On the Analysis of GAN-based Image-to-Image Translation using Gaussian Noise Injection

(ICLR 2024 Presentation)

Chaohua Shi[1], Kexin Huang[2], Lu Gan[*,3], Hongqing Liu[4]

Mingrui Zhu[1], Nannan Wang[*,1], Xinbo Gao[1]

[1]Xidian University [2]National University of Defense Technology

[3]Brunel University [4]Chongqing University of Posts Telecommunication

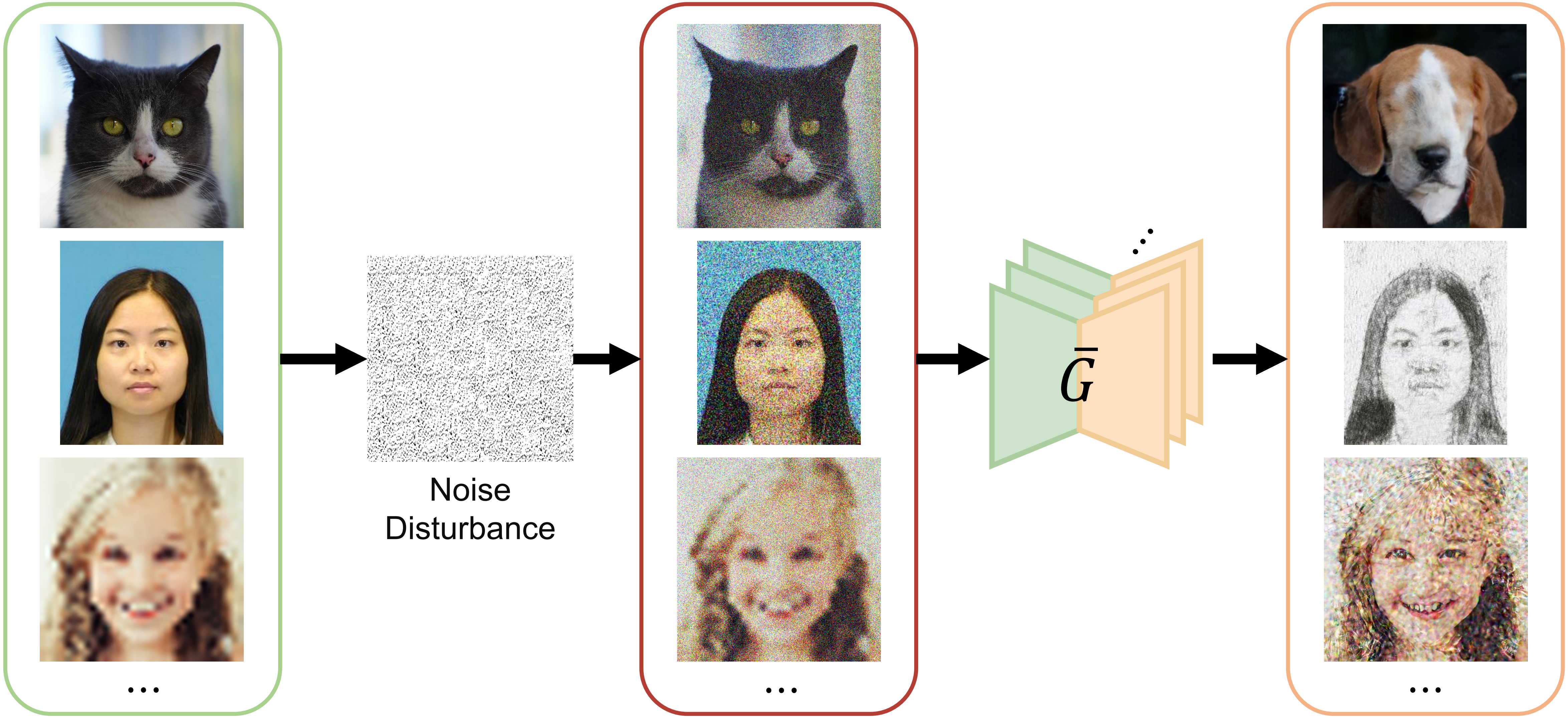
GAN-based Image-to-Image Translation Model

- G: generator
- D: discriminator
- x: an image from the source domain
- y: an image from the target domain

