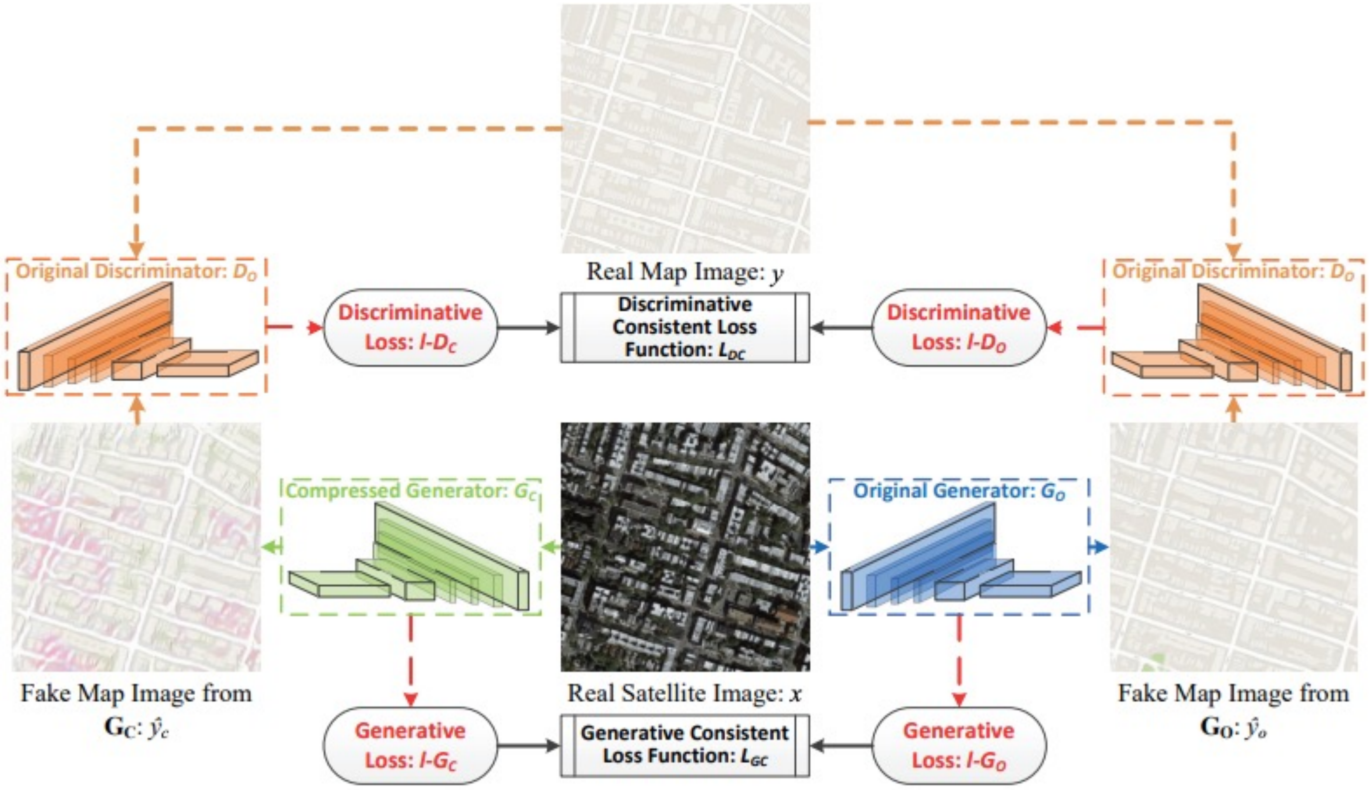
Self-supervised Pruning

- Use discriminator from the uncompressed GAN to train compressed one.

$$
\begin{aligned}
L_{GC}(l\text{-}\boldsymbol{G}_O, l\text{-}\boldsymbol{G}_C) = |l\text{-}\boldsymbol{Gen}_O - l\text{-}\boldsymbol{Gen}_C|/|l\text{-}\boldsymbol{Gen}_O| + \\
\alpha|l\text{-}\boldsymbol{Cla}_O - l\text{-}\boldsymbol{Cla}_C|/|l\text{-}\boldsymbol{Cla}_O| + \\
\beta|l\text{-}\boldsymbol{Rec}_O - l\text{-}\boldsymbol{Rec}_C|/|l\text{-}\boldsymbol{Rec}_O|
\end{aligned}
$$

$$
\begin{aligned}
L_{DC}(l\text{-}\boldsymbol{D}_O, l\text{-}\boldsymbol{D}_C) = |l\text{-}\boldsymbol{Dis}_O - l\text{-}\boldsymbol{Dis}_C|/|l\text{-}\boldsymbol{Dis}_O| + \\
\delta|l\text{-}\boldsymbol{GP}_O - l\text{-}\boldsymbol{GP}_C|/|l\text{-}\boldsymbol{GP}_O|
\end{aligned}
$$

