Image Domain Transfer

Jan Kautz, VP of Learning and Perception Research

Image Domain Transfer: enabling machines to have human-like imagination abilities

$\longrightarrow F \longrightarrow$

Input image

Domain transferred image

This image is generated by our method.

NVIDIA

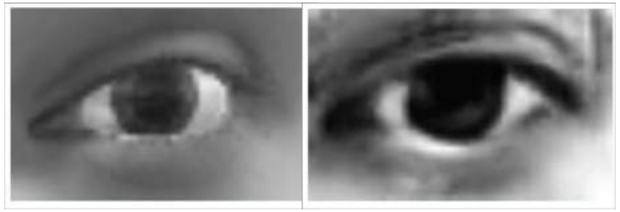
Example use cases

Low-res to high-res

Blurry to sharp

Image to painting

LDR to HDR



Synthetic to real

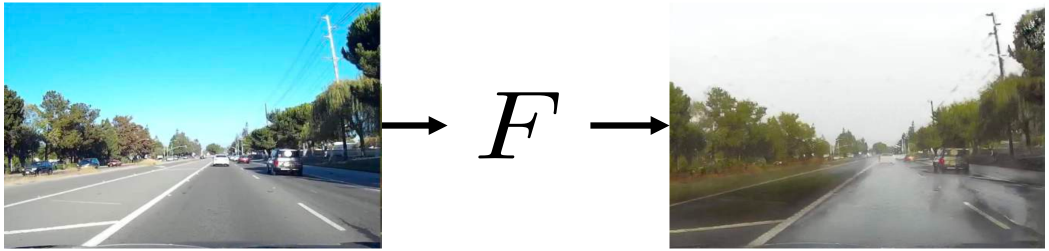
Thermal to color

Day to night

Summer to winter

Noisy to clean

Two Approaches



Input image

Domain transferred image

- Example-based
  - Non-parametric model
  - The transfer function F is defined by an example image.

$$F(\quad | x_{\text{example}})$$

- Learning-based
  - Parametric model
  - The transfer function is learned via fitting a training dataset.

$$F(\quad | \text{Training Dataset})$$

Example-based Image Domain Transfer

$$F(\quad | x_{\text{example}})$$

Example-based image domain transfer

Often referred to as **Style Transfer**

Style photo

Content photo

Stylized content photo

$$x_{\text{output}} = F(x_{\text{input}} | x_{\text{example}})$$

Example-based image domain transfer

- **Artistic** Style Transfer
    - Content: real photo;
    - Style: painting
    - *Gatys et. al. , Johnson et. al., Li et al., Huang et. al.*
- **Photo** Style Transfer
    - Content: real photo;
    - Style: real photo
    - *Luan et. al., Pitie et. al., Reinhard et. al.*


Style (painting)


Content


Output


Style


Content


Output

FastPhotoStyle

- *”A Closed-form Solution to Photorealistic Image Stylization”* by Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, Jan Kautz, ECCV 2018
- Code: https://github.com/NVIDIA/FastPhotoStyle





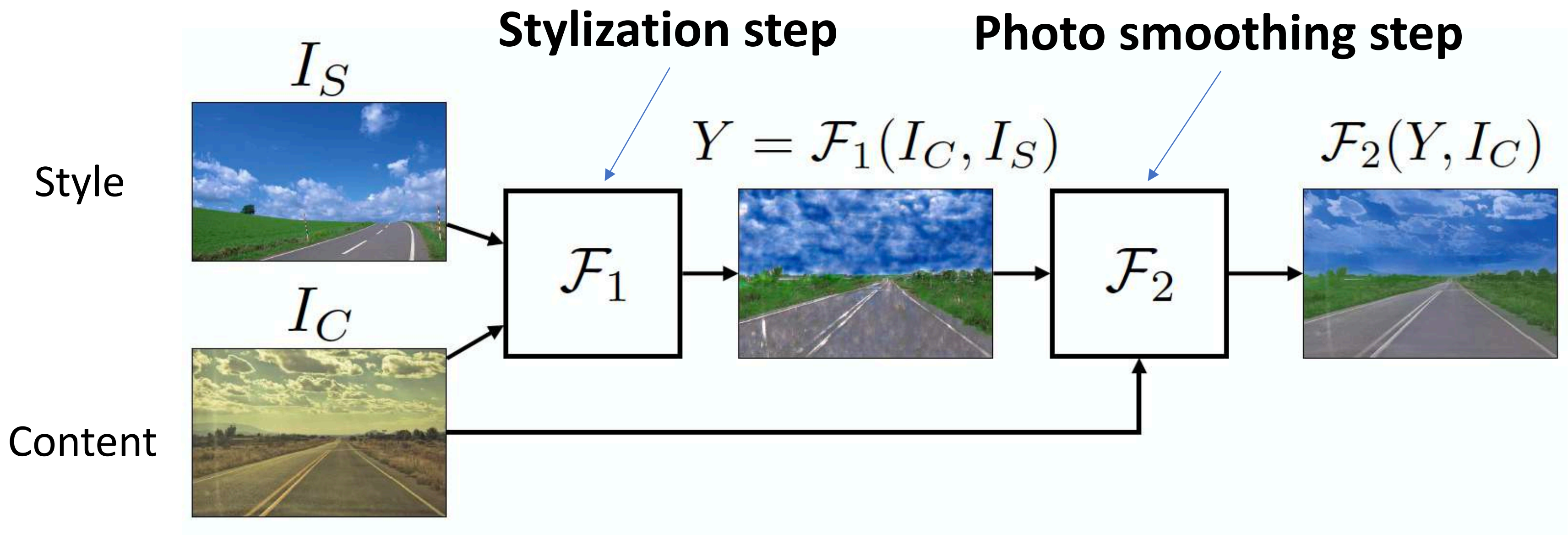
Fast Photo Style

We model photo style transfer as a close-form function mapping given by

Content image

$$\mathcal{F}_2\Big(\mathcal{F}_1(I_C, I_S), I_C\Big).$$

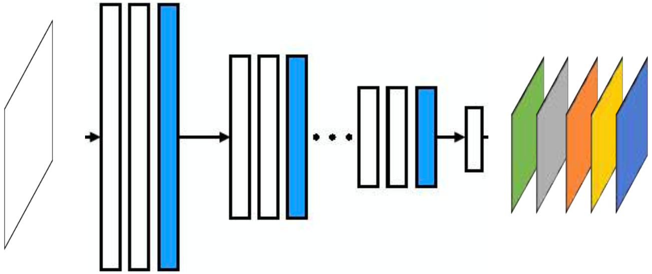
Style image

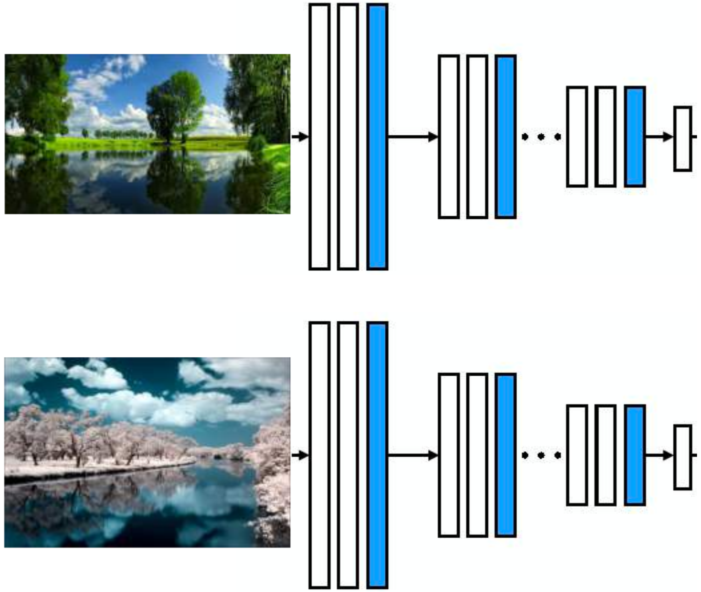


Stylization Step

Assumption: Covariance matrix of deep features encodes the style information.

VGG19 up to conv4_1

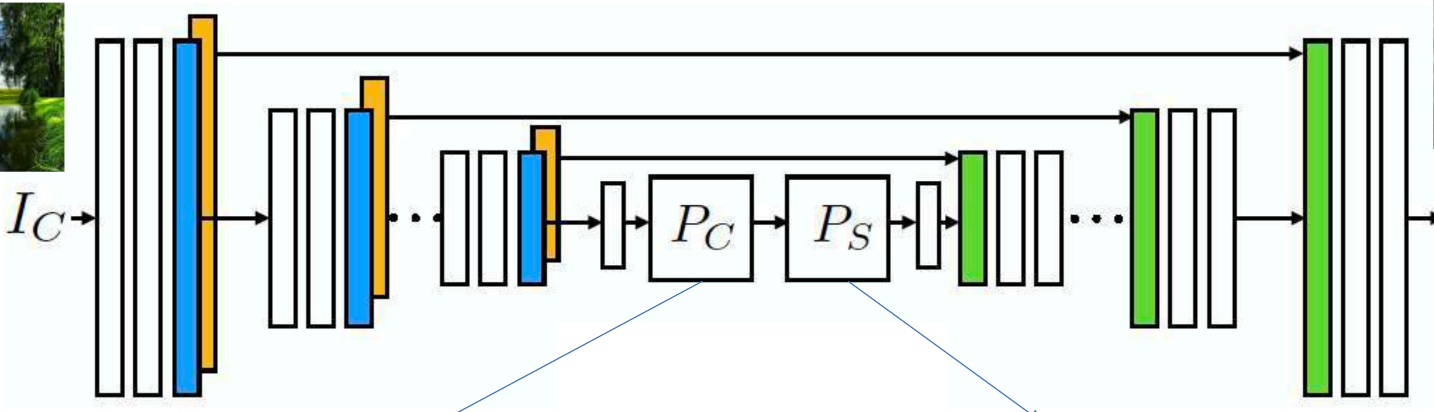




$$H_C H_C^T = E_C \Lambda_C E_C^T$$

$$H_S H_S^T = E_S \Lambda_S E_S^T$$

Stylization Step



Architecture is similar to *Li et al.*, but uses unpooling.