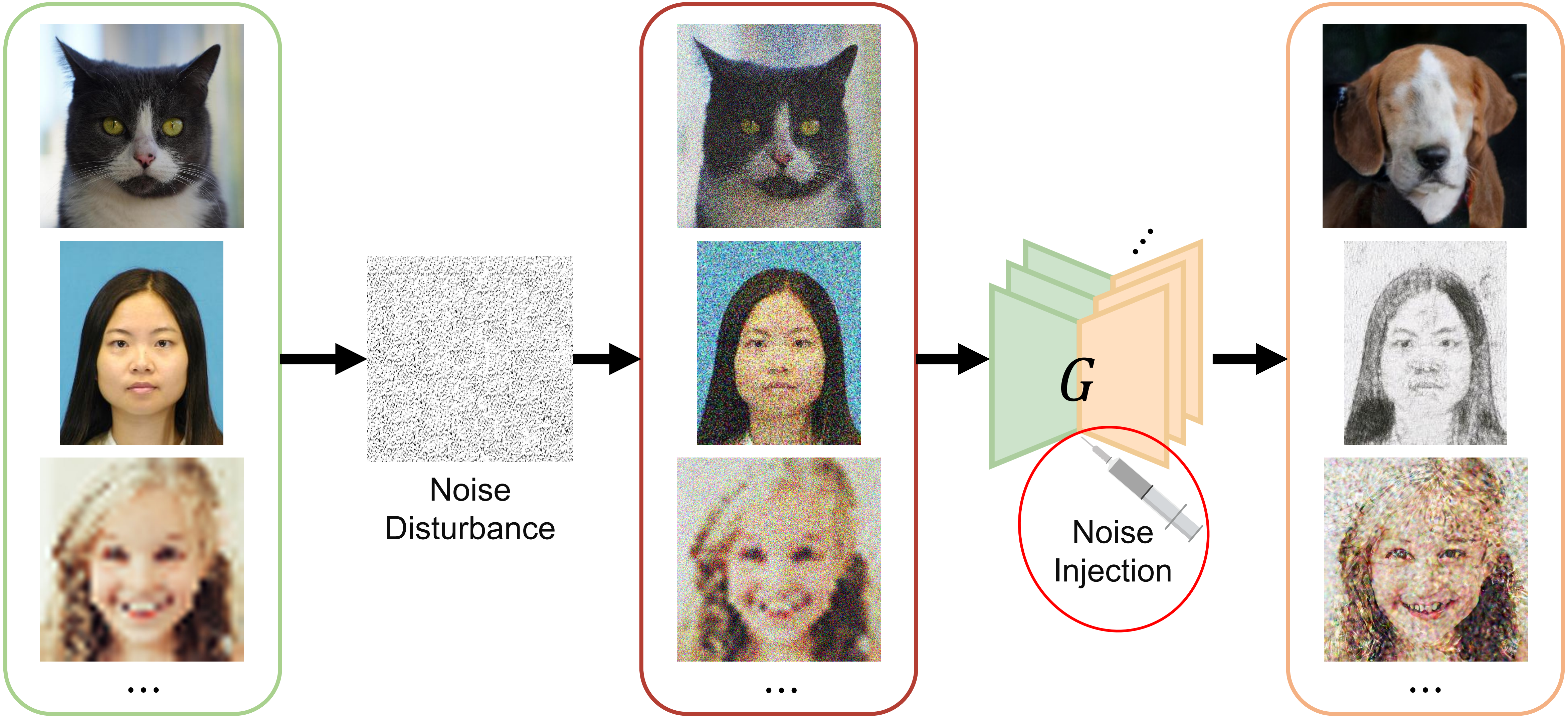
- Objective:

$$\begin{aligned}\min_G \max_D V(G, D) &= \mathbb{E}_y[\log D(y)] + \mathbb{E}_{x,y}[\log(1 - D(G(x)))] \\ &= \mathbb{E}_y[\log(D(y) - 0)] + \mathbb{E}_{x,y}[\log(1 - D(\hat{y}))]\end{aligned}$$

