Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping

Fu H, Gong M, Wang C, Batmanghelich K, Zhang K, Tao D

CVPR 2019

CV&DL for Autonomous Driving

Erkam Uyanik

Technische Universität München + Johns Hopkins Whiting School of Engineering

Outline

1. Problem Definition
2. Concepts
3. State of the Art
4. Proposed Method
5. Experiments
6. Conclusion

Problem Definition: Domain Mapping

- **Domain mapping** targets translating an image from one domain to another

- $\mathcal{X}, \mathcal{Y}$ domains
- $X \in \mathcal{X}, Y \in \mathcal{Y}$ random variables
- $x \in X, y \in Y$ samples

- Goal: learn a function G_{XY} to transform domain $\mathcal{X}$ to $\mathcal{Y}$

aerial images

→

maps

Problem Definition: Domain Mapping

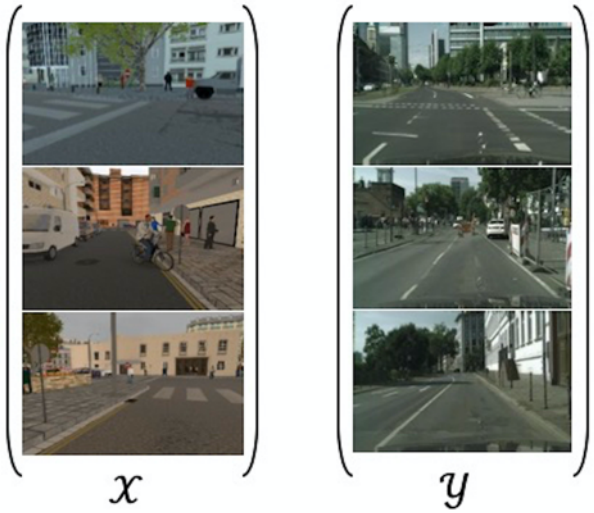
- Supervised Domain Mapping
  - significant progress from variety of literature
  - Paired data collection can be time consuming and expensive.

- Unsupervised Domain Mapping
  - Finding the optimal G_{XY} without paired data is an

    ill posed problem

    ➢ no unique solution

  - Appropriate **constraints** are needed



- GANs are popular for domain adaptation and similar tasks

Concept: Generative Adversarial Networks

- GANs learn two networks, a generator and a discriminator, in a zero-sum game setup.

- **Adversarial constraint** enforces generated images to be indistinguishable from real images.

$$\mathcal{L}_{gan} = \mathbb{E}_{y \sim P_Y}[\log D_Y(y)] + \mathbb{E}_{x \sim P_X}[\log(1 - D_Y(G_{XY}(x)))].$$

