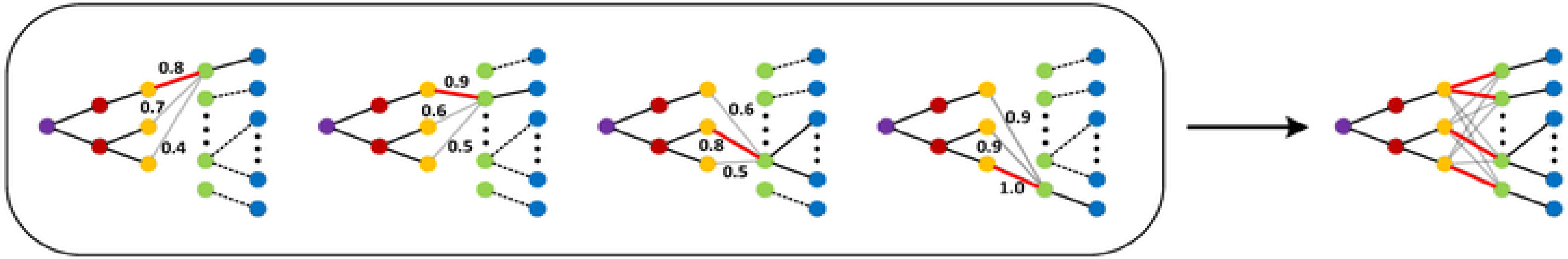
Teachers Do More Than Teach: Compressing Image-to-Image Models (CVPR 2021)

Qing Jin[1], Jian Ren[2], Oliver J. Woodford[*], Jiazhuo Wang[2], Geng Yuan[1], Yanzhi Wang[1], Sergey Tulyakov[2]
[1]Northeastern University, [2]Snap Inc.

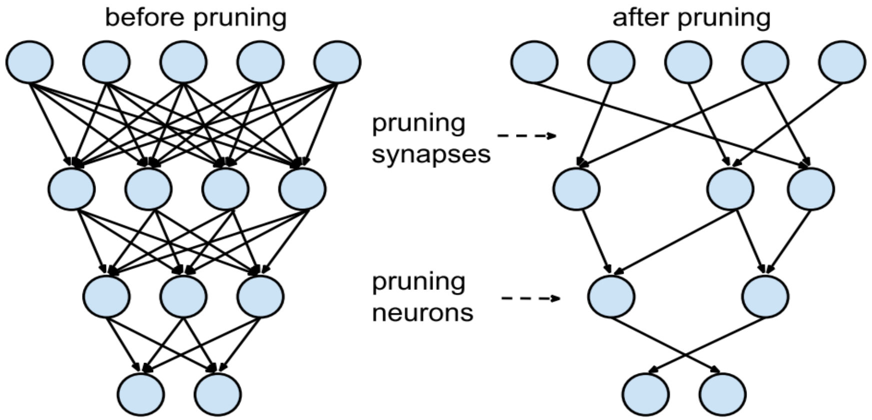
발표: 정채연

Existing Approaches

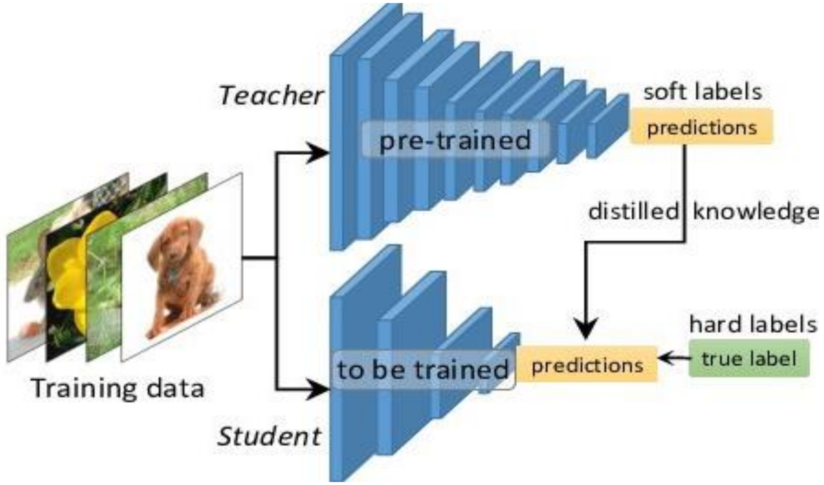
1. Network Architecture Search (NAS)



2. Pruning (e.g., weight, **channel**, layer, etc.)



3. Knowledge Distillation



Existing Approaches

1. **Co-evolutionary compression for unpaired image translation** (ICCV 2019)
   Han Shu, Yunhe Wang, Xu Jia, Kai Han, Hanting Chen, Chunjing Xu, Qi Tian, and Chang Xu

2. **AutoGAN-distiller: Searching to compress generative adversarial networks** (ICML 2020)
   Yonggan Fu, Wuyang Chen, Haotao Wang, Haoran Li, Yingyan Lin, and Zhangyang Wang

3. **Gan slimming: All-in-one GAN compression by a unified optimization framework** (ECCV 2020)
   Haotao Wang, Shupeng Gui, Haichuan Yang, Ji Liu, and Zhangyang Wang

4. **Single path one-shot neural architecture search with uniform sampling** (ECCV 2020)
   Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun

5. **Gan compression: Efficient architectures for interactive conditional GANs** (CVPR 2020)
   Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han

6. **GANs can play lottery tickets too** (ICLR 2021)
   Xuxi Chen, Zhenyu Zhang, Yongduo Sui, Tianlong Chen

→ 문제점: 높은 search cost, (original model에 비해) 낮은 경량화 model의 성능

Contributions

TL;DR: Ours is simpler but better than original.

