NEURAL INFORMATION PROCESSING SYSTEMS

CycleNet: Rethinking Cycle Consistency in Text-Guided Diffusion for Image Manipulation


Sihan Xu*[1]


Ziqiao Ma*[1]


Yidong Huang[1]

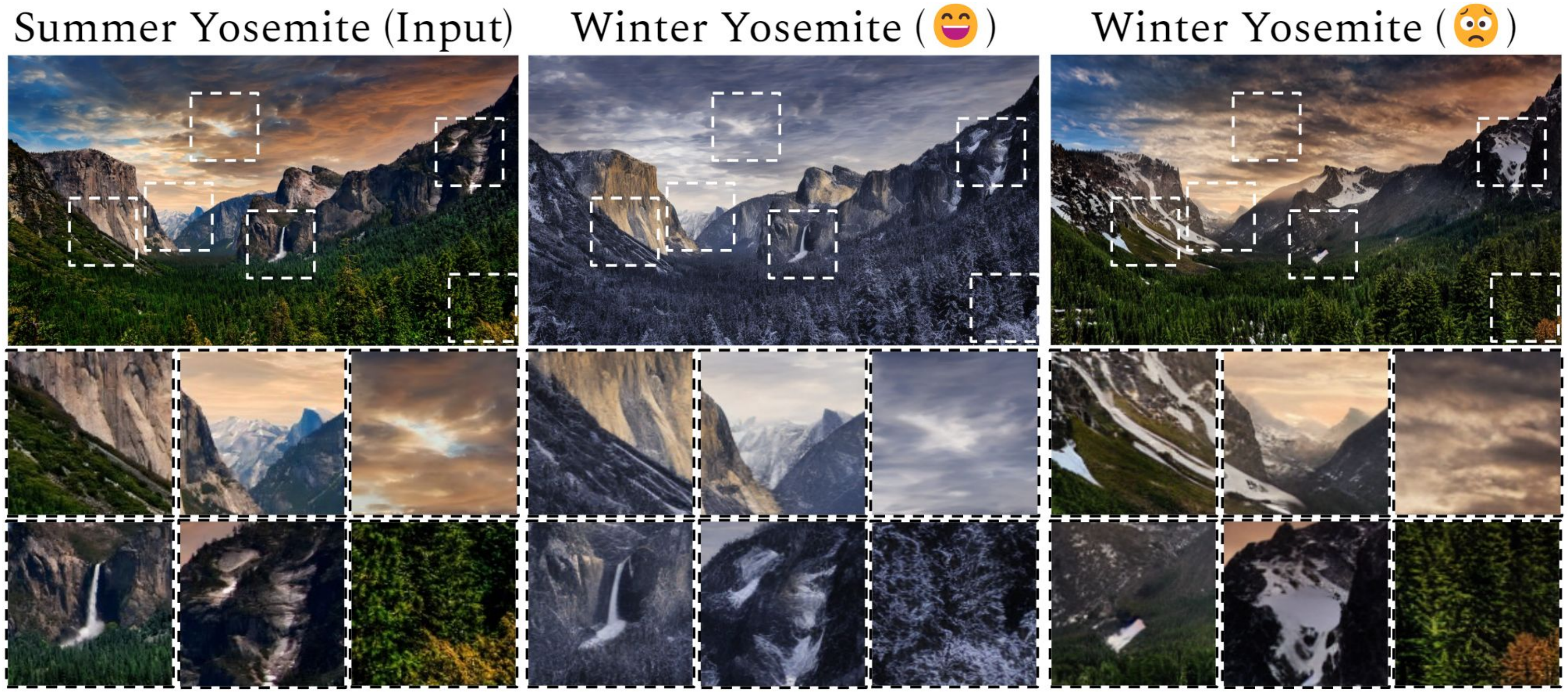

Honglak Lee[1,2]


Joyce Chai[1]

[1] UNIVERSITY OF MICHIGAN [2] LG AI Research

Background

- **The pain of consistency in image editing with diffusion models.**
  - Consistency is a desirable property in image manipulation, especially in unpaired I2I scenarios as there is no guaranteed correspondence between images in the source and target domains.
  - Pre-trained diffusion models (DMs) are effective in various image synthesis tasks.
  - Still, it remains an open challenge to adapt them in unpaired settings with a consistency guarantee.



Preliminaries

- **Diffusion models (DMs)**
  - DDPM (Ho et al., 2020) noted that the forward process allows the sampling of z_t at any time step t using a closed-form sampling function.

