Generative Reconstruction Models for Low-Quality Face Images

Chen Change Loy

MMLab@NTU, Nanyang Technological University

https://www.mmlab-ntu.com/

IAPR/IEEE Winter School on Biometrics 2023

Outline

- Introduction
- Priors for face restoration
- CodeFormer

Papers

Image Super-Resolution Using Deep Convolutional Networks

TPAMI 2015

Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang

Deep Cascaded Bi-Network for Face Hallucination

ECCV 2016

Shizhan Zhu, Sifei Liu, Chen Change Loy, Xiaoou Tang

GLEAN: Generative Latent Bank for Image Super-Resolution and Beyond

TPAMI 2022

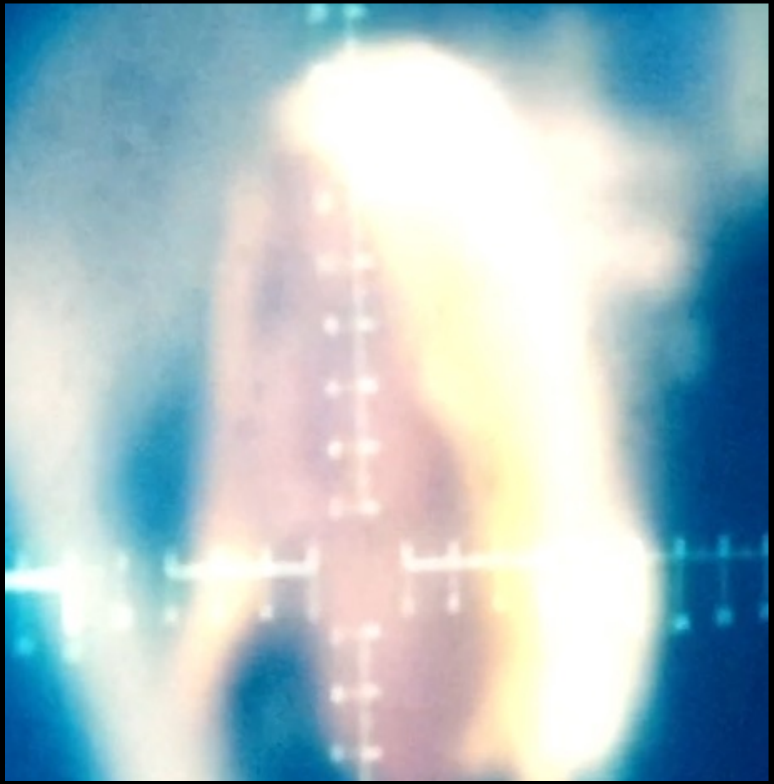
Kelvin C.K Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, Chen Change Loy

Towards Robust Blind Face Restoration with Codebook Lookup Transformer

NeurIPS 2022

Shangchen Zhou, Kelvin C.K Chan, Chongyi Li, Chen Change Loy

Introduction

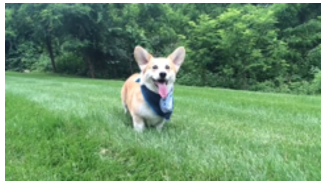


Goal of super-resolution

- Increase the resolution of images
- Produce a detailed, realistic output image.
- Be faithful to the low resolution input image.

320

180



1280

720

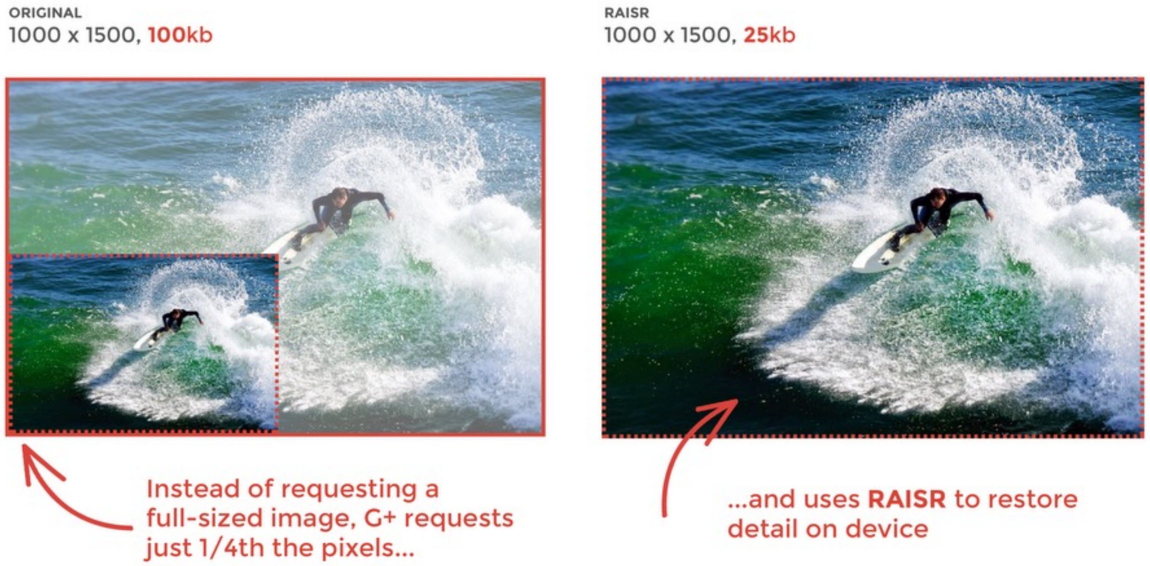
First work on this topic was published in 1984 [1] and the term "Super-resolution" itself appeared at around 1990 [2].

1. R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," in Advances in Computer Vision and Image Processing, vol. 1, chapter 7, pp. 317-339, JAI Press, Greenwich, Conn, USA, 1984.

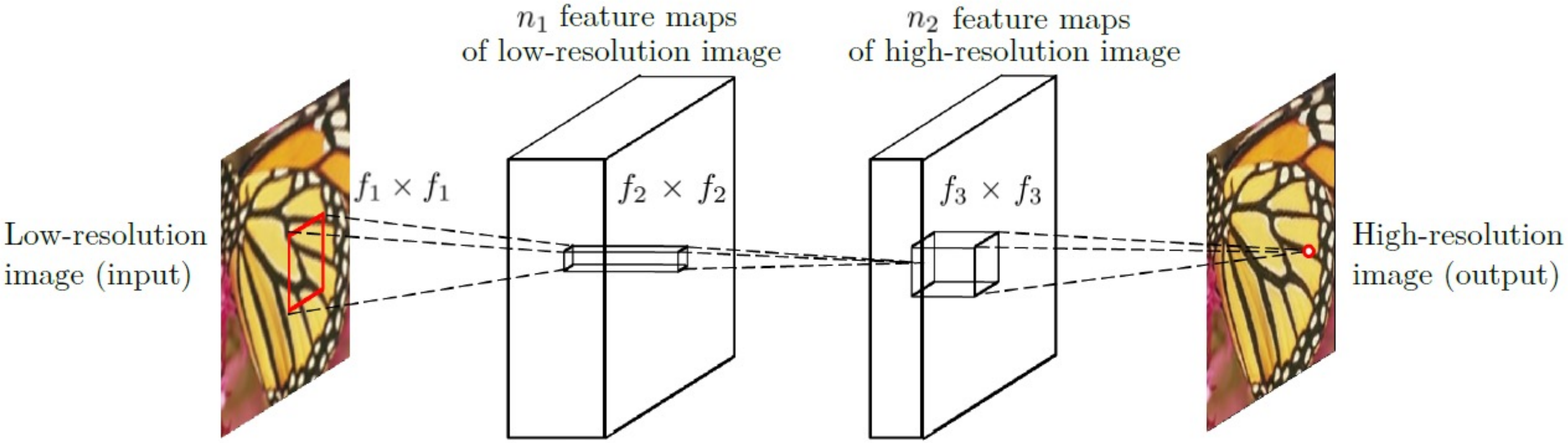
2. M. Irani and S. Peleg. 1991, "Super Resolution From Image Sequences" ICPR, 2:115--120, June 1990.

Applications

- Medical Imaging
- Satellite imaging
- CCTV surveillance (car plate or face)
- Airborne surveillance
- Saving bandwidth



SRCNN



C. Dong, C. C. Loy, K. He, X. Tang, Image Super-Resolution Using Deep Convolutional Networks, TPAMI 2015

Problem objective

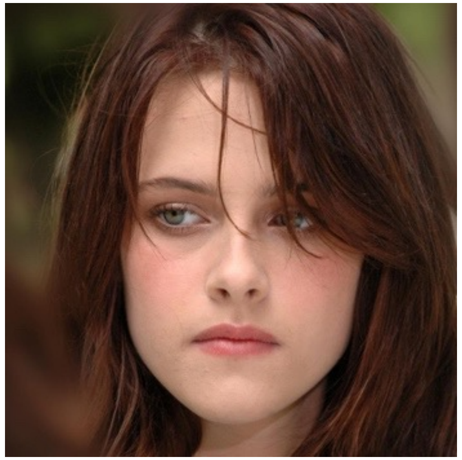
Recover the latent high-quality (HQ) faces $\mathbf{x}$ from its degraded low-quality (LQ) faces

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

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

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}_{\text{fidelity term}} + \underbrace{\lambda\Phi(\mathbf{x})}_{\text{regularization term}}$$

LQ



HQ

Problem objective

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

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

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

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}_{\text{fidelity term}} + \underbrace{\lambda\Phi(\mathbf{x})}_{\text{regularization term}}$$

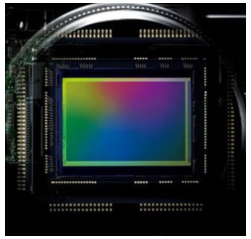
If we know the $\mathbf{H}$ and $\mathbf{v}$, then is a non-blind super-resolution. Otherwise it is a blind super-resolution

Degradation involved in real applications are typically complicated (downsampling, blur, noise, and JPEG compression) and unavailable.

Degradation in the real world

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

Take Photo



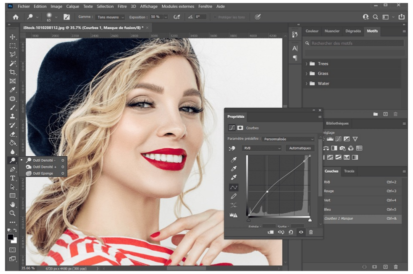
Camera Blur
(motion and defocus)