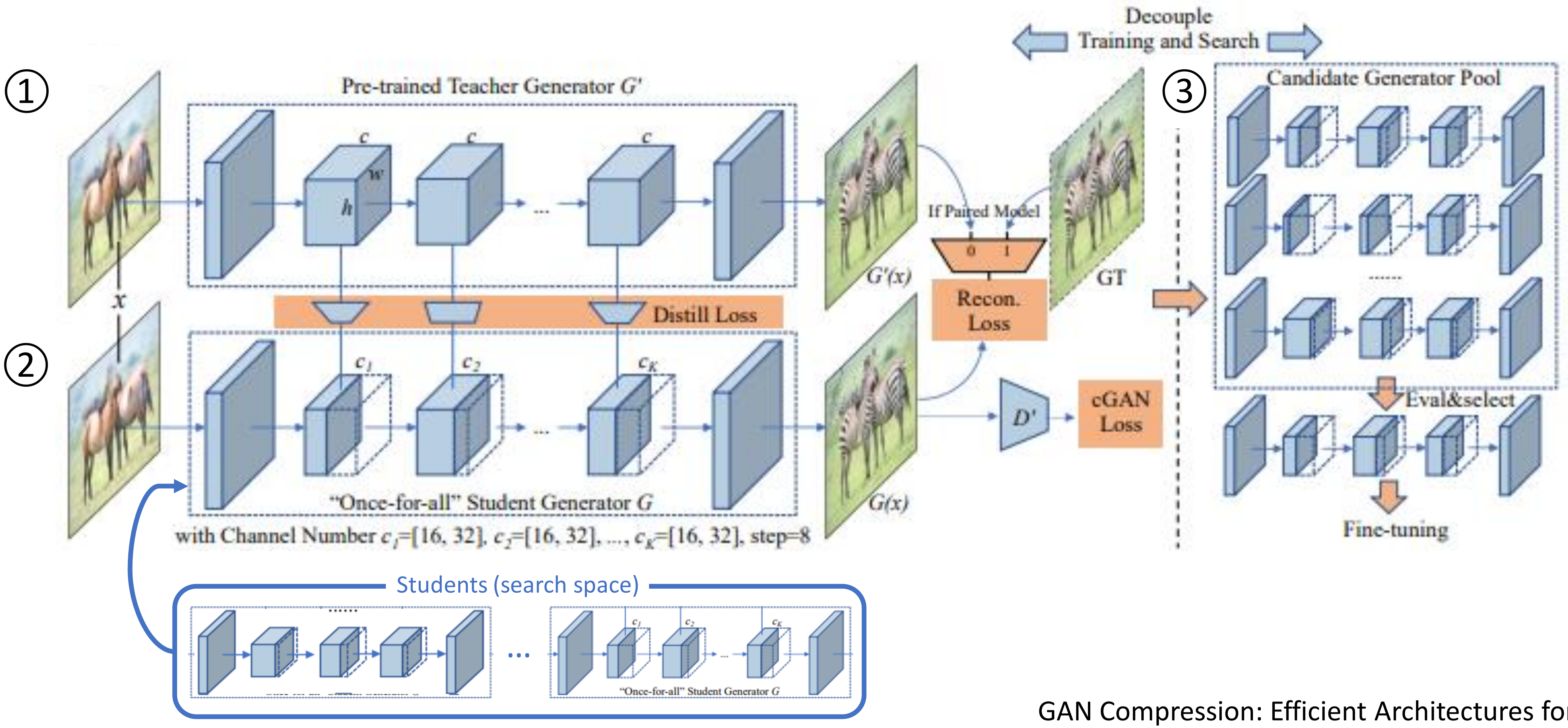
1. We introduce a new network design, **Compression And Teaching (CAT)**, which serves as both teacher network and the architecture search space of (compressed) student.

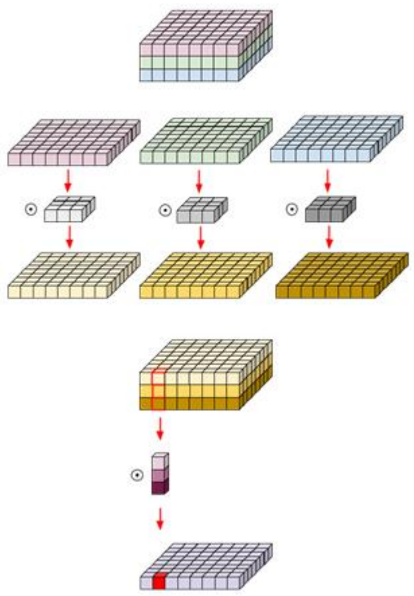
2. We propose an efficient one-step technique to directly prune the trained teacher network to achieve a target computation budget.

3. We introduce a knowledge distillation technique based on similarity between teacher and student models' feature spaces, global kernel alignment (GKA), without extra learnable layers.

Backgrounds

GAN Compression (CVPR2020)





