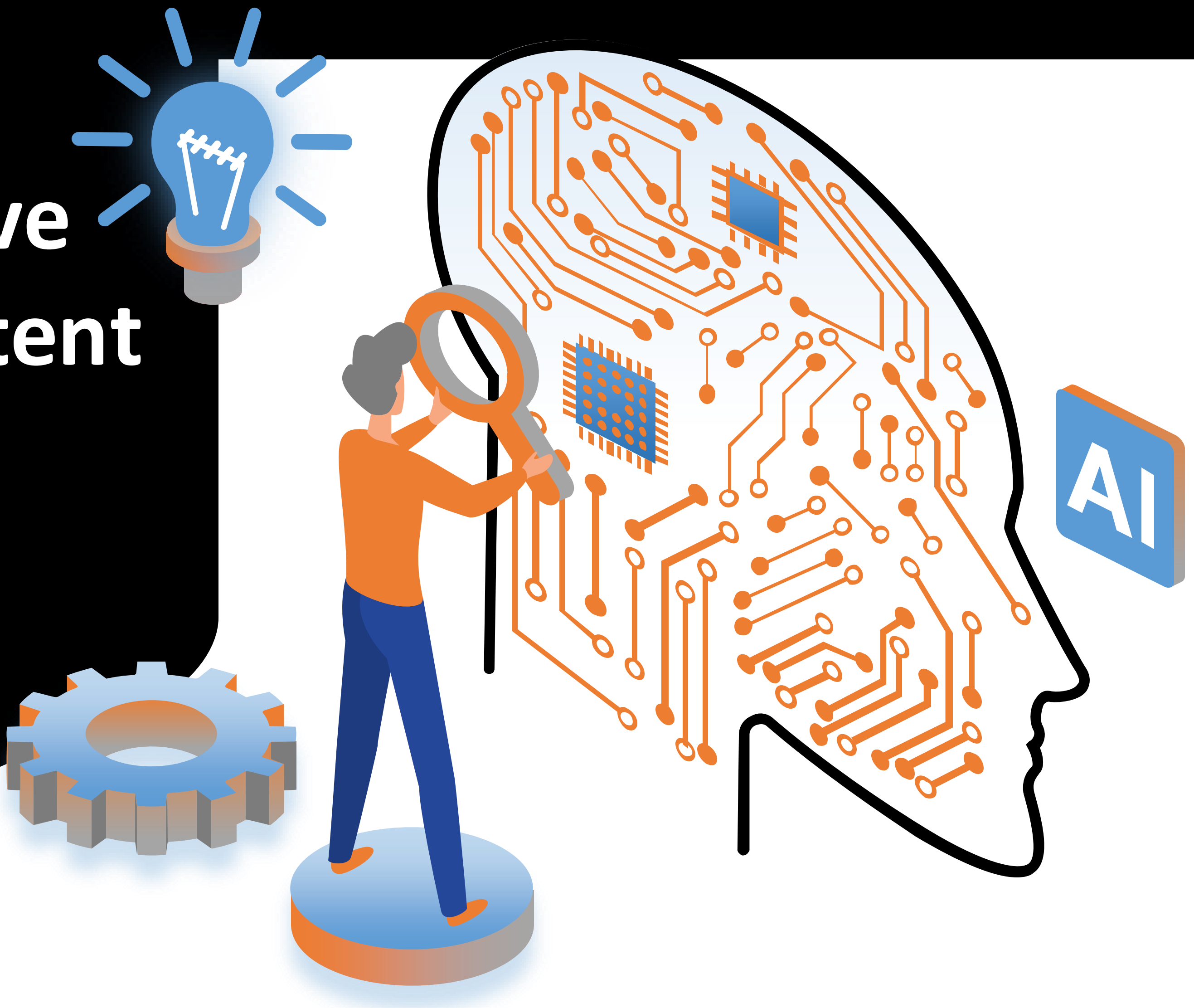
Harnessing Generative Priors for Visual Content Restoration



Chen-Change Loy

Last update: 3 Jan 2025

2025 Winter School on Biometrics

Outline

- **Introduction**
  - Problem objective
  - Challenges
  - Architectures
  - Losses
  - Handling complex degradation
  - Metric

- **Types of Prior for Restoration**

- **Diffusion Prior**

Introduction

Problem objective

Recover the latent high-quality (HQ) faces $\mathbf{x}$ from its degraded low-quality (LQ) faces $\mathbf{y}$

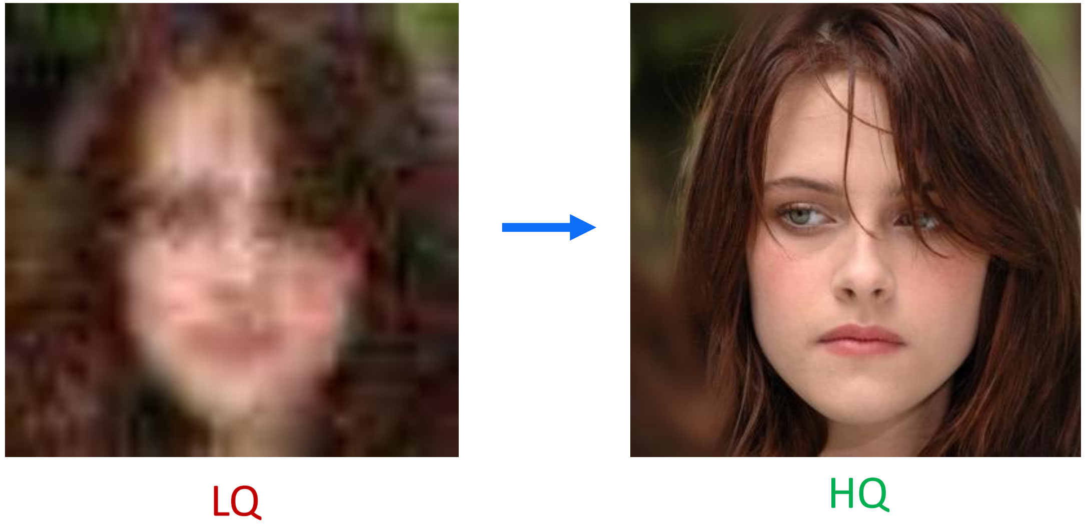
$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}$$

where $\mathbf{H}$ is a degradation matrix, $\mathbf{v}$ is additive noise

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \lambda\Phi(\mathbf{x})$$

fidelity term

regularization term



Problem objective

Recover the latent high-quality (HQ) faces $\mathbf{x}$ from its degraded low-quality (LQ) faces $\mathbf{y}$

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}$$

where $\mathbf{H}$ is a degradation matrix, $\mathbf{v}$ is additive noise

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \lambda\Phi(\mathbf{x})$$

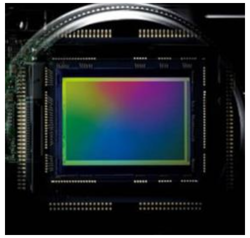
fidelity term regularization term

If we know the $\mathbf{H}$ and $\mathbf{v}$, then is a non-blind super-resolution. Otherwise it is a blind super-resolution (how to deal with this problem?).

Challenges

Real-world degradations usually come from complicate processes, such as **imaging system of cameras**, **image editing**, and **Internet transmission**.

Take Photo



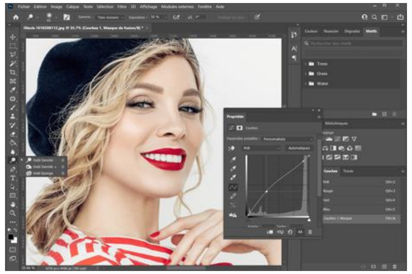
Camera Blur
(motion and defocus)

Sensor Noise

JPEG compression

Editing

Downsampling,
Artifacts



Social Media Sharing

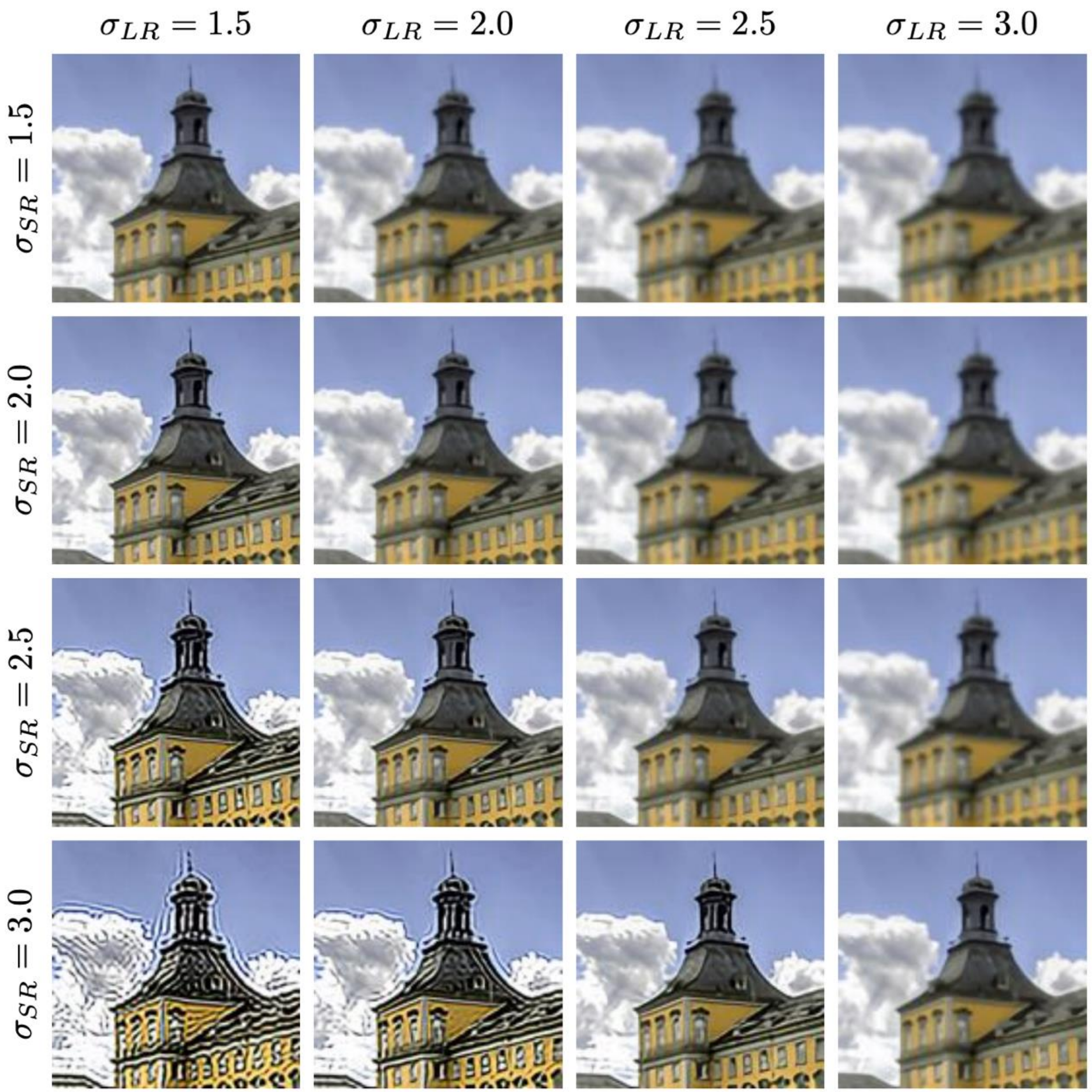
Compression,
Downsampling, Noises

…

Challenges

- Learning-based methods will suffer severe performance drop when the pre-defined degradation is different from the real one

- This phenomenon of kernel mismatch will introduce undesired artifacts to output images



SR sensitivity to the kernel mismatch.
σ_{LR} denotes the kernel used for downsampling and σ_{SR} denotes the kernel used for SR.

Figure credit: J. Gu et al., Blind Super-Resolution With Iterative Kernel Correction, CVPR 2019

Challenges

- Highly ill-posed problem - one LQ image corresponds to infinite number of HQ images

LQ

HQ

...

Challenges

- Vice versa - one HQ image corresponds to infinite number of LQ images

LQ

...

HQ

Architectures – some examples

- Convolutional neural networks
  - SRCNN
  - FSRCNN
  - VDSR
- Generative adversarial network
  - SRGAN
  - ESRGAN
- Transformers
  - SwinIR
  - Uformer
  - Restormer
- Diffusion models
  - StableSR
  - DiffBIR
  - ResShift
  - SeeSR
  - CoSeR
  - SUPIR

Losses

Mean squared error

- Minimizing the loss between the reconstructed images $F\ (\mathbf{Y};\ \Theta)$ and the corresponding ground truth high-resolution images X

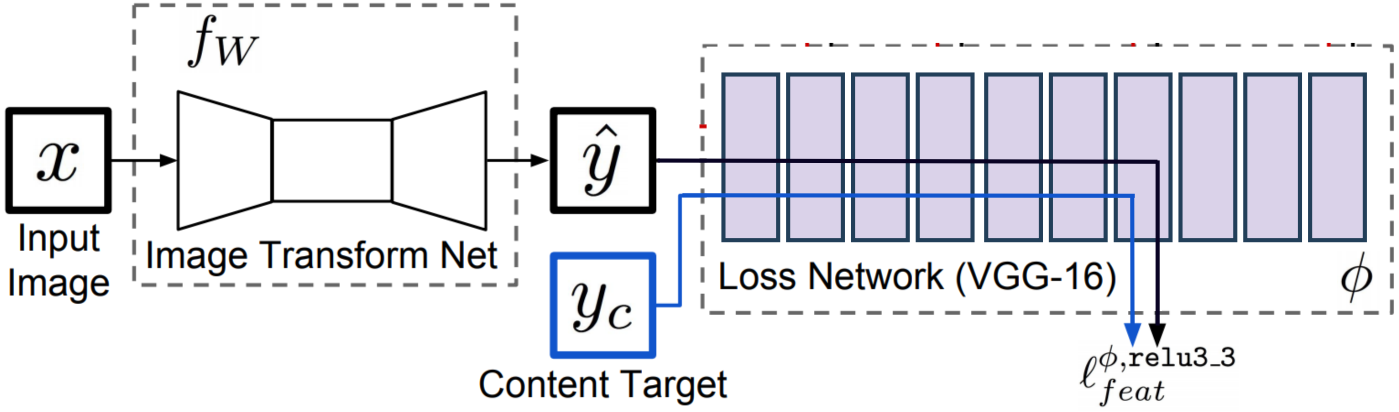
$$L(\Theta) = \frac{1}{n}\sum_{i=1}^{n} ||F(\mathbf{Y}_i;\Theta) - \mathbf{X}_i||^2$$

- The loss is minimized using stochastic gradient descent with the standard backpropagation

Losses

Perceptual loss

Encourages the output image to be perceptually similar to the target image, but does not force them to match exactly



The feature reconstruction loss is the (squared, normalized) Euclidean distance between feature representations

activations of the j-th layer of target image

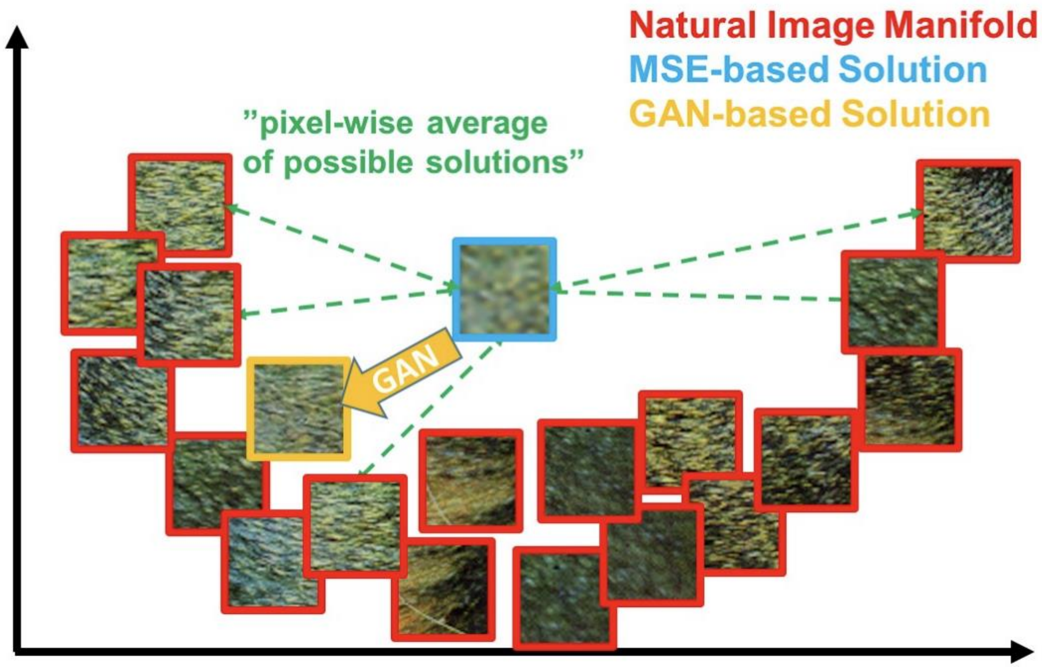
$$\ell_{feat}^{\phi,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \|\phi_j(\hat{y}) - \phi_j(y)\|_2^2$$

feature map of shape $C_j \times H_j \times W_j$

activations of the j-th layer of output image

Justin Johnson et al., [Perceptual Losses for Real-Time Style Transfer and Super-Resolution](), ECCV 2016

Losses



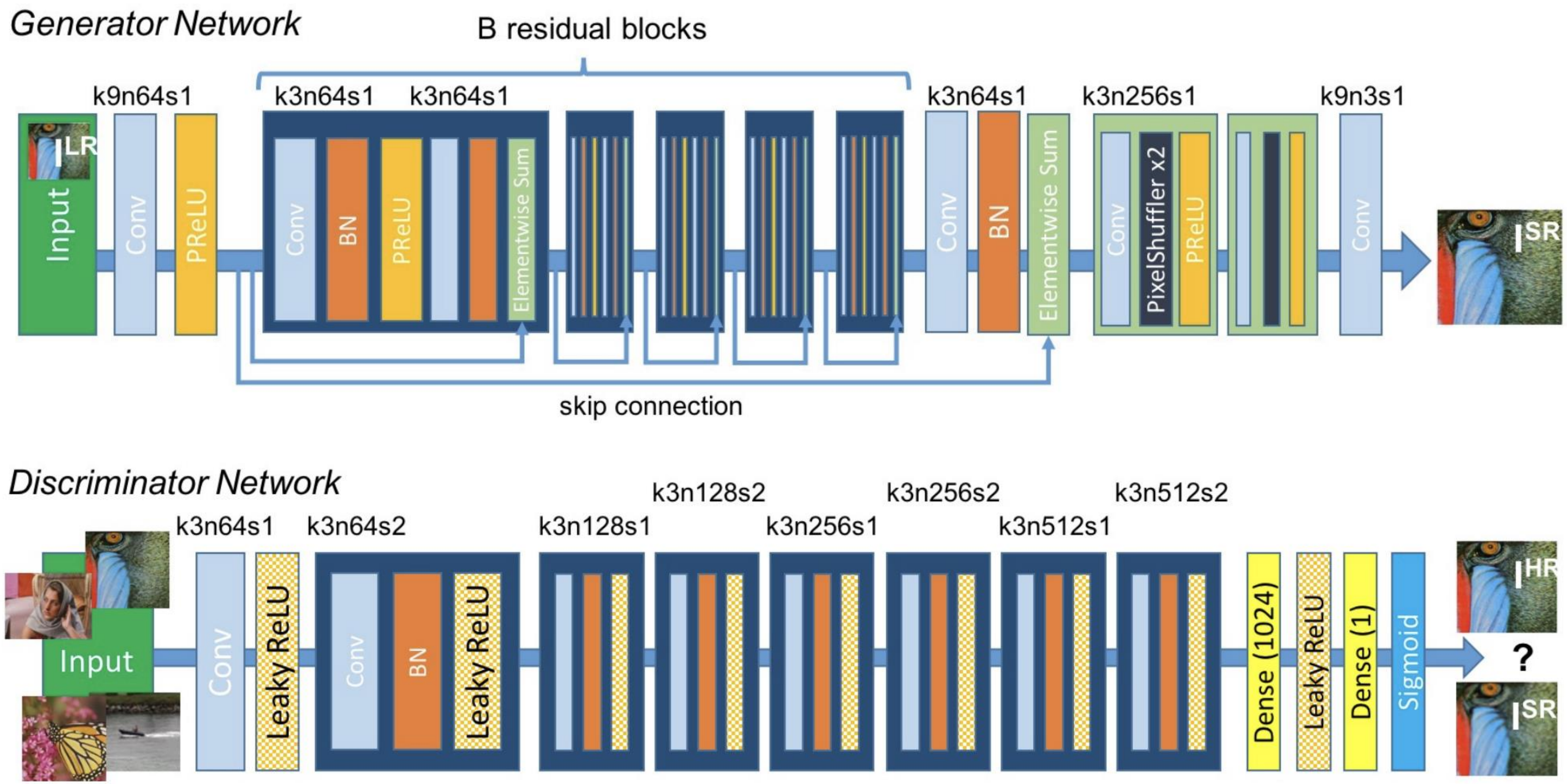
Adversarial loss

The MSE-based solution appears overly smooth due to the pixel-wise average of possible solutions in the pixel space