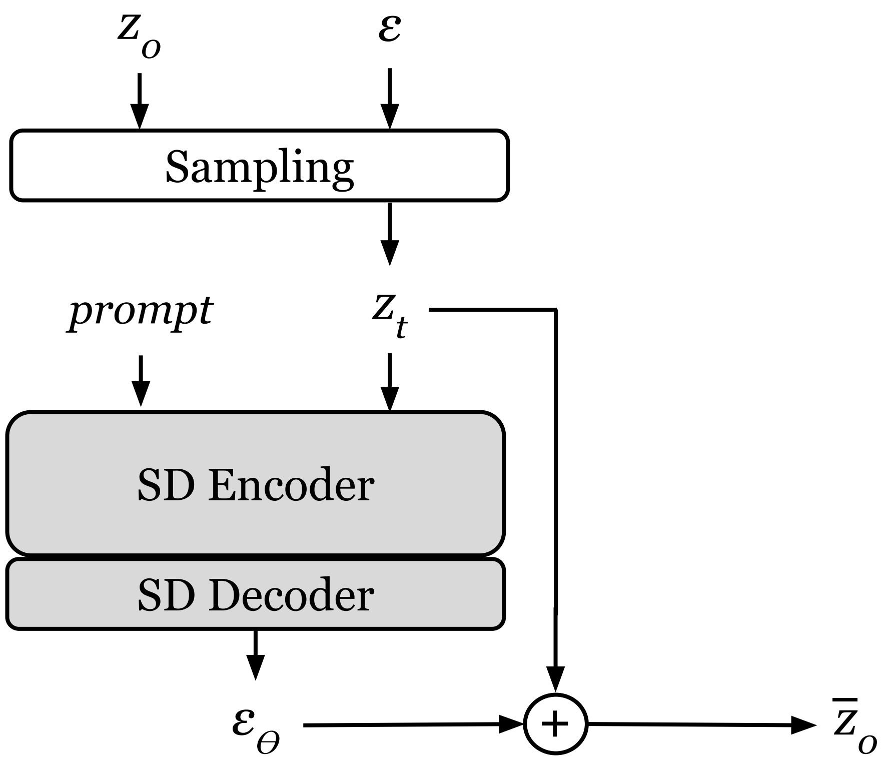
$$z_t = S(z_0, \varepsilon, t) := \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \mathbf{I}) \text{ and } t \sim [1, T]$$

  - The reverse process can be carried out with a network ε_θ that predicts the noise ε.

$$\min_\theta \mathbb{E}_{z_0, \varepsilon, t} \, ||\varepsilon - \varepsilon_\theta(z_t, t)||_2^2$$

  - One could estimate the original source image z_0 given a noised latent z_t.

$$\bar{z}_0 = G(z_t, t) := \left[z_t - \sqrt{1 - \bar{\alpha}_t}\varepsilon_\theta(z_t, t)\right] / \sqrt{\bar{\alpha}_t}$$

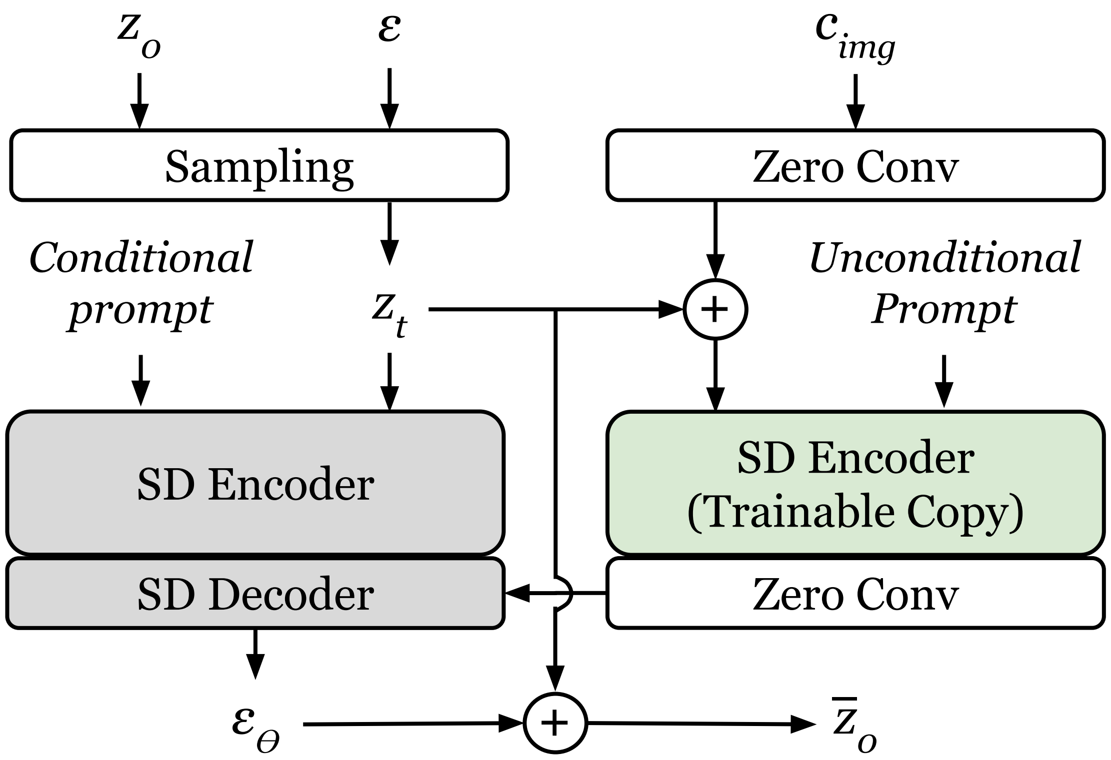


Preliminaries

- **Conditioning in DMs**
  - Under text and image conditioning, the reconstructed image can be written as:

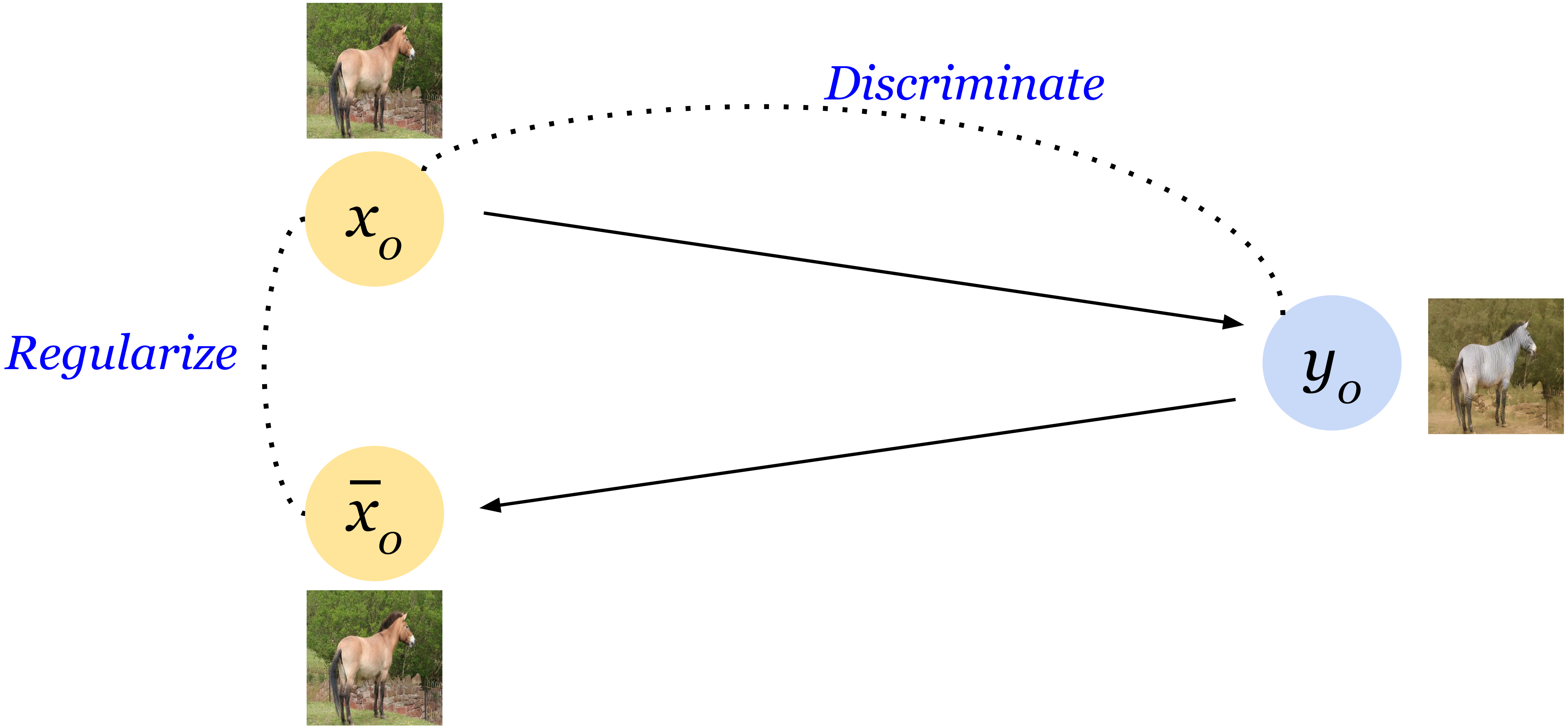
$$\bar{z}_0 = G(z_t, c_{\text{text}}, c_{\text{img}}) := \left[z_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_\theta(z_t, c_{\text{text}}, c_{\text{img}})\right] / \sqrt{\bar{\alpha}_t}$$

  - We adopt ControlNet (Zhang et al., 2023) as the adapter for conditioning.
  - We keep the conditional prompt in the frozen SD encoder and the unconditional prompt in the ControlNet, so that the LDM backbone focuses on the translation and the side network looks for the semantics that needs modification.