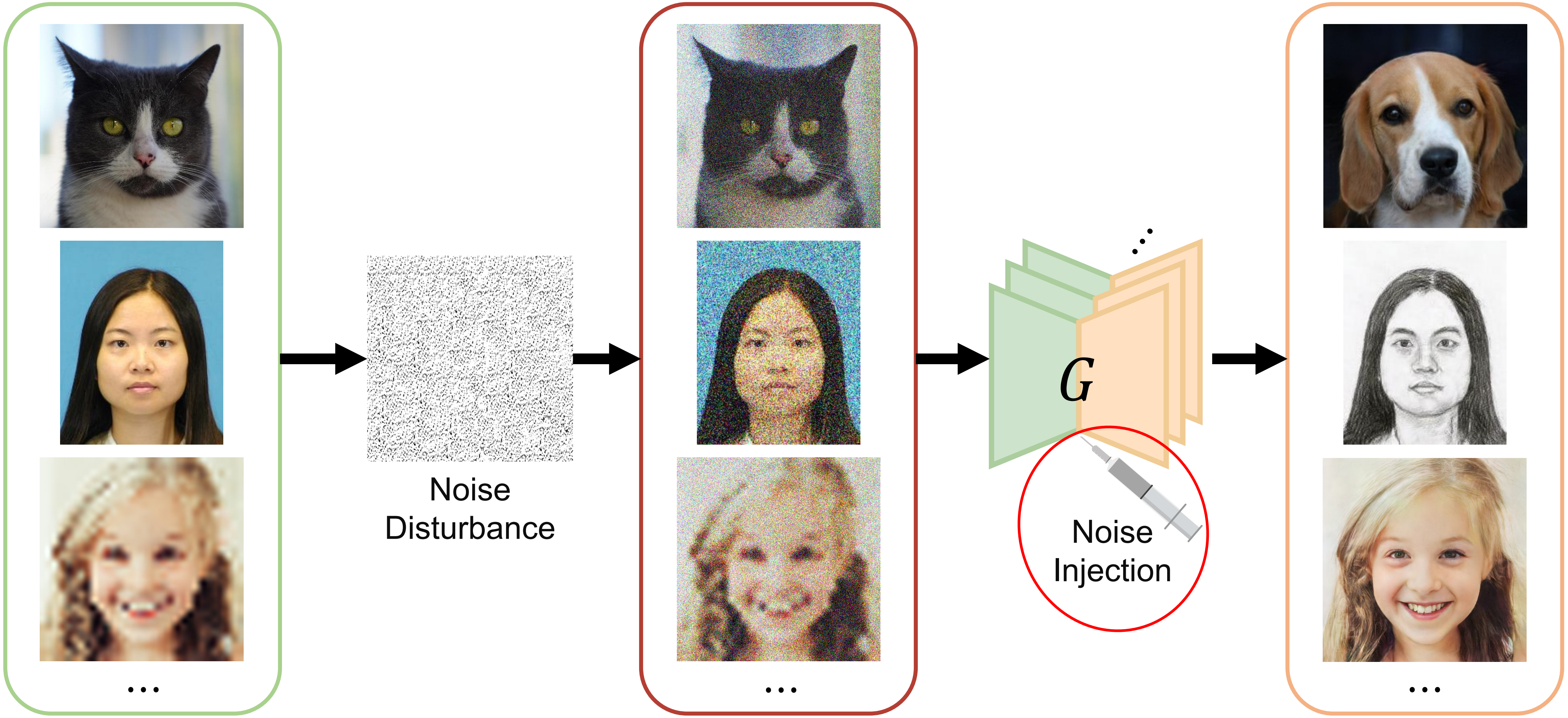
Motivation: Poor Noise Robustness



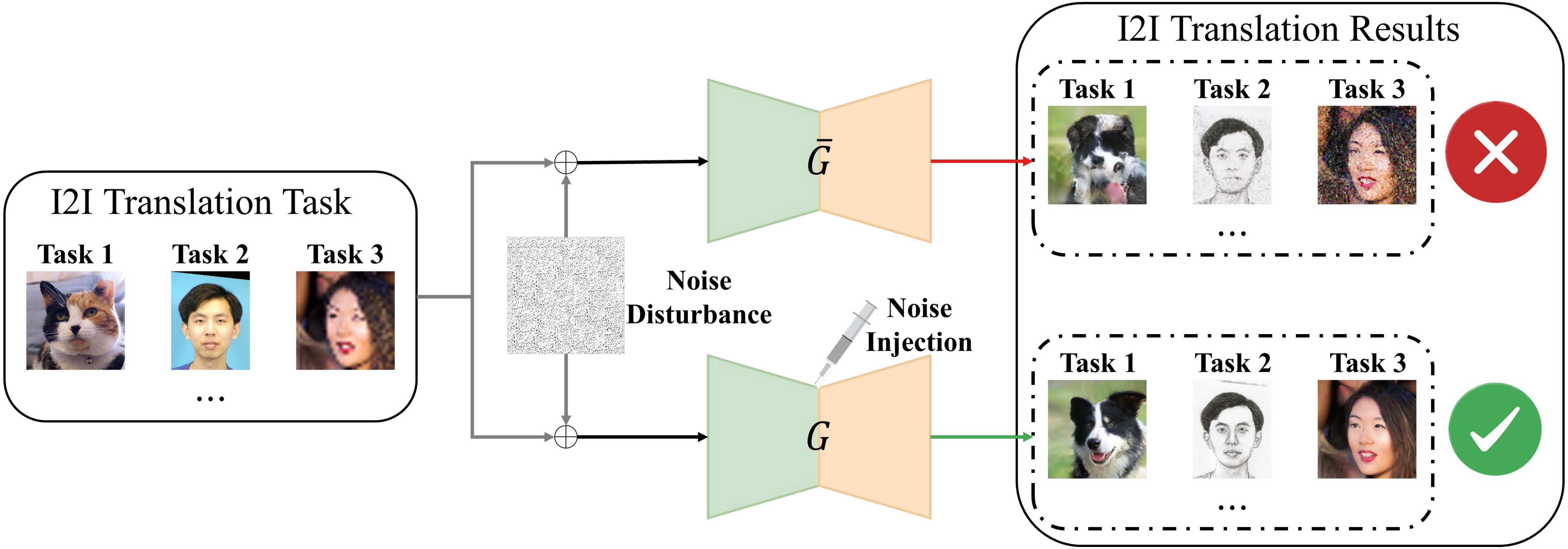
Motivation: Poor Noise Robustness



Motivation: Poor Noise Robustness



Method: Noise Injection



Theoretical Analysis: Problem Formulation

- How does the variance of Gaussian noise used in training affect the difference between the real and generated distributions?

- How does the presence of Gaussian noise in training data influence the model's ability to handle unseen noise during inference?

- Is it possible to identify an optimal noise intensity during training that guarantees consistent performance across diverse noise intensities during inference?

Theoretical Analysis: f Divergence

Theorem 1. *Let $P_{\mathbf{X},\mathbf{Y}}$ and $Q_{\mathbf{X},\mathbf{Y}}$ be two joint distributions on $\mathcal{X} \times \mathcal{Y}$ representing real data and the data generated by a model, respectively. Define $\bar{\mathbf{X}} = \mathbf{X} + \sigma\mathbf{N}$, where $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ is standard d-dimensional isotropic Gaussian noise. Let $\bar{P}_{\bar{\mathbf{X}},\mathbf{Y}}$ and $\bar{Q}_{\bar{\mathbf{X}},\mathbf{Y}}$ represent the corresponding distributions after Gaussian noise injection with their respective probability densities $\bar{p}(\bar{\boldsymbol{x}}, \boldsymbol{y})$ and $\bar{q}(\bar{\boldsymbol{x}}, \boldsymbol{y})$. For