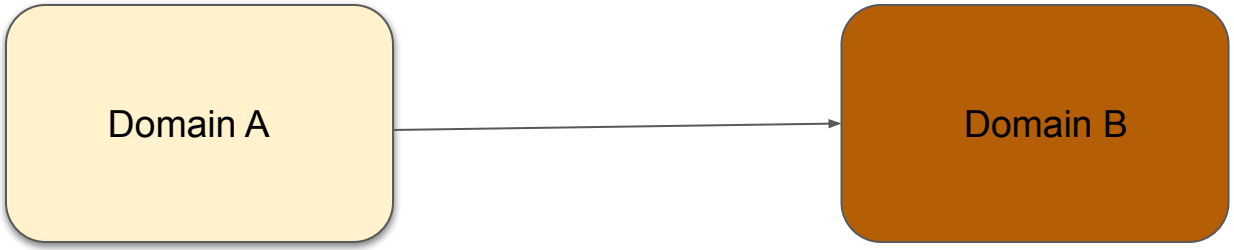
CPSC 532s Paper Reading 5a

# Diverse Image-to-Image Translation via Disentangled Representations

# Image to Image Translation (I2I)

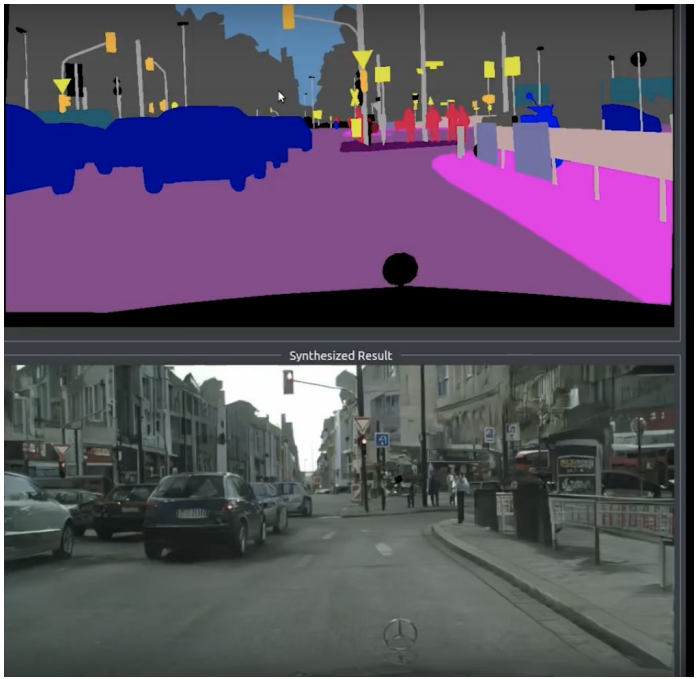
Learn mapping between two visual domains:



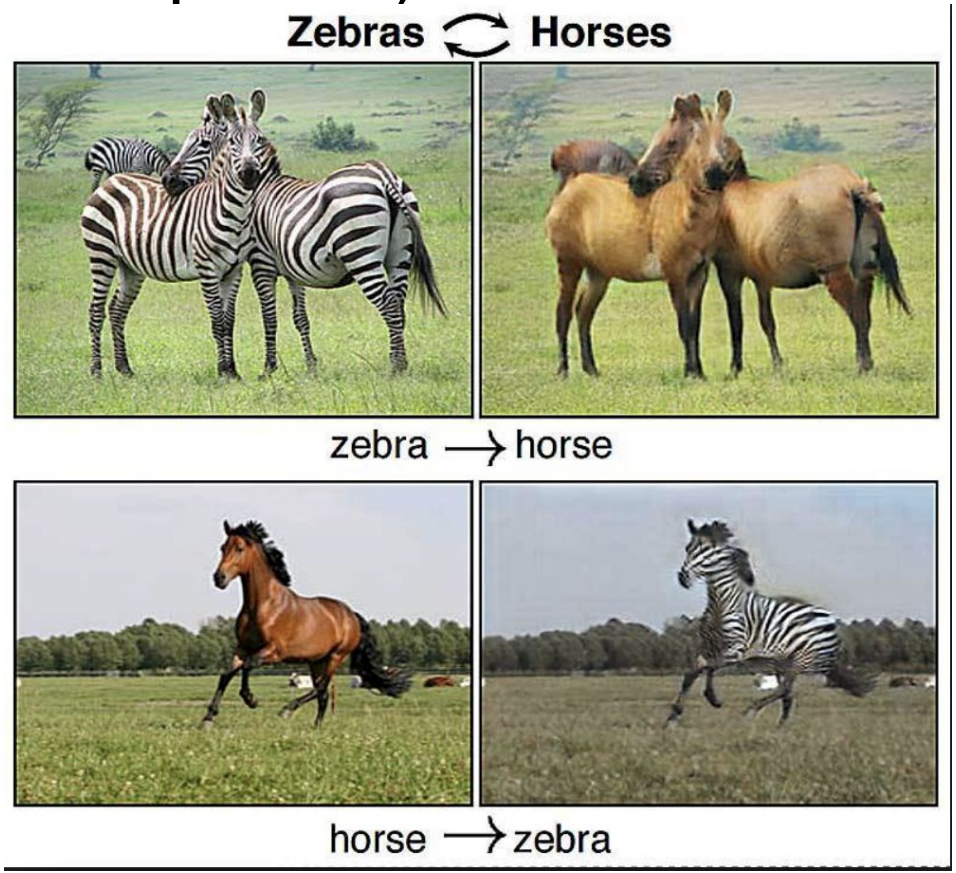

Many vision and graphic problems can formulated as Image to Image translation

# I2I (colourization)

# I2I (Image Synthesis)



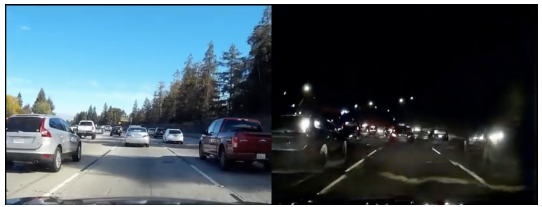
# I2I (domain adaptation)



# Challenges of Image to Image Translation

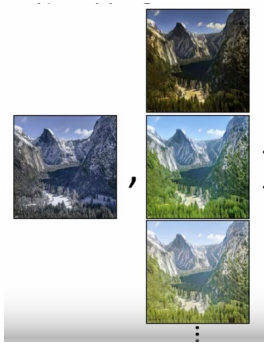
- The absence or difficulty of collecting aligned training pairs

Scarce



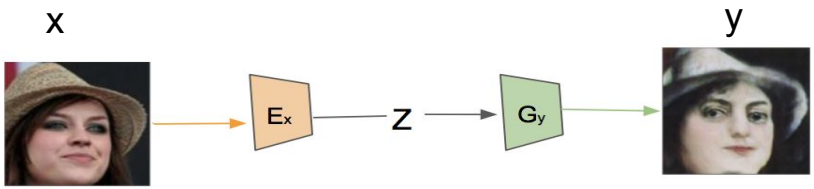
Not Exist


- Multiple possible outputs from the single image

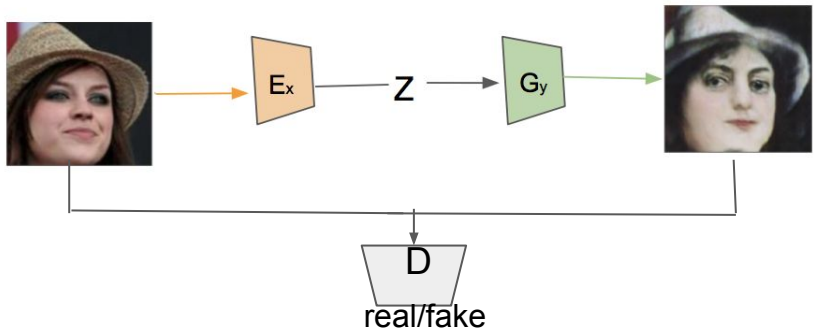


# Related work (Pix2pix: Isola et al. 2017)

Assuming to have image pairs, a model can learn mapping function between them:




Train model with GAN so discriminator can distinguish fake or real pairs.