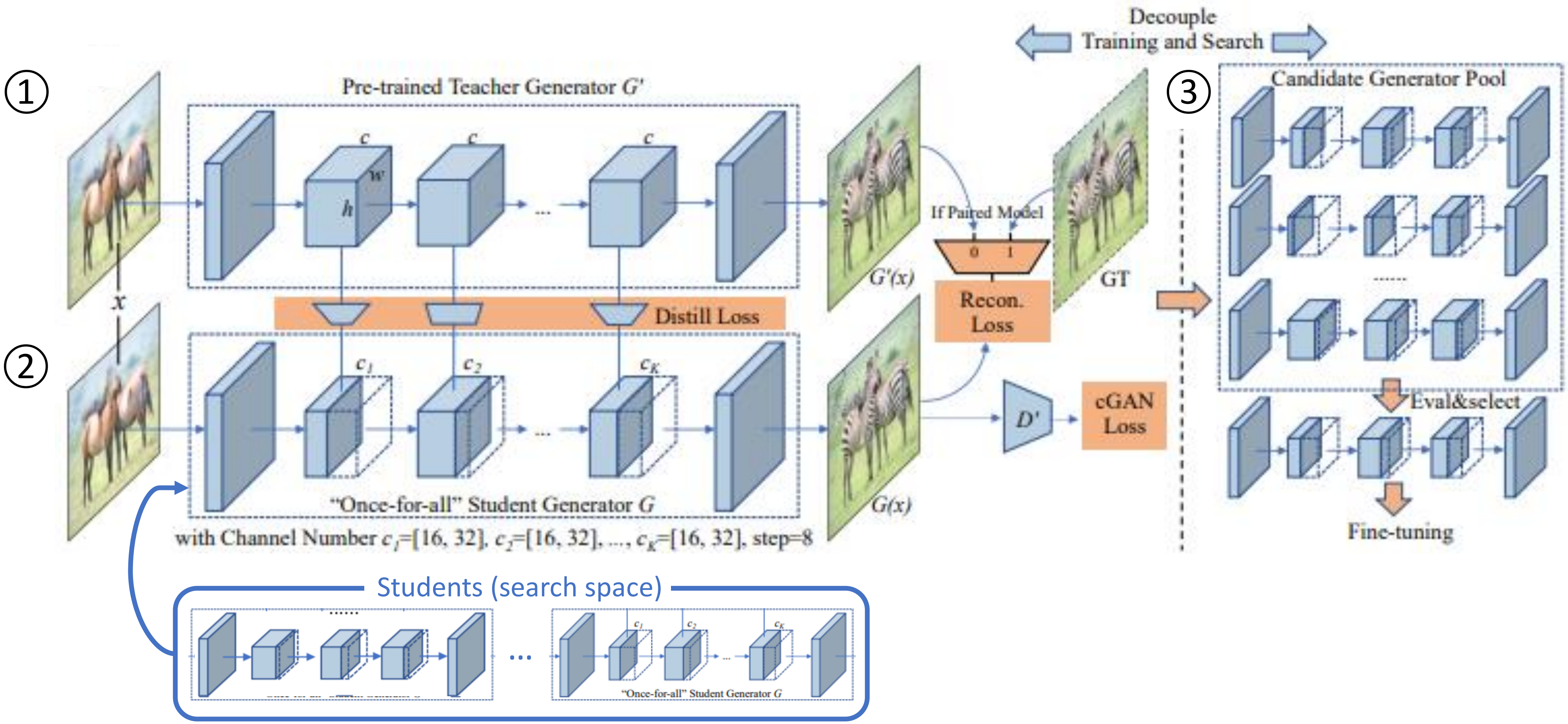
Depth-wise
Separable Conv

GAN Compression: Efficient Architectures for Interactive Conditional GANs link

[Muyang Li](), [Ji Lin](), [Yaoyao Ding](), [Zhijian Liu](), [Jun-Yan Zhu](), [Song Han]()