Sensor Noise

JPEG compression

Editing

Downsampling,
Artifacts



Social Media Sharing

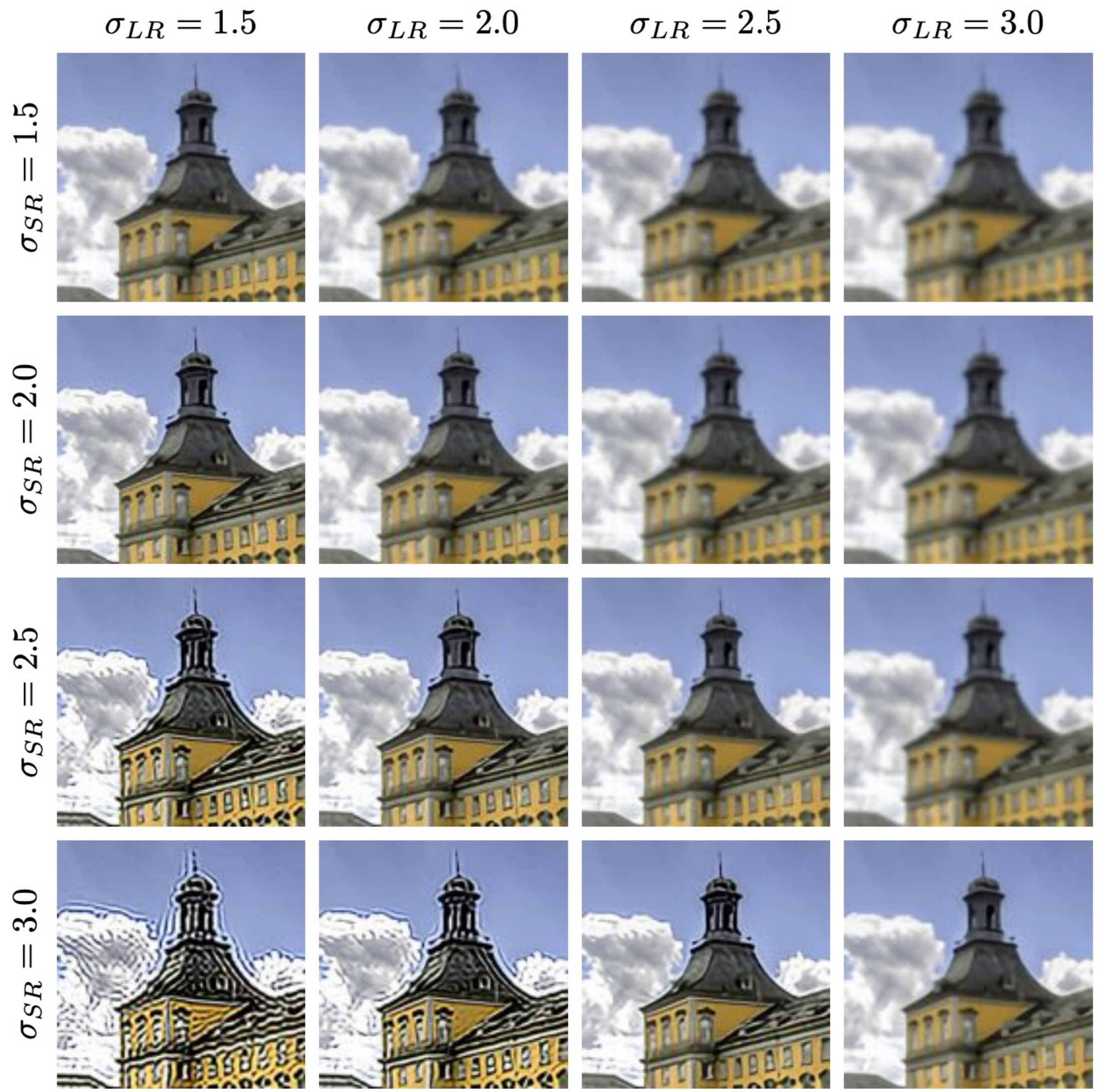
Compression,
Downsampling, Noises

…

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

Challenges

- Facial details are lost and degraded in the LQ images

LQ

Challenges

- Identity inconsistency between output and GT

Input LQ Possible Outputs HQ GT

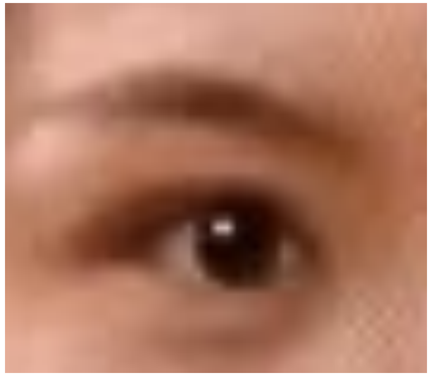
A good solution

i. Reduce the uncertainty and ambiguity of LQ-to-HQ mapping.

ii. Complement high-quality details lost in the LQ inputs.

iii. Be robust against heavy degradations while maintaining identity consistency.

How to achieve this?



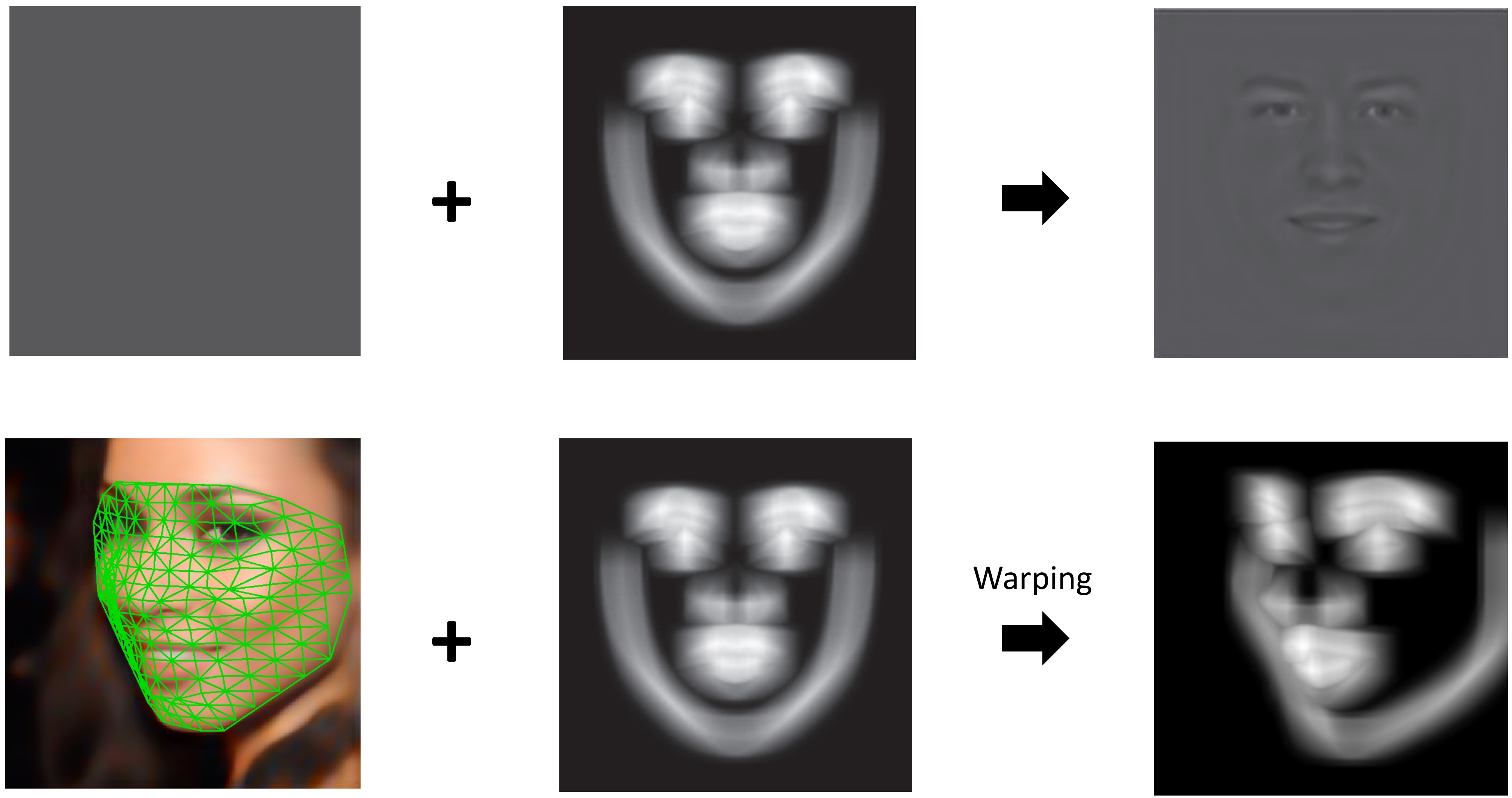
S. Zhou, K. C. K. Chan, C. Li, C. C. Loy, Towards Robust Blind Face Restoration with Codebook Lookup TransFormer, NeurIPS 2022

Priors for Face Restoration

Existing priors for face restoration

- 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

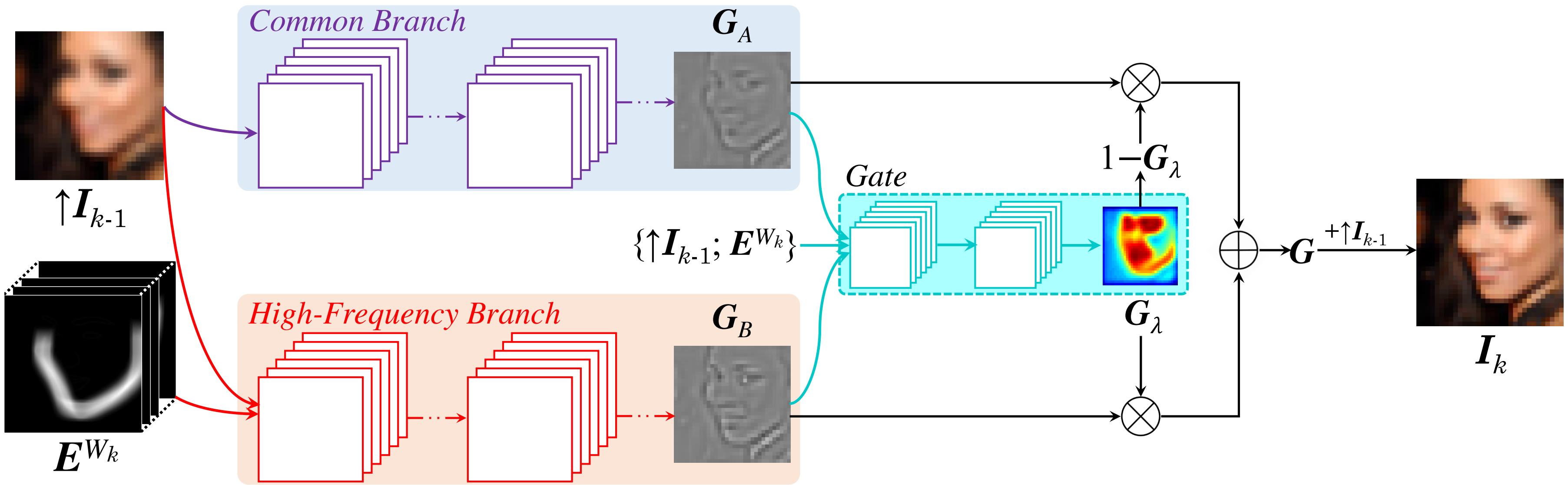


Dense correspondence field

Face prior

Geometric prior

Face restoration conditioned on prior



(a) Bicubic (b) Common (c) High-Freq. (d) **CBN** (e) Original

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

Geometric prior

Bicubic

[a] Sparse coding prior

[b] Patch-wise mixture of probabilistic PCA prior

Ours - Pixel-wise dense face correspondence field

Bicubic

[a] Sparse coding prior

[b] Patch-wise mixture of probabilistic PCA prior

Ours - Pixel-wise dense face correspondence field

[a] Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.: Deep networks for image super-resolution with sparse prior, ICCV 2015
[b] Jin, Y., Bouganis, C.S.: Robust multi-image based blind face hallucination. CVPR, 2015

