Super-resolution in-the-wild

Daniil Shlenskii

Problem formulation

Image super-resolution (SR) aims at recovering the corresponding high-resolution (HR) images from the low-resolution (LR) images.

Given: $I_y = D(I_x)$, where D is a degradation mapping (generally, unknown),
I_x is the corresponding HR image.

Goal: to get an HR approximation $\hat{I}_x = S(I_y)$ of the ground truth HR image I_x
given LR image I_y,
where S is the super-resolution model.

Evaluation

PSNR

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

LPIPS

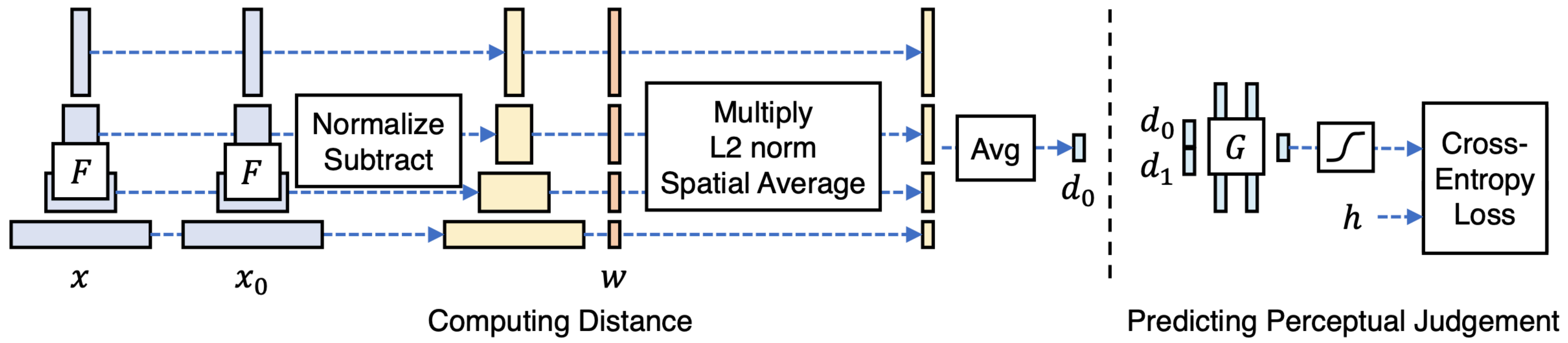


Figure 3: **Computing distance from a network** (Left) To compute a distance d_0 between two patches, x, x_0, given a network $\mathcal{F}$, we first compute deep embeddings, normalize the activations in the channel dimension, scale each channel by vector w, and take the ℓ_2 distance. We then average across spatial dimension and across all layers. (Right) A small network $\mathcal{G}$ is trained to predict perceptual judgment h from distance pair (d_0, d_1).

SSIM

$$SSIM(I, \hat{I}) = (l(I, \hat{I}))^{\alpha} \cdot (c(I, \hat{I}))^{\beta} \cdot (s(I, \hat{I}))^{\gamma},\ \alpha, \beta, \gamma > 0,$$

l, c, s are similarities in *luminance, contrast* and *structure* respectively.

CLIP-IQA

$$s(\hat{I}) = \frac{e^{s_1}}{e^{s_1} + e^{s_2}},$$

s_1 is a CLIP-score given prompt “Good photo”

s_2 is a CLIP-score given prompt “Bad photo”

RealSRGAN degradation

Classical degradation setup:

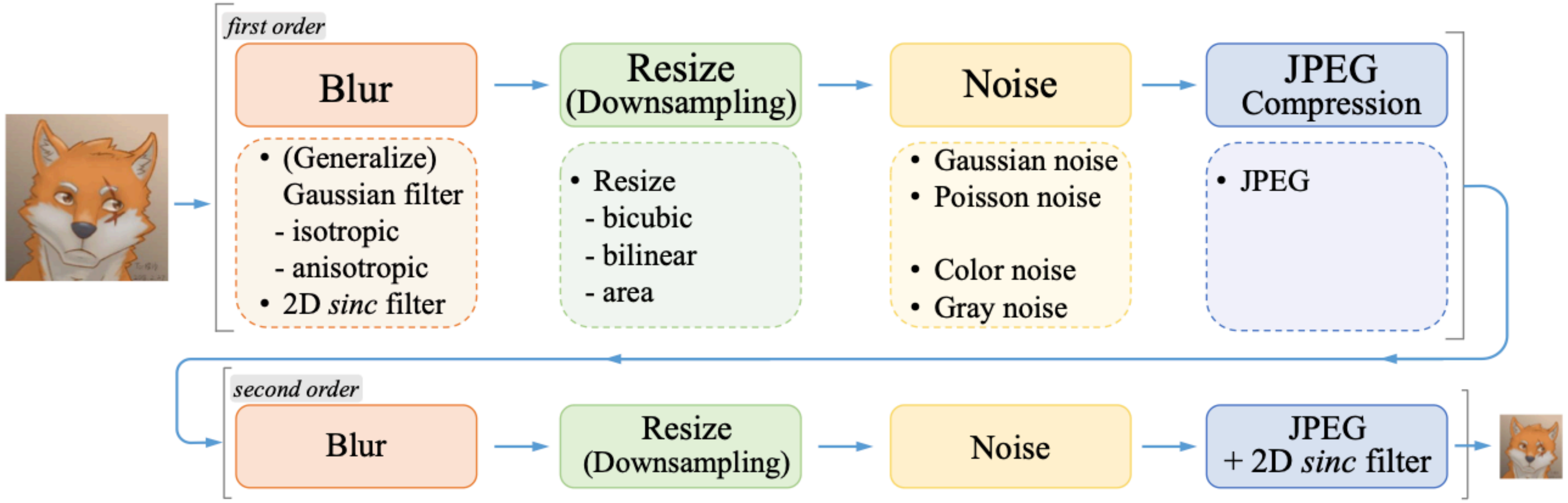
$$I_y = D(I_x) = (I_x) \downarrow_s ,$$

where $\downarrow_s$ is a downsampling operation with the scaling factor s.

RealSRGAN degradation setup:

$$I_y = D^n(I_x) = (D_n \circ \cdots \circ D_2 \circ D_1)(I_x),$$

where D_i is a set of degradations



Xintao Wang, et al, Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data, CVPR

Pixel-wise losses

$$L_1(I, \hat{I}) = \frac{1}{hwc} \sum_{i,j,k} | I_{i,j,k} - \hat{I}_{i,j,k} |$$

$$L_2(I, \hat{I}) = \frac{1}{hwc} \sum_{i,j,k} | I_{i,j,k} - \hat{I}_{i,j,k} |^2$$

Do not take into account image quality (e.g., perceptual quality, textures), so results often are perceptually unsatisfying with overshoots textures.

Generative Adversarial Network (GAN)

Beside the correspondence between LR and super-resolved images,
we want the last to be from the distribution of HR image.

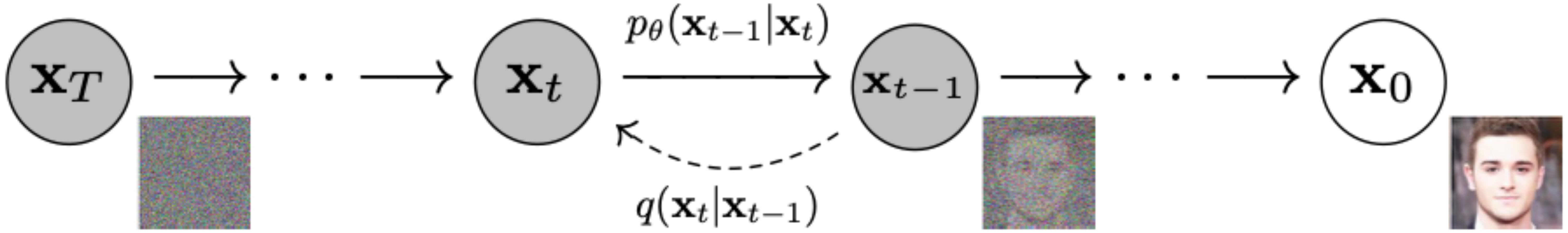
GANs are good at learning complex distributions.
GAN-objective:

$$\min_{D} \max_{G} \ \mathbb{E}_{x \sim p_{data}} \log(D(x)) + \mathbb{E}_{x \sim p_G} \log(1 - D(x))$$

Employing pixel-wise losses with this one gives perceptually better results.