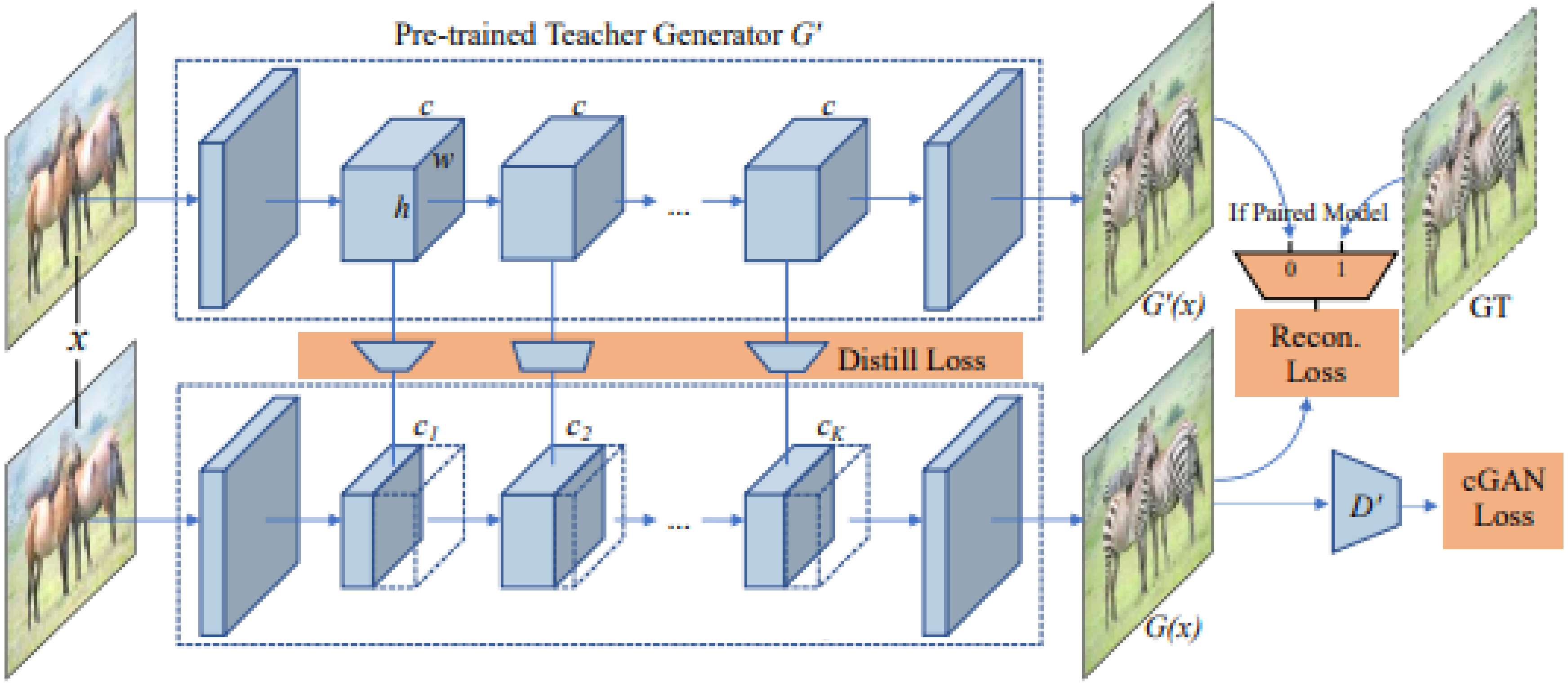
Method

Overview



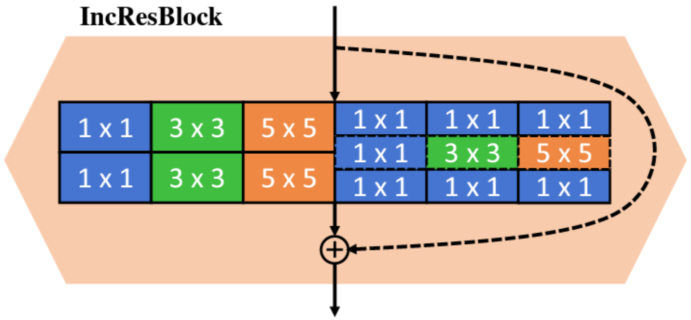
Method

Overview

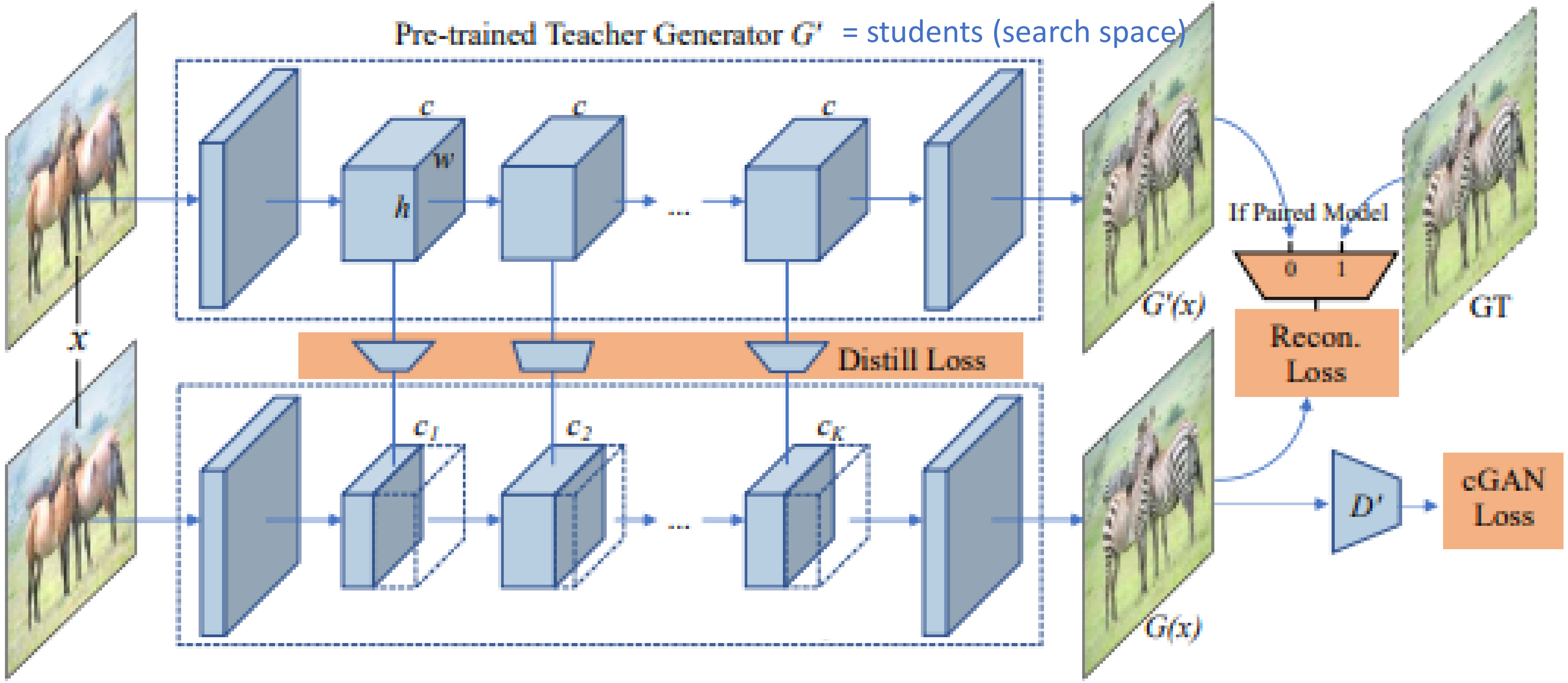


Method

Overview

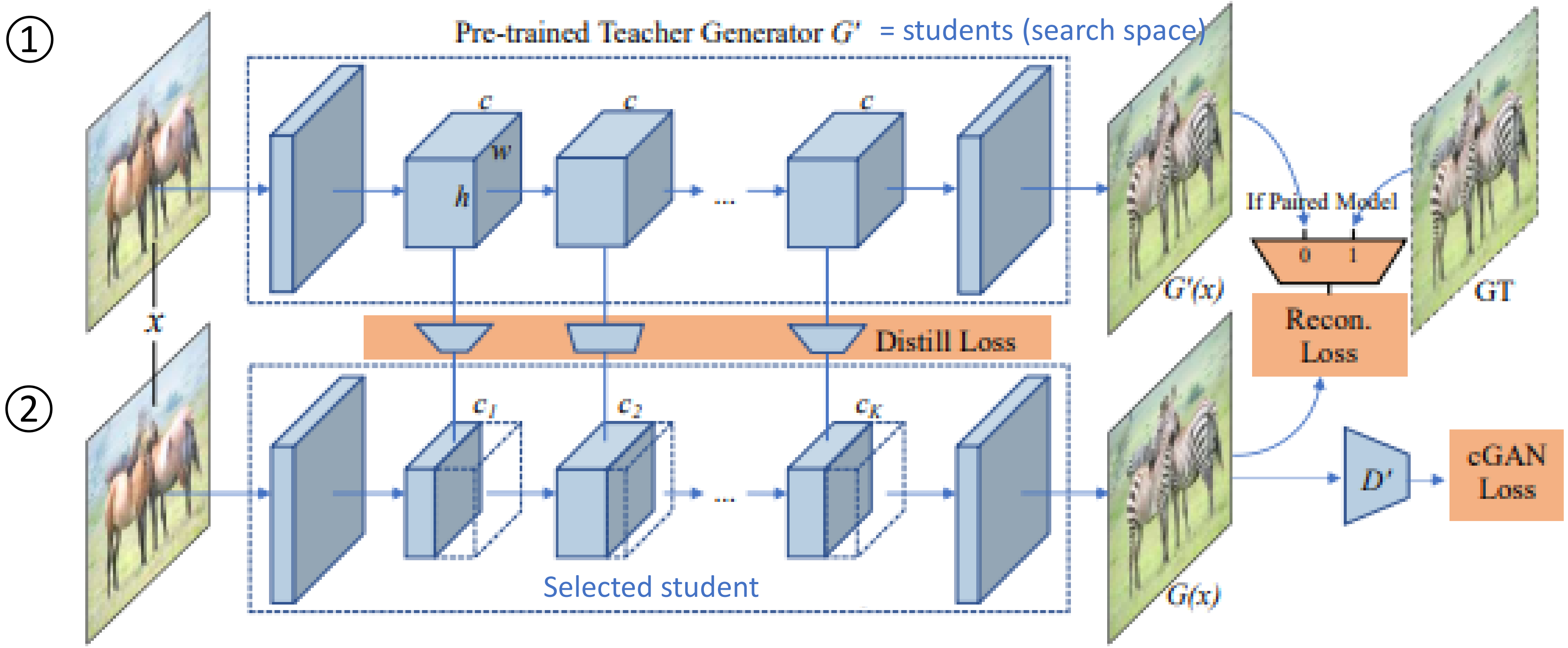


①



Method

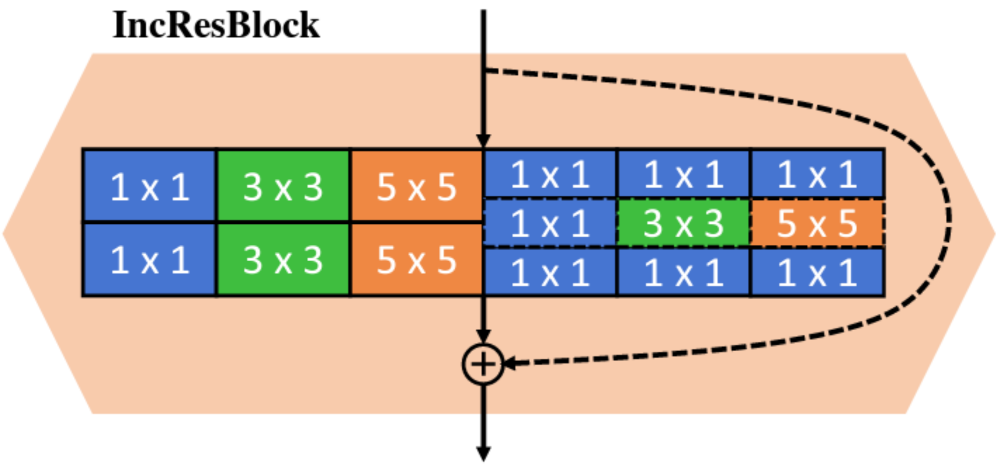
Overview



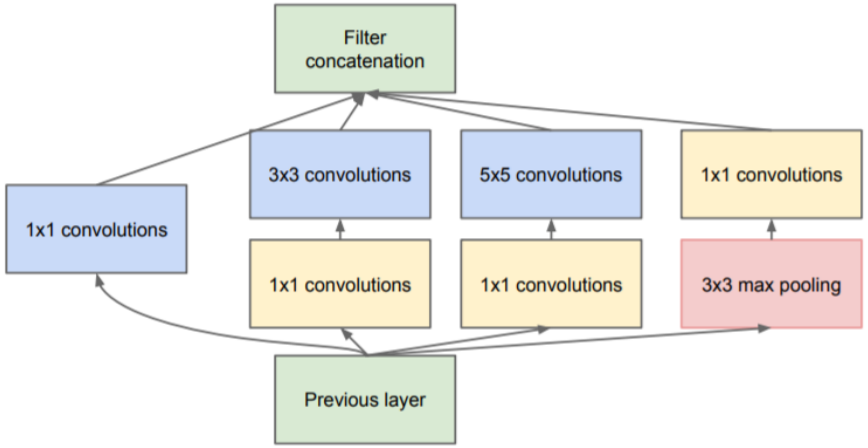
Method

1. Training Teacher Generator as Search Space of Students

- Residual block with 6 different operations
  - Conv with kernel size 1, 3, 5
  - Depth-wise conv with kernel size 1, 3, 5
- After 6 operations, sum all the features
- Each operations are followed by batch norm or instance norm layer



Inception Residual Block

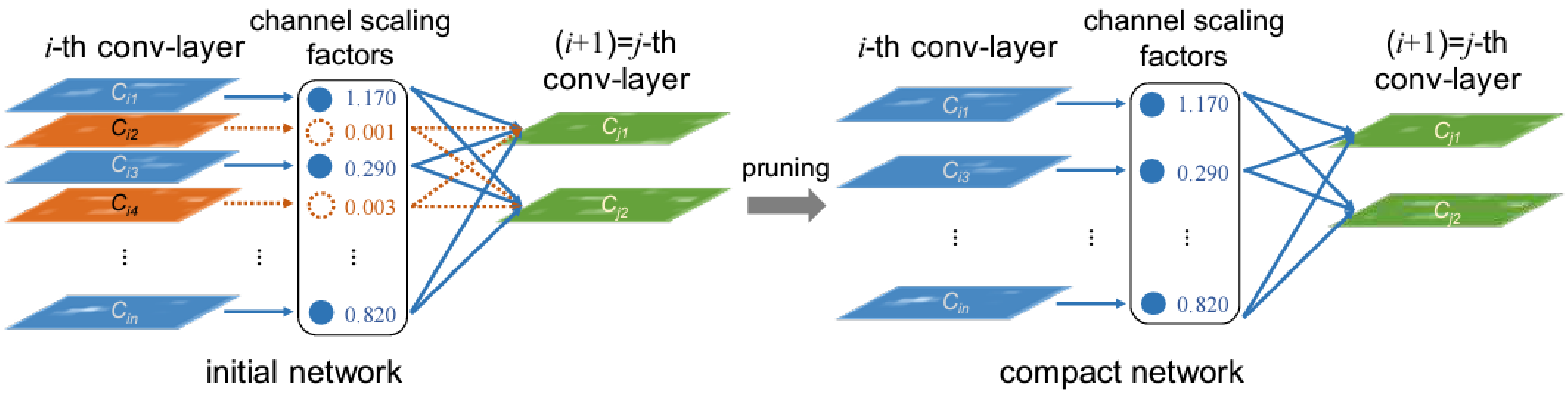


Original Inception Module

Method

2. Compressing Trained Teacher Generator via Pruning

- Channel pruning based on <u>scale factor of normalization layers</u>
- The larger the scale is, the more important the channel is
- Pruning until we satisfy our computation budget



Learning Efficient Convolutional Networks through Network Slimming (ICCV2017)
Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, Changshui Zhang

Algorithm 1 Searching via One-Step Pruning.

Require: Computational budget T_b, teacher model $\mathrm{G_T}$, scaling factors $\gamma_i^{(l)}$ (used for pruning) of the i-th channel in normalization layers $N^{(l)} \in \mathrm{G_T}$, minimum # output channels c_{lb} for convolution layers (outside the IncResBlock).

Ensure: pruned student architecture $\mathrm{G_S}$.

Initialize scale lower bound γ_{lo}: $\gamma_{lo} \leftarrow \min_{i,l} |\gamma_i^{(l)}|$.
Initialize scale upper bound γ_{hi}: $\gamma_{hi} \leftarrow \max_{i,l} |\gamma_i^{(l)}|$.
while $\gamma_{lo} < \gamma_{hi}$ **do**
  $\gamma_{th} \leftarrow (\gamma_{lo} + \gamma_{hi})/2$
  Prune channels satisfying $|\gamma_i^{(l)}| < \gamma_{th}$ on $\mathrm{G_T}$ while keep c_{lb} to get $\mathrm{G_S}$
  $T \leftarrow$ computational cost of $\mathrm{G_S}$
  if $T > T_\mathrm{b}$ **then**
    $\gamma_{lo} \leftarrow \gamma_{th}$
  else
    $\gamma_{hi} \leftarrow \gamma_{th}$
  end if
end while