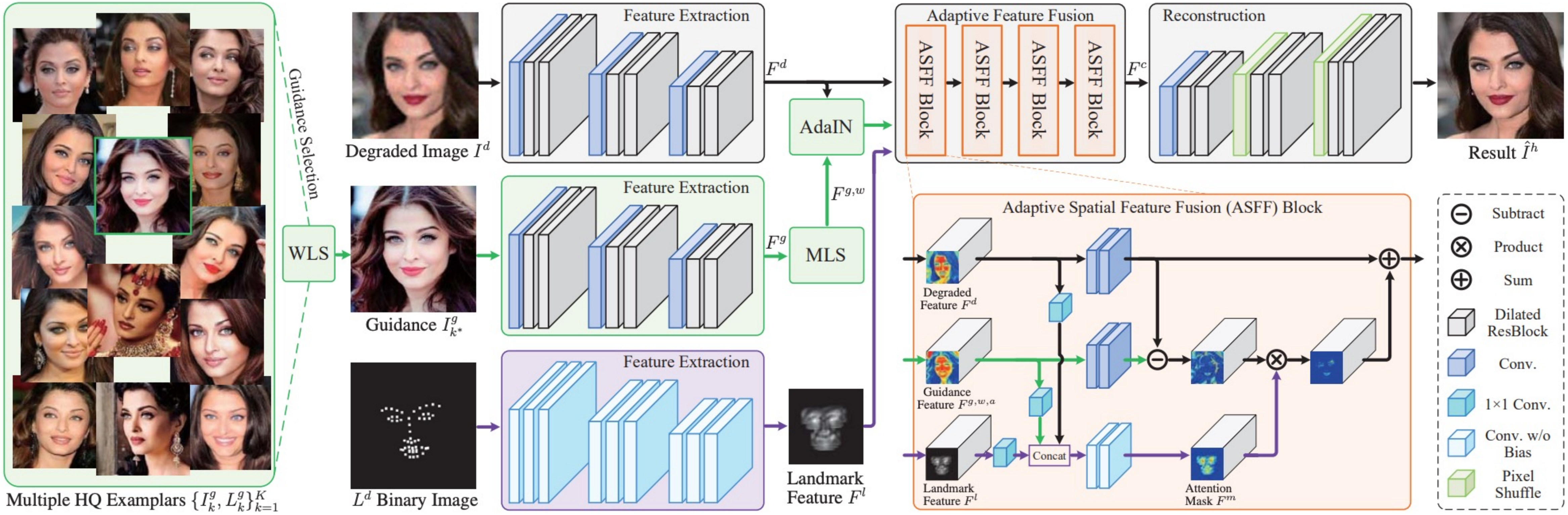
Existing priors for face restoration

- 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
  - …

Reference prior

Face restoration conditioned on exemplars

Reference prior



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

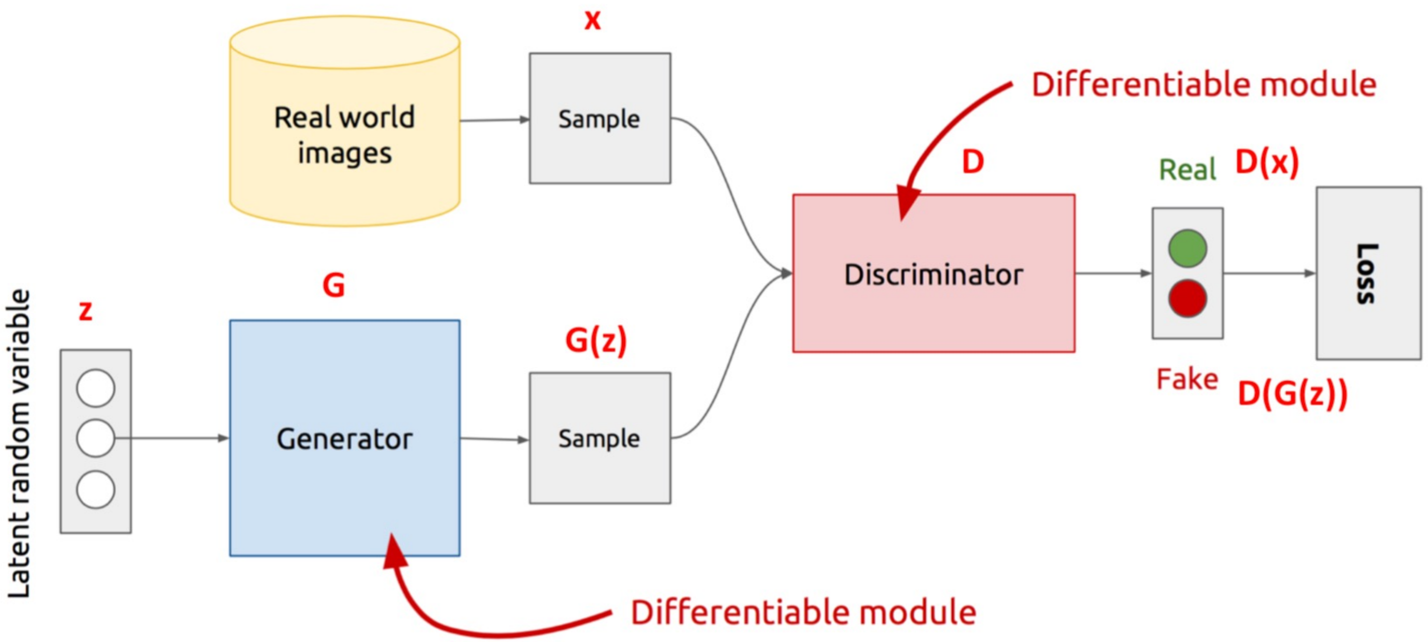
Existing priors for face restoration

- 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
  - …

Generative prior

Generative Adversarial Network

- Generative model G:
  - Captures data distribution
  - Fool $D(G(z))$
  - Generate an image $G(z)$ such that $D(G(z))$ is wrong (i.e. $D(G(z)) = 1$)
- Discriminative model D:
  - Distinguishes between real and fake samples
  - $D(x) = 1$ when x is a real image, and otherwise



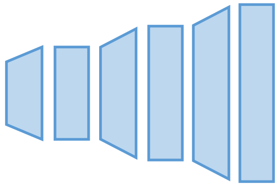
z is some random noise (Gaussian/Uniform).

z can be thought as the latent representation of the data.

Generative prior

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

Latent space



Generator

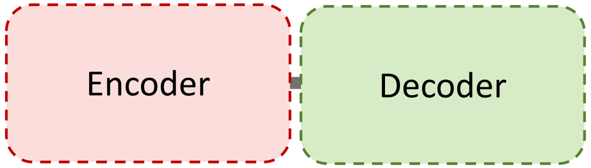
Can we leverage a GAN trained on large-scale natural images for richer priors?

GAN is a good approximator for natural image manifold.

Generative prior

Using GAN as latent bank

Encoder-Decoder Structure



A common architecture

It is typically trained from scratch using a combined objective function consisting of a fidelity term and an adversarial loss

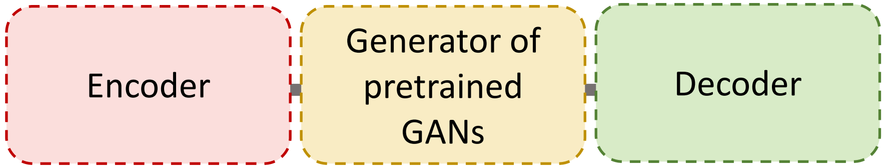
The generator is responsible for both capturing the natural image characteristics and maintaining the fidelity to the ground-truth.

This inevitably limit its capability of approximating the natural image manifold.

Generative prior

Using GAN as latent bank

Encoder-Bank-Decoder Structure



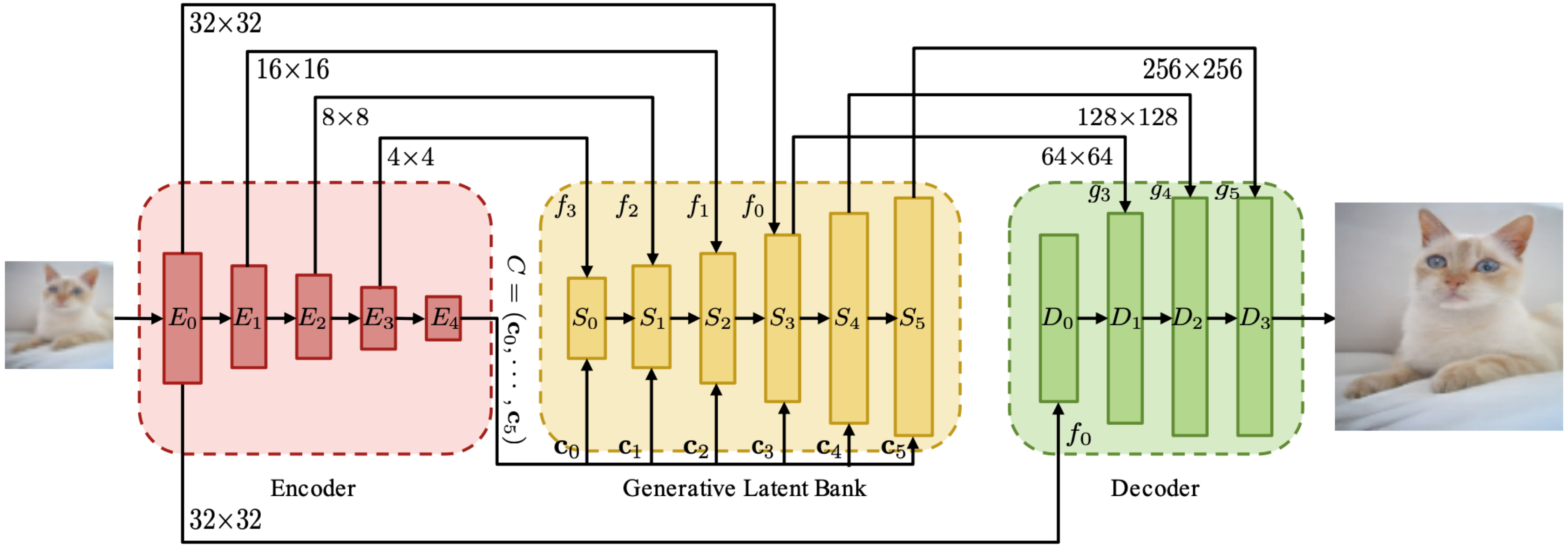
Lifts the burden of learning both fidelity and texture generation simultaneously

Does not involve image-specific optimization at runtime

Needs a single forward pass to perform image restoration

Inspired by the classic notion of dictionary but exploit GAN as a more effective way for storing priors