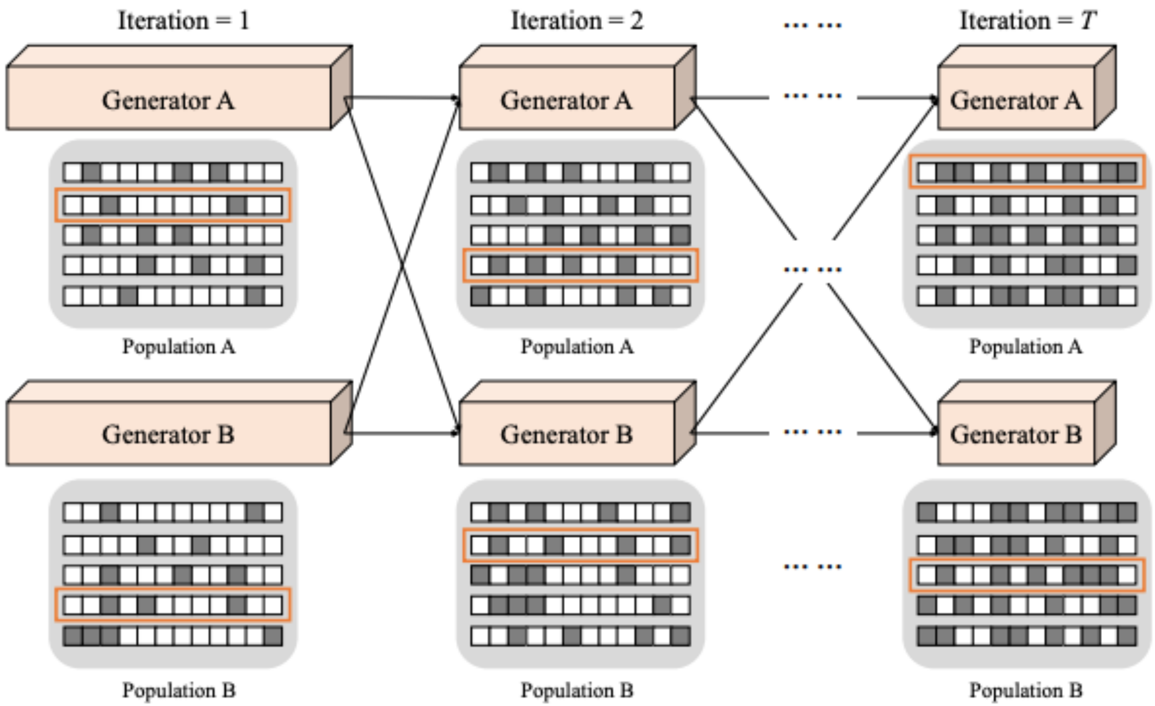
$$
L_{Overall} = \boldsymbol{L}_{GC}(l\text{-}\boldsymbol{G}_O, l\text{-}\boldsymbol{G}_C) + \lambda \boldsymbol{L}_{DC}(l\text{-}\boldsymbol{D}_O, l\text{-}\boldsymbol{D}_C)
$$