- D_Y and G_{XY} simultaneously optimize each other

State of the Art: CycleGAN

- Cycle consistency assumption

- a mapping G_{XY} and its inverse G_{YX} should be **bijections**

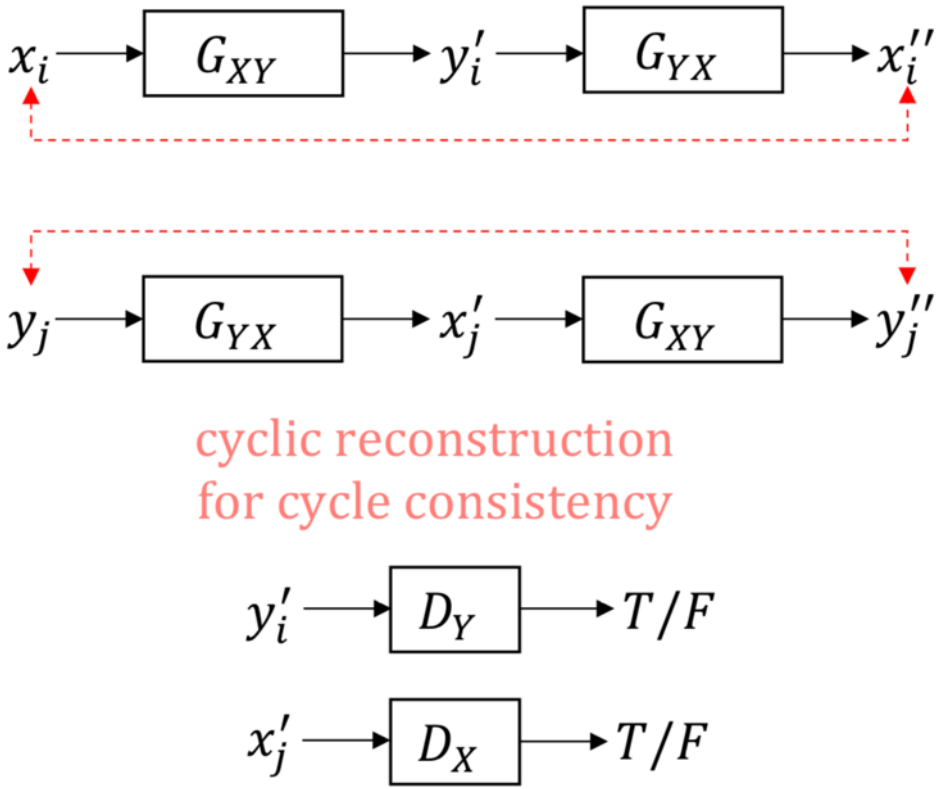
$$\mathcal{L}_{cyc} = \mathbb{E}_{x \sim P_X}\left[\|G_{YX}(G_{XY}(x)) - x\|_1\right] + \mathbb{E}_{y \sim P_Y}\left[\|G_{XY}(G_{YX}(y)) - y\|_1\right].$$

  - $x \approx G_{YX}\big(G_{XY}(x)\big)$
  - $y \approx G_{XY}\big(G_{YX}(y)\big)$

- **two-sided**: G_{XY} and G_{YX} needs to be jointly learned

❖ CycleGAN (Zhu et al. 2017), DiscoGAN (Kim et al. 2017), DualGAN (Yi et al. 2017)

State of the Art: CycleGAN



State of the Art: DistanceGAN

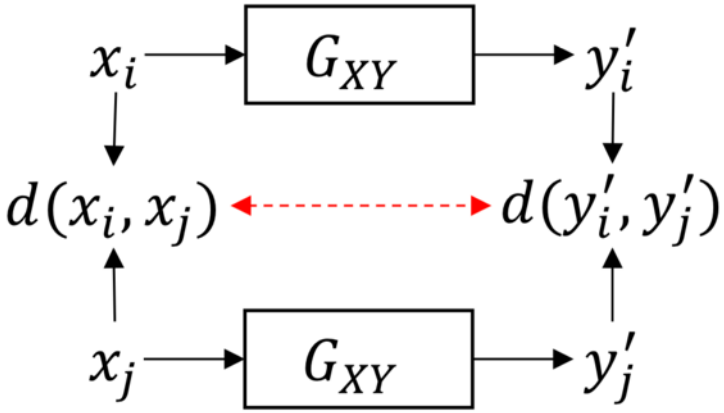
- Assumption: Distance between two examples x_i and x_j in domain $\mathcal{X}$ should be preserved after mapping to domain $\mathcal{Y}$.

$$\mathcal{L}_{dis} = \mathbb{E}_{x_i, x_j \sim P_X}[|\phi(x_i, x_j) - \psi(x_i, x_j)|],$$
$$\phi(x_i, x_j) = \frac{1}{\sigma_X}(\|x_i - x_j\|_1 - \mu_X),$$
$$\psi(x_i, x_j) = \frac{1}{\sigma_Y}(\|G_{XY}(x_i) - G_{XY}(x_j)\|_1 - \mu_Y),$$

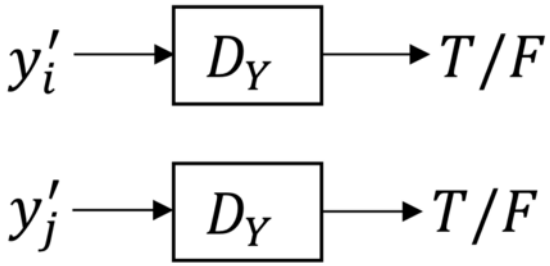
  - μ and σ are mean and stddev of distances of all possible pairs within domain