Common GAN drawbacks for data generation problem, which are not not drawbacks at all for super-resolution problem:
- Mode collapse (now we want ONE high-quality upscale)
- Unstable training (pixel-wise losses make it more stable)

Diffusion-based methods. Theory



Given:

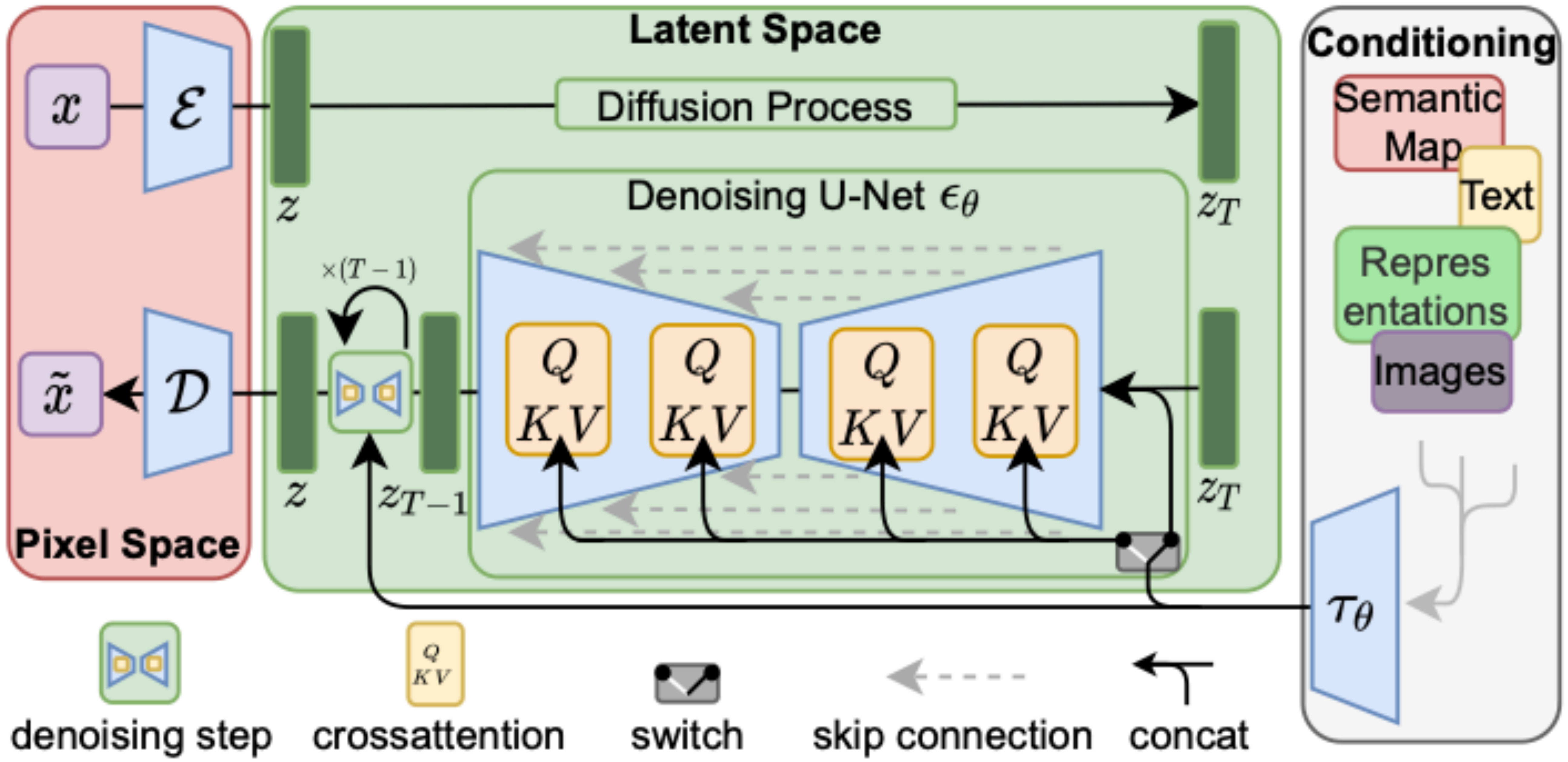
Data: $x_0 \sim q(x)$

Diffusion process: $q(x_t \,|\, x_{t-1}) = \mathcal{N}(x_t \,|\, \sqrt{1-\beta_t} x_{t-1}, \beta_t I)$, $x_T \sim \mathcal{N}(0, I)$

Goal: to learn how to reverse this process:

$$q(x_{t-1} \,|\, x_t) \approx p_\theta(x_{t-1} \,|\, x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

Latent Diffusion Model



Robin Rombach, et al, High-Resolution Image Synthesis with Latent Diffusion Models, CVPR

Diffusion-based methods. SR3

The main idea is to condition diffusion model on the LR image.

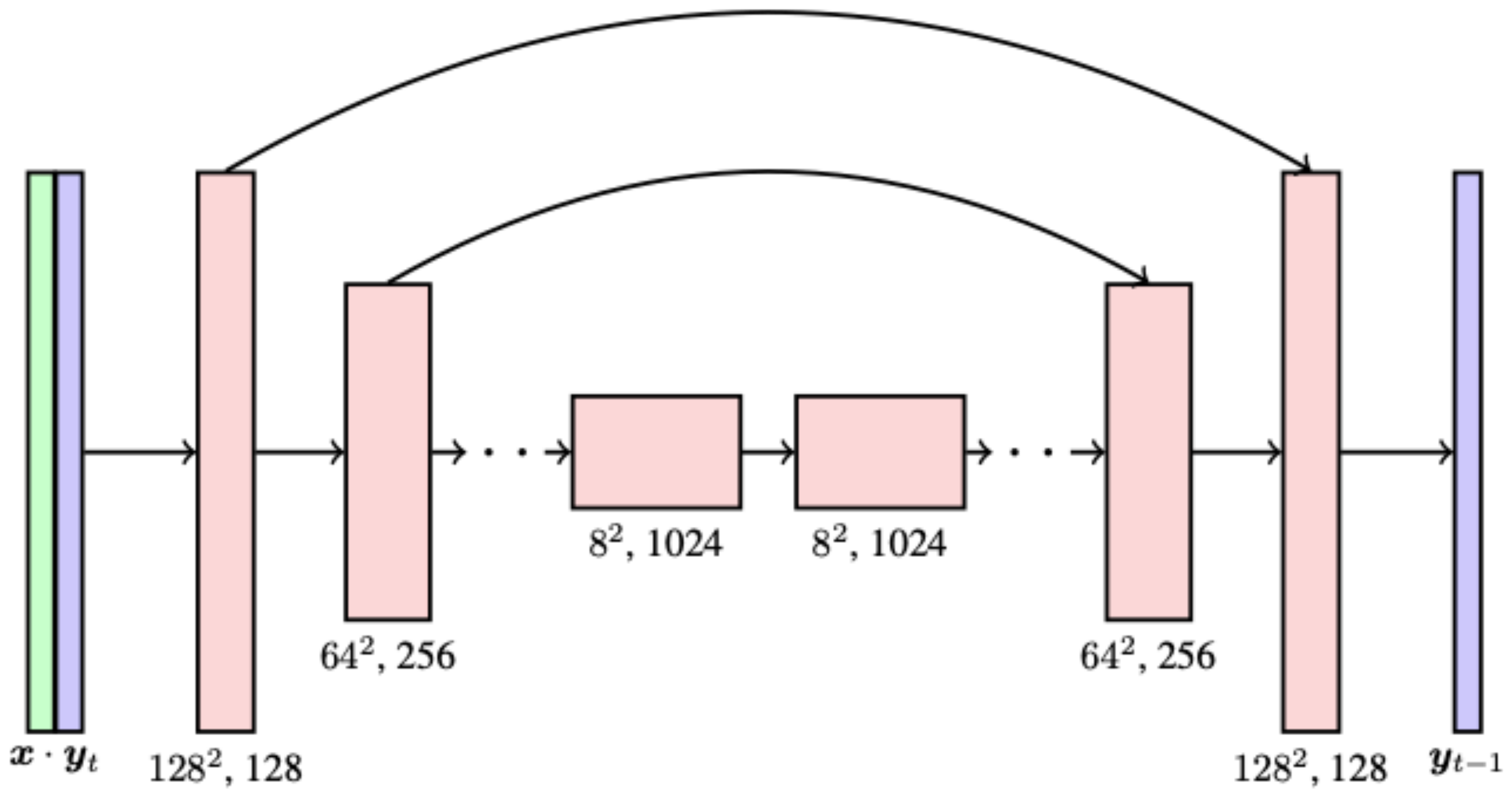


Figure A.1: Description of the U-Net architecture with skip connections. The low resolution input image $\boldsymbol{x}$ is interpolated to the target high resolution, and concatenated with the noisy high resolution image $\boldsymbol{y}_t$. We show the activation dimensions for the example task of $16\times16 \rightarrow 128\times128$ super resolution.

Chitwan Saharia, et al, Image Super-Resolution via Iterative Refinement, CVPR

Diffusion-based methods. StableSR

The main idea is to utilize generative prior of pretrained StableDiffusion.