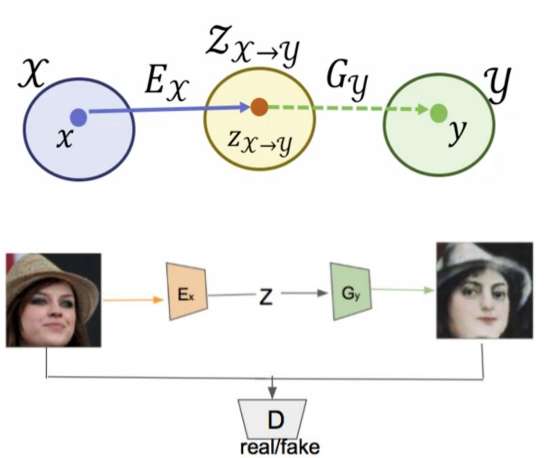


# Comparison of unsupervised I2I translation Models

One to One mapping Models:

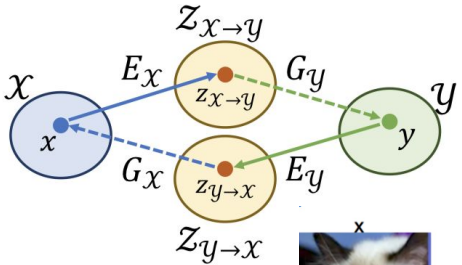
## *Paired Images*

Pix2Pix Approach

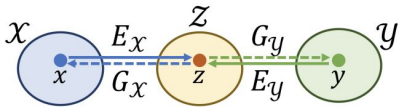


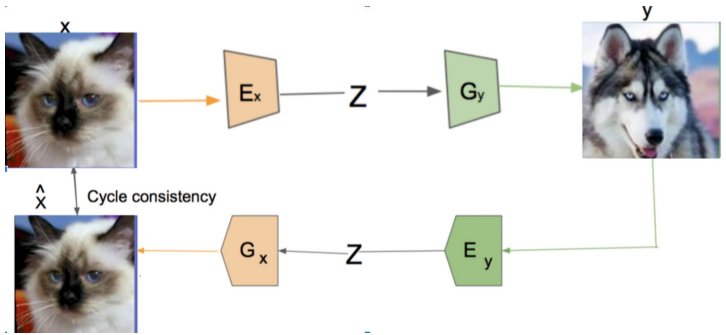

## *Unpaired Images*

CycleGan Approach



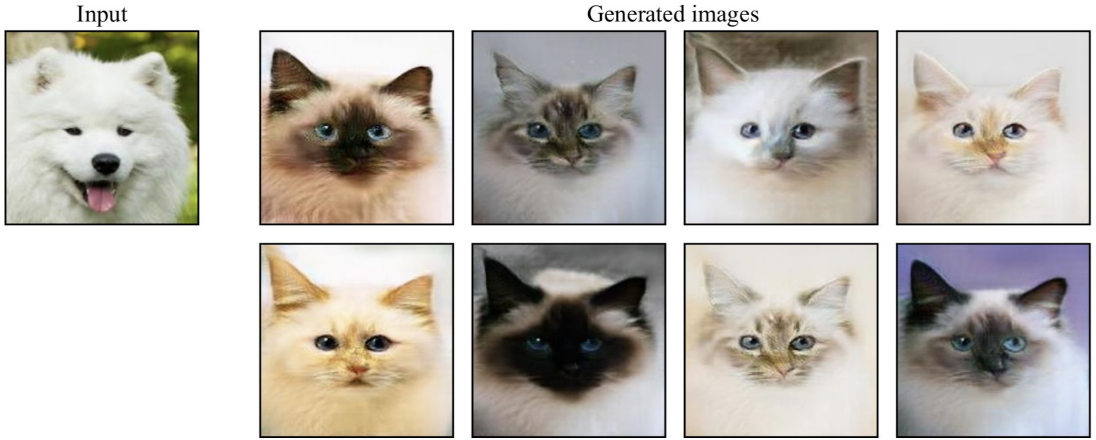

Unit Approach



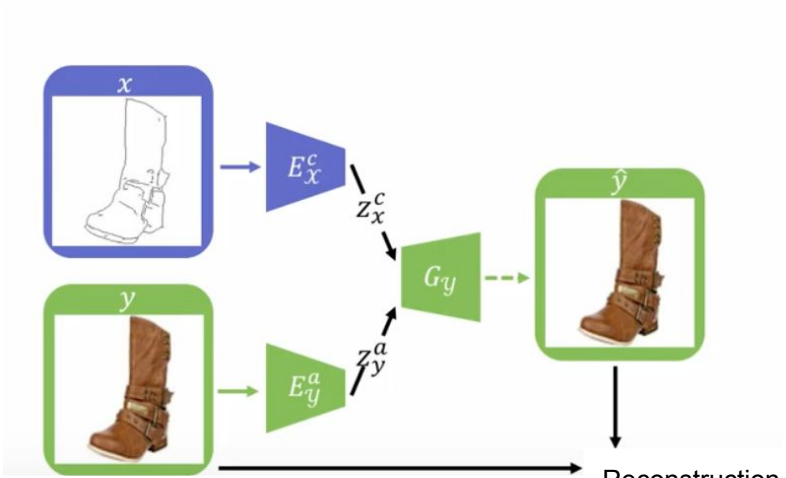
# A straightforward approach to handle multimodality

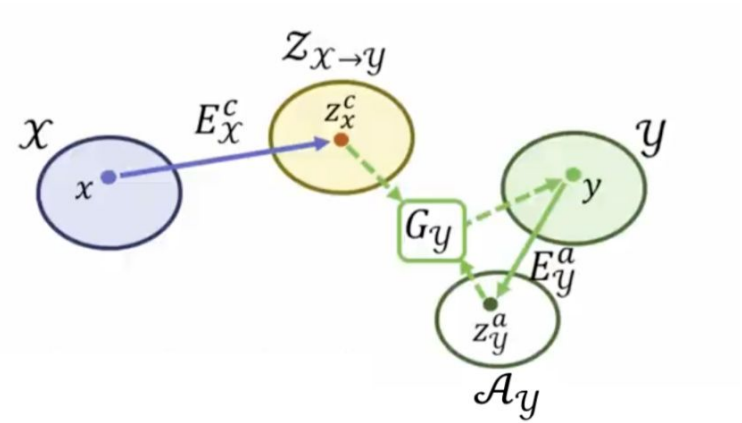
- take random noise vectors along with the conditional contexts as inputs, where the contexts determine the main content and noise vectors are responsible for variations.