C. Yu et al., Self-Supervised GAN compression, NeurIPS'20

Co-evolutionary Pruning

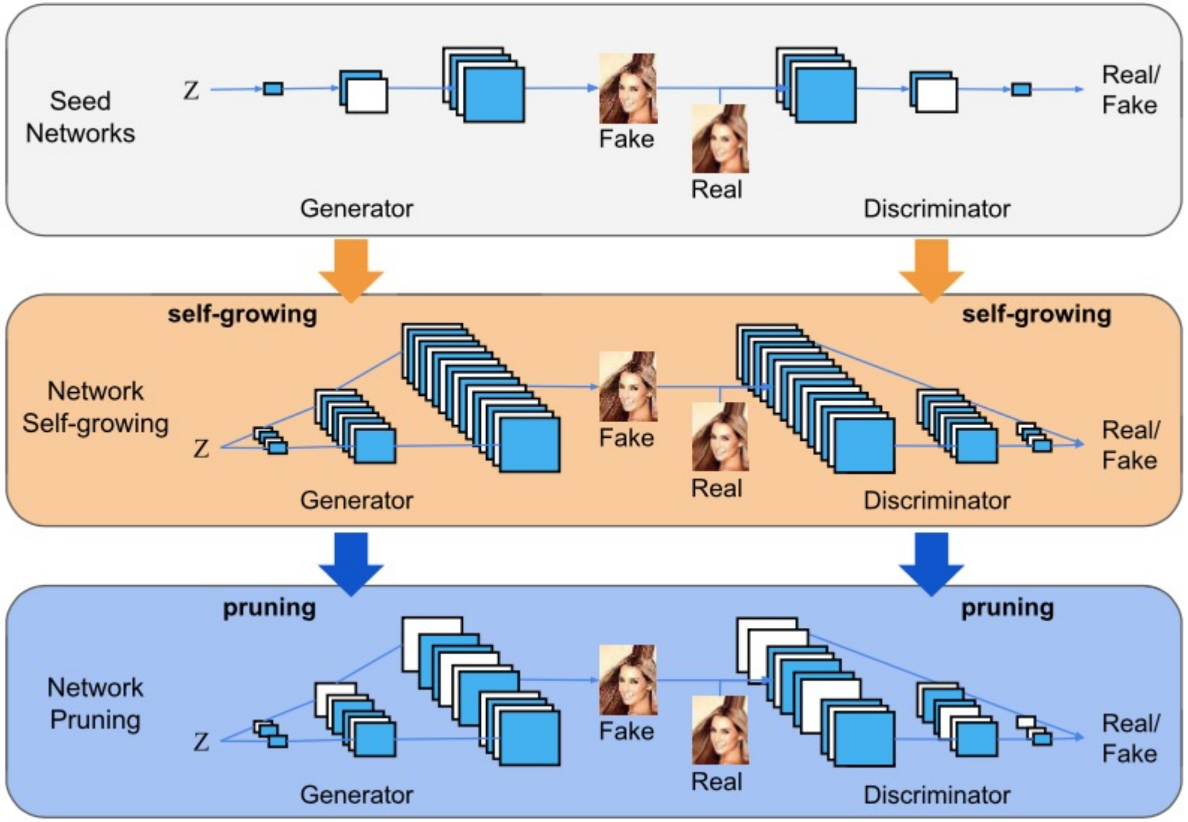
- Generator is a bitstream where each bit corresponds to a filter if the bit = 0 then the filter is pruned.
- Fitness:
  - the size of the network,
  - the compression distance,
  - the cycle loss.



H. Shu, ICCV 2019

Pruning during training

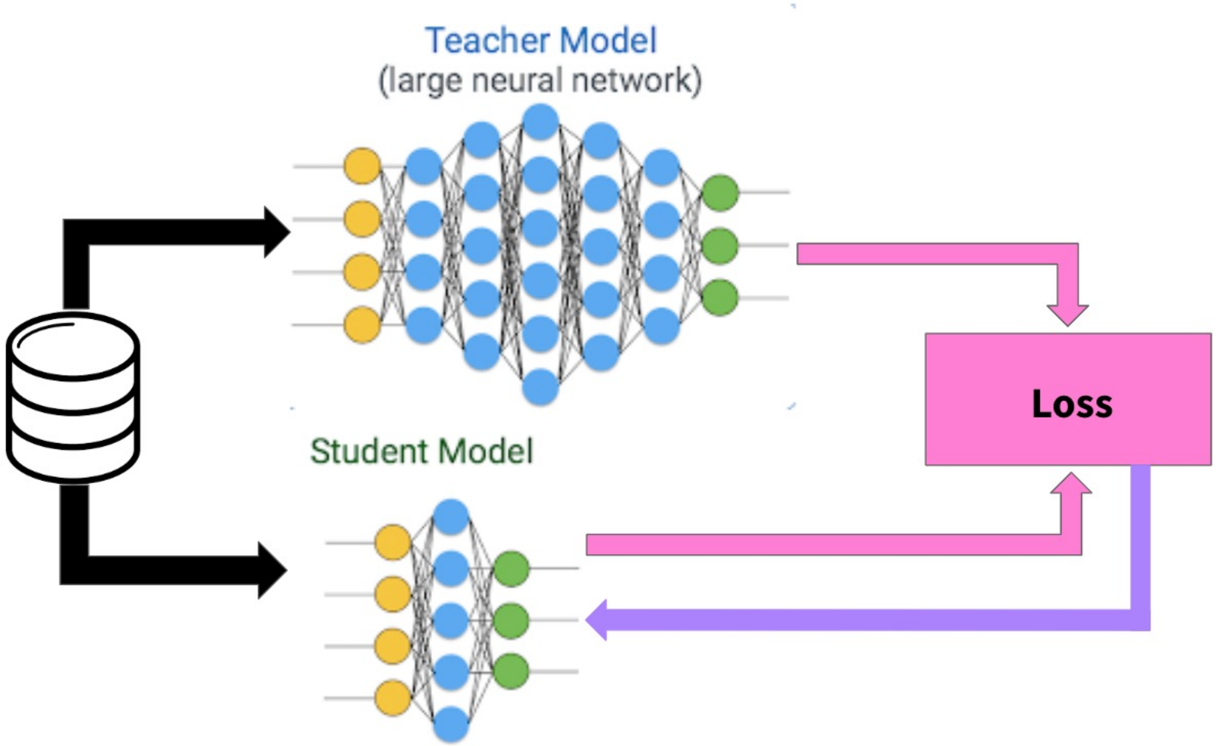
- Start from light-weight seed network.
- Grow width.
- Remove similar filters.