$$P_C = E_C \Lambda_C^{-\frac{1}{2}} E_C^\top$$

whitening

$$P_S = E_S \Lambda_S^{\frac{1}{2}} E_S^\top$$

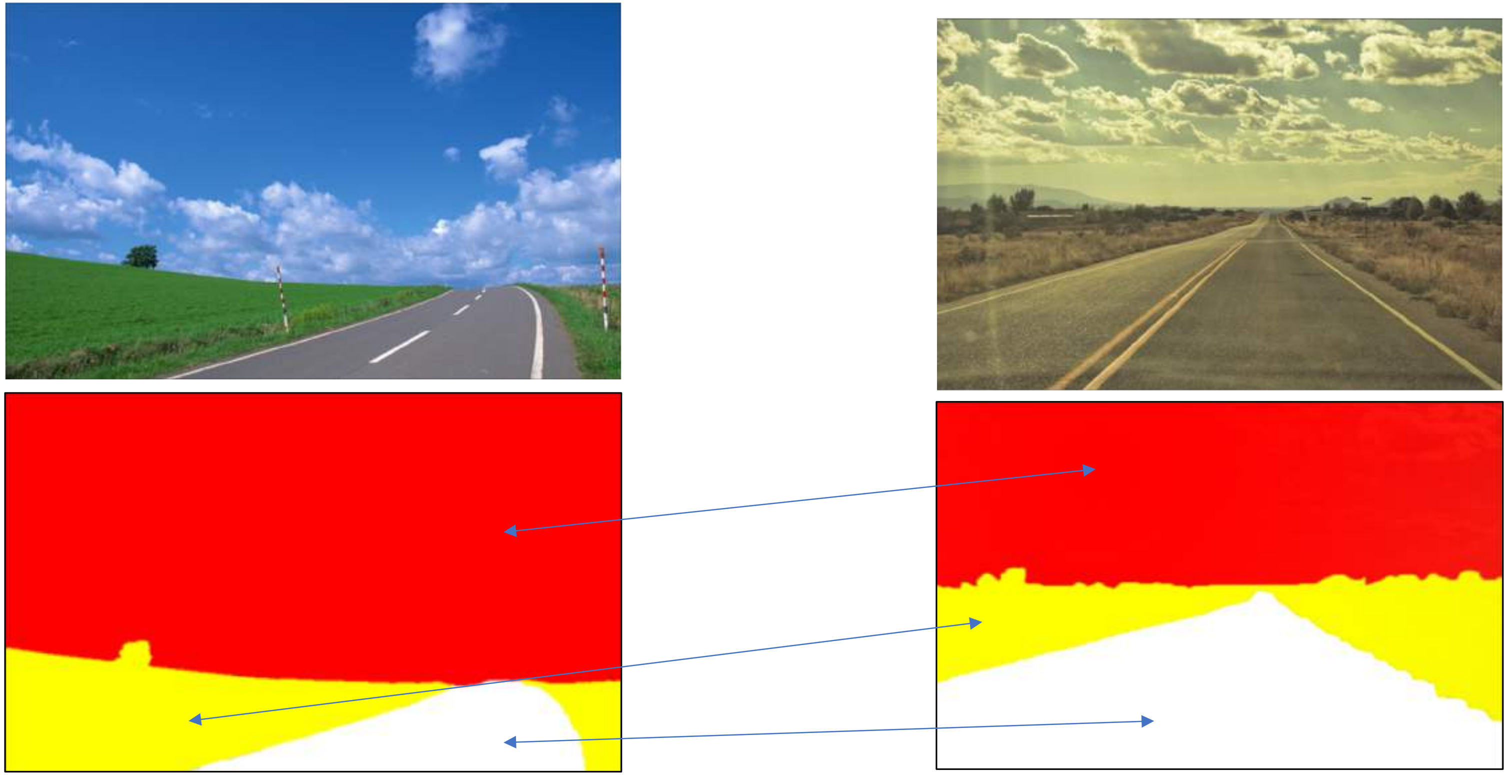
coloring

$$H_{CS} H_{CS}^\top = H_S H_S^\top$$

When semantic label maps are available

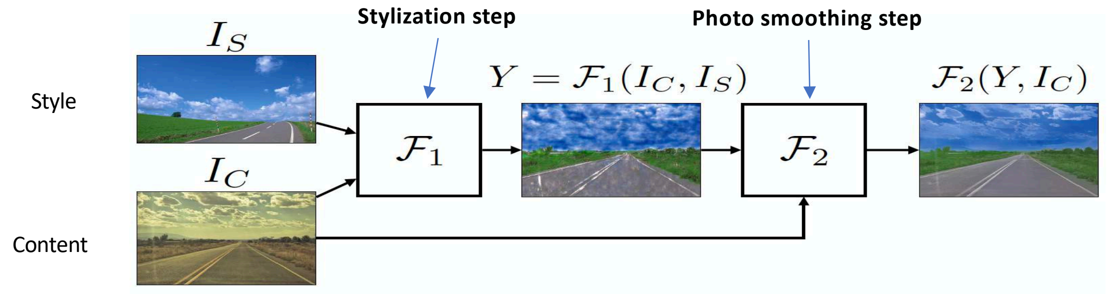
Content

Style



Style Content Output

Photo Smoothing



Assumption: If we can compute a new image where the **image pixel values resemble those in the intermediate image** but **the similarities between neighboring pixels resemble those in the content image**, then we have a photorealistic stylization outputs.

$$\operatorname*{argmin}_{r} \frac{1}{2}\left(\sum_{i,j=1}^{N} w_{ij} \left\| \frac{r_i}{\sqrt{d_{ii}}} - \frac{r_j}{\sqrt{d_{jj}}} \right\|^2 + \lambda \sum_{i=1}^{N} \| r_i - y_i \|^2\right),$$

Similarity between neighboring pixels (Gaussian/matting affinity)

Intermediate image pixel values

Close-form solution: $R^* = (1-\alpha)(I - \alpha D^{-\frac{1}{2}} W D^{-\frac{1}{2}})^{-1} Y$

Style Content Output

Style Content Output

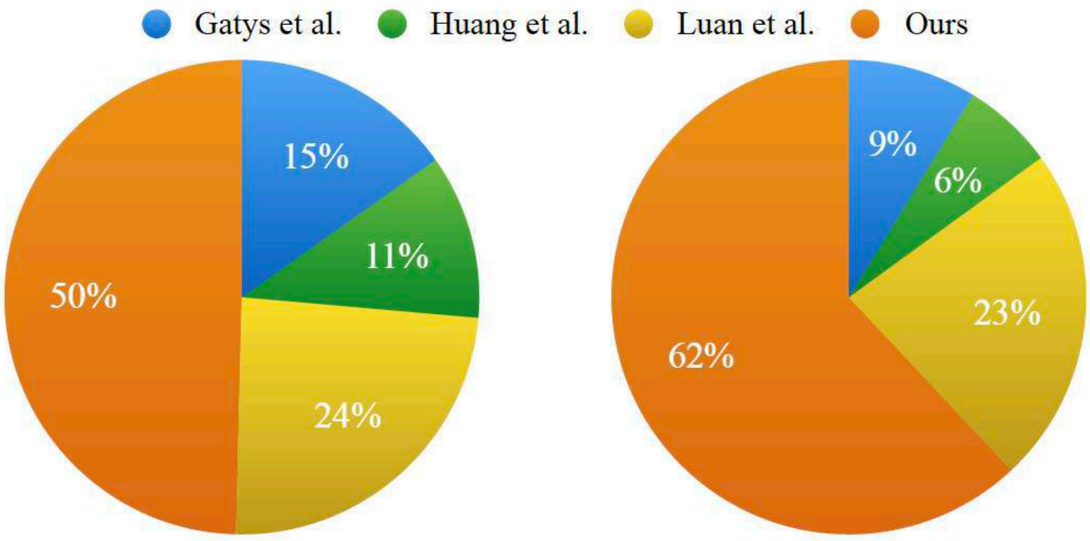
Comparison

(a) Content (b) Style (c) Gatys et al. [6] (d) Luan et al. [21] (e) Ours

Comparison

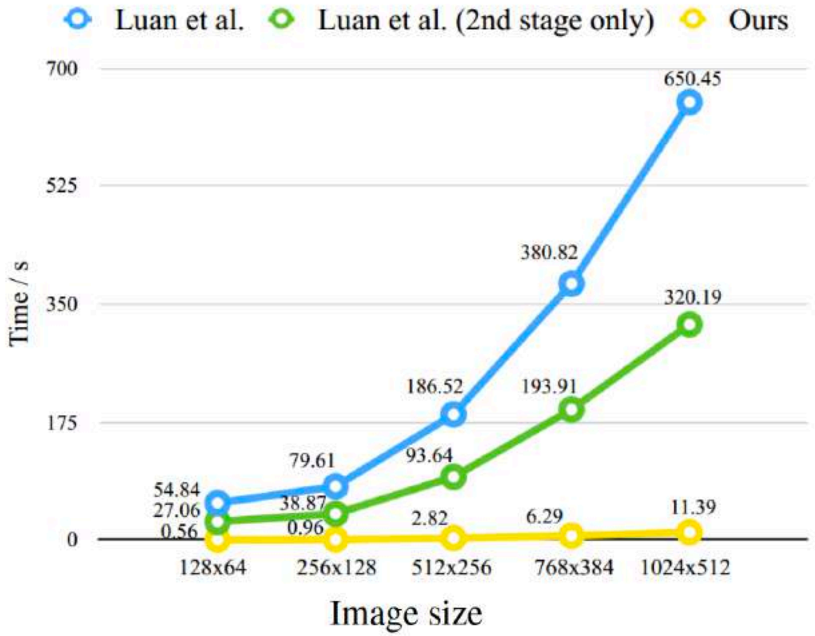
(a) Content (b) Style (c) Pitié et al. [24] (d) Luan et al. [21] (e) Ours

Quantitative results



(a) Stylization effects

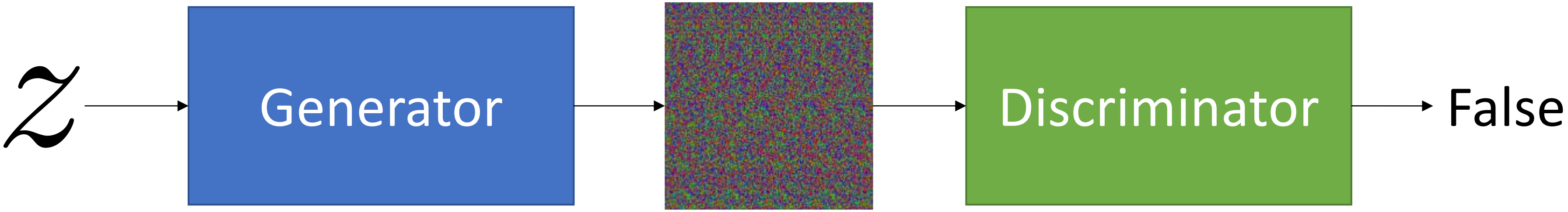
(b) Photorealism



Learning-based Image Domain Transfer

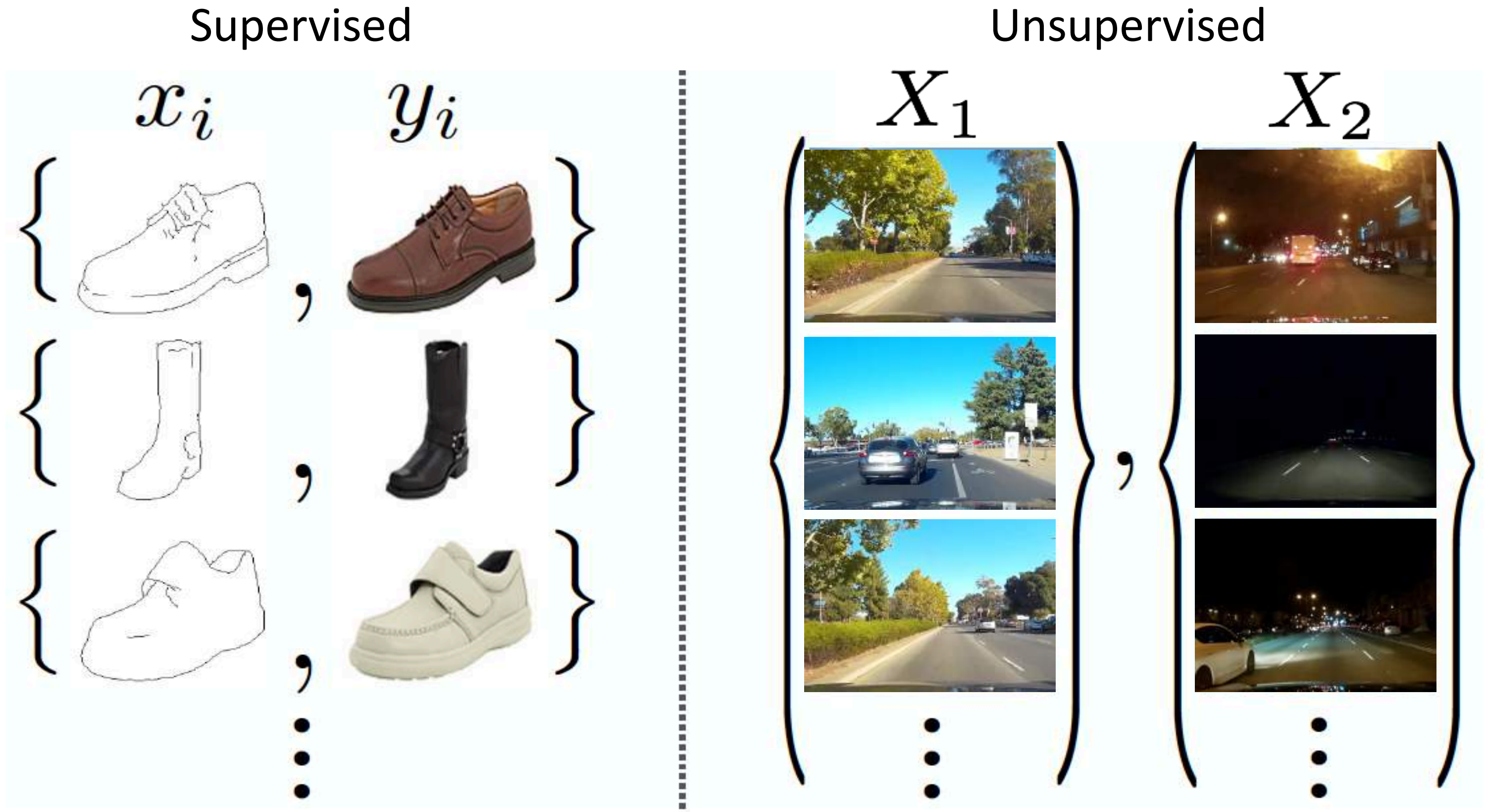
$$F(\quad|\text{Training Dataset})$$

Generative Adversarial Networks (GANs)



$$p(G(z)) \to p(X)$$

Supervised vs Unsupervised

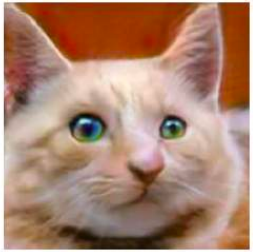


Unimodal vs Multimodal

Unimodal $p(Y|X) = \delta(F(X))$

$F($  $) =$ 

Multimodal $p(Y|X) = F(X, S)$

$F($  $) =$   

Categorization

	Supervised	Unsupervised
Unimodal	pix2pix, CRN, SRGAN	UNIT, Coupled GAN, DTN, DiscoGAN, CycleGAN, DualGAN, StarGAN
Multimodal	pix2pixHD, vid2vid, BiCycleGAN	MUNIT