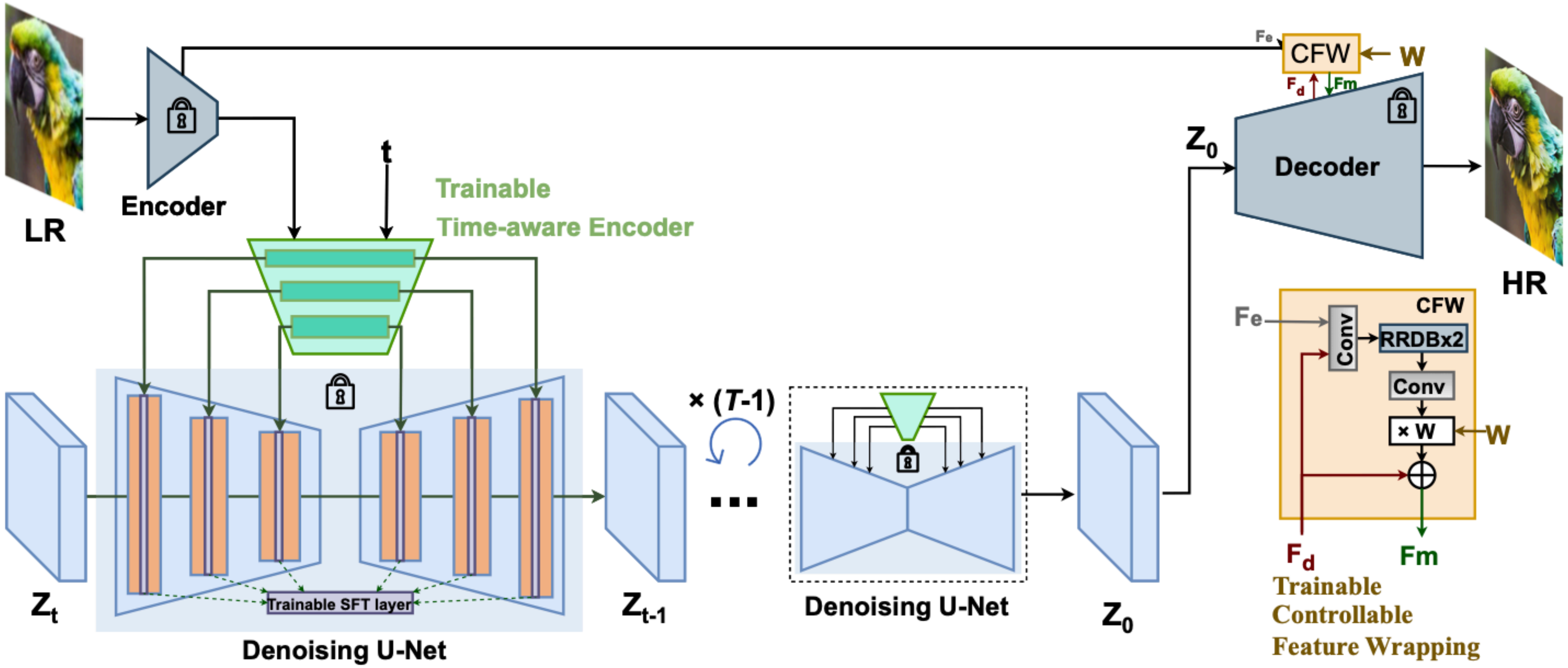
Two requirements for such an approach:

1. The resulting model must generate a plausible HR corresponding to given LR image.

2. The model should introduce minimal alterations to the original Stable Diffusion model to prevent disrupting the prior encapsulated within.

Jianyi Wang, et al, Exploiting Diffusion Prior for Real-World Image Super-Resolution, CVPR

Diffusion-based methods. StableSR



SFT: $\hat{F}^n_{diff} = (1 + \alpha^n) \odot F^n_{diff} + \beta^n; \;\; \alpha^n, \beta^n = \mathcal{M}^n_\theta(F^n)$

Jianyi Wang, et al, Exploiting Diffusion Prior for Real-World Image Super-Resolution, CVPR

Diffusion-based methods. StableSR

CFW (realism-fidelity trade-off)**:** $F_m = F_d + C_\theta(F_e, F_d) \times w$

Zoomed LR StableSR (w = 0.0) StableSR (w = 0.5) StableSR (w = 1.0)

Jianyi Wang, et al, Exploiting Diffusion Prior for Real-World Image Super-Resolution, CVPR

Diffusion-based methods. StableSR

Color-shifting:

$$x^c = \frac{\hat{x}^c - \mu_{\hat{x}}^c}{\sigma_{\hat{x}}^c} \cdot \sigma_y^c + \mu_y^c,$$

where $c \in \{r, g, b\}$ and

$\mu_{\hat{x}}^c, \sigma_{\hat{x}}^c$ (or μ_y^c, σ_y^c) are the mean and std estimated from c-th channel of $\hat{x}$ (or y)

Jianyi Wang, et al, Exploiting Diffusion Prior for Real-World Image Super-Resolution, CVPR

Diffusion-based methods. StableSR

Main takeaways:

- ControlNet

- Color-shifting problem

- Quality-fidelity tradeoff

- Prompts are null during training

Jianyi Wang, et al, Exploiting Diffusion Prior for Real-World Image Super-Resolution, CVPR

Diffusion-based methods. DiffBIR

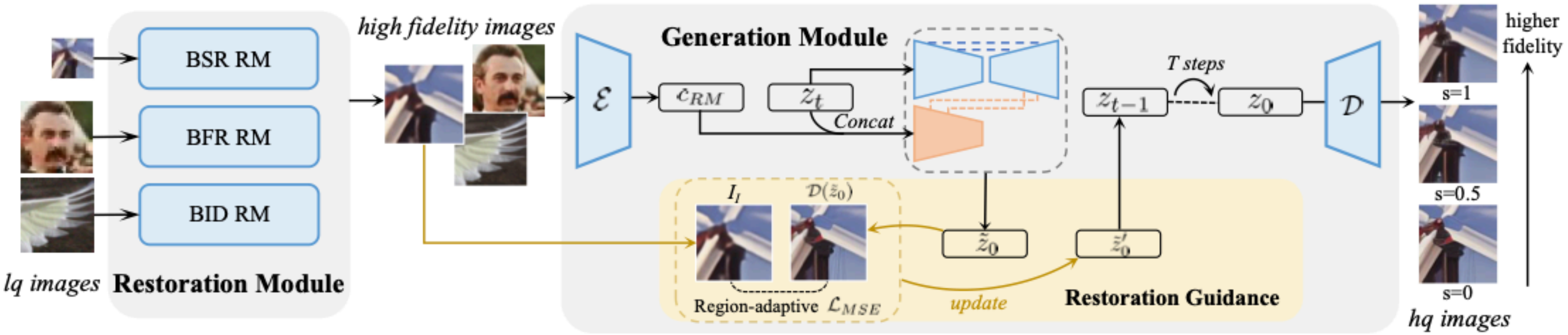
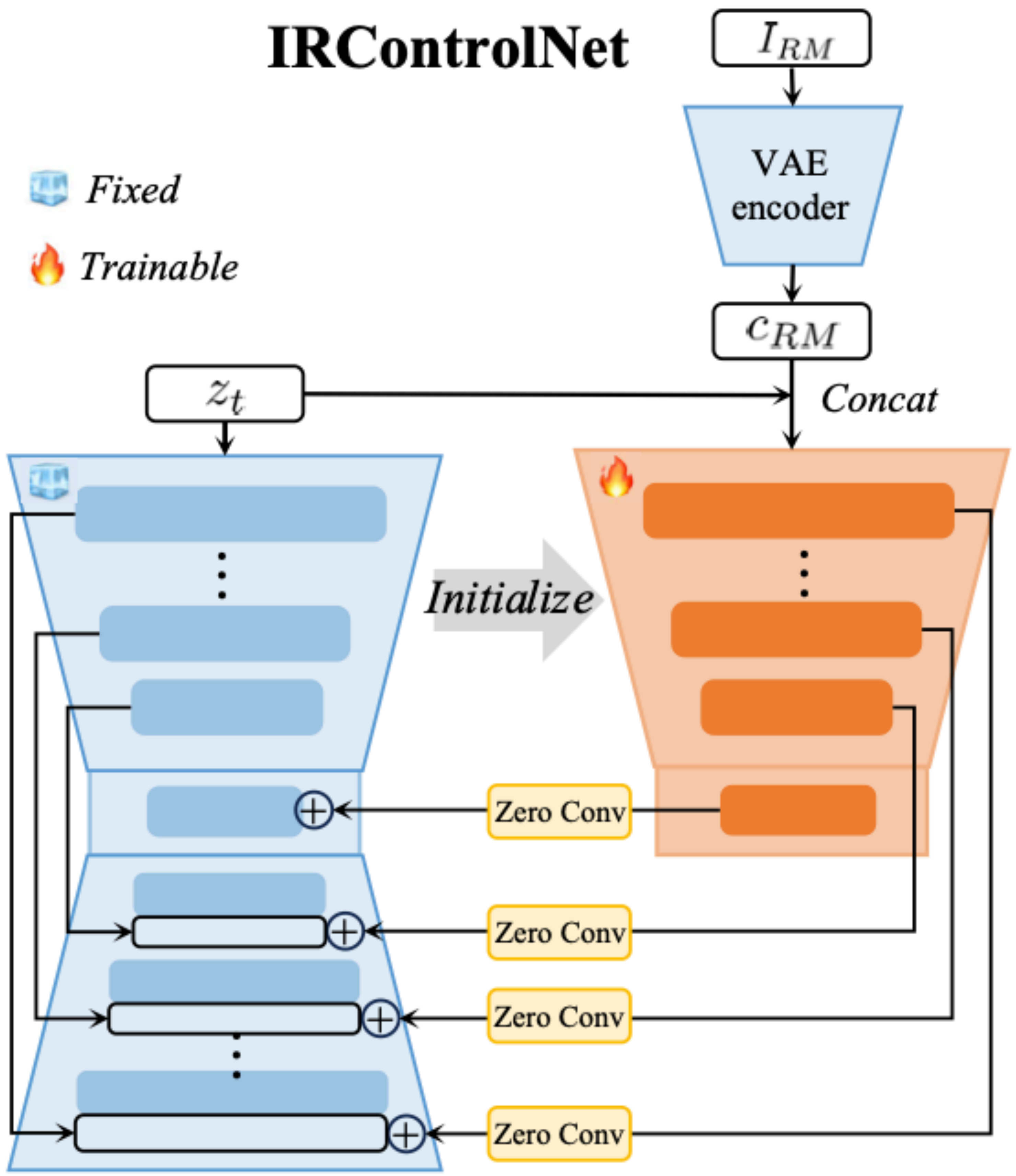


Figure 3. The two-stage pipeline of DiffBIR. 1) Restoration Module (RM) for degradation removal; 2) Generation Module (GM) for realistic image reconstruction with optional region-adaptive restoration guidance for a trade-off between *quality* and *fidelity*.

