Paper Review

Palette: Image-to-Image Diffusion Models

YeongHyeon Park

Department of Electrical and Computer Engineering

SungKyunKwan University

Some figures and illustrations included in this presentation are edited from the original.
The reference for the original source is located in the upper right corner of each page.

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

Google Research

Palette: Image-to-Image Diffusion Models

Chitwan Saharia
Google Research, Brain Team
Toronto, ON, Canada
sahariac@google.com

William Chan
Google Research, Brain Team
Toronto, ON, Canada
williamchan@google.com

Huiwen Chang
Google Research
New York, NY, USA
huiwenchang@google.com

Chris A. Lee
Google Research
Mountain View, CA, USA
chrisalee@google.com

Jonathan Ho
Google Research, Brain Team
New York, NY, USA
jonathanho@google.com

Tim Salimans
Google Research, Brain Team
Amsterdam, Netherlands
salimans@google.com

David Fleet
Google Research, Brain Team
Toronto, ON, Canada
davidfleet@google.com

Mohammad Norouzi
Google Research, Brain Team
Toronto, ON, Canada
mnorouzi@google.com

SIGGRAPH 2022
VANCOUVER+ 8-11 AUG

Google Research

Palette: Image-to-Image Diffusion Models

Chitwan Saharia
Google Research, Brain Team
Toronto, ON, Canada
sahariac@google.com

William Chan
Google Research, Brain Team
Toronto, ON, Canada
williamchan@google.com

Huiwen Chang
Google Research
New York, NY, USA
huiwenchang@google.com

Chris A. Lee
Google Research
Mountain View, CA, USA
chrisalee@google.com

Jonathan Ho
Google Research, Brain Team
New York, NY, USA
jonathanho@google.com

Tim Salimans
Google Research, Brain Team
Amsterdam, Netherlands
salimans@google.com

David Fleet
Google Research, Brain Team
Toronto, ON, Canada
davidfleet@google.com

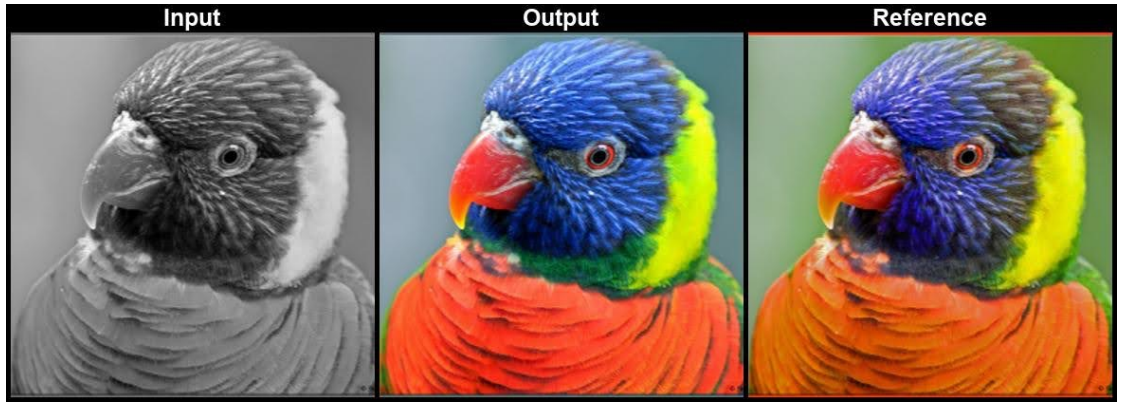
Mohammad Norouzi
Google Research, Brain Team
Toronto, ON, Canada
mnorouzi@google.com