Method

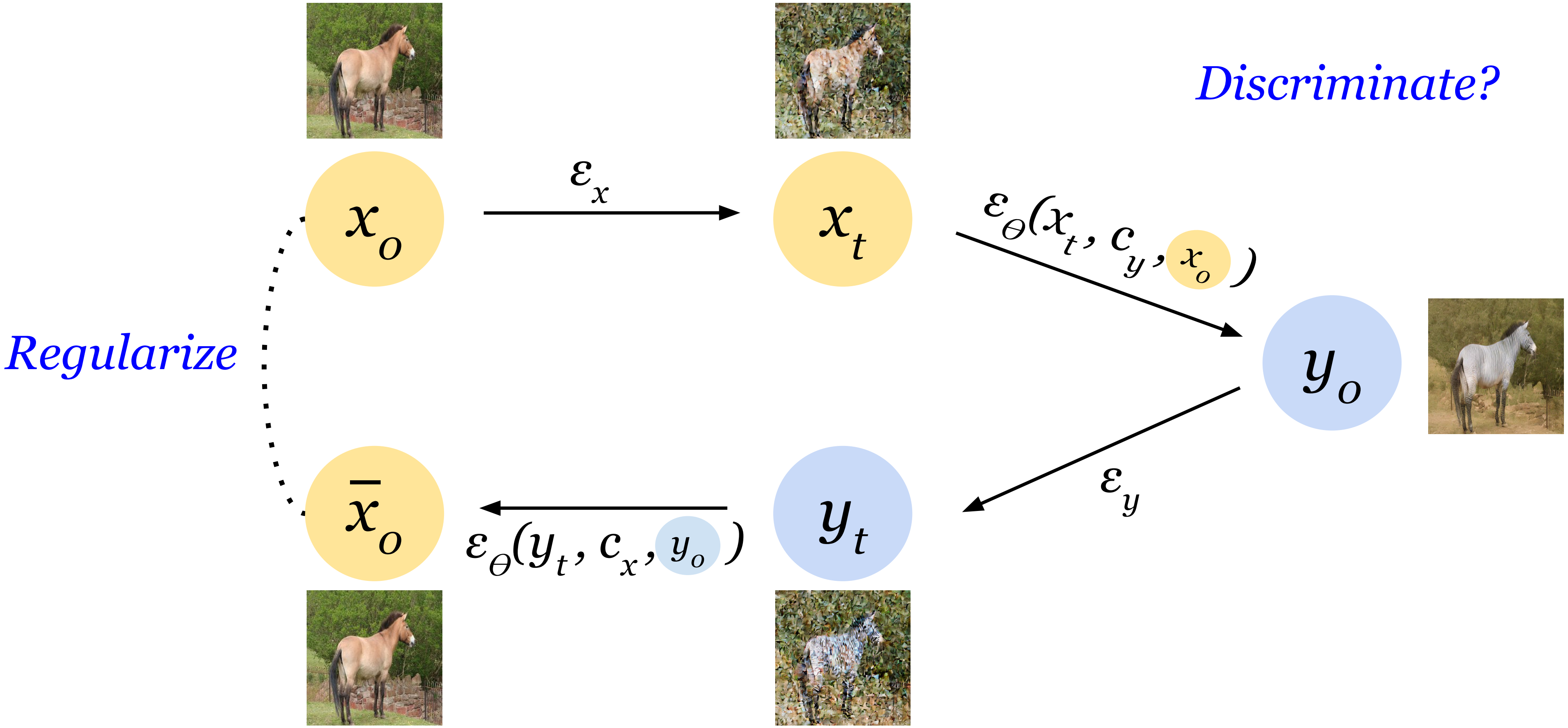
- **Cycle Consistency for GAN-based Unpaired I2I Translation**



CycleGAN (Zhu et al., 2017)