Generative Adversarial Network (GAN) drives the reconstruction towards the natural image manifold producing perceptually more convincing solutions

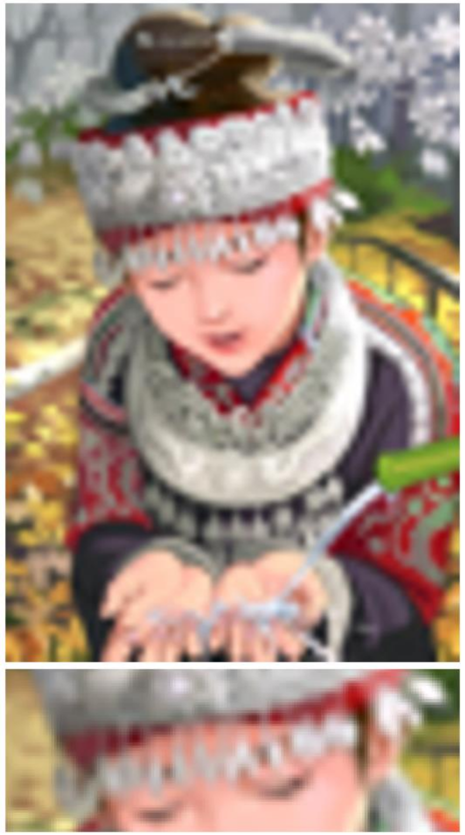
C. Ledig et al., Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, CVPR 2017

Losses

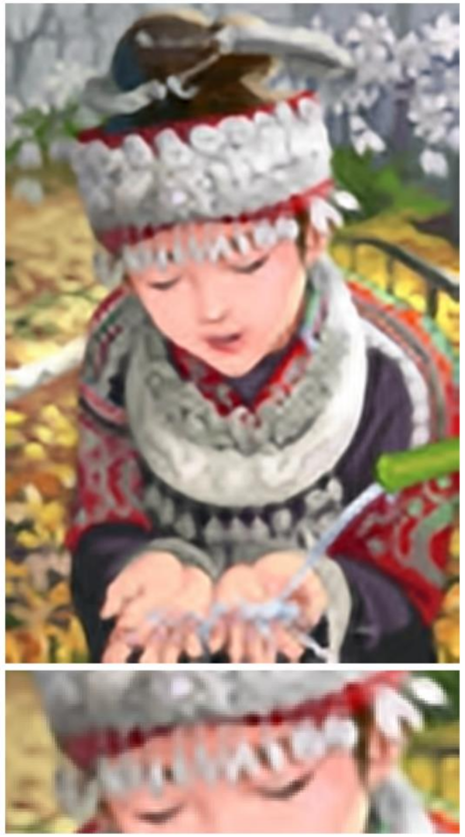


$$\min_{\theta_G} \max_{\theta_D} \; \mathbb{E}_{I^{HR} \sim p_{\text{train}}(I^{HR})}[\log D_{\theta_D}(I^{HR})] +$$
$$\mathbb{E}_{I^{LR} \sim p_G(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR}))]$$

Losses



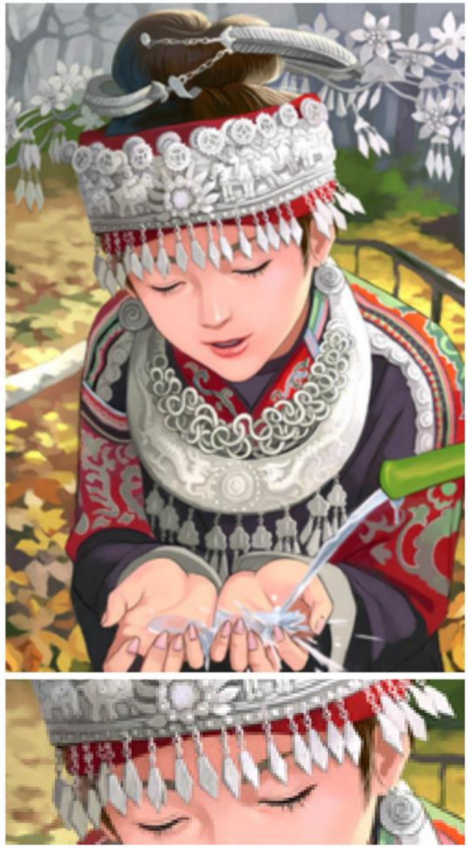
Input



MSE Loss

Perceptual Loss

Adversarial Loss

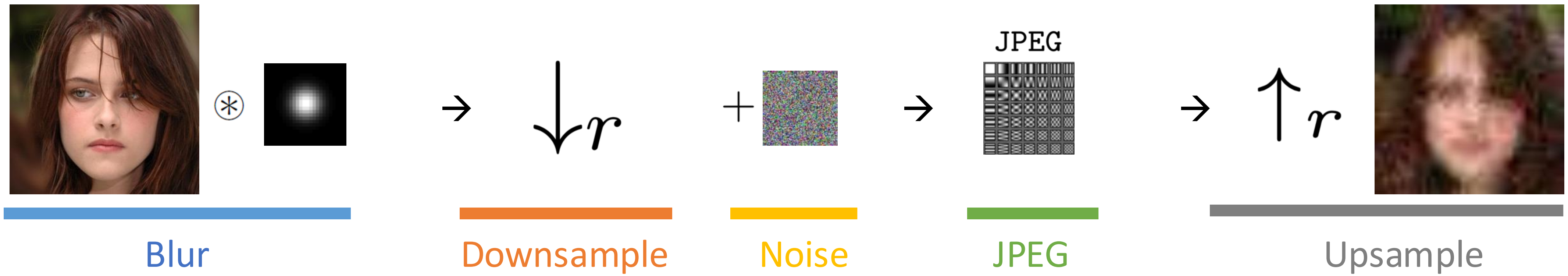


Ground Truth

Handling complex degradation

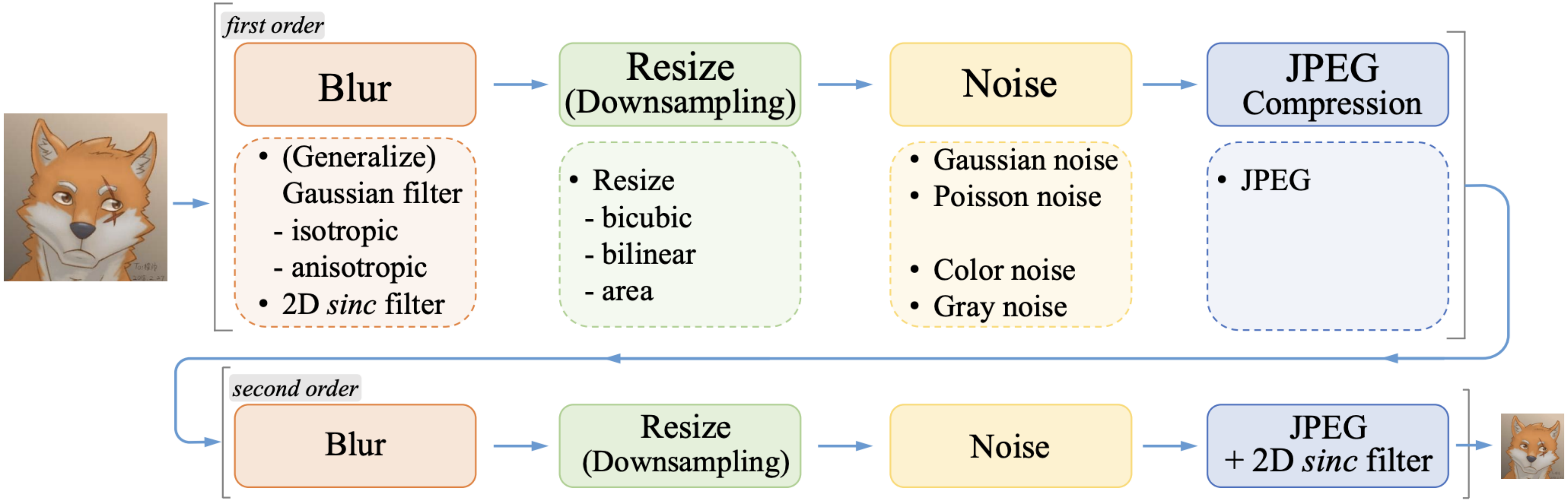
Degradation model

$$I_l = \{[(I_h \otimes k_\sigma)_{\downarrow_r} + n_\delta]_{\mathrm{JPEG}_q}\}_{\uparrow_r}$$



Handling complex degradation

Degradation model



Not a silver bullet - merely extends the solvable degradation boundary of previous blind SR methods through modifying the data synthesis process

X. Wang et al. [Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data](), AIM 2021

Metrics

Peak signal-to-noise ratio (**PSNR**) is an expression for the ratio between the maximum possible value (power) of a signal and the power of distorting noise that affects the quality of its representation

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

$$= 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

$$= 20 \cdot \log_{10}(MAX_I) - 10\log_{10}(MSE)$$

MAX_I = Maximum possible pixel value of the image. For 8 bits image, this is 255

Cons: Doesn't reflect human perception well

Metrics

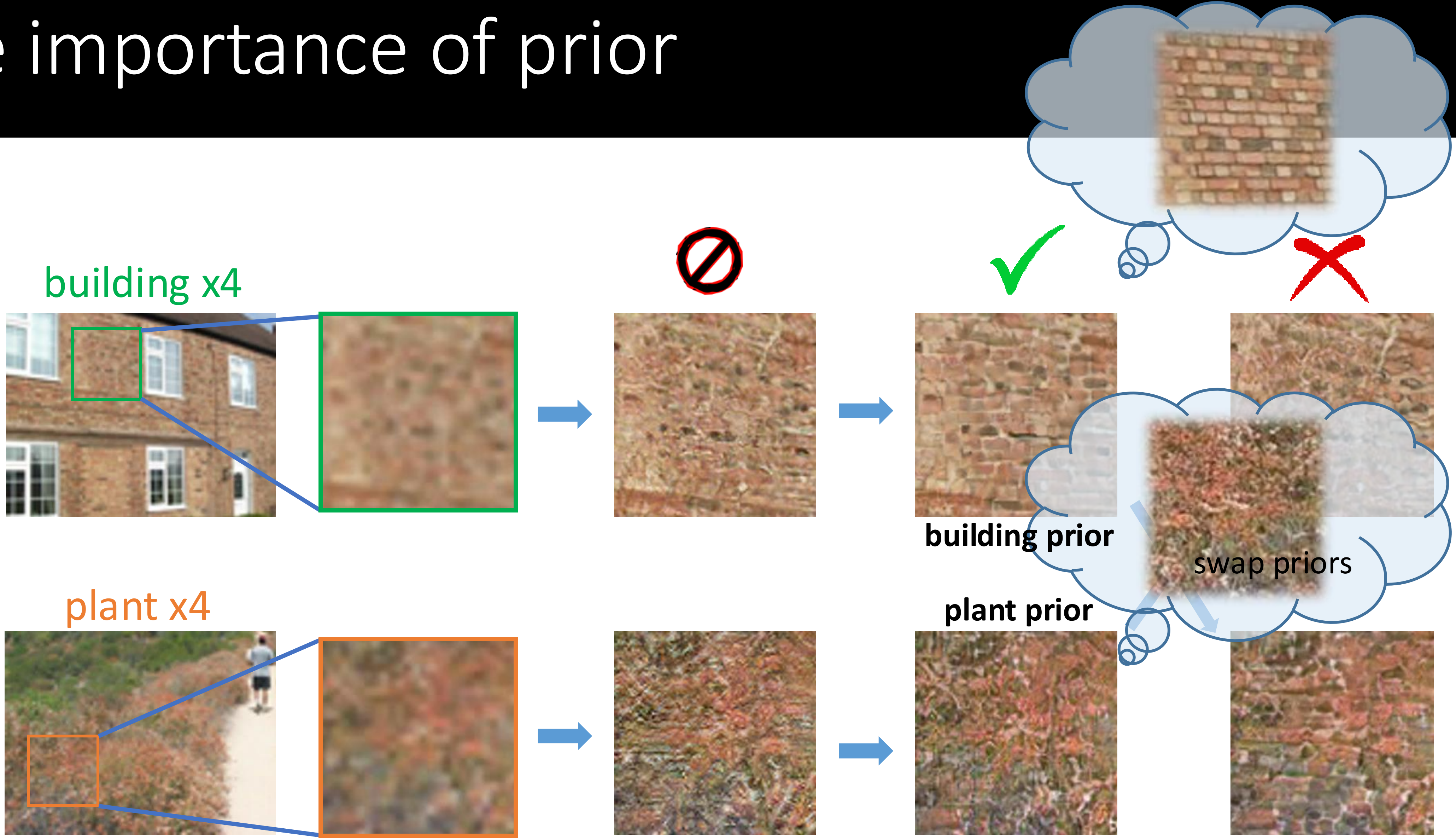
- Perceptual metric
  - FID (Heusel et al., 2017)
  - LPIPS (Zhang et al., 2018a)
  - MUSIQ (Ke et al., 2021)
  - CLIP-IQA (Wang et al., 2023)

Example:

Datasets	Metrics	RealSR	BSRGAN	DASR	Real-ESRGAN+	FeMaSR	LDM	SwinIR-GAN	IF_III	**StableSR**
DIV2K Valid	PSNR ↑	**24.62**	24.58	24.47	24.29	23.06	23.32	23.93	23.36	23.26
	SSIM ↑	0.5970	0.6269	0.6304	**0.6372**	0.5887	0.5762	0.6285	0.5636	0.5726
	LPIPS ↓	0.5276	0.3351	0.3543	**0.3112**	0.3126	0.3199	0.3160	0.4641	0.3114
	FID ↓	49.49	44.22	49.16	37.64	35.87	26.47	36.34	37.54	**24.44**
	CLIP-IQA ↑	0.3534	0.5246	0.5036	0.5276	0.5998	0.6245	0.5338	0.3980	**0.6771**
	MUSIQ ↑	28.57	61.19	55.19	61.05	60.83	62.27	60.22	43.71	**65.92**
RealSR	PSNR ↑	**27.30**	26.38	27.02	25.69	25.06	25.46	26.31	25.47	24.65
	SSIM ↑	0.7579	0.7651	0.7707	0.7614	0.7356	0.7145	**0.7729**	0.7067	0.7080
	LPIPS ↓	0.3570	0.2656	0.3134	0.2709	0.2937	0.3159	**0.2539**	0.3462	0.3002
	CLIP-IQA ↑	0.3687	0.5114	0.3198	0.4495	0.5406	0.5688	0.4360	0.3482	**0.6234**
	MUSIQ ↑	38.26	63.28	41.21	60.36	59.06	58.90	58.70	41.71	**65.88**
DRealSR	PSNR ↑	**30.19**	28.70	29.75	28.62	26.87	27.88	28.50	28.66	28.03
	SSIM ↑	0.8148	0.8028	**0.8262**	0.8052	0.7569	0.7448	0.8043	0.7860	0.7536
	LPIPS ↓	0.3938	0.2858	0.3099	0.2818	0.3157	0.3379	**0.2743**	0.3853	0.3284
	CLIP-IQA ↑	0.3744	0.5091	0.3813	0.4515	0.5634	0.5756	0.4447	0.2925	**0.6357**
	MUSIQ ↑	26.93	57.16	42.41	54.26	53.71	53.72	52.74	30.71	**58.51**
DPED-iphone	CLIP-IQA ↑	0.4496	0.4021	0.2826	0.3389	**0.5306**	0.4482	0.3373	0.2962	0.4799
	MUSIQ ↑	45.60	45.89	32.68	42.42	49.95	44.23	43.30	37.49	**50.48**

Types of Prior for Restoration

The importance of prior

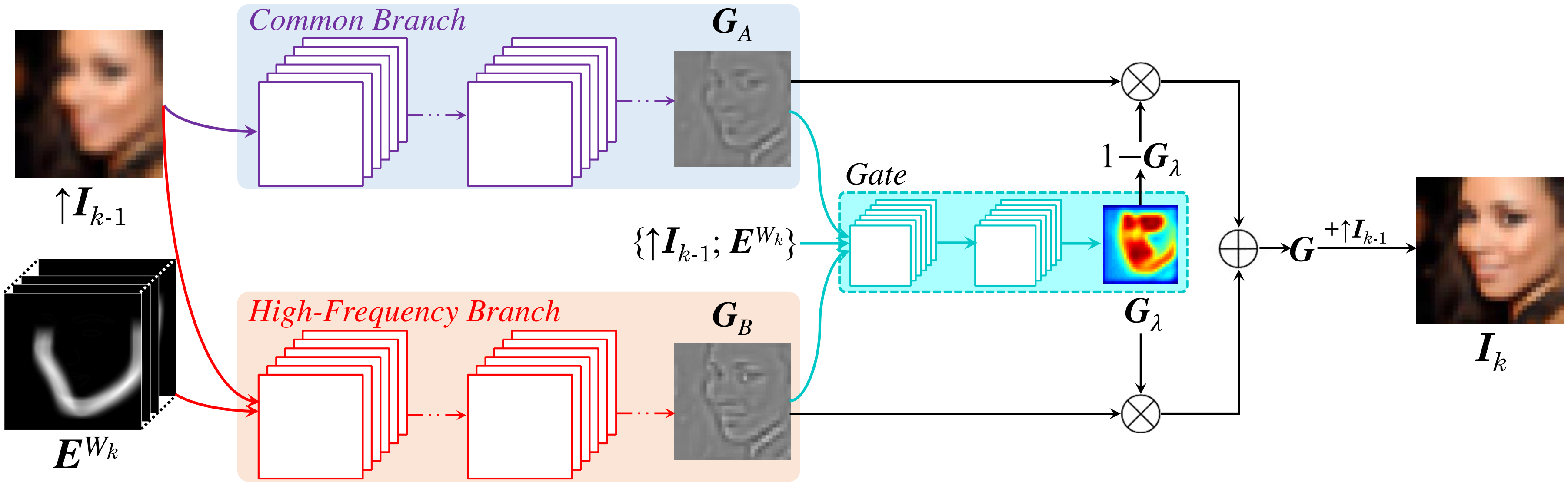


Xintao Wang et al. Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform, CVPR 2018

Existing priors (using face restoration as example)

- Geometric priors
  - Facial semantic map
  - Facial component heatmap
  - Facial 3D shape
  - …
- Reference priors
  - Similar faces
  - Facial component dictionaries
  - …
- Generative priors
  - Pre-trained face generator, e.g., StyleGAN2
  - …