X. Song, SP-GAN: Self-Growing and Pruning Generative Adversarial Networks, IEEE TNN 2021

Knowledge Distillation Choices

- Teacher model acquiring
- Student model reconstruction
  - Pruning/architecture search/progressive growth
- Training Architecture
  - G/G+D/G+D+NN
- Loss functions

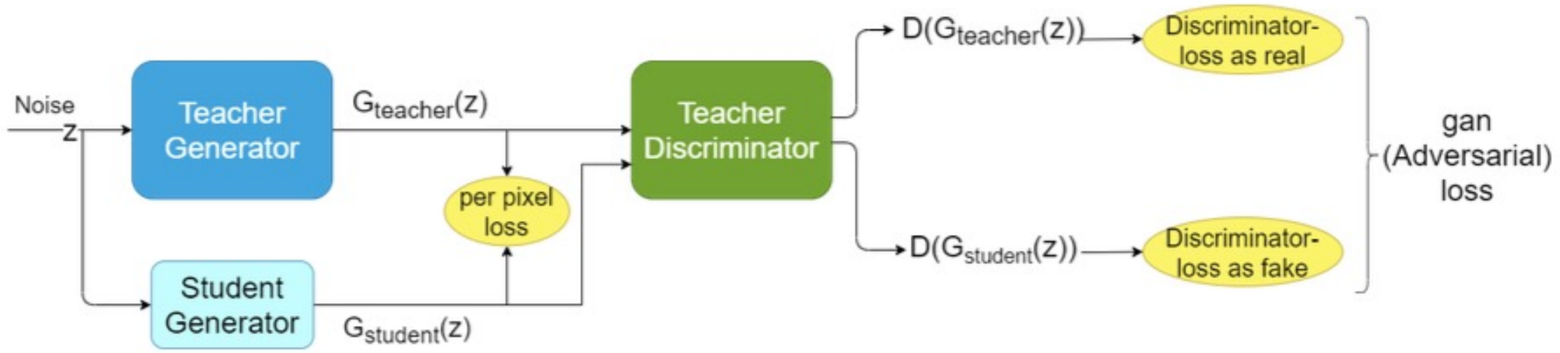


Direct Distillation

- Use pretrained teacher.
- Need per-pixel loss to stabilize student.

