Ultra-High-Definition Reference-Based Landmark Image Super-Resolution with Generative Diffusion Prior

Zhenning Shi[1, 2], Zizheng Yan[2], Yuhang Yu[2], Clara Xue[2], Jingyu Zhuang[2], Qi Zhang[2], Jinwei Chen[2], Tao Li[1, 3*], Qingnan Fan[2*]

College of Computer Science, Nankai University[1]
Vivo Mobile Communication Co. Ltd[2]
Key Laboratory of Data and Intelligent System Security Ministry of Education, China[3]
litao@nankai.edu.cn, fqnchina@gmail.com

Abstract

Reference-based Image Super-Resolution (RefSR) aims to restore a low-resolution (LR) image by utilizing the semantic and texture information from an additional reference high-resolution (reference HR) image. Existing diffusion-based RefSR methods are typically built upon ControlNet, which struggles to effectively align the information between the LR image and the reference HR image. Moreover, current RefSR datasets suffer from limited resolution and poor image quality, resulting in the reference images lacking sufficient fine-grained details to support high-quality restoration. To overcome the limitations above, we propose *TriFlowSR*, a novel framework that explicitly achieves pattern matching between the LR image and the reference HR image. Meanwhile, we introduce *Landmark-4K*, the first RefSR dataset for Ultra-High-Definition (UHD) landmark scenarios. Considering the UHD scenarios with real-world degradation, in TriFlowSR, we design a Reference Matching Strategy to effectively match the LR image with the reference HR image. Experimental results show that our approach can better utilize the semantic and texture information of the reference HR image compared to previous methods. To the best of our knowledge, we propose the first diffusion-based RefSR pipeline for ultra-high definition landmark scenarios under real-world degradation. Our code and model will be available at https://github.com/nkicsl/TriFlowSR.

Introduction

Single Image Super-Resolution (SISR) (Dong et al. 2014; Kim, Lee, and Lee 2016a,b; Lai et al. 2017) aims to reconstruct high-resolution (HR) details from a given low-resolution (LR) image. Recently, diffusion-based models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Song et al. 2020; Liu, Gong, and Liu 2022) have emerged as powerful frameworks in generating high-fidelity data. Due to their strong generative priors, diffusion-based models have been widely adopted in SISR tasks (Wang et al. 2024; Yu et al. 2024; Dong et al. 2025), offering significant improvements in perceptual quality and detail reconstruction. However, due to the inherently ill-posed nature of SISR, the reconstructed images often suffer from visual artifacts and may be inconsistent with the original scene, particularly when faced with heavily degraded images.

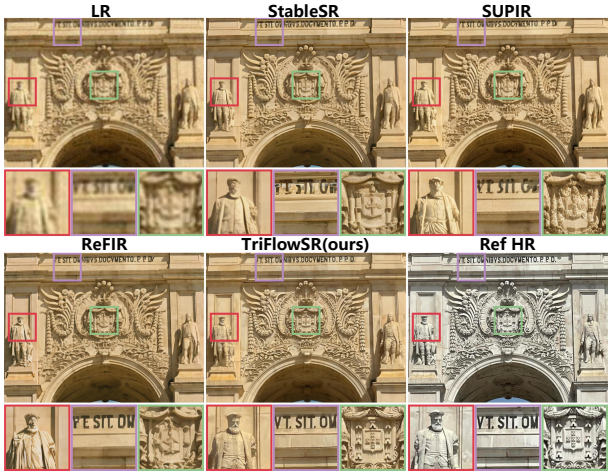

Figure 1: We propose TriFlow to enable diffusion-based RefSR in UHD landmark scenarios under the real-world degradation. Through explicit pattern matching between the LR image and the reference HR image, TriFlow fully leverages the semantic and texture information contained in the reference HR image. Please zoom in for a better view.

To address the limitations of SISR, Reference-based Super-Resolution (RefSR) has been proposed. Unlike SISR, which relies solely on a single LR image, RefSR introduces a reference HR image to leverage additional structural and texture information, thereby generating more realistic and detailed HR outputs. However, most existing RefSR approaches (Yang et al. 2020; Jiang et al. 2021; Cao et al. 2022) are built upon relatively simple CNN architectures and with idealized degradation assumptions (e.g. bicubic downsampling), which significantly limits their performance under real-world degradations.

To resolve this problem, recent studies (Sun et al. 2024; Guo et al. 2024) have explored integrating diffusion-based models into the RefSR pipeline. These methods aim to combine the strong generation capabilities of diffusion models with the contextual guidance provided by reference images. Nevertheless, they introduce the reference information via mechanisms like ControlNet (Zhang, Rao, and Agrawala 2023), without performing explicit pattern matching be-

*they are corresponding authors

tween the LR input and the reference HR image. As a result, they still suffer from generative artifacts. In addition, existing RefSR datasets (Zhang et al. 2019; Jiang et al. 2021) are limited in both resolution and image quality (*e.g.*, CUFED5 has an average resolution of 418 $\times$ 418) , leading to reference HR images that lack sufficient fine-grained detail. This is in conflict with the goal of RefSR: (1) RefSR aims to utilize high-quality reference HR images to enhance the restoration results. If the reference HR image itself is of low resolution and poor quality, the restored image quality will not be high either. (2) Images captured by modern smartphones are predominantly Ultra-High-Definition (UHD) images, and low-resolution datasets do not match the actual scenarios. UHD images, with their richer structural and textural details, are inherently better suited for RefSR, enabling more accurate reconstructions that align with current imaging standards and user expectations. Deploying RefSR methods using UHD references could potentially unlock more accurate reconstructions, better aligning with contemporary imaging standards and user expectations.

To overcome these challenges, we propose the **TriFlowSR** framework. Through the Patch-Ref Attention mechanism, we can explicitly achieve pattern matching between the LR image and the reference HR image, and realize RefSR in real-world degradation scenarios. Meanwhile, we introduce the **Landmark-4K** dataset, which is the first RefSR dataset in the UHD scenario, consisting of 185 high-quality landmark images covering 49 landmark categories worldwide. Furthermore, to enable the RefSR pipeline in UHD scenarios with real-world degradation, we design the **Reference Matching Strategy** to effectively match the LR image with the reference HR image.

Our contributions can be summarized as follows:

- To the best of our knowledge, we propose the first diffusion-based reference-based super-resolution pipeline for ultra-high definition landmark scenarios under real-world degradation.
- We propose the TriFlowSR framework, which can explicitly achieve pattern matching between the LR image and the reference HR image, thereby effectively utilizing the semantic and texture information of the reference HR image.
- We propose the first RefSR dataset for UHD scenarios, Landmark-4K. Additionally, we design a Reference Matching Strategy to effectively match the LR image with the reference HR image.

Related Works

Single Image Super-Resolution.

Single Image Super-Resolution (SISR) aims to recover the HR image with only a single LR image as input. SRCNN (Dong et al. 2014) utilized a deep learning-based convolutional neural network for Super-Resolution tasks. Various CNN-based methods have been proposed (Kim, Lee, and Lee 2016a,b; Lai et al. 2017; Lim et al. 2017; Zhang et al. 2018b) to improve the accuracy of reconstructed images. However, CNN-based methods tend to produce overly smooth results and lack high-frequency details. To generate more perceptually realistic images, GANs (Goodfellow et al. 2020; Ledig et al. 2017; Zhang et al. 2021; Wang et al. 2021) have been introduced into Super-Resolution tasks. While generating more perceptually realistic details, the training of GANs is often unstable and suffers from unnatural visual artifacts. Recently, diffusion-based models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Song et al. 2020; Liu, Gong, and Liu 2022) have emerged as powerful and effective conditional generative models, demonstrating remarkable success in synthesizing high-fidelity data and being widely applied to super-resolution tasks (Wu et al. 2024b,a). StableSR (Wang et al. 2024) fine-tuned a time-aware encoder to incorporate low-resolution (LR) image information. SUPIR (Yu et al. 2024) improved generation by integrating large diffusion backbones with high capacity adapters. TSD-SR (Dong et al. 2025) distilled a multistep SD3 model (Esser et al. 2024) into a single-step model by distilling the target score. Although these methods can effectively generate high-fidelity images, their outputs often exhibit noticeable generative artifacts in scenarios with strong degradation, resulting in hallucinations.