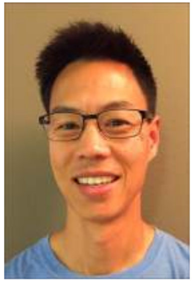
Categorization

	Supervised	Unsupervised
Unimodal	pix2pix, CRN, SRGAN	UNIT, Coupled GAN, DTN, DiscoGAN, CycleGAN, DualGAN, StarGAN
Multimodal	pix2pixHD, vid2vid, BiCycleGAN	MUNIT

pix2pixHD: *Supervised* and *multimodal* image domain transfer

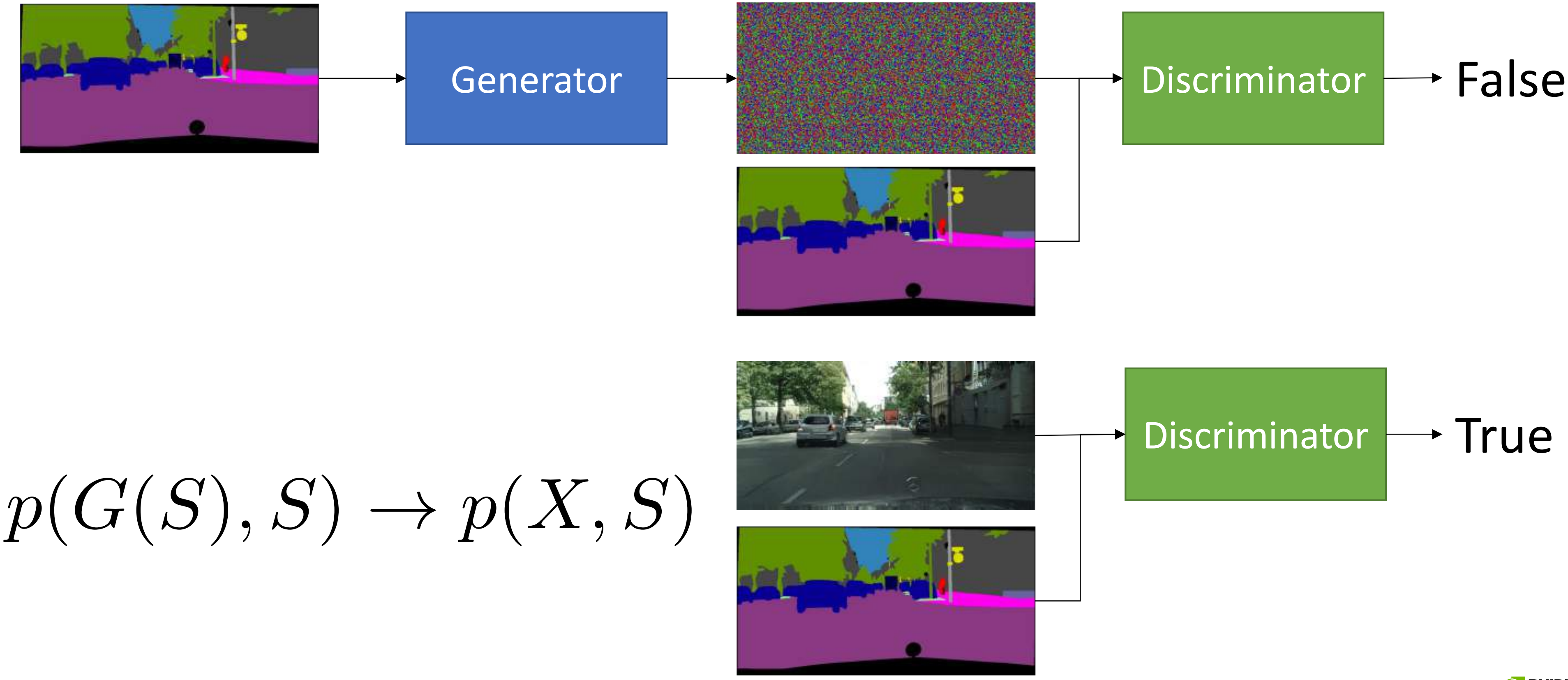
- ”*High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs*” by
  Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, Bryan Catanzaro, CVPR 2018
- Code: https://github.com/NVIDIA/pix2pixHD



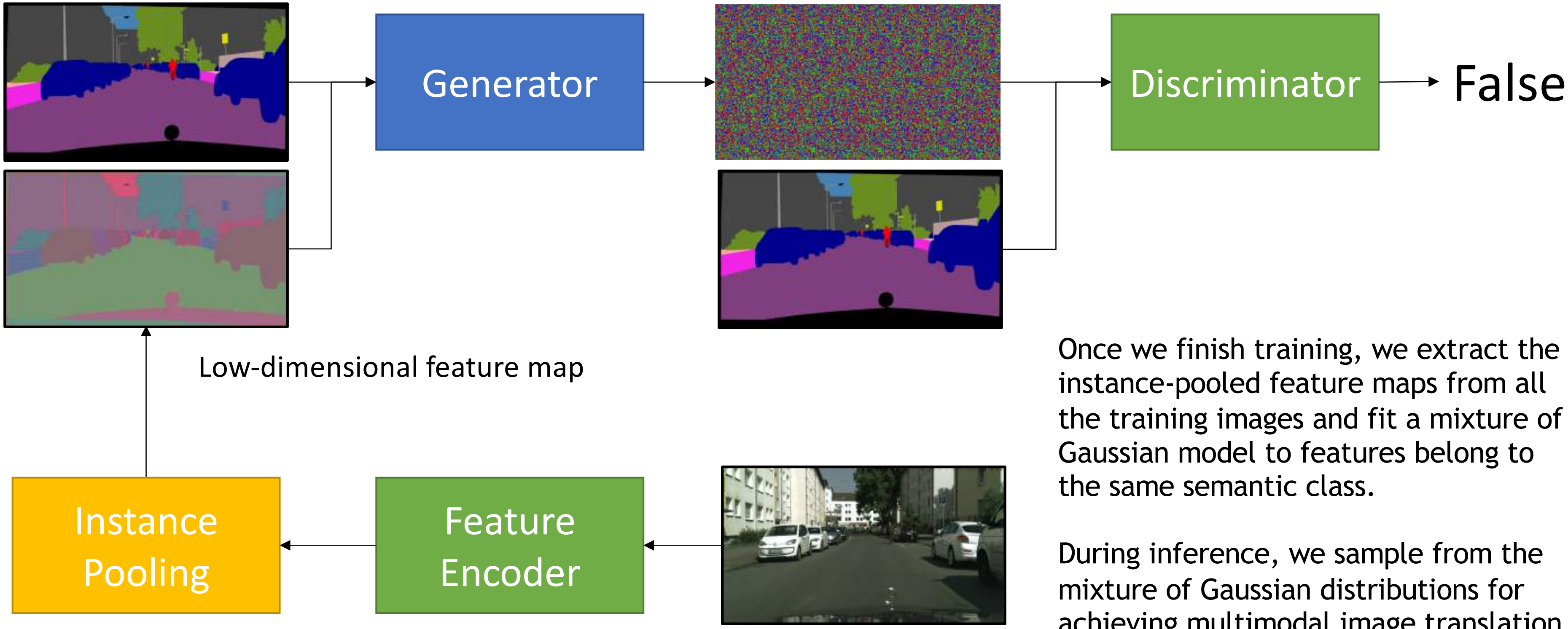
pix2pixHD

pix2pix



$$p(G(S), S) \rightarrow p(X, S)$$

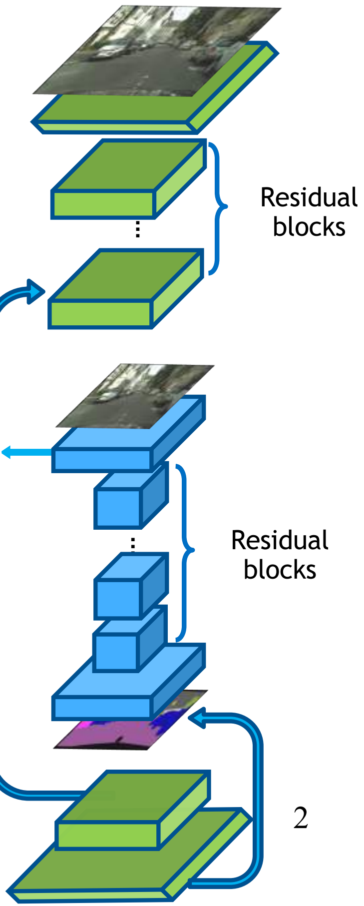
pix2pixHD



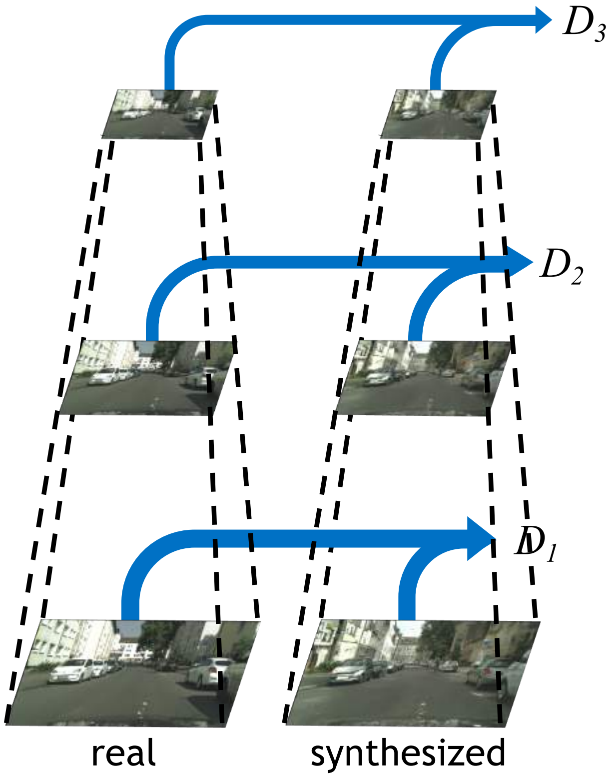
Once we finish training, we extract the instance-pooled feature maps from all the training images and fit a mixture of Gaussian model to features belong to the same semantic class.

During inference, we sample from the mixture of Gaussian distributions for achieving multimodal image translation.