# BicycleGan Approach:

*Many to many* mapping model with *paired images*:

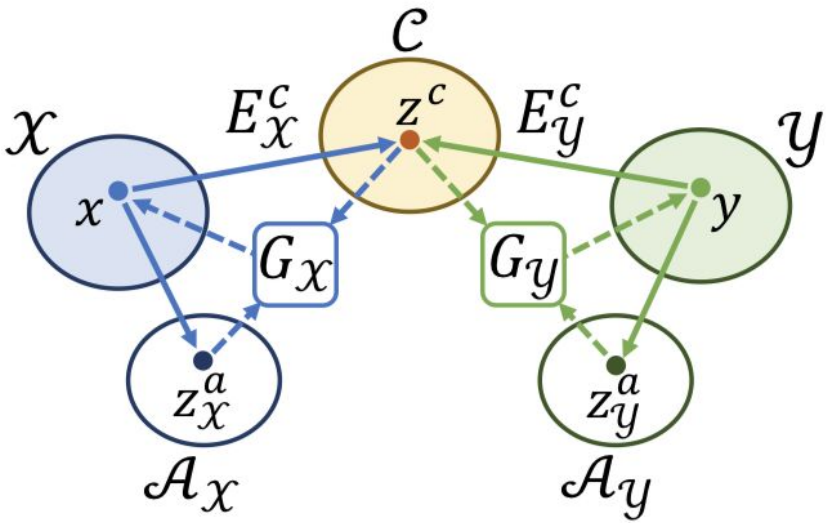


# Problem Statement

Problem:

- Idea is to create a disentangled representation to generate *<u>diverse</u> outputs* with *<u>unpaired training data.</u>*



# Methods

Disentangled Representations

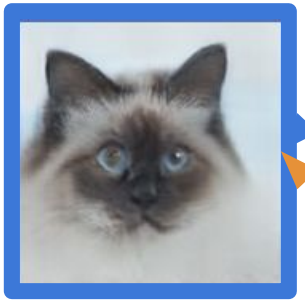
- Content space - Shared
- Attribute space - Domain-specific

A cat

*Attribute vector*

- Looking to the **LEFT**
- **DARK** background

***Content vector***

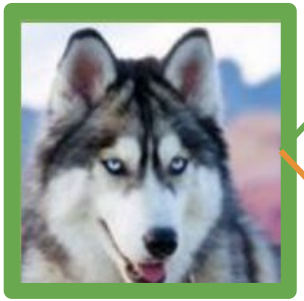


A cat

*Attribute vector*

- Looking at **FRONT**
- **BRIGHT** background

***Content vector***



A dog

*Attribute vector*

- Looking at **FRONT**
- **BRIGHT** background

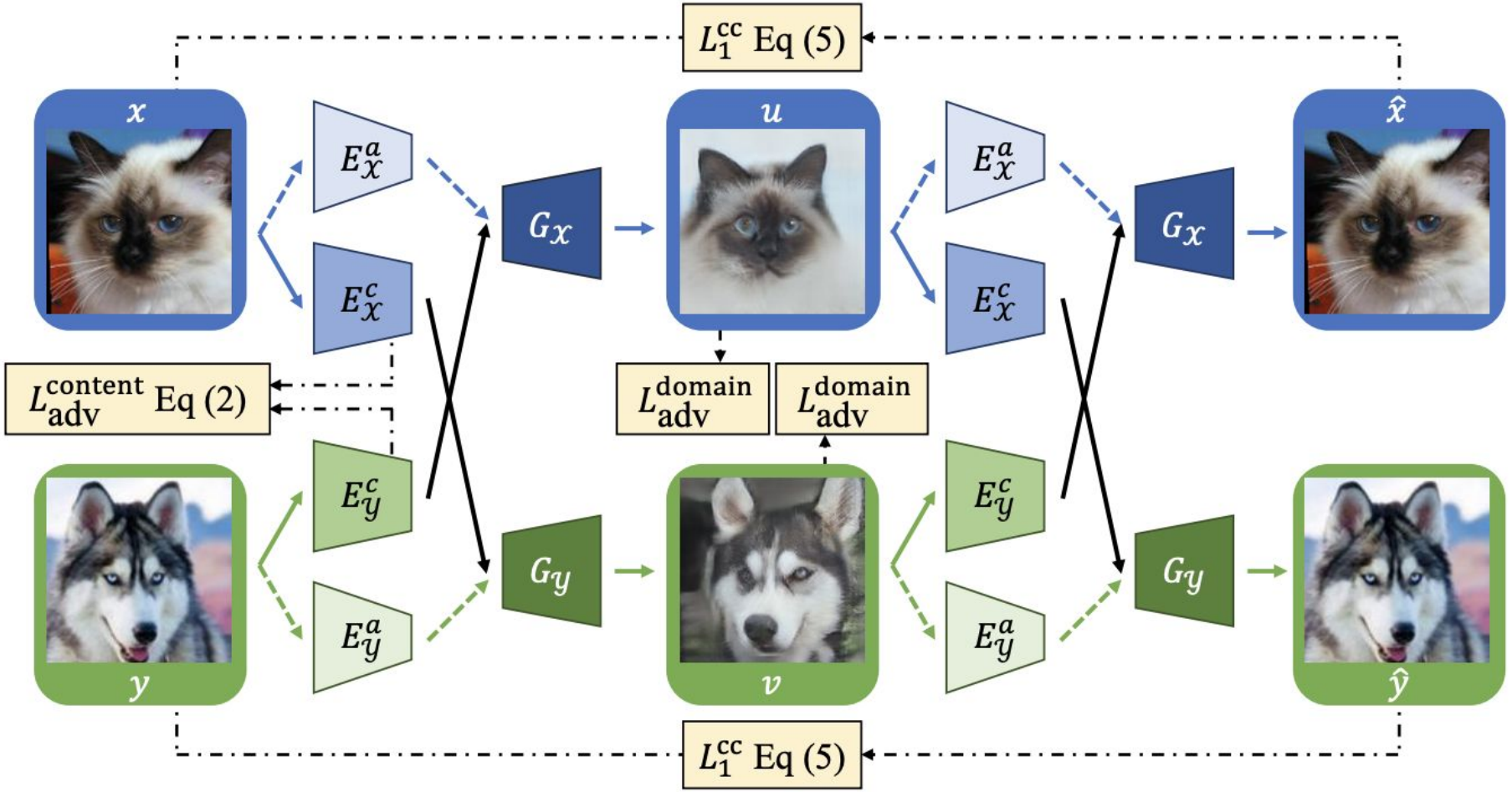
***Content vector***

A dog

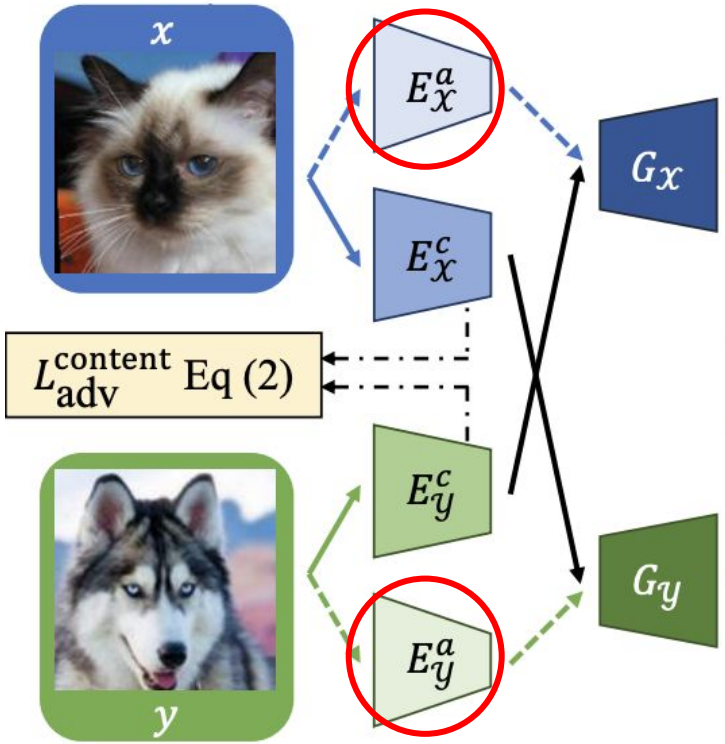
*Attribute vector*

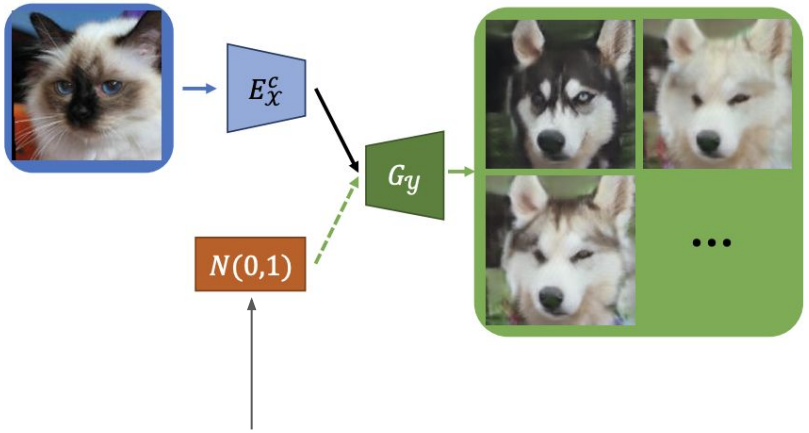
- Looking to the **LEFT**
- **DARK** background