Reference-based Image Super-Resolution.

Reference-based Image Super-Resolution (RefSR) is dedicated to leveraging the semantic and texture information of an additional reference HR image to guide the restoration of the LR image to the HR image. CrossNet (Zheng et al. 2018) aligned the reference and LR images by the flow estimation. Later, some works introduced techniques such as multi-scale (Zhang et al. 2019), cross-scale (Yang et al. 2020), deformable convolution (Shim, Park, and Kweon 2020; Dai et al. 2017) and coarse-to-fine patch matching (Lu et al. 2021). C2-Matching (Jiang et al. 2021) obtained more accurate pre-offsets of reference features to LR features through a teacher-student correlation distillation. Based on C2-Matching, some works have explored multi-scale (Xia et al. 2022), multi-reference (Zhang et al. 2023), and decoupling (Huang et al. 2022). DATSR (Cao et al. 2022) integratesd deformable convolution with the Swin Transformer (Liu et al. 2021). However, most of these methods are simple CNN models, and they assume simple and known degradations (bicubic), which limits their effectiveness when dealing with complex and unknown degradations in the real world. Recent works have attempted to combine RefSR with diffusion-based models, aiming to retain high-fidelity generation capability while effectively utilize the information from the reference HR image. CoSeR (Sun et al. 2024) combined the information from the LR image and the reference HR image by inputting them into two separate ControlNets. ReFIR (Guo et al. 2024) balanced the reference HR features and the super-resolution (SR) image features through Spatial Adaptive Gating. However, these methods introduce information through ControlNet without performing explicit pattern matching between the LR features and the reference HR features, which still results in generative artifacts.

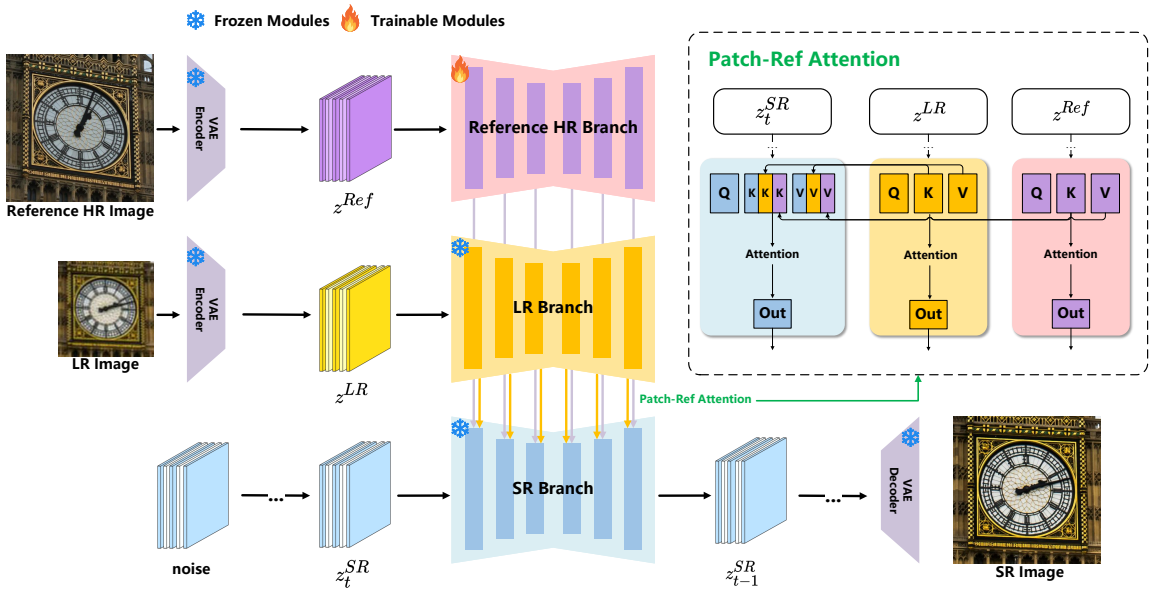

Figure 2: The framework of our proposed TriFlowSR. TriFlowSR comprises three branches: the SR branch, the LR branch, and the Reference HR branch. During the RefSR training process, the SR and LR branches are frozen, while the LR image and the reference HR image process pattern matching via the Patch-Ref Attention. This enables the transfer of semantic and texture information from the reference HR image.

Methodology

TriFlowSR

In the RefSR task, we are given two inputs: the LR image I^{LR} and the reference HR image I^{Ref}. Our goal is to generate the super-resolution image I^{SR} that closely approximates the distribution of the ground truth HR image I^{HR}. Previous diffusion-based RefSR methods introduce the information from the reference HR image into the backbone network through the ControlNet structure, without explicitly performing pattern matching between LR image and the reference HR image. As a result, they struggle to effectively transfer the semantic and texture information from the reference HR image, which will easily result in generative artifacts. To address this issue, we propose the **TriFlowSR** architecture, which consists of three branches: the Super-Resolution (SR) branch, the Low-Resolution (LR) branch, and the Reference High-Resolution (HR) branch. The training is completed in two stages.

The SR branch is a pre-trained text-to-image (T2I) diffusion model (Rombach et al. 2022; Esser et al. 2024), which functions by sampling ϵ from a Gaussian distribution in the latent space to generate high-fidelity images. This branch remains frozen throughout all stages, preserving the diffusion priors and attention mechanisms of the existing pre-trained model. In the first stage, we pre-train the LR branch on the SISR datasets, so that the model has a basic SISR super-resolution capability. The structure of the LR branch is identical to the SR branch. Through the VAE Encoder (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021), the LR image I^{LR} is mapped to $z^{LR} = Encoder(I^{LR})$ in the latent space, which transmits the information of the LR image to the SR branch. Let the parameters of the SR branch and the LR branch be θ and Θ, we have $I^{SISR} = Decoder(\mathcal{F}(\epsilon, z^{LR}; \theta, \Theta))$.

In the second stage, the LR branch and the SR branch are frozen, and training is restricted to the Reference HR branch alone. Similar to the LR branch, the Reference HR branch has the same structure as the SR branch. We have $z^{Ref} = Encoder(I^{Ref})$. To effectively transfer the semantic and textural features of the reference HR image to the backbone network, inspired by Hu (2024), we utilize a cross-attention mechanism to convey the information. We propose the **Patch-Ref Attention**, where the LR image features and the reference HR image features are integrated into the main branch as follows:

$$\begin{aligned} PatchRefAttn = & softmax(\frac{Q^{SR}[K^{SR}, K^{LR}, K^{Ref}]^T}{\sqrt{d_k}}) \\ & \cdot [V^{SR}, V^{LR}, V^{Ref}], \end{aligned} \tag{1}$$

where Q, K and V are the query, key and value from the attention layer of the branches. $[\cdot, \cdot]$ denotes the concatenation operation, and d_k represents the channel dimension. Patch-Ref Attention enables explicit feature matching between the LR image and Reference image at the patch scale. The resulting attention score matrix is then used to selectively retain beneficial reference features while suppressing detrimental features. This interaction enables the model to adaptively select and transfer textures and structural information from the reference HR image rather than relying solely on coarse semantic alignment. Denoting the parameters of the Reference HR branch as ψ, we have $I^{RefSR} = Decoder(\mathcal{F}(\epsilon, z^{LR}, z^{Ref}; \theta, \Theta, \psi))$. Following

the Rectified Flow(Liu, Gong, and Liu 2022), our diffusion loss at the second stage can be represented as Eq.2 where $z_t^{HR} = (1-t)z^{HR} + t\epsilon$.

$$\mathcal{L}_\psi = \int_0^1 \mathbb{E}[||v_\psi(z_t^{HR}, z^{LR}, z^{Ref}, t) - (\epsilon - z^{HR})||^2]dt. \tag{2}$$

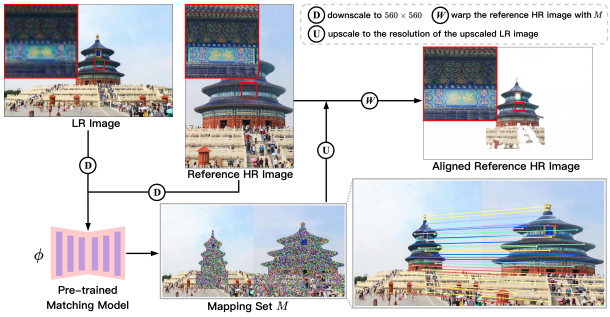

Figure 3: Illustration of the Reference Matching Strategy. We obtain the coarse mapping from the reference HR image to the LR image at a low resolution, and upscale the mapping to the upscaled LR image resolution. Then we warp the reference HR image to align it with the upscaled LR image.