❖ Benaim et al. 2017

State of the Art: DistanceGAN



preserving $d(\cdot)$
for distance consistency



Proposed Method: GcGAN

- *Simple geometric transformations* do not change *semantic structure* of the image
  - Transformations without shape deformation
  - The information that distinguishes object classes is preserved

- Example:

  90° clockwise rotation

Proposed Method: GcGAN

- A geometric transformation $f(\cdot)$ between input images should be preserved by translators G_{XY} and $G_{\tilde{X}\tilde{Y}}$
- $\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}$: transformed domains by applying $f(\cdot)$
- **Geometry Consistency Constraint**

$$\mathcal{L}_{geo} = \mathbb{E}_{x\sim P_X}[\|G_{XY}(x) - f^{-1}(G_{\tilde{X}\tilde{Y}}(f(x)))\|_1] + \mathbb{E}_{x\sim P_X}[\|G_{\tilde{X}\tilde{Y}}(f(x)) - f(G_{XY}(x))\|_1].$$

  - $f\big(G_{XY}(x)\big) \approx G_{\tilde{X}\tilde{Y}}\big(f(x)\big)$
  - "reconstruction loss"

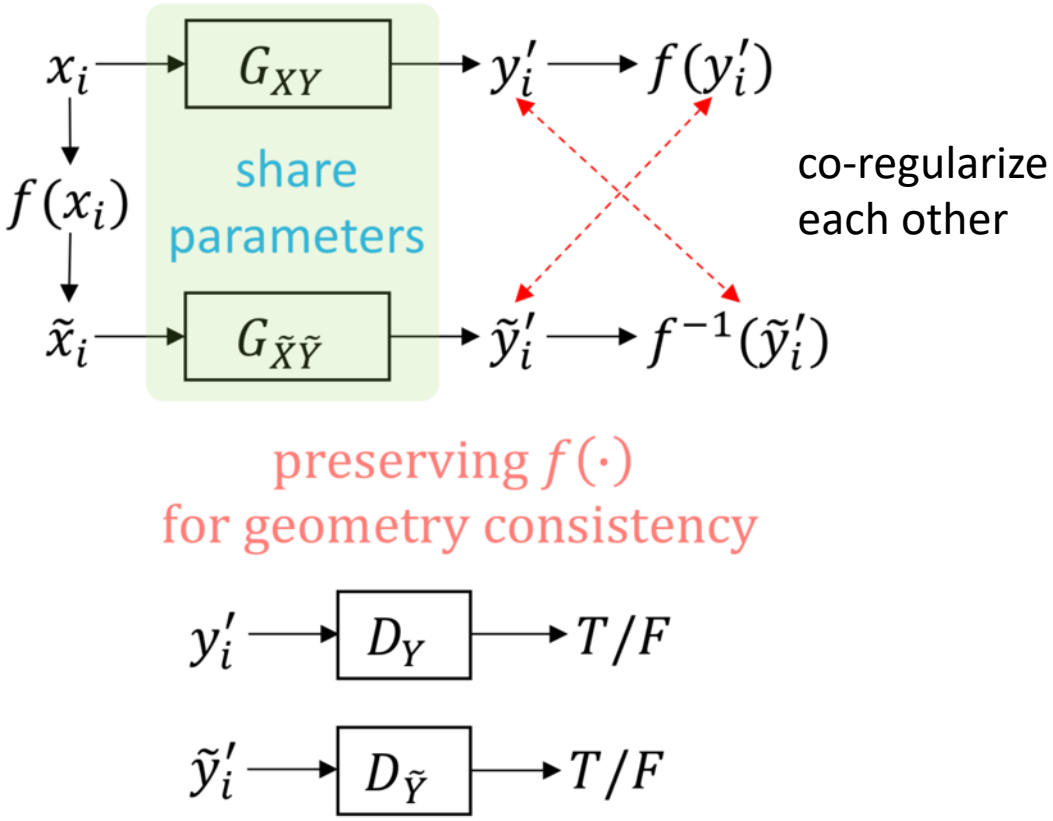
Proposed Method: GcGAN

$$\begin{aligned}\mathcal{L}_{GcGAN} = &\mathcal{L}_{gan}(G_{XY}, D_Y, X, Y) \\ &+ \mathcal{L}_{gan}(G_{\tilde{X}\tilde{Y}}, D_{\tilde{Y}}, X, Y) \\ &+ \lambda \mathcal{L}_{geo}(G_{XY}, G_{\tilde{X}\tilde{Y}}, X, Y)\end{aligned}$$