the generator function f, if its second order derivative f'' exists and $D_f\left(P_{\mathbf{X},\mathbf{Y}} \parallel Q_{\mathbf{X},\mathbf{Y}}\right)$ is finite, then $D_f\left(\bar{P}_{\bar{\mathbf{X}},\mathbf{Y}} \parallel \bar{Q}_{\bar{\mathbf{X}},\mathbf{Y}}\right)$ satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2} D_f\left(\bar{P}_{\bar{\mathbf{X}},\mathbf{Y}} \parallel \bar{Q}_{\bar{\mathbf{X}},\mathbf{Y}}\right) = -\frac{1}{2}\eta_f(\sigma^2), \tag{1}$$

in which $\eta_f(\sigma^2)$ represents the weighted mean square error between two score functions

$$\eta_f(\sigma^2) = \mathbb{E}_{\bar{P}_{\bar{\mathbf{X}},\mathbf{Y}}}\left\{\frac{\bar{p}(\bar{\boldsymbol{x}}, \boldsymbol{y})}{\bar{q}(\bar{\boldsymbol{x}}, \boldsymbol{y})} f''\left(\frac{\bar{p}(\bar{\boldsymbol{x}}, \boldsymbol{y})}{\bar{q}(\bar{\boldsymbol{x}}, \boldsymbol{y})}\right) \left\|\nabla_{\bar{\boldsymbol{x}}} \log \bar{p}(\bar{\boldsymbol{x}}, \boldsymbol{y}) - \nabla_{\bar{\boldsymbol{x}}} \log \bar{q}(\bar{\boldsymbol{x}}, \boldsymbol{y})\right\|^2\right\}, \tag{2}$$

where $\nabla_{\bar{\boldsymbol{x}}} \log \bar{p}(\bar{\boldsymbol{x}}, \boldsymbol{y})$ and $\nabla_{\bar{\boldsymbol{x}}} \log \bar{q}(\bar{\boldsymbol{x}}, \boldsymbol{y})$ are the score functions of $\bar{p}(\bar{\boldsymbol{x}}, \boldsymbol{y})$ and $\bar{q}(\bar{\boldsymbol{x}}, \boldsymbol{y})$, respectively.