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$$

Method

3. Training (Compressed) Student Network

- Centered Kernel alignment (CKA)
  = Centering + KA

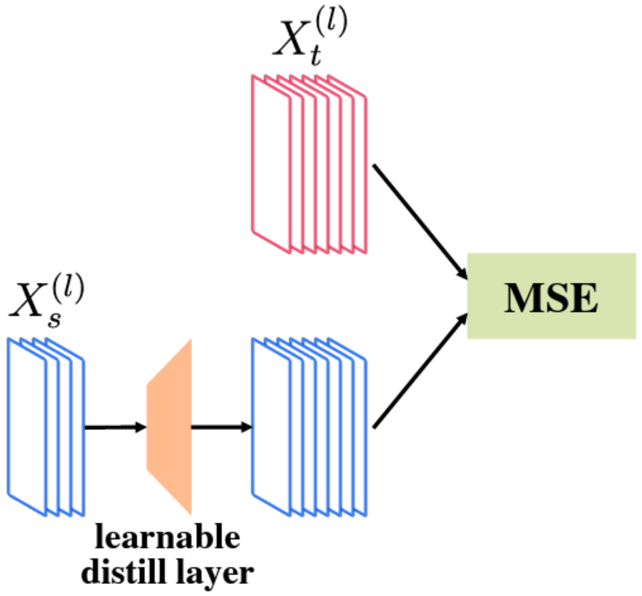
$$\mathrm{KA}(X,Y) = \frac{\|Y^{\mathrm{T}}X\|_{\mathrm{F}}^2}{\|X^{\mathrm{T}}X\|_{\mathrm{F}}\|Y^{\mathrm{T}}Y\|_{\mathrm{F}}}$$

$$\|Y^{\mathrm{T}}X\|_{\mathrm{F}}^2 = \langle \mathrm{vec}(XX^{\mathrm{T}}), \mathrm{vec}(YY^{\mathrm{T}})\rangle$$

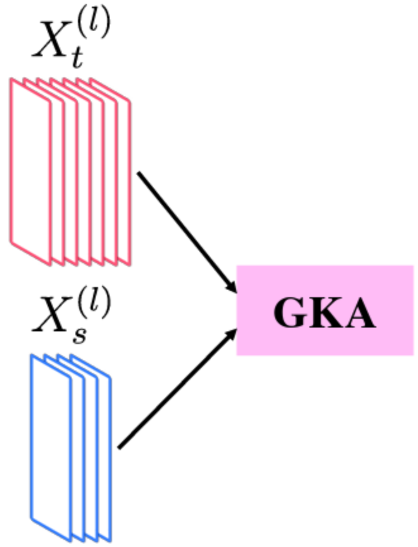
- Global KA (GKA)

$$\mathrm{GKA}(X,Y) = \mathrm{KA}(\rho(X), \rho(Y))$$

$$\rho : \mathbb{R}^{n\times hwc} \to \mathbb{R}^{nhw\times c}$$



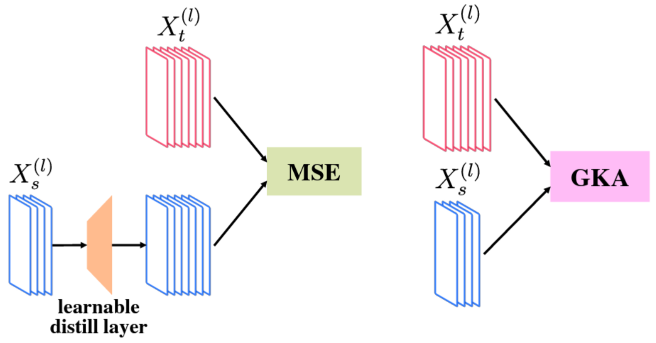
원래 GAN Compression의
Knowledge distillation loss



Newly-proposed
Knowledge distillation loss

Method

3. Training (Compressed) Student Network



- Centered Kernel alignment (CKA)
  = Centering + KA

$$\mathrm{KA}(X,Y) = \frac{\|Y^{\mathrm{T}}X\|_{\mathrm{F}}^2}{\|X^{\mathrm{T}}X\|_{\mathrm{F}}\|Y^{\mathrm{T}}Y\|_{\mathrm{F}}}$$

$$\|Y^{\mathrm{T}}X\|_{\mathrm{F}}^2 = \langle \mathrm{vec}(XX^{\mathrm{T}}), \mathrm{vec}(YY^{\mathrm{T}}) \rangle$$

- Global KA (GKA)

$$\mathrm{GKA}(X,Y) = \mathrm{KA}(\rho(X), \rho(Y))$$

$$\rho : \mathbb{R}^{n \times hwc} \rightarrow \mathbb{R}^{nhw \times c}$$

- Total loss

$$\mathcal{L}_{\mathrm{T}} = \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{recon}}\mathcal{L}_{\mathrm{recon}} + \lambda_{\mathrm{dist}}\mathcal{L}_{\mathrm{dist}}$$

$$\mathcal{L}_{\mathrm{adv}} = \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\log D(\mathbf{x},\mathbf{y})\right] + \mathbb{E}_{\mathbf{x}}\left[\log(1 - D(\mathbf{x}, G(\mathbf{x})))\right]$$

$$\mathcal{L}_{\mathrm{dist}} = -\sum_{l \in \mathcal{S}_{\mathrm{KD}}} \mathrm{GKA}(X_t^{(l)}, X_s^{(l)})$$

Experiments

Quantitative Comparison with Baselines

Table 1: Quantitative comparison between different compression techniques for Image-to-Image models. We use mIoU to evaluate the generation quality of Cityspaces and FID for other datasets. Higher mIoU or lower FID indicates better performance.

Model	Dataset	Method	MACs	FID↓	mIoU↑
CycleGAN	Horse→Zebra	Original [83, 36]	56.8B	61.53	-
		Shu *et al.* [63]	13.4B	96.15	-
		AutoGAN Distiller [20]	6.39B	83.60	-
		GAN Slimming [68]	11.25B	86.09	-
		GAN Lottery [11]	~11.35B†	~83.00†	-
		Li *et al.* [36]	2.67B	71.81	-
		CAT (Ours)	**2.55B**	**60.18**	-
	Zebra→Horse	Original [83, 68]	56.8B	148.81	-
		GAN Slimming [68]	11.81B	120.01	-
		CAT (Ours)	2.59B	142.68	-
Pix2pix	Cityscapes	Original [29, 36]	56.8B	-	42.06
		Li *et al.* [36]	5.66B	-	40.77
		CAT (Ours)	**5.57B**	-	**42.53**
	Map→Aerial photo	Original [29, 36]	56.8B	47.76	-
		Li *et al.* [36]	4.68B	48.02	-
		CAT (Ours)	**4.59B**	**44.96**	-
GauGAN	Cityscapes	Original [57, 36]	281B	-	62.18
		Li *et al.* [36]	31.7B	-	61.22
		CAT-A (Ours)	**29.9B**	-	**62.35**
		CAT-B (Ours)	5.52B	-	54.71

Table 2: Further quantitative comparison on KID between different compression techniques for Image-to-Image models, where lower KID indicates better performance.