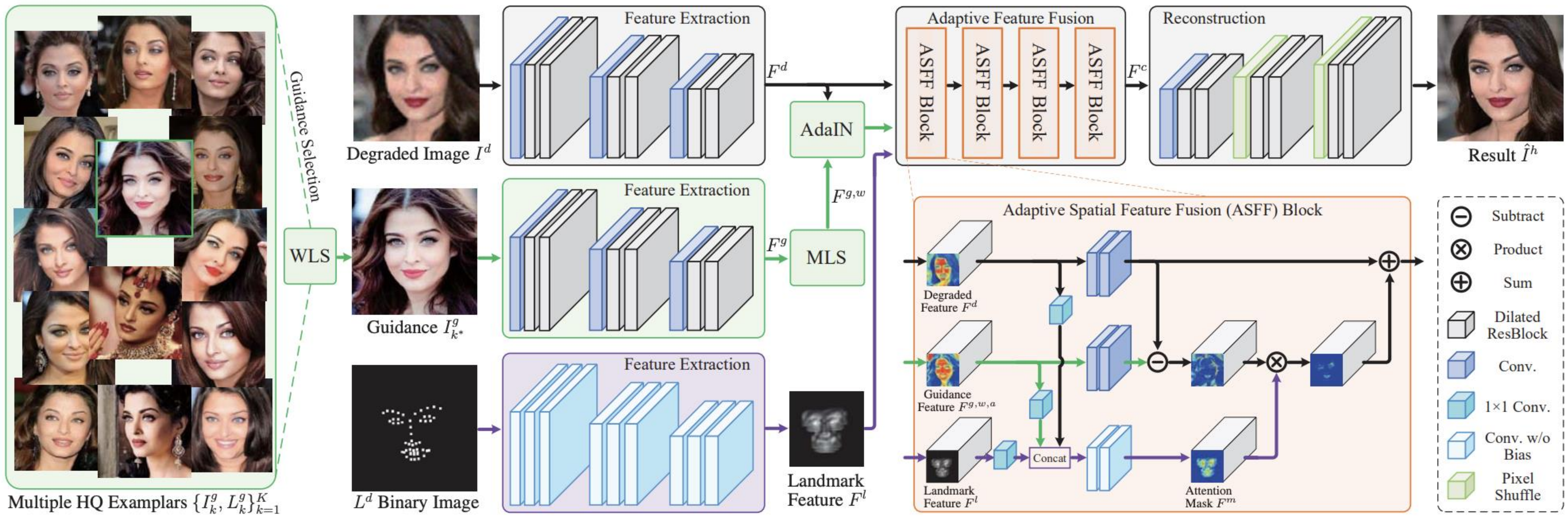
Geometric prior



High-frequency prior indicates the location with high-frequency details

S. Zhu, S. Liu, C. C. Loy, X. Tang, Deep Cascaded Bi-Network for Face Hallucination, ECCV 2016

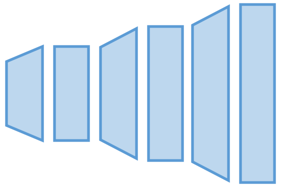
Reference prior



X. Li et al., Enhanced Blind Face Restoration with Multi-Exemplar Images and Adaptive Spatial Feature Fusion, CVPR 2020

Generative prior from GAN

$\mathbf{z} \sim \mathcal{N}(0, I)$



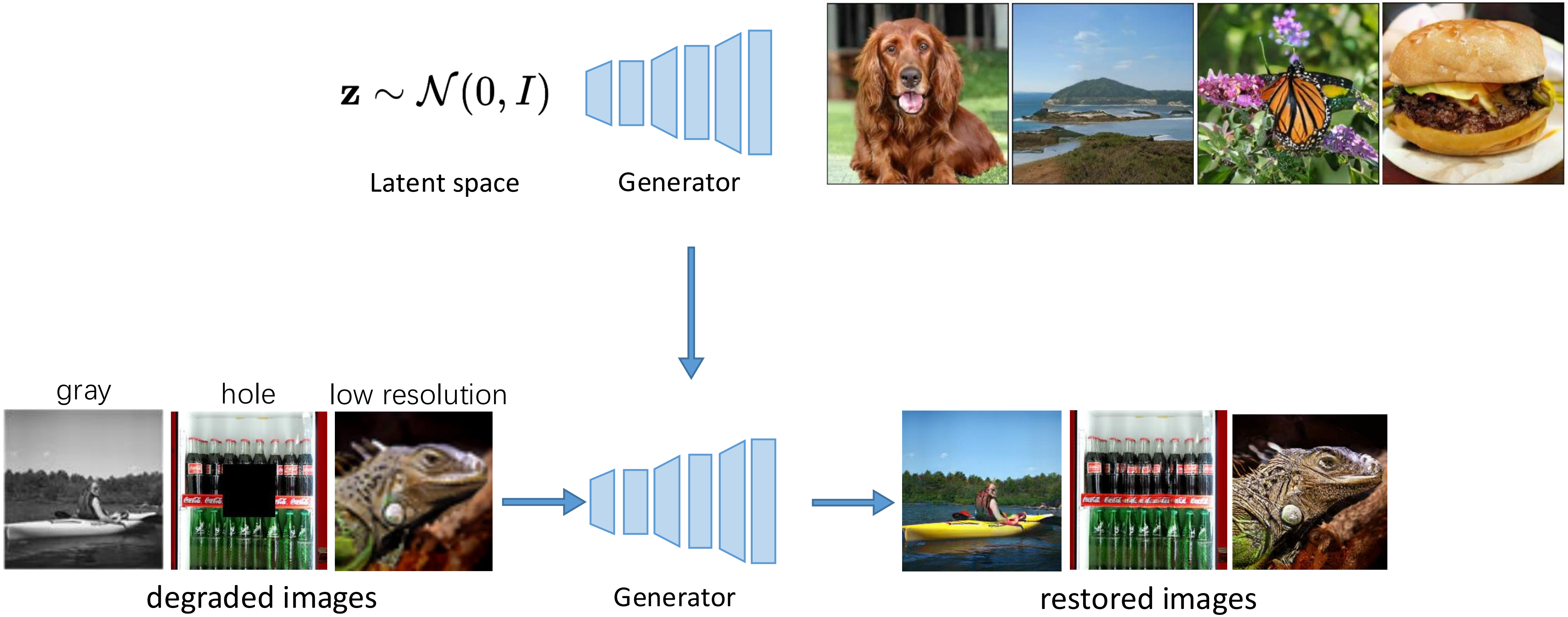
Latent space

Generator

Can we leverage a GAN trained on large-scale natural images for richer priors?

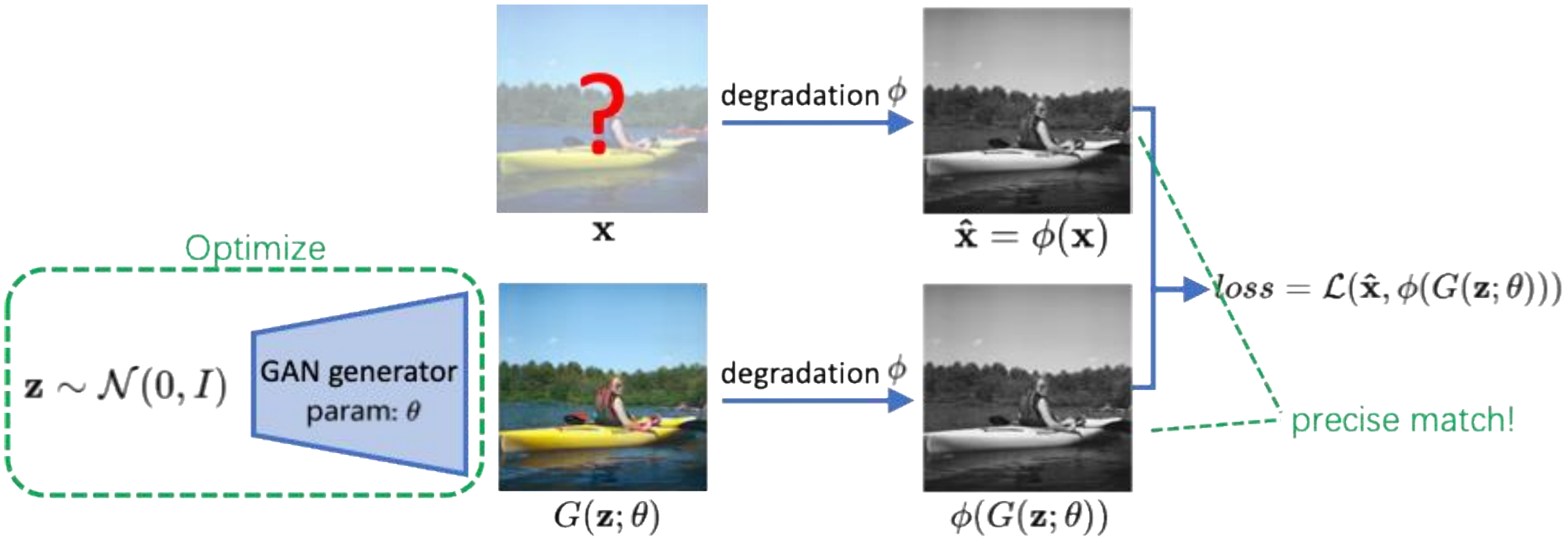
GAN is a good approximator for natural image manifold.

Generative prior from GAN



Xingang Pan et al., Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation, ECCV 2020 (Oral)

Generative prior from GAN

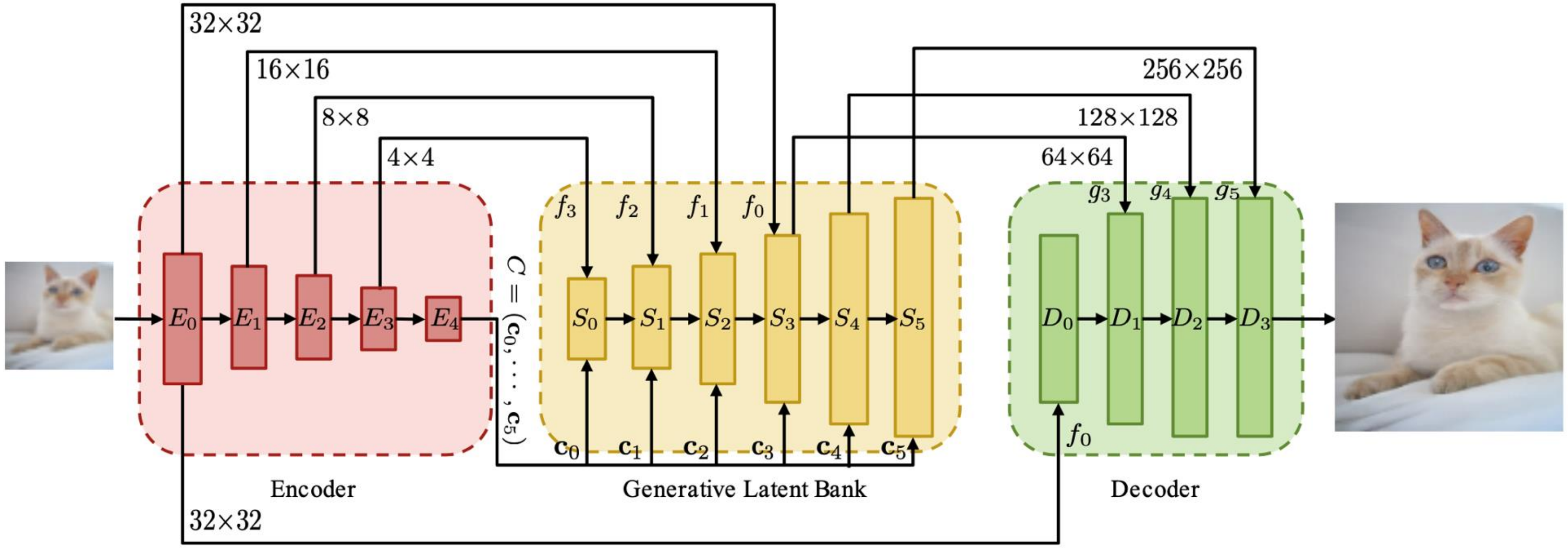


$$\theta^*, \mathbf{z}^* = argmin_{\theta,\mathbf{z}} \mathcal{L}(\hat{\mathbf{x}}, \phi(G(\mathbf{z};\theta)))$$ (Relaxed GAN-inversion)

Need to solve an expensive optimization problem

Xingang Pan et al., Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation, ECCV 2020 (Oral)

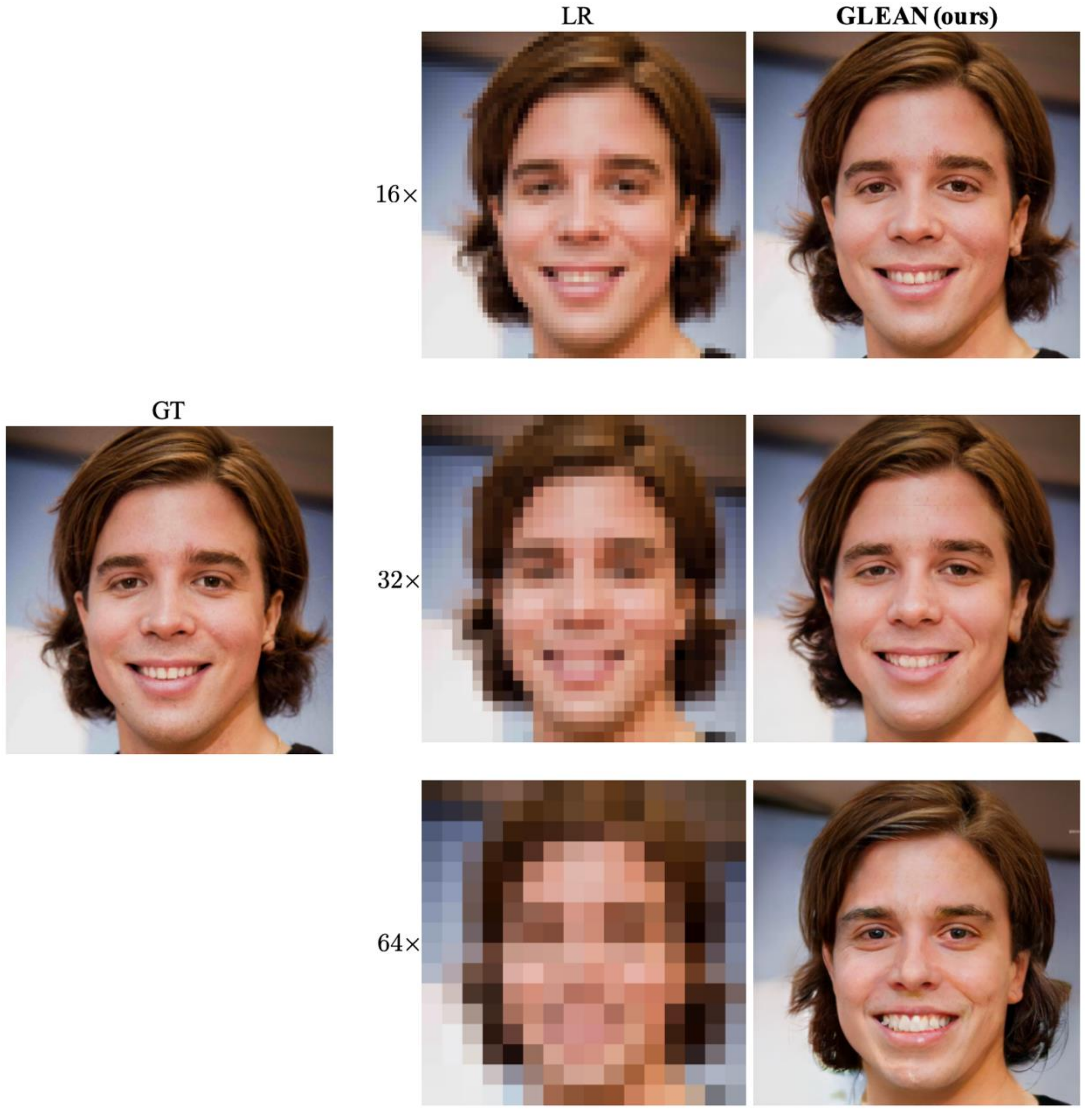
Generative prior from GAN

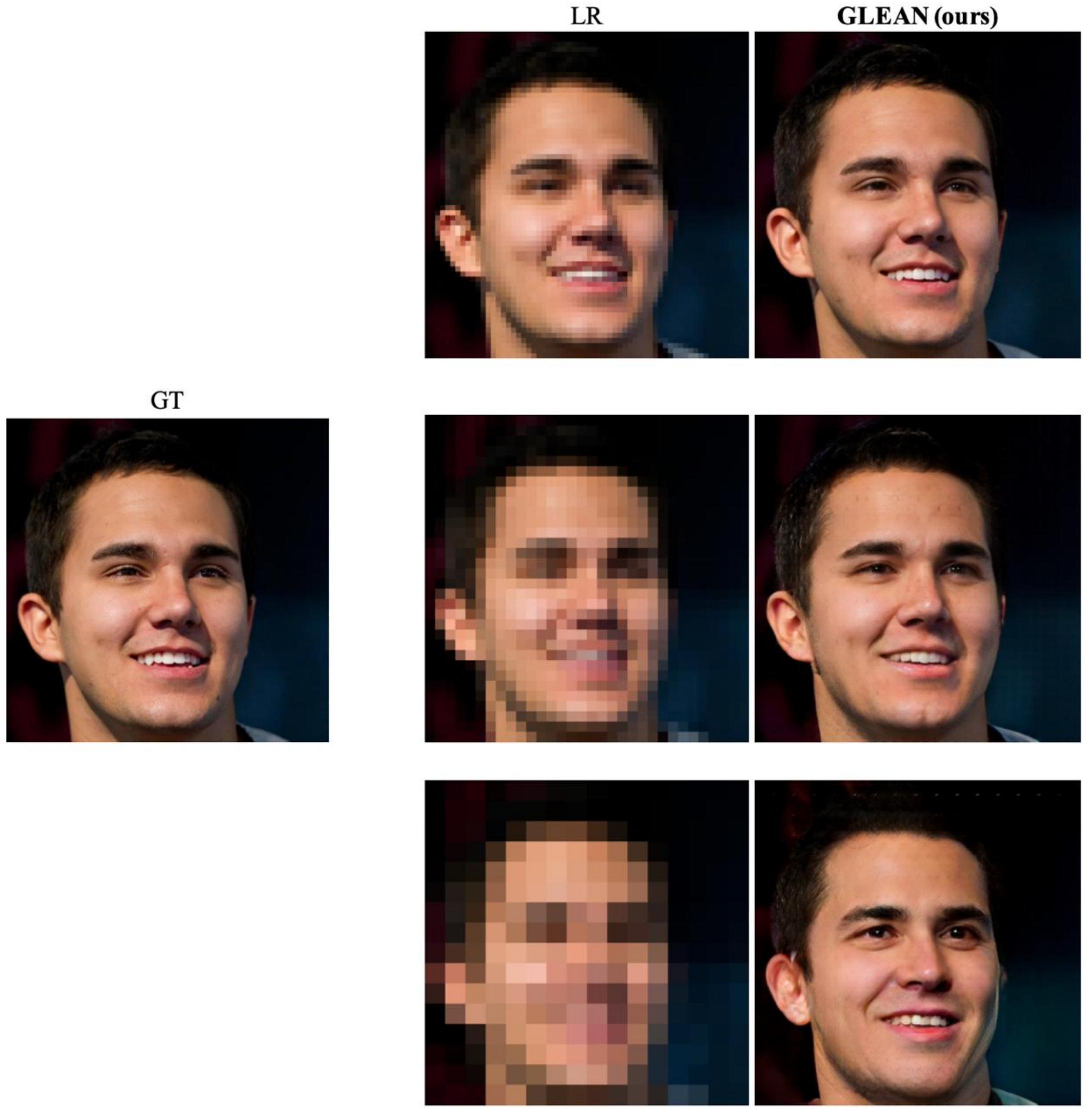


Condition the bank by passing both the latent vectors and multi-resolution convolutional features from the encoder to achieve high-fidelity results. Symmetrically, multi-resolution cues need to be passed from the bank to the decoder.

K. C. K. Chan, X. Wang, X. Xu, J. Gu, C. C. Loy, GLEAN: Generative Latent Bank for Image Super-Resolution and Beyond, TPAMI 2022

Generative prior from GAN





Generative prior from GAN

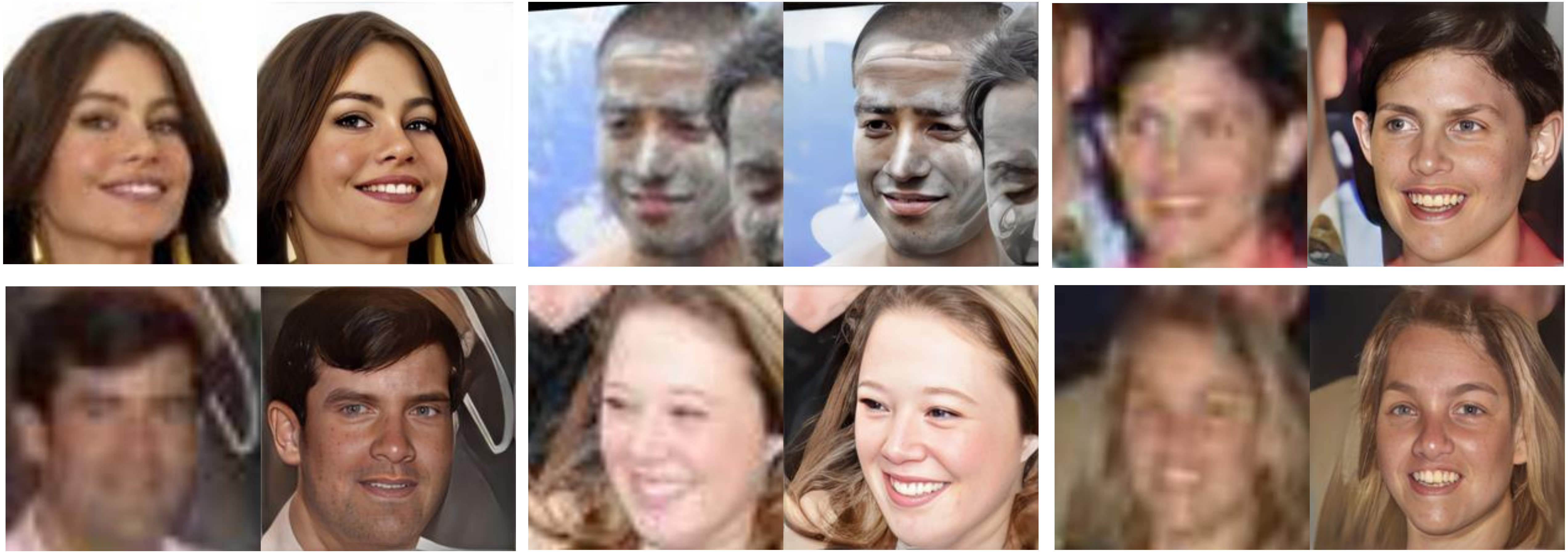
484x484

242x242

121x121

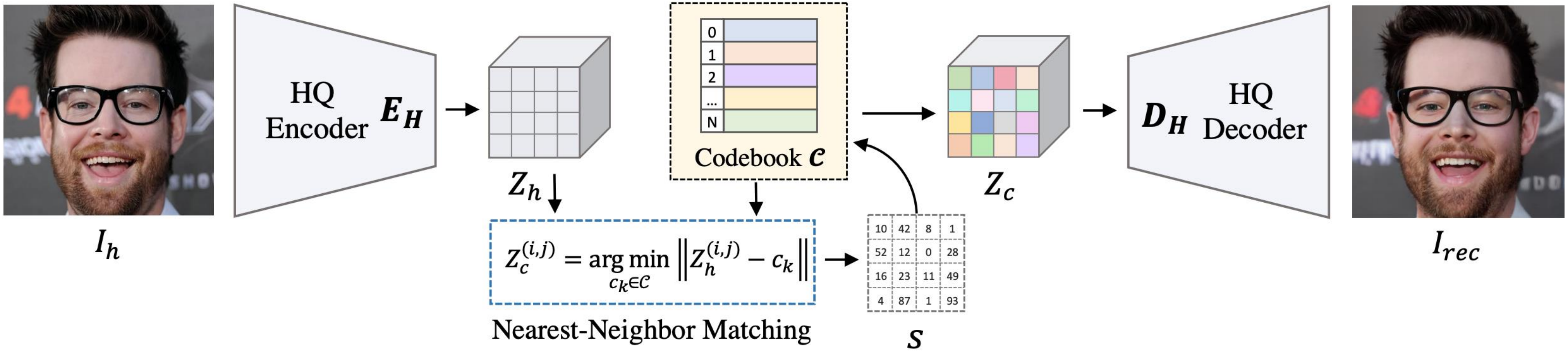
60x60

Generative prior from GAN



Discrete codebook prior

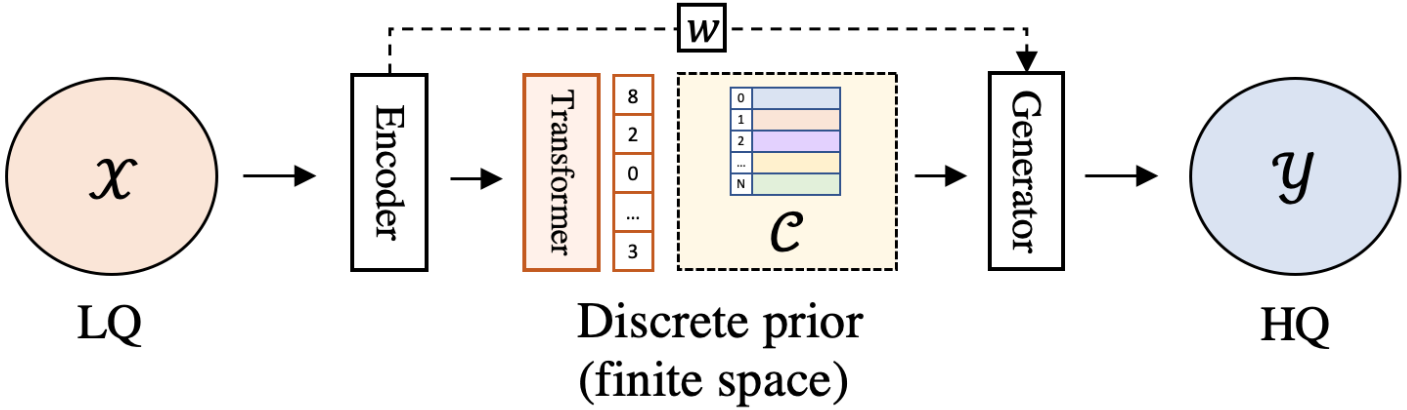
VQ-VAE is a type of variational autoencoder that uses vector quantisation to obtain a discrete latent representation. It differs from VAEs in two key ways: the encoder network outputs discrete, rather than continuous, codes; and the prior is learnt rather than static (the posteriors and priors in VAEs are assumed normally distributed with diagonal covariance).



