Learning to Balance: Decoupled Siamese Diffusion Transformer for Reference-Based Remote Sensing Image Super-Resolution

Bin Luo[1], Runmin Dong[2⋆], Zhaoyang Luo[1], Jinxiao Zhang[4], Jiyao Zhao[3], Fan Wei[4], and Haohuan Fu[1,3,4⋆]

[1] Tsinghua Shenzhen International Graduate School, Shenzhen, China
[2] Sun Yat-sen University, Zhuhai, China
[3] National Supercomputing Center in Shenzhen, Shenzhen, China
[4] Tsinghua University, Beijing, China
luob21@mails.tsinghua.edu.cn, dongrm3@mail.sysu.edu.cn

Abstract. Diffusion-based methods demonstrate significant potential for remote sensing image super-resolution at large scaling factors, particularly in reference-based super-resolution (RefSR), where high-resolution reference images provide critical fine-grained texture priors. However, existing methods often suffer from a trade-off between over-reliance on reference information, which leads to texture artifacts, and under-utilization of such information, which results in insufficient detail recovery. To address these issues, we propose DS-DiT, a Decoupled Siamese Diffusion Transformer that decouples the interaction between low-resolution (LR) and reference (Ref) conditions within the attention mechanism. By allowing LR structural priors and Ref texture information to independently interact with the noisy latent, the framework effectively mitigates competition between the two conditional sources. To further compensate for the limited local modeling ability of global attention, we introduce a Patch-Level Weighting (PLW) module that adaptively modulates the fusion of conditional sources. In addition, the siamese architecture enables an inference-time autoguidance strategy that exploits the prediction discrepancy between strong and weak Ref conditions to improve generation quality without additional training. Experimental results across multiple datasets and scaling factors show that DS-DiT outperforms existing methods in both quantitative metrics and visual fidelity. The source code is available at https://github.com/B1nary-L/DS-DiT.

Keywords: Image Super-Resolution · Diffusion Model · Remote Sensing

1 Introduction

High-resolution (HR) remote sensing (RS) imagery is indispensable for a wide range of applications, including urban planning, environmental monitoring, and

⋆ Corresponding authors.

disaster assessment [4, 13]. However, due to inherent constraints in sensor hardware and imaging conditions, acquired RS images often suffer from a trade-off between spatial and temporal resolutions [22], frequently failing to meet the requirements of fine-grained analysis. Image super-resolution (SR) techniques have emerged as an effective solution to bridge this gap by reconstructing HR images from easily accessible low-resolution (LR) observations with short revisit cycles, thereby effectively filling the temporal gaps in high-resolution data sequences.

Despite their utility, single-image super-resolution (SISR) methods face severe ill-posedness at large scaling factors [44], such as $\times 8$ or $\times 16$, where they struggle to accurately recover lost high-frequency details. Reference-based super-resolution (RefSR) addresses this limitation by introducing an additional HR reference (Ref) image from the same geographic area but at a different acquisition time. The rich texture priors provided by these reference images effectively alleviate the information deficiency inherent in SISR. While LR and Ref images are typically multi-temporal observations of the same location and share aligned spatial extents, factors such as varying viewpoints, spectral differences, and land cover changes make the effective exploitation of reference features challenging. Consequently, early RefSR methods [14, 46] focused primarily on feature alignment between LR and Ref to facilitate the transfer of reference textures.

The emergence of diffusion models [23, 42] has revolutionized RS RefSR by introducing multi-condition generative frameworks. Although diffusion models can generate remarkably rich details, especially by leveraging reference textures, they are prone to over-reliance on Ref features, which often leads to artifacts or false textures. Therefore, the core challenge in RS RefSR lies in effectively utilizing reference textures in unchanged regions while suppressing erroneous transfers in areas where land cover has changed. To this end, methods like Ref-Diff [5] introduce ideal land cover change priors as explicit conditions to guide the diffusion process. However, the performance of such methods degrades significantly in real-world applications where precise change detection masks are unavailable. Even the joint use of independent change detection and diffusion models often fails to produce stable and high-quality super-resolution results.

Meanwhile, recent advances in generative modeling have explored Multi-modal Diffusion Transformer (MM-DiT) architectures [8, 17], which enable effective cross-modal fusion through joint self-attention mechanisms. Nevertheless, for RS imagery, the standard text modality is insufficient for describing complex land cover features, rendering traditional text-guidance mechanisms ineffective. The optimal architectural design for integrating high-resolution reference images within the MM-DiT framework remains an open research challenge. A straightforward approach would be to replace the text branch in MM-DiT with LR and Ref images as new conditioning modalities. However, we observe that unconstrained interaction among these conditions leads to information competition, causing certain features to be over-exploited or under-utilized.