Unveils how the rate of change of $D_f(\bar{P} \parallel \bar{Q})$ concerning σ^2 is portrayed through $n_f(\sigma^2)$!

Theoretical Analysis: f Divergence

For small $\sigma = \sigma_t$, a Taylor series expansion yields:

$$D_f\left(P_{\mathbf{X},\mathbf{Y}} \parallel Q_{\mathbf{X},\mathbf{Y}}\right) = \boxed{D_f\left(\bar{P}_{\mathbf{X}+\sigma_t\mathbf{N},\mathbf{Y}} \parallel \bar{Q}_{\mathbf{X}+\sigma_t\mathbf{N},\mathbf{Y}}\right)} + \frac{\sigma_t^2}{2}\boxed{\eta_f(\sigma_t^2)} + o(\sigma_t^2)$$

Minimizing divergence between $\bar{p}(x + \sigma_t \mathbf{N}, y)$ and $\bar{q}(x + \sigma_t \mathbf{N}, y)$

Tends to decrease

Hence, by injecting Gaussian noise with small σ_t^2 and aligning the noise-perturbed distributions during training, the model

- be guided to align the original, noise-free distributions as well
- results in a coherent I2I translation

Theoretical Analysis: Mismatched Noisy Inputs

Gaussian source

Theorem 2. *Consider the KL-divergences denoted by $\rho(\sigma_t^2, \boldsymbol{\Sigma}_e)$ in (6) for general noise, and $\rho_g(\sigma_t^2, \boldsymbol{\Sigma}_e)$ in (7) for Gaussian noise. Under these definitions, the following properties hold:*

1. *Let $\boldsymbol{\Sigma}_e = \sigma_e^2 \boldsymbol{\Sigma}_{\tilde{e}}$, in which $\boldsymbol{\Sigma}_{\tilde{e}}$ is normalized covariance matrix with $\mathrm{Tr}(\boldsymbol{\Sigma}_{\tilde{e}}) = d$. Then, $\rho(\sigma_t^2, \sigma_e^2 \boldsymbol{\Sigma}_{\tilde{e}})$ is convex with respect to σ_e^2. Additionally, for small σ_e^2 with $\sigma_e^2 \ll 1$, the following approximation is valid:*

$$\rho(\sigma_t^2, \sigma_e^2 \boldsymbol{\Sigma}_{\tilde{e}}) = \rho_g(\sigma_t^2, \sigma_e^2 \boldsymbol{\Sigma}_{\tilde{e}}) + o(\sigma_e^2). \tag{8}$$

2. *If $\boldsymbol{\Sigma}_e \geq \frac{\sigma_t^2}{2} \boldsymbol{I}_d$, the inequality $\rho(\sigma_t^2, \boldsymbol{\Sigma}_e) < \rho(0, \boldsymbol{\Sigma}_e)$ is satisfied.*

Notation:

- X be d-dimensional random variable with normal distribution $\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$
- Training: $\bar{X}, \bar{Y} = G(\bar{X})$ denote the training noisy counterparts(input and corresponding output)
- Inference: $\hat{X} = X + E, \hat{Y} = G(\hat{X})$ denote the inference noisy counterparts
- Marginal distributions: $\bar{P}_{\bar{X}}, \hat{P}_{\hat{X}}, \bar{Q}_{\bar{Y}}, \hat{Q}_{\hat{Y}}$
- Joint distribution: $\bar{Q}_{\bar{X},\bar{Y}}, \hat{Q}_{\hat{X},\hat{Y}}$