Method

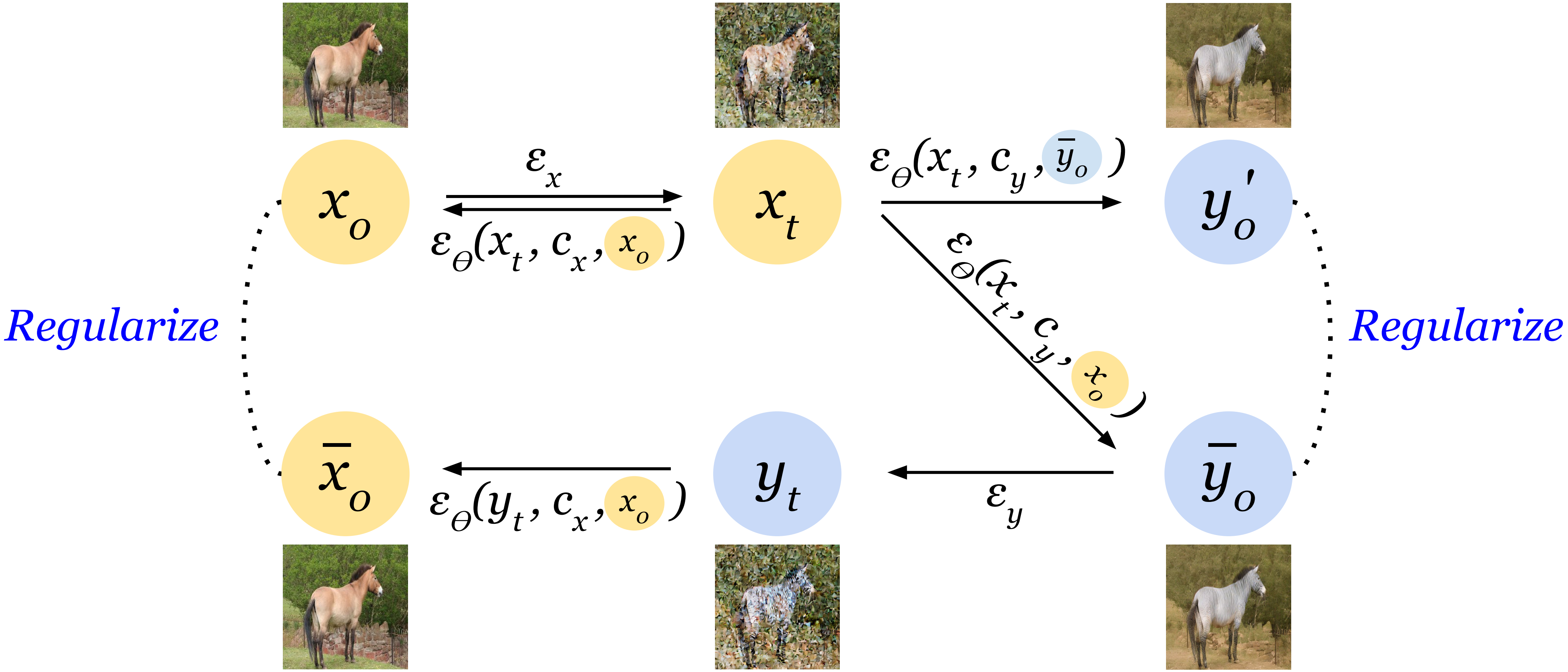
- **Cycle Consistency for Diffusion-based Unpaired I2I Translation?**



CycleNet (Ours)

Method

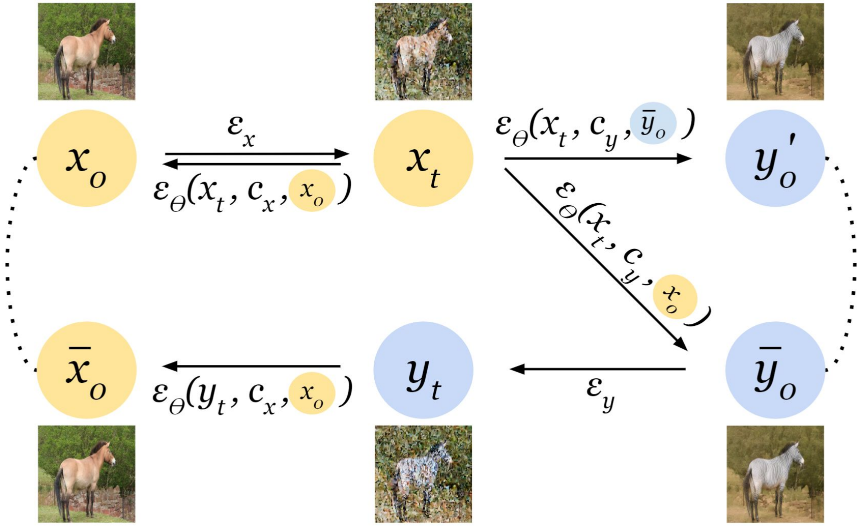
- **Cycle Consistency for Diffusion-based Unpaired I2I Translation?**



CycleNet (Ours)

Method

- **Cycle Consistency for Diffusion-based Unpaired I2I Translation**
  - With the translation cycle, a set of consistency losses is given as follows (see paper for proofs).



$$\mathcal{L}_{x\to x} = \mathbb{E}_{x_0,\varepsilon_x} ||\varepsilon_\theta(x_t, c_x, x_0) - \varepsilon_x||_2^2$$

$$\mathcal{L}_{y\to y} = \mathbb{E}_{x_0,\varepsilon_x,\varepsilon_y} ||\varepsilon_\theta(y_t, c_y, \bar{y}_0) - \varepsilon_y||_2^2$$

$$\mathcal{L}_{x\to y\to x} = \mathbb{E}_{x_0,\varepsilon_x,\varepsilon_y} ||\varepsilon_\theta(y_t, c_x, x_0) + \varepsilon_\theta(x_t, c_y, x_0) - \varepsilon_x - \varepsilon_y||_2^2$$