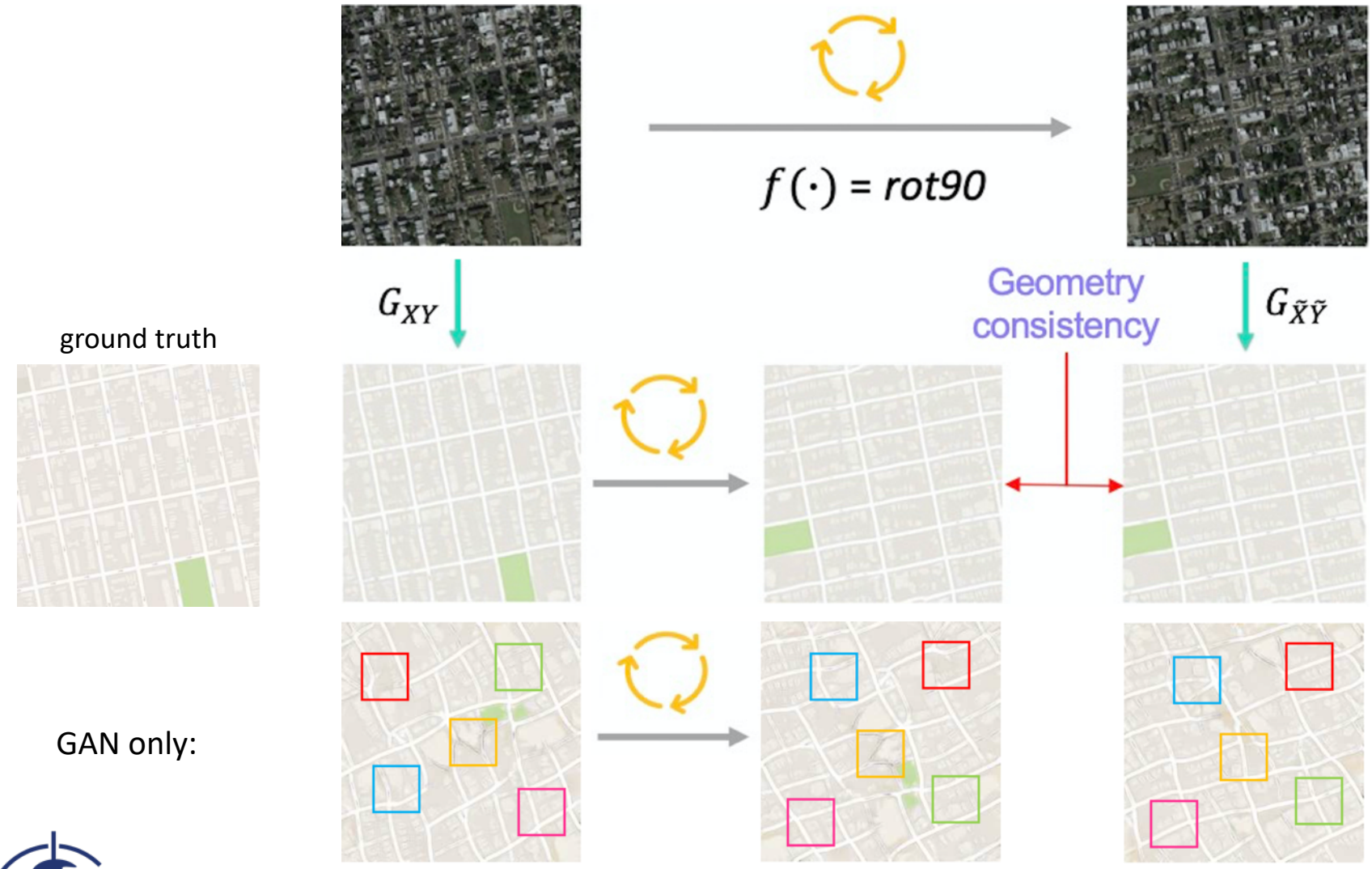
- May need to tune λ for the specific task
- G_{XY} and $G_{\tilde{X}\tilde{Y}}$ have the same architecture and **share** all the parameters!
- **One sided**: G_{XY} can be trained independently from G_{YX}