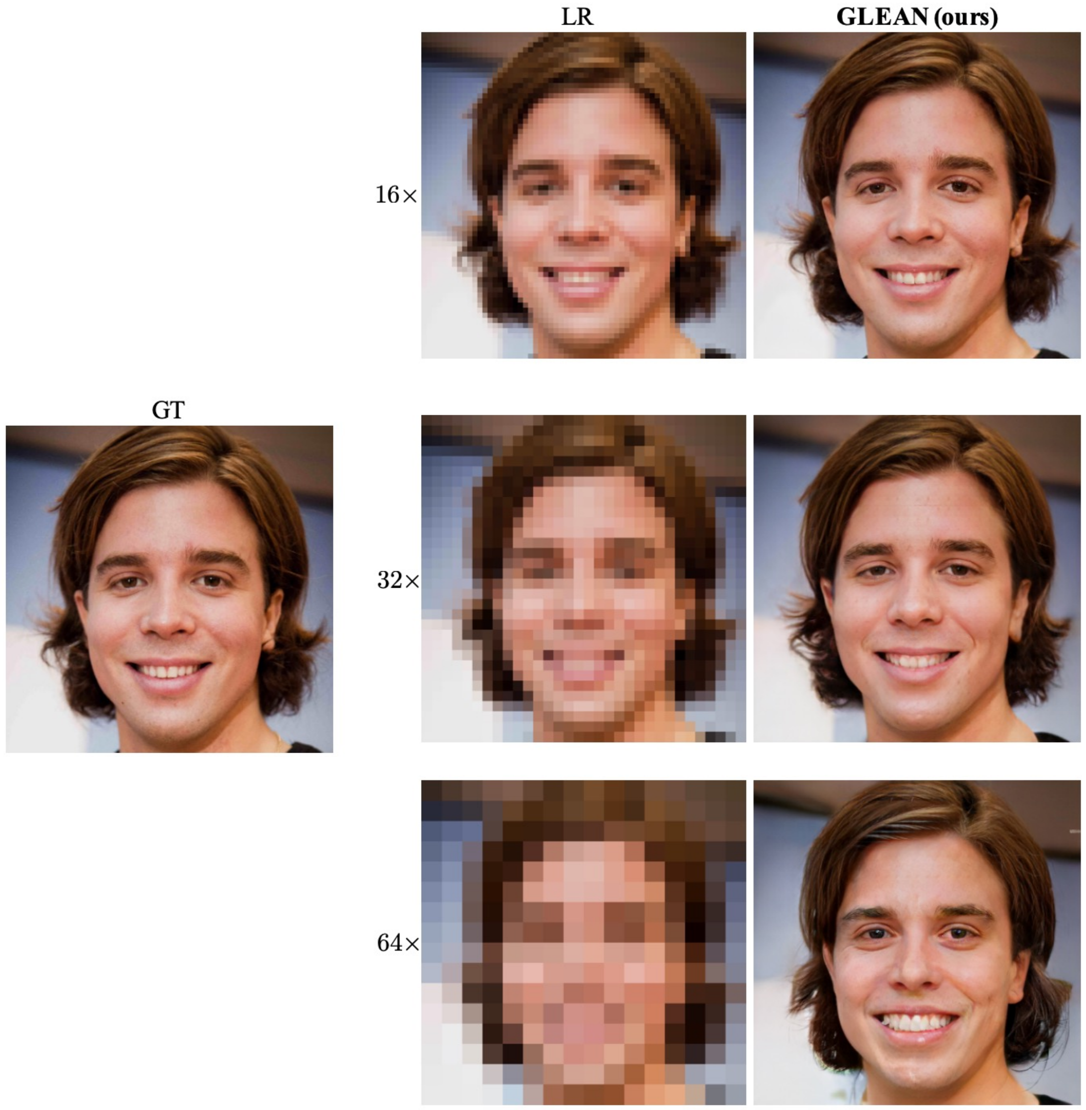
Generative prior

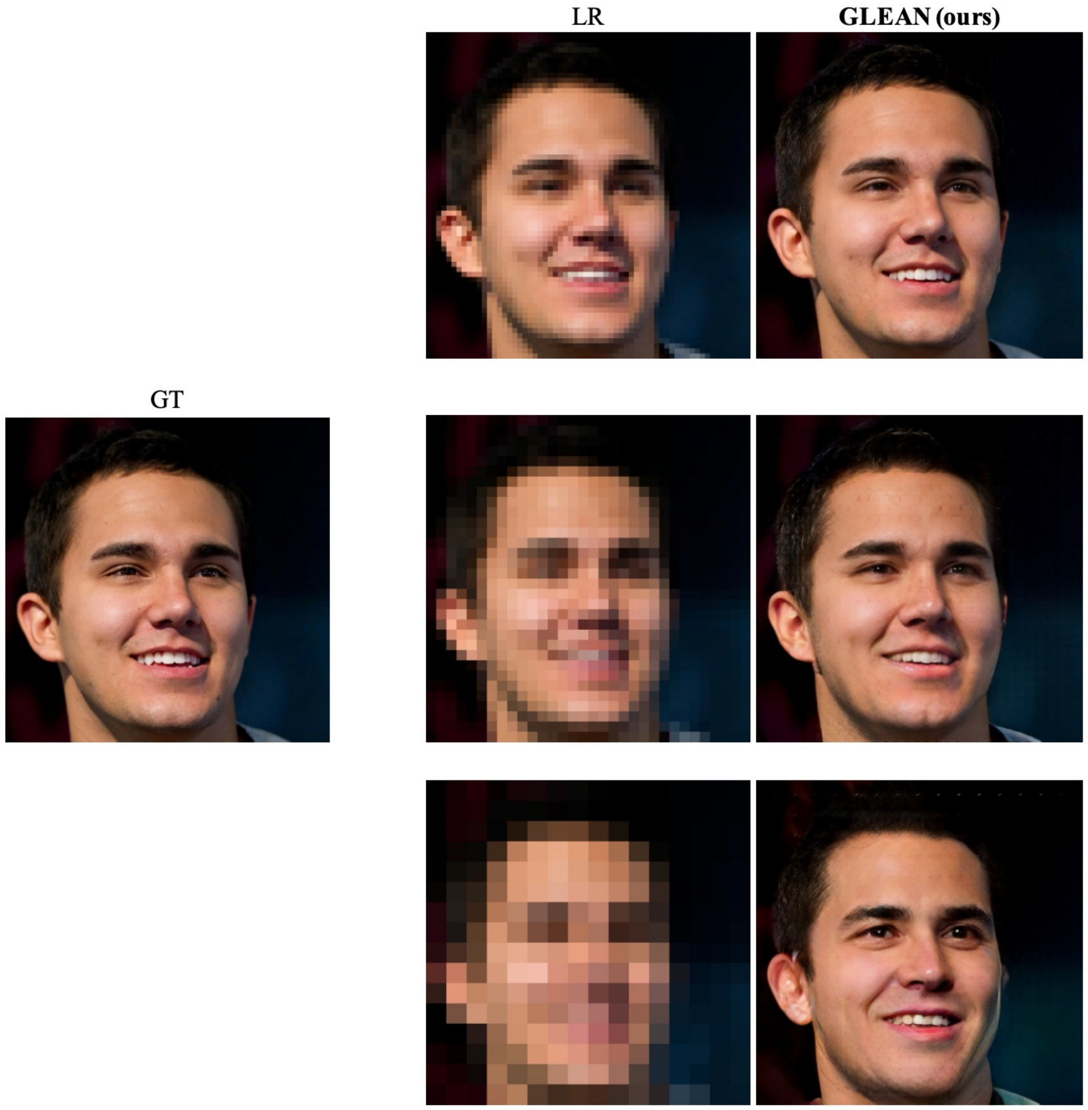


Condition the bank by passing both the latent vectors and multi-resolution convolutional features from the encoder to achieve high-fidelity results. Symmetrically, multi-resolution cues need to be passed from the bank to the decoder.