Coarse-to-fine Generator

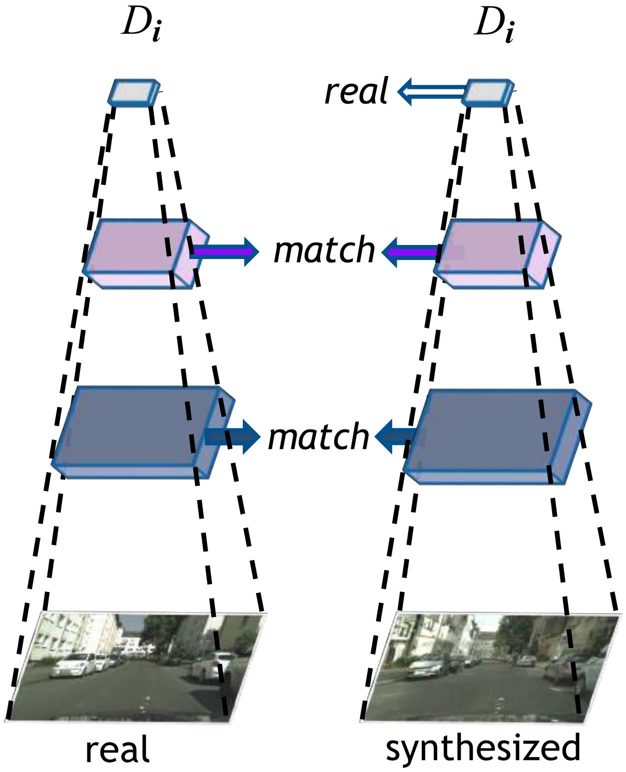


Multi-scale Discriminators

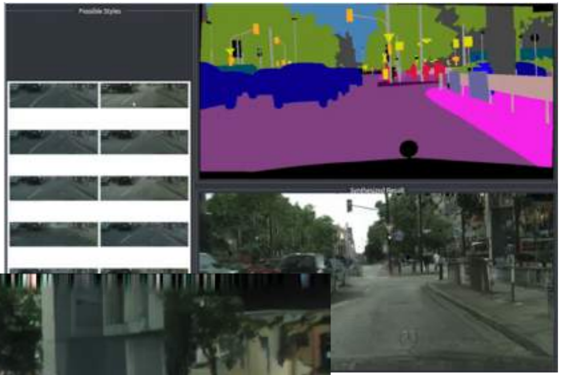


Robust Objective

(GAN + discriminator feature matching loss)

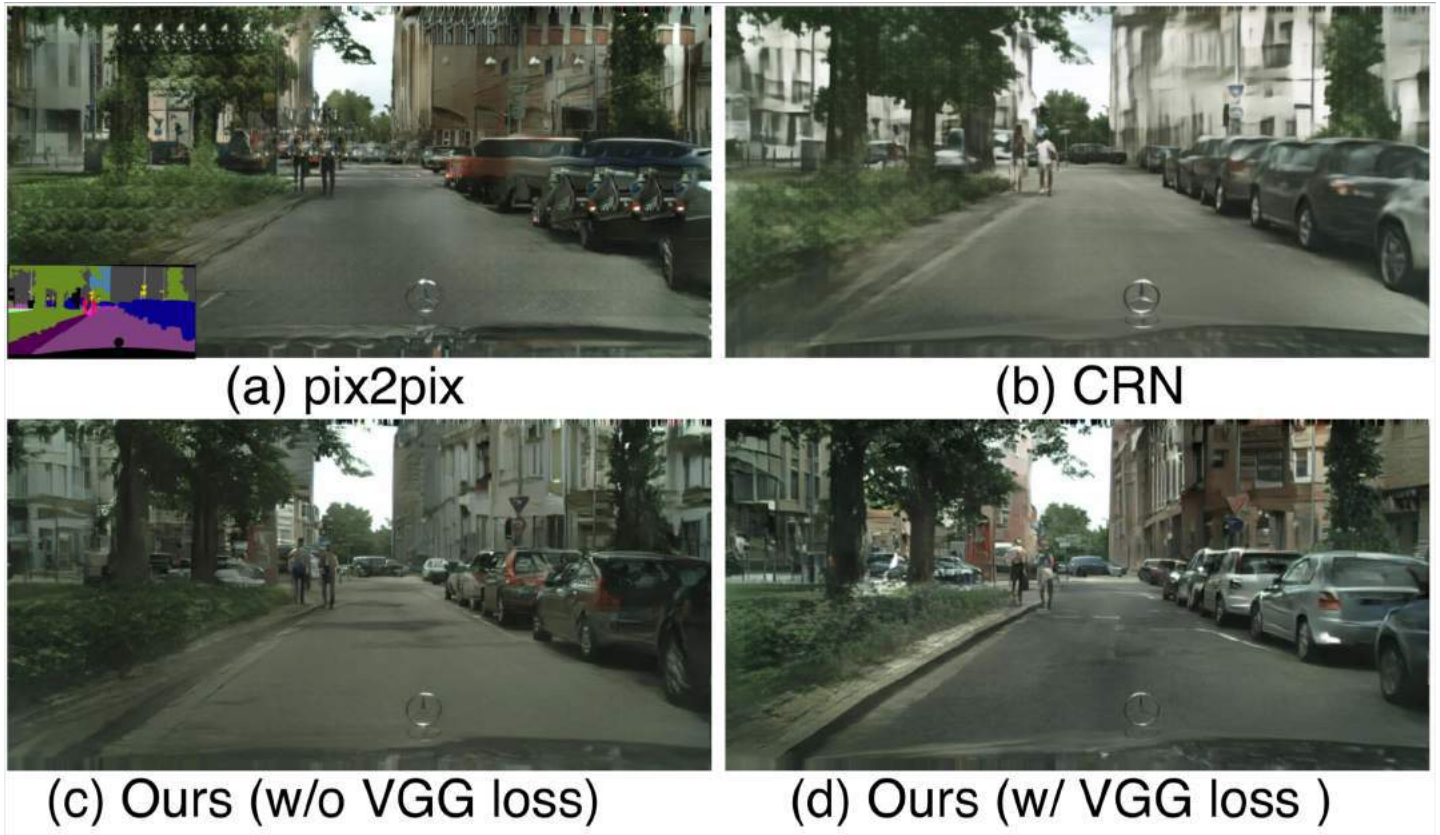


pix2pixHD multimodal results



pix2pixHD label changes

Comparison



Categorization

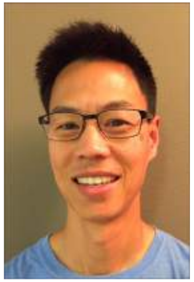
	Supervised	Unsupervised
Unimodal	pix2pix, CRN, SRGAN	UNIT, Coupled GAN, DTN, DiscoGAN, CycleGAN, DualGAN, StarGAN
Multimodal	pix2pixHD, vid2vid, BiCycleGAN	MUNIT

vid2vid: Video-to-Video Synthesis

- ”*Video-to-Video Synthesis*” by Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan Catanzaro, NIPS 2018
- Code: https://github.com/NVIDIA/vid2vid