Reference Matching Strategy

For Ultra-High-Definition (UHD) images, directly processing the entire image for super-resolution is computationally expensive. To address this, previous SISR methods (Wang et al. 2024; Wu et al. 2024b) typically adopt a tiling strategy, where the input image is divided into smaller tiles, each processed independently and subsequently stitched together to reconstruct the full-resolution image. However, for RefSR, this tiling strategy encounters unique challenges. Due to the inherent differences in scale, perspective, and content between the reference HR image and the LR input image, the tiles from the reference image usually do not correspond directly to the patches of the LR image. To overcome this issue, we propose the **Reference Matching Strategy** that aligns the reference HR image with the LR image at the pixel level.

Given the bicubic-upscaled LR image I^{LR} and the reference HR image I^{Ref}, we first downscale them to 560×560. We then employ a pre-trained matching model ϕ to establish a coarse correspondence between the reference HR image and the upscaled LR image, where $(u, v) \in I^{LR}_{down}, (x, y) \in I^{Ref}_{down}$:

$$M, C \leftarrow \mathcal{F}(I^{LR}_{down}, I^{Ref}_{down}; \phi), \tag{3}$$

$$M(u, v) = (u, v) \mapsto (x, y), C(u, v) \in [0, 1] \tag{4}$$

Here, M represents the mapping set of points between the downscaled reference HR image and the downscaled LR image, and C represents the certainty of each correspondence. Then, we upscale the M and C to the resolution of upscaled LR image. Based on the upscaled M, we use $F.grid\_sample$ to warp the corresponding reference HR image to I^{Ref}_{warped}. Then we utilize upscaled C to suppress low-confidence textures, as $C \odot I^{Ref}_{warped} + (1-C) \odot mask$, where $\odot$ represents the dot product operation and $mask$ represents a pure white image in Figure 3. As shown in Figure 3, by leveraging effective feature matching and alignment techniques, our strategy enables the reference image to provide more accurate and spatially consistent texture information.

Figure 4: Illustration of the RefSR datasets. (a) represents the CUFED5 dataset (Zhang et al. 2019), (b) represents the WR-SR dataset (Jiang et al. 2021), and (c) represents our proposed Landmark-4K dataset. We apply a center crop to these images, extracting the central region that accounts for 20% of the original image area. Notably, the Landmark-4K dataset offers higher resolution and image quality compared to existing RefSR datasets. Please zoom in for a better view.

Datasets	Numbers	Average Resolutions
CUFED5	126	418×418
WR-SR	80	770×770
Landmark-4K (ours)	185	3295×3295

Table 1: A quantitative comparison of the RefSR datasets. The Landmark-4K dataset outperforms the CUFED5 dataset and the WR-SR dataset in terms of both resolution and dataset size.

Landmark-4K Dataset

Reference-based Super-Resolution (RefSR) aims to leverage the semantic structures and fine-grained texture details from the reference HR image to guide the reconstruction of low-resolution images, thereby producing more realistic super-resolved results. However, existing RefSR datasets (Zhang et al. 2019; Jiang et al. 2021) generally suffer from low resolution and poor image quality, resulting in the reference HR images lacking sufficient detail. This limitation restricts the performance of pre-trained text-to-image (T2I) diffusion models to generate high-fidelity images within the RefSR framework. It is evident that Ultra-High-Definition

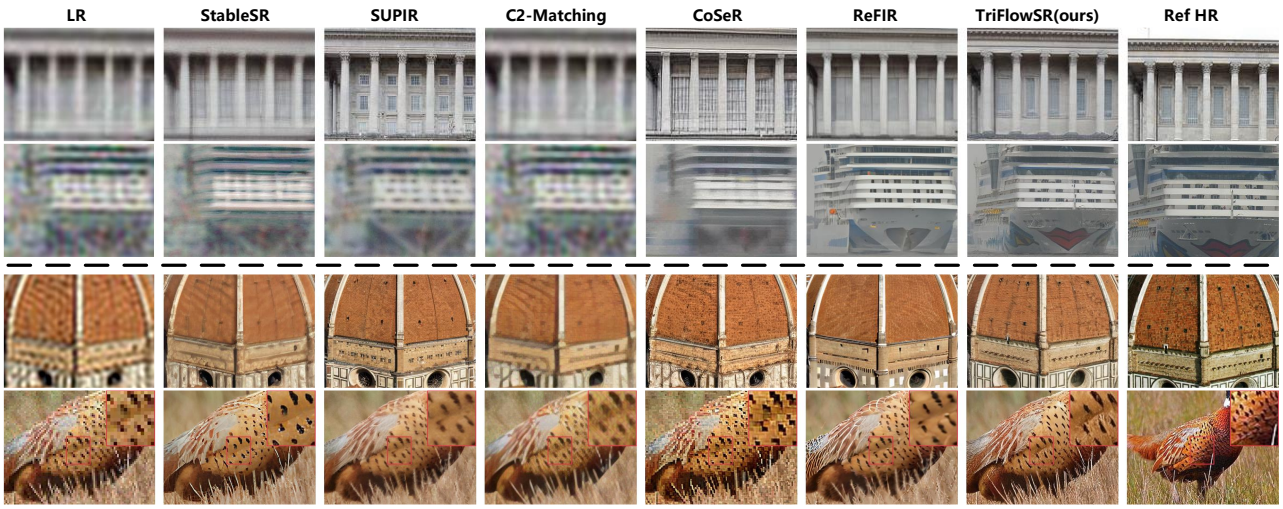

Figure 5: Visual comparisons of the Super-Resolution results by different methods on the CUFED5 dataset and the WR-SR dataset. The top side of the dashed line represents the results on the CUFED 5 dataset, and the bottom side represents the results on the WR-SR dataset.

Methods	CUFED5 (Zhang et al. 2019)					WR-SR (Jiang et al. 2021)				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
StableSR	**20.79**	0.529	0.515	201.29	0.303	23.17	0.597	0.409	88.85	0.246
DiffBIR	18.43	0.361	0.508	206.33	0.246	21.35	0.468	0.451	102.94	0.227
SUPIR	18.97	0.467	0.481	168.26	0.279	21.82	0.517	0.381	64.67	0.197
TSDSR	19.18	0.491	0.327	152.64	0.192	22.04	0.577	0.333	77.71	0.195
C2-Matching	20.77	0.517	0.728	282.43	0.372	23.82	0.613	0.658	142.79	0.318
DATSR	20.75	0.513	0.730	282.19	0.370	**23.83**	**0.615**	0.667	142.71	0.315
CoSeR	19.94	0.503	0.393	158.70	0.220	22.71	0.573	0.430	112.38	0.277
ReFIR	20.32	0.529	0.334	134.62	0.186	20.99	0.532	0.373	65.48	0.195
TriFlowSR	20.21	**0.535**	**0.275**	**114.93**	**0.166**	22.11	0.578	**0.313**	**54.76**	**0.174**

Table 2: Quantitative comparisons with other methods on the CUFED5 dataset and the WR-SR dataset. We report PSNR, SSIM, LPIPS, FID and DISTS. The best and second-best results are highlighted in **bold** and underlined. "↑" (resp. "↓") means the larger (resp. smaller), the better.