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





Generative prior

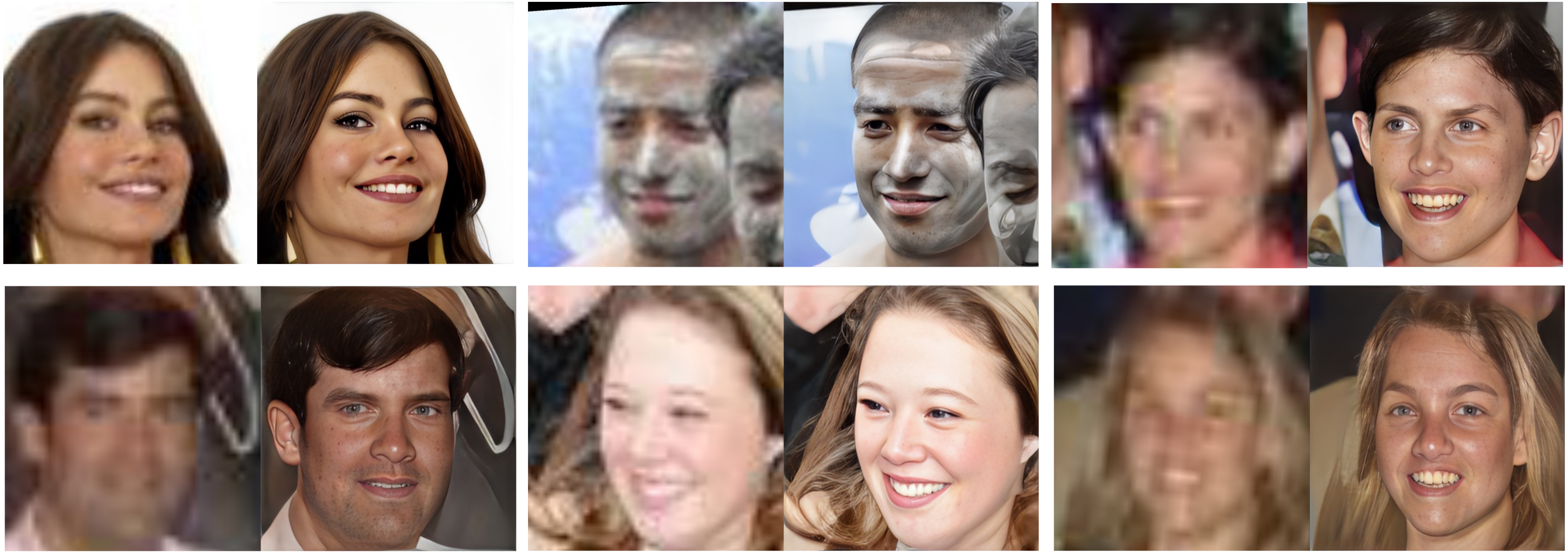
484x484

242x242

121x121

60x60

Generative prior



Generative prior

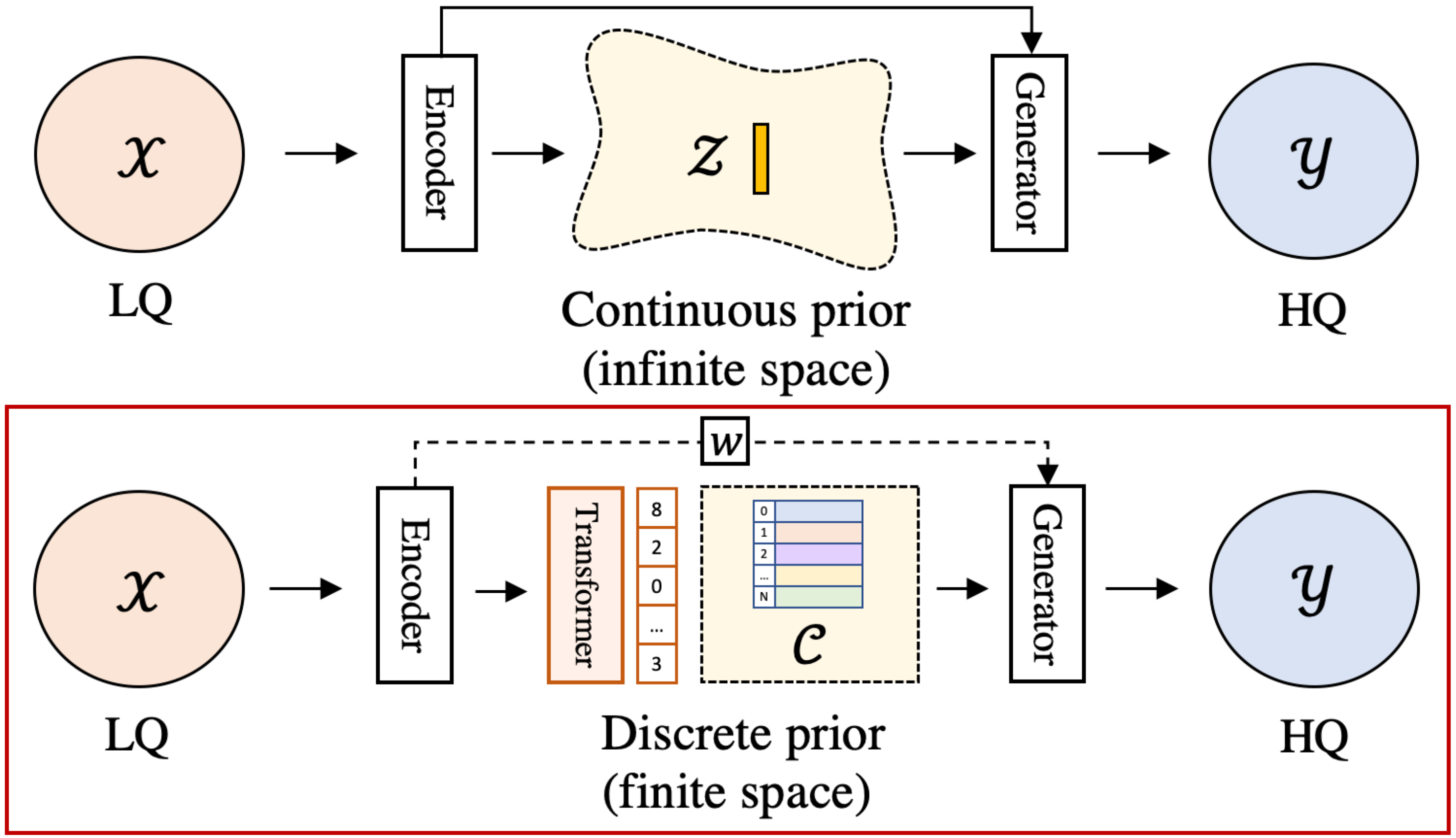
Generative prior

CodeFormer

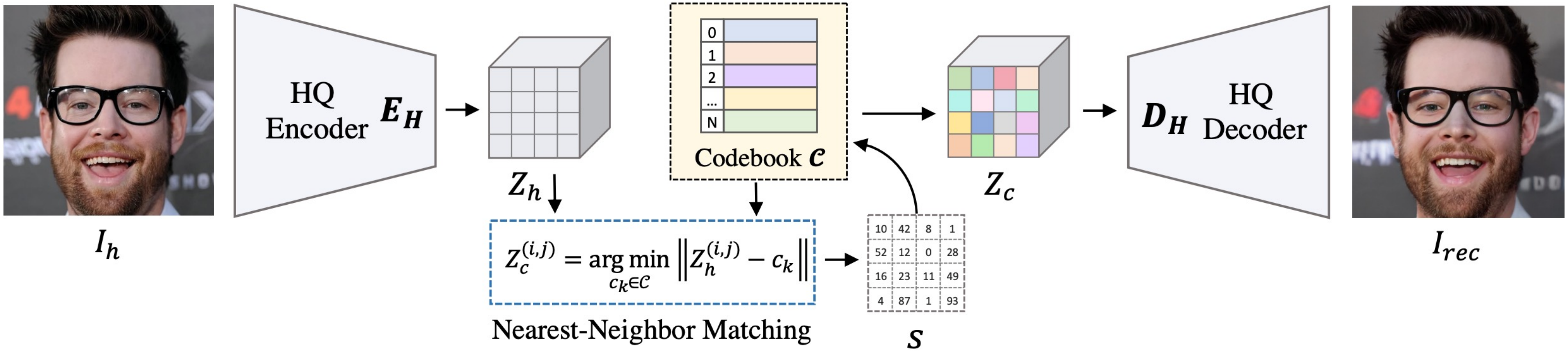
Continuous prior *v.s.* discrete prior

StyleGAN-based frameworks

CodeFormer



VQGAN



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

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

Continuous prior *v.s.* discrete prior

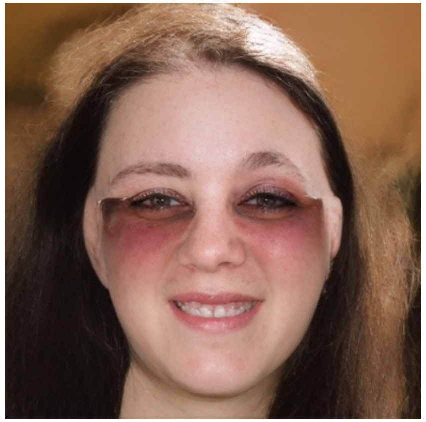
A. LQ-HQ mapping √
B. Details √
C. Identity

Input

PULSE
(continuous, w/o connection)

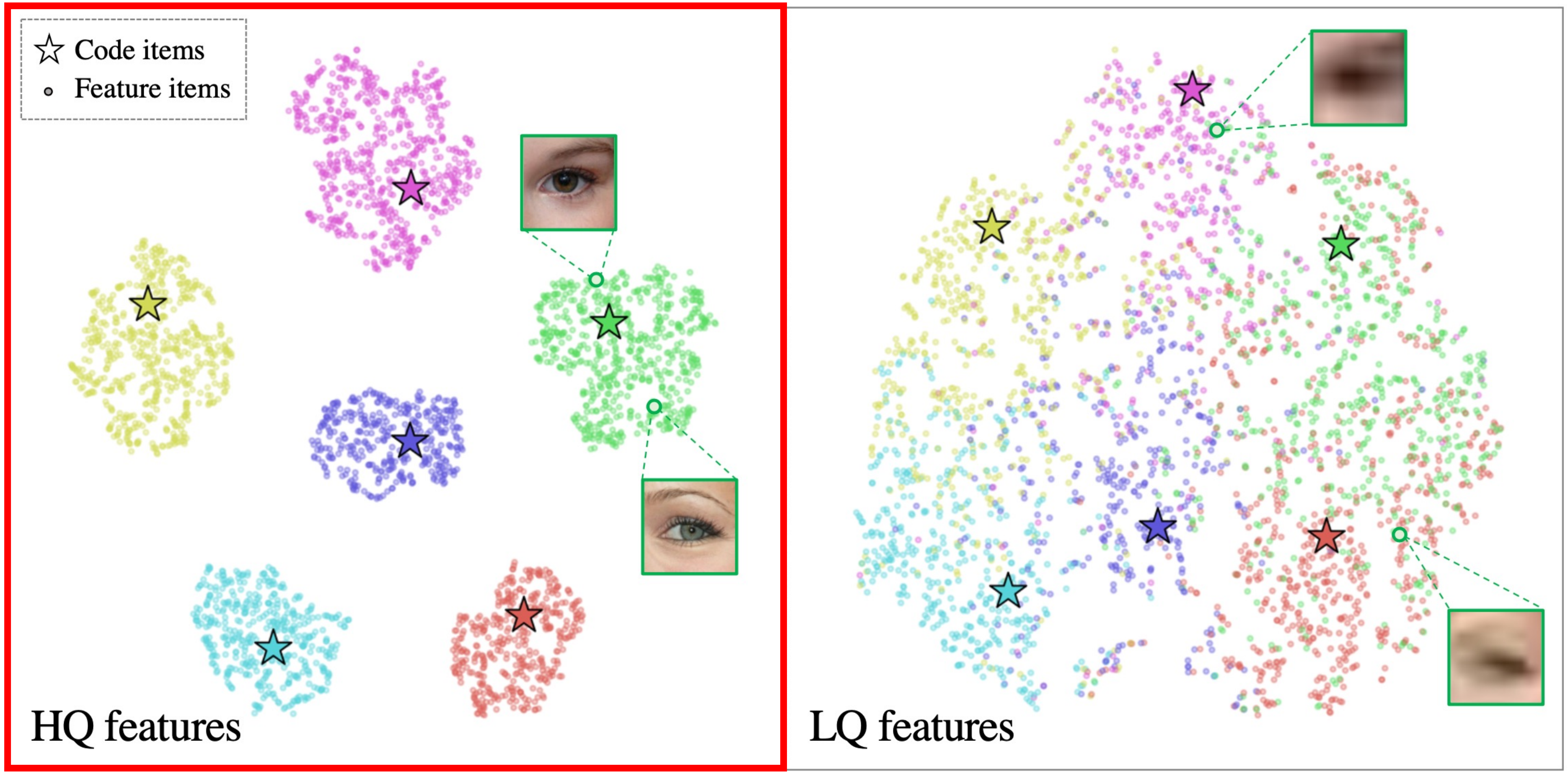
GFP-GAN
(continuous, w/ connection)

Ground Truth



Nearest Neighbor
(discrete, w/o connection)

Codebook lookup



(b) Distributions of HQ (left) / LQ (right) features and the codebook items

Continuous prior *v.s.* discrete prior

A. LQ-HQ mapping

B. Details √

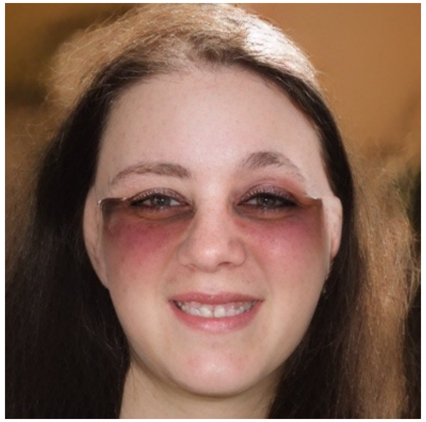
C. Identity √

Input

PULSE
(continuous, w/o connection)

GFP-GAN
(continuous, w/ connection)

Ground Truth



Nearest Neighbor
(discrete, w/o connection)

CodeFormer
(discrete, w/o connection/w=0)