To mitigate these issues, and drawing inspiration from the architectural paradigms explored in CreatiLayout [40] and TODSynth [38], we propose DS-DiT, a Decoupled Siamese Diffusion Transformer that decouples the LR-Ref in-

teraction within the attention mechanism. Unlike existing siamese designs that replicate noisy features, our design allows the noisy image tokens to generate a single shared set of query, key, and value (Q, K, and V) projections that are dispatched to two parallel joint attention paths. This configuration ensures that LR structural priors and Ref texture information interact independently yet consistently with the same noisy image latent, effectively alleviating inter-source competition while maintaining global coherence. Furthermore, since global attention mechanisms may overlook critical local details, we introduce a Patch-Level Weighting (PLW) module that adaptively regulates the fusion of multi-source information, enhancing fine-grained detail recovery.

Importantly, the siamese architecture design provides a consistent latent anchor for the parallel attention paths. This consistency allows us to introduce an autoguidance strategy during inference, which is analogous to classifier-free guidance [12]. By exploiting the prediction discrepancy of the same model under strong and weak reference conditions, we explicitly amplify reference utilization and improve generation quality without the need for additional training or auxiliary models. Extensive experiments on multiple remote sensing datasets across large scaling factors demonstrate that DS-DiT achieves superior results even in the absence of ideal change detection priors.

The main contributions of this work are summarized as follows:

- We propose a Decoupled Siamese Diffusion Transformer (DS-DiT), which mitigates inter-source information competition by decoupling the LR-Ref interaction within the attention mechanism. This decoupled architecture further enables an inference-time autoguidance strategy, which improves generation quality without requiring additional training or auxiliary models.
- We design a Patch-Level Weighting (PLW) module to adaptively modulate the fusion of structural and textural features on a per-patch basis. By mitigating the detail loss inherent in global attention, this module enhances fine-grained detail recovery.
- Extensive experiments on multiple remote sensing datasets demonstrate that DS-DiT outperforms state-of-the-art methods in both quantitative metrics and visual fidelity across diverse evaluation settings.

2 Related Work

2.1 Reference-Based Image Super-Resolution

Compared to single-image super-resolution (SISR), RefSR achieves higher-quality reconstruction by jointly exploiting a low-resolution (LR) image and an additional high-resolution reference (Ref) image. Early RefSR methods focus primarily on feature alignment between LR and Ref images to enable texture transfer. CrossNet [46] employs optical flow estimation for image alignment, while SRNTT [45] performs multi-scale matching in the feature space. TTSR [35] transfers textures via cross-attention, and AMSA [34] adopts a coarse-to-fine progressive matching strategy. C^2-Matching [14] leverages contrastive learning

and knowledge distillation to handle cross-transformation and cross-resolution matching. DATSR [2] further introduces deformable attention for more flexible feature alignment.

Unlike natural images, LR and Ref images in remote sensing scenarios are typically acquired from the same location at different times, and can thus be considered spatially aligned. Consequently, the core challenge of remote sensing RefSR lies in effectively exploiting reference textures in unchanged regions while suppressing the erroneous transfer of irrelevant textures in changed regions. However, existing remote sensing RefSR approaches [6, 41] remain insufficient in utilizing effective texture information from Ref images. At large scaling factors, GAN-based methods struggle to faithfully reconstruct fine details in changed regions, leading to reconstructed images that lack high-frequency details. While recent diffusion-based works like Ref-Diff [5] attempt to use explicit land cover change priors to guide fusion, their reliance on additional change detection annotations limits their practical applicability. In contrast, our DS-DiT framework explores effective reference utilization without requiring any external change detection priors.

2.2 Conditional Diffusion Models for Super-Resolution

In recent years, diffusion models have achieved remarkable progress in image super-resolution owing to their powerful generative priors. SR3 [25] conditions on the LR image and performs super-resolution through iterative refinement. StableSR [31] leverages the generative prior of a pretrained text-to-image diffusion model for blind super-resolution. DiffBIR [18] adopts a two-stage strategy that first removes degradations and then supplements fine details via a diffusion model. PASD [37] introduces pixel-aware cross-attention, enabling the diffusion model to better perceive image structures. SeeSR [33] employs a degradation-aware prompt extractor to generate more realistic image details while preserving semantic consistency. DiT4SR [7] explores the capability of DiT-based diffusion models for real-world image super-resolution.