(UHD) images are inherently more suitable for the RefSR task, as they provide significantly richer structural and textural information for high-quality reconstruction. Therefore, building a high-quality UHD RefSR dataset is of great importance for advancing the development of RefSR methods and unlocking the full capacity of generative models in high-fidelity image restoration.

To construct a high-quality Ultra-High-Definition (UHD) dataset for Reference-based Image Super-Resolution (RefSR), we first collected approximately 1,000 landmark images from publicly available online sources[1]. We then performed an initial filtering step by removing all images with a resolution lower than 2K. Subsequently, we resized images with a resolution larger than 4096 by scaling them proportionally to a maximum size of 4096 pixels, in order to prevent excessive image sizes while preserving essential details. We further refined the dataset by removing images with noticeable blur or noise. Lastly, we categorized the images according to landmark types and removed images with significant viewpoint variations within each category to ensure effective semantic and texture correspondence between the reference and target images. Following this rigorous data curation pipeline, we propose the Landmark-4K dataset, which consists of 185 high-quality landmark images covering 49 landmark categories from the worldwide. Each landmark category contains 3 to 4 images captured from different viewpoints. Specifically, one image from each category is designated for evaluating the self-reference capability of RefSR models, while the remaining images are used for testing their cross-reference generalization ability. The Landmark-4K dataset is designed to serve as a high-quality, challenging, and practical benchmark for advancing research in the field of Reference-based Image Super-resolution.

[1]pexels.com

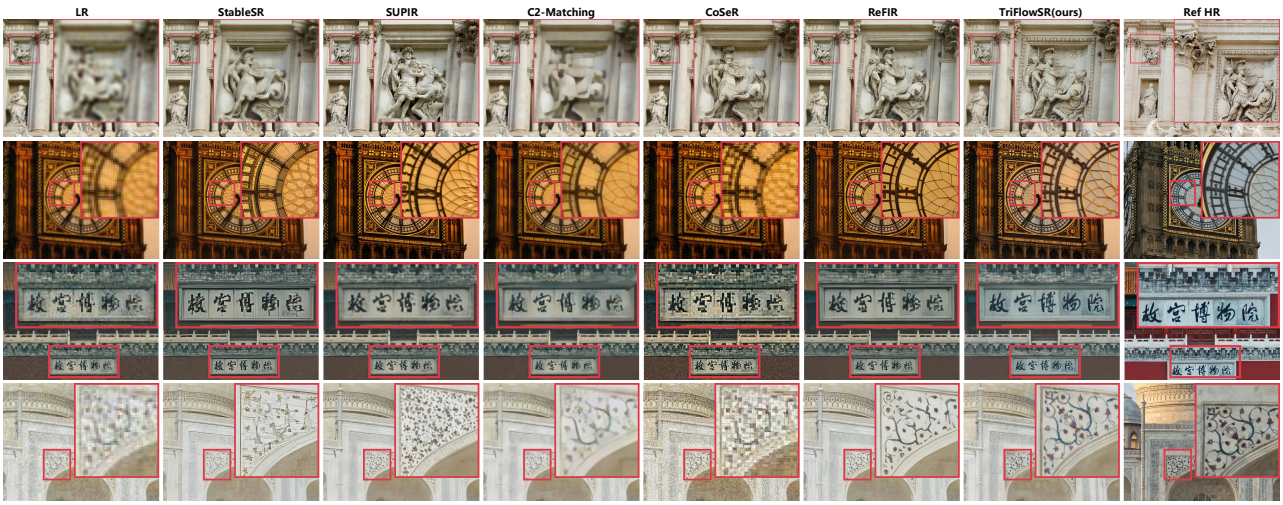

Figure 6: Visual comparisons of the Super-Resolution results by different methods on our proposed Lanmark-4K dataset. Please zoom in for a better view. These images are the results of manual cropping and alignment on the original outputs. In reality, the LR image and the reference HR image may have misalignments in terms of scale, position, and orientation. More visual comparisons are available in the Appendix due to space limitations.

Methods	Landmark-4K			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
StableSR	24.90	0.722	0.308	0.153
DiffBIR	23.99	0.610	0.463	0.208
SUPIR	24.44	0.687	0.312	<u>0.137</u>
TSDSR	23.26	0.692	0.295	0.154
C2-Matching	25.63	0.721	0.578	0.272
DATSR	<u>25.64</u>	0.722	0.597	0.269
CoSeR	24.72	0.697	0.359	0.220
ReFIR	25.21	<u>0.735</u>	<u>0.280</u>	0.149
TriFlowSR	**25.89**	**0.776**	**0.230**	**0.115**

Table 3: Quantitative comparisons with other methods on our proposed Landmark-4K dataset. We report PSNR, SSIM, LPIPS and DISTS. The best and second-best results are highlighted in **bold** and <u>underlined</u>. "↑" (resp. "↓") means the larger (resp. smaller), the better.

Evaluation

Datasets and Baselines.

To evaluate our method, we adopt three benchmarks: CUFED5 (Zhang et al. 2019) testing dataset, WR-SR (Jiang et al. 2021) dataset and our proposed Landmark-4K dataset. CUFED5 testing dataset consists of 126 image pairs, and each input image has 5 reference images with different similarity levels. Following previous methods (Yang et al. 2020; Lu et al. 2021; Xia et al. 2022; Cao et al. 2022), we use the most similar image as the reference HR image. WR-SR dataset consists of 80 images, and each image has one reference image searched through Google Image. The Landmark-4K dataset consists of 185 high-quality landmark images covering 49 landmark categories from the worldwide, with each image having a corresponding high-quality reference HR image from the same landmark scenario. Following previous methods (Guo et al. 2024; Zhang et al. 2024), we use the Real-ESRGAN (Wang et al. 2021) degradation pipeline (same configuration as StableSR (Wang et al. 2024)) with $\times 4$ down-sampling scale to generate the real-world degraded images. To validate the performance of our model, we compare the proposed method with both the popular SISR and RefSR methods. The SISR methods include StableSR (Wang et al. 2024), DiffBIR (Lin et al. 2024), SUPIR (Yu et al. 2024) and TSDSR (Dong et al. 2025). The RefSR methods include C2-Matching (Jiang et al. 2021), DATSR (Cao et al. 2022), CoSeR (Sun et al. 2024) and ReFIR (Guo et al. 2024).

Implementation Details and Evaluation Metrics.

In the experiment, we train the LR branch based on the LSDIR dataset (Li et al. 2023) and the first 10K face images from the FFHQ (Karras, Laine, and Aila 2019) dataset in the first stage, following common Super-Resolution settings (Wang et al. 2024; Dong et al. 2025). In the second stage, we train the Reference HR branch based on the DL3DV (Ling et al. 2024) dataset and the Inter4K (Stergiou and Poppe 2022) dataset. We use Stable Diffusion 3 (Esser et al. 2024) as the pre-trained diffusion backbone for the SR Branch, and kept it frozen throughout all training stages. We use Roma (Edstedt et al. 2024) as the pre-trained matching model ϕ for the Reference Matching Strategy. More training details are provided in the Appendix. To verify the performance of different models, we conduct quantitative evaluations, including both structural metrics, perceptual metrics, and distribution consistency metrics. We use the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) (Wang et al. 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018a) and Deep Image Structure and Texture Similarity (DISTS) (Ding et al. 2020) to assess

the quality of the generated images. We employ the Frechet Inception Distance (FID) (Heusel et al. 2017) to evaluate the realism of the results. Since FID involves downsampling before calculation and is not suitable for UHD images, we did not quantify the FID metric on the Landmark-4K dataset.

Qualitative and Quantitative Results.