Nearest Neighbor v.s. CodeFormer

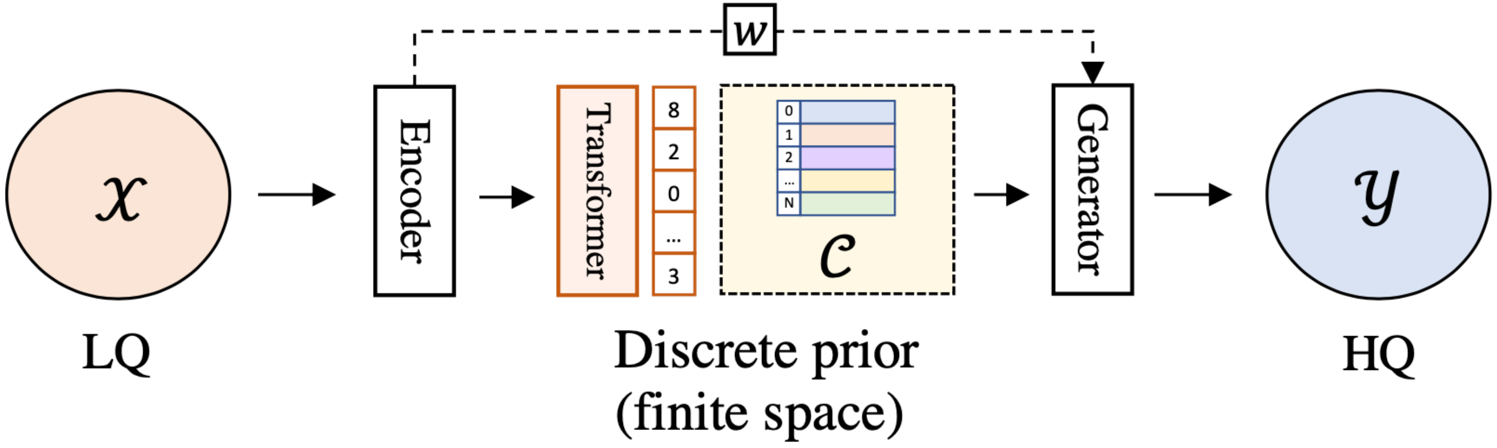
Real Input　Nearest Neighbor　CodeFormer

Controllability

A. LQ-HQ mapping

B. Details

C. Identity √

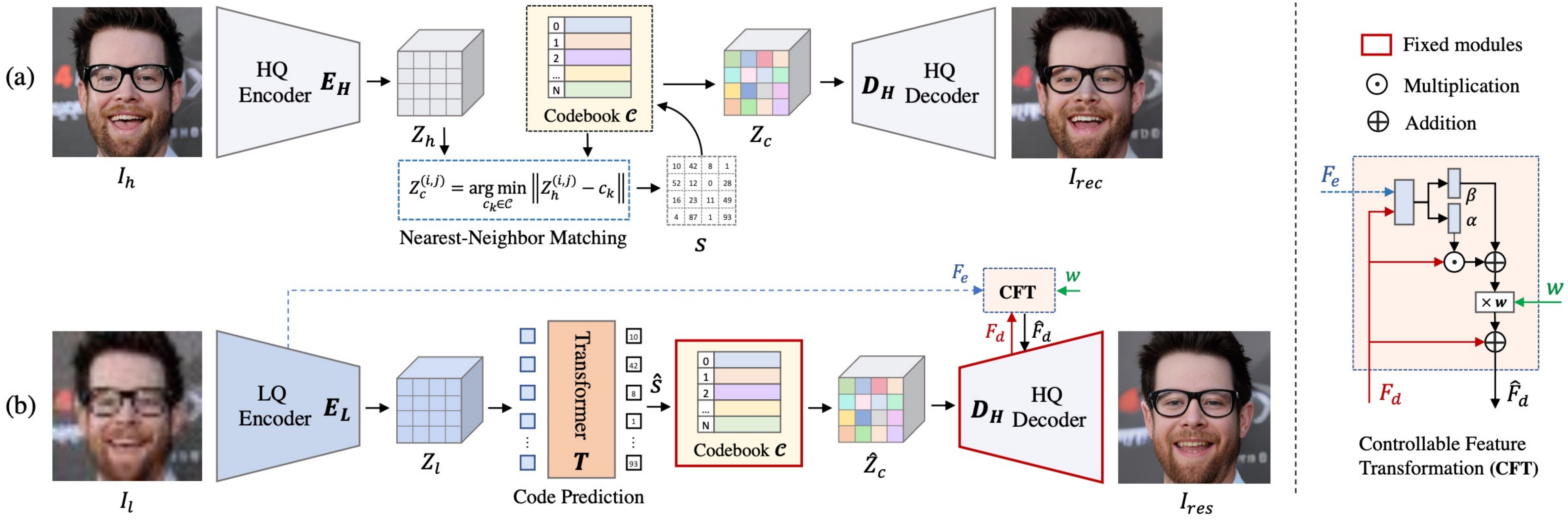


Addressing the challenges

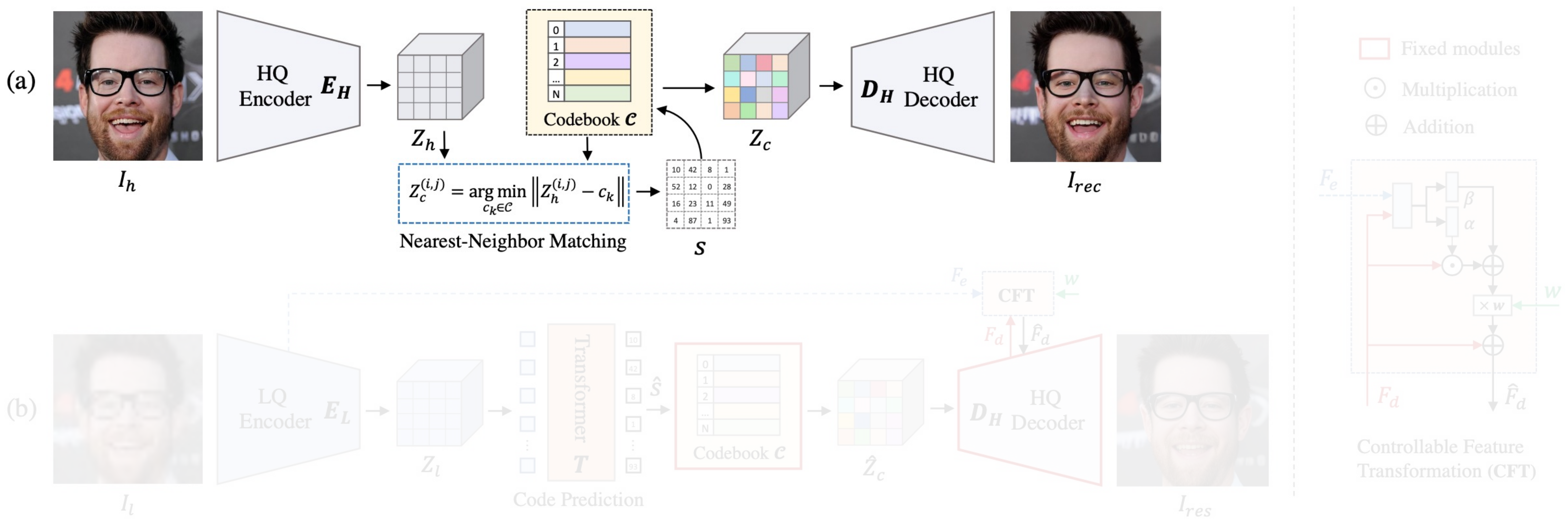
Challenges	**CodeFormer**
A. LQ-HQ mapping	❑ Discrete Codebook Prior
B. Details	❑ Transformer Module
C. Identity	❑ Controllable Module

Framework of CodeFormer

It contains three training stages



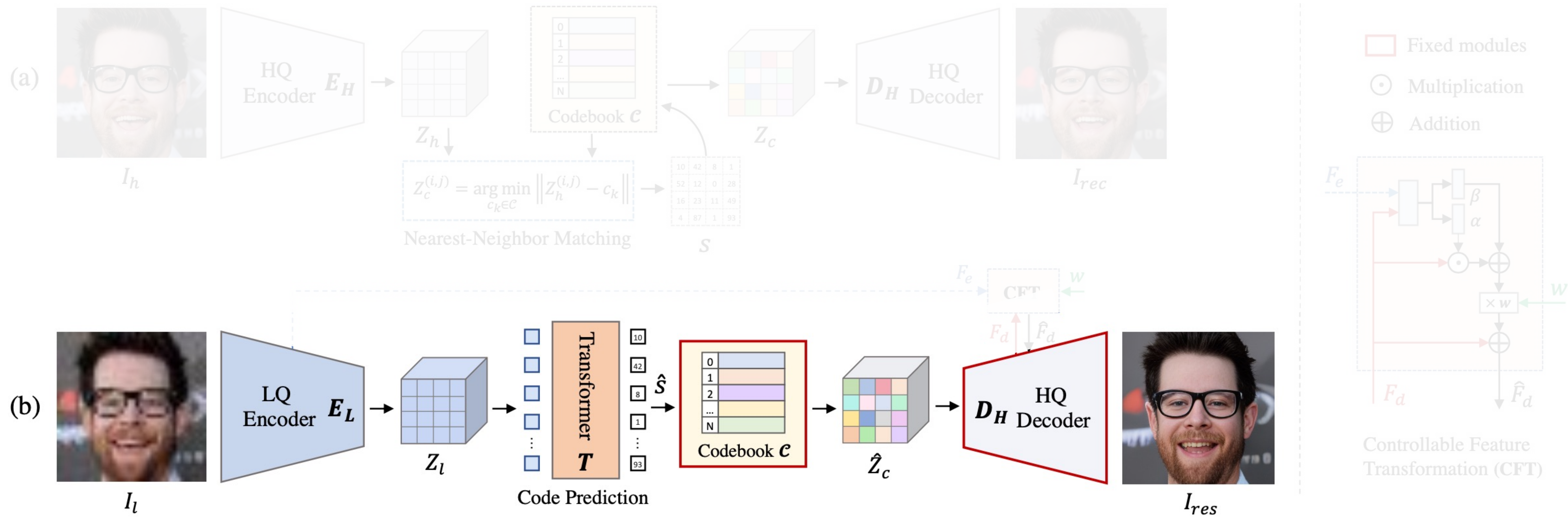
Stage I: Codebook Learning (VQGAN)



$$\mathcal{L}_1 = \|I_h - I_{rec}\|_1; \quad \mathcal{L}_{per} = \|\Phi(I_h) - \Phi(I_{rec})\|_2^2; \quad \mathcal{L}_{adv} = [\log D(I_h) + \log(1 - D(I_{rec}))]$$

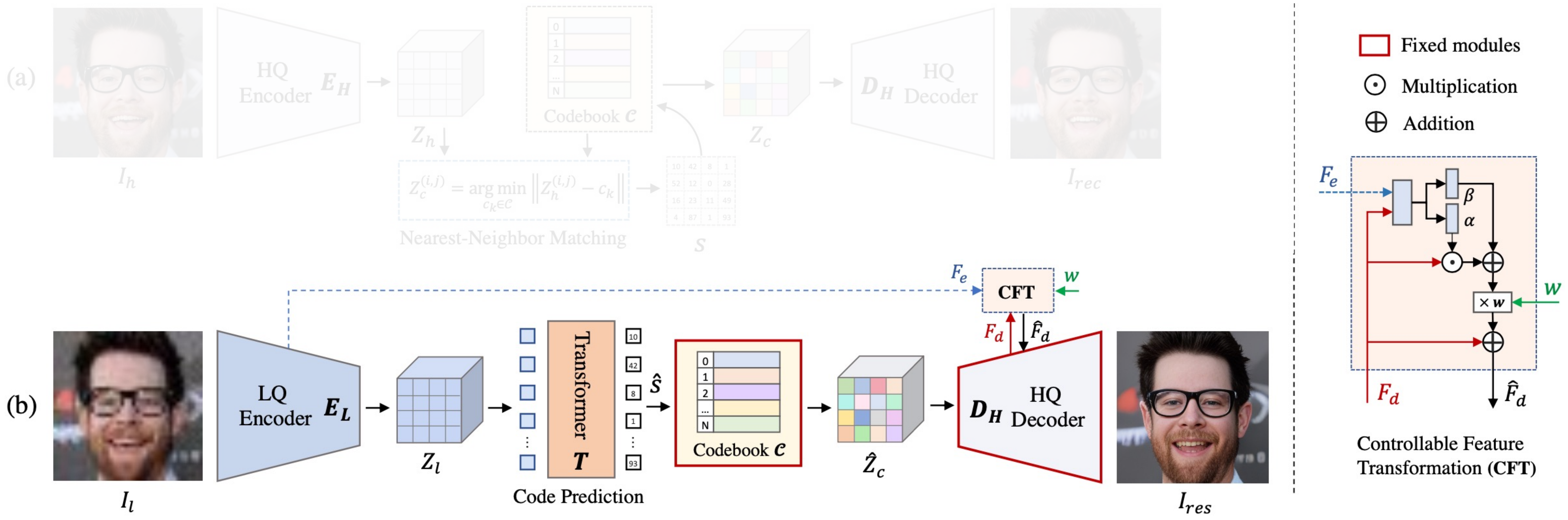
$$\mathcal{L}_{code}^{feat} = \|\mathrm{sg}(Z_h) - Z_c\|_2^2 + \beta\|Z_h - \mathrm{sg}(Z_c)\|_2^2$$

Stage II: Codebook Lookup Transformer



$$\mathcal{L}_{code}^{token} = \sum_{i=0}^{mn-1} -s_i \log(\hat{s}_i); \quad \mathcal{L}_{code}^{feat'} = \| Z_l - \mathrm{sg}(Z_c) \|_2^2$$

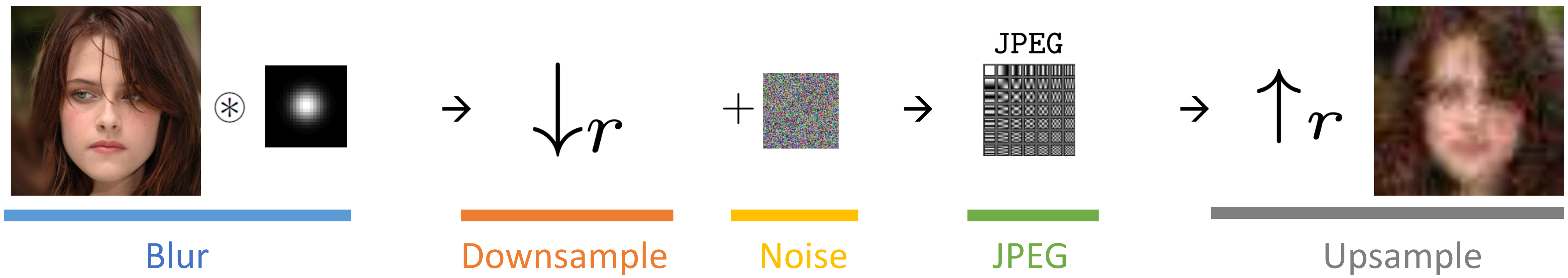
Stage III: Controllable Feature Transformation



$$\hat{F}_d = F_d + (\alpha \odot F_d + \beta) \times w; \quad \alpha, \beta = \mathcal{P}_\theta(c(F_d, F_e))$$

Degradation model

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



Blur
- (Generalize) Gaussian filter
  - isotropic
  - anisotropic

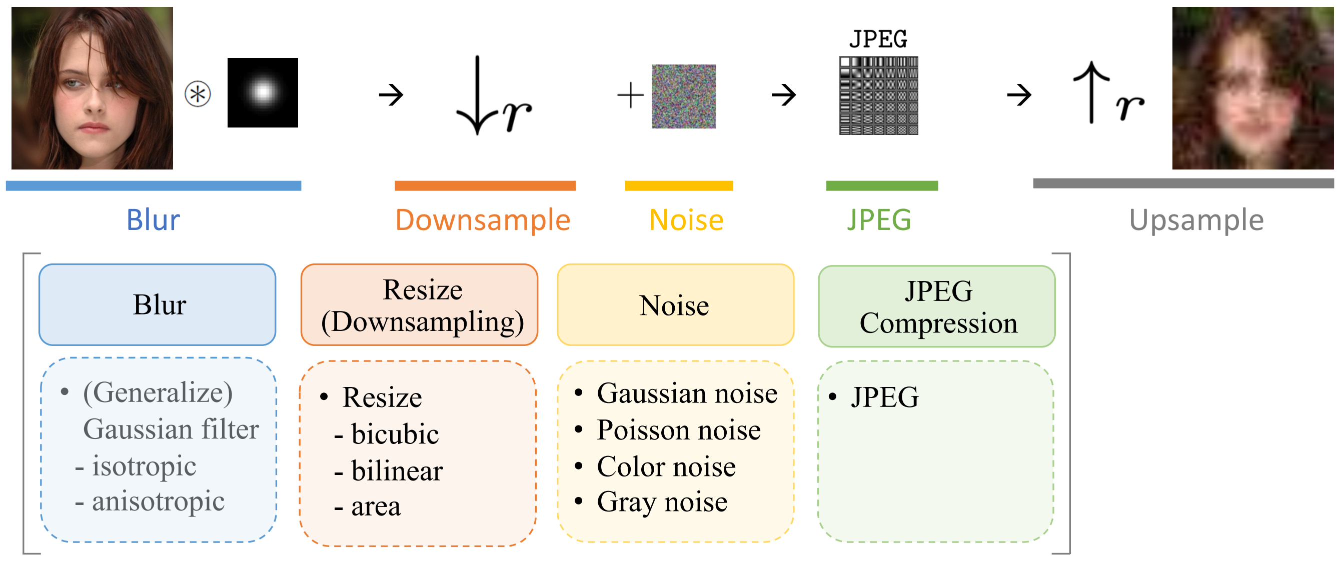
Resize (Downsampling)
- Resize
  - bicubic
  - bilinear
  - area