Ho et al. 'Denoising Diffusion …'
NeurIPS, 2020

Preview of Palette

https://diffusion-palette.github.io/

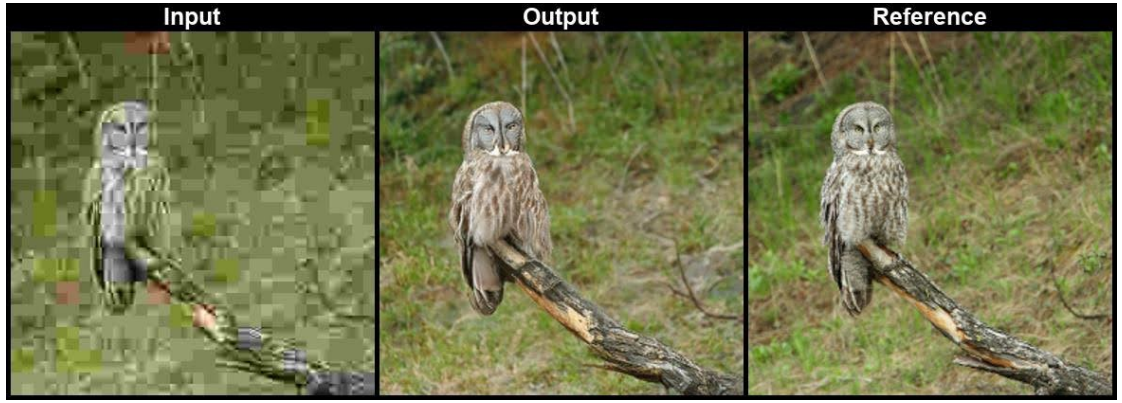
Image-to-image Translation



Colorization

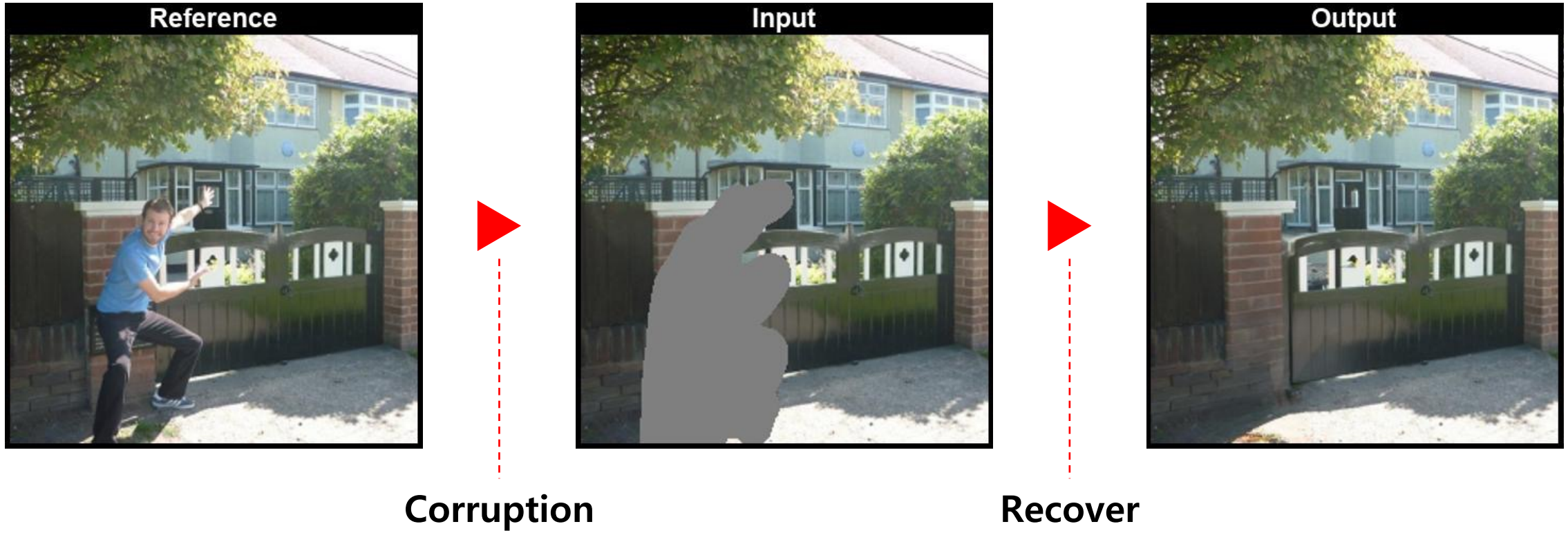
Inpainting

Uncropping (=outpainting)



JPEG restoration

How to Translate?



Contribution: Unified X

- A unified diffusion-based image-to-image translation (i2i) framework for four tasks.
  - No hyperparameter tuning required
  - No task-specific neural networks required

- A unified evaluation protocol on i2i translation tasks based on visual perception.
  - Inception Score (IS)
  - Fréchet Inception Distance (FID)
  - Classification Accuracy (CA)
  - Perceptual Distance (PD)

Why Diffusion?

Xiao, Zhisheng et al. "Tackling the generative learning trilemma with denoising diffusion GANs." *arXiv preprint*. 2021.

Why Diffusion?

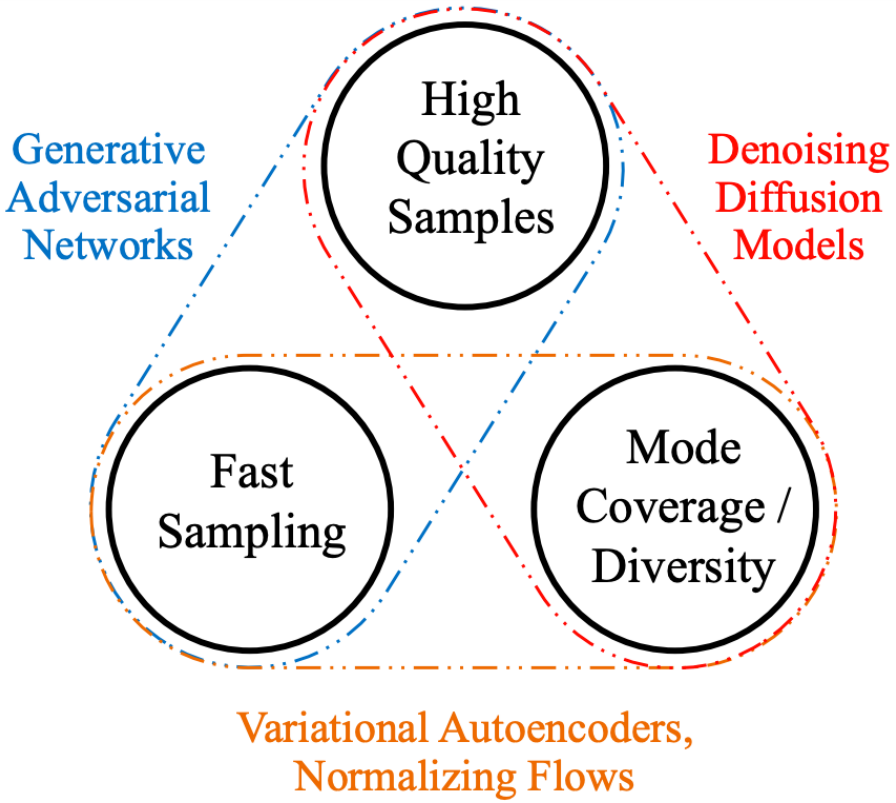


Figure 1: Generative learning trilemma.

Xiao, Zhisheng et al. "Tackling the generative learning trilemma with denoising diffusion GANs." *arXiv preprint*. 2021.

Why Diffusion?