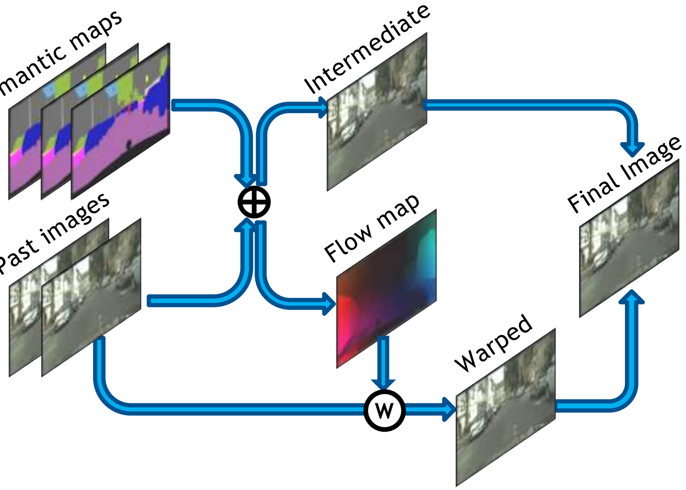




Motivation

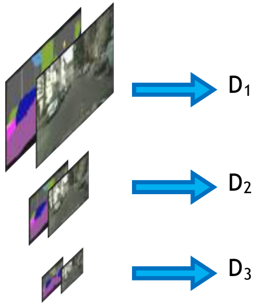
Using pix2pixHD

Sequential Generator

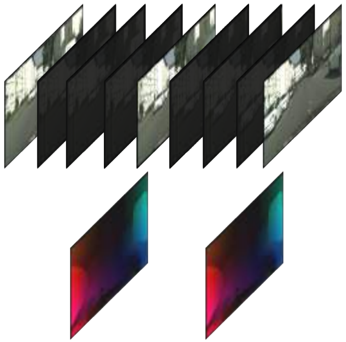


Multi-scale Discriminators

Image Discriminator

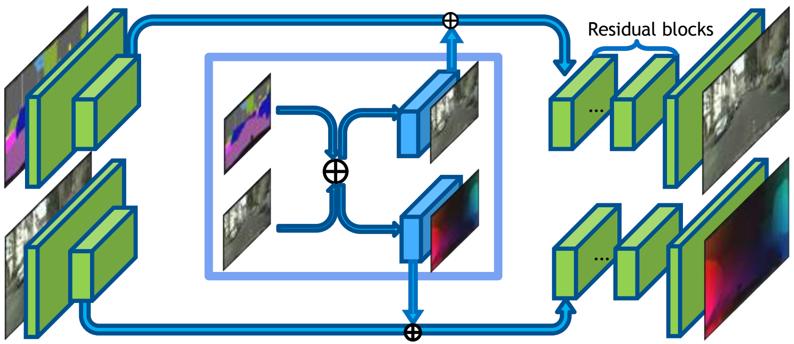


Video Discriminator



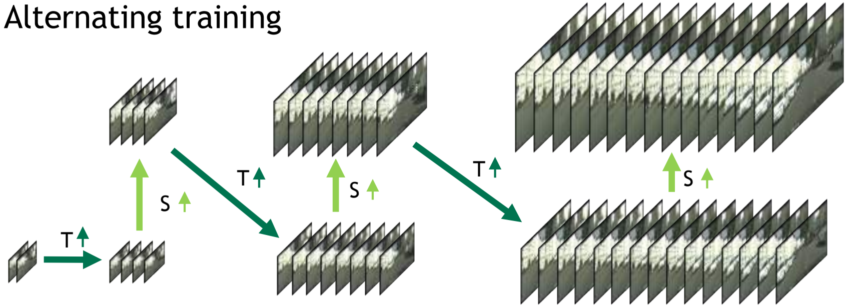
Spatio-temporally Progressive Training

Spatially progressive

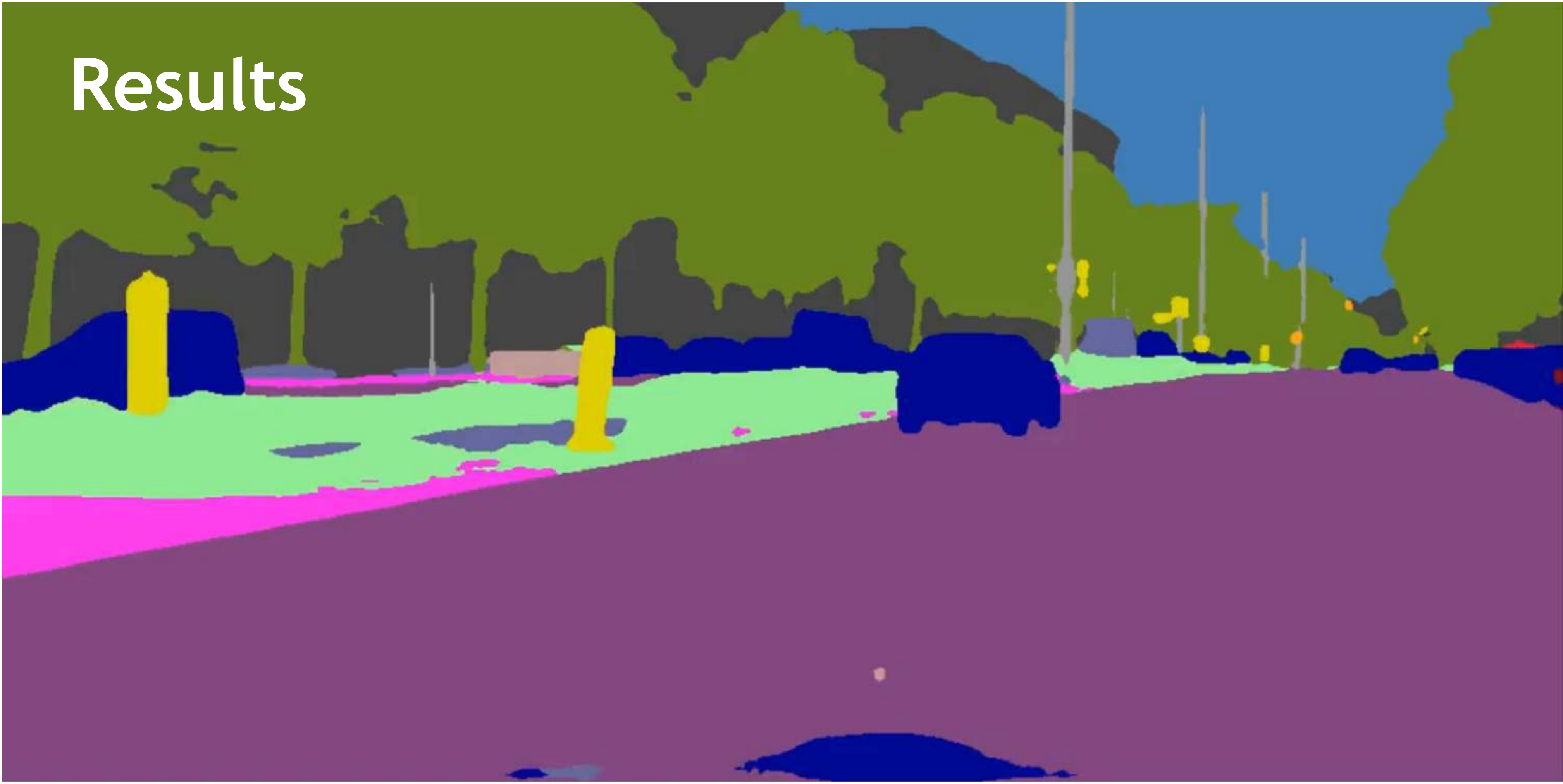


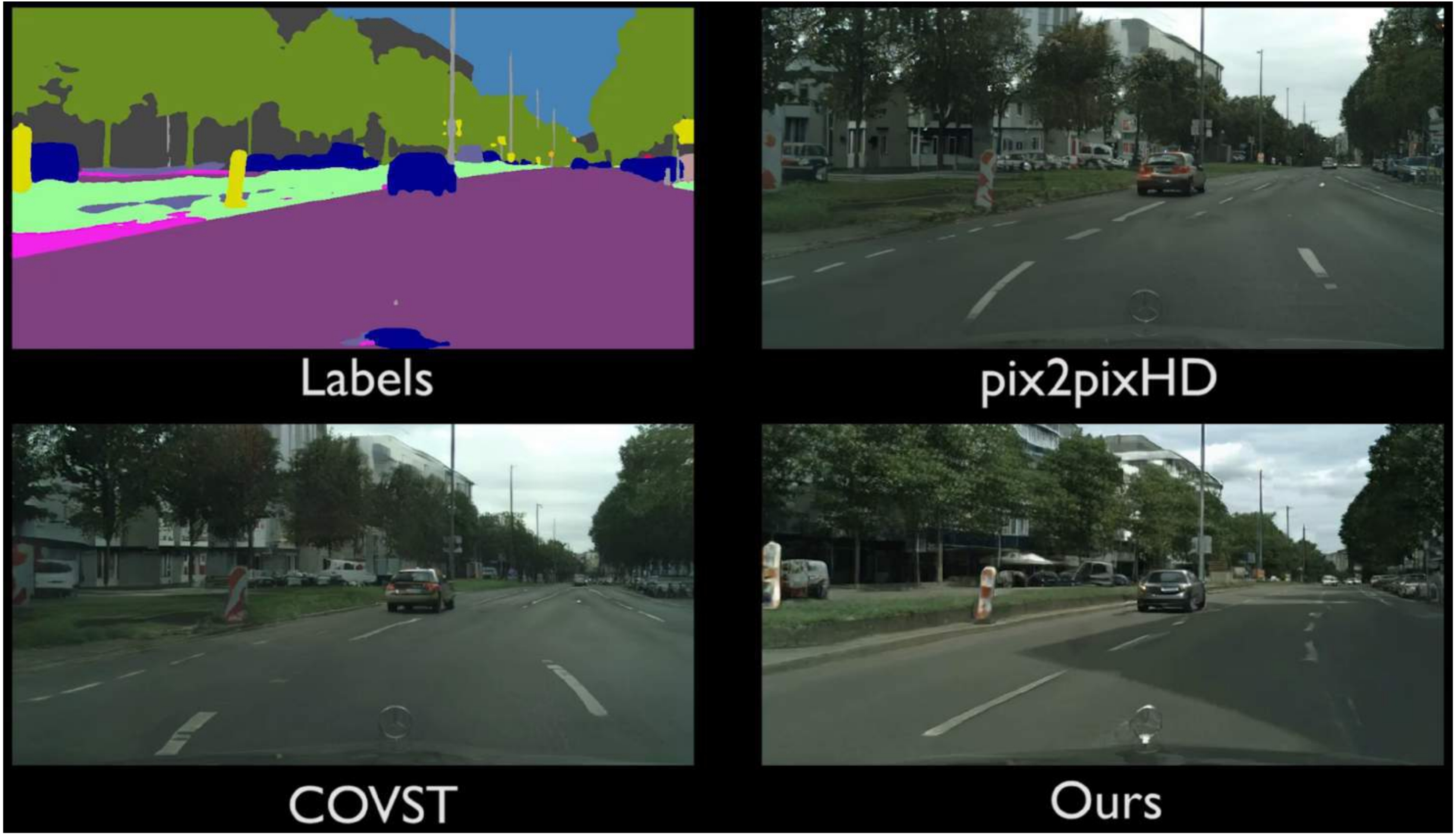
Temporally progressive

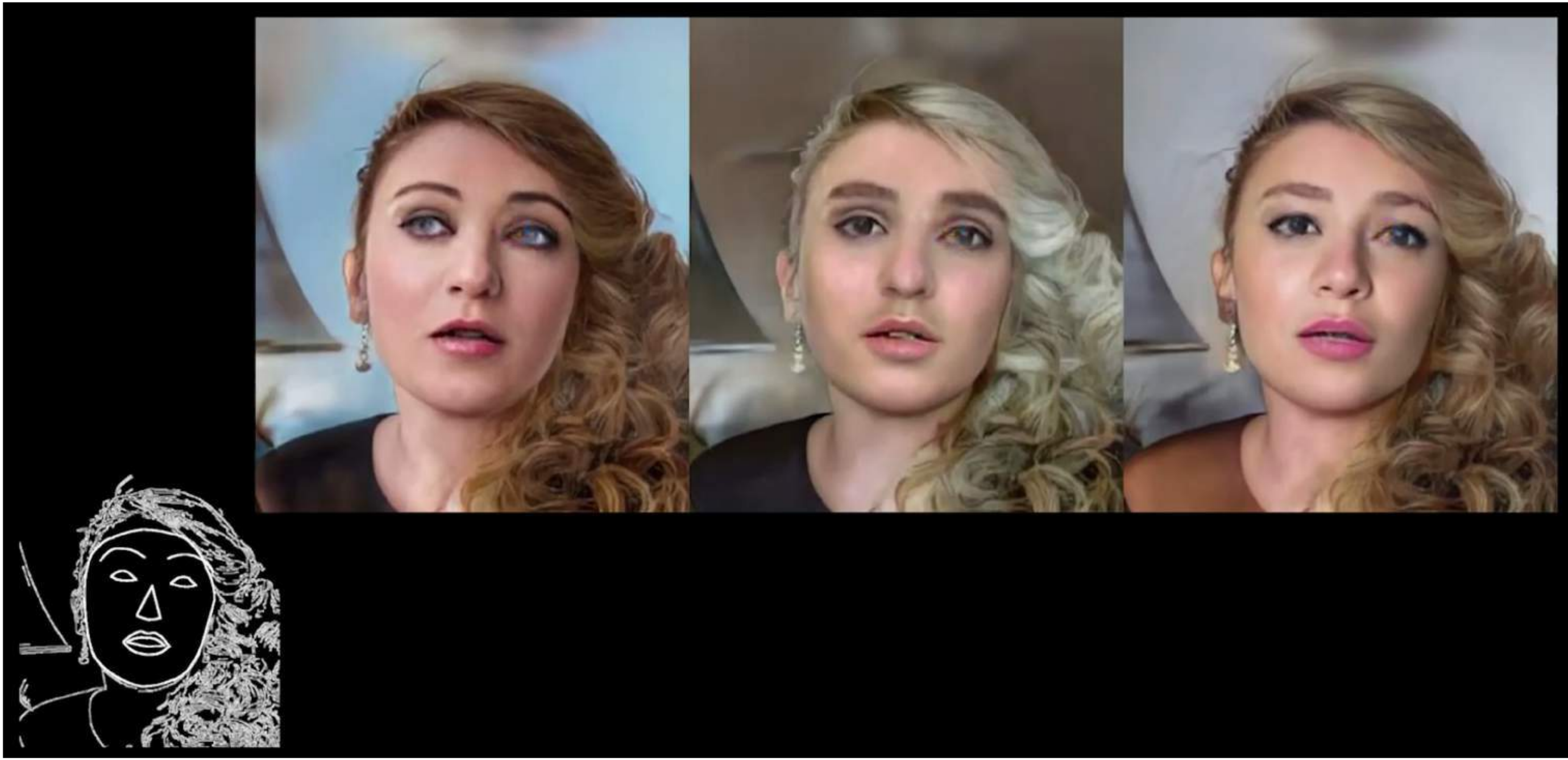
Alternating training



Results




Labels
pix2pixHD
COVST
Ours




AI Rendered Game

Categorization

	Supervised	Unsupervised
Unimodal	pix2pix, CRN, SRGAN, …	UNIT, Coupled GAN, DTN, DiscoGAN, CycleGAN, DualGAN, StarGAN
Multimodal	pix2pixHD, vid2vid, BiCycleGAN	MUNIT

UNIT: *Unsupervised* and *unimodal* image domain transfer

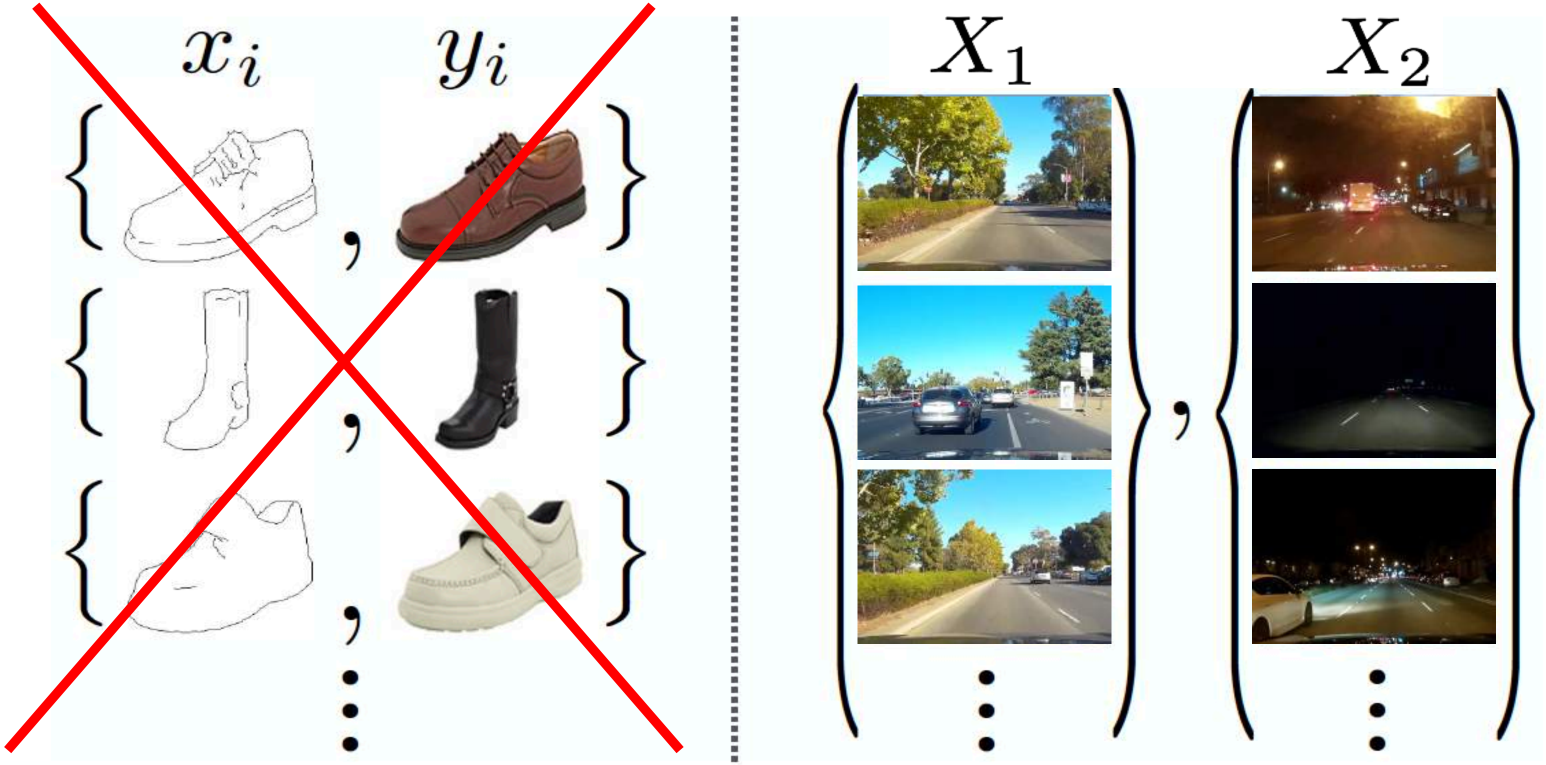
- ”*Unsupervised Image-to-image Translation Networks*” by Ming-Yu Liu, Thomas Breuel, Jan Kautz, NIPS 2017
- Code: https://github.com/mingyuliutw/UNIT

