DreamSR: Towards Ultra-High-Resolution Image Super-Resolution via a Receptive-Field Enhanced Diffusion Transformer

Supplementary Material

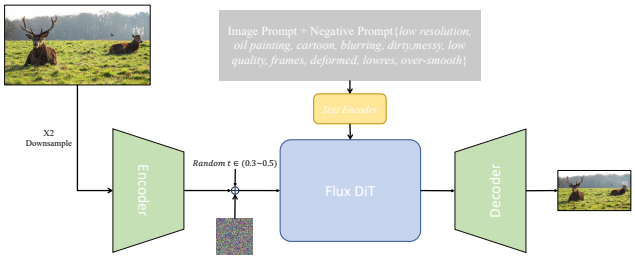


Figure 7. Image to image degradation pipeline for DreamSR.

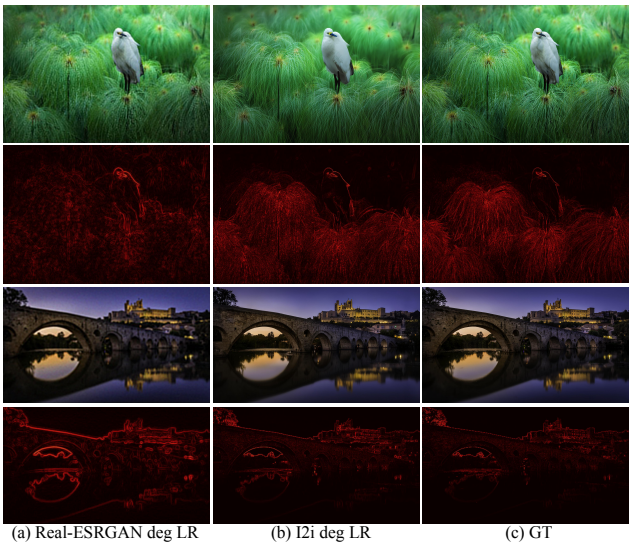
6. Detail of training scheme

The pseudo-code of our DreamSR's training scheme is summarized as Algorithm 1. To stabilize the training period, we first train the Patch Context aware MM-ControlNet for 20,000 steps. After that, we switch to a probabilistic training scheme, where the MM-ControlNet and Restoration Acceleration LoRA modules are updated with probabilities of 0.6 and 0.4, respectively, in conjunction with our stage-specific degradation pipelines. During data processing, we employ a Receptive-Field Enhancement strategy that randomly crops image patches to enrich spatial diversity. To support both patch-based high-resolution restoration and full-image SR for smaller inputs, we resize images to a 1024-pixel shorter side with 0.2 probability while maintaining local text consistency with global text.

7. Detail of i2i degradation pipeline

As Sec. 3.3 illustrated, we propose stage-specific degradation pipelines and introduce an image-to-image (i2i) degradation pipeline that utilizes the FLUX model to systematically erase details from high-quality images. Specifically, as shown in Figure 7, using real high-quality images I_{hq}, image prompts P_{img} and negative prompts P_{neg}, we first downsample the image and encode it through VAE encoder and get $\mathbf{z}_{hq}$. Then the $\mathbf{z}_{hq}$ is added Gaussian noise at random range $0.3 \sim 0.5$. Subsequently, iterative inference is performed using the negative prompts to generate a low-resolution image with erased details.

Compared with the Real-ESRGAN degradation, our i2i degradation produces structurally coherent low-quality images with substantially reduced textures, allowing the model to better learn detail generation. To further validate this property, we apply a DWT-based frequency decomposition to the ground-truth images, the i2i-degraded outputs, and the Real-ESRGAN degraded outputs. As shown in Figure 8, the structural information of the i2i-degraded images



(a) Real-ESRGAN deg LR (b) I2i deg LR (c) GT

Figure 8. Comparison of GT images, LR images from different degradation methods, and their high-frequency components map.