As shown in Table 2, our method achieves the best SSIM, LPIPS, FID, and DISTS on the public CUFED dataset. On the public WR-SR dataset, our method obtains the best LPIPS, FID and DISTS. Since the reference HR images in WR-SR dataset are retrieved from Google Image Search, their alignment with the LR images is relatively poor (as illustrated in Figure 5). As a result, our method may yield slightly lower PSNR and SSIM compared to some existing methods. However, our method significantly outperforms them in perceptual quality and distribution consistency metrics, including LPIPS, FID and DISTS. Furthermore, as shown in Table 3, our method achieves the best PSNR, SSIM, LPIPS, and DISTS on our proposed Landmark-4K dataset. As shown in Figure 5 and Figure 6, benefiting from explicitly pattern matching the LR image and the reference HR image, our approach better leverages the semantic and texture information from the reference HR image compared to other methods, resulting in more realistic textures and finer details.

Ablation Study

Effectiveness of the Components

Notably, the LR and Ref features are only concatenated during cross-attention operations, making them completely decoupled. First, we attempt to perform inference with only the LR branch and SR branch, which degenerates our model into a SISR model. Then, we introduce the Reference branch, but only utilize the relative position of the LR image tiles to map to the reference HR image tiles. Next, we use the pre-trained matching model to obtain the matching map and simply resize the corresponding reference region to the input tile size without using warping to align the images. Finally, we apply the warping operation to align the LR image with the reference HR image. As shown in Table 4, compared to pure SISR, introducing the Reference branch can significantly improve PSNR, SSIM, LPIPS, and DISTS. Introducing the Reference Matching Strategy can solve the problem of incorrect tile retrieval due to scale mismatch and positional differences between the LR image and the reference HR image. Visual comparisons are provided in the Appendix due to space constraints.

Control Scale of the Reference branch

Similar to ControlNet (Zhang, Rao, and Agrawala 2023) and IP-adapter (Ye et al. 2023), TriFlow can also control the influence of the reference branch by adjusting the weight of the attention map. Specifically, during the Pat-Ref Attention operation, we can control the attention weight corresponding to the reference branch feature by multiplying K^{Ref} by a coefficient $kscale$, i.e., $K^{Ref} \mapsto kscale \times K^{Ref}$. Since we only performed concatenation on the K and V features, directly controlling the weight of K^{ref} can control the corresponding attention weight $Q[kscale \times K^{ref}]^T$ of V^{ref}, thereby achieving the result of controlling the influence of the reference branch. As shown in Table 5, when $kscale = 1$, ours is a RefSR model. When the value of $kscale$ is between $0 - 1$, we can consider it as a trade-off between SISR and RefSR. As $kscale$ gradually decreases to zero, ours gradually degenerates into a SISR model. More perspective comparison results are detailed in the Appendix due to space constraints.

RB	M	W	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
			24.07	0.664	0.355	0.166
✓			<u>25.67</u>	<u>0.756</u>	0.241	<u>0.113</u>
✓	✓		25.37	0.750	<u>0.239</u>	**0.110**
✓	✓	✓	**25.89**	**0.776**	**0.230**	0.115

Table 4: Ablation study on our proposed Landmark-4K dataset. We report PSNR, SSIM, LPIPS and DISTS. RB represents the Reference branch, M represents the Matching operation and W represents the Warping operation. The best and second-best results are highlighted in **bold** and <u>underlined</u>. "↑" (resp. "↓") means the larger (resp. smaller), the better.

$kscale$	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
0	24.07	0.664	0.355	0.166
0.2	24.36	0.712	0.299	0.181
0.4	24.57	0.709	0.287	0.159
0.6	24.65	0.710	0.285	0.150
0.8	<u>24.79</u>	<u>0.716</u>	<u>0.280</u>	<u>0.145</u>
1	**25.89**	**0.776**	**0.230**	**0.115**

Table 5: Quantitative comparisons about the control scale ($kscale$) of the Reference branch on our proposed Landmark-4K dataset. We report PSNR, SSIM, LPIPS and DISTS. The best and second-best results are highlighted in **bold** and <u>underlined</u>. "↑" (resp. "↓") means the larger (resp. smaller), the better.

Conclusion

In this paper, we address the problem of Ultra-High-Definition Reference-Based Landmark Image Super-Resolution. To this end, we propose TriflowSR, a novel framework that explicitly performs pattern matching between LR and reference HR features. To address the issues of low quality and relatively low resolution in existing RefSR datasets, we propose the first UHD RefSR dataset based on landmark scenes and design a Reference Matching Strategy to match corresponding tiles. We successfully implement a diffusion-based RefSR pipeline in UHD landmark scenarios under the real-world degradation. Experimental results show that our approach can better utilize the semantic and texture information of the reference HR image than previous methods, resulting in more realistic texture and details.

References

Cao, J.; Liang, J.; Zhang, K.; Li, Y.; Zhang, Y.; Wang, W.; and Gool, L. V. 2022. Reference-based image super-resolution with deformable attention transformer. In *European conference on computer vision*, 325–342. Springer.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2567–2581.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, 184–199. Springer.

Dong, L.; Fan, Q.; Guo, Y.; Wang, Z.; Zhang, Q.; Chen, J.; Luo, Y.; and Zou, C. 2025. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23174–23184.

Edstedt, J.; Sun, Q.; Bökman, G.; Wadenbäck, M.; and Felsberg, M. 2024. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19790–19800.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Guo, H.; Dai, T.; Ouyang, Z.; Zhang, T.; Zha, Y.; Chen, B.; and Xia, S.-t. 2024. Refir: Grounding large restoration models with retrieval augmentation. *Advances in Neural Information Processing Systems*, 37: 46593–46621.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Huang, Y.; Zhang, X.; Fu, Y.; Chen, S.; Zhang, Y.; Wang, Y.-F.; and He, D. 2022. Task decoupled framework for reference-based super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5931–5940.

Jiang, Y.; Chan, K. C.; Wang, X.; Loy, C. C.; and Liu, Z. 2021. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2103–2112.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016a. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646–1654.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016b. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1637–1645.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 624–632.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Li, Y.; Zhang, K.; Liang, J.; Cao, J.; Liu, C.; Gong, R.; Zhang, Y.; Tang, H.; Liu, Y.; Demandolx, D.; et al. 2023. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1775–1787.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136–144.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Qiao, Y.; Ouyang, W.; and Dong, C. 2024. Diffbir: Toward blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, 430–448. Springer.

Ling, L.; Sheng, Y.; Tu, Z.; Zhao, W.; Xin, C.; Wan, K.; Yu, L.; Guo, Q.; Yu, Z.; Lu, Y.; et al. 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22160–22169.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lu, L.; Li, W.; Tao, X.; Lu, J.; and Jia, J. 2021. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6368–6377.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4296–4304.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shim, G.; Park, J.; and Kweon, I. S. 2020. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8425–8434.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Stergiou, A.; and Poppe, R. 2022. Adapool: Exponential adaptive pooling for information-retaining downsampling. *IEEE Transactions on Image Processing*, 32: 251–266.

Sun, H.; Li, W.; Liu, J.; Chen, H.; Pei, R.; Zou, X.; Yan, Y.; and Yang, Y. 2024. Coser: Bridging image and language for cognitive super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25868–25878.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2024. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12): 5929–5949.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1905–1914.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, R.; Sun, L.; Ma, Z.; and Zhang, L. 2024a. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553.

Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; and Zhang, L. 2024b. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25456–25467.

Xia, B.; Tian, Y.; Hang, Y.; Yang, W.; Liao, Q.; and Zhou, J. 2022. Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2768–2776.

Yang, F.; Yang, H.; Fu, J.; Lu, H.; and Guo, B. 2020. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5791–5800.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, F.; Gu, J.; Li, Z.; Hu, J.; Kong, X.; Wang, X.; He, J.; Qiao, Y.; and Dong, C. 2024. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25669–25680.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4791–4800.

Zhang, L.; Li, X.; He, D.; Li, F.; Ding, E.; and Zhang, Z. 2023. LMR: a large-scale multi-reference dataset for reference-based super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13118–13127.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, 286–301.