[VQGAN] *Esser et al.*, Taming Transformers for High-Resolution Image Synthesis, CVPR 2021

[VQVAE] *Oord et al.*, Neural Discrete Representation Learning, NeurIPS 2017

Discrete codebook prior - CodeFormer



Learn discrete codebook prior in a small proxy space to reduce the uncertainty and ambiguity of restoration mapping by, while providing rich visual atoms for generating high-quality faces.

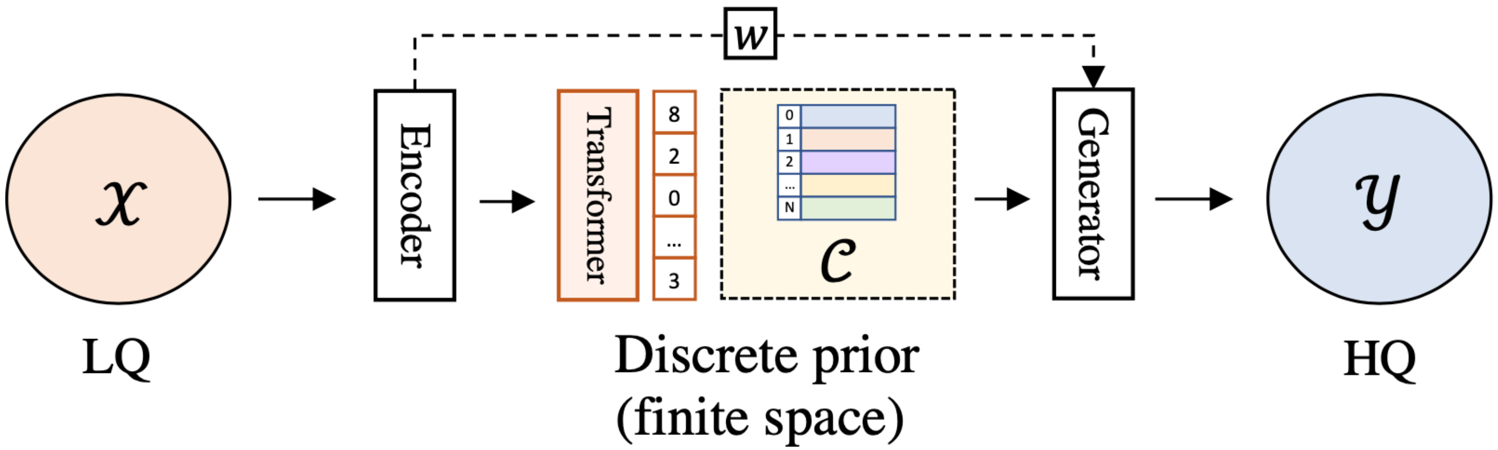
Cast blind face restoration as a code prediction task

A Transformer-based prediction network to model the global composition and context of the low-quality faces for code prediction

Enable the discovery of natural faces that closely approximate the target faces even when the inputs are severely degraded

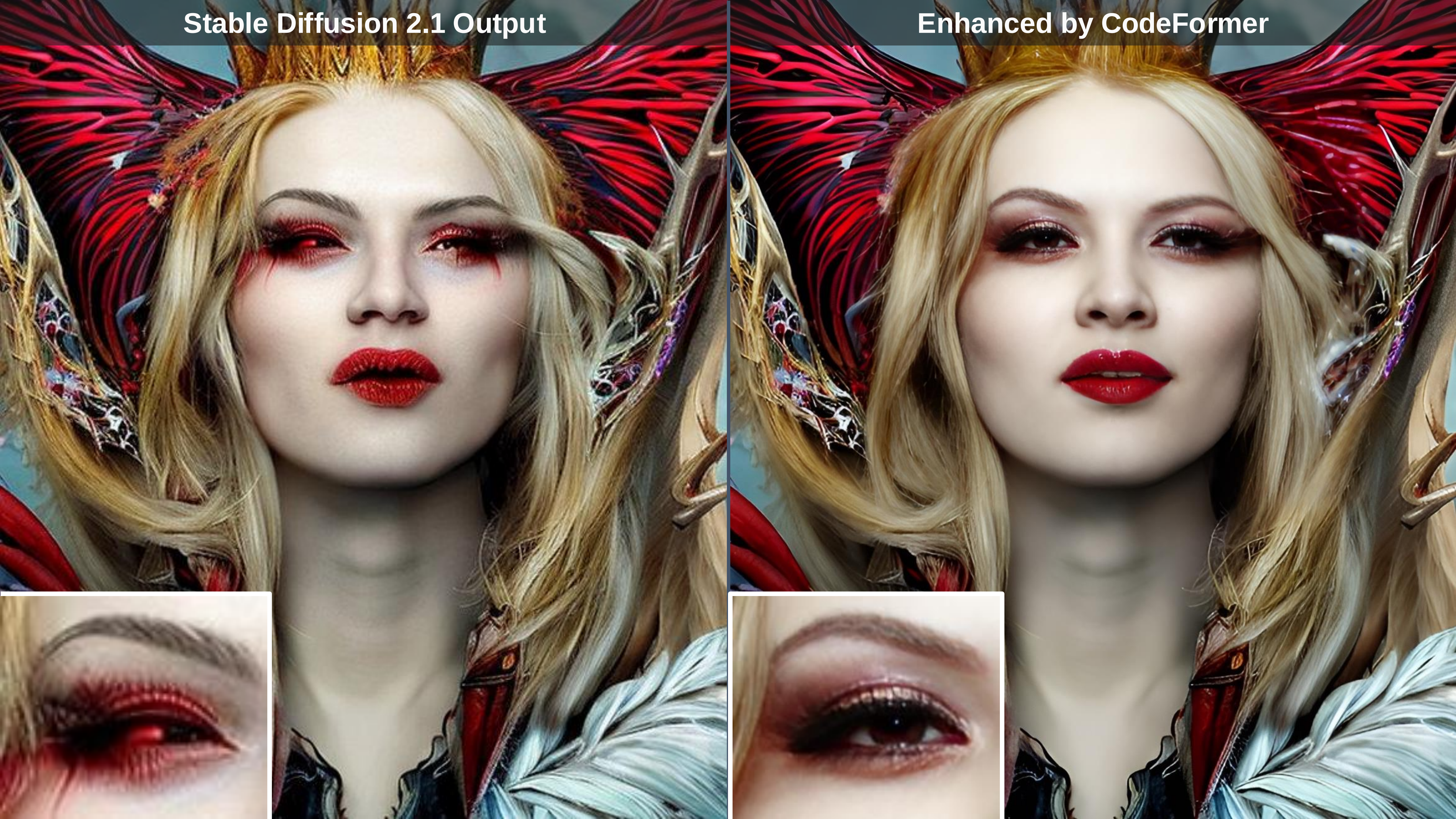
Shangchen Zhou et al., Towards Robust Blind Face Restoration with Codebook Lookup Transformer, NeurIPS 2022

Discrete codebook prior - CodeFormer



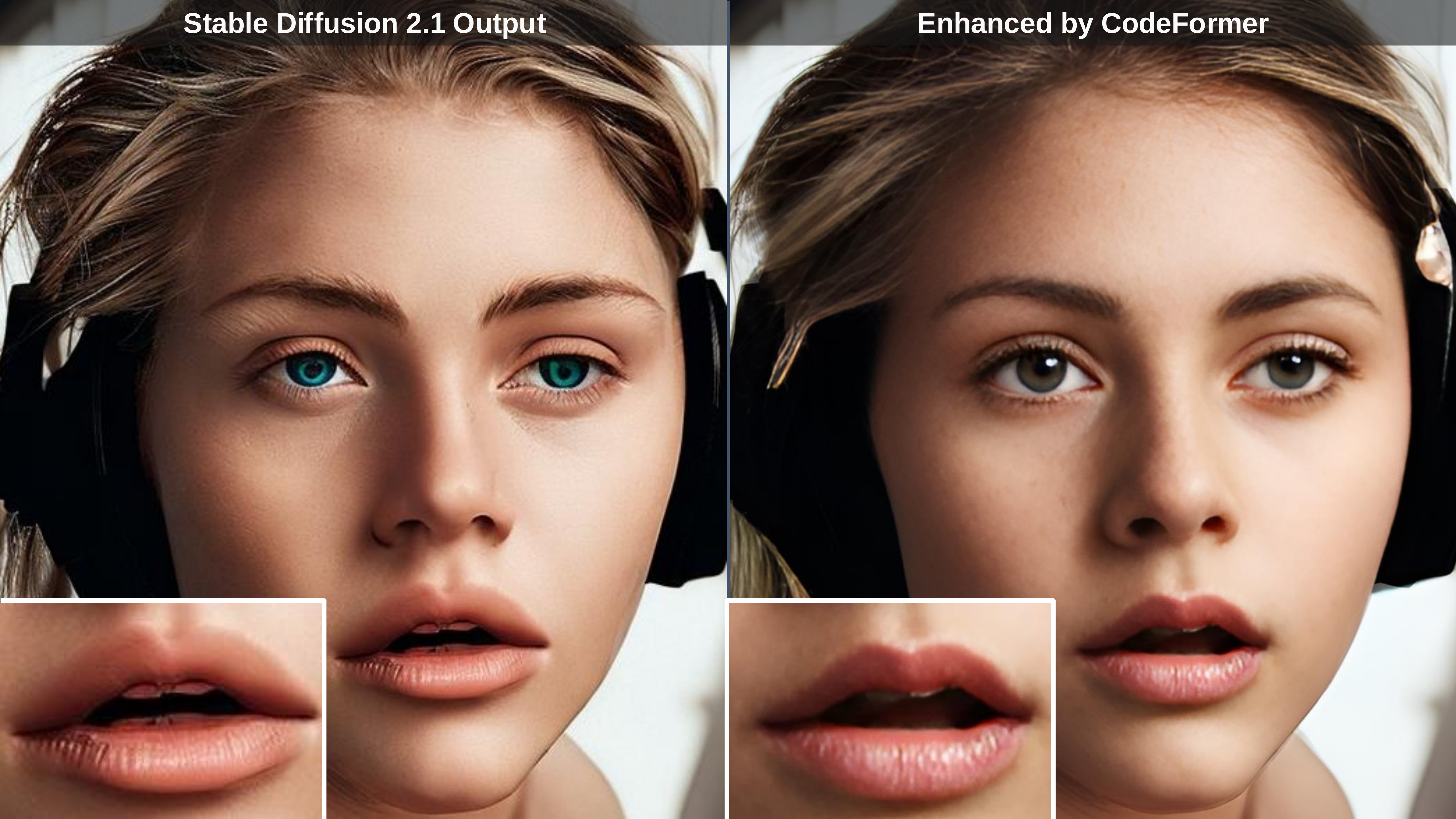
Shangchen Zhou et al., Towards Robust Blind Face Restoration with Codebook Lookup Transformer, NeurIPS 2022