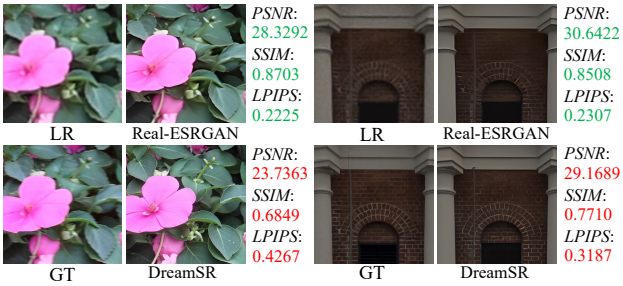


Figure 9. Example of misalignment between reference metrics and perceptual quality.

remain largely consistent with those of the ground truth, whereas Real-ESRGAN introduces strong structural distortion, leading to significant deviations in the high-frequency bands. This demonstrates that our i2i degradation preserves global layout while removing more recoverable textures, ultimately facilitating superior detail generation.

8. Discussion of reference metrics

In Sec. 4.2, we point out that diffusion-based SR methods tend to underperform on reference metrics owing to their inherently strong detail generation capability. Moreover, many widely used real-world SR datasets were collected years ago, and the quality of their ground-truth (GT) images is inherently constrained by the limitations of the acquisition devices of that time. As the generative power of super-resolution models continues to improve, their outputs may

Algorithm 1 Training Scheme of DreamSR

Input: Training dataset $\mathcal{S}\{(I_{hq}, I_{lq}, P_{global})\}$, pretrained Flux DiT model ϵ_θ, Text Encoder $\mathcal{E}_{text}$, VAE Encoder $\mathcal{E}$ and VAE Decoder $\mathcal{D}$, LLaVA model M, Patch Context aware MM-ControlNet $\mathcal{F}_\theta$, Restoration Acceleration LoRA ϕ, training iteration N

Initialize $\mathcal{F}_\theta$ parameterized by ϵ_θ

for $i \leftarrow 1$ **to** N **do**

  /* Data Preprossing */

  Sample I_{hq}, I_{lq}, P_{global} from $\mathcal{S}$

  $r_d \leftarrow \text{random}(0, 1)$

  if $r_d > 0.2$ **then**

    $I_{hq}, I_{lq} \leftarrow \text{RandomCrop}(I_{hq}, I_{lq}, 512)$

    $P_{patch} \leftarrow M(I_{lq})$

  else

    $I_{hq}^{\text{resize}} \leftarrow \text{ResizeShortEdge}(I_{hq}, 1024)$

    $I_{lq}^{\text{resize}} \leftarrow \text{ResizeShortEdge}(I_{lq}, 1024)$

    $I_{hq}, I_{lq} \leftarrow \text{RandomCrop}(I_{hq}^{\text{resize}}, I_{lq}^{\text{resize}}, 1024)$

    $P_{patch} \leftarrow P_{global}$

  $r_t \leftarrow \text{random}(0, 1)$

  if $i < 20000$ **or** $r_t > 0.6$ **then**

    /* ControlNet training */

    Sample t from $(0, 1)$

    Sample $\mathbf{z}_1$ from $\mathcal{N}(0, \boldsymbol{I})$

    $\mathbf{z}_{lq} \leftarrow \mathcal{E}(I_{lq})$

    $E_{patch} \leftarrow \mathcal{E}_{text}(P_{patch})$

    $E_{global} \leftarrow \mathcal{E}_{text}(P_{global})$

    $\mathbf{z}_t \leftarrow (1 - t) \cdot \mathbf{z}_{lq} + t \cdot \mathbf{z}_1$

    $\{f_c^{img}, f_c^{txt}\} \leftarrow \mathcal{F}_\theta(\mathbf{z}_{lq}, \mathbf{z}_t, t, E_{patch})$

    $v_p \leftarrow \epsilon_\theta(\mathbf{z}_t, t, E_{global}, f_c^{img}, f_c^{txt})$

    /* Compute ControlNet loss */

    $v_{gt} \leftarrow \mathbf{z}_1 - \mathbf{z}_0$

    $\mathcal{L}_v \leftarrow ||v_p - v_{gt}||_2^2$

    if $t < 0.2$ **then**

      $\hat{\mathbf{z}}_0 = \mathbf{z}_t - t \cdot v_p$

      $\mathcal{L}_p \leftarrow ||\mathcal{D}(\hat{\mathbf{z}}_0) - I_{gt}||_2^2 + \lambda LPIPS(\mathcal{D}(\hat{\mathbf{z}}_0), I_{gt})$

