ECCV 2024

Efficient Diffusion-Driven Corruption Editor for Test-Time Adaptation

Yeongtak Oh*, Jonghyun Lee*, Jooyoung Choi, Dahuin Jung, Uiwon Hwang† , Sungroh Yoon†

Data Science & Artificial Intelligence Laboratory
Electrical and Computer Engineering
Seoul National University

* Equal contribution, † Corresponding authors

Code : https://github.com/oyt9306/Decorruptor

TL;DR

- We propose image editing pipelines to revert unknown corrupted images into clean ones at test-time
- Performance
  - To robustify a diffusion model, we propose corruption modeling scheme for inductive fine-tuning
- Efficiency
  1. We leverage pixel space diffusion → latent space diffusion : Improve time & memory efficiency
  2. We distill diffusion model → consistency model : Improve time efficiency (20 steps -> ~4 steps)

Corrupted Reverted

Ours Decorruptor-DPM 20 NFEs

Ours Decorruptor-CM ~4 NFEs

Only Universal Prompt

Degradation: Snow, severity 5

DSAIL Data Science & AI Laboratory

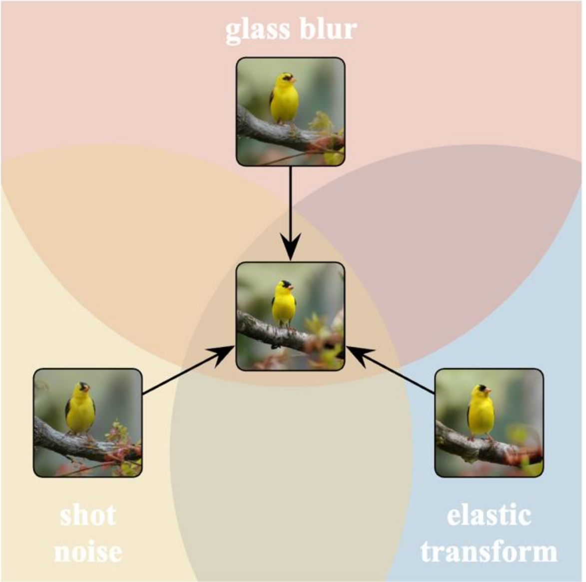
Contents

- **Introduction**
- **Related Works**
- **Proposed Method**
  - 1) Decorruptor-DPM
  - 2) Decorruptor-CM
- **Experimental Results**
  - 1) Qualitative Results
  - 2) Quantitative Results
  - 3) Further Analysis
- **Conclusions**

Introduction

- **Test-Time Adaptation (TTA)**
  - Training data: $\mathcal{D}^{tr} = \{x^{tr}, y^{tr}\}$, Test data: $\mathcal{D}^{te} = \{x^{te}, y^{te}\}$
  - However, y^{te} is not accessible and the source model is pre-trained only with $\mathcal{D}^{tr}$
  - Using small-set of $\mathcal{D}^{te}$, TTA aims to efficiently boost the model performance at test-time

- **Diffusion-based TTA Methods** * NFE : Neural Function Evaluation
  - DDA aims to update the input images from all targets to the source domain
  - DDA enables direct use of the source classifier without adaptation
  - ➢ It only supports pixel-level, large NFEs to revert the corrupt images



Introduction

- **Diffusion-based Image Editing Methods**
  - InstructPix2Pix(IP2P) enables image editing through instruction fine-tuning
  - However, SD-based <u>IP2P cannot generate de-corrupted (test-time corrupted) image</u>
  - Thus, we <u>robustify diffusion models</u> to be ready for the incoming unknown corruptions at the test-time