$$\mathcal{L}_{per\_pixel} = loss(G_{student}(z), G_{teacher}(z))$$

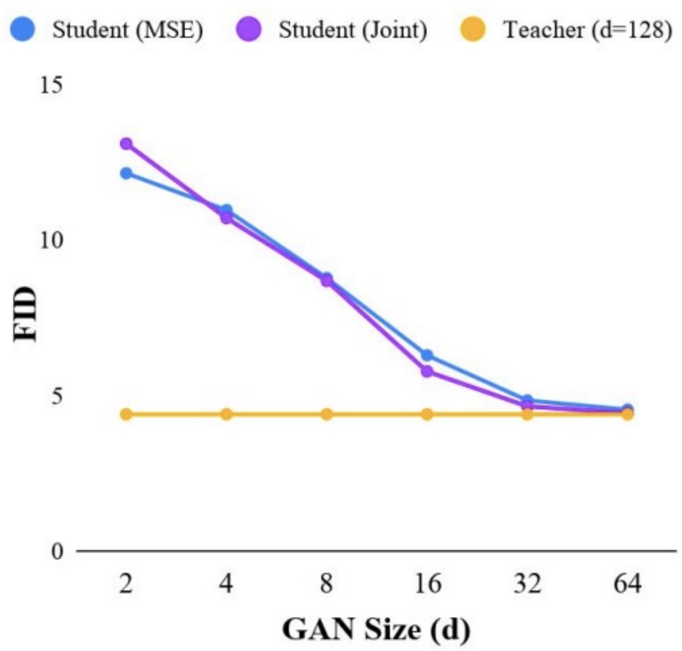
$$\mathcal{L}_{recon} = \mathcal{L}_{gan} + \lambda \mathcal{L}_{per\_pixel}$$

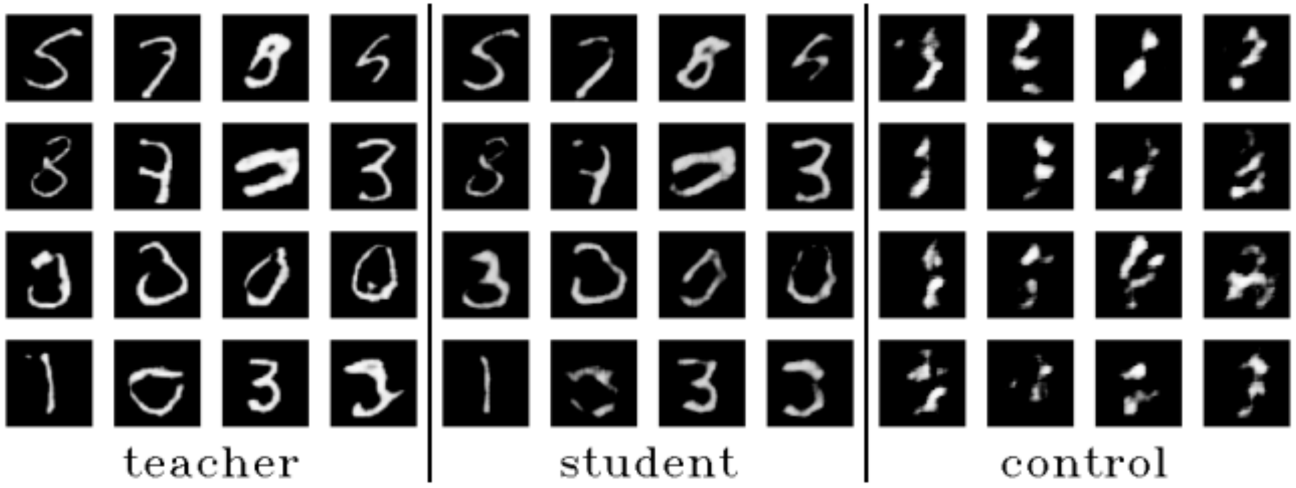


Aguinaldo et al., arXiv:1902.00159

Direct Distillation

- Clear dependence of quality on the complexity of student.
- Almost independent on loss-type.
- Results are much better than for direct training of small generator.