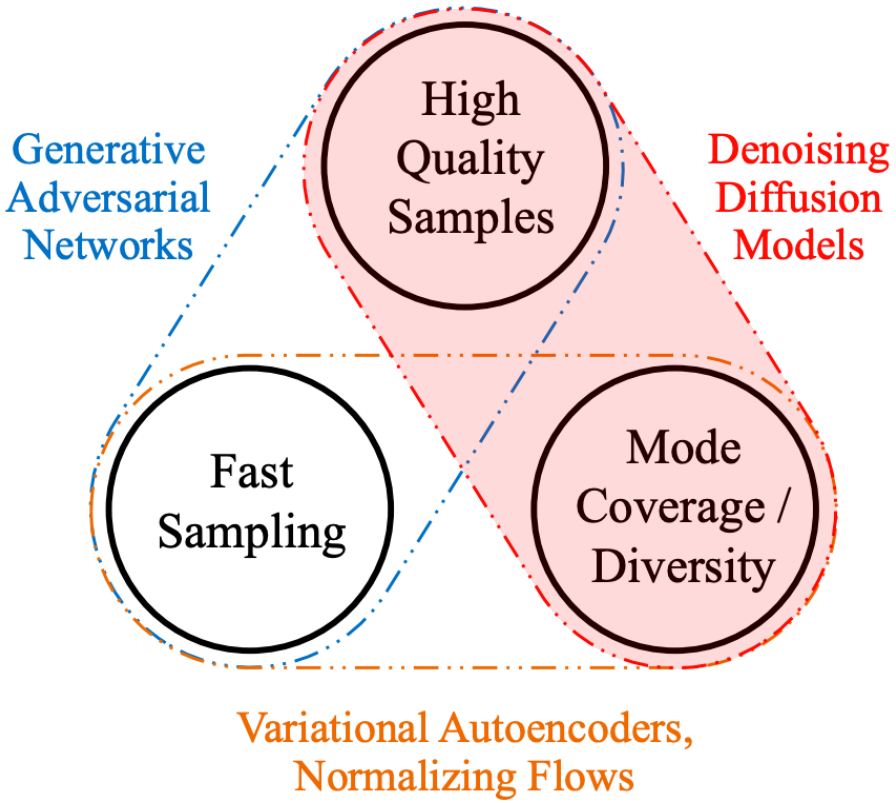


Figure 1: Generative learning trilemma.

Xiao, Zhisheng et al. "Tackling the generative learning trilemma with denoising diffusion GANs." *arXiv preprint*. 2021.

Mode Coverage / Diversity

- **Coverage** : Infer each mode without missing.
- **Diversity** : Infer all the different modes without exception.

Diffusion

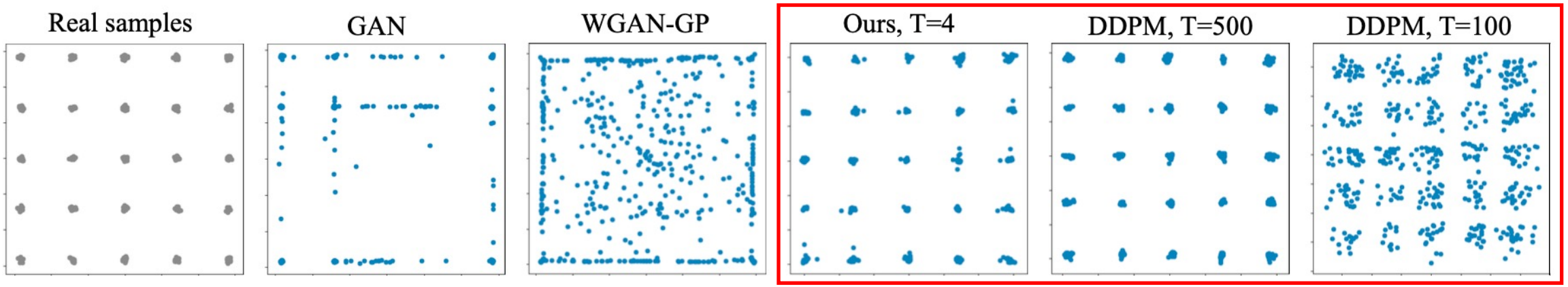
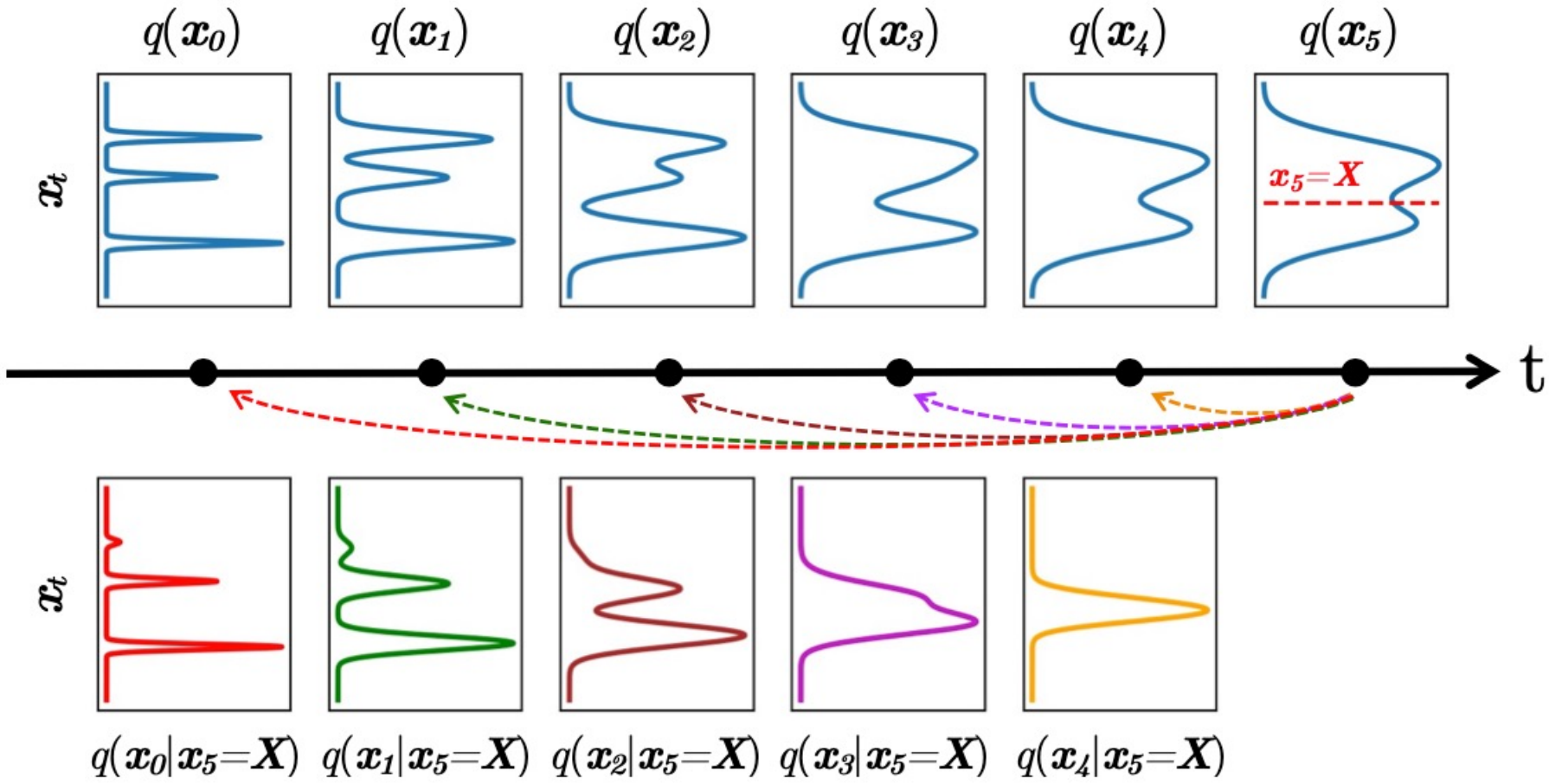


Figure 6: Qualitative results on the 25-Gaussians dataset.