Noise
- Gaussian noise
- Poisson noise
- Color noise
- Gray noise

JPEG Compression
- JPEG

Degradation model

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



Gaussian noise: Gaussian noise has a probability density function equal to that of the Gaussian distribution

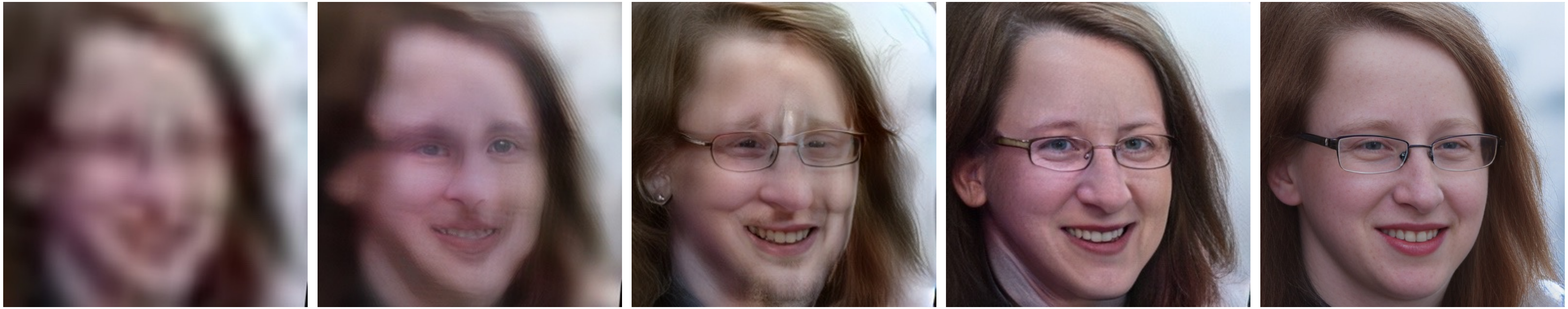
Poisson noise: model the sensor noise caused by statistical quantum fluctuations, that is, variation in the number of photons sensed at a given exposure level

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

Evaluation on blind face restoration

Real Input | DFDNet | GFP-GAN | GPEN | CodeFormer (Ours)

Evaluation on blind face restoration

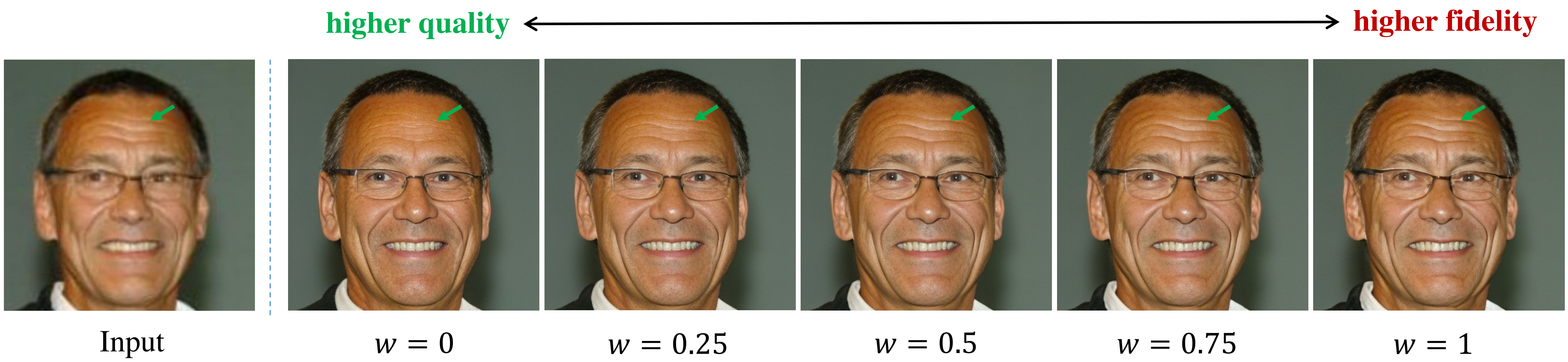


Real Input | DFDNet | GFP-GAN | GPEN | CodeFormer (Ours)

Evaluation on blind face restoration

Real Input DFDNet GFP-GAN GPEN CodeFormer (Ours)

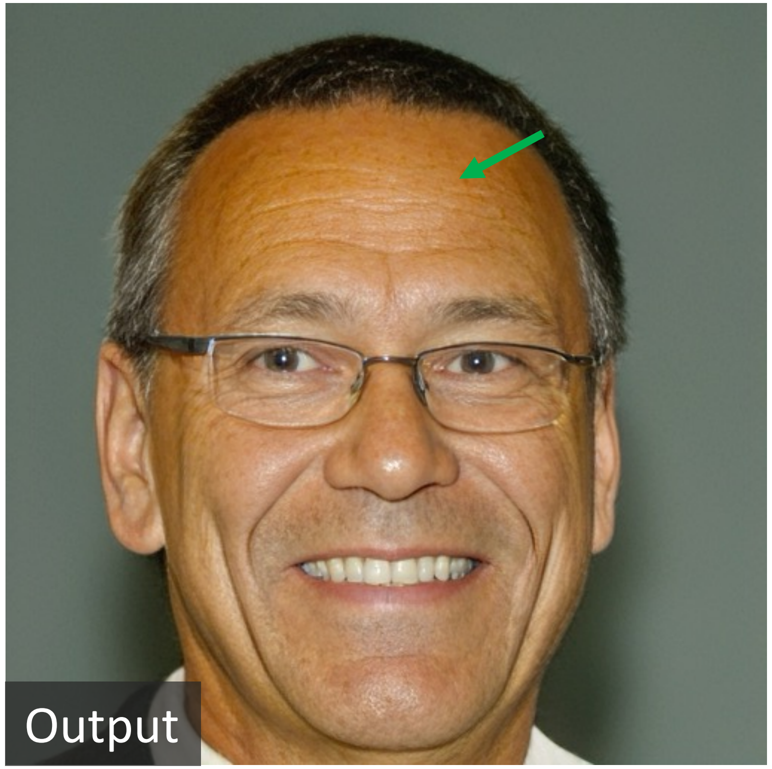
Evaluation on CFT module



Continuous Transitions between Image **Quality** and **Fidelity** via **Controllable Feature Transformation Module**