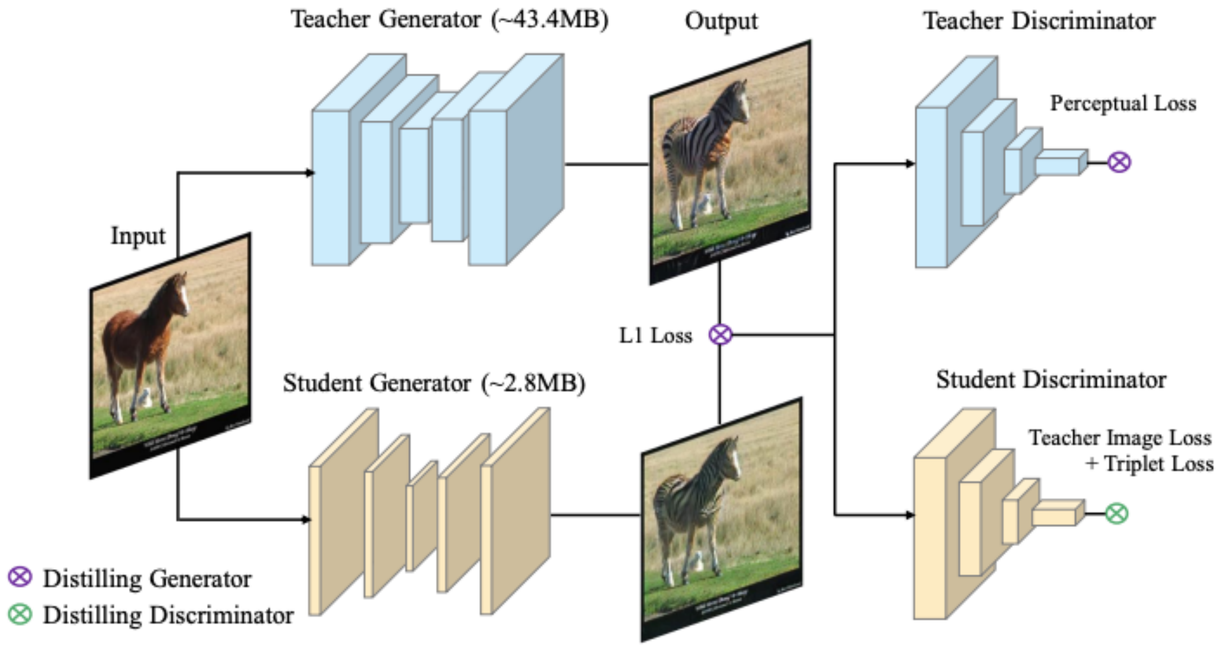


Aguinaldo et al., arXiv:1902.00159

Simultaneous Distillation

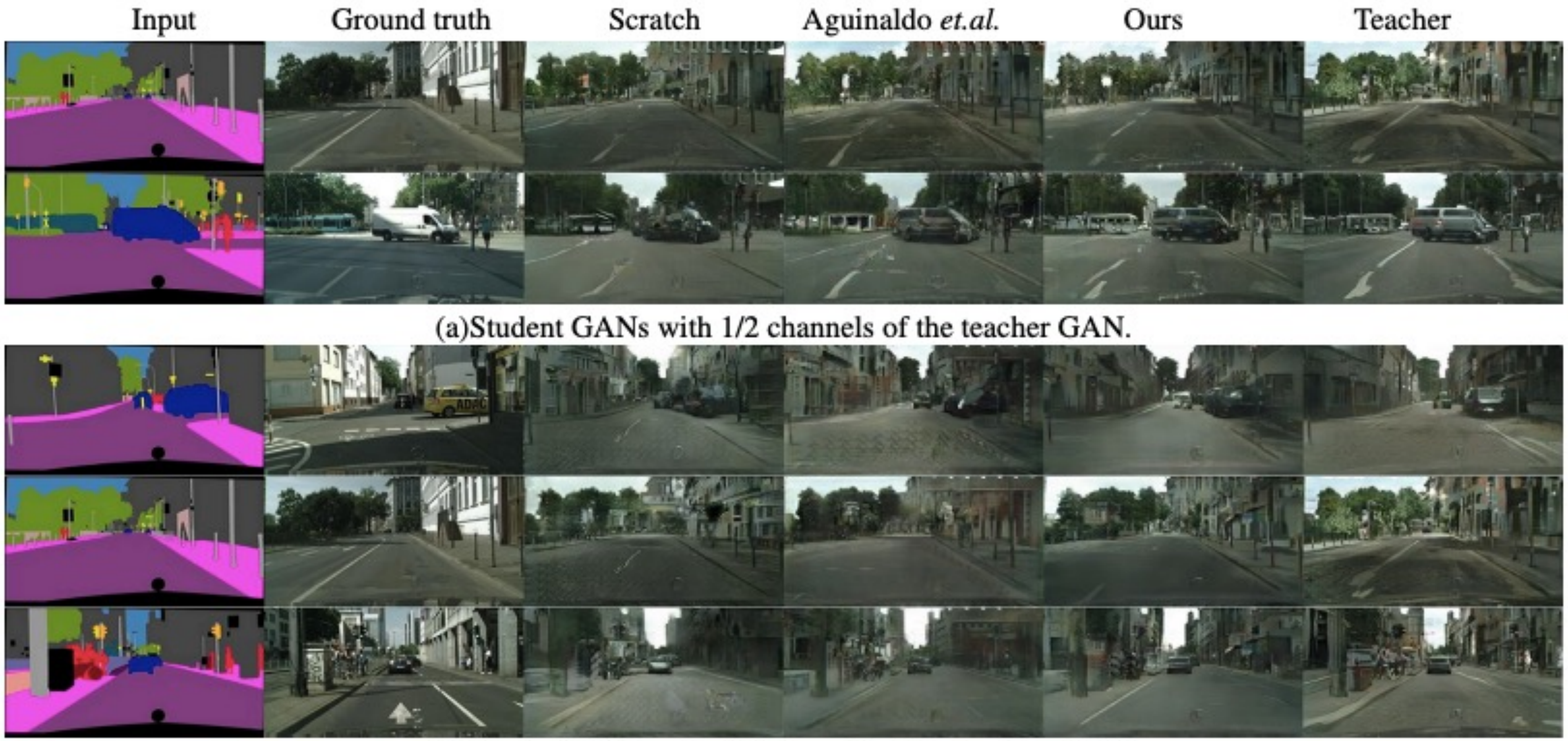
- Triplet loss: to consider the fact that student is closer to teacher than teacher to reality.