Xinqi Lin, et al, DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior

Diffusion-based methods. DiffBIR

ControlNet



Xinqi Lin, et al, DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior

Diffusion-based methods. DiffBIR

Quality-Fidelity tradeoff

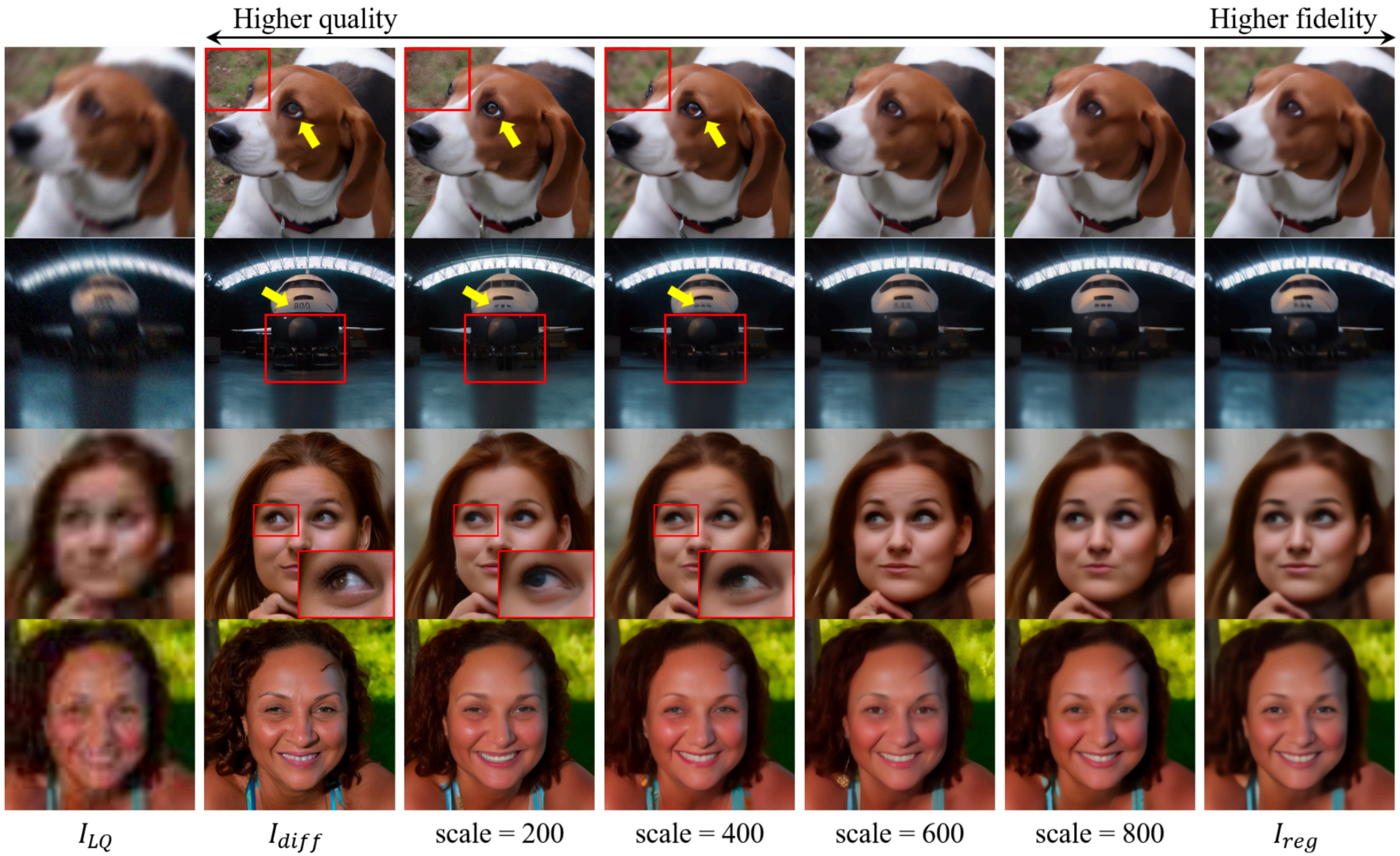
Classifier guidance: $\nabla_{x_t} \log p(x_t \mid y) = \nabla_{x_t} \log p(x_t) + s \cdot \nabla_{x_t} \log p(y \mid x_t)$

Restoration guidance: $p(y \mid x_t) = \dfrac{1}{Z} \exp(-L(D[x_0^\theta(x_t)], y))$

Xinqi Lin, et al, DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior

Diffusion-based methods. DiffBIR

Quality-Fidelity tradeoff



Xinqi Lin, et al, DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior

Diffusion-based methods. DiffBIR

Main takeaways:

- Restoration Module

- ControlNet initialization

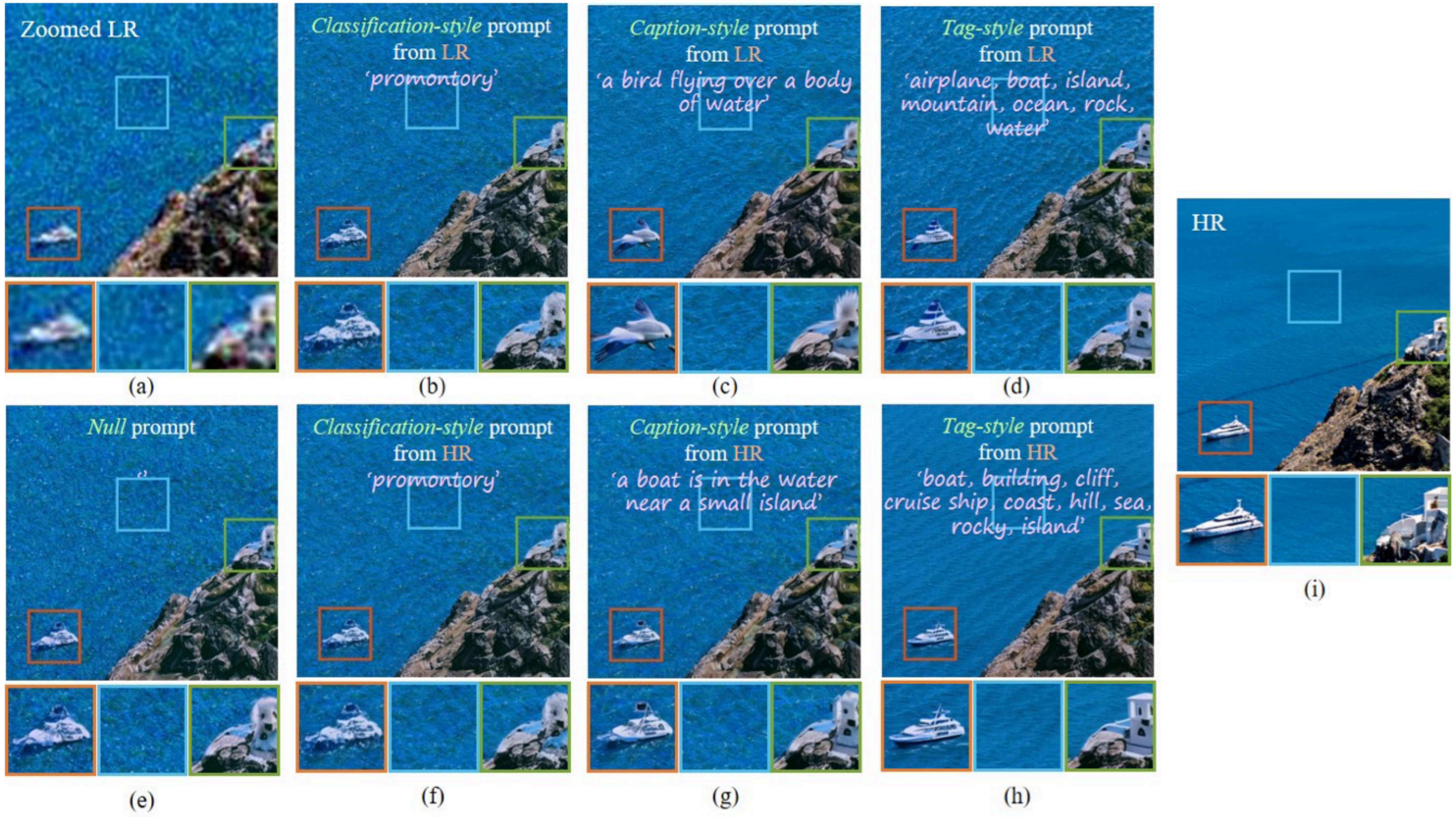
- Restoration guidance

- Prompts on inference:
    positive: null
    negative: [“low quality”, “blurry”]

Xinqi Lin, et al, DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior

Diffusion-based methods. SeeSR

Stable Diffusion is a text-to-image model.
The main idea of SeeSR model is in preparation the most appropriate text prompts.