    $\mathcal{L} \leftarrow \mathcal{L}_v + \mathcal{L}_p$

    Update $\mathcal{F}_\theta$ with $\mathcal{L}$

  else

    /* LoRA training */

    $I_{lq} \leftarrow RealESGRANDeg(I_{hq})$

    $\epsilon_\phi \leftarrow \phi, \epsilon_\theta$

    $\mathbf{z}_{lq} \leftarrow \mathcal{E}(I_{lq})$

    $E_{patch} \leftarrow \mathcal{E}_{text}(P_{patch})$

    $E_{global} \leftarrow \mathcal{E}_{text}(P_{global})$

    $\{f_c^{img}, f_c^{txt}\} \leftarrow \mathcal{F}_\theta(\mathbf{z}_{lq}, \mathbf{z}_t, t_{fix}, E_{patch})$

    $v_p \leftarrow \epsilon_\phi(\mathbf{z}_t, t_{fix}, E_{global}, f_c^{img}, f_c^{txt})$

    $\hat{\mathbf{z}}_{dr} \leftarrow \mathbf{z}_{lq} - t_{fix} \cdot v_p$

    $\mathcal{L}_{lora} \leftarrow ||\hat{\mathbf{z}}_{dr} - \mathbf{z}_0||_2^2 + ||\mathcal{D}(\hat{\mathbf{z}}_{dr}) - I_{gt}||_2^2 + \lambda_1 LPIPS(\mathcal{D}(\hat{\mathbf{z}}_{dr}), I_{gt}) + \lambda_2 GAN(\mathcal{D}(\hat{\mathbf{z}}_{dr}))$

    Update ϕ with $\mathcal{L}_{lora}$

end

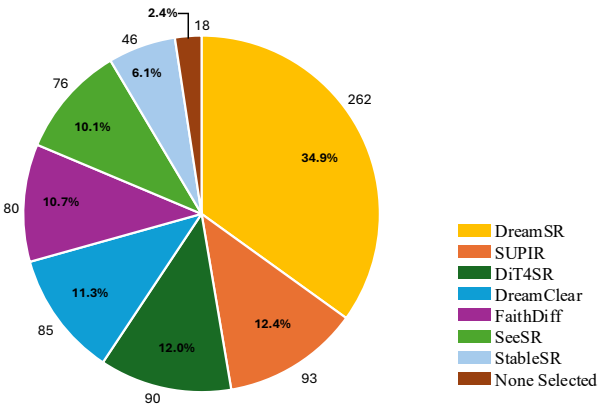


Figure 10. The voting results from 15 volunteers. The percentages and numerical counts are displayed with the pie chart.

even surpass the GT quality, making conventional objective metrics less indicative of true perceptual performance. As shown in Figure 9, Real-ESRGAN often produces overly smooth outputs that lack fine details, yet still obtains relatively high reference metric scores. In contrast, our method generates results with richer textures and better perceptual fidelity, but consequently receives lower scores under these traditional metrics.

9. User Study

To further assess the perceptual effectiveness of our method, we conducted a user study using 50 test images selected from our benchmark, covering a wide range of resolutions and diverse real-world scenarios. We invited 15 volunteers with substantial experience in image quality evaluation. For each LQ input, the corresponding HQ results produced by all competing methods were presented side-by-side. Participants were instructed to select the best result based on two equally weighted criteria: (1) overall perceptual quality and (2) consistency with the input image in terms of structure and texture. If none of the results was deemed satisfactory, the participants were allowed to abstain from making a selection. As shown in Figure 10, our method achieves the highest selection rate among all evaluated approaches, demonstrating its superior perceptual fidelity and stronger alignment with the input image content.

10. More Visual Comparisons

Figs. 11 to 13 presents additional visual comparisons between DreamSR and other diffusion-based methods across multiple resolutions. As shown, DreamSR is able to generate richer fine-grained details while maintaining stronger structural consistency. This advantage becomes even more pronounced in high-resolution scenarios, where our method demonstrates superior generative capability and structural preservation, leading to notably improved visual results.