Theoretical Analysis: Mismatched Noisy Inputs

Non-Gaussian source

Theorem 3. *Let $\mathbf{X}$ be a d-dimensional random vector with an arbitrary probability distribution and finite entropy $h(\boldsymbol{X})$. Denote $\theta(\sigma_t^2, \sigma_e^2 \boldsymbol{\Sigma}_{\widetilde{e}}) \triangleq D_{KL}\left(\hat{P}_{\mathbf{X}+\mathbf{E}} \| \bar{P}_{\mathbf{X}+\boldsymbol{\sigma}_t \mathbf{N}}\right)$, where the definitions of $\boldsymbol{E}$, $\boldsymbol{\Sigma}_{\widetilde{e}}$, $\boldsymbol{N}$, σ_t and σ_e^2 are the same as those in Theorem 2. Let $\theta_g(\sigma_t^2, \sigma_e^2 \boldsymbol{\Sigma}_{\widetilde{e}})$ denotes the special case of $\theta(\sigma_t^2, \sigma_e^2 \boldsymbol{\Sigma}_{\widetilde{e}})$ when $\boldsymbol{E}$ is Gaussian noise. Then,*

1. *For small σ_e^2 with $\sigma_e^2 \ll 1$,*

$$\theta(\sigma_t^2, \sigma_e^2 \boldsymbol{\Sigma}_{\widetilde{e}}) = \theta_g(\sigma_t^2, \sigma_e^2 \boldsymbol{\Sigma}_{\widetilde{e}}) + o(\sigma_e^2); \tag{9}$$

2. *When $\boldsymbol{E}$ is also iid Gaussian, $\theta_g(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d) \triangleq D_{KL}\left(\hat{P}_{\mathbf{X}+\boldsymbol{\sigma}_e \mathbf{N}} \| \bar{P}_{\mathbf{X}+\boldsymbol{\sigma}_t \mathbf{N}}\right)$ satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}\sigma_e^2}\theta_g(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d) = \mathbb{E}_{\hat{p}(\hat{\boldsymbol{x}})}\left\{-\frac{1}{2}\left\|\nabla_{\hat{\boldsymbol{x}}} \log \hat{p}(\hat{\boldsymbol{x}})\right\|^2 + \frac{1}{2}\nabla_{\hat{\boldsymbol{x}}} \log \hat{p}(\hat{\boldsymbol{x}}) \cdot \nabla_{\bar{\boldsymbol{x}}} \log \bar{p}(\bar{\boldsymbol{x}})\right\} \tag{10}$$

Notation:

- X be d-dimensional random variable with an arbitrary probability distribution and finite entropy $h(X)$
- Training: $\bar{X}, \bar{Y} = G(\bar{X})$ denote the training noisy counterparts(input and corresponding output)
- Inference: $\hat{X} = X + E, \hat{Y} = G(\hat{X})$ denote the inference noisy counterparts
- Marginal distributions: $\bar{P}_{\bar{X}}, \hat{P}_{\hat{X}}, \bar{Q}_{\bar{Y}}, \hat{Q}_{\hat{Y}}$
- Joint distribution: $\bar{Q}_{\bar{X},\bar{Y}}, \hat{Q}_{\hat{X},\hat{Y}}$

Theoretical Analysis: Training Noise Intensity

Given an i.i.d. Gaussian noise e with $\boldsymbol{\Sigma}_e = \sigma_e^2 \boldsymbol{I}_d$ $(0 \leq \sigma_e^2 \leq \lambda_{\max})$, define $\sigma_{t,o}^2$ as the optimal noise level that minimizes the worst-case KL distance $\rho(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d)$