- 2 geometric transformations:
  - vertical flipping (vf)
  - 90° clockwise rotation (rot)

Proposed Method: GcGAN



Proposed Method: GcGAN

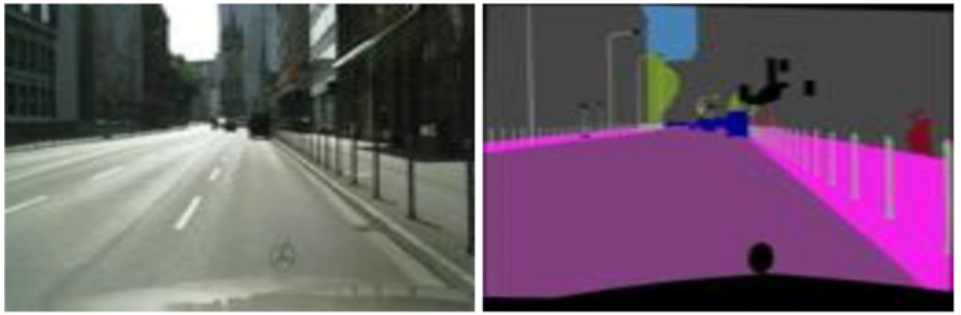


Experiments: Cityscapes

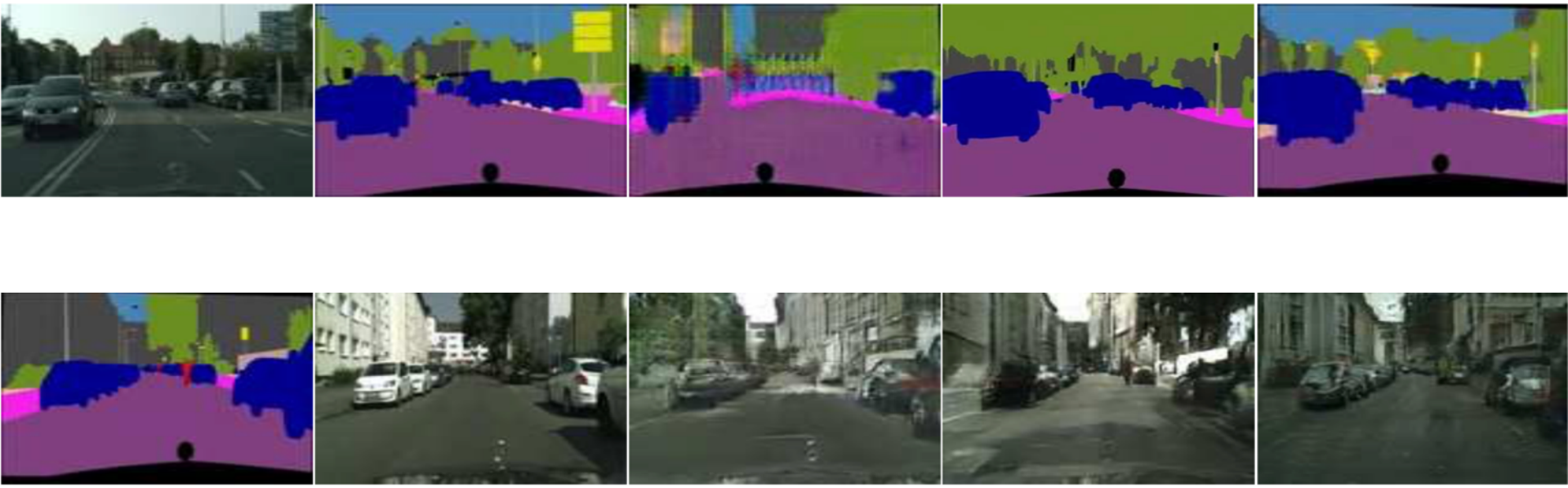
- Cityscapes
  - 3975 image – label pairs: 2975 training, 500 validation (used for testing)
  - 19 + 1 label classes