***Content vector***


$L_1^{cc}$ Eq (5)
$x$
$E_x^a$
$E_x^c$
$G_x$
$u$
$E_x^a$
$E_x^c$
$G_x$
$\hat{x}$
$L_{adv}^{content}$ Eq (2)
$L_{adv}^{domain}$
$L_{adv}^{domain}$
$E_y^c$
$E_y^a$
$G_y$
$v$
$E_y^c$
$E_y^a$
$G_y$
$y$
$\hat{y}$
$L_1^{cc}$ Eq (5)

# Introducing Diversity

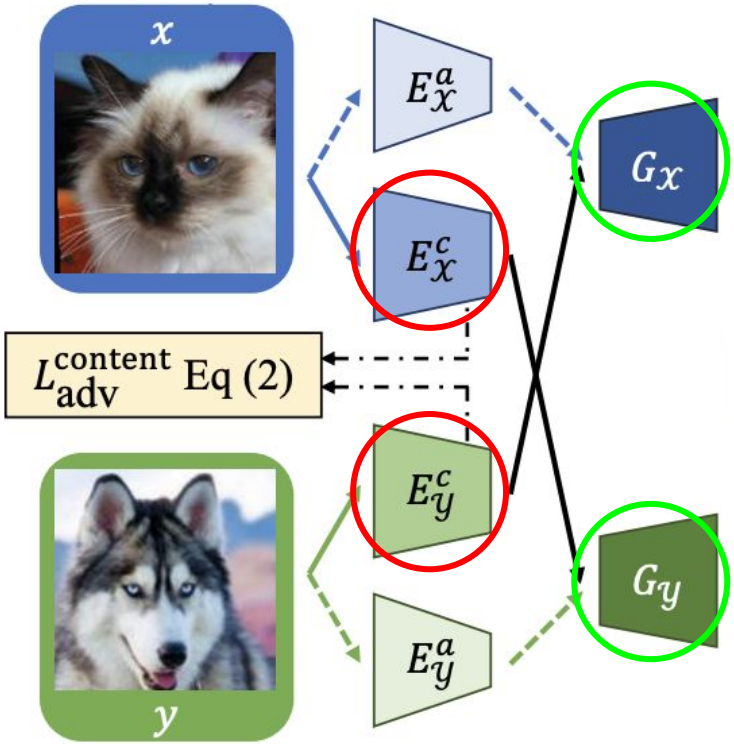



Randomly sampled from attribute space

# Strategies of Representation Disentanglement

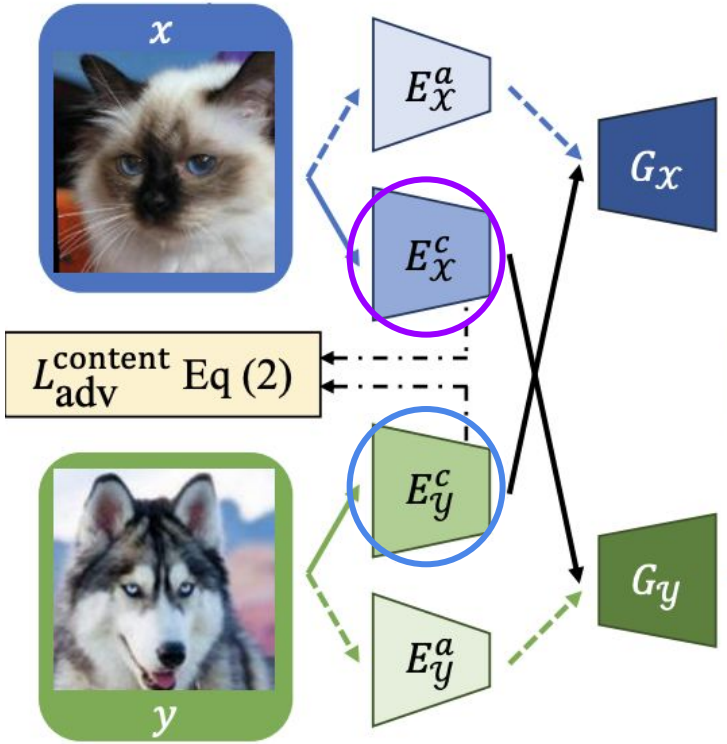
- Weight sharing



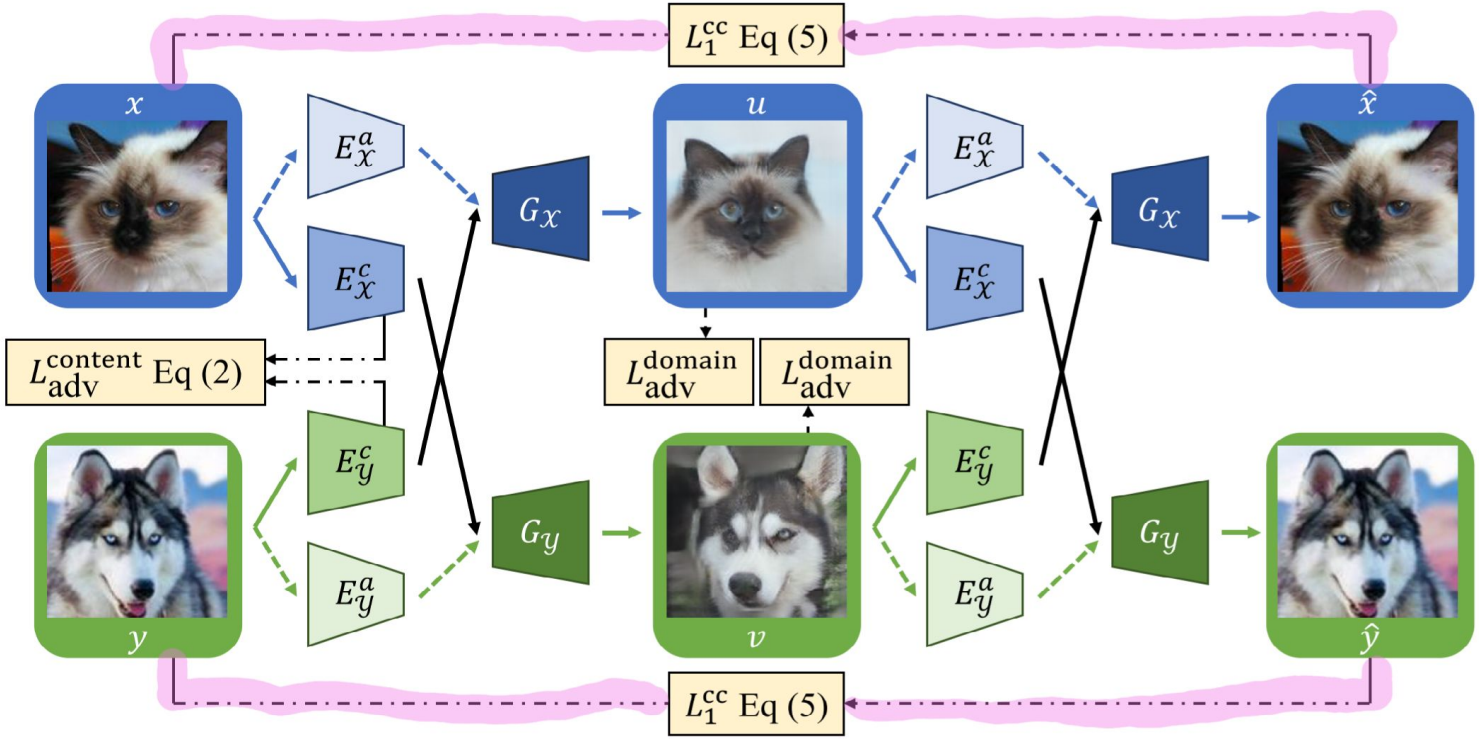
# Strategies of Representation Disentanglement