Stable Diffusion 2.1 Output
Enhanced by CodeFormer


Stable Diffusion 2.1 Output
Enhanced by CodeFormer

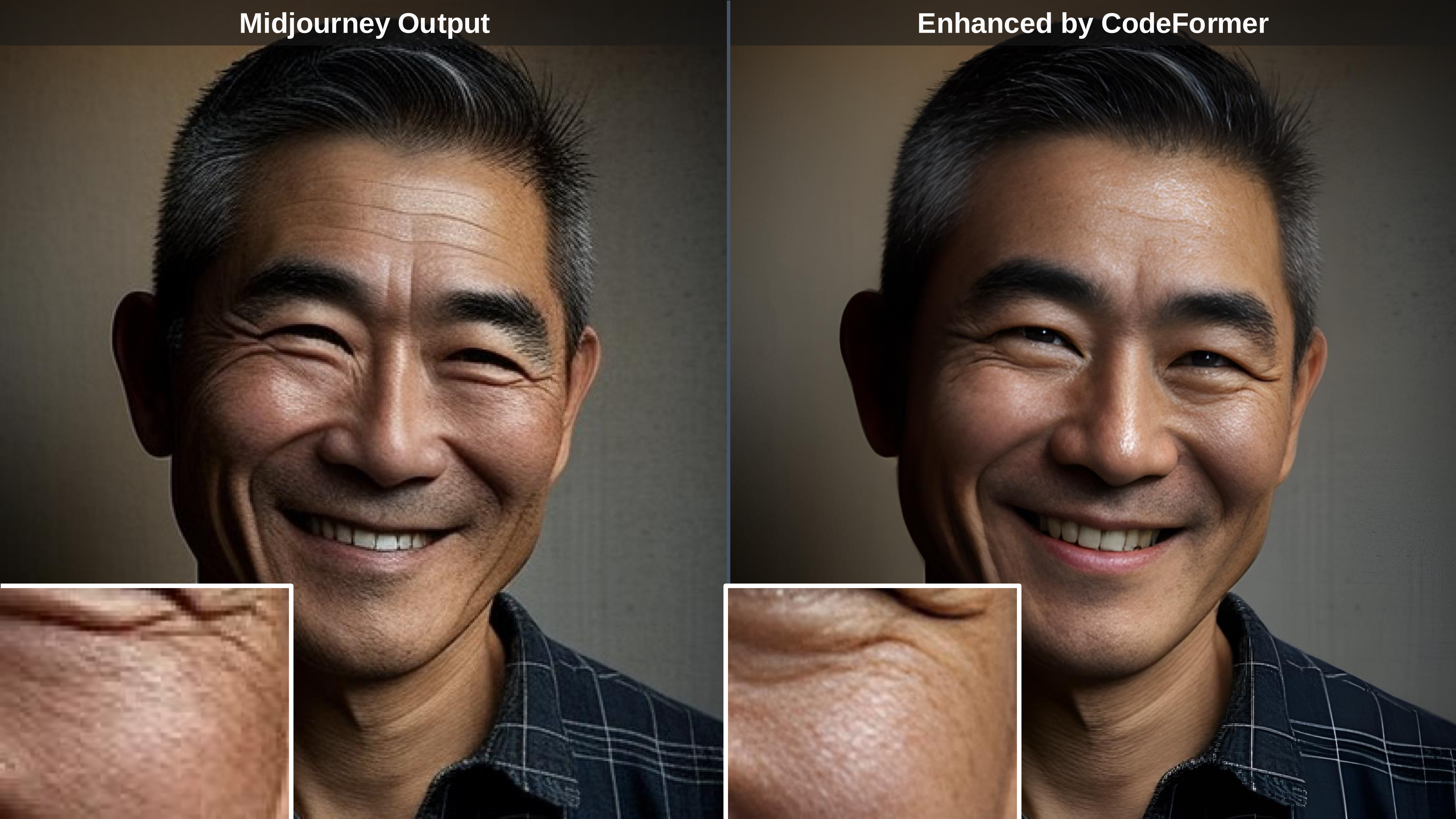
Stable Diffusion 2.1 Output

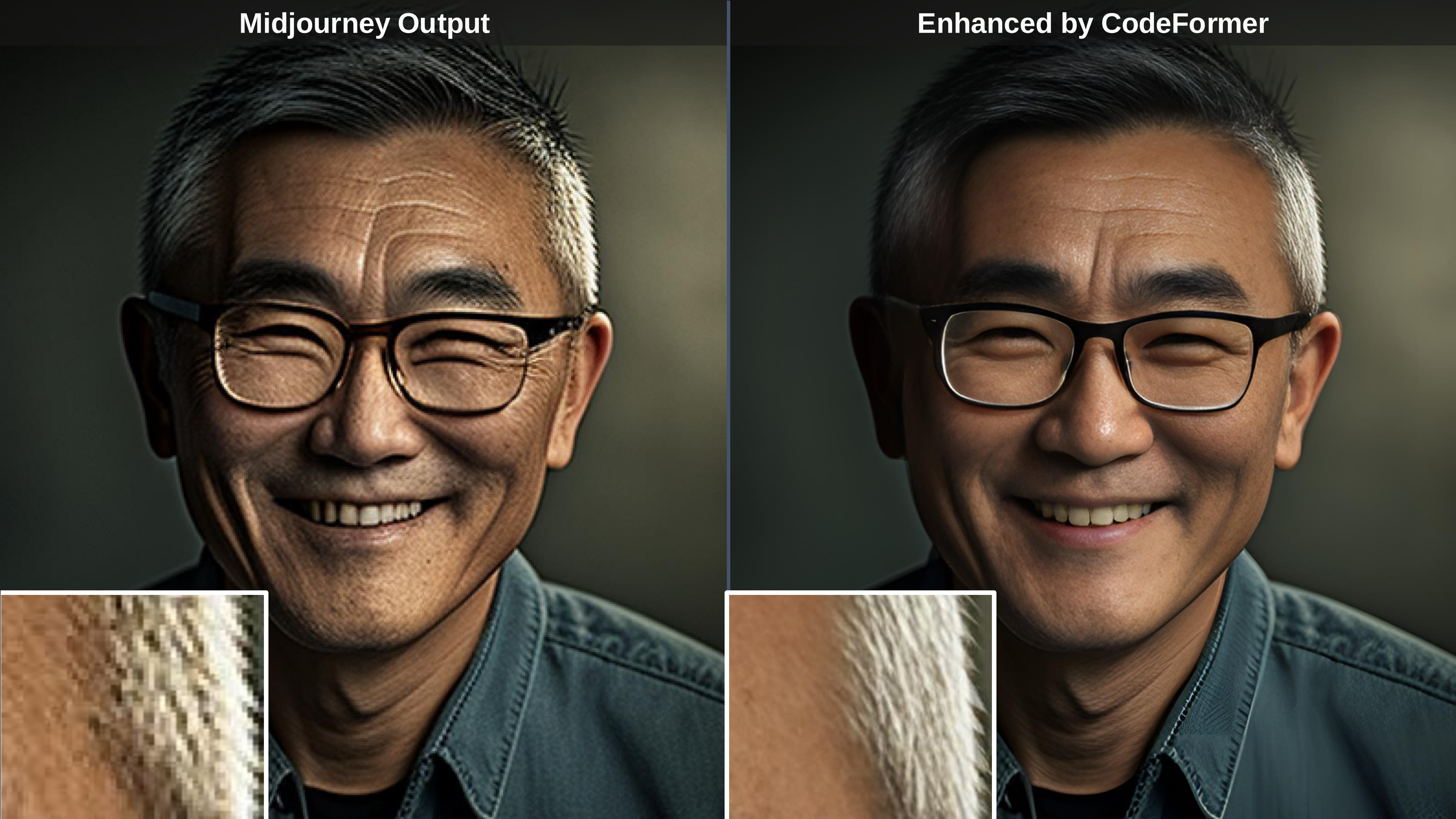
Enhanced by CodeFormer


Stable Diffusion 2.1 Output
Enhanced by CodeFormer

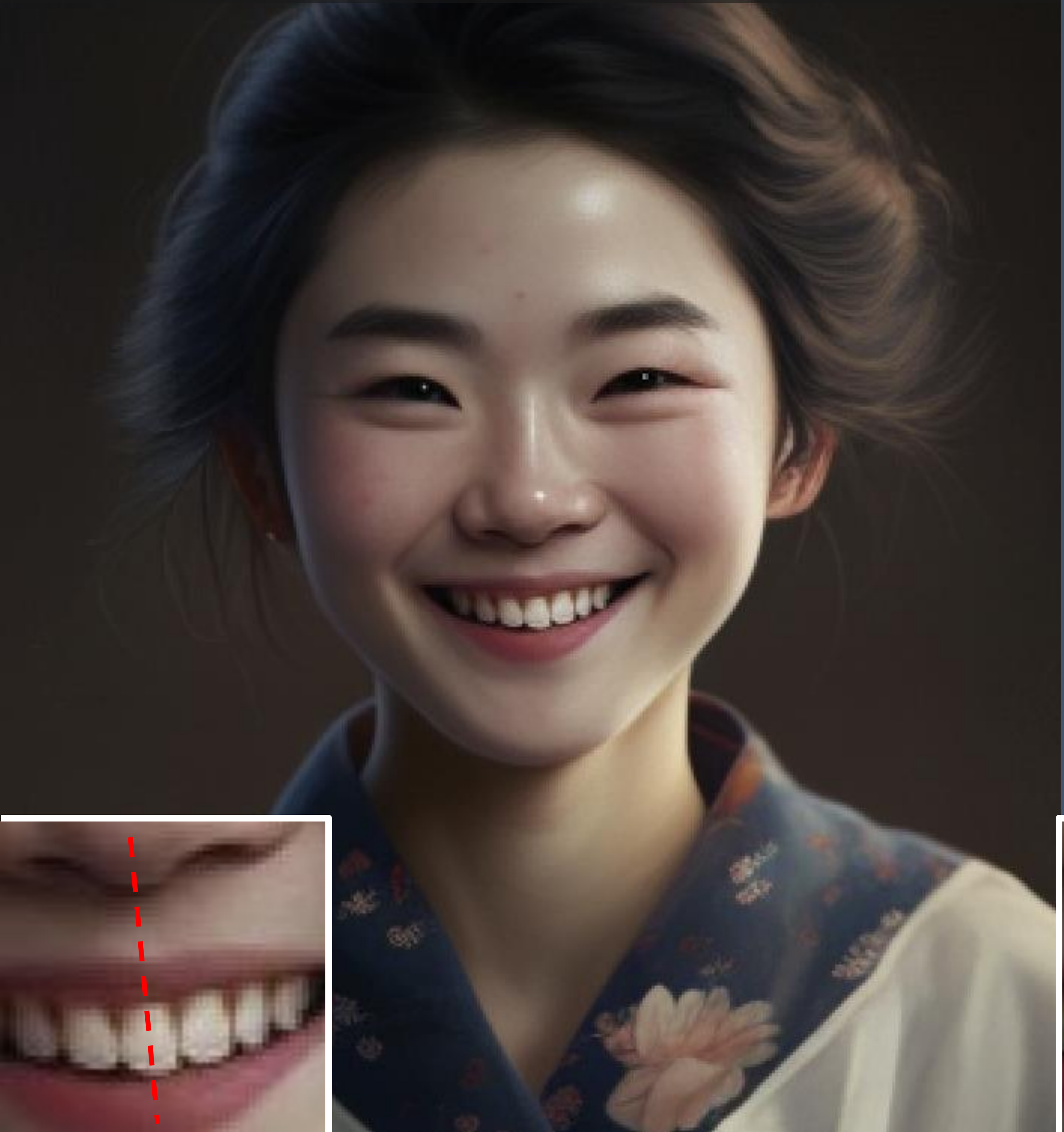
Midjourney Output

Enhanced by CodeFormer


Midjourney Output
Enhanced by CodeFormer


Midjourney Output
Enhanced by CodeFormer

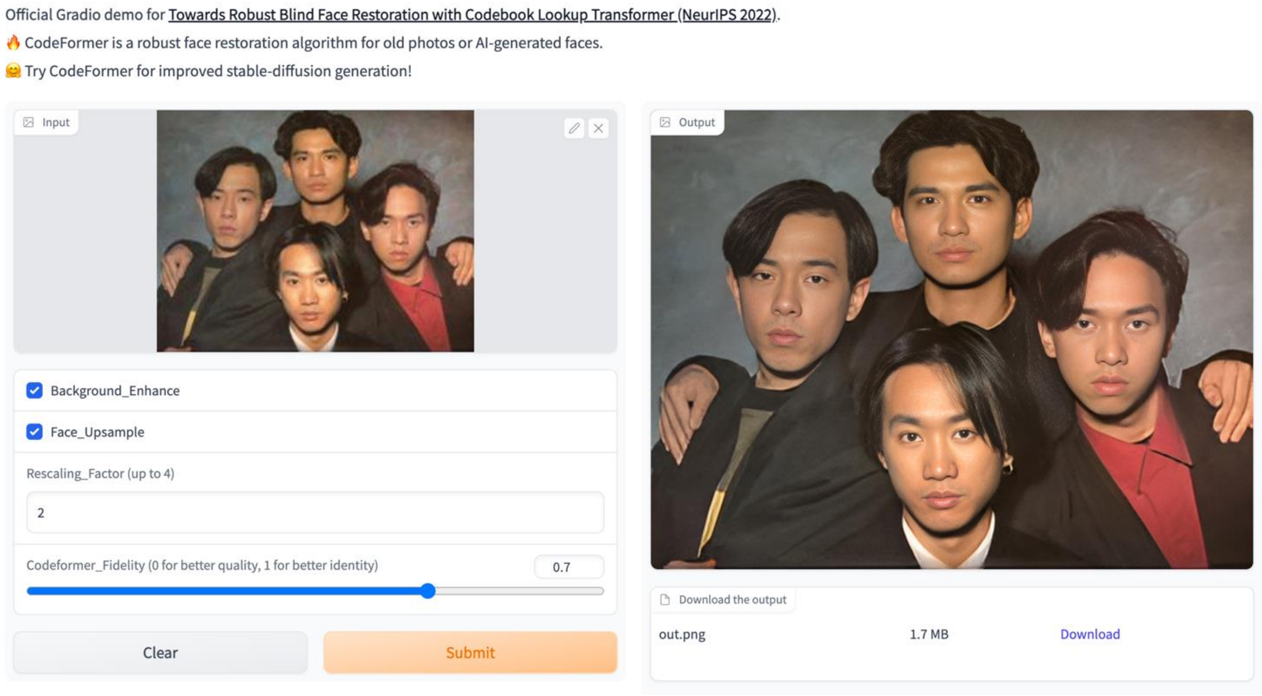
Midjourney Output



Enhanced by CodeFormer

Over 62,000,000 World-wide Access

Code and demo

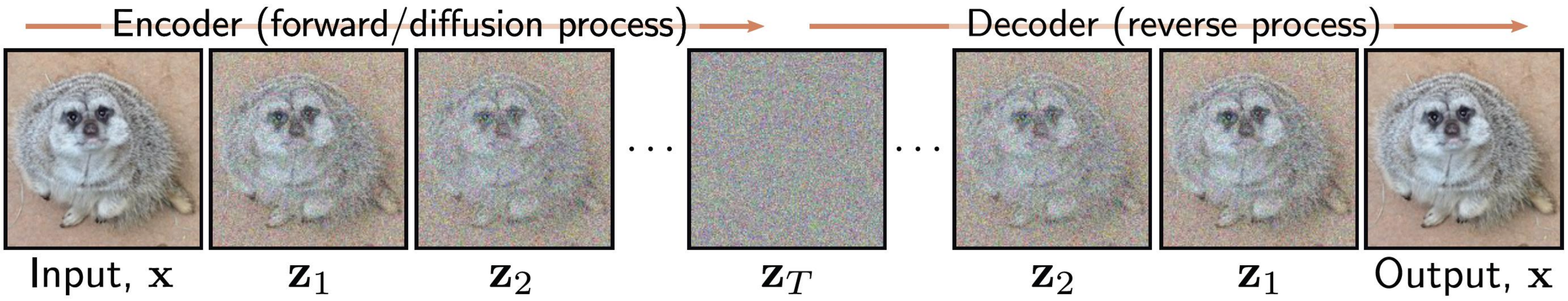


https://github.com/sczhou/CodeFormer

https://huggingface.co/spaces/sczhou/CodeFormer

Diffusion Prior

More Generic Prior from Diffusion Models?



It is unclear how restoration can be achieved via diffusion model

- Diffusion model is stochastic! How to keep the prior and maintain fidelity?
- Diffusion model hasn't seen relevant degradations! How to handle complex degradations?
- Diffusion model is slow! How to improve inference efficiency?

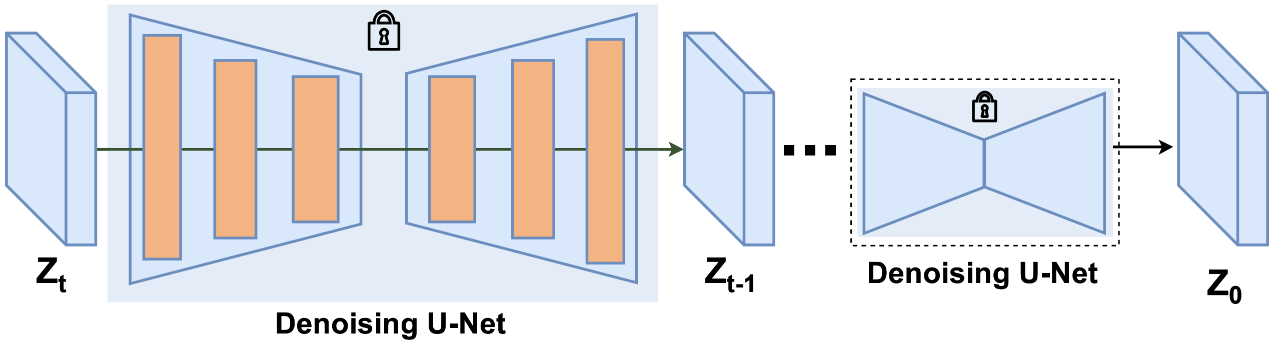
Image Credit: Simon J.D. Prince, Understanding Deep Learning, 2023

StableSR | Framework

Keeping the prior and fidelity

- Frozen stable diffusion model as a backbone

- Minimal alterations to prevent disrupting the prior



Jianyi Wang et al. Exploiting Diffusion Prior for Real-World Image Super-Resolution. IJCV 2024

StableSR | Framework

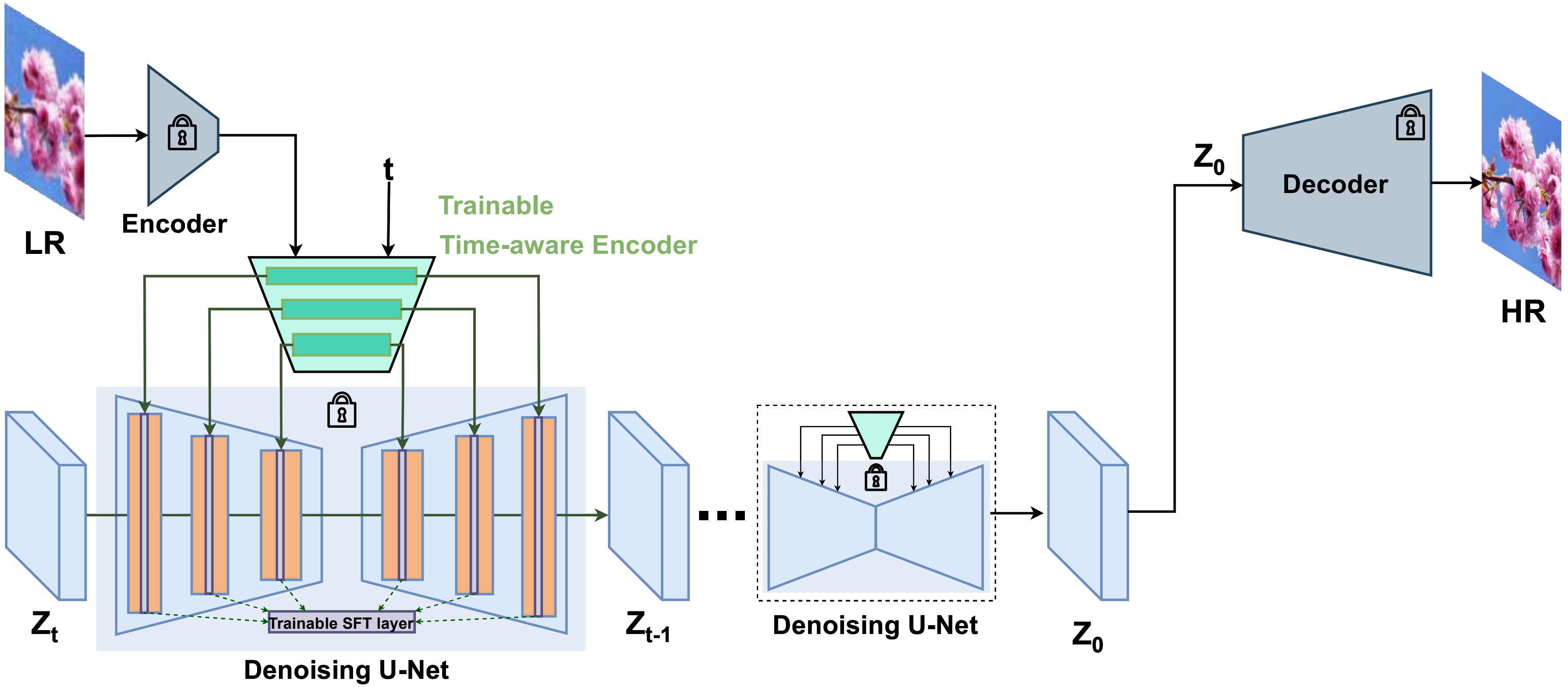
Keeping the prior and fidelity

- Train only the time-aware encoder and spatial feature transformation layer

$$\boldsymbol{\alpha}^n, \boldsymbol{\beta}^n = \mathcal{M}_\theta^n(\mathbf{F}^n)$$

$$\hat{\mathbf{F}}^n_{\text{dif}} = (1 + \boldsymbol{\alpha}^n) \odot \mathbf{F}^n_{\text{dif}} + \boldsymbol{\beta}^n$$

- Adaptively adjust the condition strength derived from the LR feature through *t*

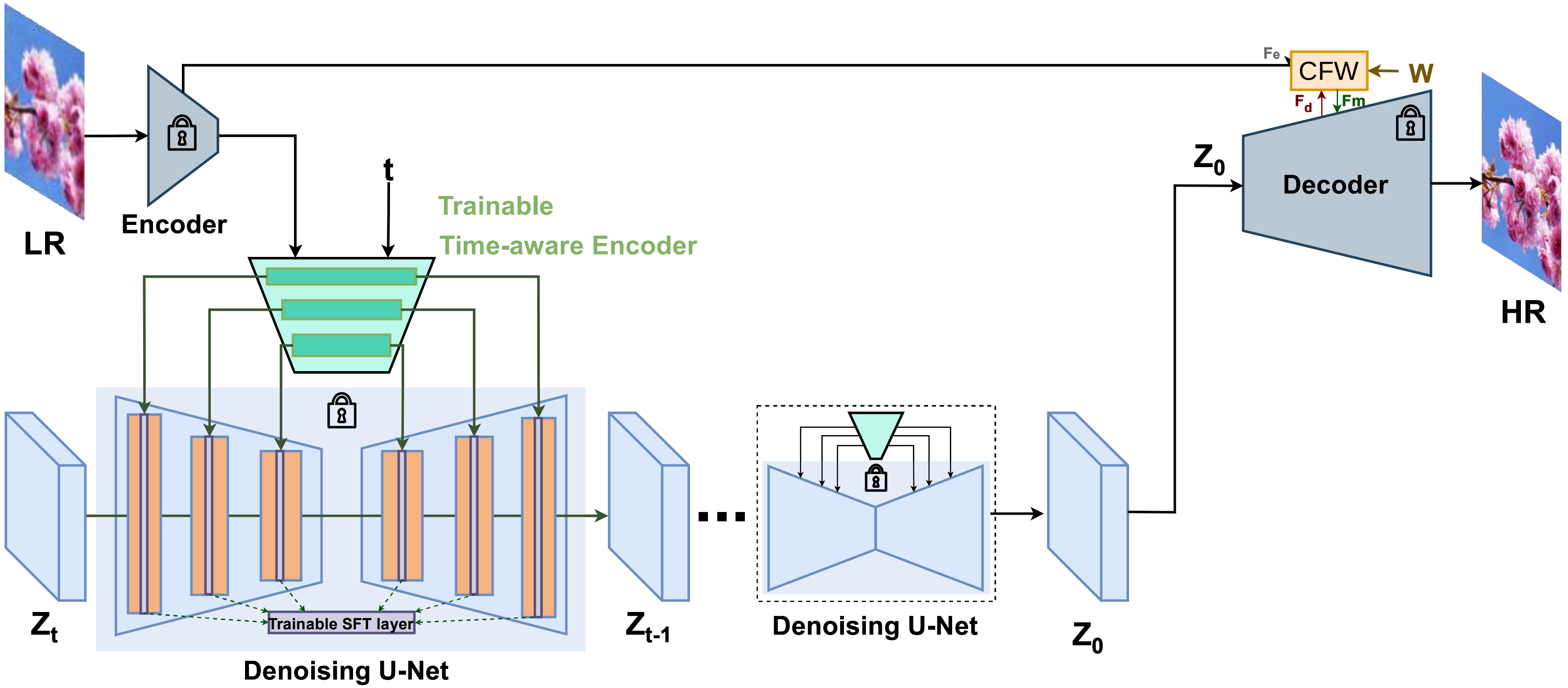


Xintao Wang et al. Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform. CVPR 2018

StableSR | Fidelity-Realism Trade-off

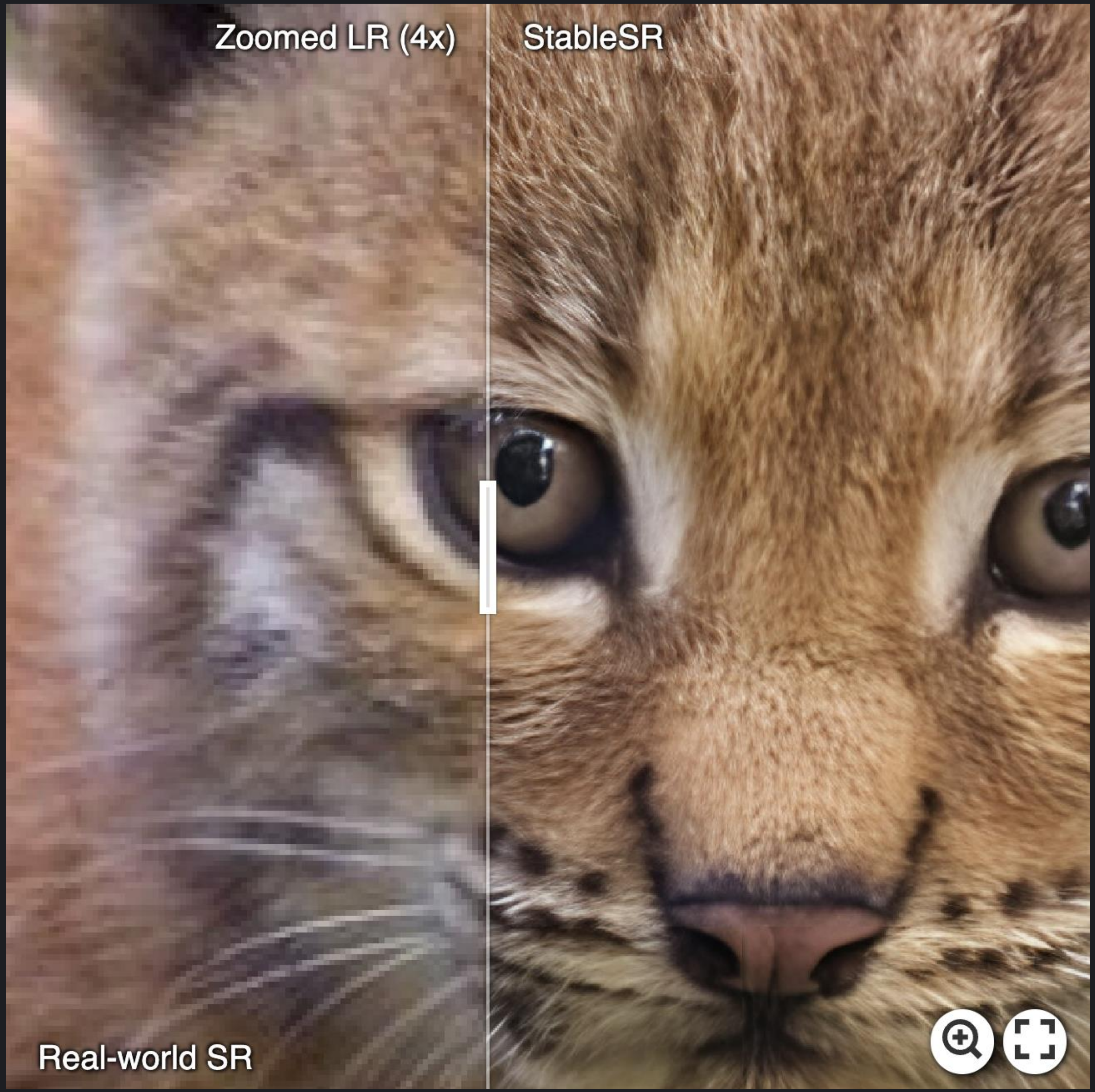
Keeping the prior and fidelity

- Add a controllable skip connection to benefit from structural guidance from the LR image, enhancing fidelity
- Control the modulation strength through **w**
- A larger **w** allows stronger structural guidance