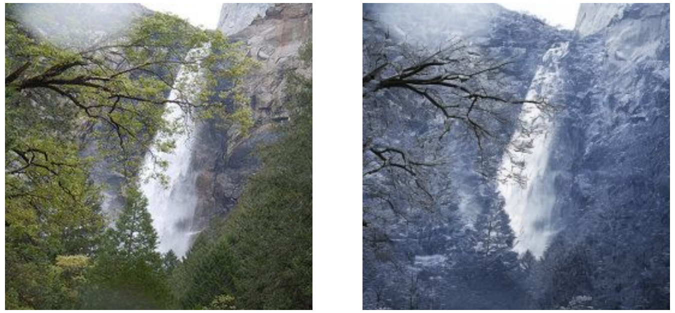
$$\mathcal{L}_{x\to y\to y} = \mathbb{E}_{x_0,\varepsilon_x} ||\varepsilon_\theta(x_t, c_y, x_0) - \varepsilon_\theta(x_t, c_y, \bar{y}_0)||_2^2$$

CycleNet (Ours)

Benchmarks

- **I2I Translation on Different Granularities:**
  - Global Scene level
  - Object type level

Summer <-> Winter



Apple <-> Orange

Horse <-> Zebra

Benchmarks

- **I2I Translation on Different Granularities:**
  - Object state level: ManiCups (this work), a dataset of state-level image manipulation that tasks models to manipulate cups by filling or emptying liquid to/from containers, formulated as a multi-domain I2I translation dataset for object state changes.