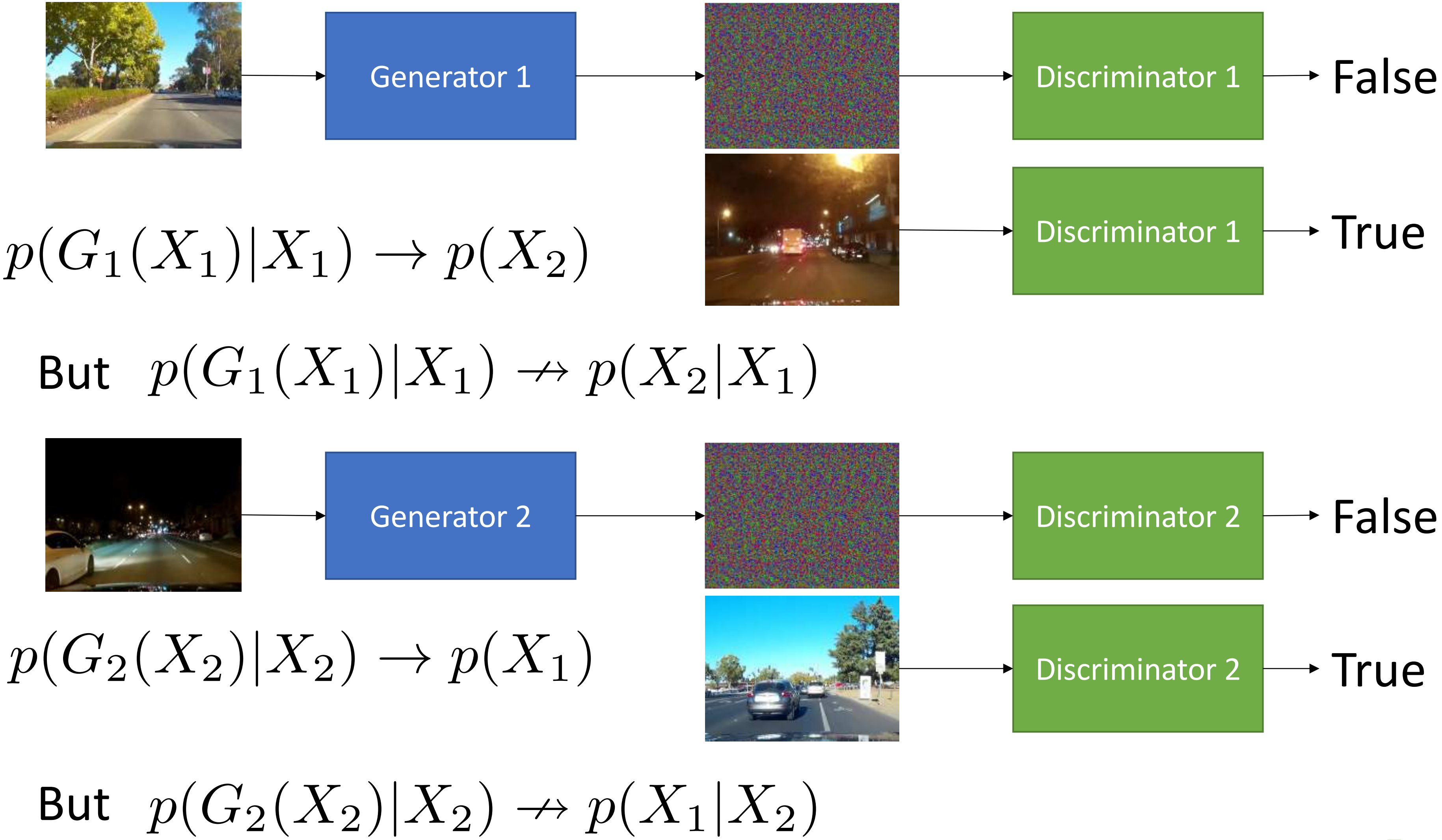


Supervised vs Unsupervised

Supervised

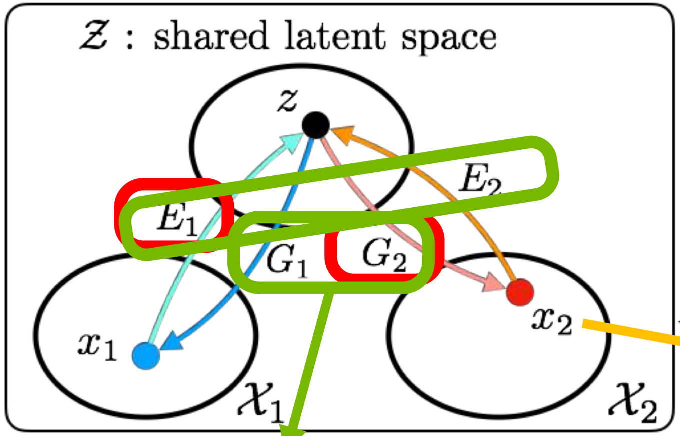
Unsupervised

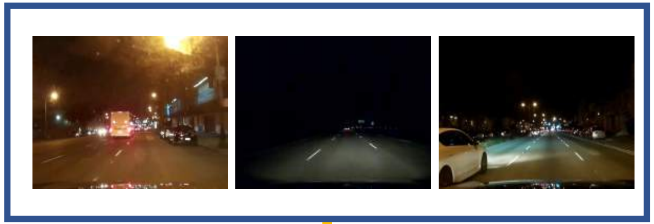




Generator 1
Discriminator 1
False
Discriminator 1
True
$p(G_1(X_1)|X_1) \to p(X_2)$
But $p(G_1(X_1)|X_1) \nrightarrow p(X_2|X_1)$
Generator 2
Discriminator 2
False
Discriminator 2
True
$p(G_2(X_2)|X_2) \to p(X_1)$
But $p(G_2(X_2)|X_2) \nrightarrow p(X_1|X_2)$

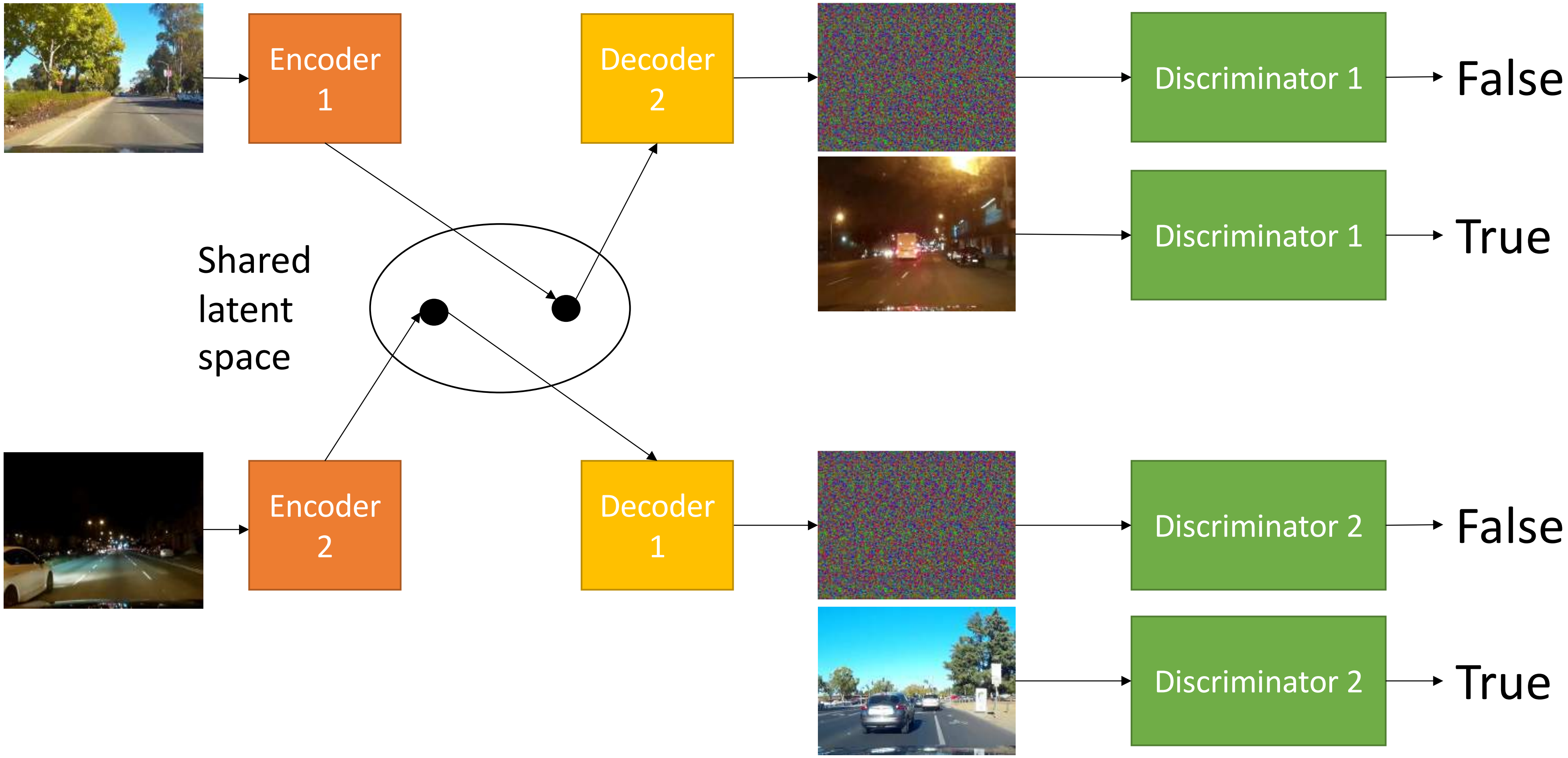
UNIT assumption: Shared Latent Space



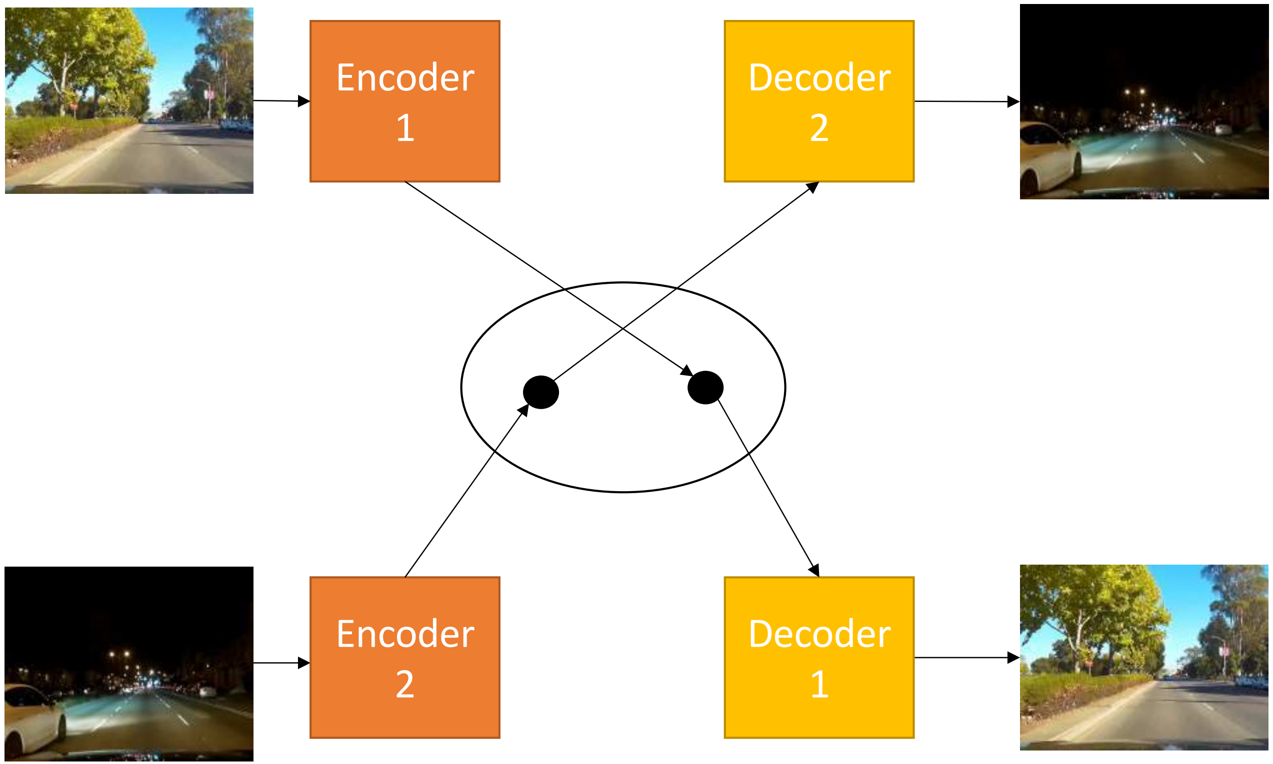


Domain 2 GAN Discriminator

Coupling the mapping function via weight-sharing


Encoder 1
Decoder 2
Discriminator 1
False
Discriminator 1
True
Shared latent space
Encoder 2
Decoder 1
Discriminator 2
False
Discriminator 2
True