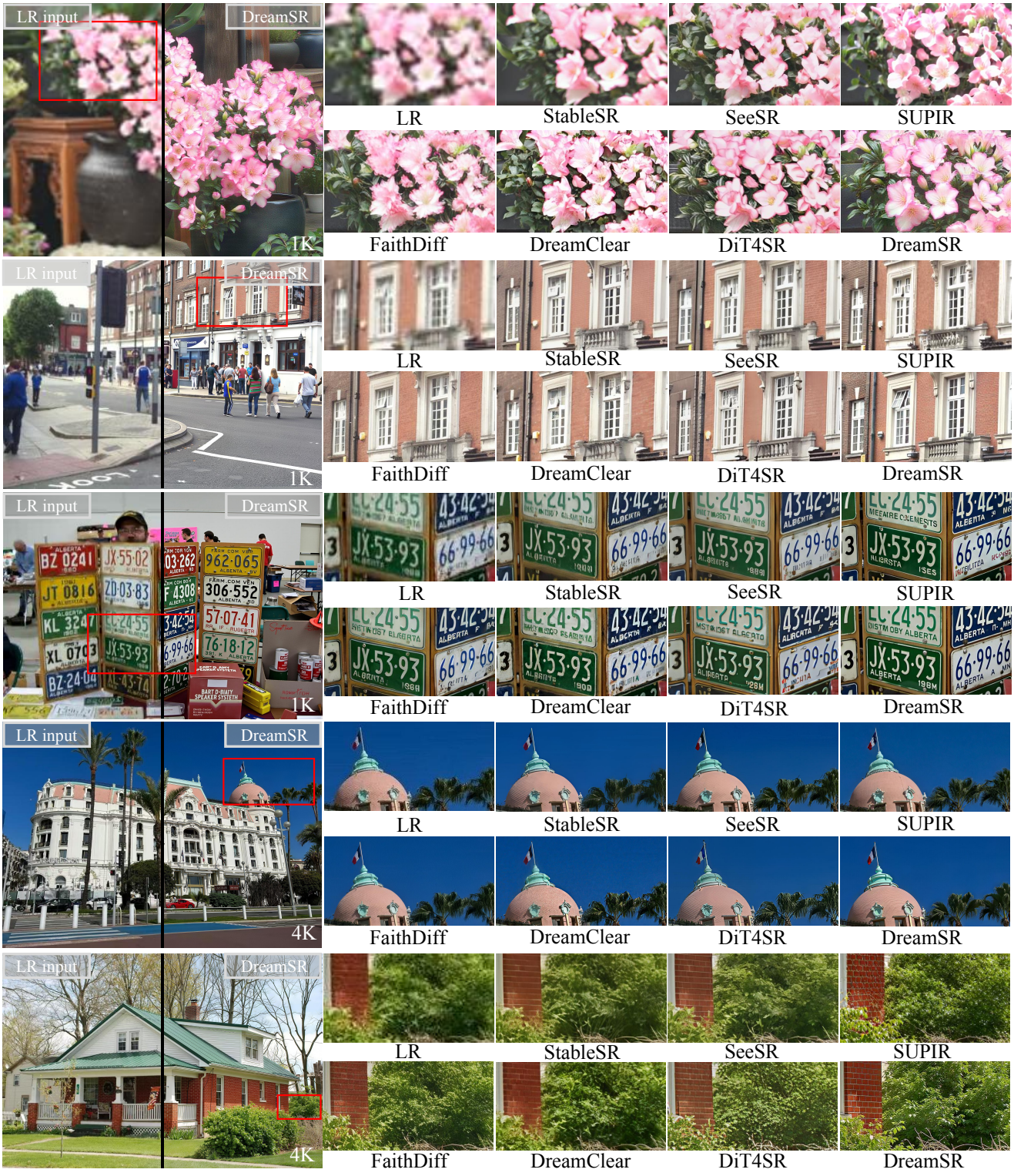


Figure 11. Qualitative comparisons with different methods on real-world datasets. Zoom in for better view.

Figure 12. Qualitative comparisons with different methods on a high resolution real-world image ($1095 \times 720 \rightarrow 4380 \times 2880$). Zoom in for better view.