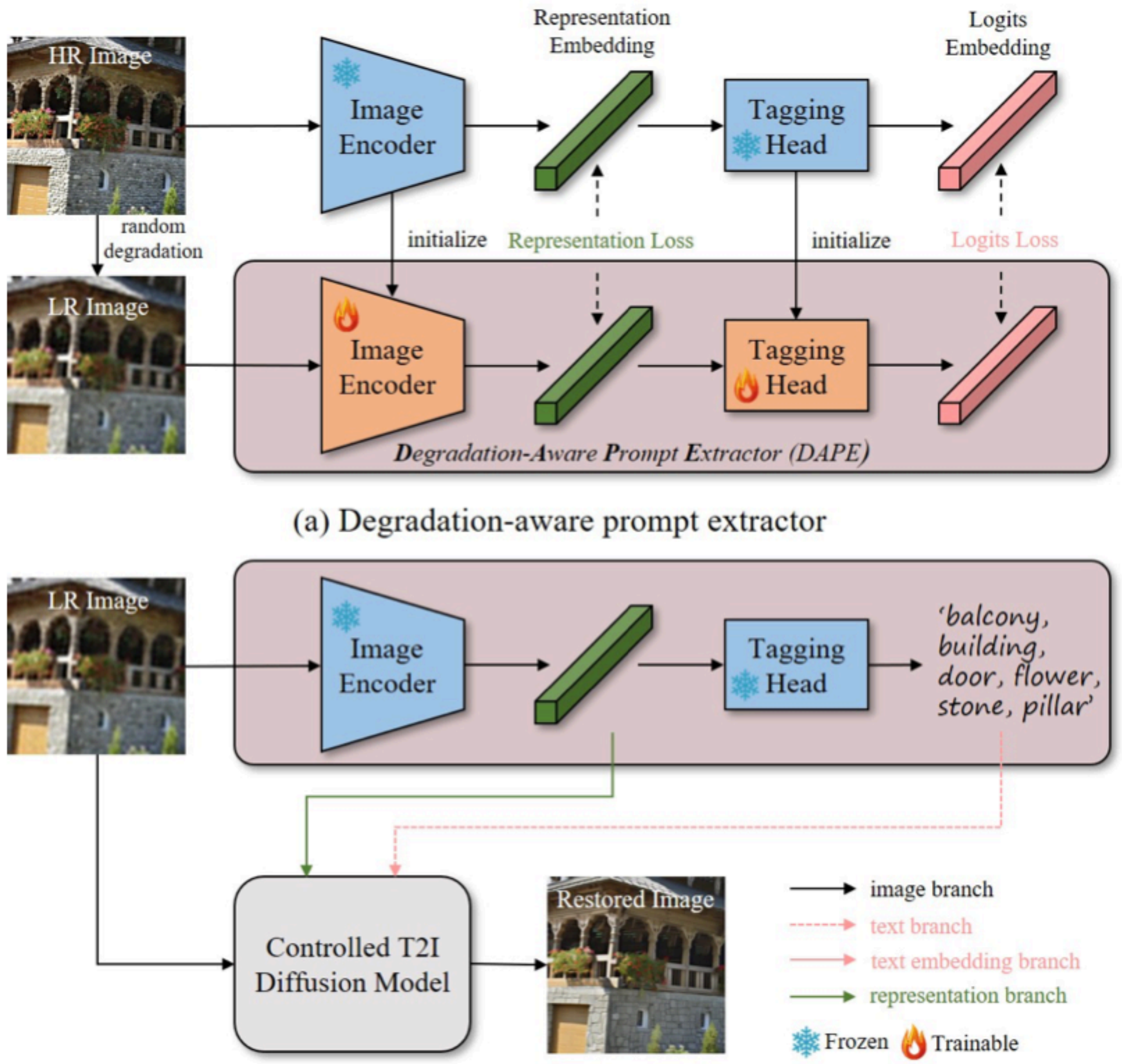


	Rich Objects	Concise Description	Degradation Aware
Classification-style	✗	✓	✓
Caption-style	✓	✗	✗
Tag-style	✓✓	✓	✗
Our DAPE	✓✓	✓	✓

Rongyuan Wu, et al, SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution, CVPR

Diffusion-based methods. SeeSR

Tagging Module



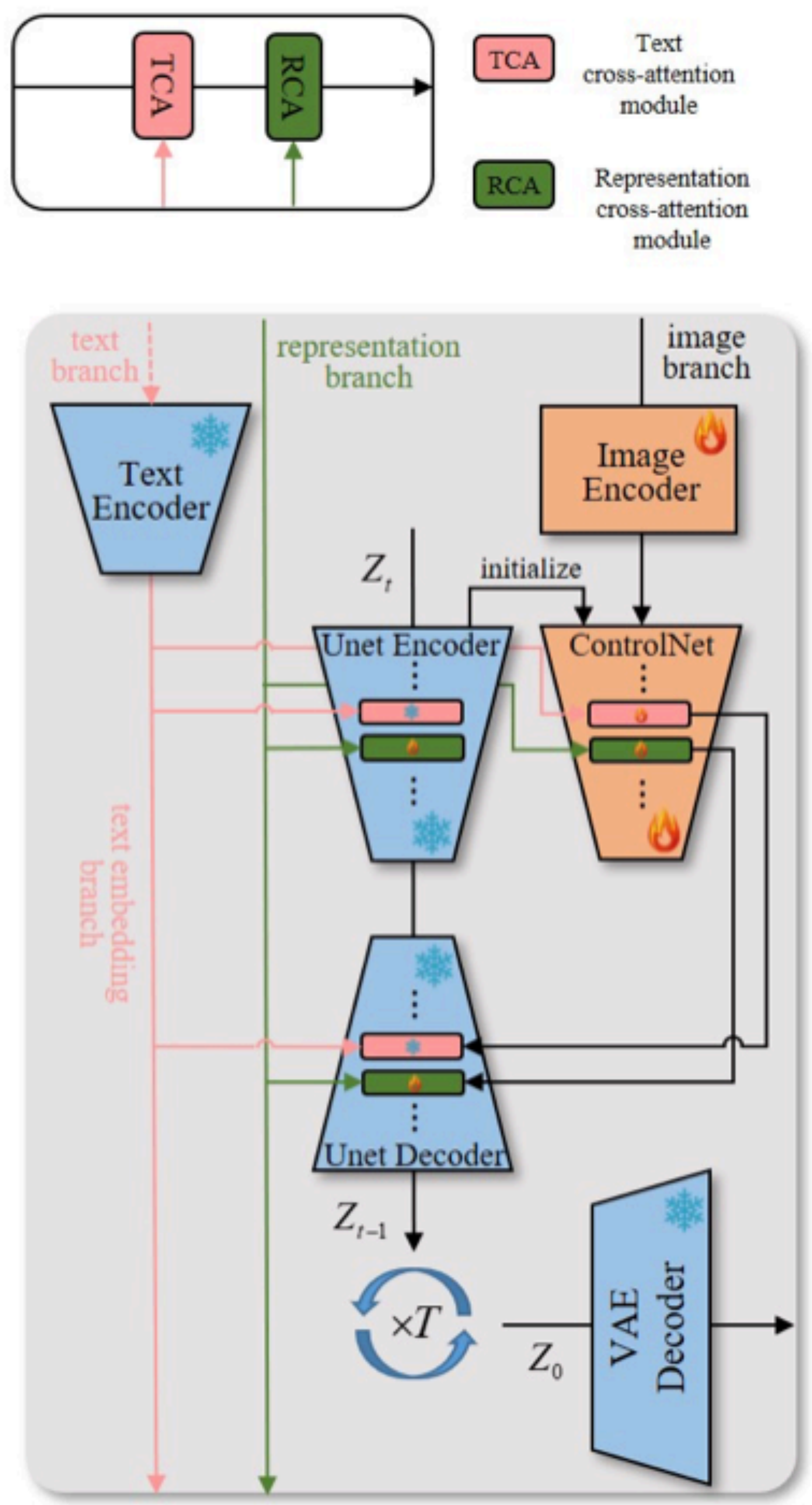
(a) Degradation-aware prompt extractor

(b) Real-ISR with DAPE

Rongyuan Wu, et al, SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution, CVPR

Diffusion-based methods. SeeSR

Learnable LR encoder



(c) Controlled T2I diffusion model

Rongyuan Wu, et al, SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution, CVPR

Diffusion-based methods. SeeSR

Main takeaways:

- Tagging module

- Trainable Encoder

- Importance of text

Rongyuan Wu, et al, SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution, CVPR

Diffusion-based methods. Takeaways

Main takeaways:
- Utilizing Stable Diffusion generative prior via ControlNet
- Image Restoration Module
- Importance of text prompts
- Techniques for quality-fidelity tradeoff

Pros:
- Good image quality
- Text driven super-resolution
- Small number of trainable parameters

Cons:
- Expensive inference:
    - many steps
    - heavy model
- Not high enough fidelity

Rongyuan Wu, et al, SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution, CVPR