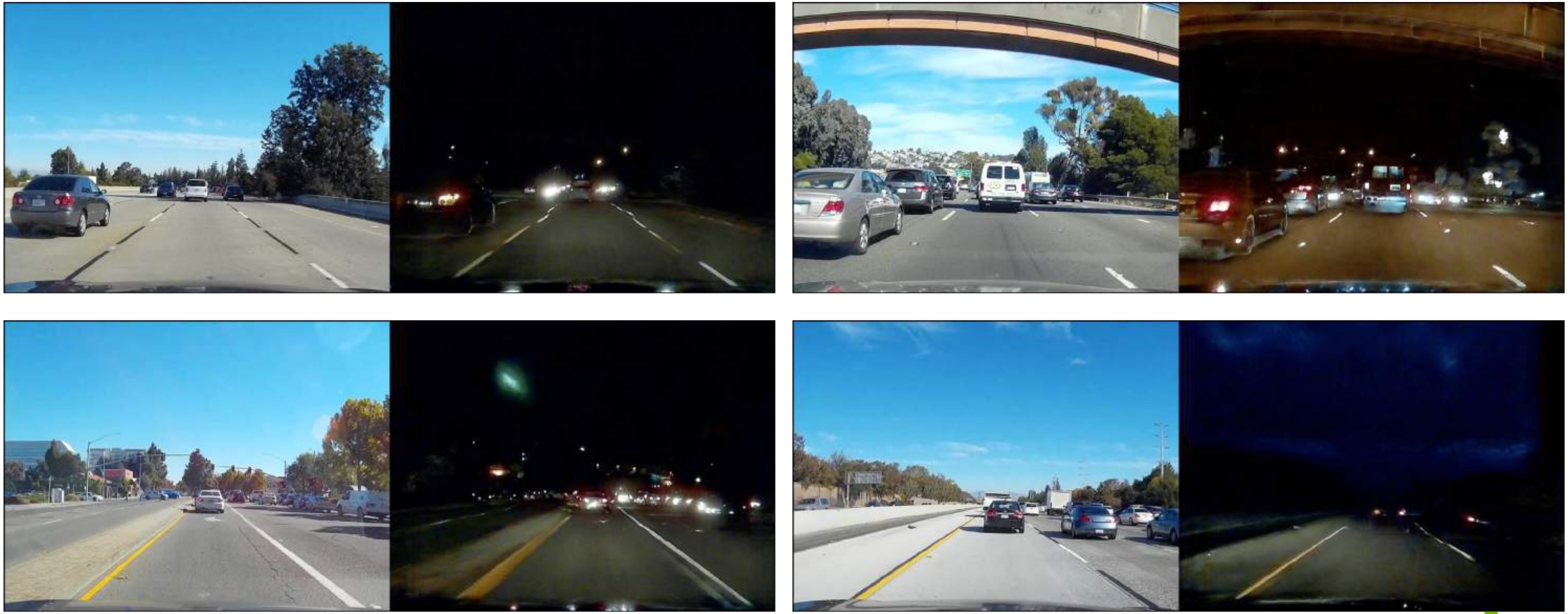
NVIDIA


Encoder 1
Decoder 2
Encoder 2
Decoder 1

Day to Night Translation

Resolution
640x480

Input | Translated | Input | Translated



Snowy to Summery Translation

Resolution
640x480

Input	Translated	Input	Translated

Sunny to Rainy Translation

Resolution
640x480

Input | Translated | Input | Translated

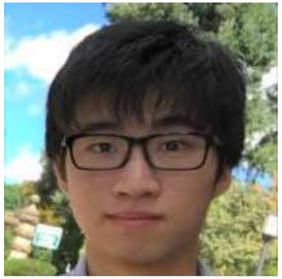



Categorization

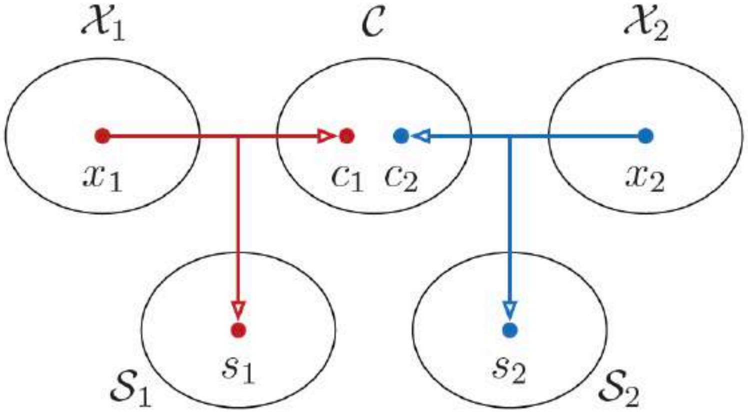
	Supervised	Unsupervised
Unimodal	pix2pix, CRN, SRGAN	UNIT, Coupled GAN, DTN, DiscoGAN, CycleGAN, DualGAN, StarGAN
Multimodal	pix2pixHD, BiCycleGAN	MUNIT

MUNIT: *Unsupervised* and *multimodal* image domain transfer

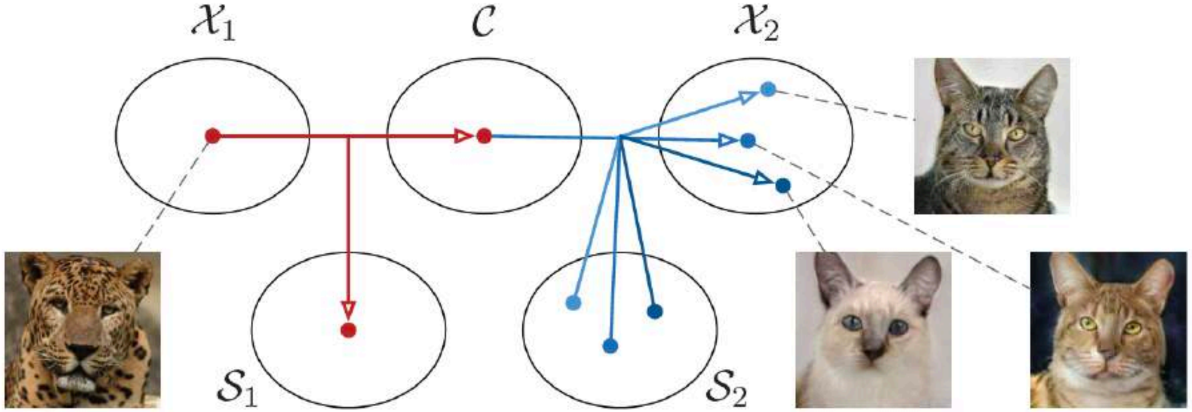
- ”*Multimodal Unsupervised Image-to-image Translation*” by Xun Huang, Ming-Yu Liu, Serge Belongie, Jan Kautz, ECCV 2018
- Code: https://github.com/NVlabs/MUNIT