Despite their generative capability, diffusion models suffer from an inherent hallucination problem in SISR due to the lack of additional reference information, frequently producing textures inconsistent with the actual scene. Recent studies therefore combine RefSR with diffusion models, aiming to suppress hallucinations by leveraging Ref image information while generating high-quality images. CoSeR [28] uses a diffusion model to synthesize a corresponding reference image for the LR input, and then injects both into the denoising process via ControlNet. ReFIR [9] proposes a retrieval-augmented image restoration framework that exploits retrieved reference images to generate details faithful to the original scene. CRefDiff [29] fuses local and global information from the Ref image and leverages the generative prior of a diffusion model for real-world remote sensing image super-resolution. Ada-RefSR [32] follows a trust-but-verify principle, adaptively exploiting reliable reference information while suppressing unreliable parts.

2.3 Classifier-Free Guidance

Classifier-Free Guidance (CFG) [12] is a widely used inference-time technique for improving generation quality by extrapolating between conditional and unconditional predictions. Applying CFG typically requires jointly training models by dropping conditions with a fixed probability. However, in our framework, the text branch has been removed, and simply replacing the Ref image with an unrelated image to simulate unconditional generation can degrade reference utilization during normal inference. To circumvent the need for separate unconditional training, techniques like ICG [24] approximate unconditional predictions by sampling random independent conditions. Furthermore, Karras *et al.* [15] suggest using a smaller or under-trained version of the model in place of the unconditional model for guidance. Inspired by these self-guidance strategies, we propose an autoguidance strategy specifically tailored for the decoupled siamese architecture. This strategy leverages the prediction discrepancy between strong and weak reference conditions to explicitly amplify reference utilization without additional training.

3 Methodology

In this work, we adopt conditional flow matching as the training paradigm to improve the reconstruction quality of RefSR at large scaling factors. The overall workflow of the proposed method is shown in Fig. 1. We modify the MM-DiT architecture and design a Decoupled Siamese Diffusion Transformer, which introduces LR and Ref images as conditioning modalities on equal footing. In the joint attention mechanism, the guidance from LR and Ref branches is explicitly decoupled to prevent direct interaction between them, thereby alleviating inter-source competition. Furthermore, to capture local information that is easily overlooked by global attention, we adaptively fuse LR and Ref information at the patch level and inject the fused features into the denoising branch. During inference, we exploit the prediction discrepancy of the same model under different reference conditioning strengths to guide the denoising direction, further improving reconstruction quality without any additional training.

3.1 Preliminary

Unlike early diffusion models [11,26] that rely on stochastic differential equations (SDEs) to describe the diffusion process, Flow Matching [19] directly models sample trajectories using ordinary differential equations (ODEs), resulting in more efficient training and faster sampling. Rectified Flow [20] further simplifies this modeling process by constraining the transport paths between samples to linear interpolation trajectories:

$$\mathbf{x}_t = (1-t)\,\mathbf{x}_0 + t\,\mathbf{x}_1, \quad t \in [0,1]. \tag{1}$$