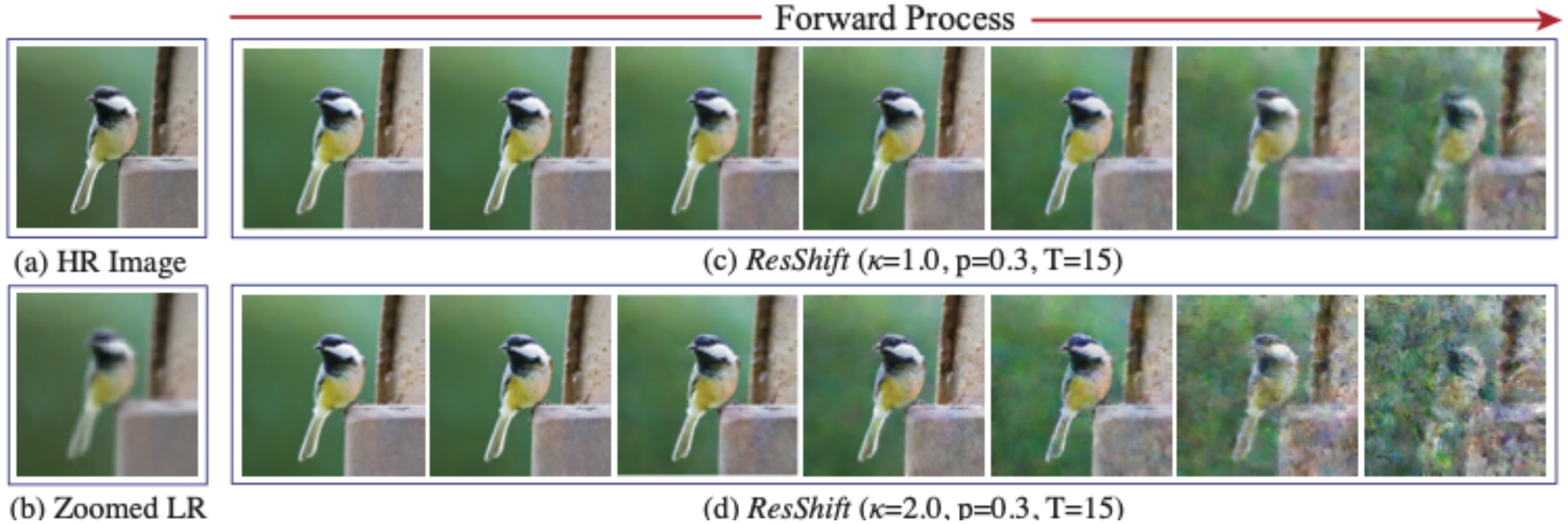
DMs acceleration. ResShift

Let's consider SR3 approach.
Starting sampling from pure gaussian noise seems redundant:
  since we have LR image we already have a content of HR image,
  so first steps of sampling process are redundant.

ResShift proposes forward diffusion process, which starts at the clean image and ends at the noised LR image:

$$q(x_t | x_{t-1}, y) = \mathcal{N}(x_t | x_{t-1} + \alpha_t(y - x_0), \alpha_t),$$

$$q(x_t | x_0, y) = \mathcal{N}(x_t | x_0 + \eta_t(y - x_0), \eta_t I), \quad \alpha_t = \eta_t - \eta_{t-1}, \alpha_1 = \eta_1$$



Zongsheng Yue, et al, ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting, CVPR

DMs acceleration. ResShift

Main takeaways:

- SR specific diffusion process

- Uses 15 steps

- Better than baseline using 100 steps

Zongsheng Yue, et al, ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting, CVPR

DMs acceleration. AddSR

The main idea is to unite Diffusion and GAN paradigms:

$$L_{ADD} = L_{dis}(\hat{x}_\theta(x_s, s), \hat{x}_\phi(x_t, t)) + \lambda L_{adv}(\hat{x}_\theta(x_s, s), x_0, \psi)$$

$$T_{student} = \{t_1, t_2, t_3, t_4\}$$

$$s \in T_{student}$$

$$x_0 \xrightarrow{diffuse} x_s \xrightarrow[prediction]{student} \hat{x}_\theta(x_s, s) \xrightarrow{diffuse} x_t \xrightarrow[prediction]{teacher} \hat{x}_\phi(x_t, t)$$

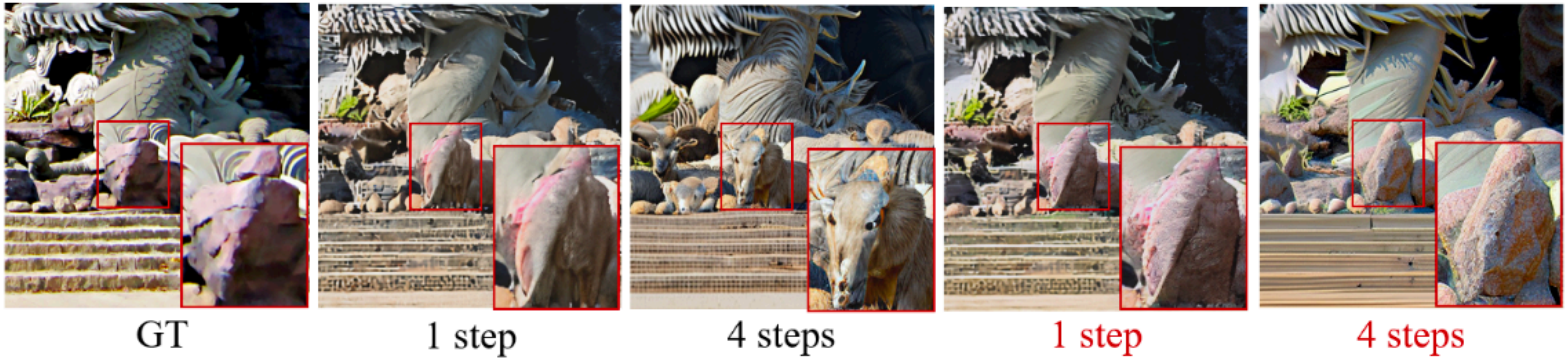
Rui Xie, et al, AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation, CVPR

DMs acceleration. AddSR

Time-adapting loss

$$L_{ADD} = \gamma_s L_{dis}(\hat{x}_\theta(x_s, s), \hat{x}_\phi(x_t, t)) + \lambda L_{adv}(\hat{x}_\theta(x_s, s), x_0, \psi)$$