* SD: Stable Diffusion

(a) Instruction-based Editing

(b) Limitations

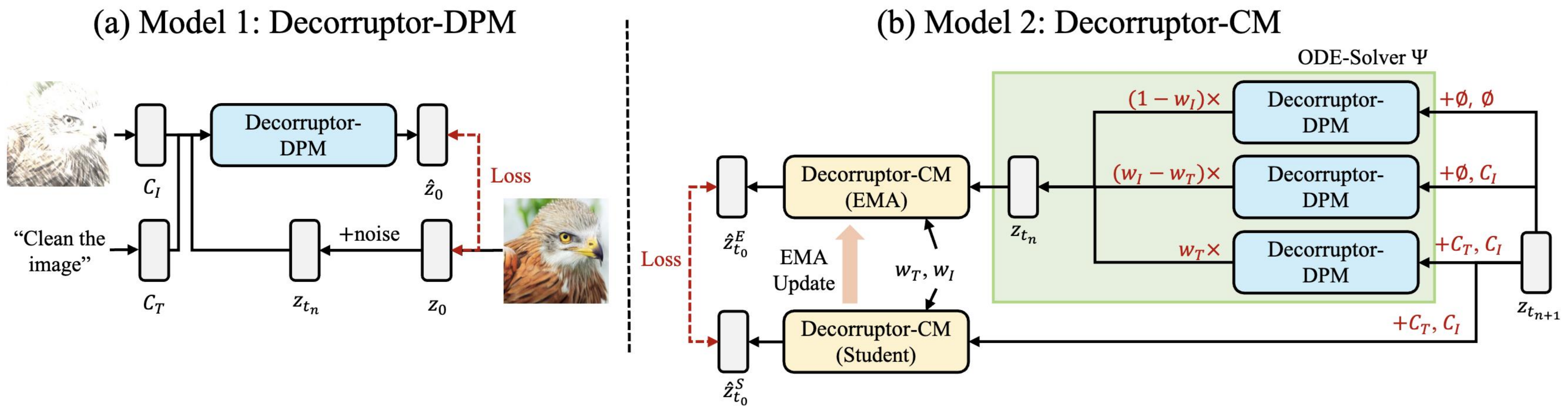
- **Comparisons with Other Diffusion-based Variants**
  - Editing) We enable image editing for incoming unknown corruption
  - IR) We do not require any pre-defined corruptions or degradation kernels at test time
  - TTA) We support highly efficient instant decorruption

TTA requirements		Image Editing	Image Reconstruction		Image Decorruption	
		InstructPix2Pix [5]	DDRM [27]	DPS [8]	DDA [14]	Ours (DPM / CM)
Efficiency (Minimal overhead)	NFEs	20	20	1000	50	20 / 4
	Noise space	Latent space	Pixel space	Pixel space	Pixel space	Latent space
Generalization	Degradation type	✗	Pre-defined	Pre-defined	Unseen	Unseen
Performance	IN-C Acc (%)	✗	✗	✗	29.7	30.5 / **32.8**
	IN-$\bar{\text{C}}$ Acc (%)	✗	✗	✗	29.4	41.8 / **47.1**

DSAIL Data Science & AI Laboratory

Proposed Method : Overview of Decorruptor

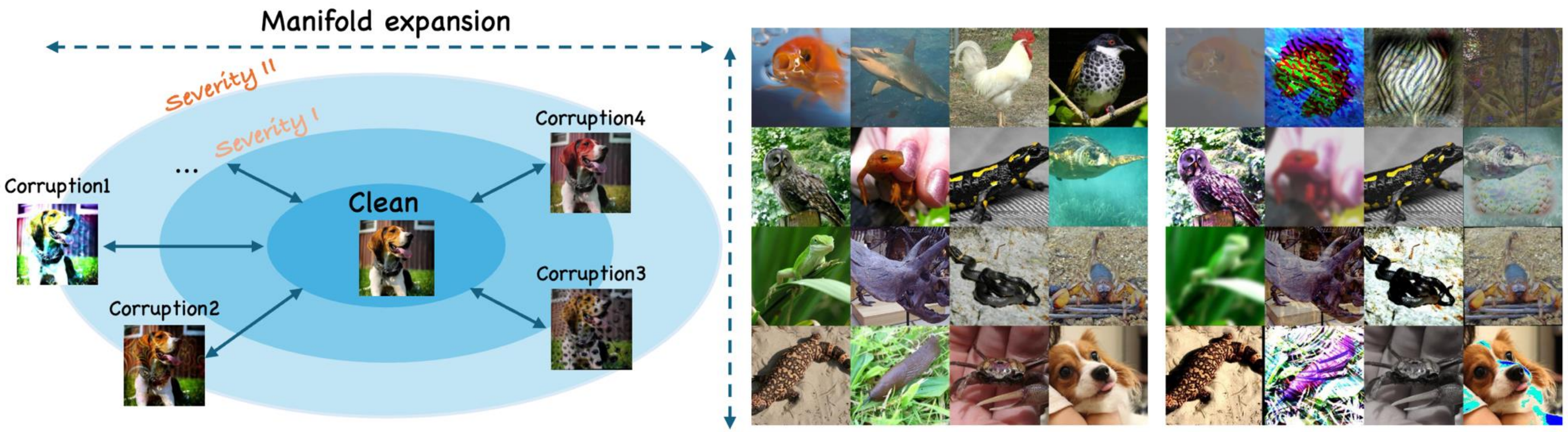
- Overall Pipelines



- Key Takeaways
  a) We adopt *IP2P pipeline to receive both text and image inputs*
  b) Universal prompt remains fixed *during both training/inference phases*
  c) Decorruptor-DPM/CM model can be considered as an *img2img translation model*