Zhang, Y.; Yang, Q.; Chandler, D. M.; and Mou, X. 2024. Reference-Based Multi-Stage Progressive Restoration for Multi-Degraded Images. *IEEE Transactions on Image Processing*.

Zhang, Z.; Wang, Z.; Lin, Z.; and Qi, H. 2019. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7982–7991.

Zheng, H.; Ji, M.; Wang, H.; Liu, Y.; and Fang, L. 2018. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European conference on computer vision (ECCV)*, 88–104.

Appendix

Detail Experimental Setting

In the experiment, we train the LR branch based on the LSDIR dataset (Li et al. 2023) and the first 10K face images from the FFHQ (Karras, Laine, and Aila 2019) dataset in the first stage, following common Super-Resolution settings (Wang et al. 2024; Dong et al. 2025). The I_{LR} is upsampled to the desired size using bicubic interpolation during the pre-processing stage, following common practice. In the second stage, we train the Reference HR branch based on the DL3DV (Ling et al. 2024) dataset and the Inter4K (Stergiou and Poppe 2022) dataset. Both stages of training are conducted on 4 NVIDIA H20 GPUs. We utilized the AdamW optimizer with a learning rate set to $5e^{-5}$. We use Stable Diffusion 3 (Esser et al. 2024) as the pre-trained diffusion backbone for the SR Branch, and kept it frozen throughout all training stages. We use Roma (Edstedt et al. 2024) as the pre-trained matching model ϕ for the Reference Matching Strategy. In practical implementation, since the mask can be any available image, we will use the result of the SISR (Single Image Super-Resolution) model as the actual mask used. During the two-stage training phase, we employed mixed training with resolutions of 1024×1024 and 512×512 to enhance the model's ability to handle images of varying resolutions. Following previous methods (Yang et al. 2020; Lu et al. 2021; Xia et al. 2022; Cao et al. 2022), we used the second-order degradation model from Real-ESRGAN (Wang et al. 2021) (with the same configuration as StableSR (Wang et al. 2024)) with a $\times 4$ down-sampling scale to generate real-world degraded images. Simultaneously, we applied data augmentation techniques such as random flipping, random cropping, ColorJitter, and Homography transformation to improve the model's robustness. During the inference phase, for the CUFED5 dataset, we padded both the LR image and the reference HR image to a size of 512×512 to achieve alignment. For the WR-SR dataset, we padded the images to a square shape with the maximum width and height to align them. For the Landmark-4K dataset, we employed a sliding window approach with a tile size of 1024 and a tile step of 256 for tile-based inference, and used the c2-blending strategy to stitch the results back to the original size. We used BF16 precision during training and inference. For detailed hyperparameter configurations, please refer to Table 6.

	Hyperparameter
Batch size	16
Learning rate	$5e^{-5}$
Warp-up steps	100
Training steps	$60k$
Max grad norm	1.0
Precision	BF16

Table 6: Experimental settings for our TriFlow during the training stage.

Effectiveness of the Components

RB	M	W	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
			24.07	0.664	0.355	0.166
✓			<u>25.67</u>	<u>0.756</u>	0.241	<u>0.113</u>
✓	✓		25.37	0.750	<u>0.239</u>	**0.110**
✓	✓	✓	**25.89**	**0.776**	**0.230**	0.115

Table 7: Ablation study on our proposed Landmark-4K dataset. We report PSNR, SSIM, LPIPS and DISTS. RB represents the Reference branch, M represents the Matching operation and W represents the Warping operation. The best and second-best results are highlighted in **bold** and <u>underlined</u>. "↑" (resp. "↓") means the larger (resp. smaller), the better.

Notably, the LR and Ref features are only concatenated during cross-attention operations, making them completely decoupled. First, we attempt to perform inference with only the LR branch and SR branch, which degenerates our model into a SISR model. Then, we introduce the Reference branch, but only utilize the relative position of the LR image tiles to map to the reference HR image tiles. Next, we use the pre-trained matching model to obtain the matching map and simply resize the corresponding reference region to the input tile size without using warping to align the images. Finally, we apply the warping operation to align the LR image with the reference HR image. As shown in Table 4, compared to pure SISR, introducing the Reference branch can significantly improve PSNR, SSIM, LPIPS, and DISTS. Introducing the Reference Matching Strategy can solve the problem of incorrect tile retrieval due to scale mismatch and positional differences between the LR image and the reference HR image. As shown in Figure 7, without the Reference branch, the model is unable to utilize any information from the reference HR image, and the result is generated solely based on the diffusion prior. Due to the scale mismatch and positional differences between the LR image and the reference HR image, without using the Reference Matching Strategy, the LR image tile is likely to fail to correspond to the correct reference HR tile, resulting in incorrect reference information. When the Warping operation is not used, the LR image and the reference HR image may fail to align effectively, resulting in difficulty in matching some detailed information (such as fingers and face).

Control of the Reference branch

Similar to ControlNet (Zhang, Rao, and Agrawala 2023) and IP-adapter (Ye et al. 2023), TriFlow can also control the influence of the reference branch by adjusting the weight of the attention map. Specifically, during the Pat-Ref Attention operation, we can control the attention weight corresponding to the reference branch feature by multiplying K^{Ref} by a coefficient $kscale$, i.e., $K^{Ref} \mapsto kscale \times K^{Ref}$. Since we only performed concatenation on the K and V features, directly controlling the weight of K^{ref} can control the corresponding attention weight $Q[kscale \times K^{ref}]^T$ of V^{ref}, thereby achieving the result of controlling the in-

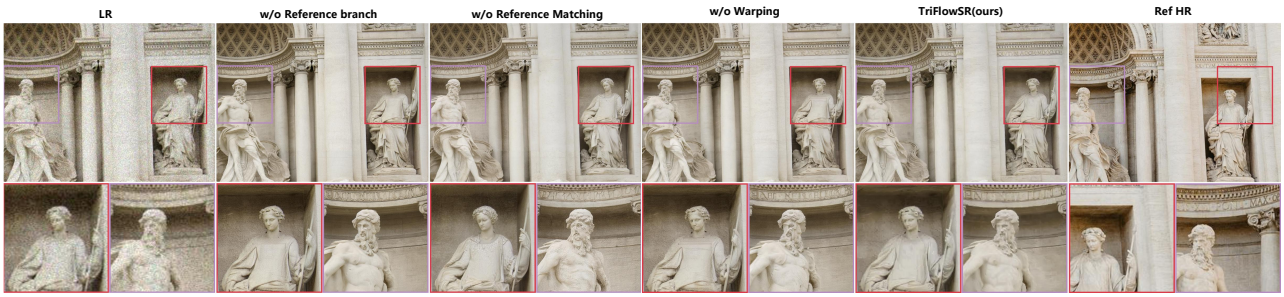

Figure 7: Visual comparisons of the Super-Resolution results of the ablation experiments on our proposed Lanmark-4K dataset. Without the Reference branch, ours will degenerate into a SISR model. If the Reference Matching Strategy & Warping is not used, the LR image and the reference HR image may fail to align effectively, resulting in difficulty in matching some detailed information (such as fingers and face). Please zoom in for a better view.