Empty <->

Coffee

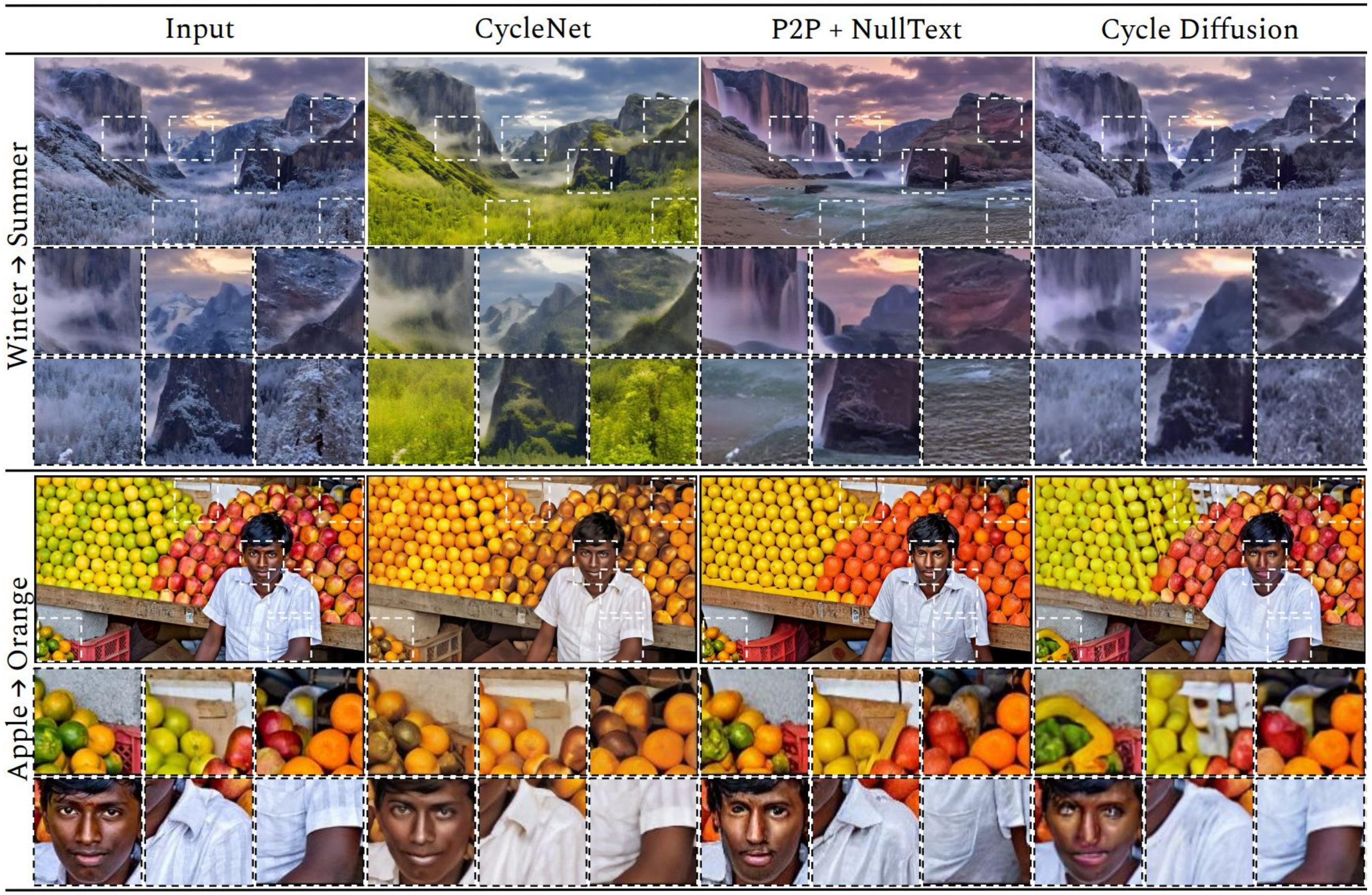
Juice

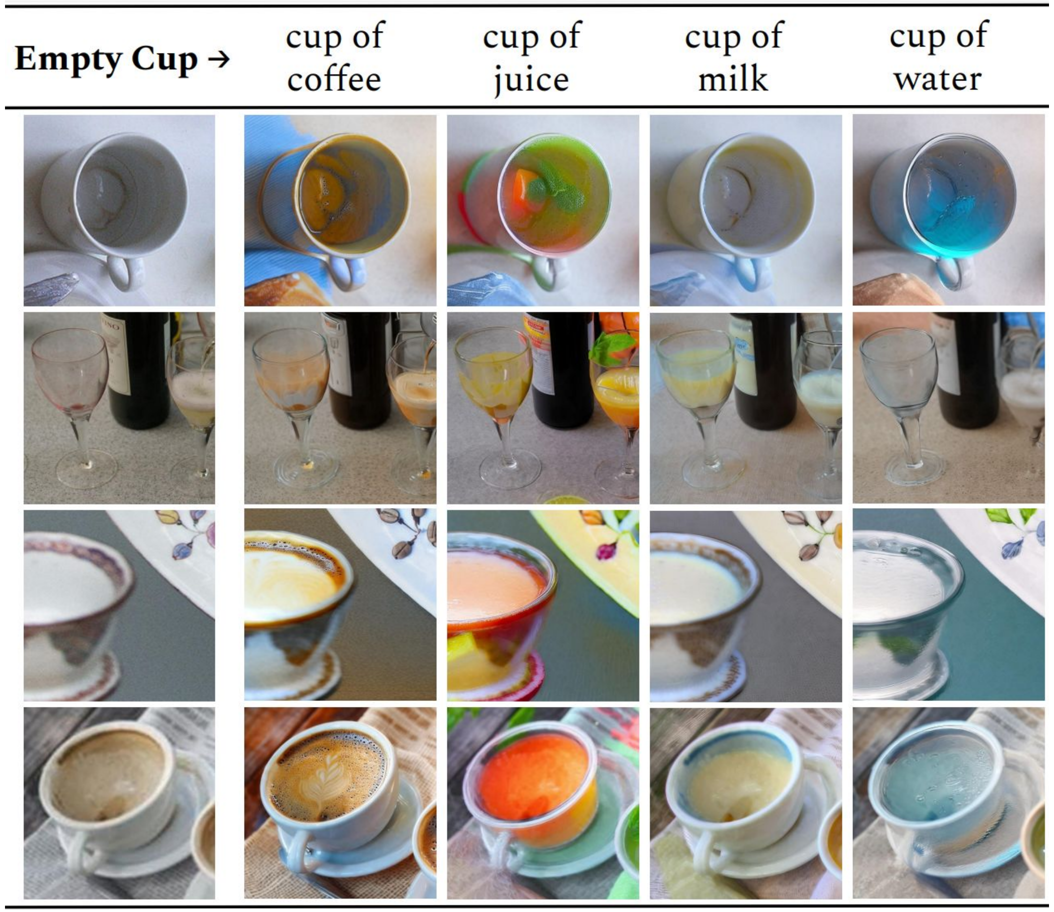
Milk

Water

Experiment and Results

- **Qualitative Examples**





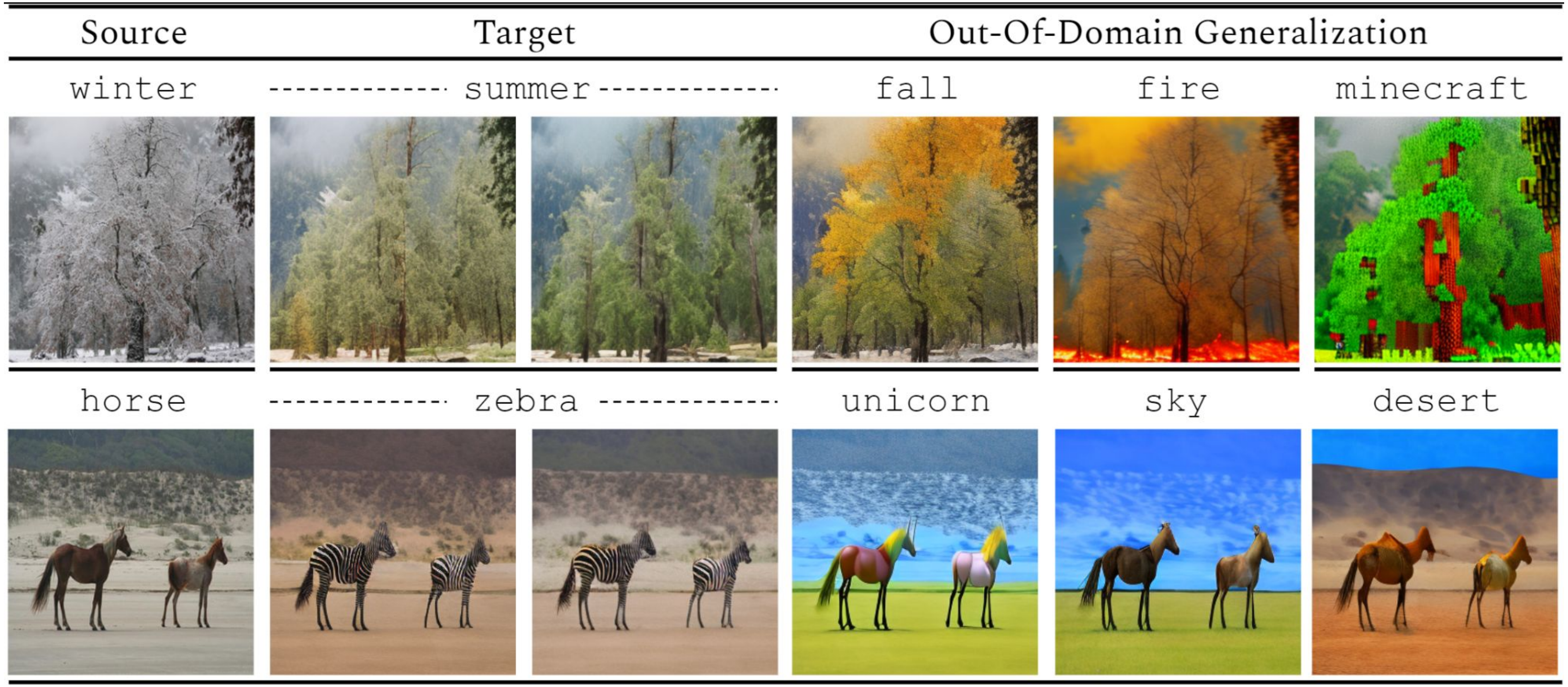
Experiment and Results

- **Quantitative Evaluation**
  - Image quality
  - Translation quality
  - Translation consistency

Tasks	summer→winter (Scene level, 256 × 256)							horse→zebra (Object level, 256 × 256)						
Metrics	FID↓	FID_{clip}↓	CLIP ↑	LPIPS↓	PSNR↑	SSIM↑	$L2^{\times 10^4}$ ↓	FID↓	FID_{clip}↓	CLIP ↑	LPIPS↓	PSNR↑	SSIM↑	$L2^{\times 10^4}$ ↓
GAN-based Methods														
CycleGAN	133.16	18.85	22.07	0.20	16.27	0.39	3.62	77.18	27.69	28.07	0.25	18.53	0.67	1.39
CUT	180.09	23.45	24.21	0.19	20.05	0.71	1.15	45.50	21.00	29.15	0.46	13.71	0.35	2.44
Mask-based Diffusion Methods														
Inpaint + ClipSeg	246.56	79.70	21.85	0.57	12.63	0.19	2.83	187.63	40.03	26.32	0.30	15.45	0.43	2.31
Text2LIVE	100.63	22.59	26.03	0.22	16.51	0.67	1.74	128.21	24.46	30.51	0.14	21.05	0.81	1.03
Mask-free Diffusion Methods														