Proposed Method : Decorruptor-DPM

- Corruption Modeling Scheme



(a) Connection Modeling with Various Corruptions (b) Clean Images (c) Corrupted Images

- Key Takeaways
  a) We perform data augmentation in an *on-the-fly manner with class-agnostic images* (i.e., fractals)
  b) This robustification of diffusion models has not been previously explored (using PIXMIX, SimSiam aug)
  c) We broaden diffusion model's manifold by *fine-tuning to edit corrupted inputs into clean ones*

DSAIL Data Science & AI Laboratory

PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures (CVPR 2022)
Exploring Simple Siamese Representation Learning (CVPR 2021)

Proposed Method : Decorruptor-DPM

- Scheduling Image Guidance Scale

Training $$\mathcal{L}(\theta) = \mathbb{E}_{z\sim\mathcal{E}(x),c_T,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c_T, c_I)\|^2\right]$$

Inference $$\begin{aligned}\hat{\epsilon}_\theta(z_t, t, c_I, c_T) = \epsilon_\theta(z_t, t, \emptyset, \emptyset) + \omega_I(t)(\epsilon_\theta(z_t, t, c_I, \emptyset) - \epsilon_\theta(z_t, t, \emptyset, \emptyset)) \\ + \omega_T(\epsilon_\theta(z_t, t, c_I, c_T) - \epsilon_\theta(z_t, t, c_I, \emptyset)).\end{aligned}$$

- Key Takeaways
  a) At training time, we fine-tune U-Net of the diffusion model *initialized from the checkpoint of SD-v1.5*
  b) At inference time, we utilized *20 DDIM steps* and *sqrt noise scheduling* for the image guidance
  c) As we receive corrupted input images, ω_I is sampled from 1.8 to 0 for $t \in [T, 0]$

DSAIL Data Science & AI Laboratory

Proposed Method : Decorruptor-CM

- Integrate Multi-Modal Guidance while Diffusion Distillation

Training

$$\mathcal{L}_{LCD}(\theta, \theta^-; \Psi) = \mathbb{E}_{z_t,\omega,n}\left[d\left(f_\theta(z_{t_{n+1}}, \omega, c, t_{n+1}), f_{\theta^-}(\hat{z}_{t_n}^{\Psi,\omega_I,\omega_T}, \omega, c, t_n)\right)\right]$$