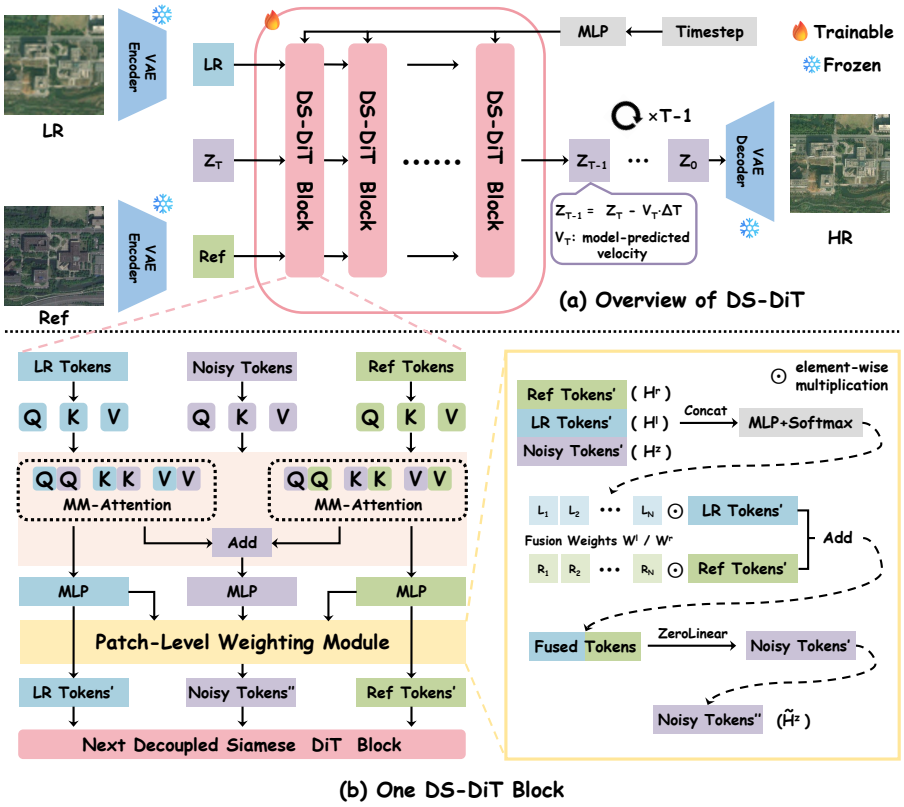

Fig. 1: Overall architecture of the proposed Decoupled Siamese Diffusion Transformer (DS-DiT). Our framework employs a decoupled siamese interaction with a shared set of Q, K, and V to mitigate inter-source competition, complemented by a Patch-Level Weighting (PLW) module for adaptive local feature fusion.

The training objective is to let the model-predicted velocity field $\mathbf{v}_t^\theta$ fit the ground-truth velocity field $\mathbf{v} = \mathrm{d}\mathbf{x}_t/\mathrm{d}t = \mathbf{x}_1 - \mathbf{x}_0$:

$$\mathcal{L}_{\mathrm{RF}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1} \left\| \mathbf{v}_t^\theta - (\mathbf{x}_1 - \mathbf{x}_0) \right\|_2^2 . \tag{2}$$

Current mainstream text-to-image models [1, 8, 27] adopt the MM-DiT architecture to predict the velocity field. Previous diffusion models typically rely on cross-attention to fuse image and text information, whereas MM-DiT designs independent Transformer branches for each modality. After encoding image and text features in their respective feature spaces, MM-DiT concatenates the resulting tokens along the sequence dimension and performs joint self-attention. This design enables image and text features to interact fully within a unified representation space, yielding superior multimodal fusion. Inspired by this, we treat LR and Ref images as conditioning modalities and model them in a manner similar

to the text modality in MM-DiT, enabling the model to exploit structural priors from the LR image and texture information from the Ref image for high-quality reconstruction.

3.2 Decoupling LR and Ref Interaction in MM-Attention

The text modality is unable to accurately describe land cover features in low-resolution remote sensing images, making it ineffective for remote sensing super-resolution tasks. We therefore completely discard the text branch and model LR and Ref as new conditioning modalities. A straightforward idea is to construct an M^3-DiT architecture that allows unconstrained interactions among all three modalities in the attention mechanism. Details of M^3-DiT are provided in the supplementary material. However, in this setting, the softmax operation in self-attention forces information from different sources to compete for limited attention weights within a single unified sequence, leading to feature dilution where structural priors or texture information are under-utilized. To alleviate this issue, we adopt a decoupled siamese architecture that separates the interaction between LR and Ref conditions.

In CreatiLayout [40], the image branch generates independent sets of queries (Q), keys (K), and values (V) for the joint attention with each modality. In contrast, in our design, the noisy image tokens $\mathbf{h}^z$ produce only a single set of Q, K, and V in each block, which is shared by both LR and Ref branches. The interaction between the LR tokens and the noisy image tokens is realized through MM-Attention:

$$\mathbf{h}_l^z,\ \mathbf{h}_{\text{new}}^l = \text{Attention}\big([\mathbf{h}^z\mathbf{W}_q^z,\ \mathbf{h}^l\mathbf{W}_q^l],\ [\mathbf{h}^z\mathbf{W}_k^z,\ \mathbf{h}^l\mathbf{W}_k^l],\ [\mathbf{h}^z\mathbf{W}_v^z,\ \mathbf{h}^l\mathbf{W}_v^l]\big)\,, \quad (3)$$

where $[\cdot,\cdot]$ denotes concatenation along the sequence dimension, $\mathbf{W}_q^z, \mathbf{W}_k^z, \mathbf{W}_v^z$ are the learnable projection matrices for the noisy image branch, $\mathbf{W}_q^l, \mathbf{W}_k^l, \mathbf{W}_v^l$ are the corresponding projection matrices for the LR branch, and $\mathbf{h}_l^z$ and $\mathbf{h}_{\text{new}}^l$ are the updated noisy image tokens and LR tokens after interaction, respectively. Meanwhile, the Ref tokens and the noisy image tokens undergo the same joint attention mechanism:

$$\mathbf{h}_r^z,\ \mathbf{h}_{\text{new}}^r = \text{Attention}\big([\mathbf{h}^z\mathbf{W}_q^z,\ \mathbf{h}^r\mathbf{W}_q^r],\ [\mathbf{h}^z\mathbf{W}_k^z,\ \mathbf{h}^r\mathbf{W}_k^r],\ [\mathbf{h}^z\mathbf{W}_v^z,\ \mathbf{h}^r\mathbf{W}_v^r]\big)\,. \quad (4)$$