Model	Dataset	Method	MACs	KID↓
CycleGAN	Horse→Zebra	Original [83]	56.8B	0.020±0.002
		CAT (Ours)	**2.55B**	**0.017±0.002**
	Zebra→Horse	Original [83]	56.8B	0.030±0.002
		CAT (Ours)	2.59B	0.036±0.002
Pix2pix	Map→Aerial	Original [29]	56.8B	0.154±0.010
		CAT (Ours)	**4.6B**	**0.009±0.002**
GauGAN	Cityscapes	Original [57]	281B	0.026±0.003
		CAT-A (Ours)	**29.9B**	**0.014±0.002**
		CAT-B (Ours)	**5.5B**	**0.013±0.002**

Experiments

Qualitative Comparison with Baselines

Input

CAT (Ours)
MACs: 2.56B
FID: 53.48
KID: 0.015±0.001

CycleGAN
MACs: 56.8B
FID: 61.53
KID: 0.020±0.002

Experiments

Qualitative Comparison with Baselines

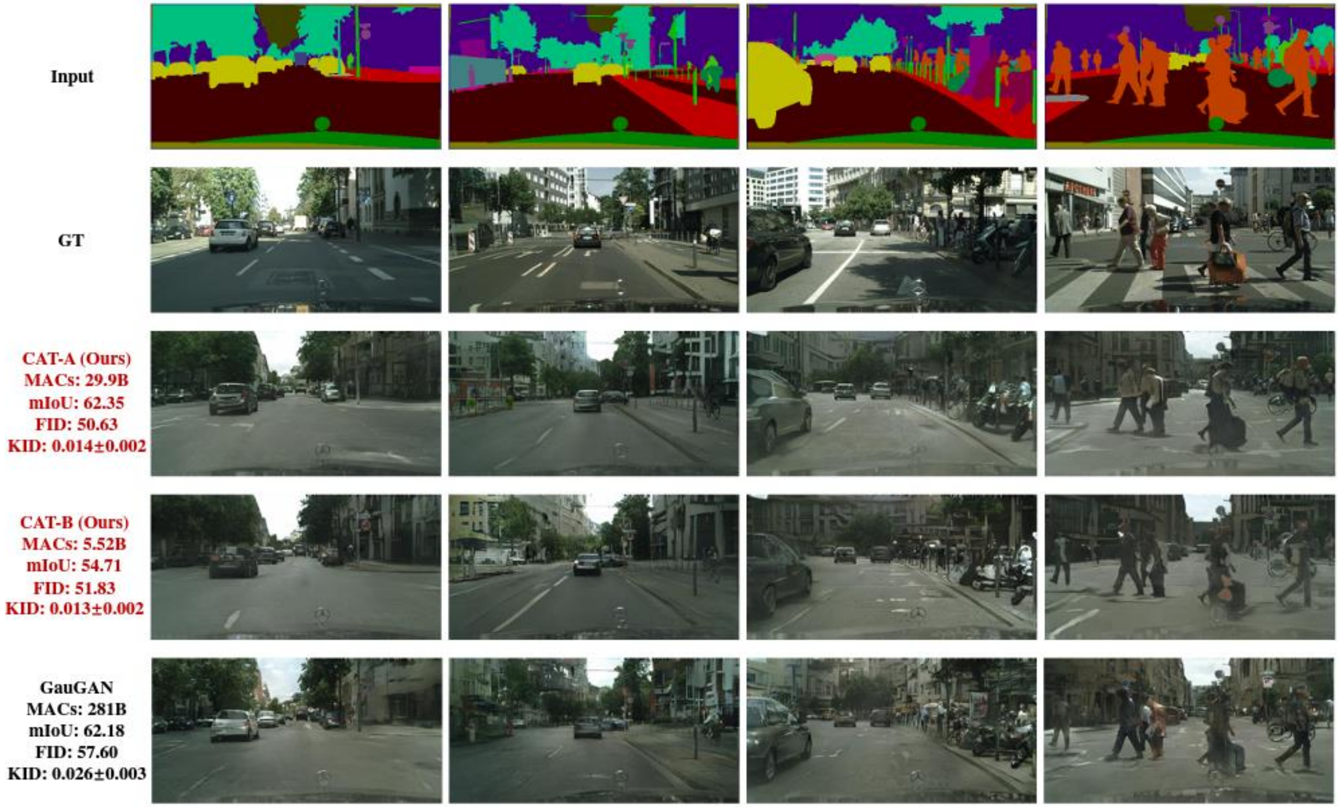
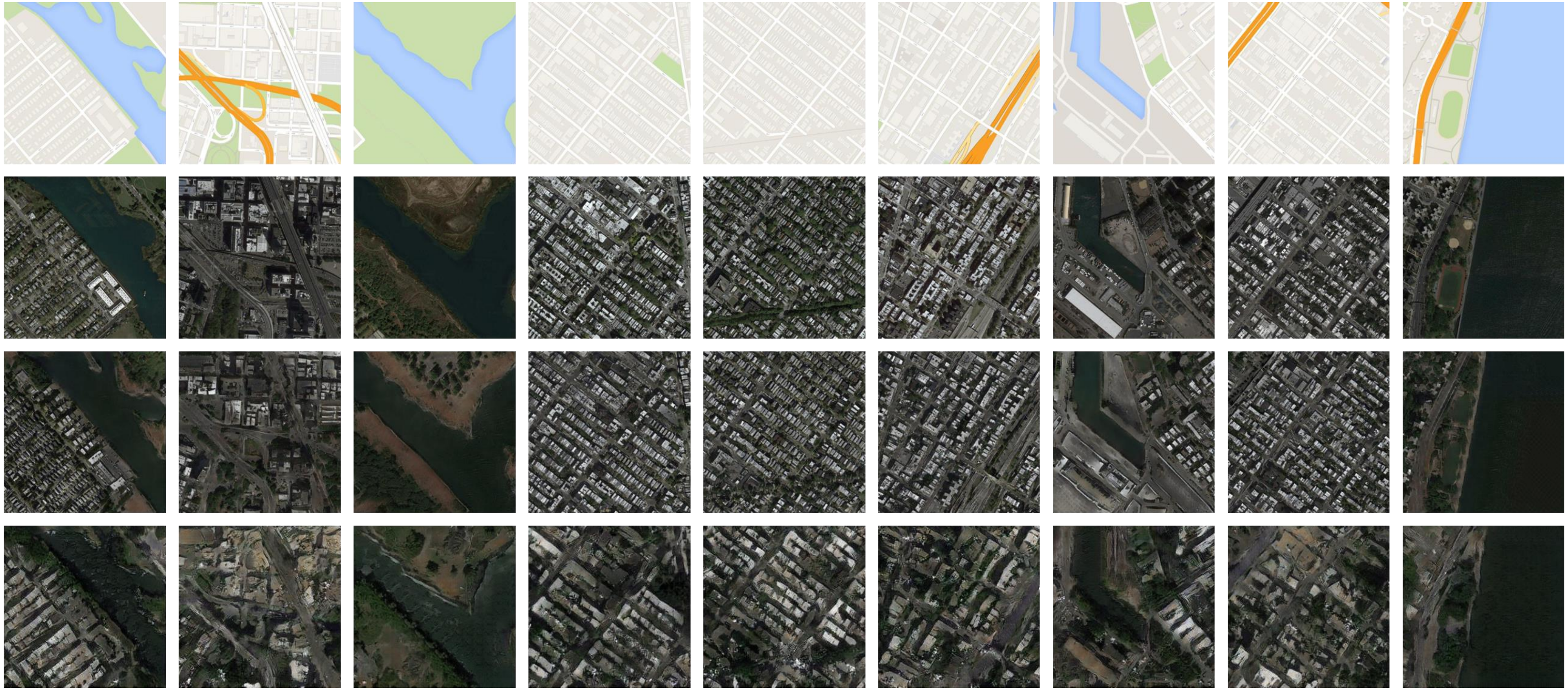