Xiao, Zhisheng et al. "Tackling the generative learning trilemma with denoising diffusion GANs." *arXiv preprint*. 2021.

Diffusion Model

- x_0 : original sample
- x_t : noise injected sample (diffused sample)

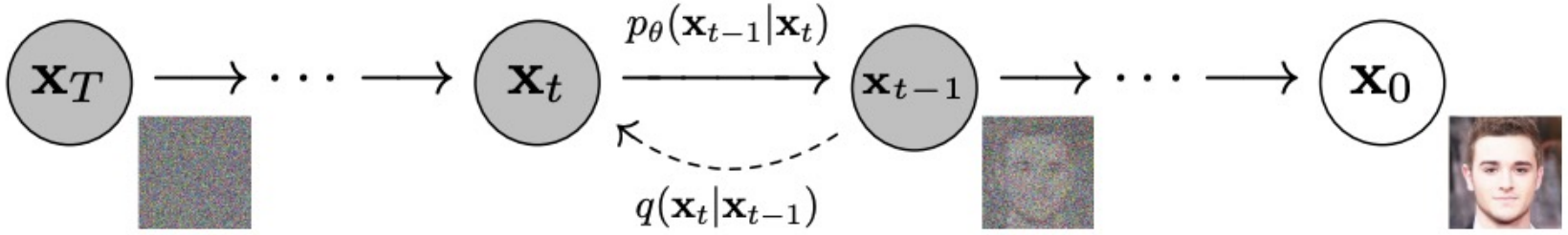


Ho, Jonathan et al. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

Denoising Diffusion (Ho et al, 2020)

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Reverse Denoising Process



Diffusion Process
(Noise Injection)

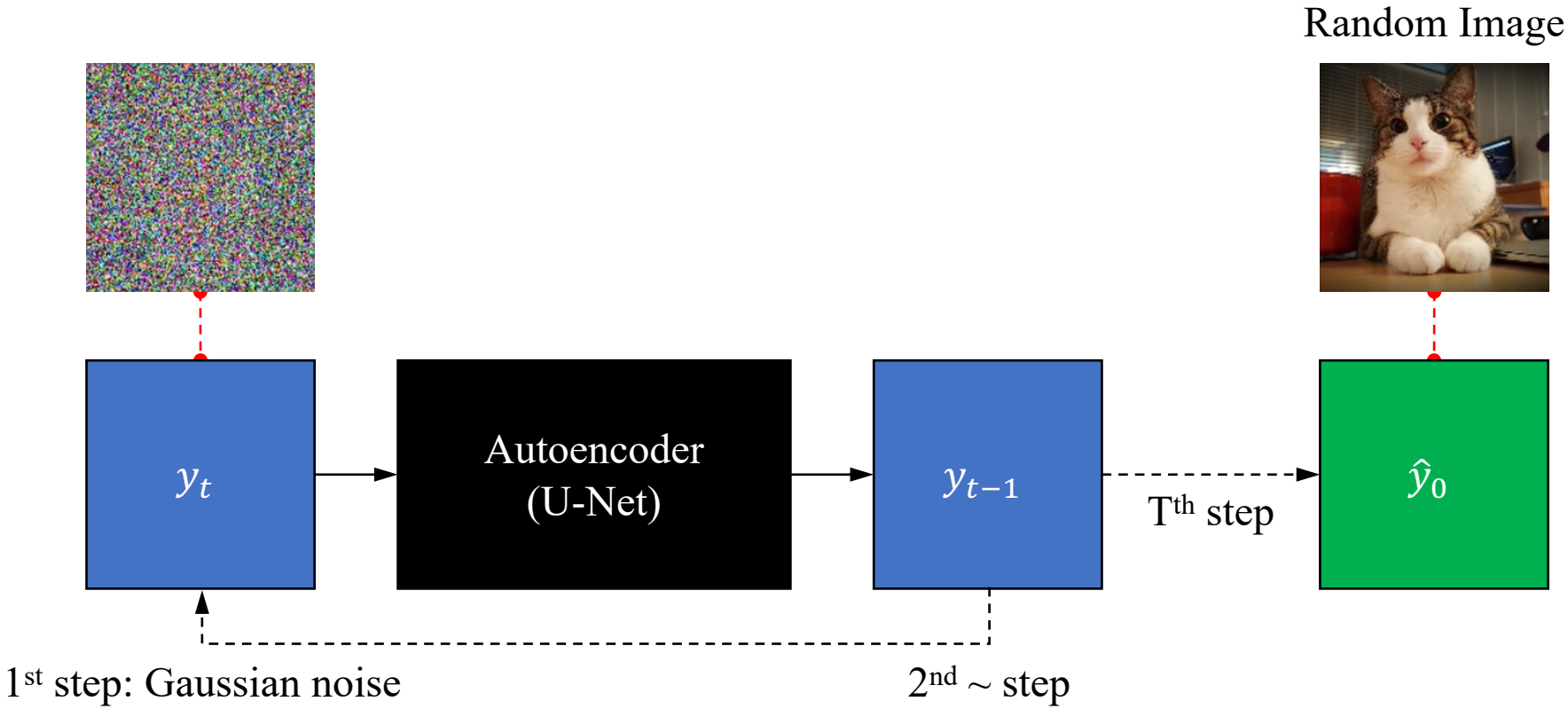
Gaussian Noise

Reference Image

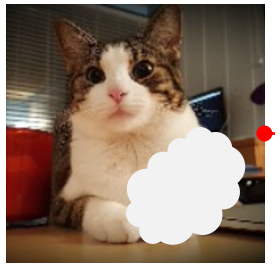
Palette

Image-to-Image Diffusion Models

Unconditional Diffusion



Conditional Diffusion



x

Condition

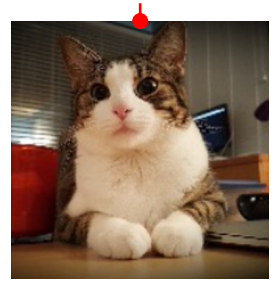
x

Autoencoder
(U-Net)