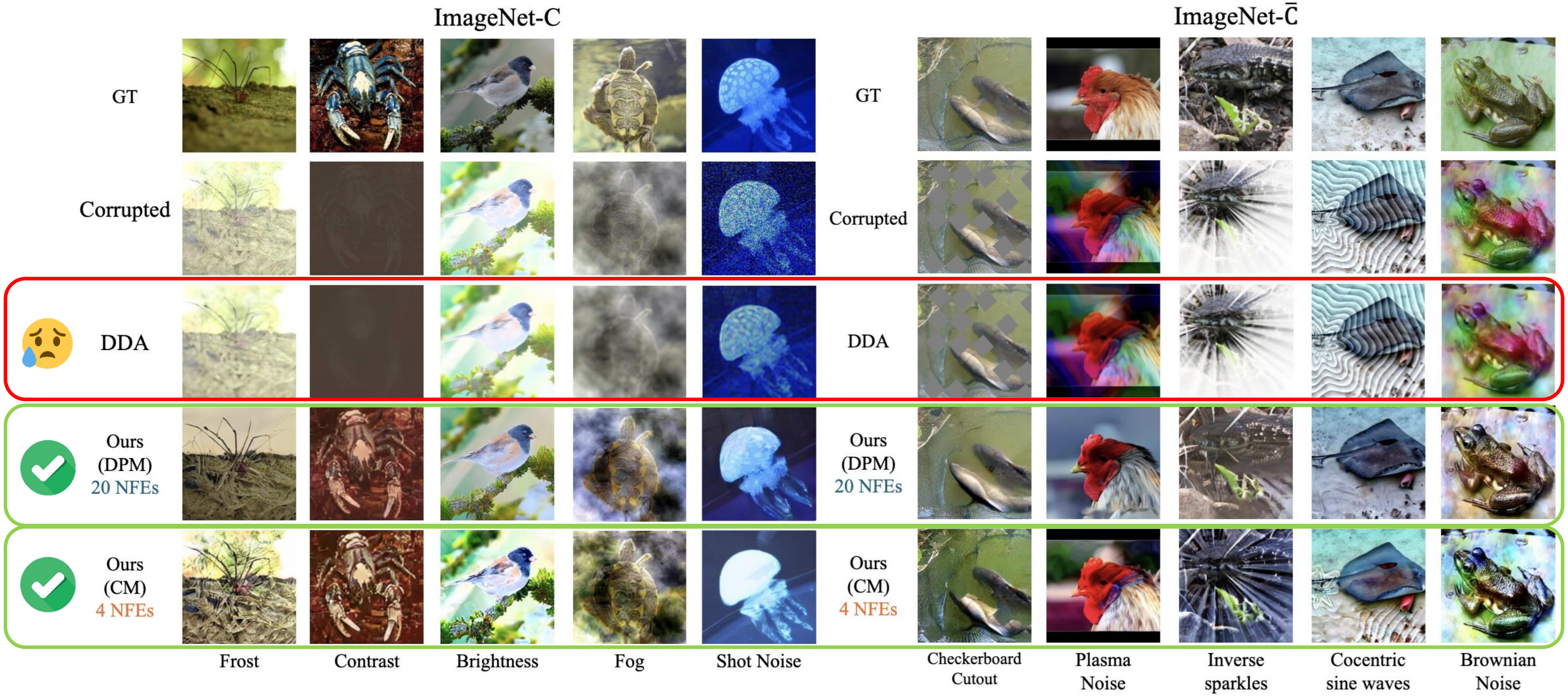
Inference

$$\begin{aligned}\hat{z}_{t_n}^{\Psi,\omega_I,\omega_T} - z_{t_{n+1}} &\approx \Psi(z_{t_{n+1}}, t_{n+1}, t_n, \emptyset, \emptyset) \\ &+ \omega_I(\Psi(z_{t_{n+1}}, t_{n+1}, t_n, c_I, \emptyset) - \Psi(z_{t_{n+1}}, t_{n+1}, t_n, \emptyset, \emptyset)) \\ &+ \omega_T(\Psi(z_{t_{n+1}}, t_{n+1}, t_n, c_I, c_T) - \Psi(z_{t_{n+1}}, t_{n+1}, t_n, c_I, \emptyset))\end{aligned}$$

- Key Takeaways
  a) We distill CM *on the same dataset (ImageNet)* employed during the DPM training phase
  b) At training time, we condition both learnable multi-modal guidance scales
  c) At inference time, multi-modal CFG scales ω_I and ω_T *obviates the need for guidance scheduling*

Experimental Results

- **Qualitative Results of Image Editing for Unknown Corruptions**



DSAIL Data Science & AI Laboratory

Experimental Results

- **Quantitative Results of TTA on Corruption Benchmarks**

140x faster!

Method	Runtime (s/sample)↓	Memory (MB)↓	IN-C Acc. (%)↑	IN-$\bar{C}$ Acc. (%)↑
MEMO	<u>0.41</u>	7456	24.7	-
DDA	19.5	10320 + 2340	29.7	29.4
Decorruptor-DPM	0.42	**4602 + 2340**	<u>30.5</u>	<u>41.8</u>
4×Decorruptor-CM	**0.14**	4958 + 2383	**32.8**	**47.1**

Method	ResNet-50	Swin-T	ConvNeXt-T	Swin-B	ConvNeXt-B
Source-Only	18.7	33.1	39.3	40.5	45.6
MEMO (0.41s)	24.7	29.5	37.8	37.0	45.8
DiffPure (27.3s)	16.8	24.8	28.8	28.9	32.7
DDA (19.5s)	29.7	<u>40.0</u>	<u>44.2</u>	44.5	<u>49.4</u>
Decorruptor-DPM (0.42s)	30.5	37.8	42.2	42.5	46.6
4×Decorruptor-CM (0.14s)	<u>32.8</u>	39.7	44.0	<u>44.7</u>	48.6
8×Decorruptor-CM (0.25s)	**34.2**	**41.1**	**45.2**	**46.1**	**49.8**