- Weight sharing
- Content discriminator

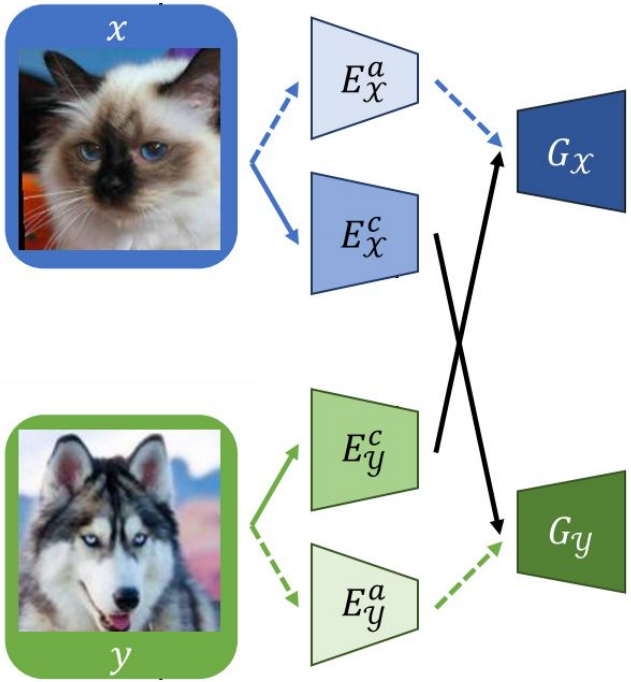
$$L_{\text{adv}}^{\text{content}}(E_{\mathcal{X}}^c, E_{\mathcal{Y}}^c, D^c) = \mathbb{E}_x[\frac{1}{2}\log D^c(E_{\mathcal{X}}^c(x)) + \frac{1}{2}\log(1 - D^c(E_{\mathcal{X}}^c(x)))] + \mathbb{E}_y[\frac{1}{2}\log D^c(E_{\mathcal{Y}}^c(y)) + \frac{1}{2}\log(1 - D^c(E_{\mathcal{Y}}^c(y)))]$$



# Cross-Cycle Consistency Loss

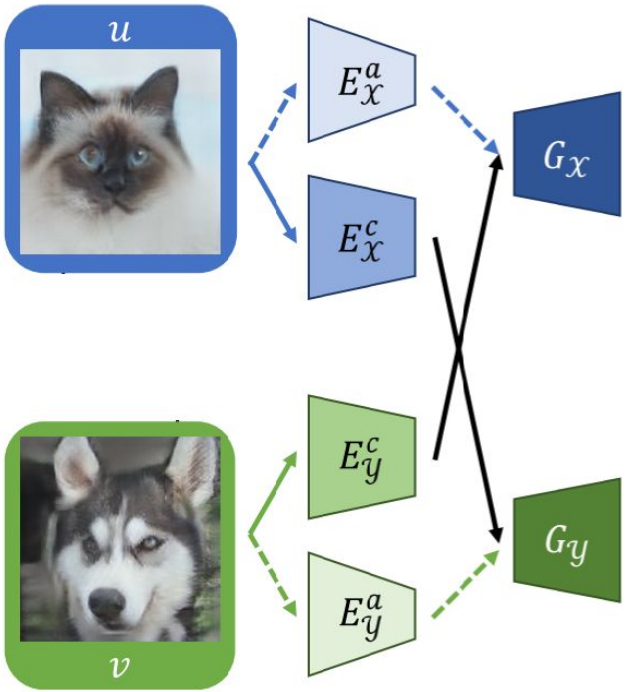


# Cross-Cycle Consistency Loss




- An image from each domain is decomposed into attribute and content representations
- The content vectors are then swapped with each other
- The new content-attribute pairs are then fed into the generator corresponding to their attribute representations

# Cross-Cycle Consistency Loss

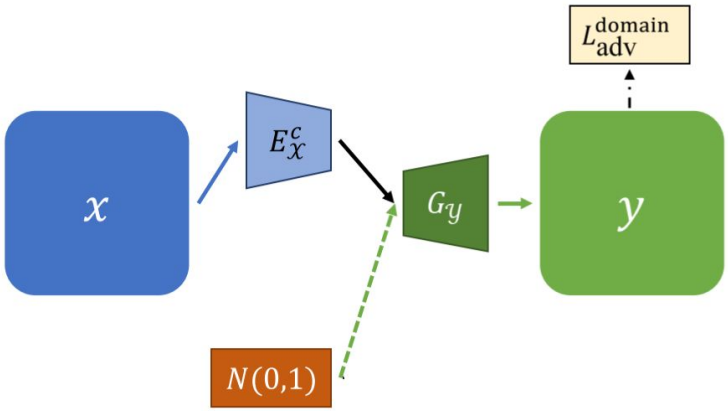



The process is the repeated on the generated images:

- The content and attribute vectors are encoded
- The content vectors are swapped
- The new content-attribute representation pairs are fed into their generators

The generators should produce something similar to the original images

# Domain Adversarial Loss

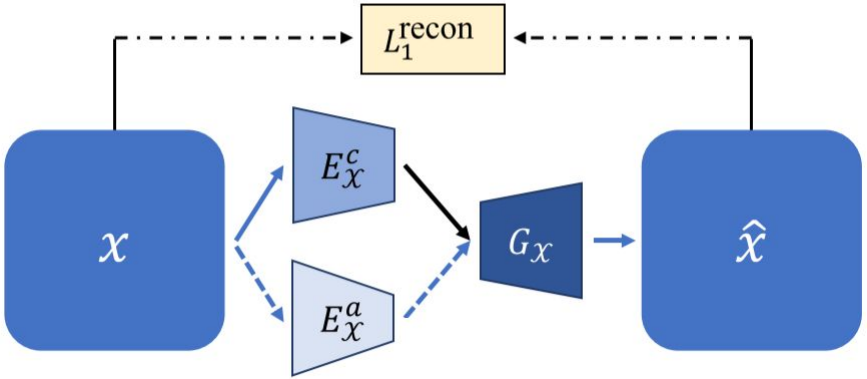



- Domain specific generators and discriminators are trained using this loss
- Discriminators must distinguish real images from generated images
- Attribute representations are pulled from a standard normal distribution
- A KL loss is used to coerce the attribute space to follow a standard normal distribution