Shangchen Zhou et al. Towards Robust Blind Face Restoration with Codebook Lookup Transformer. NeurIPS 2022

Zoomed LR (4x)
StableSR
Real-world SR

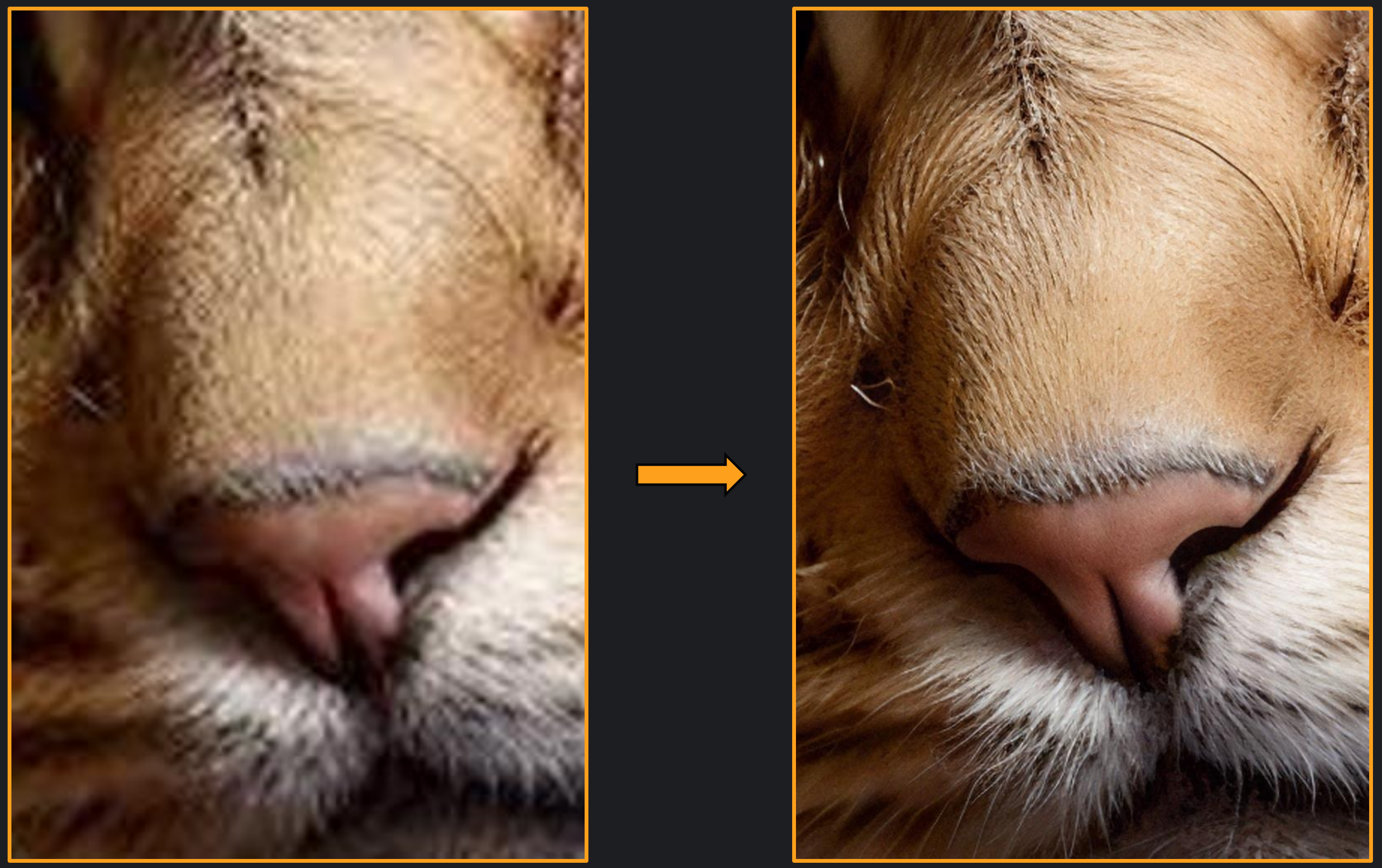


Zoomed LR (4x)
StableSR
Real-world SR


Zoomed LR (4x)
StableSR
Old Photo SR

Zoomed LR (4x)
StableSR
Old Photo SR


Zoomed LR (4x)
StableSR
AIGC SR (1024x1536 to 4096x6144)



PGDiff | A Versatile Solution

(a) Blind Face Restoration

(b) Face Colorization

(c) Face Inpainting

(d) Old Photo Restoration (w/o scratches)

(e) Old Photo Restoration (w/ scratches)

Peiqing Yang. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. NeurIPS 2023

PGDiff | Motivation

- Modelling degradation is hard
- Just model easily accessible properties, e.g., image structure and color statistics of high-quality images
- Apply this guidance during the reverse diffusion process
- Inspired by *classifier guidance*, which is originally used by class-conditional generation (see next two slides)

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. NeurIPS, 2021.

Conditional generation

If the data has associated labels c, these can be exploited to control the generation.

How about modifying the denoising update from $\mathbf{z}_t$ to $\boldsymbol{z}_{t-1}$ to take into account class information c?

Adding an extra term into the update step during the reverse process to bias the denoising update toward that class

Dhariwal & Nichol, [Diffusion Models Beat GANs on Image Synthesis](), NeurIPS 2021.

Conditional generation

Algorithm - Sampling

Input: Model, $\mathbf{g}_t[\bullet, \phi_t]$
Output: Sample, $\mathbf{x}$
$\mathbf{z}_T \sim \text{Norm}_{\mathbf{z}}[\mathbf{0}, \mathbf{I}]$ // Sample last latent variable
for $t = T \ldots 2$ **do**
  $\hat{\mathbf{z}}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\mathbf{g}_t[\mathbf{z}_t, \phi_t]$ // Predict previous latent variable
  $\epsilon \sim \text{Norm}_{\epsilon}[\mathbf{0}, \mathbf{I}]$ // Draw new noise vector
  $\boxed{\mathbf{z}_{t-1} = \hat{\mathbf{z}}_{t-1} + \sigma_t \epsilon}$ // Add noise to previous latent variable
$\mathbf{x} = \frac{1}{\sqrt{1-\beta_1}}\mathbf{z}_1 - \frac{\beta_1}{\sqrt{1-\alpha_1}\sqrt{1-\beta_1}}\mathbf{g}_1[\mathbf{z}_1, \phi_1]$ // Generate sample from $\mathbf{z}_1$ without noise

---

Classifier guidance

A classifier learns to identify the category of object being synthesized at each step

This is used to bias the denoising update toward that class

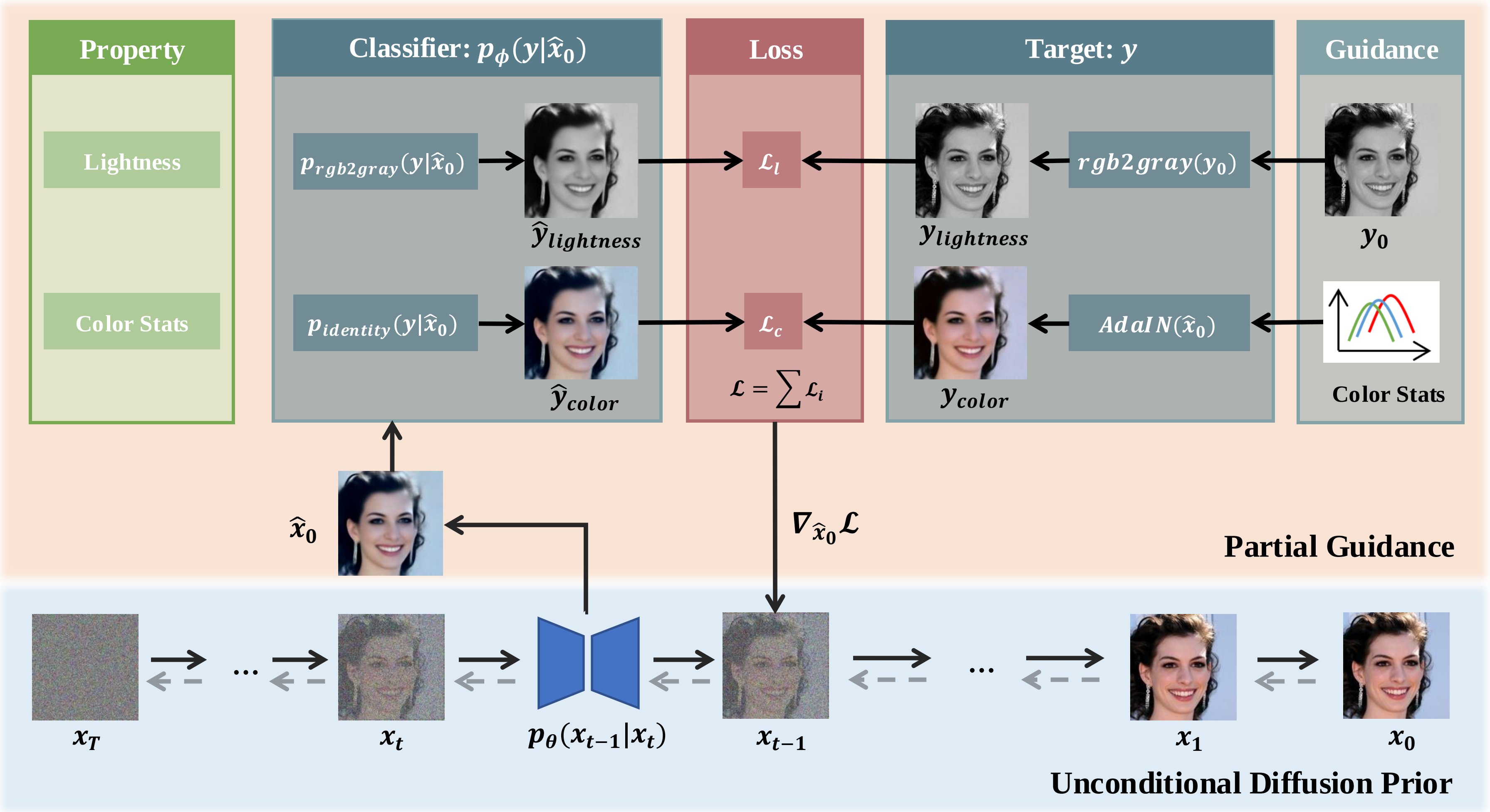
$$\mathbf{z}_{t-1} = \hat{\mathbf{z}}_{t-1} + \sigma_t^2 \underbrace{\frac{\partial \log\left[\Pr(c|\mathbf{z}_t)\right]}{\partial \mathbf{z}_t}}_{\text{gradient of the log likelihood of an auxiliary classifier model}} + \sigma_t \epsilon.$$

The update from $\mathbf{z}_t$ to $\mathbf{z}_{t-1}$ now makes the class c more likely

Like the U-Net, it is usually shared across all time steps and takes time as an input.

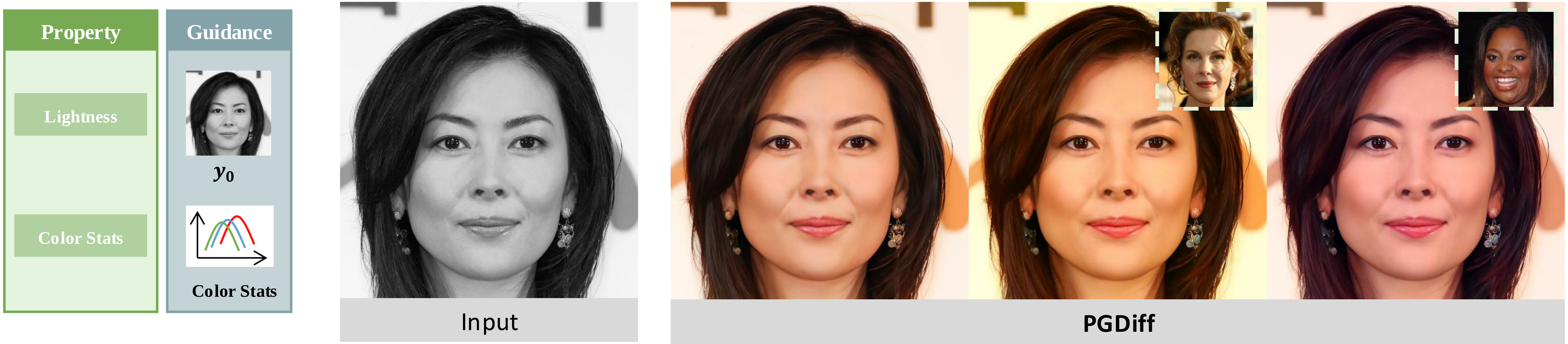
Dhariwal & Nichol, Diffusion Models Beat GANs on Image Synthesis, NeurIPS 2021.

PGDiff | Framework



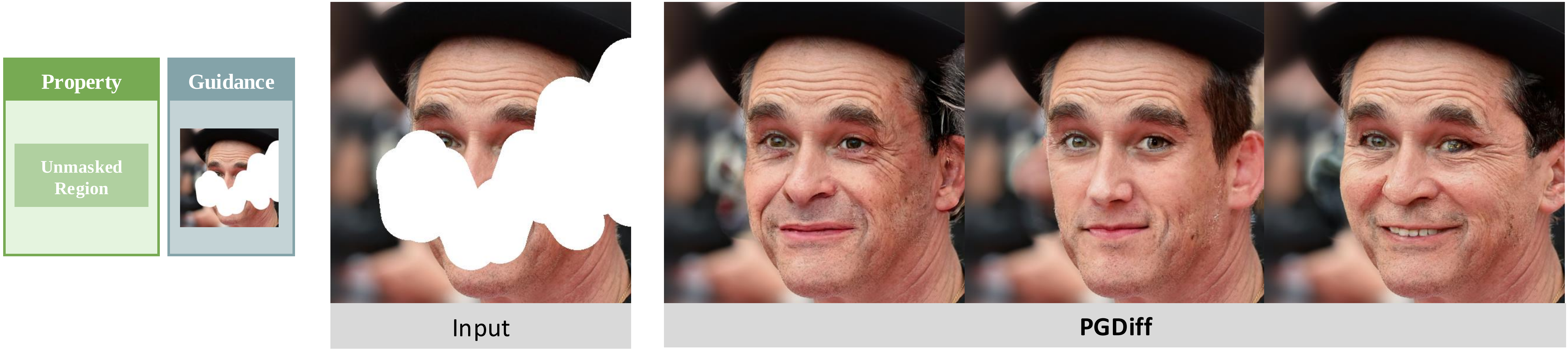
Peiqing Yang. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. NeurIPS 2023

PGDiff | Face Colorization



Peiqing Yang. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. NeurIPS 2023

PGDiff | Face Inpainting



Peiqing Yang. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. NeurIPS 2023

PGDiff | Blind Face Restoration

Property	Guidance
Smooth Semantics	

Input

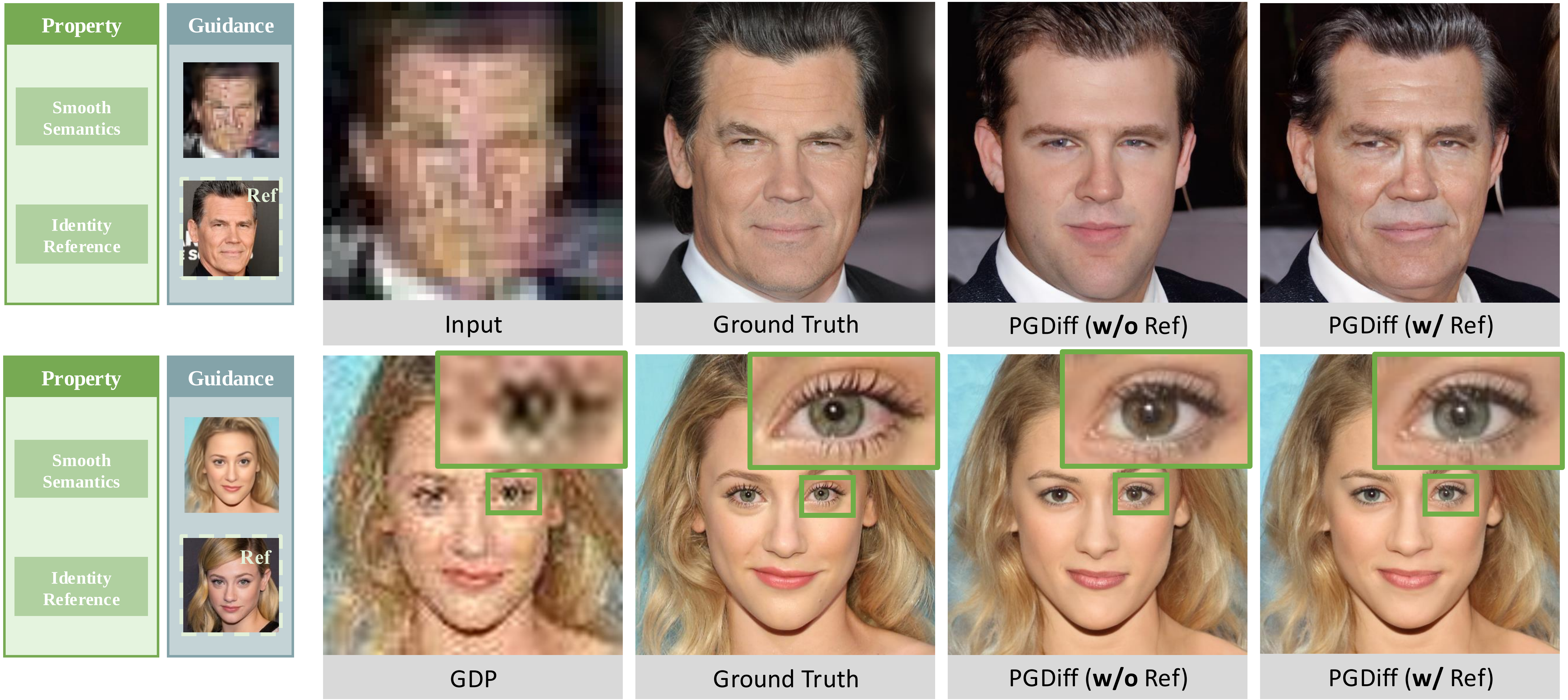
PGDiff

Peiqing Yang. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. NeurIPS 2023

PGDiff | Blind Face Restoration

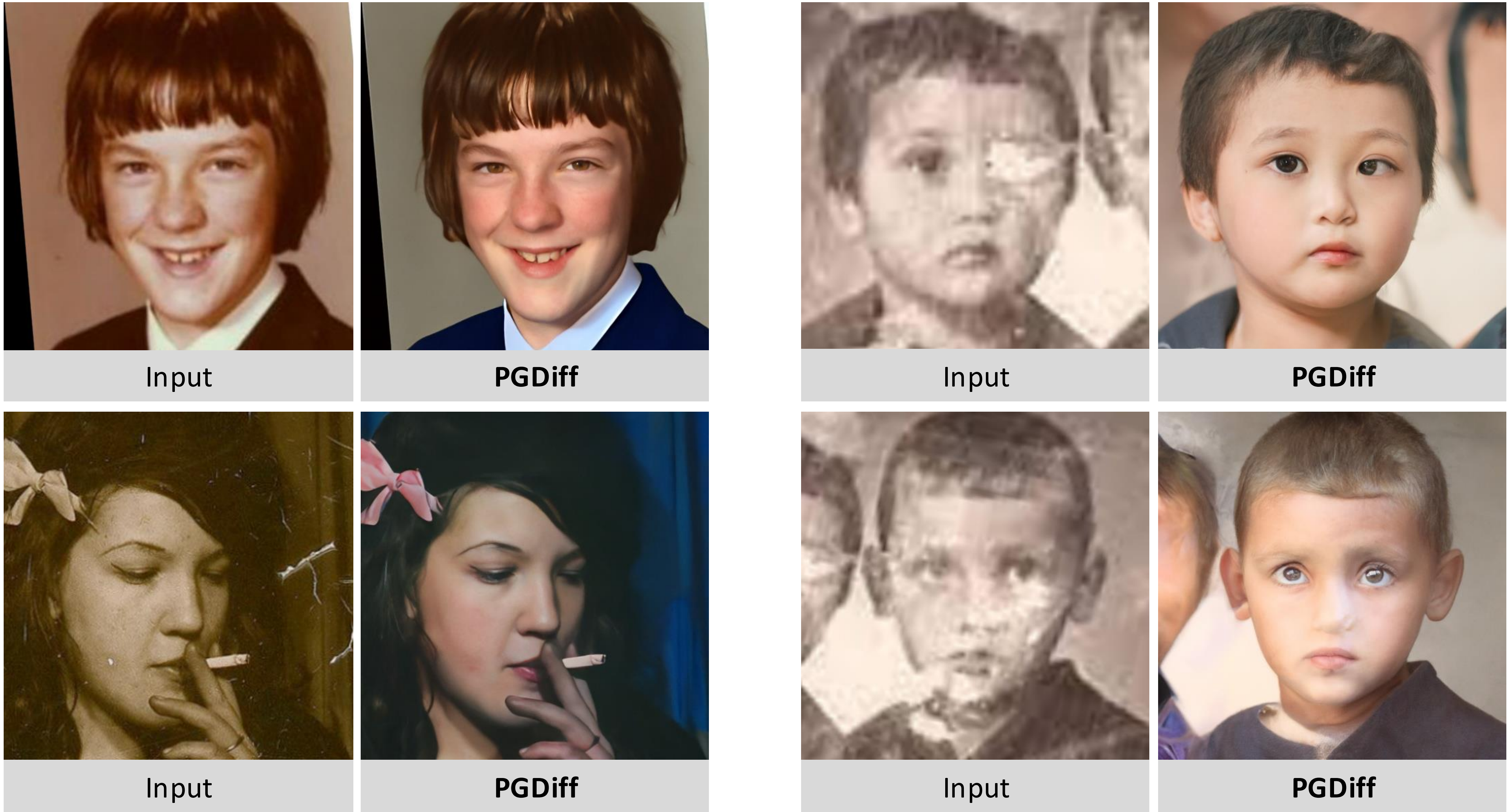
Peiqing Yang. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. NeurIPS 2023