y_{t-1}

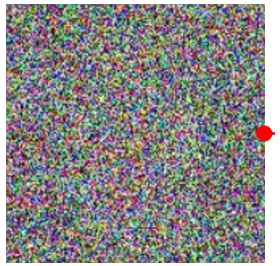
Tth step

$\hat{y}_0$



Conditional Output

2nd ~ step



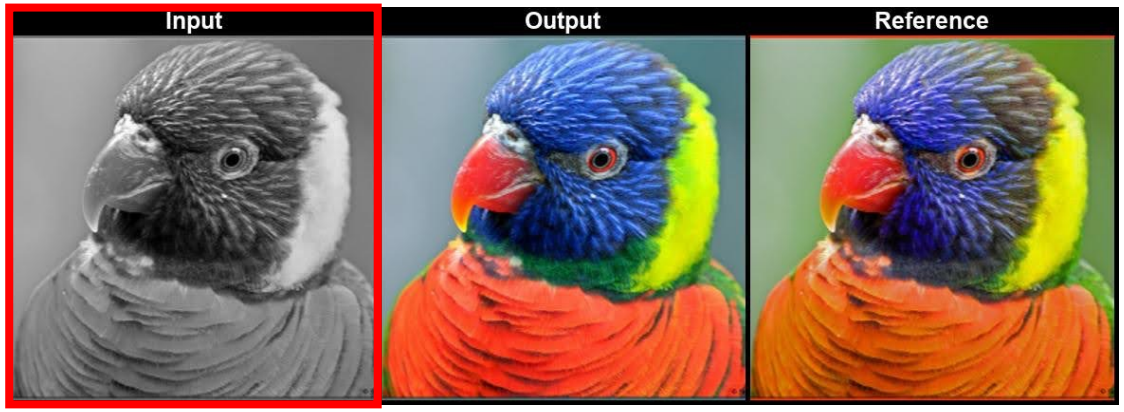
y_t

1st step: Gaussian noise

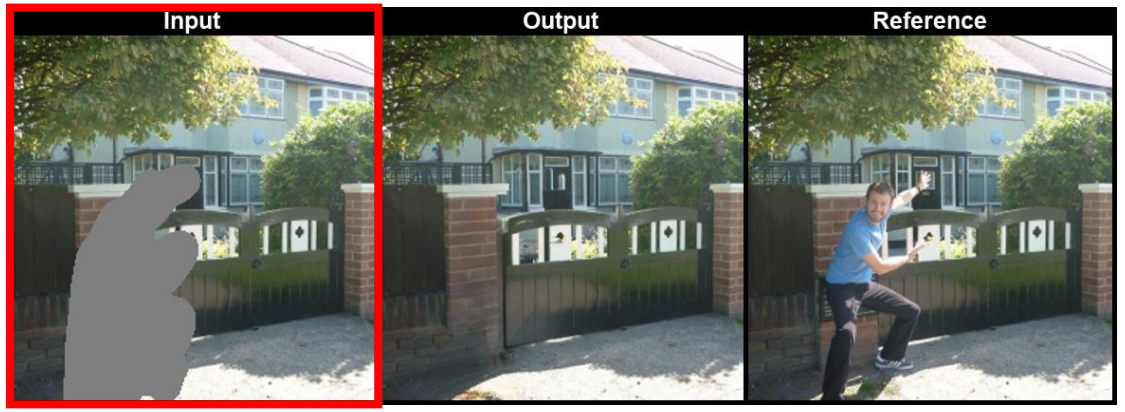
The only difference is the conditional information!

Image-to-image Translation

Conditional information in each task.

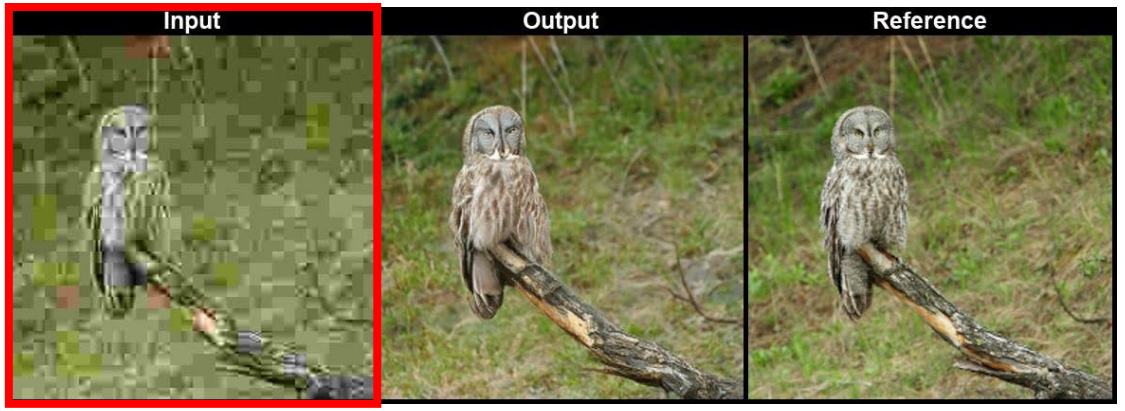


Colorization



Inpainting

Uncropping (=outpainting)



JPEG restoration

Evaluation Protocols

- **Fréchet Inception Distance (FID)** ↓
  - Measure the feature map similarity (e.g. covariance) between ground truth and generated sample.
  - Basically, a pre-trained neural network is used for measurement.

- **Perceptual Distance (PD)** ↓
  - Measure the feature map distance between ground truth and generated sample.
  - Basically, a pre-trained neural network is used for measurement.

- **Inception Score (IS)** ↑
  - Measure that the generated samples have both diversity and clarity.

- **Classification Accuracy (CA)** ↑
  - Confirm that the generated sample contains proper information to classify accurately.
  - The classification will be conducted with a pre-trained classifier.

Kumar, Manoj, Dirk Weissenborn, and Nal Kalchbrenner. "Colorization transformer." *International Conference on Learning Representations (ICLR)*. 2021.
Yu, Jiahui, et al. "Free-form image inpainting with gated convolution." *Proceedings of the IEEE/CVF international conference on computer vision (CVPR)*. 2019.
Teterwak, Piotr, et al. "Boundless: Generative adversarial networks for image extension." *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*. 2019.