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

# Self-Reconstruction Loss

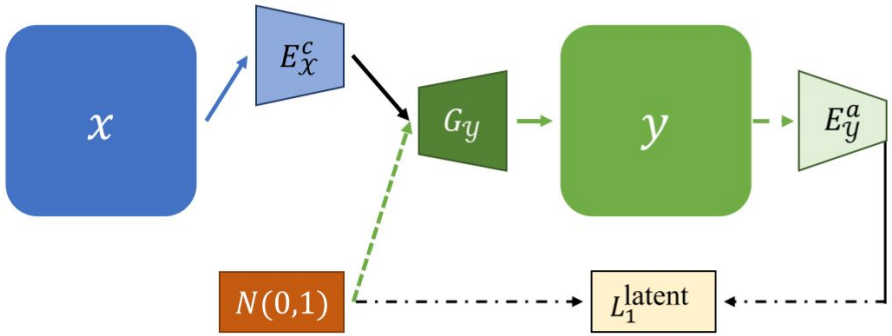
- The process of encoding and immediately decoding an image should produce the initial image
- Ensures that image information is preserved through the encoding and decoding phases




$$\hat{x} = G_{\mathcal{X}}(E^c_{\mathcal{X}}(x), E^a_{\mathcal{X}}(x))$$
$$\hat{y} = G_{\mathcal{Y}}(E^c_{\mathcal{Y}}(y), E^a_{\mathcal{Y}}(y))$$

# Latent Regression Loss

- Encourages an invertible mapping between the attribute space and images




$$\hat{x} = G_{\mathcal{X}}(E^c_{\mathcal{X}}(x), E^a_{\mathcal{X}}(x))$$
$$\hat{y} = G_{\mathcal{Y}}(E^c_{\mathcal{Y}}(y), E^a_{\mathcal{Y}}(y))$$

# Key Implementation Details

Input image size: 216*216

Content encoder $E^C$ : 3 conv + 4 residual blocks

Attribute encoder $E^a$ : 4 conv + FCs

Generator G: 4 residual blocks + 3 fractionally strided conv

# Evaluation

## Datasets:

- Yosemite
- Artworks
- Edge-to-shoes
- Photo-to-portrait (WikiArt)
- CelebA

## Compare DRIT with methods of:

- DRIT w/o Dc
- CycleGAN, UNIT, BicycleGAN
- Cycle/Bicycle (Baseline)

## Domain adaptation:

- MNIST to MNIST-M

- Synthetic Cropped LineMod to Cropped LineMod

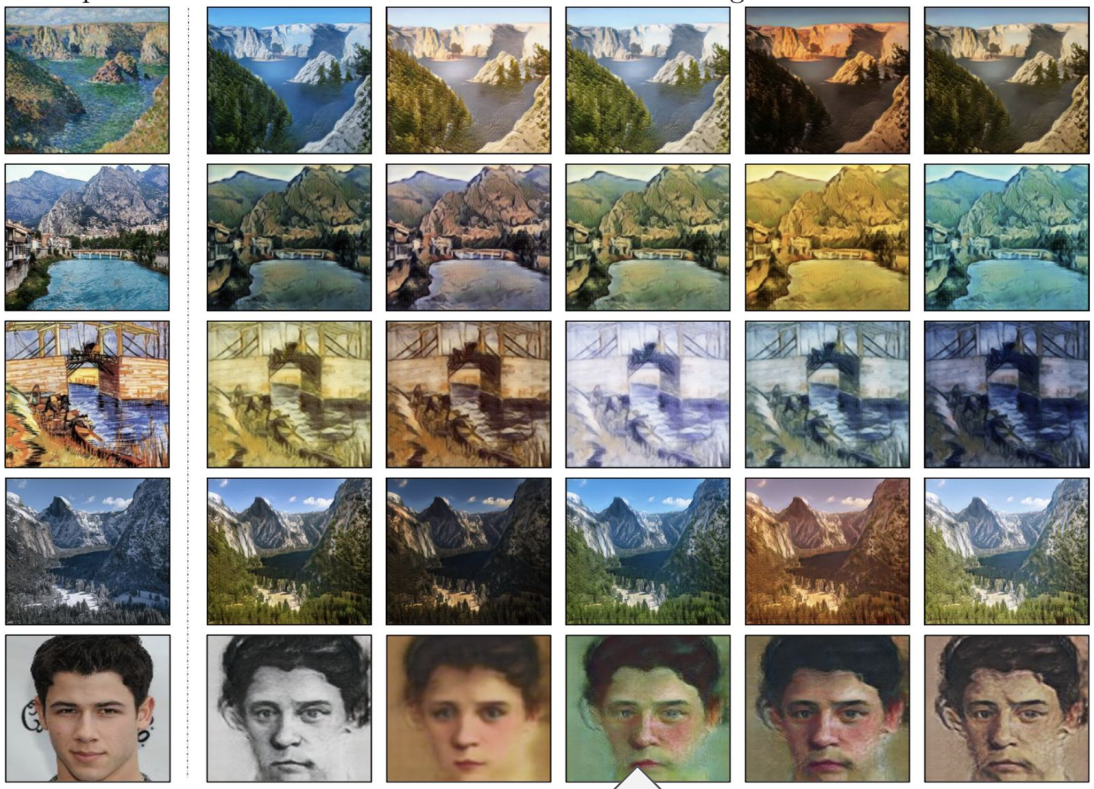
# Evaluation

- Qualitative evaluation
- Quantitative evaluation

We want to show:

- Diverse translation results with randomly sampled attributes
- Example guided translation with transferred attribute vectors from existing images