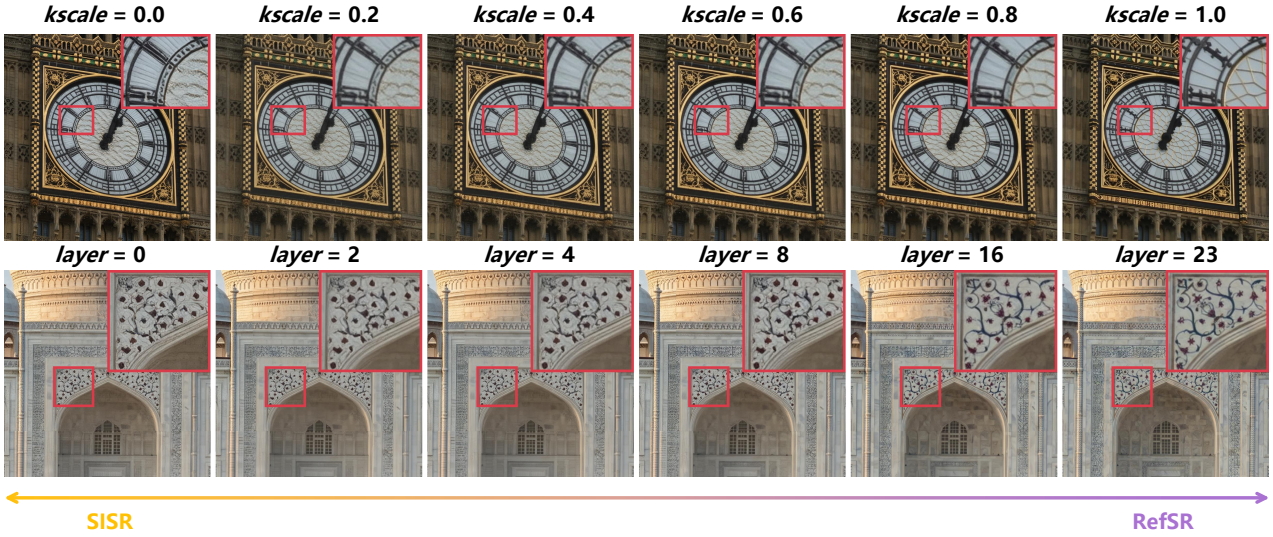

Figure 8: Visual comparisons of the Super-Resolution results of the ablation experiments on our proposed Lanmark-4K dataset. Without the Reference branch, ours will degenerate into a SISR model. If the Reference Matching Strategy & Warping is not used, the LR image and the reference HR image may fail to align effectively, resulting in difficulty in matching some detailed information (such as facial information). Please zoom in for a better view.

fluence of the reference branch. As shown in Table 5, when $kscale = 1$, ours is a RefSR model, while when $kscale$ is between $0 - 1$, we can consider it as a trade-off between SISR and RefSR. As $kscale$ gradually decreases to zero, ours gradually degenerates into a SISR model. Moreover, since the LR branch and the Reference branch are completely decoupled, the degree of control over the reference HR image can also be effectively controlled by adjusting the number of layers in the Reference branch. As shown in Table 9, when the number of layers decreases (these models are all retrained in stage 2), our model gradually degenerates from a RefSR model to a SISR model, which is consistent with the principle of $kscale$. As shown in Figure 8, we can control the degree of supervision from the reference HR image by adjusting the weight of the attention map (k_{scale}) and the number of layers in the Reference branch module, allowing our model to transition seamlessly between the SISR model and the RefSR model.

Reference Image Condition Modeling

In this chapter, we explored the current mainstream methods for introducing additional supervisory information into diffusion-based models, and compared them with our proposed Patch-Ref Attention mechanism.

Controlnet (Zhang, Rao, and Agrawala 2023) is an extension of diffusion models that enables pixel-level control over image generation using additional structural inputs, such as edge maps, depth maps, or human poses (for the RefSR task, it is the reference HR image). ControlNet introduces additional information by adding an auxiliary branch network that mirrors the structure of the backbone. A zero convolution module is used to integrate the noisy input with the conditional input within this branch. The output of the auxiliary branch is then added to the main network to guide the diffusion process. However, this approach lacks explicit pattern matching. The direct addition operation only allows for

$kscale$	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
0	24.07	0.664	0.355	0.166
0.2	24.36	0.712	0.299	0.181
0.4	24.57	0.709	0.287	0.159
0.6	24.65	0.710	0.285	0.150
0.8	<u>24.79</u>	<u>0.716</u>	<u>0.280</u>	<u>0.145</u>
1	**25.89**	**0.776**	**0.230**	**0.115**

Table 8: Quantitative comparisons about the control scale ($kscale$) of the Reference branch on our proposed Landmark-4K dataset. We report PSNR, SSIM, LPIPS and DISTS. The best and second-best results are highlighted in **bold** and <u>underlined</u>. "↑" (resp. "↓") means the larger (resp. smaller), the better.

$layers$	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
0	24.07	0.664	0.355	0.166
2	24.74	0.723	0.28	0.132
4	24.75	0.720	0.283	0.136
8	24.79	0.724	0.273	0.132
16	<u>25.38</u>	<u>0.747</u>	<u>0.252</u>	<u>0.126</u>
23	**25.89**	**0.776**	**0.230**	**0.115**

Table 9: Quantitative comparisons about the $layers$ of the Reference branch on our proposed Landmark-4K dataset. We report PSNR, SSIM, LPIPS, FID and DISTS. The best and second-best results are highlighted in **bold** and <u>underlined</u>. "↑" (resp. "↓") means the larger (resp. smaller), the better.

a coarse transfer of semantic information, making it difficult to effectively capture the structural and textural details of the reference high-resolution image. As shown in Table 10, we retrain the Reference branch based on the ControlNet structure using exactly the same settings and architecture. Experimental results demonstrate that our proposed TriFlowSR enables explicit feature matching between the LR image and the reference HR image, thereby making more effective use of the semantic and textural information in the reference HR image.

$Method$	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
Controlnet	24.55	0.717	0.271	0.121
TriFlowSR(ours)	**25.89**	**0.776**	**0.230**	**0.115**

Table 10: Quantitative comparisons about the reference image condition modeling of the Reference branch on our proposed Landmark-4K dataset. We report PSNR, SSIM, LPIPS and DISTS. The best results are highlighted in **bold**. "↑" (resp. "↓") means the larger (resp. smaller), the better.

IP-Adapter (Ye et al. 2023) is a lightweight module designed to enable image-based prompting for diffusion models. It connects a pre-trained image encoder (e.g., CLIP) to the diffusion model, allowing users to guide image generation using reference images instead of text. By injecting features from the reference image into the cross-attention layers, IP-Adapter provides effective visual control while keeping the base model frozen. However, the introduced reference information is compressed by an image encoder, which typically retains only semantic and partial structural information. This injection method severely compromises the texture information of the reference HR image, leading to strong generative artifacts.

T2I-Adapter (Mou et al. 2024) is a lightweight module designed to provide structural guidance (e.g., edges, depth, pose) to pre-trained text-to-image diffusion models. It works by learning an external adapter network that extracts features from the control input and injects them into the diffusion model without modifying its original weights. This approach enables controllable image generation with minimal computational cost and strong compatibility with existing models. The main difference between T2I-Adapter and IP-adapter is that the embedding information is introduced through addition (similar to ControlNet) rather than cross-attention. Like IP-adapter, T2I-Adapter also faces the issue of compressed reference information.

Resource efficiency

We compare the floating point of operations (FLOPs) and inference time with other methods on our proposed Lanmark-4K dataset by $torch.profiler$, using 1024×1024 images as the input. All experiments were conducted on a single NVIDIA H20. As shown in Table 11, compared to other methods, our TriFlowSR has the lowest TFLOPs and the shortest inference time.

Methods	PSNR↑	LPIPS↓	TFLOPs↓	time (s)↓
SUPIR	24.44	0.312	<u>1287.10</u>	<u>17.25</u>
CoSeR	24.72	0.359	-	53.27
ReFIR	<u>25.21</u>	<u>0.28</u>	1561.51	26.07
TriFlowSR(ours)	**25.89**	**0.23**	**491.65**	**10.59**

Table 11: Quantitative comparisons about resource efficiency on our proposed Landmark-4K dataset. We report PSNR, LPIPS, TFLOPs and inference time. The best and second-best results are highlighted in **bold** and <u>underlined</u>. "↑" (resp. "↓") means the larger (resp. smaller), the better.

More Visual Results

As shown in the Figure 9, Figure 10 and Figure 11, we provide more visual comparison results. Meanwhile, we also provide more illustrations about our proposed Landmark-4K dataset, as shown in the Figure 12.