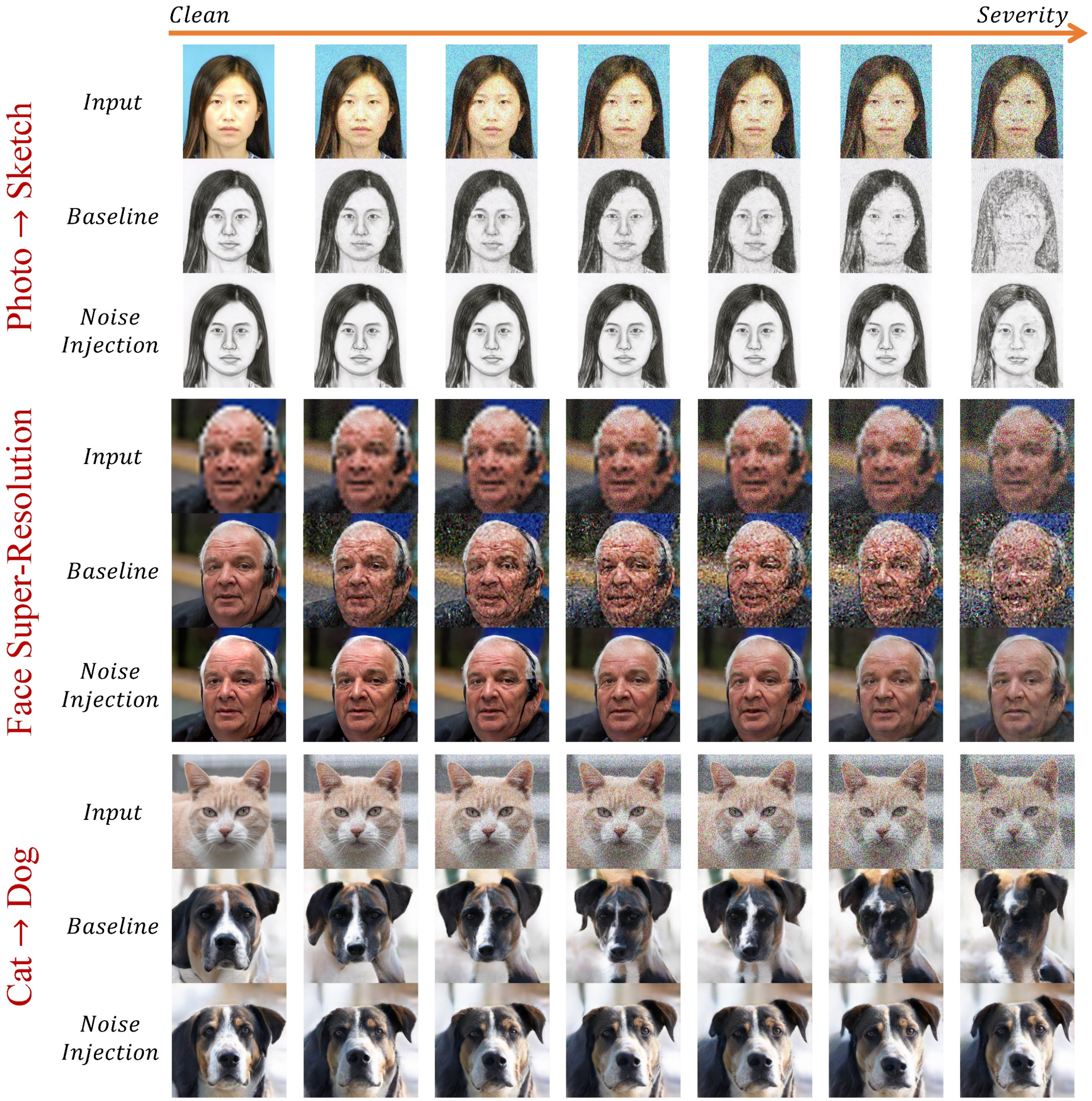
$$\sigma_{t,o}^2 = \arg\min_{\sigma_t^2} \left\{ \max_{0 \leq \sigma_e^2 \leq M} \rho(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d) \right\}$$

For this optimal level, it satisfies $\rho\left(\sigma_{t,o}^2, \mathbf{0}_d\right) = \rho\left(\sigma_{t,o}^2, \lambda_{\max} \boldsymbol{I}_d\right)$. Besides, if σ_e^2 is uniformly distributed between 0 and $\lambda_{\max}$, the optimal training noise intensity $\bar{\sigma}_{t,o}^2$ that minimizes the average KL-divergence is $\frac{1}{2}\lambda_{max}$, i.e.,

$$\bar{\sigma}_{t,o}^2 = \arg\min_{\sigma_t^2} \mathbb{E}_{\sigma_e^2 \sim \mathcal{U}(0,\lambda_{\max})} \{\rho(\sigma_t^2, \sigma_e^2 I_d)\} = \frac{1}{2}\lambda_{\max}$$

Hence, this corollary offers a theoretically sound method for determining the optimal training noise variance for an arbitrary type of i.i.d. inference noise.