- parsing → image
  - predict labels from generated images (using FCN-8s)

- image → parsing
  - convert generated labels to class-level labels using nearest neighbor search



Experiments: Cityscapes



Input | Ground Truth | GAN alone | CycleGAN | GcGAN

Experiments: Cityscapes

method	image → parsing			parsing → image		
	pixel acc	class acc	mean IoU	pixel acc	class acc	mean IoU
Benchmark Performance						
CoGAN [40]	0.45	0.11	0.08	0.40	0.10	0.06
BiGAN/ALI [15, 16]	0.41	0.13	0.07	0.19	0.06	0.02
SimGAN [54]	0.47	0.11	0.07	0.20	0.10	0.04
CycleGAN (Cycle) [66]	**0.58**	0.22	0.16	0.52	0.17	0.11
DistanceGAN [5]	-	-	-	0.53	0.19	0.11
GAN alone (baseline)	0.514	0.160	0.104	0.437	0.161	0.098
GcGAN-*rot*	0.574	**0.234**	0.170	**0.551**	**0.197**	**0.129**
GcGAN-*vf*	0.576	0.232	**0.171**	0.548	0.196	0.127

Experiments: Cityscapes

method	image → parsing			parsing → image		
	pixel acc	class acc	mean IoU	pixel acc	class acc	mean IoU
Benchmark Performance						
GAN alone (baseline)	0.514	0.160	0.104	0.437	0.161	0.098
GcGAN-*rot*	0.574	**0.234**	0.170	**0.551**	**0.197**	**0.129**
GcGAN-*vf*	0.576	0.232	**0.171**	0.548	0.196	0.127
Ablation Studies (Robustness & Compatibility)						
$\mathcal{L}_{GcGAN}$ w/o $\mathcal{L}_{geo}$ ($\lambda = 0$)	0.486	0.163	0.102	0.396	0.148	0.088
$\mathcal{L}_{GcGAN}$ w/o $\mathcal{L}_{gan}(\tilde{X}, \tilde{Y})$	0.549	0.199	0.139	0.526	0.184	0.111
GcGAN-*rot*-Seperate	0.575	0.233	0.170	0.545	0.196	0.124
GcGAN-Mix-comb	0.573	0.229	0.168	0.545	0.197	0.128
GcGAN-Mix-rand	0.564	0.217	0.156	0.547	0.192	0.123
GcGAN-*rot* + Cycle	**0.587**	**0.246**	**0.182**	**0.557**	**0.201**	**0.132**

Experiments: SVHN → MNIST

- SVHN: Street View House Numbers
  - 73257 training images
- MNIST: Handwritten Digits
  - 60000 training images
- classify generated images using pretrained network

DistanceGAN

GcGAN

Experiments: SVHN → MNIST

method	class acc (%)
Benchmark Performance	
DistanceGAN (Dist.) [5]	26.8
CycleGAN (Cycle) [66]	26.1
Self-Distance [5]	25.2
GcGAN-*rot*	32.5
GcGAN-*vf*	**33.3**
Ablation Studies (Compatibility)	
Cycle + Dist. [5]	18.0
GcGAN-*rot* + Dist.	**34.0**
GcGAN-*rot* + Cycle	33.8
GcGAN-*rot* + Dist. + Cycle	33.2