ControlNet + Canny	338.24	83.26	21.77	0.59	6.05	0.09	11.30	397.71	77.68	23.88	0.61	7.37	0.07	3.89
ILVR	105.19	37.24	22.91	0.59	10.06	0.16	3.62	148.45	40.80	25.95	0.57	10.24	0.17	3.57
EGSDE	131.00	38.74	22.96	0.44	17.68	0.27	1.53	97.61	27.79	27.31	0.41	18.05	0.29	1.44
SDEdit	330.98	79.70	21.85	0.57	12.63	0.19	2.83	398.60	83.21	24.17	0.66	9.75	0.11	4.01
Pix2Pix-Zero	311.03	81.54	22.03	0.57	14.31	0.32	5.08	377.44	86.21	24.37	0.67	11.18	0.19	3.85
MasaCtrl	106.91	52.38	20.79	0.36	16.22	0.36	3.71	333.17	68.31	21.15	0.40	16.31	0.37	1.83
P2P + NullText	160.00	41.12	23.31	0.37	16.84	0.39	1.73	287.45	48.93	23.91	0.36	17.20	0.41	1.68
CycleDiffusion	243.98	62.96	22.32	0.44	15.06	0.31	2.20	347.27	66.80	25.04	0.57	11.51	0.21	3.46
FastCycleNet	**82.48**	17.61	**23.62**	0.14	**22.45**	**0.57**	0.91	**80.75**	**27.23**	27.36	0.32	19.29	0.51	1.31
CycleNet	82.52	**17.54**	23.32	**0.13**	22.42	**0.57**	**0.90**	81.69	28.11	**28.91**	**0.27**	**20.42**	**0.52**	**1.14**

Experiment and Results

- **Diversity and Generalization to Out-of-Distribution Domains**



Thank You!

Paper

Homepage

SLED Lab