PGDiff | Reference-based Restoration



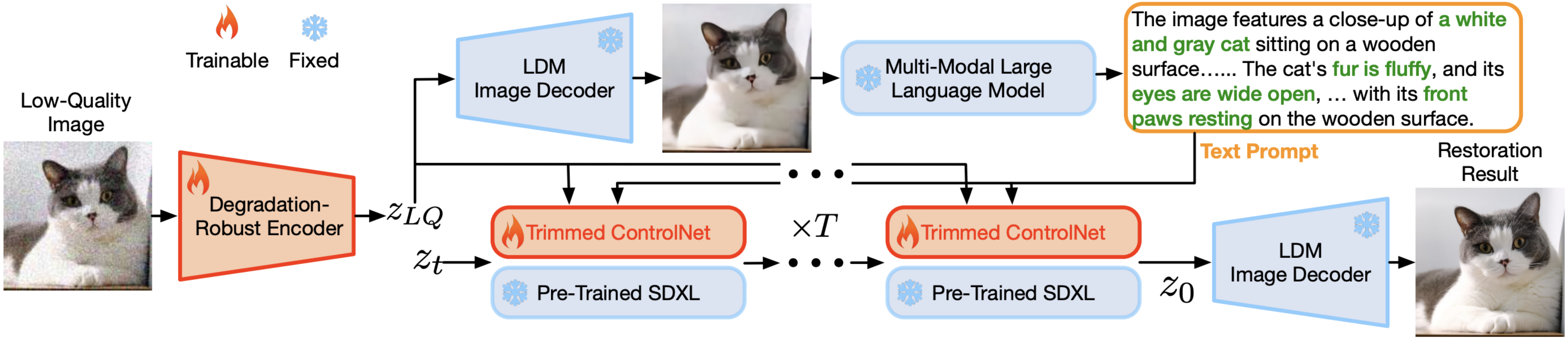
Peiqing Yang. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. NeurIPS 2023

PGDiff | Combine Multiple Guidances



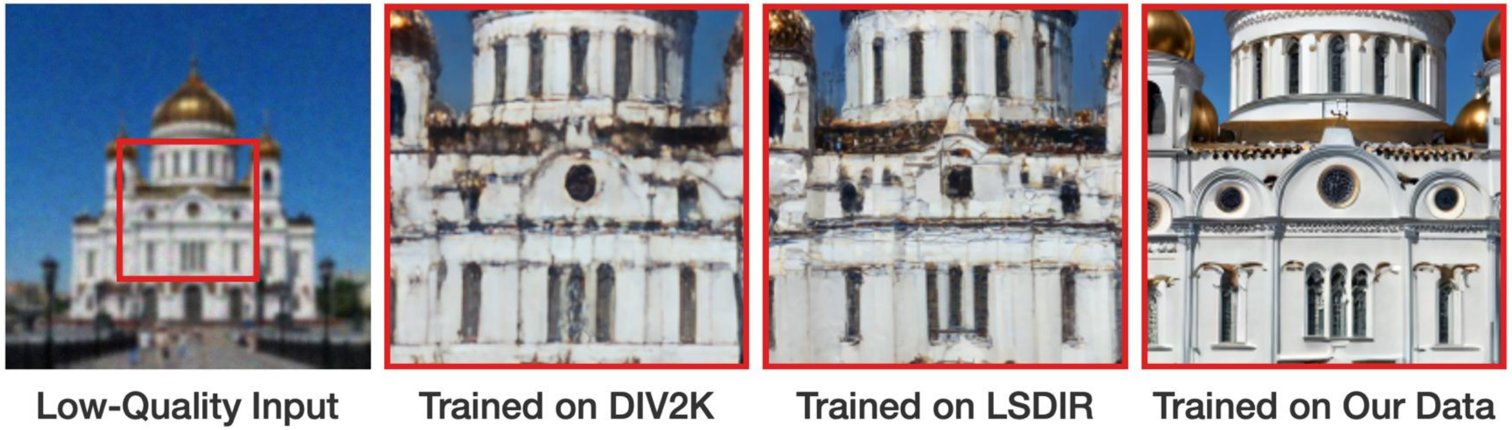
Peiqing Yang. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. NeurIPS 2023

Scaling Up Image Restoration

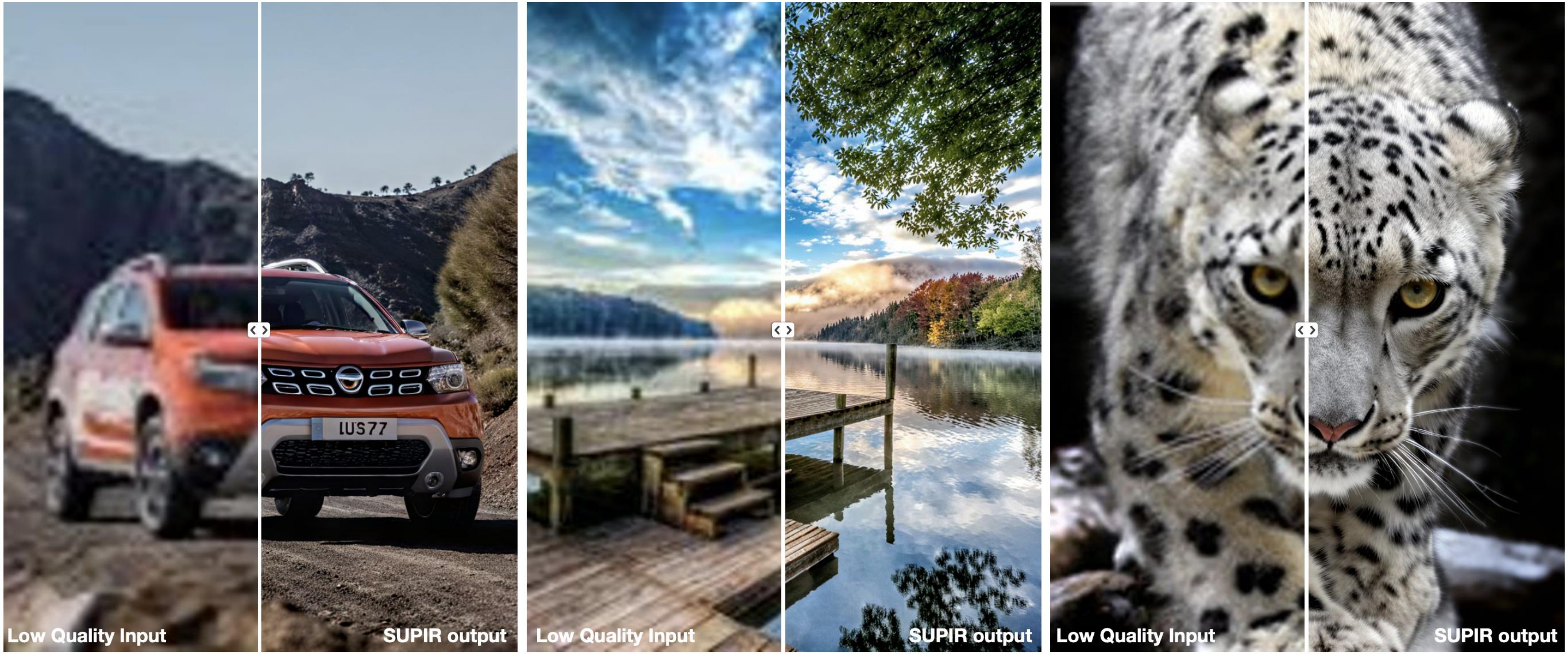


Scaling Up

- Model: SDXL
- Data: The authors collected a large-scale dataset of high-resolution images, which includes 20 million 1024×1024 high-quality, texture-rich images



Scaling Up Image Restoration



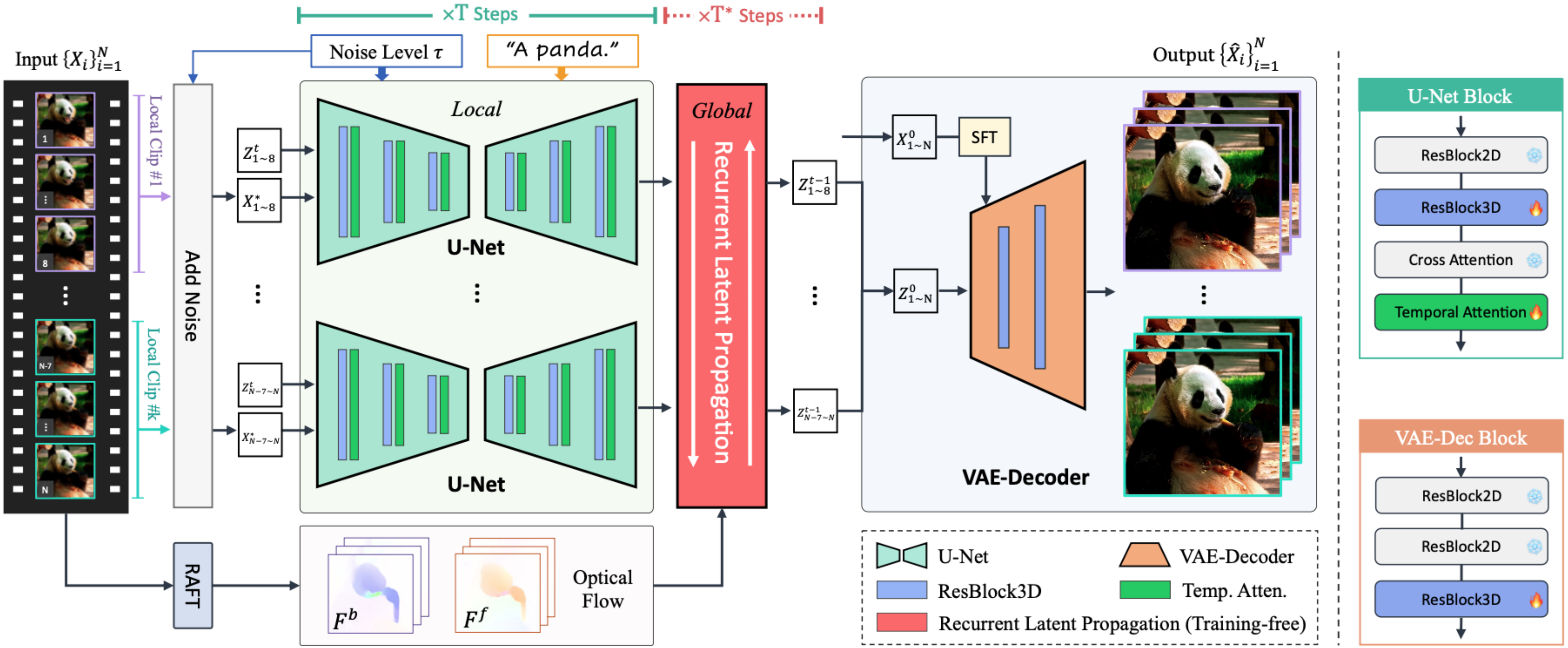
Fanghua Yu et al., Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild, CVPR 2024

Upscale-A-Video | Motivation

Temporal Consistency of Video Diffusion Models for VSR

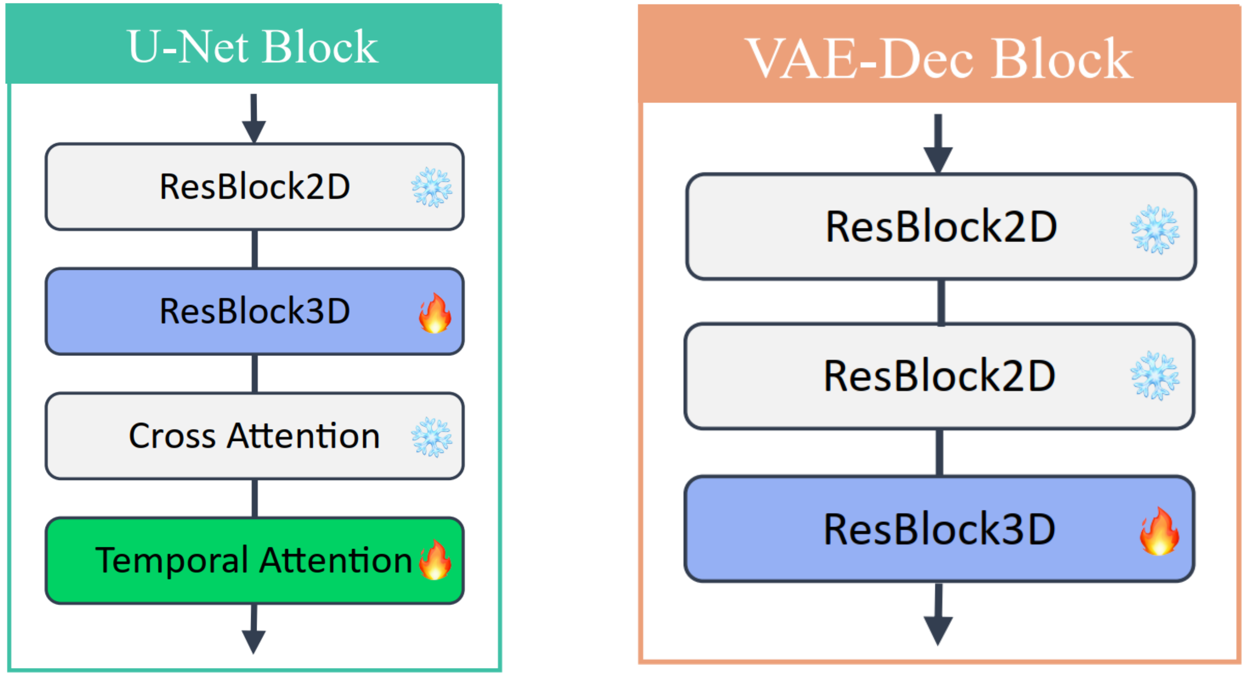
- Local low-level consistency
- Global temporal consistency in longer videos

Upscale-A-Video | Framework



Upscale-A-Video | Local Consistency within Video Segments

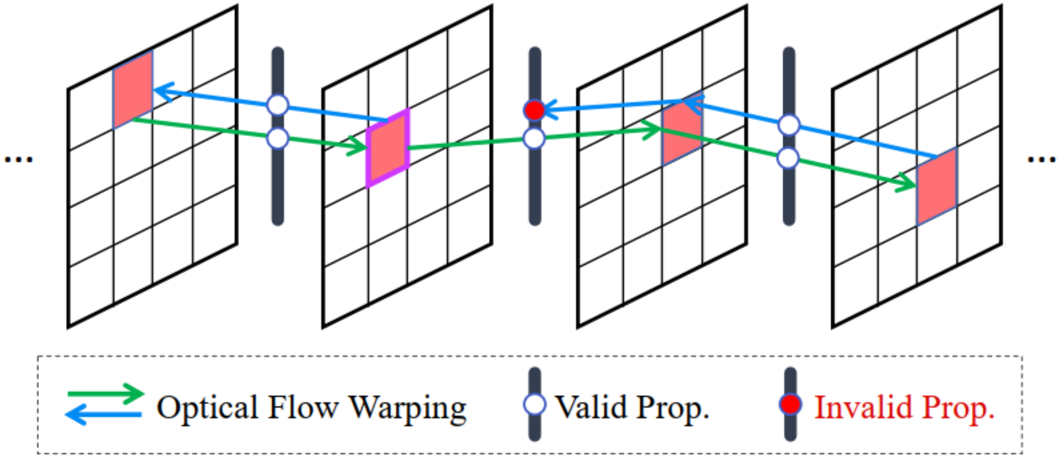
Additional temporal layers that are integrated with the existing spatial layers.



Finetuning U-Net and VAE-Decoder, while keeping the pretrained spatial layers unchanged.

Upscale-A-Video | Global Consistency cross Video Segments

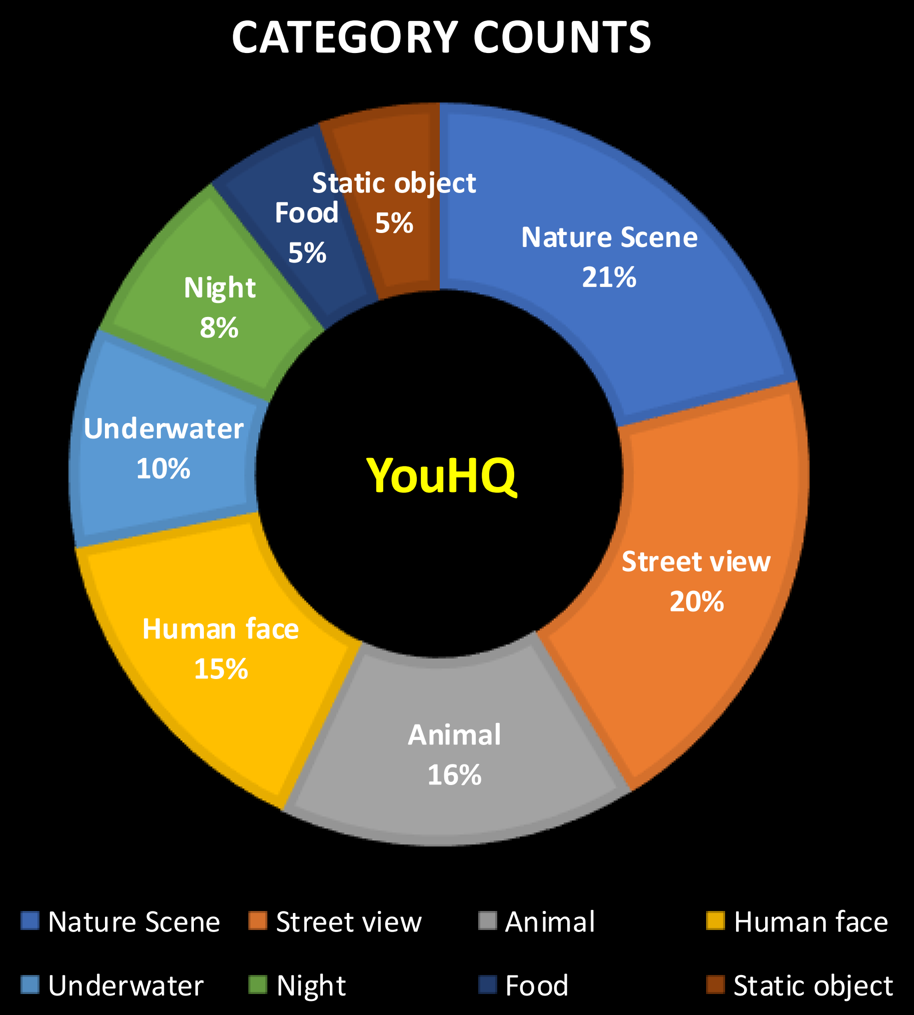
A training-free flow-guided recurrent propagation module within the latent space.



$$E_{i-1\to i}(p) = \left\| f_{i-1\to i}(p) + f_{i\to i-1}(p + f_{i-1\to i}(p)) \right\|_2^2, \quad (1)$$

$$\begin{aligned} \widetilde{\hat{z}_0^i} = {} & \left[\mathcal{W}(\widetilde{\hat{z}_0^{i-1}}, f_{i\to i-1}) * \beta + \hat{z}_0^i * (1-\beta)\right] * M \\ & + \hat{z}_0^i * (1 - M), \end{aligned} \quad (2)$$