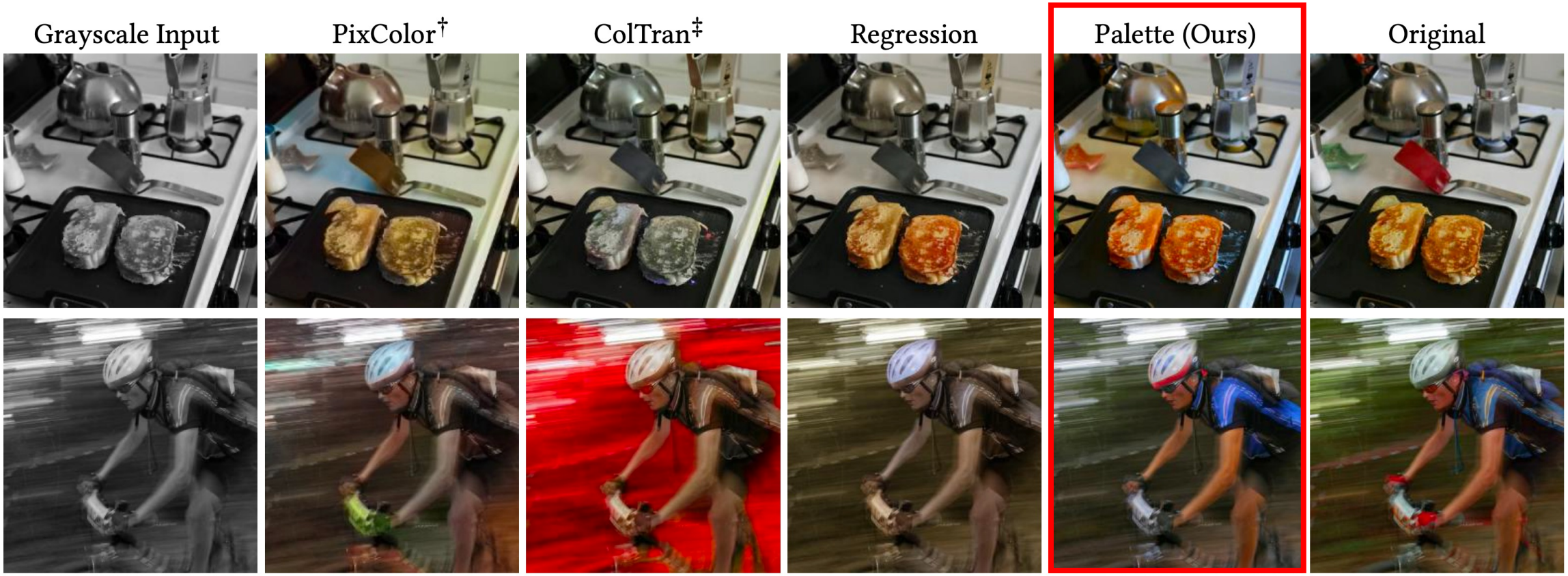
Summarized Performance

- **Task-specific** : Each model is trained only for the specific task.
- **Multi-task** : The unified model trained on all four tasks.

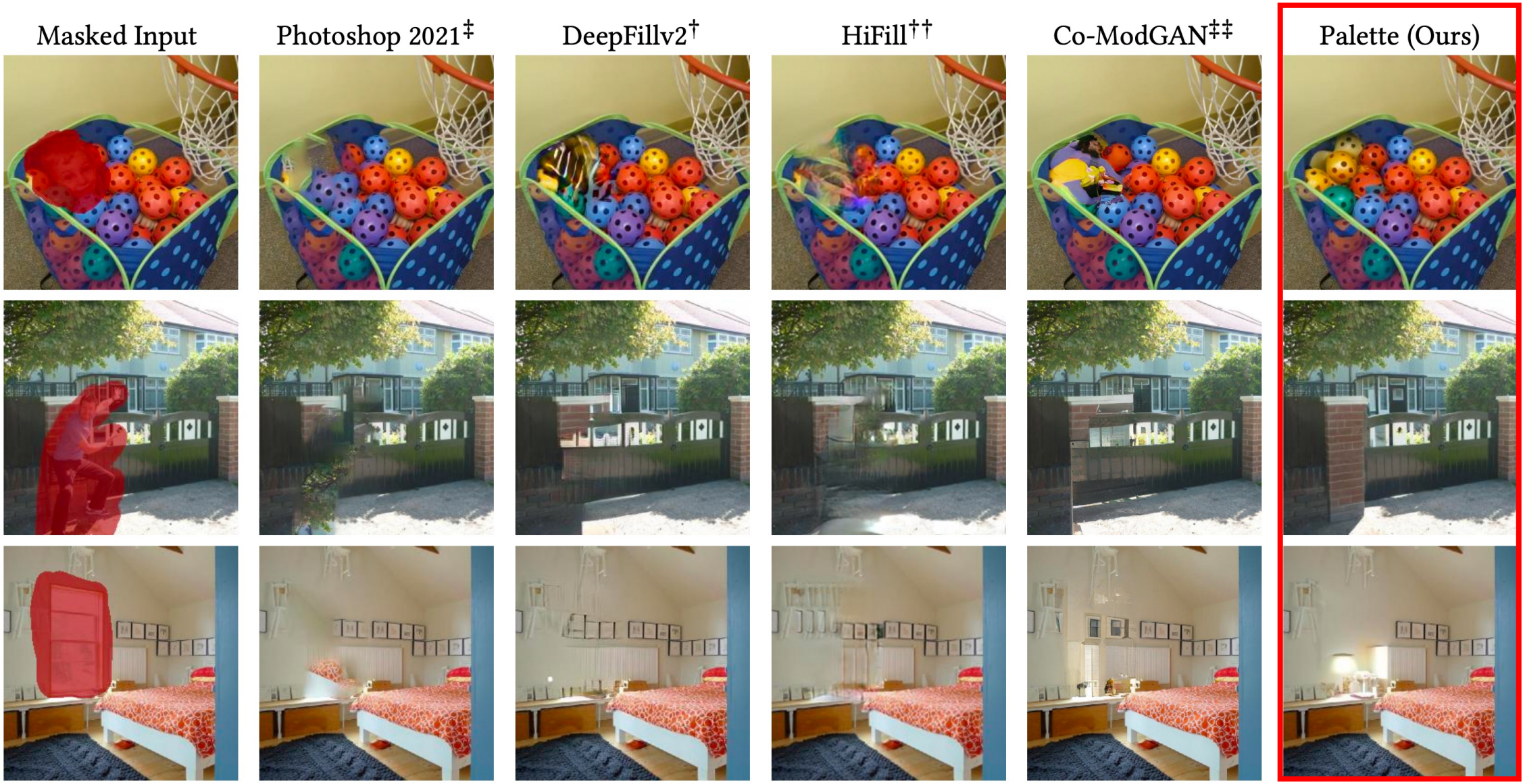
Summarized from Table 1, 2, 3, 4 and 7.