# Qualitative Evaluation - Diversity



Images generated by random vectors in the attribute space

Monet - photo

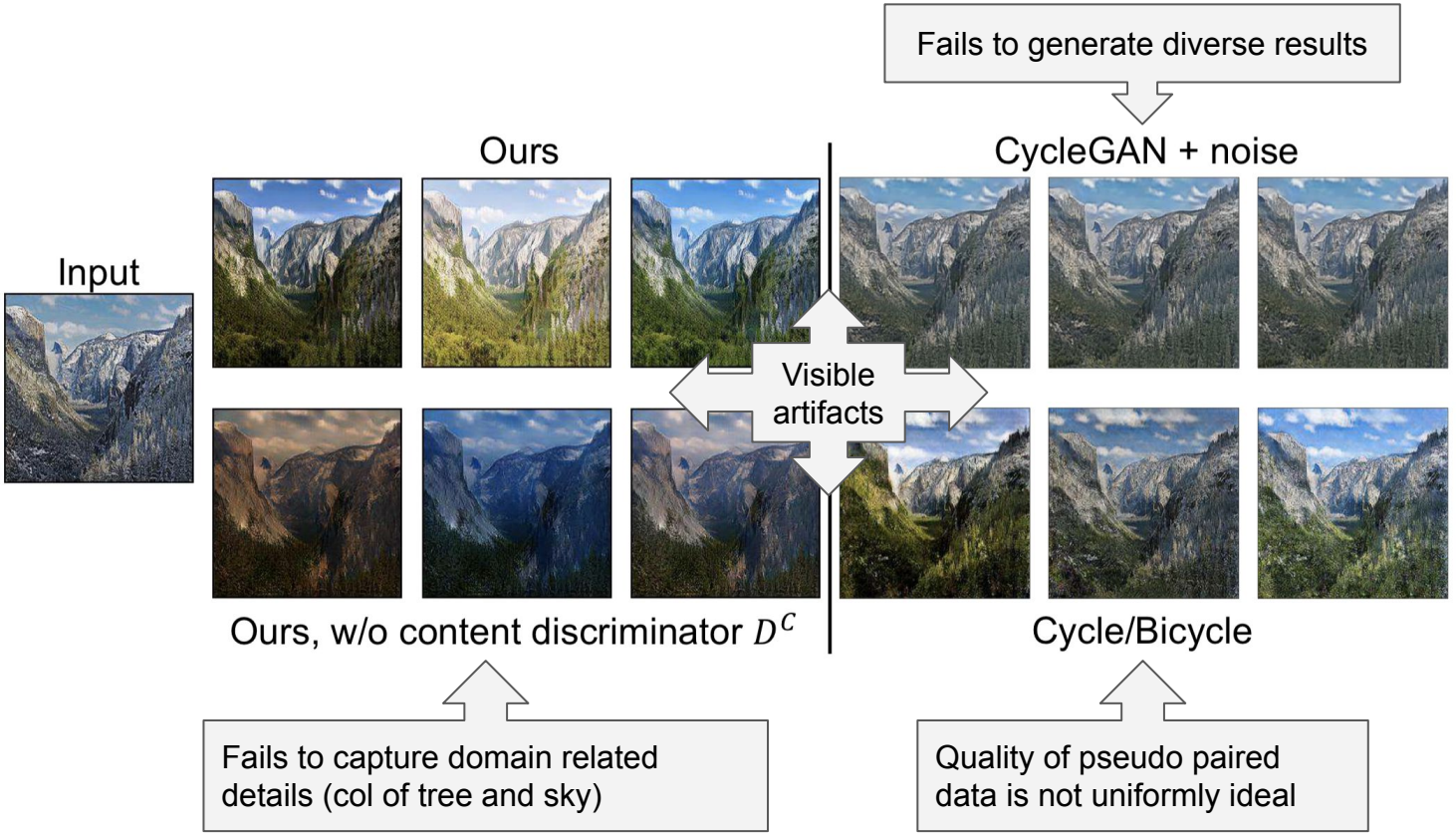
Photo - van Gogh

Van Gogh - Monet

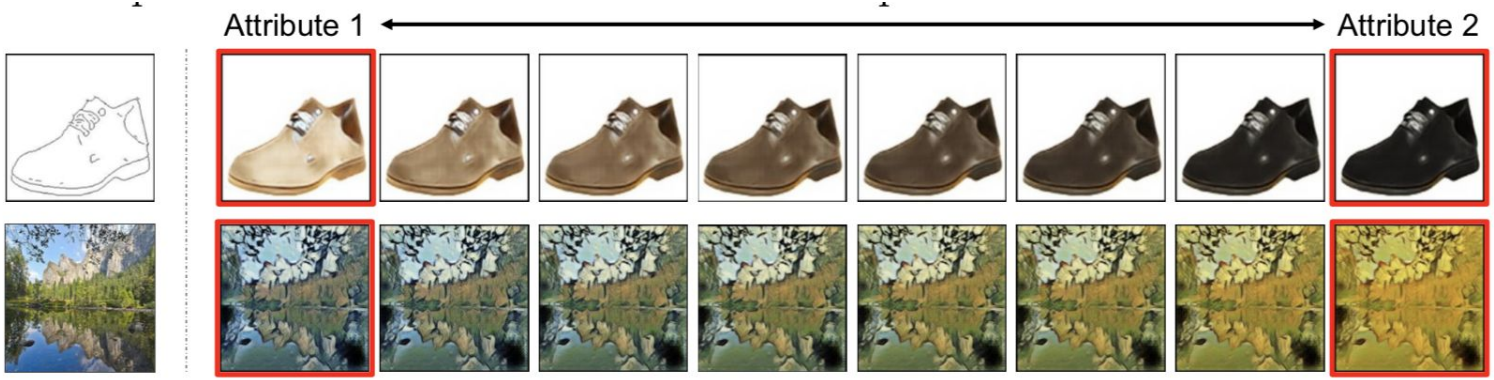
Winter - summer

Photograph - portrait

# Qualitative Evaluation - Diversity



# Qualitative Evaluation - Diversity

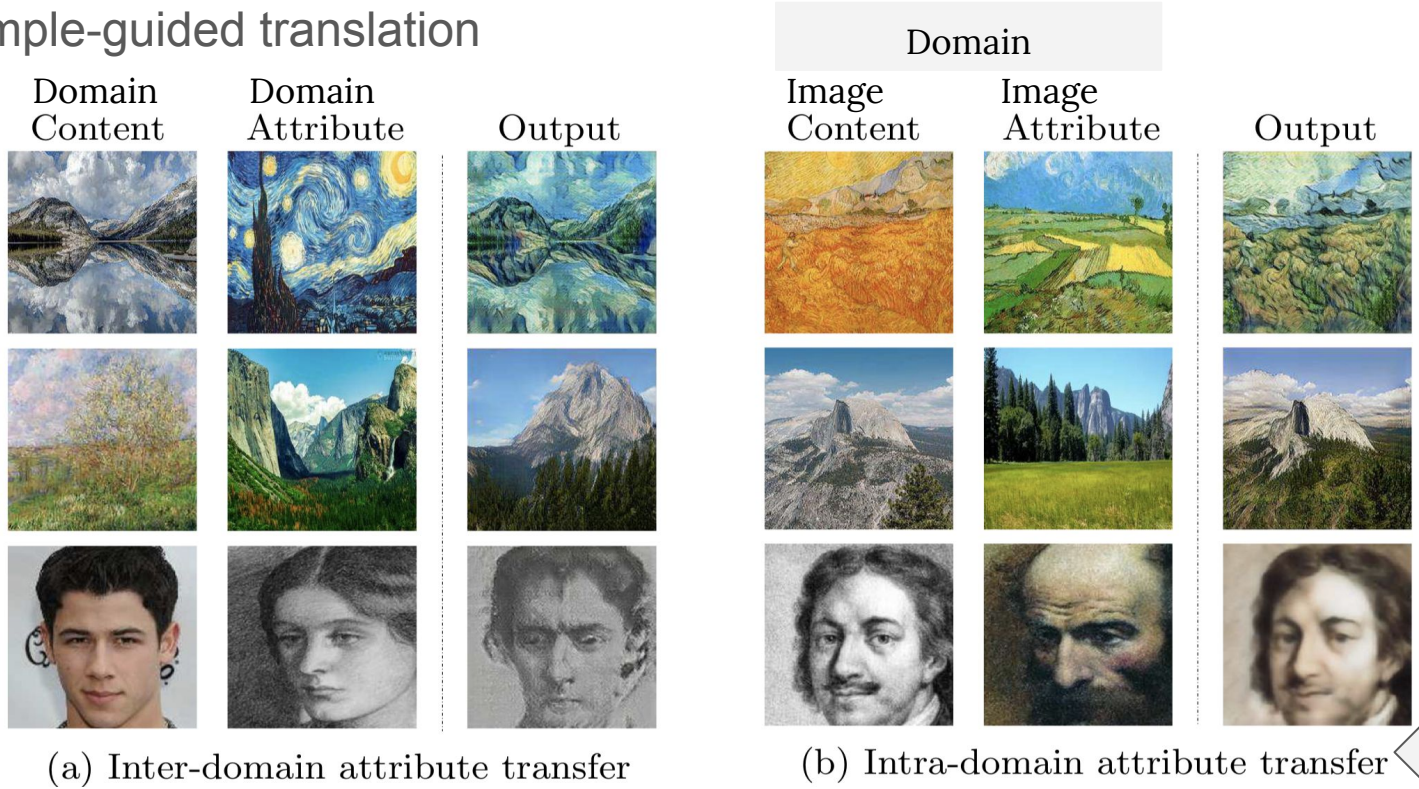



Linear interpolation between two attributes

- Verifies the continuity in the attribute space
- Model can generalize in the distribution