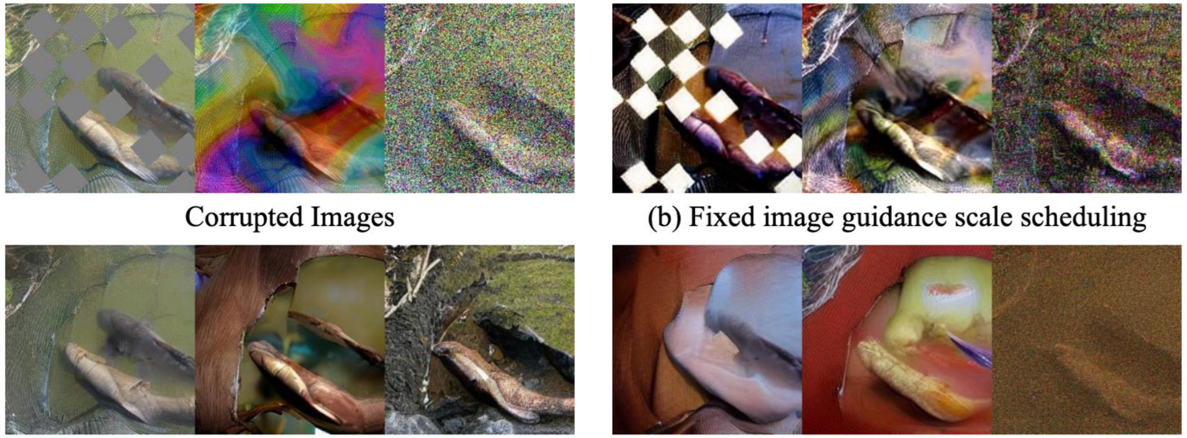
(a) ImageNet-C

- **Performances on OOD Datasets**

	Source	+ DDA	+ DPM	+ 4×CM	PIXMIX	+ DDA	+ DPM	+ 4×CM
VISDA-2021 acc (%)	35.7	40.2	<u>40.9</u>	**42.0**	44.0	45.4	<u>45.6</u>	**46.1**
ImageNet-A acc (%)	0.0	0.5	<u>1.9</u>	**2.7**	6.3	5.2 (-1.1)	<u>8.1</u>	**9.8**

DSAIL Data Science & AI Laboratory

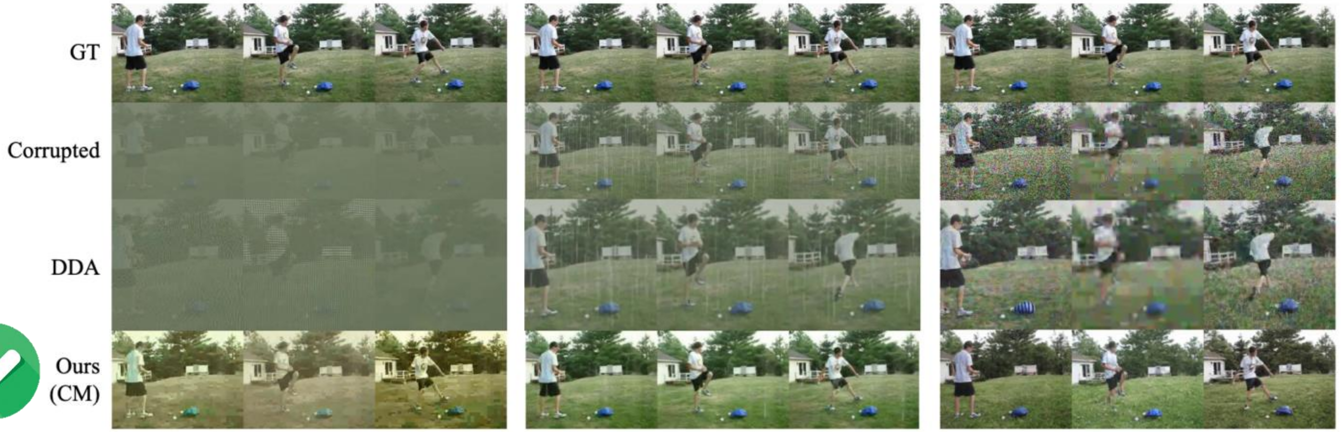
- **Further Analysis**
  a) Effects of Our Multi-Modal Guidance Scaling on Consistency Model