$$\mathcal{L}_{tri}(D_S) = \frac{1}{n}\sum_{i=1}^{n}\Big[\|\hat{D}_S(y_i) - \hat{D}_S(G_T(x_i))\|_1 - \|\hat{D}_S(y_i) - \hat{D}_S(G_S(x_i))\|_1 + \alpha\Big]_+$$



H. Chen et al., Distilling portable Generative Adversarial Networks for Image Translation, AAAI 2020

Simultaneous Distillation



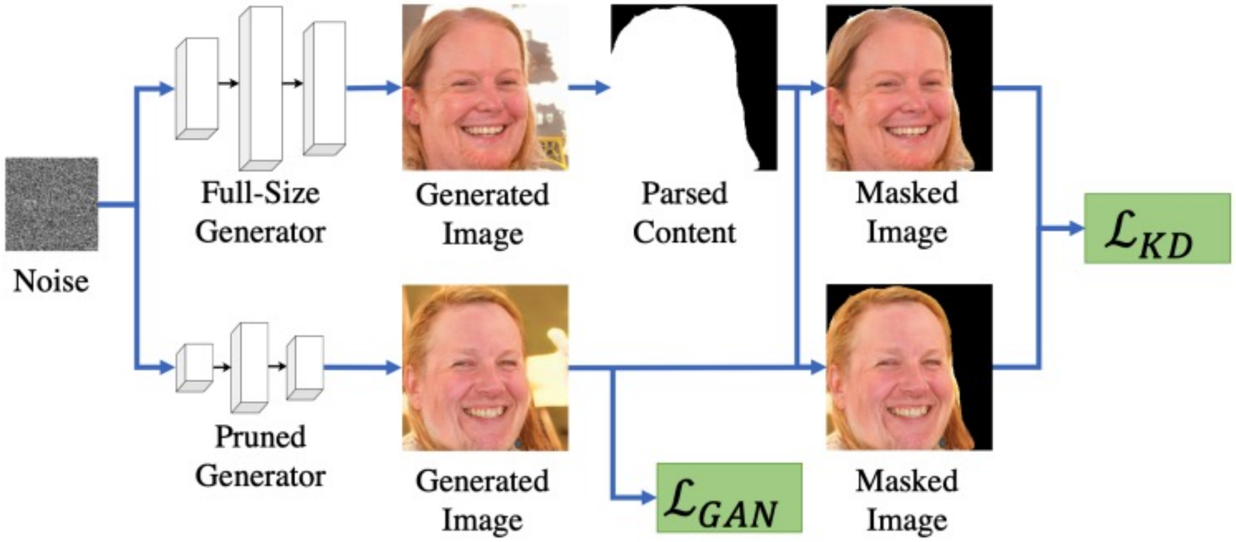
(a)Student GANs with 1/2 channels of the teacher GAN.

(b)Student GANs with 1/4 channels of the teacher GAN.

H. Chen et al., Distilling portable Generative Adversarial Networks for Image Translation, AAAI 2020

Content-aware loss

- Use masks to put correct attention to the content.
- The masked are specifically predefined.



Y. Liu et al., CVPR2021

Outlook

- More methods are available (like quantization).
- Current state-of-the-art methods use loss with several components.
- No unified approach exist.