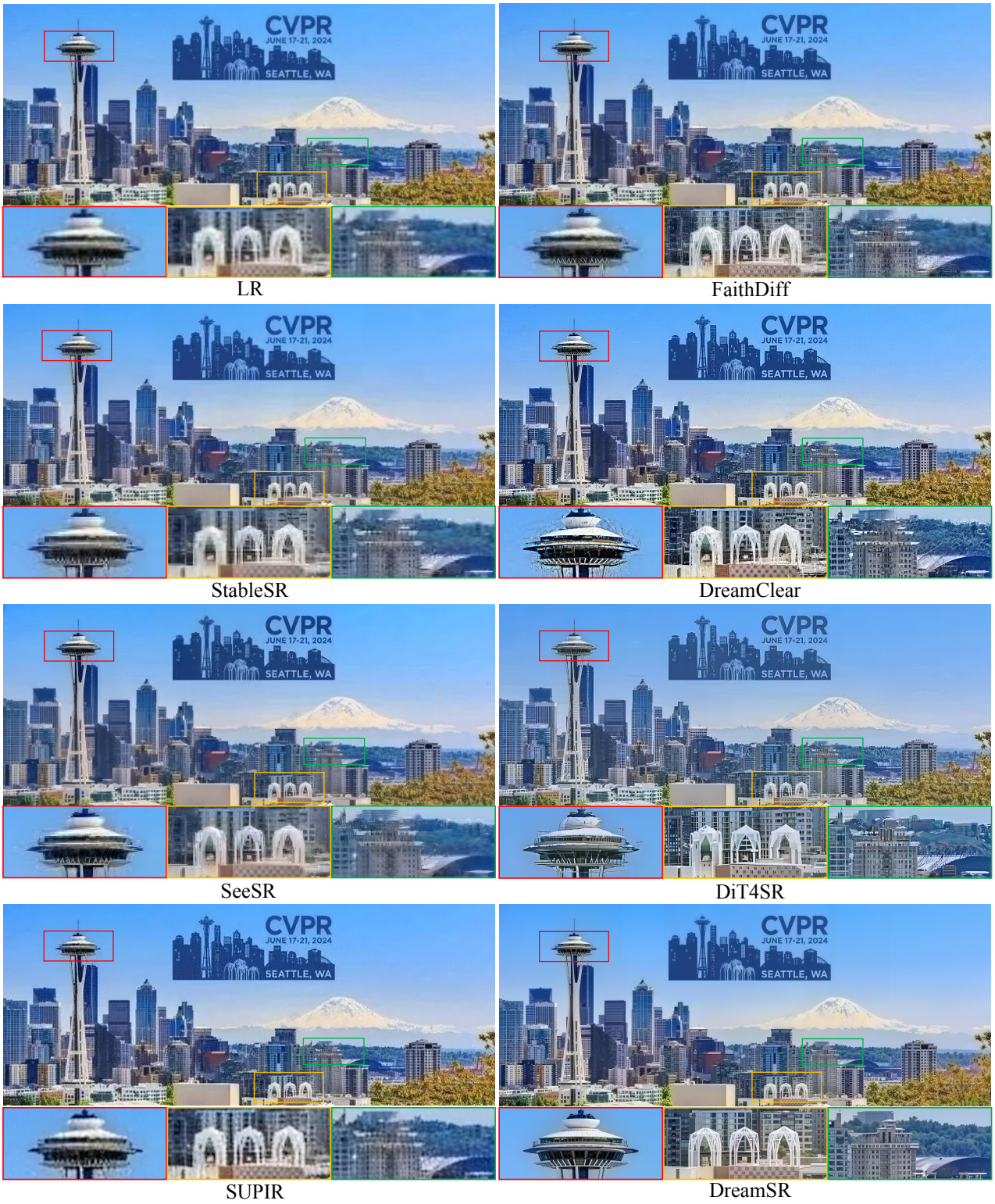


Figure 13. Qualitative comparisons with different methods on a high resolution real-world image ($1759 \times 720 \rightarrow 7036 \times 2880$). Zoom in for better view.