Evaluation on CFT module

Mild Degradation

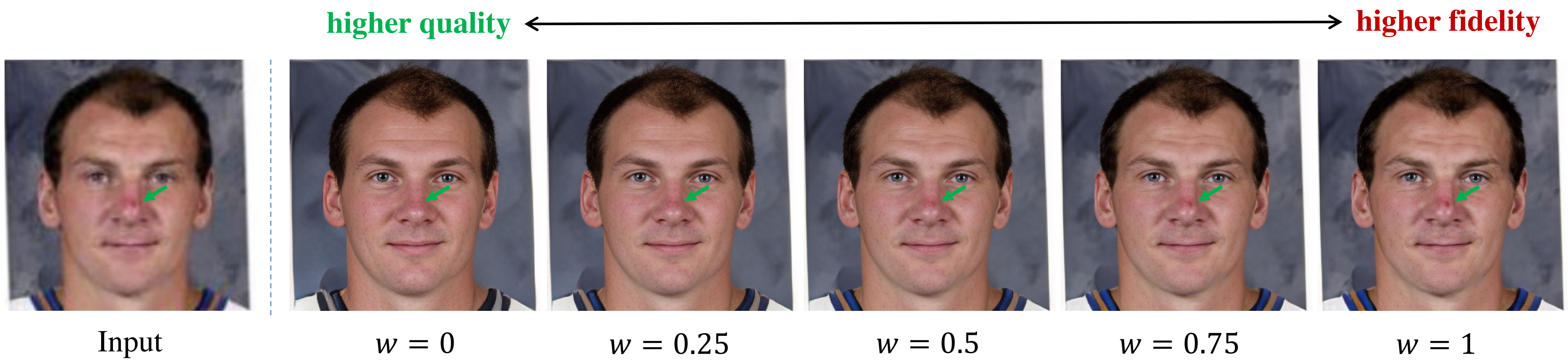


Quality **Fidelity**

w

$(w = 0)$ $(w = 1)$

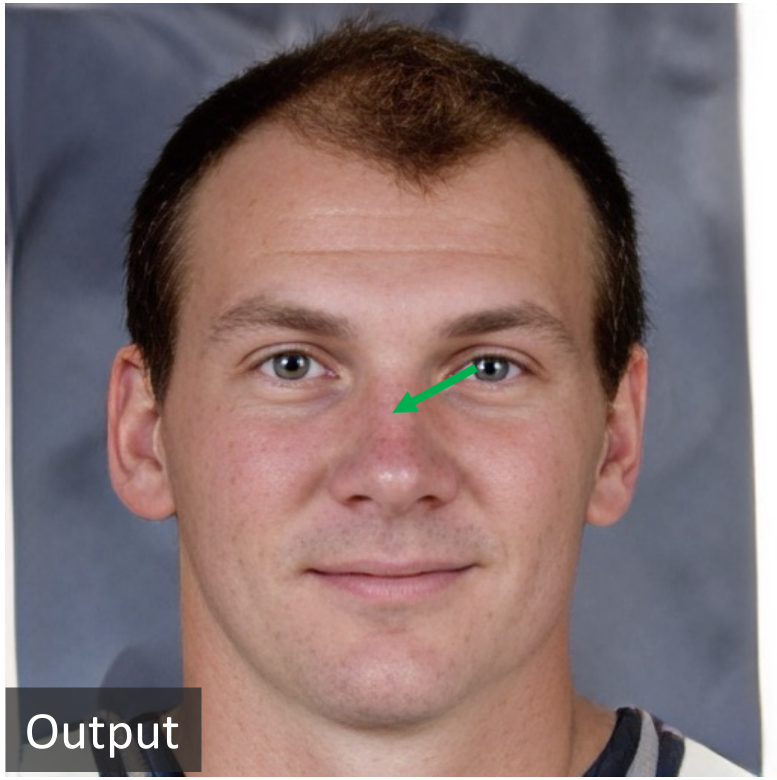
Evaluation on CFT module



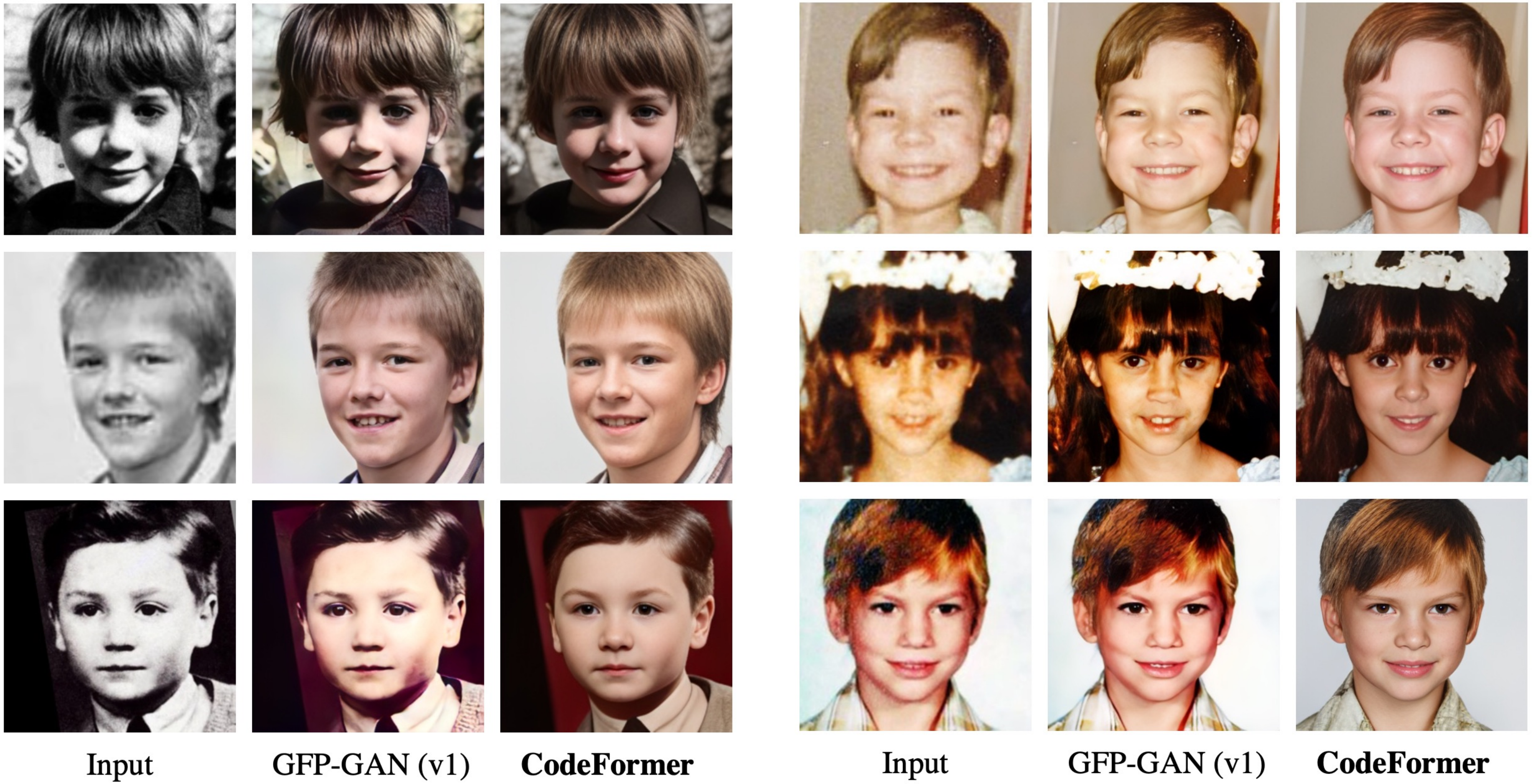
Continuous Transitions between Image **Quality** and **Fidelity** via **Controllable Feature Transformation Module**

Evaluation on CFT module

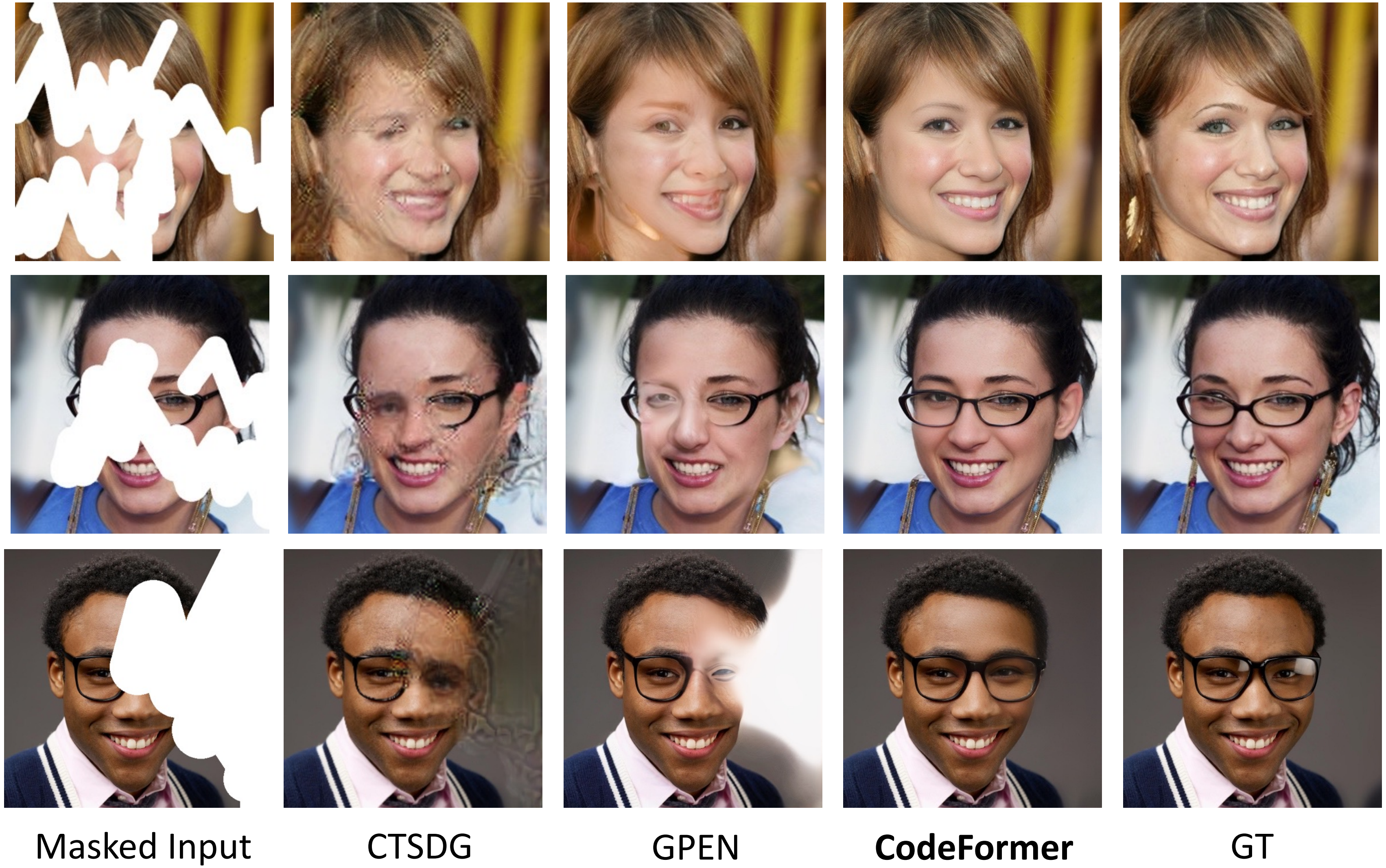
Mild Degradation



Face color enhancement



Face inpainting



Face inpainting (extremely large mask)

Masked Input (extremely large mask) | CTSDG | GPEN | **CodeFormer**

Old photo enhancement

Old Photo

CodeFormer

Old photo enhancement

Old Photo

CodeFormer

Old photo enhancement

Old Photo

CodeFormer

Old photo enhancement

Old Photo

CodeFormer

Old photo enhancement

AI-Generated Face

CodeFormer

Old photo enhancement

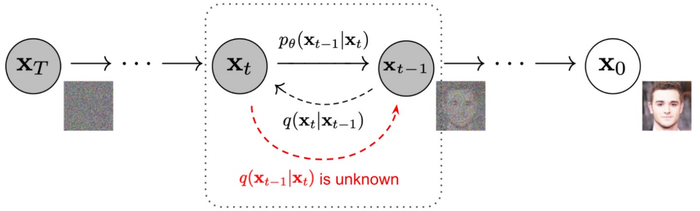
AI-Generated Face

CodeFormer

Discussions

- Next generation of generative priors

  StyleGAN2 -> VQGAN -> Diffusion Model?

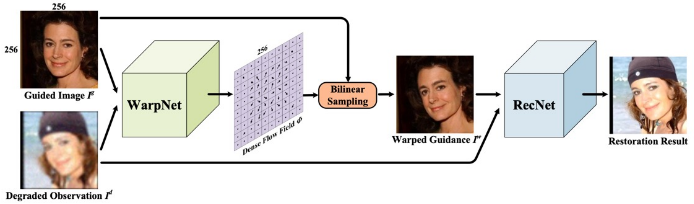


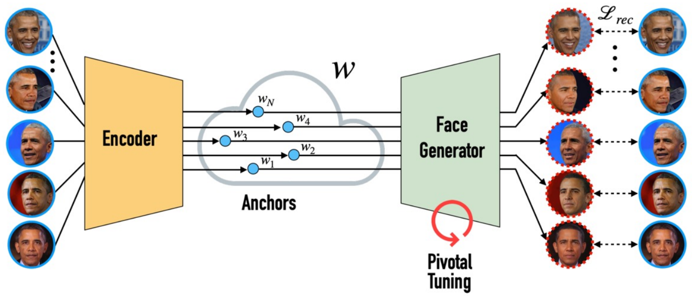
- Identity inconsistency issue

  Training Setting; Network Structure;

  Reference-based model (e.g., Li et al);

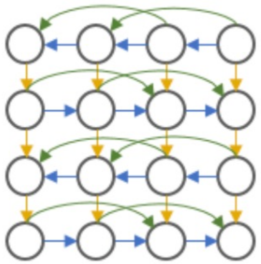
  Personalized model (e.g., MyStyle)



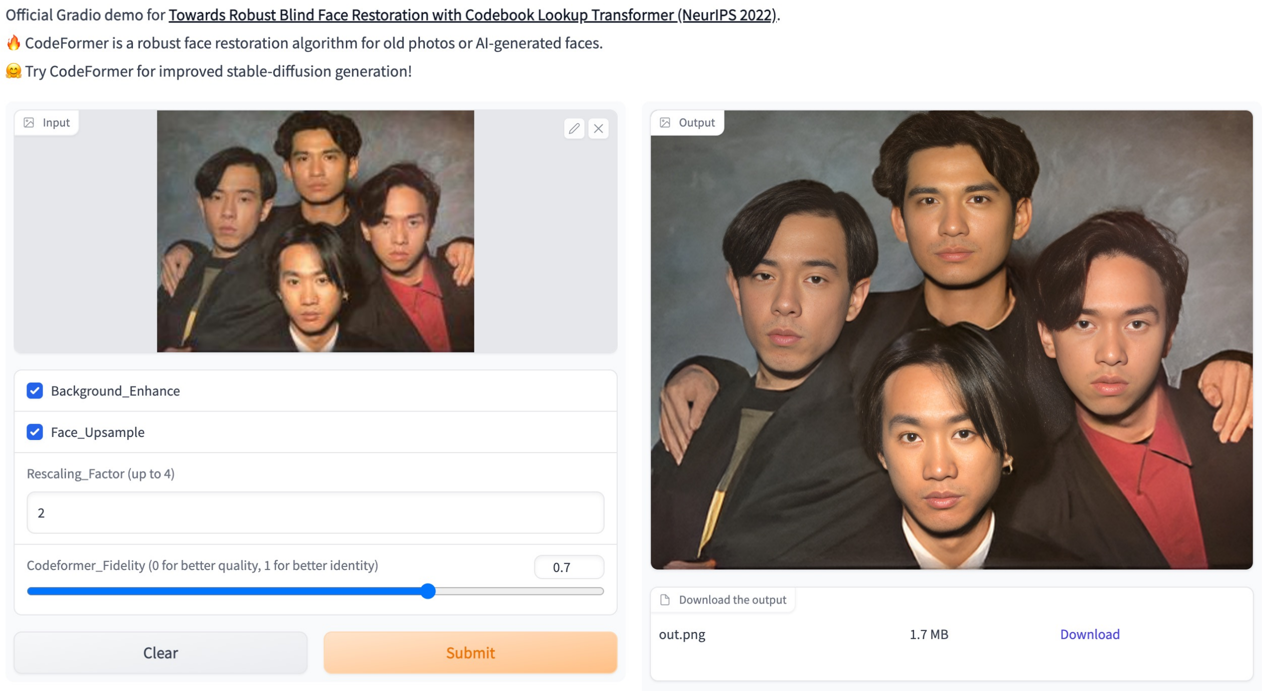


- Video face restoration

  Recurrent networks (e.g., BasicVSR series)



QA & Thanks!



https://github.com/sczhou/CodeFormer

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