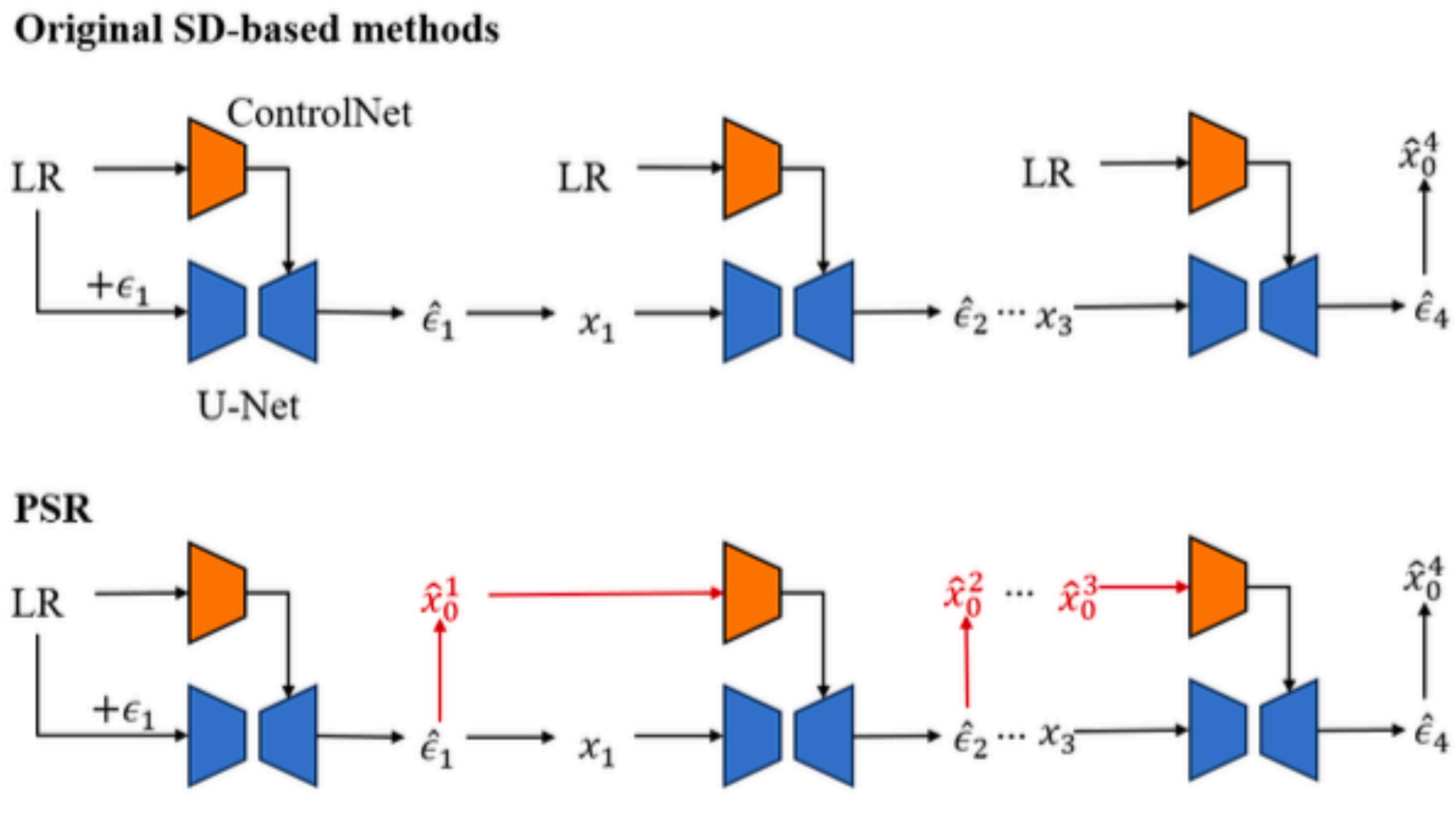
γ_s is lower for larger timestamp s

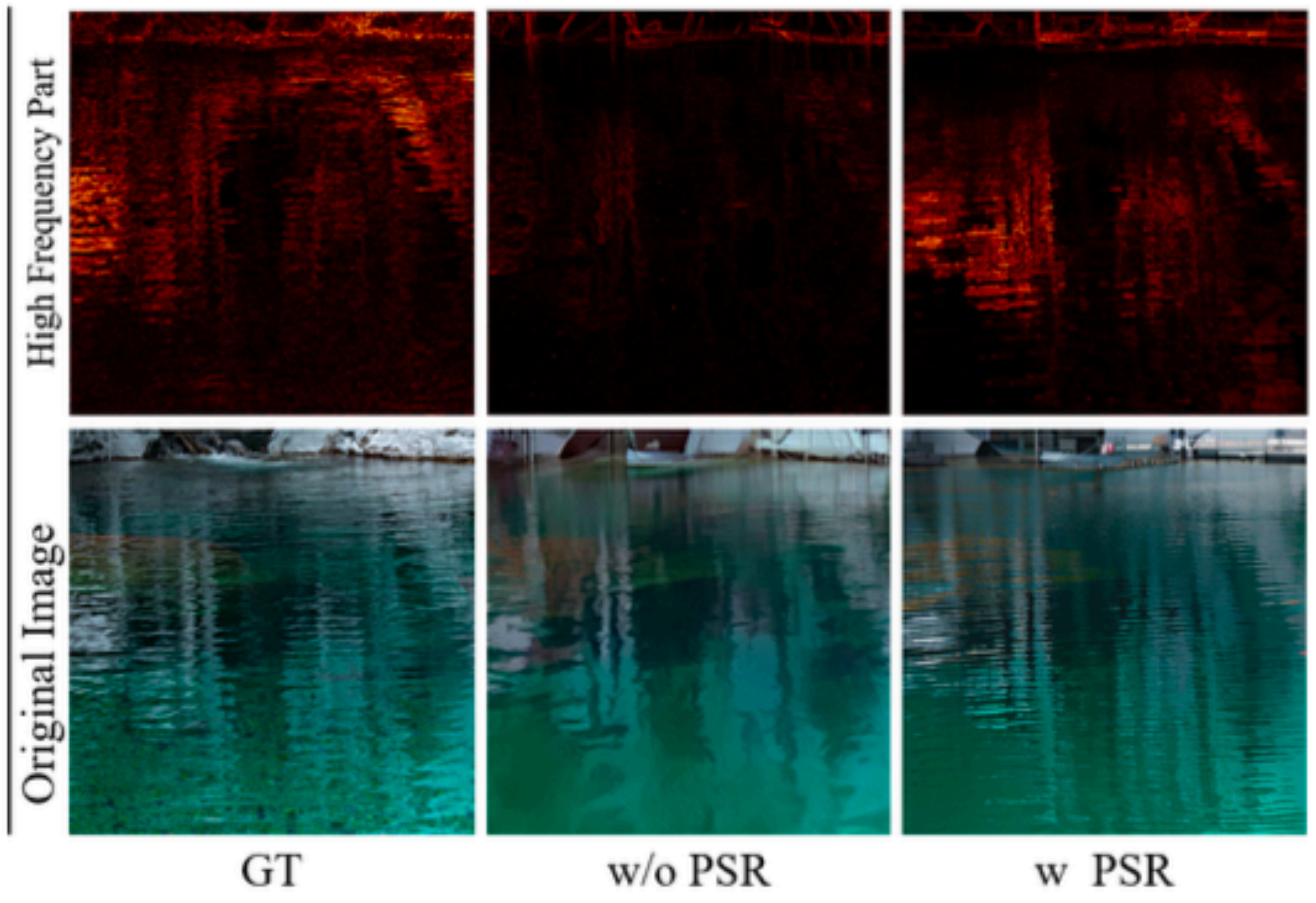


Rui Xie, et al, AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation, CVPR

DMs acceleration. AddSR

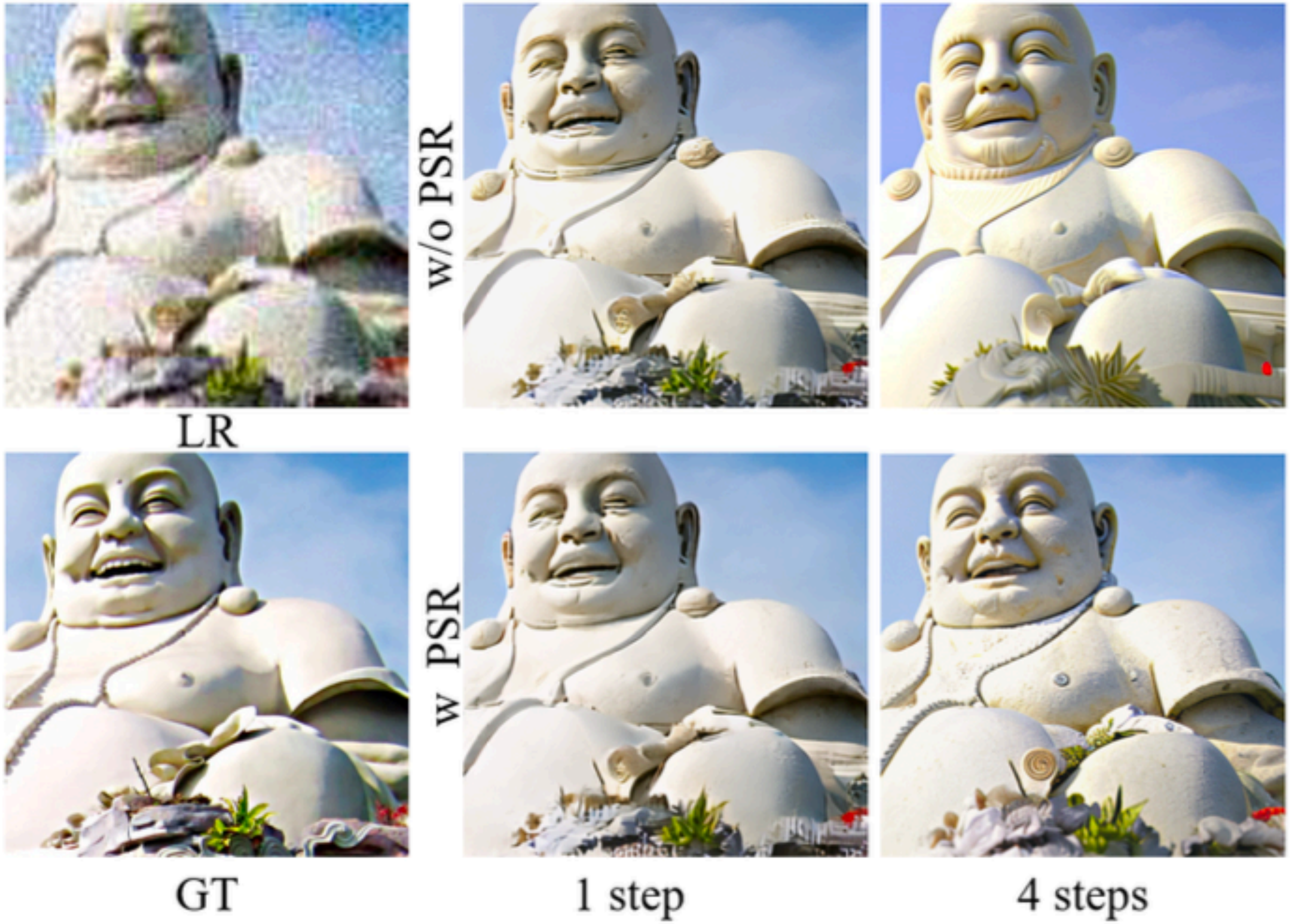
Prediction-based Self-Refinement (PSR)





Rui Xie, et al, AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation, CVPR