Corrupted Images | (b) Fixed image guidance scale scheduling

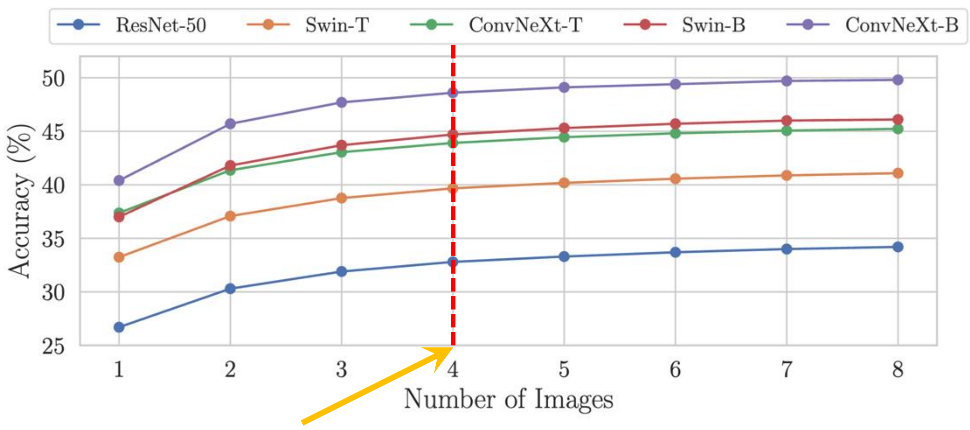
(a) Ours (w/ multi-modal guidance conditioning) | (c) Not using image guidance scale scheduling

  b) Video Decorruption Results

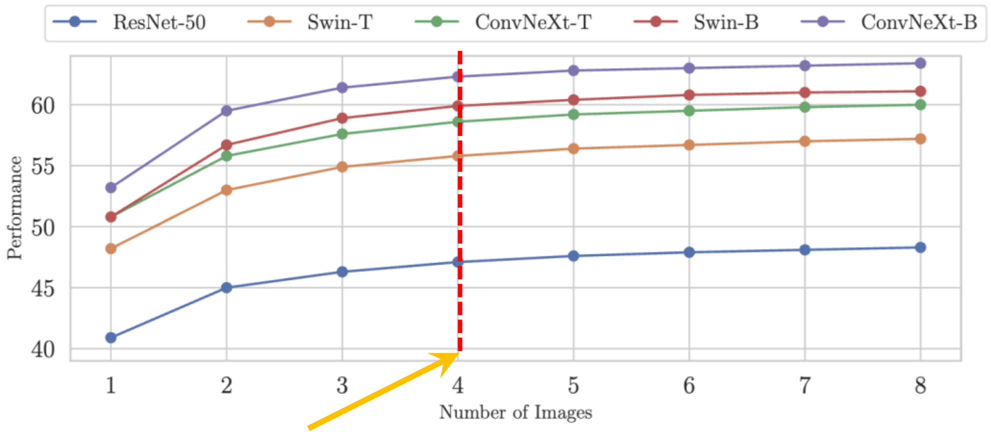


(a) Contrast | (b) Rain | (c) Gaussian

- **Further Analysis**
  - c) Ensemble Stabilities: The number of ensembles to inference on TTA benchmarks
    - ➢ The more ensembling samples, we can get the higher accuracies
    - ➢ We set 4 as the default number of ensembling samples



(a) ImageNet-C



(b) ImageNet-$\bar{C}$

DSAIL
Data Science & AI Laboratory

Conclusions

- We propose Decorruptor-DPM, a latent diffusion model with *efficient memory and time utilization*
  - Revamp previous methods impractical for real-world usage due to its slow processing speed

- We introduce Decorruptor-CM, employing consistency distillation *to accelerate input updates further*
  - Surpassing the baseline diffusion-based approach in speed by *100 times* while *delivering superior performance*.

- We expect our novel corruption editing pipeline provides *new insights for image-update TTA*

See you at Poster Session 4!

Wed 2 Oct 11:30 p.m. KST — 1:30 a.m.

< Project Page >