Task	Model	FID↓	PD↓	IS↑	CA↑
Colorization	Coltran	19.37	-	-	-
	Palette (Task-specific)	**3.4**	**48.0**	**212.9**	**72.0%**
	Palette (Multi-task)	3.7	57.1	187.4	69.4%
Inpainting	DeepFillv2	18.0	117.2	135.3	64.3%
	Palette (Task-specific)	**6.6**	**59.5**	**173.9**	**69.3%**
	Palette (Multi-task)	6.8	65.2	165.7	68.9%
Uncropping	Boundless	18.7	127.9	104.1	58.8%
	Palette (Task-specific)	**5.8**	**85.9**	**138.1**	**63.4%**
	Palette (Multi-task)	-	-	-	-
JPEG restoration (QF=5)	Regression	29.0	155.4	73.9	52.8%
	Palette (Task-specific)	8.3	95.5	133.6	64.2%
	Palette (Multi-task)	**7.0**	**92.4**	**137.8**	**64.7%**

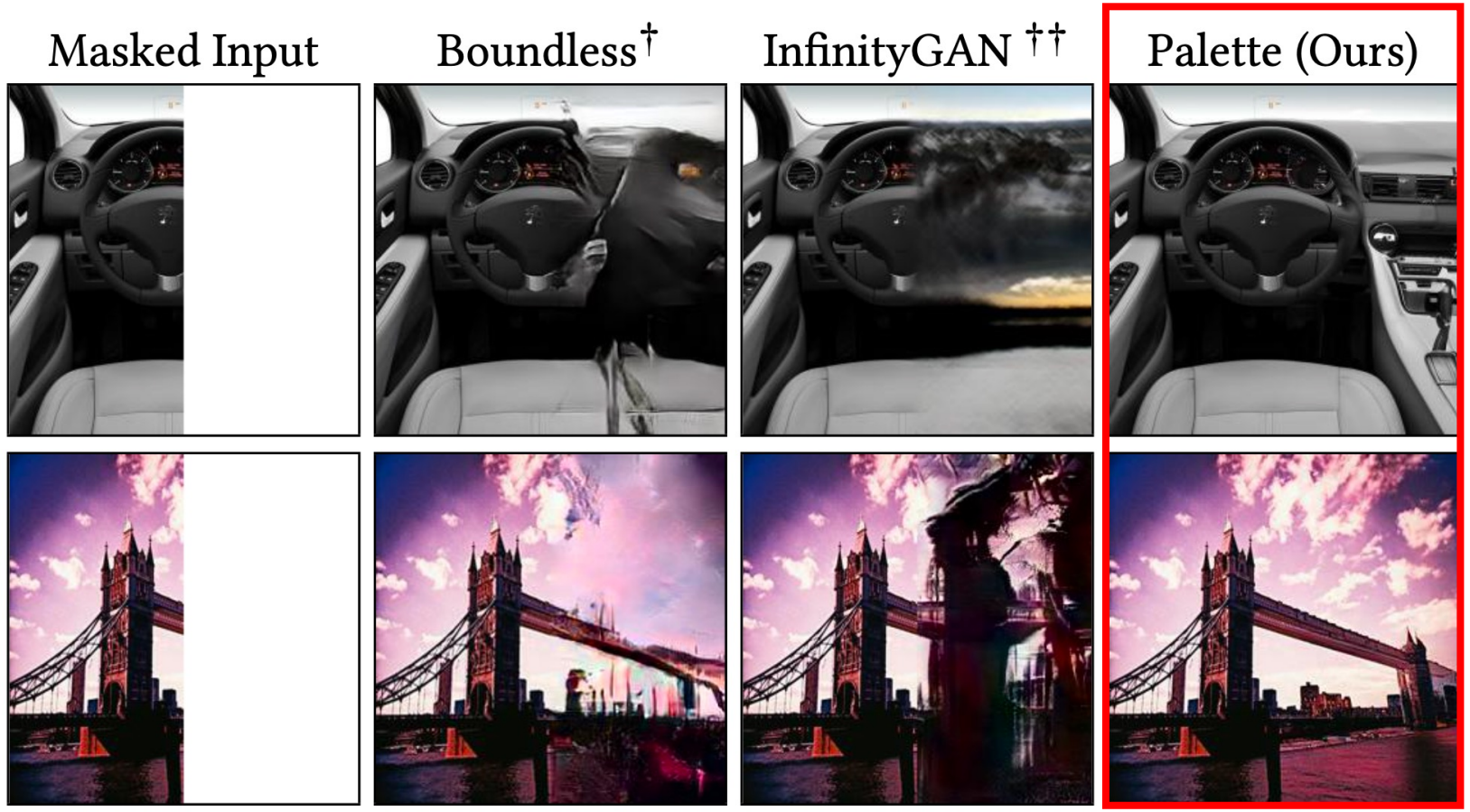
Quality – Colorization