Figure 4: Qualitative results on Cityscapes dataset. Images generated by our compressed model (CAT-A, third row) have higher mIoU and lower FID than the original GauGAN model (fifth row), even with much reduced computational cost. For our CAT-B model (fourth row, $50.9\times$ compressed than GauGAN), although it has lower mIoU, the CAT-B model can synthesize higher fidelity images (lower FID) than GauGAN.

Experiments

Qualitative Comparison with Baselines



Input

GT

CAT (Ours)
MACs: 4.59B
FID: 45.63
KID: 0.012±0.002

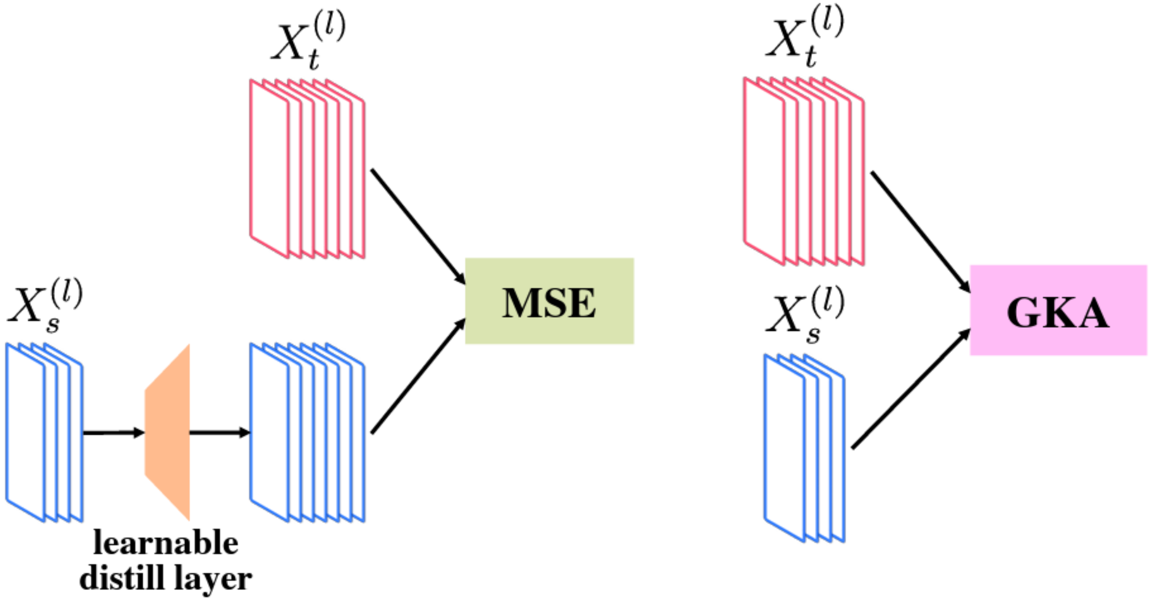
Pix2pix
MACs: 56.8B
FID: 47.76
KID: 0.154±0.010

Experiments

Ablation Analysis of Knowledge Distillation

Table S2: Analysis of knowledge distillation methods on Cityscapes dataset with the Pix2pix setting. Our methods (GKA) achieves the best result.

Method	mIoU↑
w/o Distillation	39.39
w/ MSE; Loss Weight 0.5	39.83
w/ MSE; Loss Weight 1.0	39.76
Ours	**42.53**



원래 GAN Compression의 Knowledge distillation loss

Newly-proposed Knowledge distillation loss

Experiments

Analysis of Searching Cost

Table 3: Architecture search cost, measured in seconds of GPU computation, for our method vs. Li *et al*. [36], across different models.

Model	Dataset	Method	Search Cost (GPU Seconds)
CycleGAN	Horse→Zebra	Li *et al*. [36]	$\gtrsim 7.2 \times 10^4$
		CAT (Ours)	**3.81**
	Zebra→Horse	Li *et al*. [36]	$\gtrsim 7.2 \times 10^4$
		CAT (Ours)	**3.62**
Pix2pix	Cityscapes	Li *et al*. [36]	$\gtrsim 7.2 \times 10^4$
		CAT (Ours)	**4.28**
	Map→Aerial photo	Li *et al*. [36]	$\gtrsim 7.2 \times 10^4$
		CAT (Ours)	**4.33**
GauGAN	Cityscapes	Li *et al*. [36]	$\gtrsim 1.2 \times 10^6$
		CAT-A (Ours)	**8.22**
		CAT-B (Ours)	**6.20**

Thank you