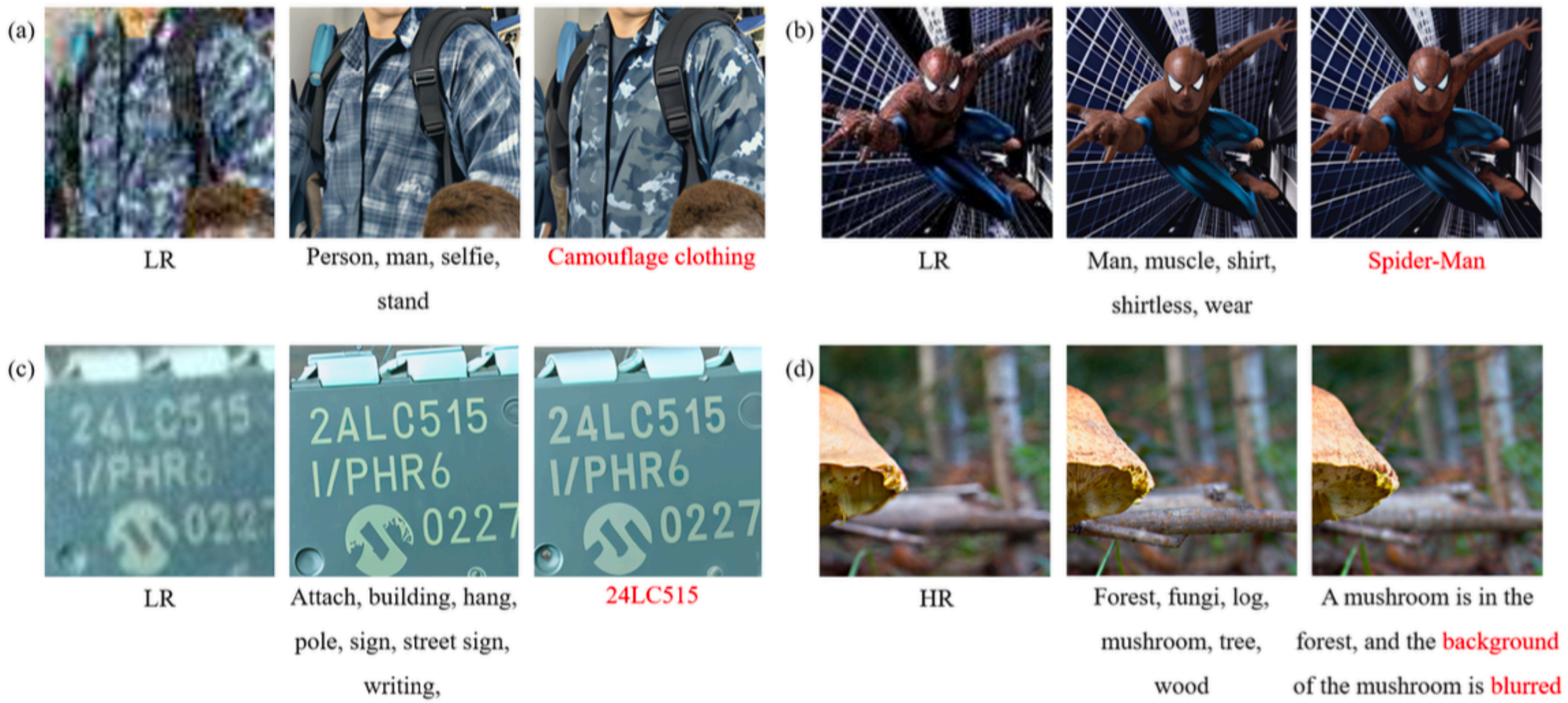
DMs acceleration. AddSR



Rui Xie, et al, AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation, CVPR

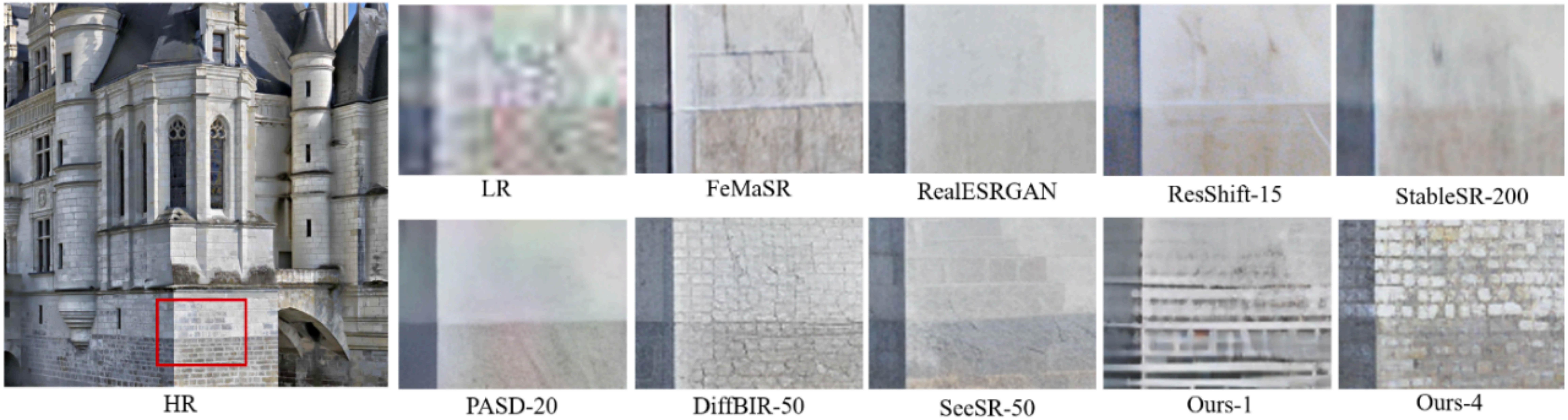
DMs acceleration. AddSR

Prompt-guided restoration



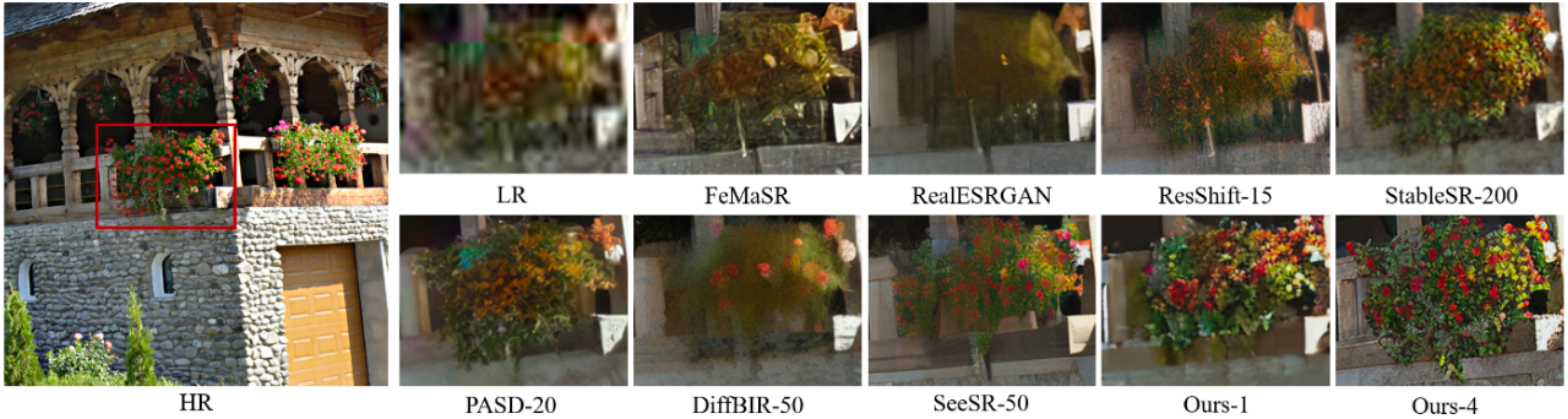
Rui Xie, et al, AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation, CVPR

DMs acceleration. AddSR



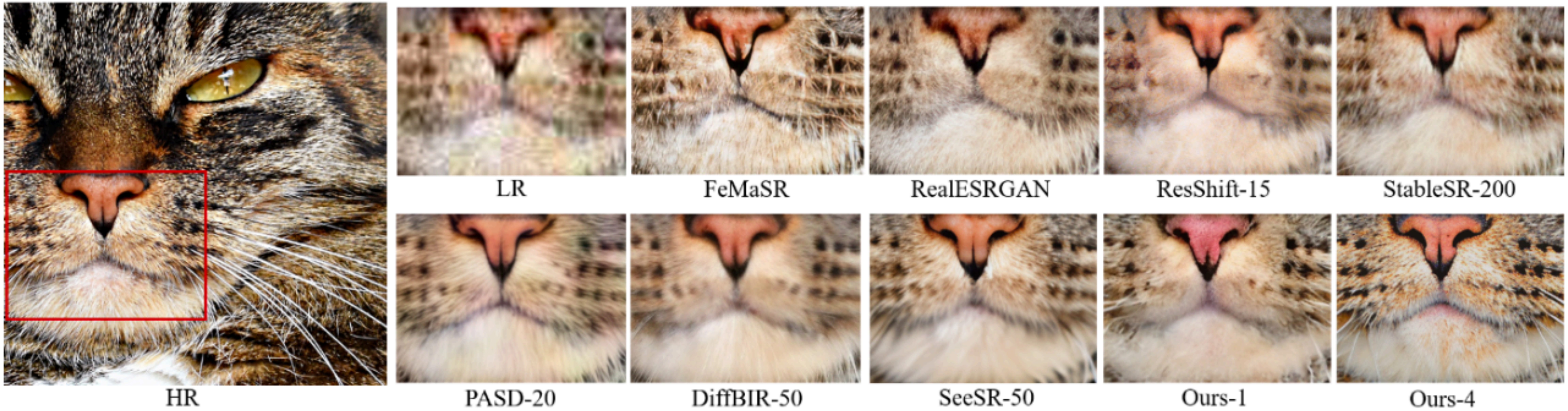
Rui Xie, et al, AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation, CVPR

DMs acceleration. AddSR



Rui Xie, et al, AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation, CVPR

DMs acceleration. AddSR



Rui Xie, et al, AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation, CVPR

DMs acceleration. AddSR

Main takeaways:

- Diffusion + GAN-loss

- Uses 4 steps

- Time-adapting loss

- Prediction-based Self-Refinement

Rui Xie, et al, AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation, CVPR

Conclusion

Best Design Choices

1. GAN-loss: beneficial for high image quality
2. Diffusion “refinement” process: {2 or 4} refining steps are better than just 1
3. SD-based methods performs better if using good enough text prompts.
4. ResShift is an SR-specific diffusion model, which seems to be the most appropriate for further distillation:
   - not starting from pure noise
   - no heavy ControlNet

Question

1. How much SD generative prior is really needed?
2. Is text prompting important for not SD-based models?