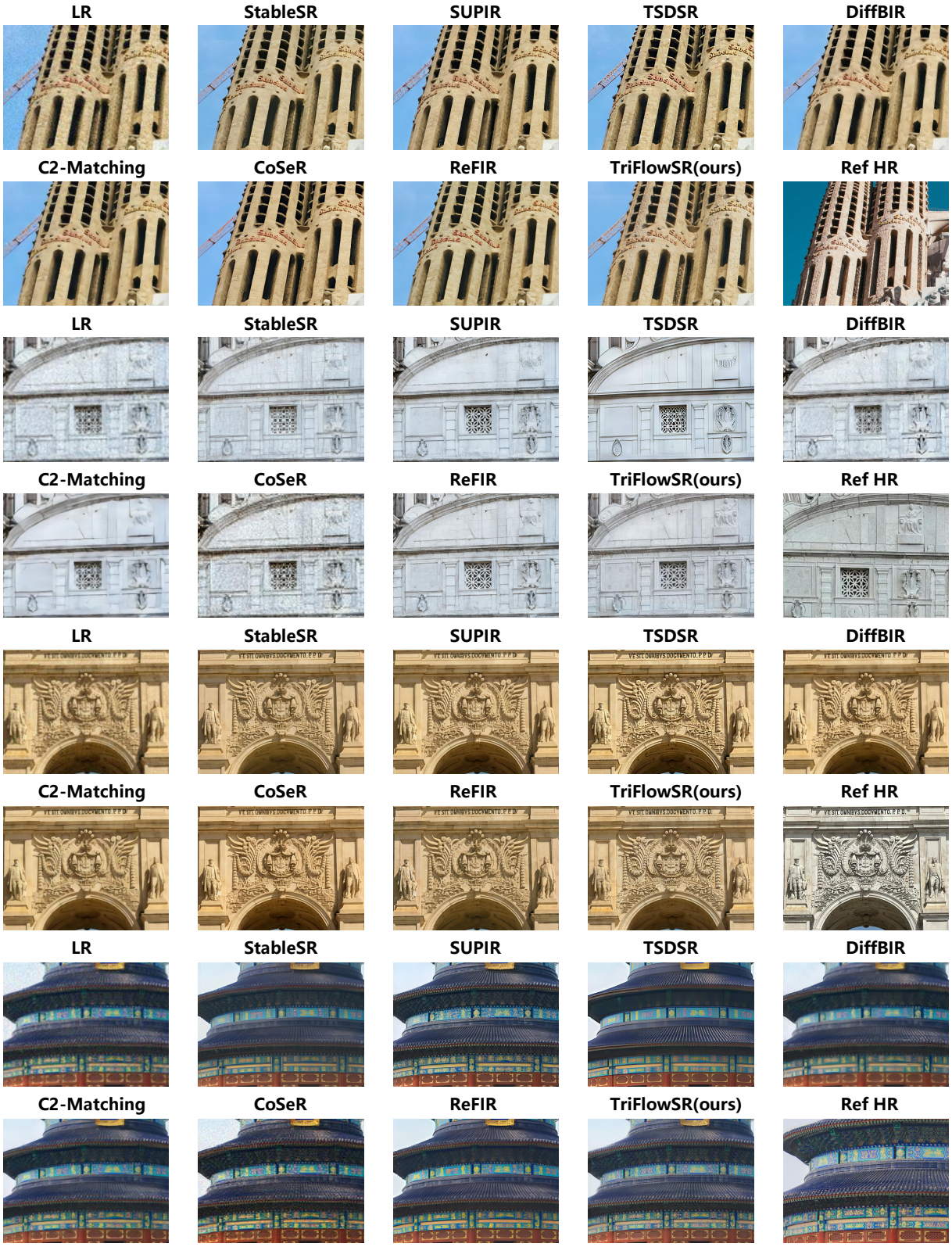

Figure 9: More visual comparisons of the Super-Resolution results by different methods on the our proposed Lanmark-4K dataset. Please zoom in for a better view.

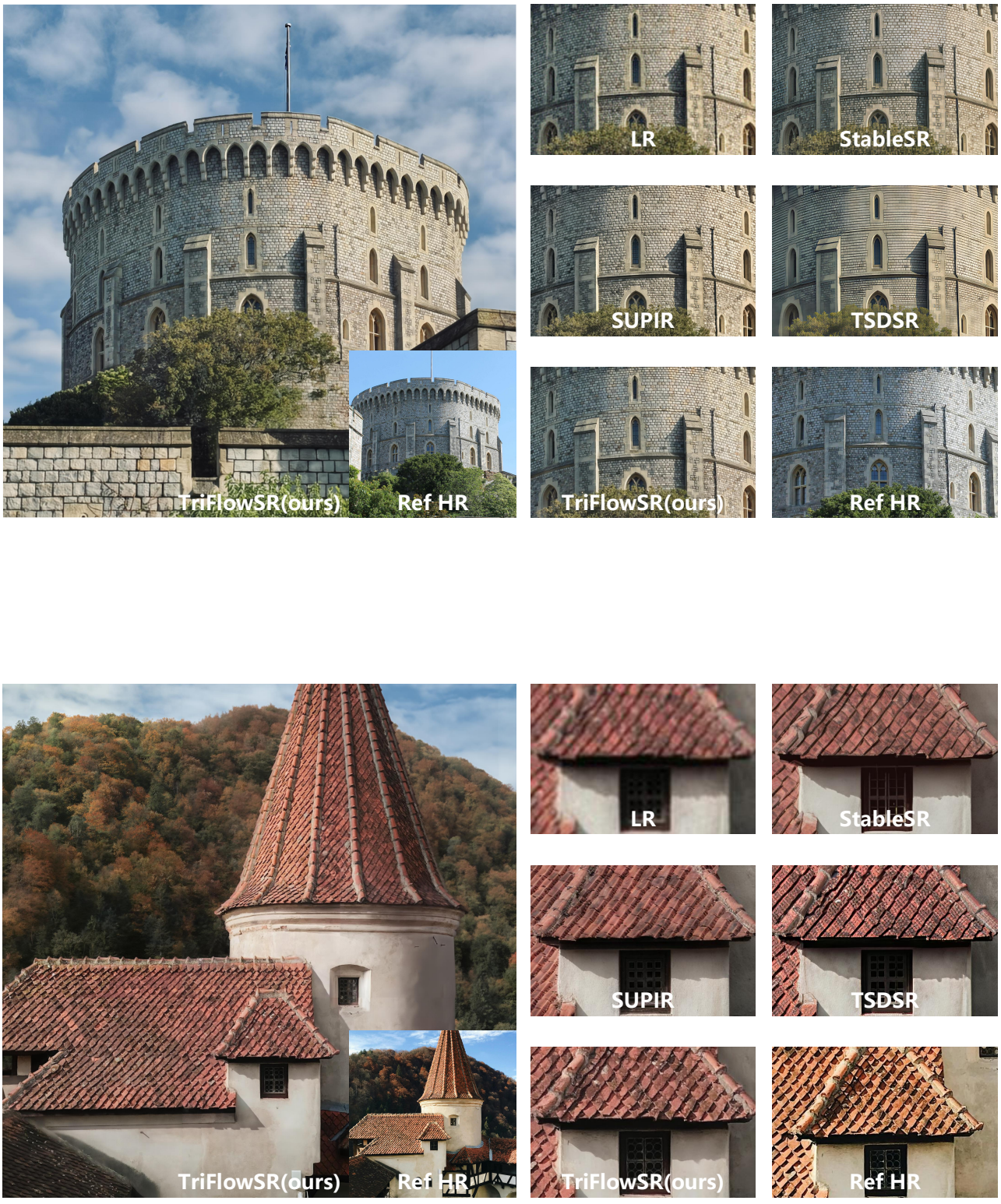

Figure 10: More visual comparisons of the Super-Resolution results by different methods on the our proposed Lanmark-4K dataset. Please zoom in for a better view.

Figure 11: More visual comparisons of the Super-Resolution results by different methods on the our proposed Lanmark-4K dataset. Please zoom in for a better view.

Figure 12: More visual demonstrations on our proposed Landmark-4K dataset. Please zoom in for a better view.