YouHQ Dataset



Features:

- Video Number: ~37,000
- Video Resolution: 1080 × 1920
- Video Length: 1s clips (i.e., 32 frames with 30fps)
- Scenes: diverse (human face, animal, street view, landscape, static object, outdoor, indoor, nighttime …)

Upscale-A-Video | Quantitative Evaluation

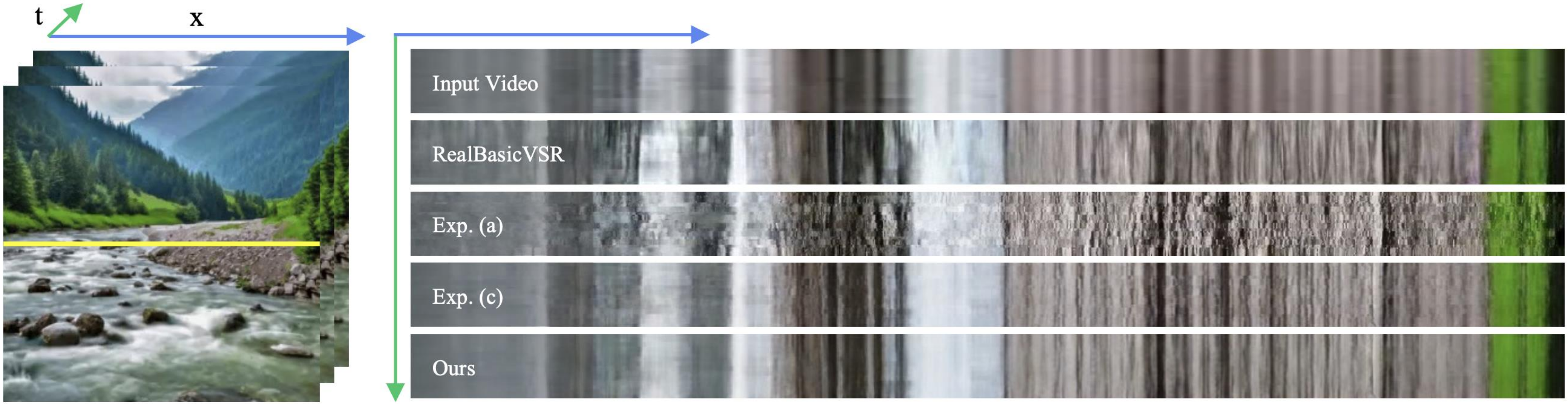
Datasets	Metrics	Real-ESRGAN [66]	SD ×4 Upscaler [2]	ResShift [84]	StableSR [63]	RealVSR [81]	DBVSR [48]	RealBasicVSR [10]	Ours
SPMCS	PSNR ↑	22.89	23.19	23.27	22.71	23.88	24.28	24.51	25.32
	SSIM ↑	0.669	0.631	0.667	0.657	0.681	0.726	0.717	0.741
	LPIPS ↓	0.238	0.304	0.257	0.231	0.437	0.302	0.198	0.222
	E^*_{warp} ↓	1.364	5.008	4.942	4.815	0.294	1.360	0.559	0.367
UDM10	PSNR ↑	27.13	28.07	27.62	26.45	27.38	29.60	29.11	30.79
	SSIM ↑	0.843	0.811	0.827	0.825	0.825	0.880	0.876	0.878
	LPIPS ↓	0.190	0.186	0.222	0.181	0.278	0.155	0.172	0.133
	E^*_{warp} ↓	1.462	1.710	2.196	2.797	0.531	1.943	0.602	0.446
REDS30	PSNR ↑	22.40	22.98	23.00	23.72	23.05	24.37	23.91	24.41
	SSIM ↑	0.591	0.572	0.580	0.635	0.603	0.633	0.636	0.631
	LPIPS ↓	0.303	0.399	0.369	0.352	0.658	0.588	0.249	0.335
	E^*_{warp} ↓	3.658	3.753	4.131	1.645	0.378	9.659	1.557	1.278
YouHQ40	PSNR ↑	24.37	19.71	23.77	24.53	24.19	25.37	24.09	25.83
	SSIM ↑	0.710	0.579	0.654	0.711	0.695	0.719	0.689	0.733
	LPIPS ↓	0.272	0.442	0.376	0.271	0.484	0.430	0.306	0.268
	E^*_{warp} ↓	1.856	3.399	4.426	1.529	0.485	1.149	1.052	0.737
VideoLQ	CLIP-IQA ↑	0.360	0.158	0.430	0.344	0.211	0.274	0.387	0.530
	MUSIQ ↑	49.48	26.21	40.95	44.23	24.52	29.15	55.33	57.99
	DOVER ↑	7.161	2.884	4.679	6.783	2.531	3.628	7.562	7.811
AIGC30	CLIP-IQA ↑	0.430	0.329	0.569	0.467	0.276	0.290	0.565	0.674
	MUSIQ ↑	47.09	35.30	43.32	44.93	24.39	27.22	58.87	57.66
	DOVER ↑	9.710	5.646	7.042	9.668	3.285	3.523	10.68	11.67

Red and blue indicate the best and the second best performance

Upscale-A-Video | Qualitative Comparisons on Real Data


AIGC Video
RealBasicVSR
StableSR
Upscale-A-Video (Ours)

Upscale-A-Video | Temporal Profile



Upscale-A-Video | Effectiveness of Text Prompt

Input

w/o Text Prompt

w/ Text Prompt

- More Results on Real-world Videos -

Remove Flickers For Old Movie

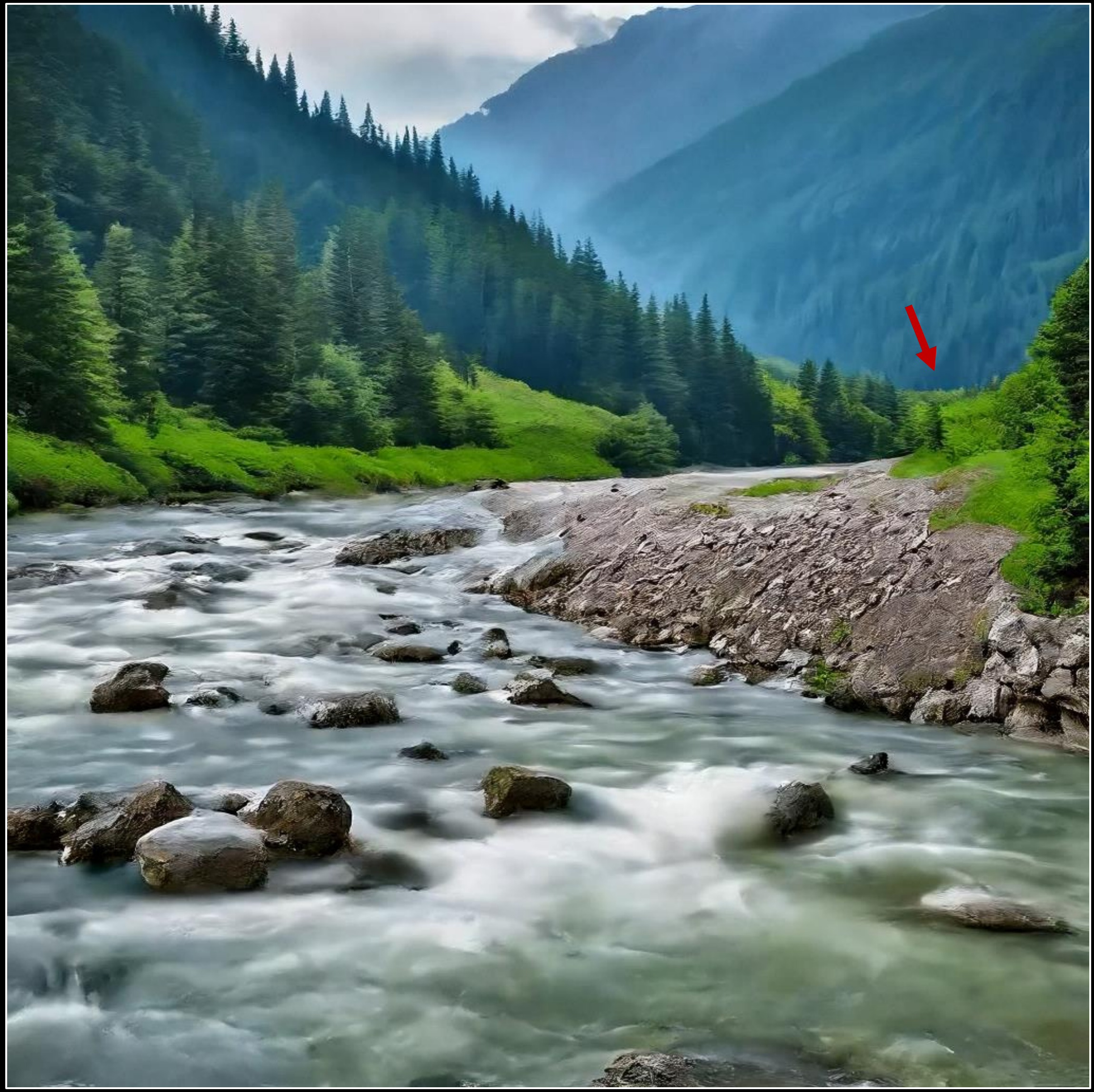
Old Movie Clip

Upscale-A-Video (Ours)


Upscale-A-Video
Real-world Video

- More Results on AIGC Videos -

AIGC Video

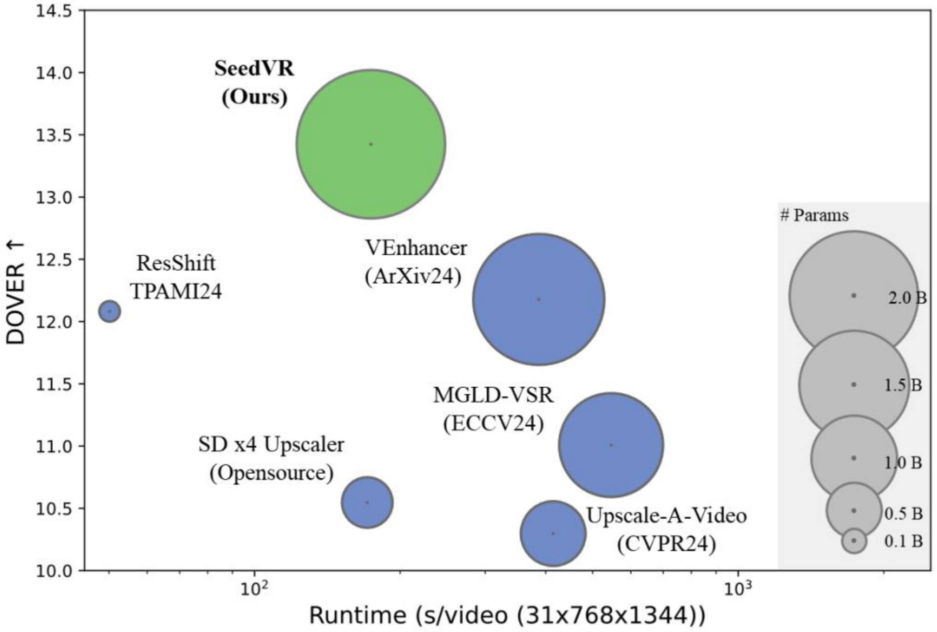


Upscale-A-Video (Ours)

"campfire at night in a snowy forest with starry sky in the background"

"A steam train moving on a mountainside by Vincent van Gogh"

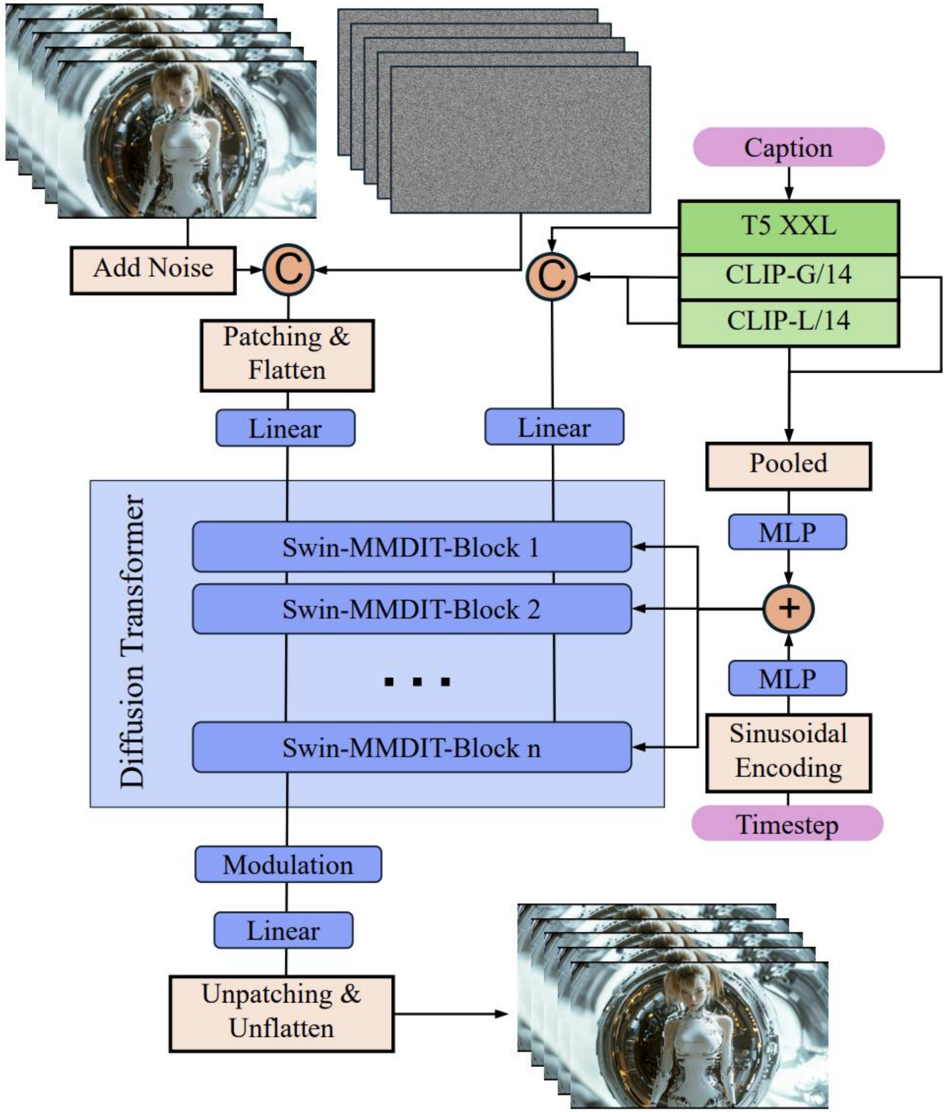
Scaling Up Video Restoration



Despite its 2.48B parameters, SeedVR is over 2× faster than existing diffusion-based video restoration approaches

Jianyi Wang et al., SeedVR: Seeding Infinity in Diffusion Transformer Towards Generic Video Restoration, arXiv 2025

Scaling Up Video Restoration



Extending the MMDiT block from SD3 with shifted window attention like Swin

Use a large non-overlapping window attention - effective for achieving competitive quality at a lower computational cost

Large-scale Training

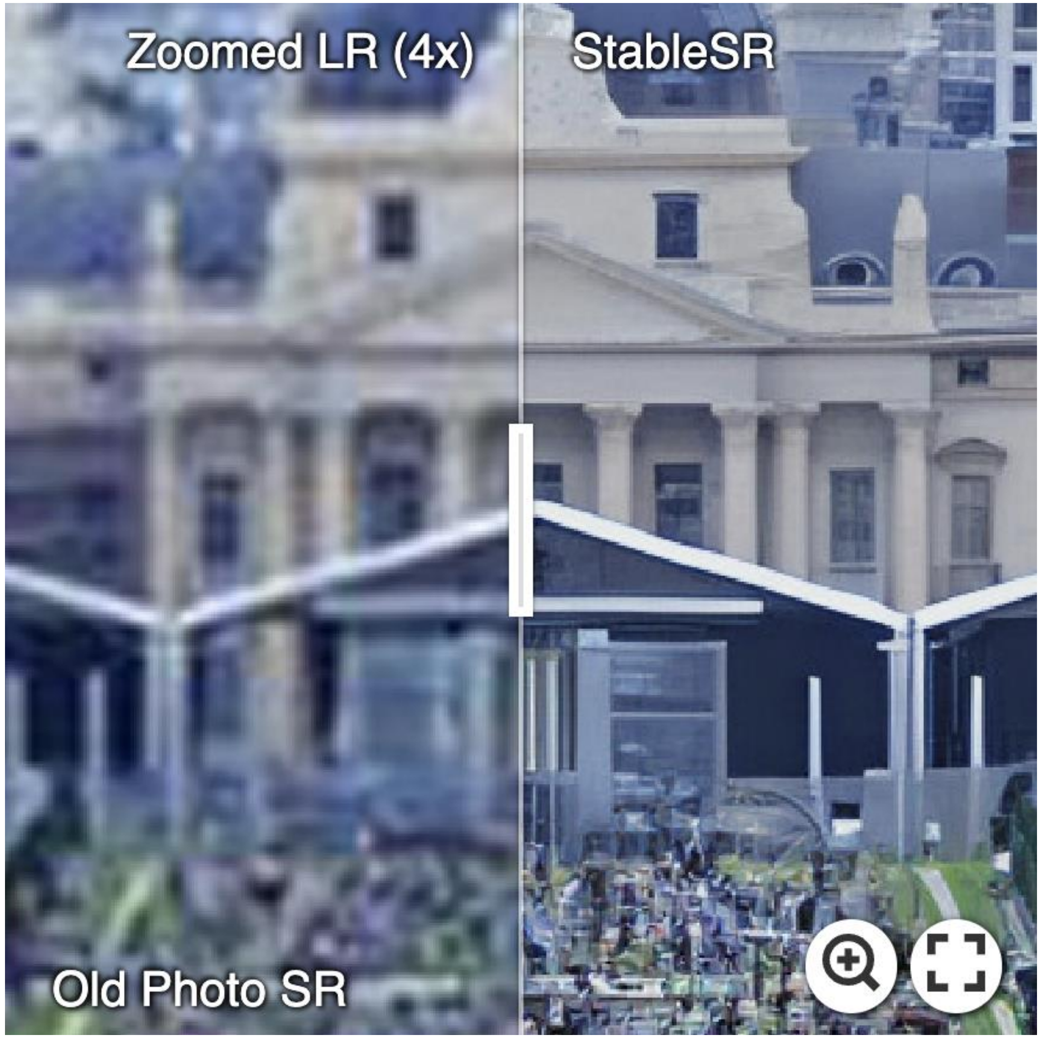
- We trained the model on image and video data simultaneously.
- We collected about 100 million images and 5 million videos

Encoding a 720p video with 21 frames takes approximately 2.9s on average. Precomputing high-quality (HQ) and LQ video latent features along with text embeddings, we can achieve a 4× speed up in training.


Bicubic
MGLD-VSR
VEnhancer
Ours

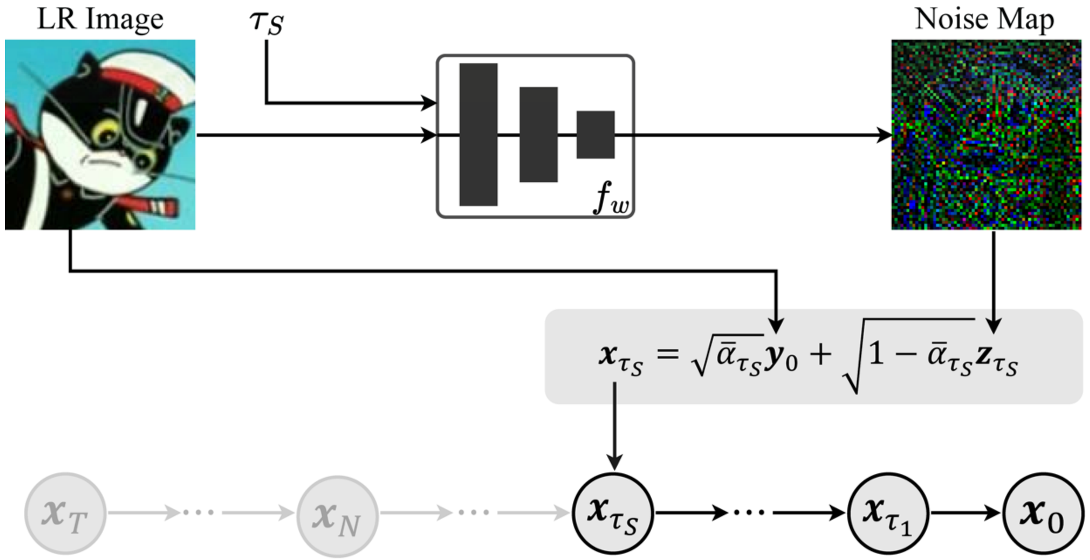
Problems to solve

- Recovering natural scene with the right semantics is hard



Problems to solve

- Diffusion model is still slow
  - *InvSR* mitigates this problem by allowing arbitrary-step restoration through diffusion inversion



Zongsheng Yue et al., Arbitrary-steps Image Super-resolution via Diffusion Inversion, arXiv 2024