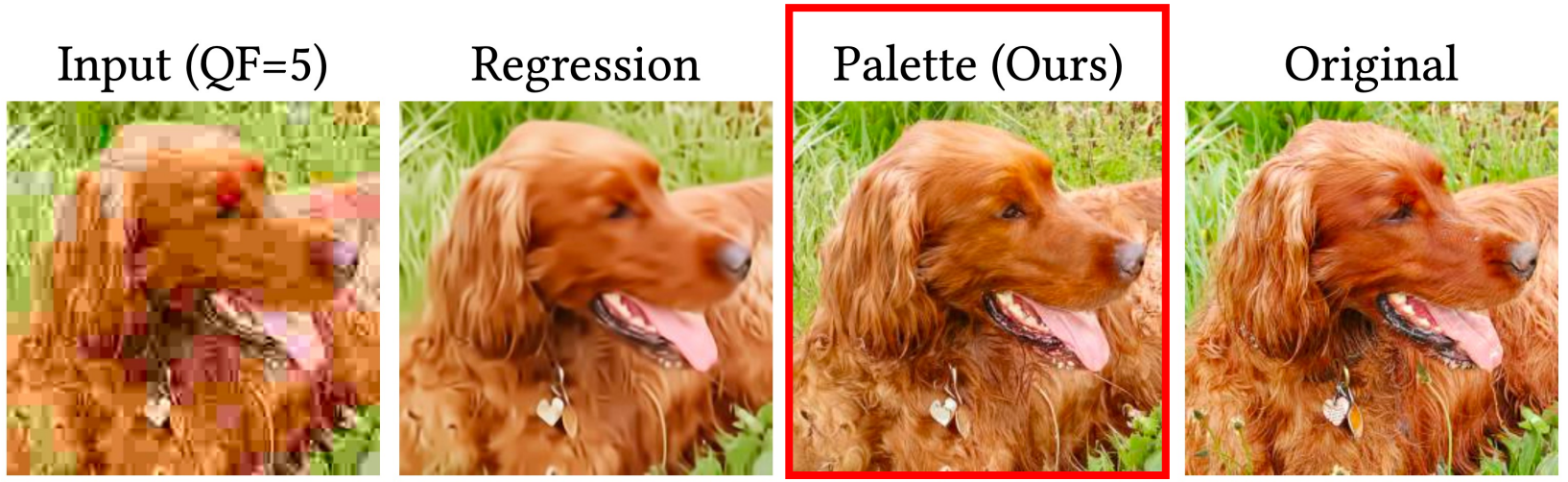
Quality – Inpainting



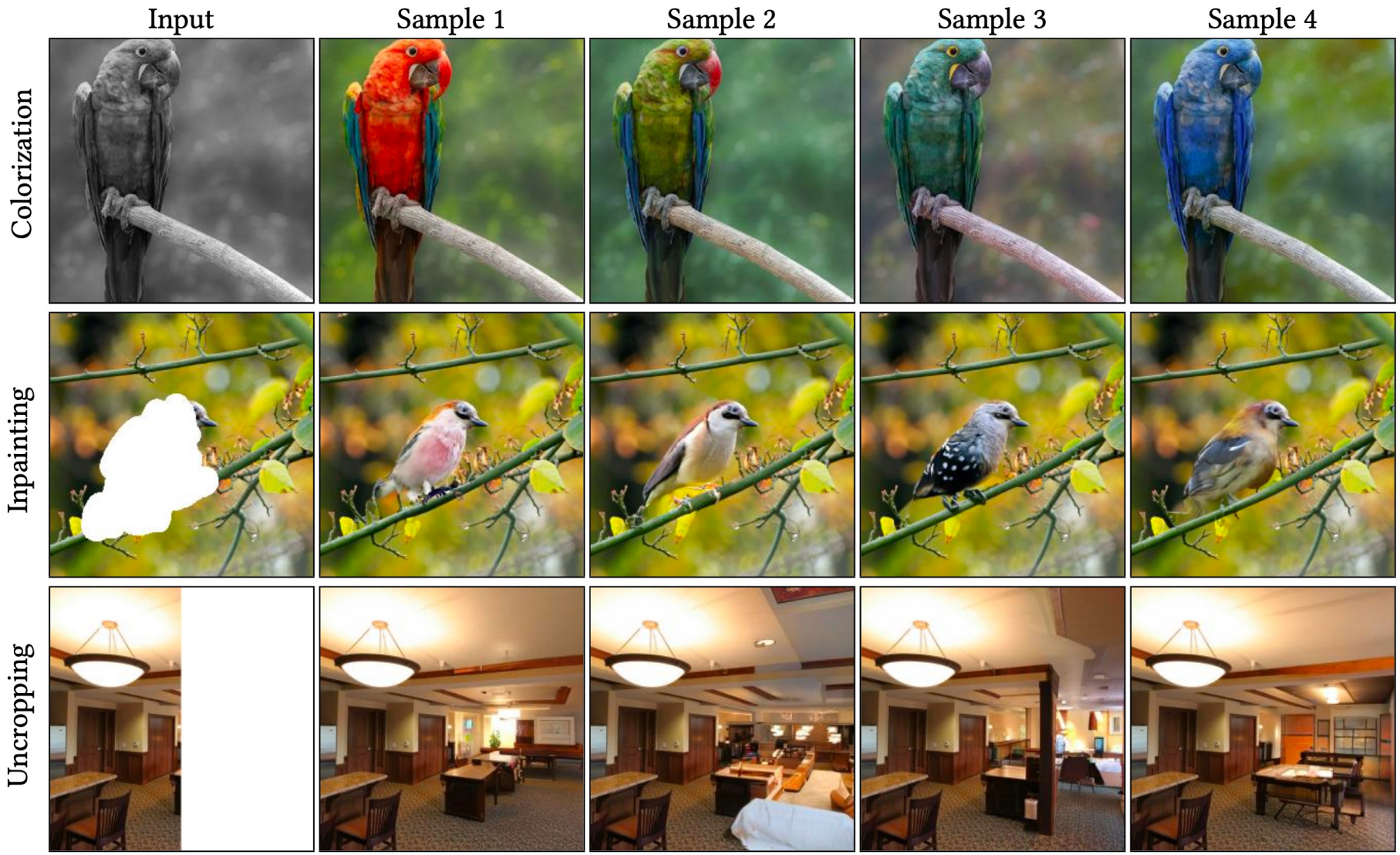
Quality – Uncropping (Outpainting)



Quality – JPEG restoration



Generative Diversity of Palette



Panorama Generation by Palette

Conclusion

Unified diffusion based i2i framework

- Diversity is improved based on diffusion.
- By conditional input, there is no need to construct each task-specific model.
- Moreover, the Palette outperforms existing task-specific models.

Unified evaluation protocol

- The performance can be evaluated from various viewpoints with the proposed protocol.
- It enables performance comparison with other studies easier and more reliable.