Results



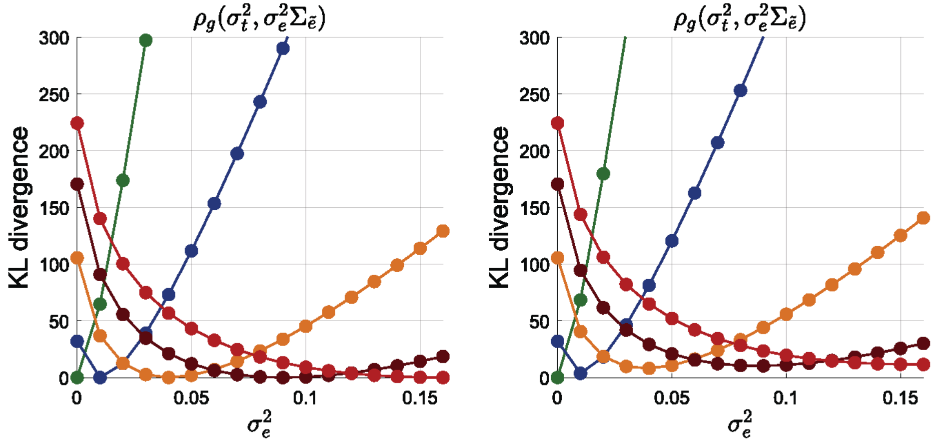
Three GAN-based I2I models are used to verify our theoretical analysis

- Sketch Transformer[1] (Photo→Sketch)
- HiFaceGAN[2] (Face Super-Resolution)
- GP-UNIT[3] (Cat→Dog)

[1] Zhu, et al. “A sketch-transformer network for face photo-sketch synthesis.” IJCAI. 2021.
[2] Yang, et al. “Hifacegan: Face renovation via collaborative suppression and replenishment.” ACM MM. 2020.
[3] Yang, et al. “GP-UNIT: Generative prior for versatile unsupervised image-to-image translation.” TPAMI. 2023

Results

Theory

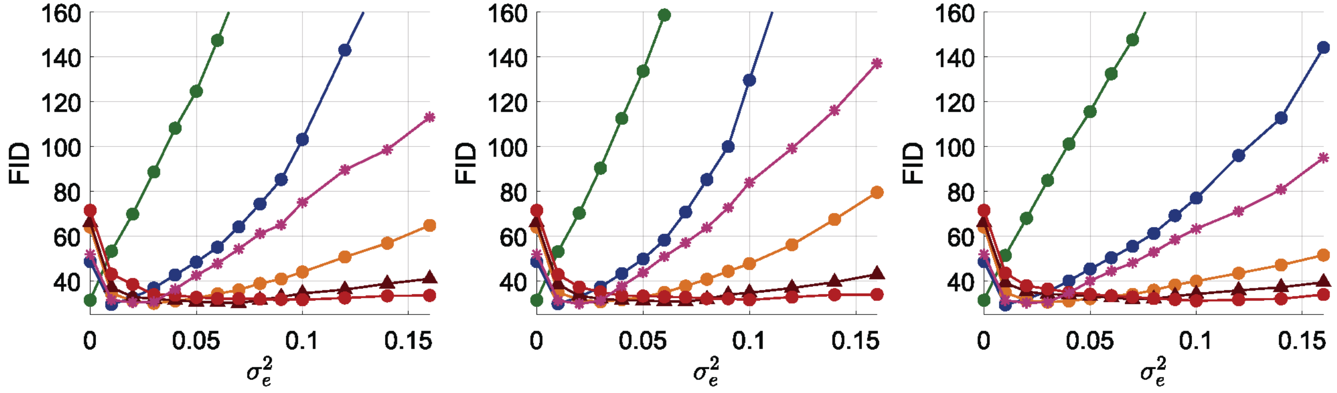


(a) iid Gaussian Noise (b) non-iid Gaussian Noise

$\sigma_{t0}^2 = 0$ $\sigma_{t1}^2 = 0.01$ $\sigma_{t2}^2 = 0.04$ $\sigma_{t3}^2 = 0.08$ $\sigma_{t4}^2 = 0.16$

Visualization $\rho_g(\sigma_t^2, \sigma_e^2\Sigma_{\tilde{e}})$ for AR(1) signal model with $d = 256$ and covariance matrix $\mathbf{\Sigma}_s(k,l) = 0.95\rho^{|k-l|}$ (for $0 \leq k, l \leq 255$).

Experiment



(a) Gaussian Noise (b) Uniform Noise (c) Laplacian Noise

$\sigma_{t0}^2 = 0$ $\sigma_{t1}^2 = 0.01$ $\sigma_{t2}^2 = 0.04$ $\sigma_{t3}^2 = 0.08$ $\sigma_{t4}^2 = 0.16$ Learnable σ_t^2

FID score comparison on noisy inputs for models trained with different Gaussian noise levels.

Consistent trends!!!

Thank You

More details + results in our paper!

Paper: https://openreview.net/forum?id=sLregLuXpn

Code: https://github.com/Alan0693/Noise-Injection

Brunel University London