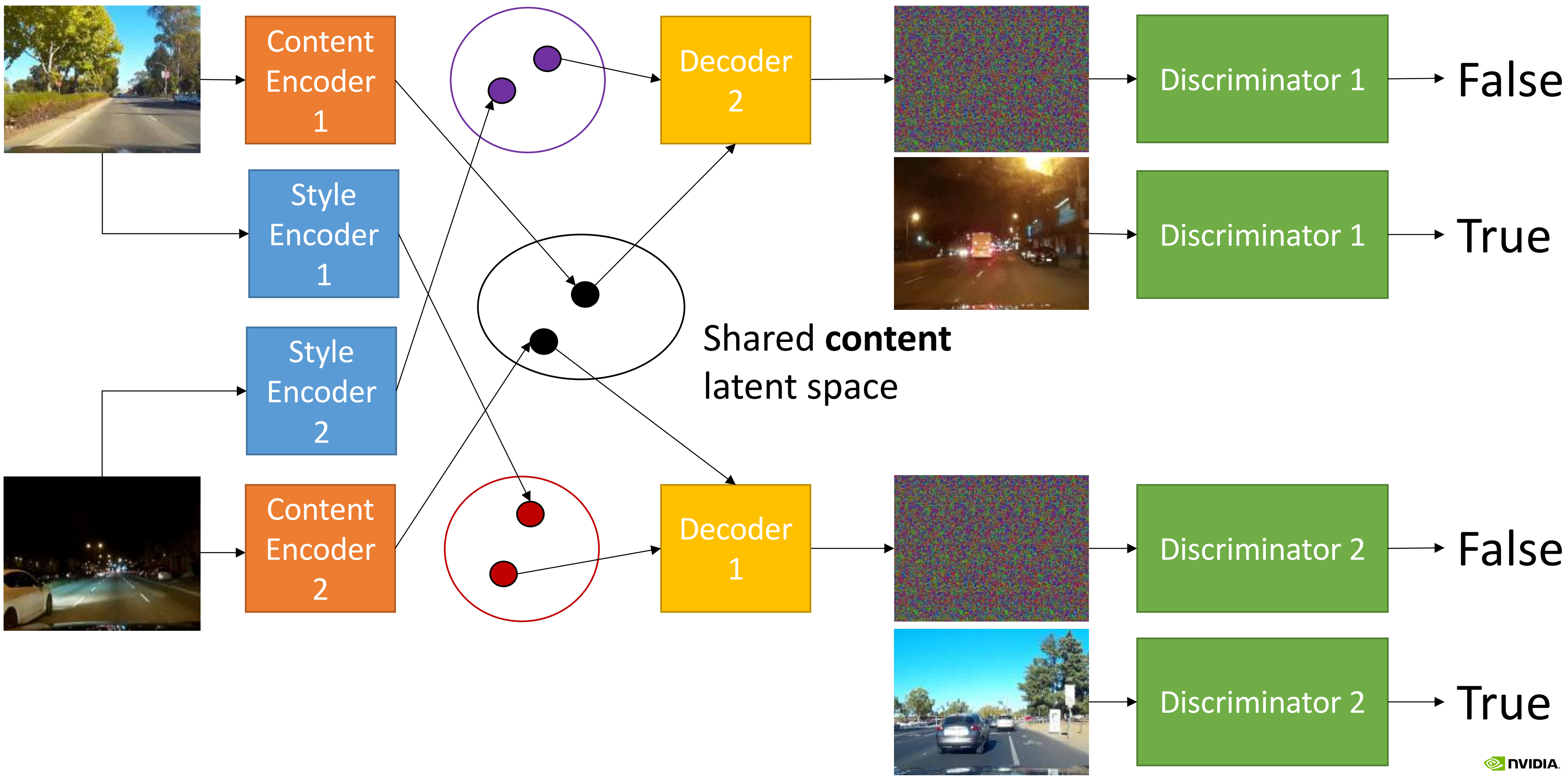

MUNIT assumption: Partially Shared Latent Space



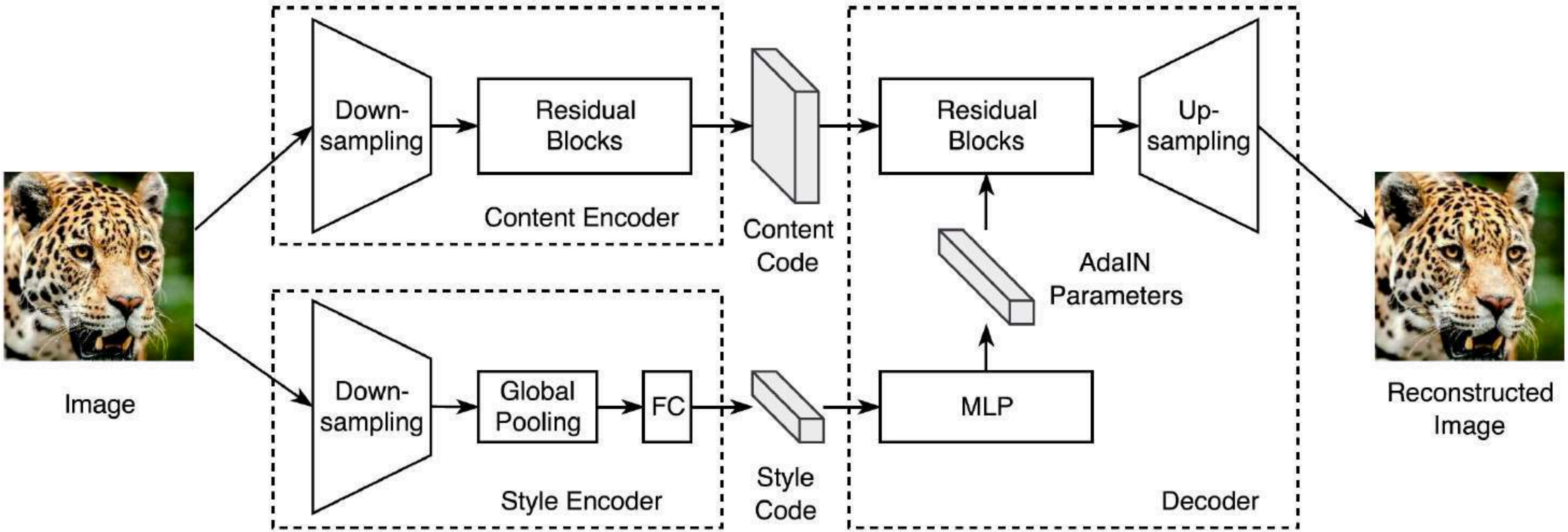
(a) Auto-encoding



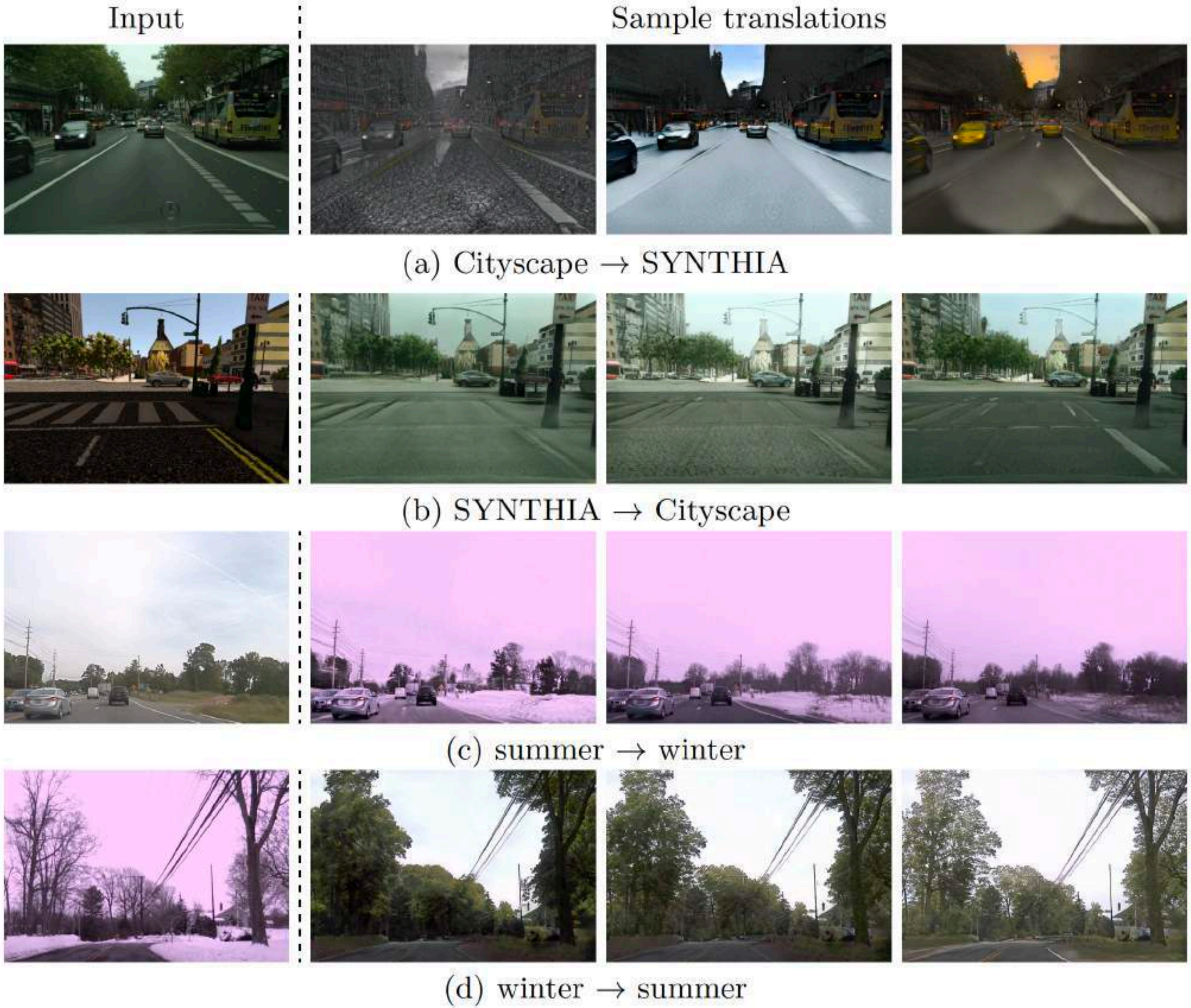
(b) Translation


Content Encoder 1
Style Encoder 1
Style Encoder 2
Content Encoder 2
Decoder 2
Decoder 1
Shared **content** latent space
Discriminator 1
False
Discriminator 1
True
Discriminator 2
False
Discriminator 2
True

MUNIT network



MUNIT results


Input
Sample translations
(a) Cityscape → SYNTHIA
(b) SYNTHIA → Cityscape
(c) summer → winter
(d) winter → summer

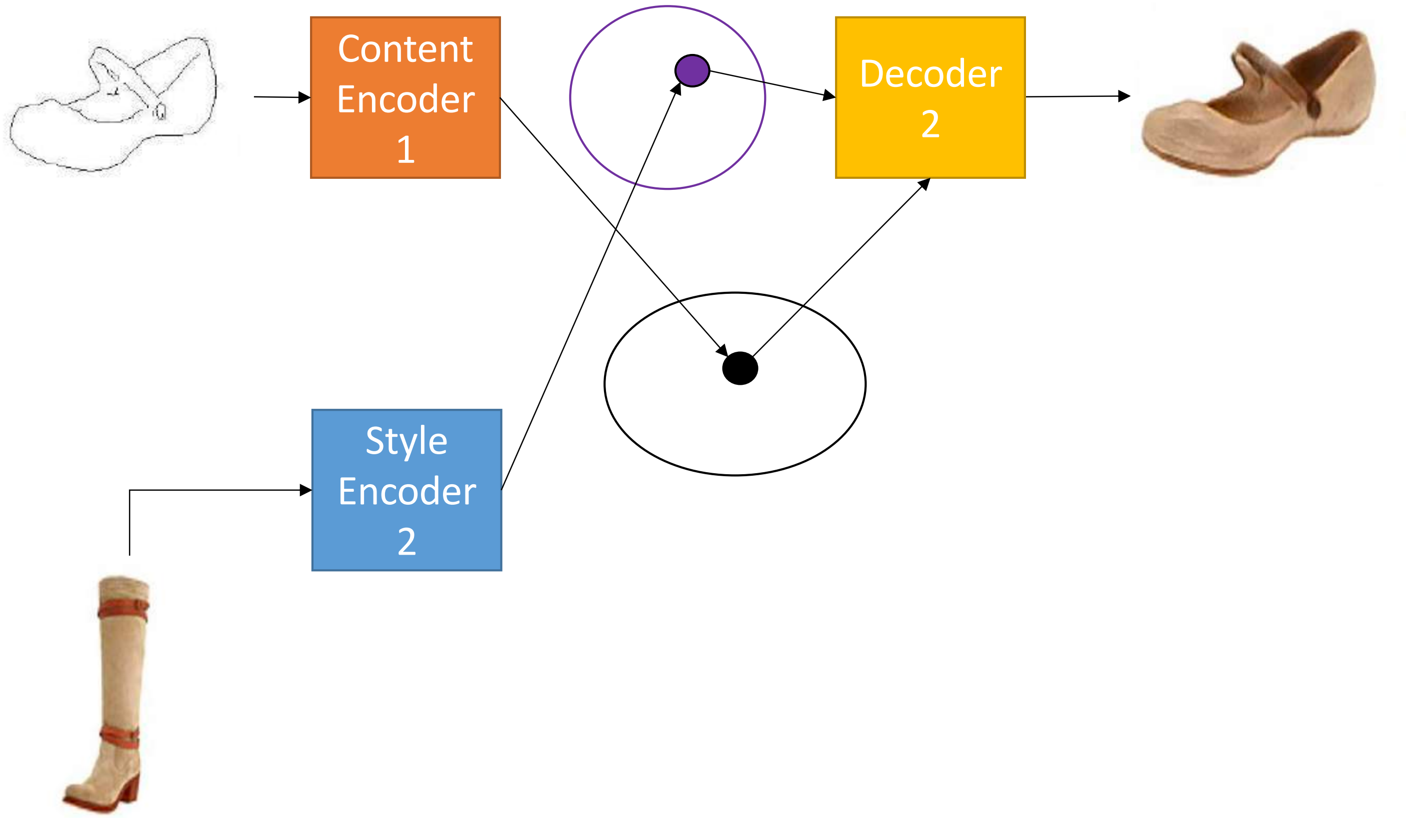
Input | Sample translations

(a) Yosemite summer → winter

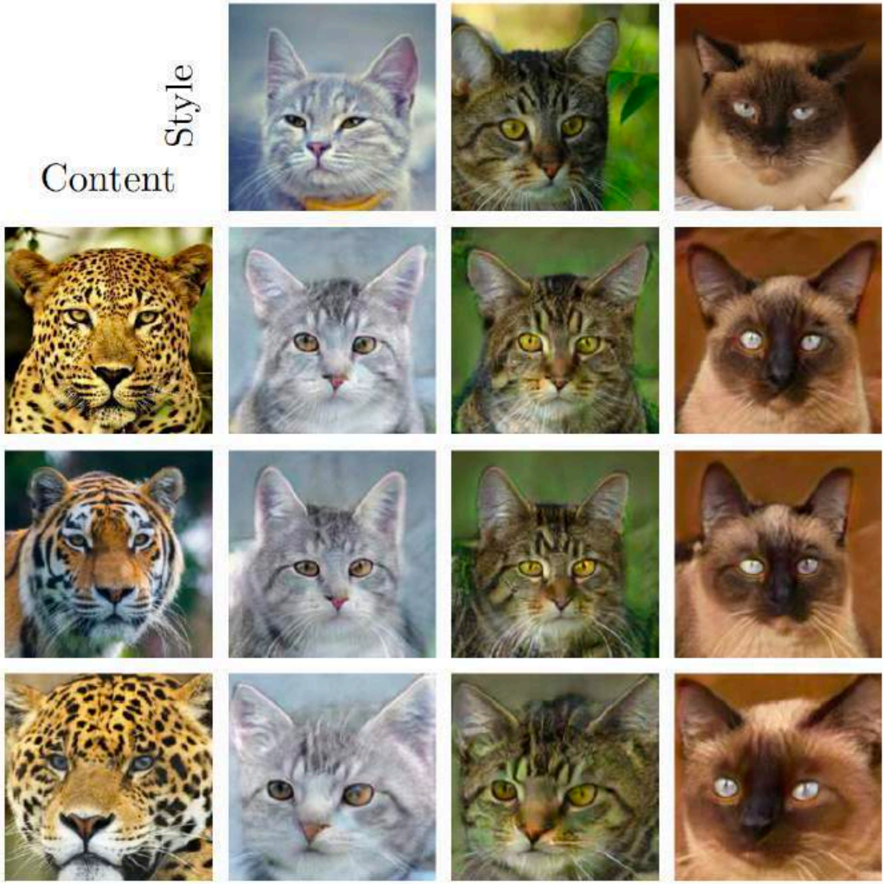
(b) Yosemite winter → summer

MUNIT Style Transfer



(b) edges → shoes

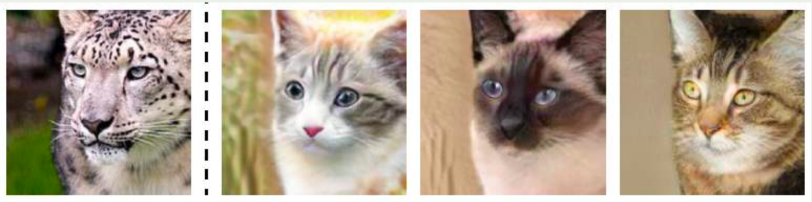


(b) big cats → house cats

Style Transfer Comparison

Conclusion

- Example-based image domain transfer
- Learning-based image domain transfer

Learning-based	Unimodal	Multimodal
Supervised		
Unsupervised		

YOUR LIFE'S WORK STARTS HERE

JOIN NVIDIA

100 Best Companies to Work For
— Fortune

Most Innovative Companies
— Fast Company

World's Most Admired Companies
— Fortune

Employees' Choice: Highest Rated CEOs
— Glassdoor

50 Smartest Companies
— MIT Tech Review

INTERESTED? Email: aijobs@nvidia.com

NVIDIA


NVIDIA