Experiments: Google Maps

- Pairs of aerial photo and maps
- 2194 pairs: 1096 training, 1098 test

- Trained in unsupervised manner

- no quantitative results for map → aerial photo

Input | Ground Truth | GAN alone | CycleGAN | GcGAN

Experiments: Google Maps

- **aerial photo → map**

$\delta_1 = 5$ $\delta_2 = 10$

method	RMSE	acc (δ_1)	acc (δ_2)
Benchmark Performance			
CycleGAN [66]	**28.15**	**41.8**	**63.7**
GAN alone (baseline)	33.27	19.3	42.0
GcGAN-*rot*	28.31	41.2	63.1
GcGAN-*vf*	28.50	37.3	58.9
Ablation Studies (Robustness & Compatibility)			
GcGAN-*rot*-Separate	30.25	40.7	60.8
GcGAN-Mix-comb	**27.98**	**42.8**	**64.6**
GcGAN-*rot* + Cycle	28.21	40.6	63.5

$$max(|r_i - r'_i|, |g_i - g'_i|, |b_i - b'_i|) < \delta$$

Experiments: Qualitative Results

object transfiguration (Horse → Zebra)

Monet paintings to photos

Input CycleGAN GcGAN

Experiments: Qualitative Results

Experiments: Qualitative Results

Night → Day

Input GcGAN Input GcGAN

Photographs Monet Cezanne Van Gogh

Conclusion

- Geometry consistency improves traning of GAN alone
  - solves mode collapse problem
- Reduces semantic distortions
- Competitive performance with state of the art
- Compatible with other constraints

- Requires predefined geometric transformation $f(\cdot)$
- May need to choose $f(\cdot)$ and λ according to the task
- No comparison with variations of CycleGAN

Thank you for listening

Network

- 256x256 input
- C: Feature channel, K: Kernel size, S: Stride

- SVHN → MNIST
  - smaller network
  - no residual block

- identity mapping loss
  - generator to be near an identity mapping when real examples are provided
  - more conservative for unknown content

Generator				
Index	Layer	C	K	S
1	Conv + ReLU	64	7	1
2	Conv + ReLU	128	3	2
3	Conv + ReLU	256	3	2
4	ResBlk + ReLU	256	3	1
5	ResBlk + ReLU	256	3	1
6	ResBlk + ReLU	256	3	1
7	ResBlk + ReLU	256	3	1
8	ResBlk + ReLU	256	3	1
9	ResBlk + ReLU	256	3	1
10	ResBlk + ReLU	256	3	1
11	ResBlk + ReLU	256	3	1
12	ResBlk + ReLU	256	3	1
12	Deconv + ReLU	128	3	2
13	Deconv + ReLU	64	3	2
14	Conv	3	7	1
15	Tanh	-	-	-
Discriminator				
1	Conv + LeakyReLU	64	4	2
2	Conv + LeakyReLU	128	4	2
3	Conv + LeakyReLU	256	4	2
4	Conv + LeakyReLU	512	4	1
5	Conv	512	4	1

Experiment Details

- Cityscapes
  - translators are trained with 128x128 images in unaligned fashion
  - a high-quality translated image should produce qualitative semantic segmentation like real images when feeding it into a scene parser.
- Horse → Zebra
  - ImageNet: wild horse, zebra
  - no parameter sharing
- Summer – Winter
  - Images are provided by CycleGAN

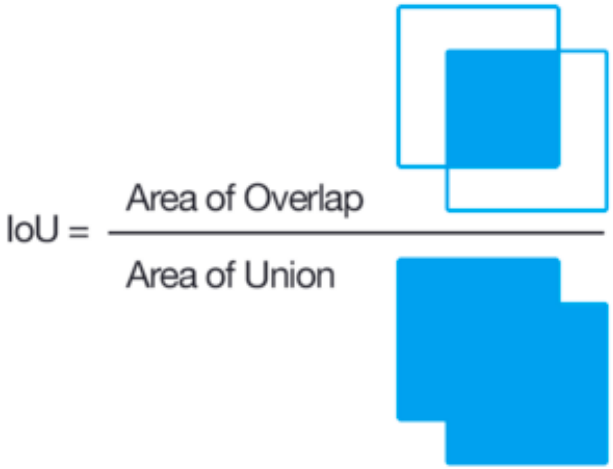
Metrics

- **Pixel Accuracy**
  - Percent of correctly classified pixels
  - Does not work well with imbalanced data!

- Class Accuracy

- mean IoU
  - Intersection over Union
  - average of IoU for all classes

- RMSE
  - root mean square error



Source: Wikipedia