# Qualitative Evaluation - Attribute Transfer

## Example-guided translation

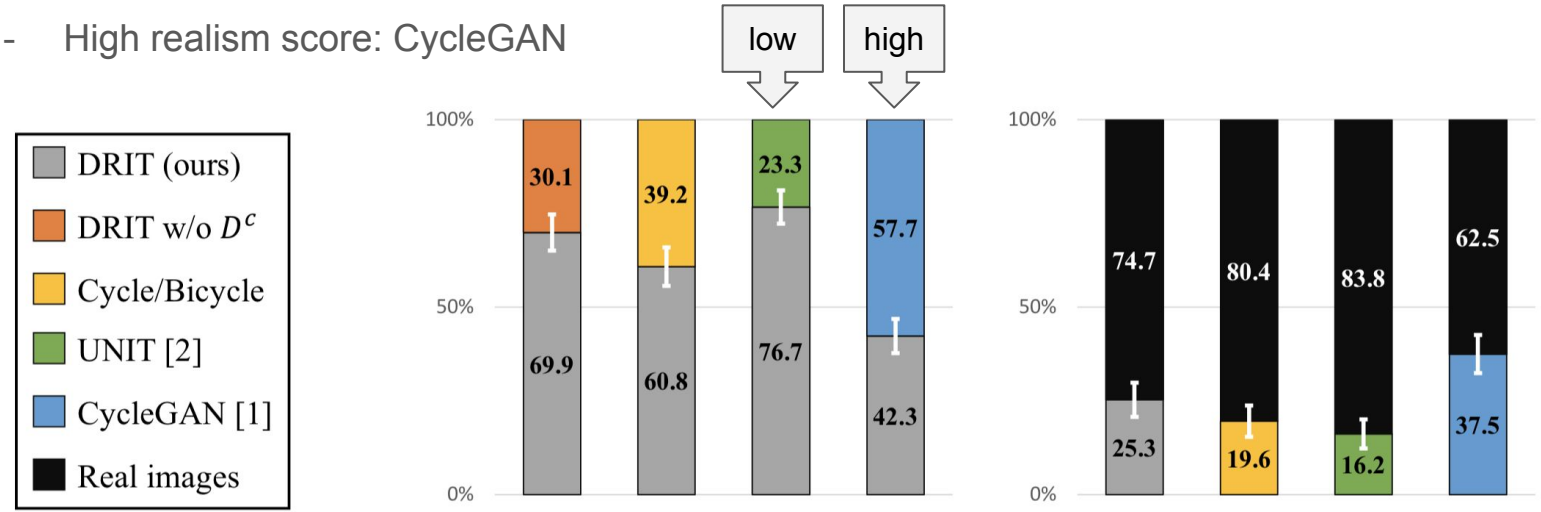


# Quantitative Evaluation - Realism

Dataset: winter-summer translation with Yosemite dataset

Realism:

- User study using pairwise comparison
- Low realism score: UNIT
- High realism score: CycleGAN



# Quantitative Evaluation - Diversity

Diversity:

- LPIPS metric to measure similarity among images
- Compute distance between 1000 pairs of random images translated from 100 real images

| Method | Diversity |
|---|---|
| real images | .448 ± .012 |
| DRIT | **.424** ± .010 |
| DRIT w/o $D^c$ | .410 ± .016 |
| UNIT [27] | .406 ± .022 |
| CycleGAN [48] | .413 ± .008 |
| Cycle/Bicycle | .399 ± .009 |

Highest in realism, but limited diversity

Constrained by data generated by CycleGAN

# Quantitative Evaluation - Reconstruction Ability

Dataset: edge-to-shoes dataset

Test: given a paired data, measure the reconstruction errors of y with

$$\hat{y} = G_y(E_X^a(x),\ E_Y^a(y))$$

Compare performance:

Trained on paired data

Trained on unpaired data

| Method | Reconstruct error |
|---|---|
| BicycleGAN [49]* | **0.0945** |
| DRIT | <u>0.1347</u> |
| DRIT, w/o $D^c$ | 0.2076 |

# Domain Adaptation

## I2I translation scheme -- domain adaptation

- Train labeled source images to the target domain
- Treat the generated labeled images as training data
- Train the classifiers of each task in the target domain

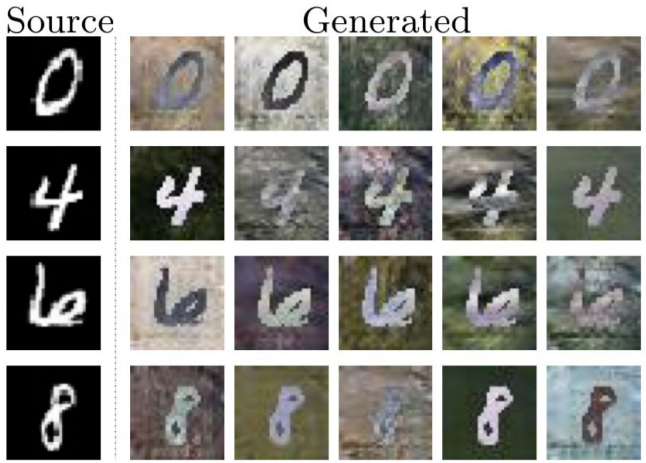
## Tasks

- MINIST to MINIST-M
- Synthetic Cropped LineMod to Cropped LineMod

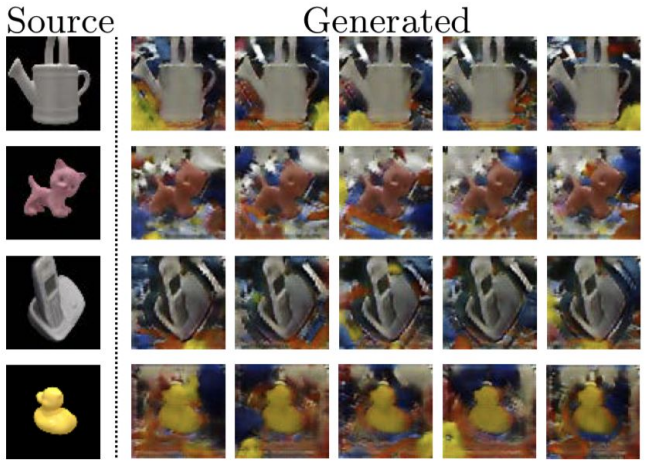
## Compare with

- CycleGAN
- PixelDA
- DANN
- DSN

# Domain Adaptation




(c) MNIST → MNIST-M




(d) Synthetic → Real Cropped LineMod

# Domain Adaptation

Generates image with similar appearances, degrading the performance of the adapted classifiers

Comparable to CycleGAN

Leverage label info during training

(a) MNIST-M

| Model | Classification Accuracy (%) |
|---|---|
| Source-only | 56.6 |
| CycleGAN [46] | 74.5 |
| Ours, ×1 | 86.93 |
| Ours, ×3 | 90.21 |
| Ours, ×5 | **91.54** |
| DANN [13] | 77.4 |
| DSN [4] | 83.2 |
| PixelDA [3] | **95.9** |
| Target-only | 96.5 |

(b) Cropped LineMod

| Model | Classification Accuracy (%) | Mean Angle Error (°) |
|---|---|---|
| Source-only | 42.9 (47.33) | 73.7 (89.2) |
| CycleGAN [46] | 68.18 | 47.45 |
| Ours, ×1 | 95.91 | 42.06 |
| Ours, ×3 | 97.04 | 37.35 |
| Ours, ×5 | **98.12** | **34.4** |
| DANN [13] | 99.9 | 56.58 |
| DSN [4] | **100** | 53.27 |
| PixelDA [3] | 99.98 | **23.5** |
| Target-only | 100 | 12.3 (6.47) |

# Conclusion

| *Positive Points* | *Negative Points* |
| --- | --- |
| A novel disentangled representation framework embeds images into domain-invariant content features and domain-specific attribute vectors | Additional components make these methods less generalizable and incur extra computational loads on training |
| The content discriminator is described to to facilitate the representation disentanglement and outputs be more diverse and realistic | Strong assumptions or regularizations including: shared latent spaces, loss on KL divergence from simple Gaussians |
| Propose a cross-cycle consistency loss for cyclic reconstruction of unpaired training data | Work limitations:<br>● Attribute space is not fully exploited<br>● Difficult when domain characteristics differ significantly |

# Future Works

- How to make generator/discriminator more aware of difference and capture high-level statistics of domains
- Video to Video translations

Thanks for Listening

# Questions?