The updated noisy image tokens are then obtained via a summation operation:

$$\mathbf{h}_{\text{new}}^z = \mathbf{h}_l^z + \mathbf{h}_r^z. \quad (5)$$

In this way, the guidance from LR and Ref is explicitly decoupled at the attention stage: each interacts with the noisy image independently, effectively mitigating inter-source competition.

3.3 Injecting LR and Ref Information via Patch-Level Weighting

In the DiT architecture, the attention mechanism operates at the global level. However, relying solely on global attention may lead to the loss of local information [7], which is equally critical for recovering fine-grained image details. It is

therefore necessary to strengthen the fusion of LR and Ref features at the local level. Notably, the information provided by Ref is not entirely reliable, as land cover in certain regions may have changed between the acquisition times of LR and Ref. For such changed regions, the guidance from LR should be strengthened to maintain fidelity, whereas for regions where land cover remains unchanged, enhancing the guidance from Ref is more beneficial for recovering fine textures. Based on this analysis, we adaptively fuse LR and Ref through Patch-Level Weighting and inject the fused features into the denoising branch.

Specifically, after the joint attention computation, the LR, Ref, and noisy image tokens $\mathbf{h}^l_{\text{new}}$, $\mathbf{h}^r_{\text{new}}$, and $\mathbf{h}^z_{\text{new}}$ are each passed through their respective MLPs, yielding feature tokens $\mathbf{H}^l, \mathbf{H}^r, \mathbf{H}^z \in \mathbb{R}^{N\times C}$. We first concatenate these three along the feature dimension and feed the result into a three-layer MLP, which outputs a per-patch weight map $\mathbf{W} \in \mathbb{R}^{N\times 2}$. A softmax operation is applied along the last dimension of $\mathbf{W}$ to assign specific fusion weights for each patch, and the resulting weight map is then split into LR and Ref weights. The weighted LR and Ref features are then fused and injected into the noisy tokens as follows:

$$\tilde{\mathbf{H}}^z = \mathbf{H}^z + \text{ZeroLinear}\big(\mathbf{W}^l \odot \mathbf{H}^l + \mathbf{W}^r \odot \mathbf{H}^r\big), \tag{6}$$

where $\mathbf{W}^l$ and $\mathbf{W}^r$ are the per-patch weights for LR and Ref, respectively, ZeroLinear denotes a linear projection layer with zero-initialized weights to ensure training stability, $\odot$ denotes element-wise multiplication with broadcasting, and $\tilde{\mathbf{H}}^z$ is the final output of the denoising branch in the current block.

3.4 Improving Image Quality with Autoguidance

In conditional rectified flow, the ideal velocity field $\mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0$ is a constant that depends only on the initial noise $\mathbf{x}_1$ and the target data $\mathbf{x}_0$. In the RefSR task, however, given LR and Ref as two conditions, the model-predicted velocity field $\mathbf{v}^\theta_t(\mathbf{x}_t \mid c_{\text{lr}}, c_{\text{ref}})$ deviates from the ideal velocity field and varies with the current state $\mathbf{x}_t$. If the predicted velocity field can be corrected through guidance at each sampling step, the entire sampling trajectory can be optimized toward the target distribution, leading to higher-quality reconstruction.

In our siamese attention mechanism, the noisy image tokens are updated via $\mathbf{h}^z_{\text{new}} = \mathbf{h}^z_l + \mathbf{h}^z_r$. During sampling, we introduce a scaling coefficient λ for the tokens produced by the reference interaction path:

$$\mathbf{h}^z_{\text{new}} = \mathbf{h}^z_l + \lambda\, \mathbf{h}^z_r. \tag{7}$$

When $\lambda = 1$, the model predicts the velocity field $\mathbf{v}^\theta_t(\mathbf{x}_t \mid c_{\text{lr}}, c_{\text{ref}})$ in its original direction. To construct a weak reference condition, we set $\lambda = 0$, which eliminates the direct contribution of the Ref branch to the noisy tokens and significantly reduces the guidance strength. The model then predicts a weakened velocity field $\mathbf{v}^{\theta^-}_t(\mathbf{x}_t \mid c_{\text{lr}}, c^-_{\text{ref}})$. Analogous to the extrapolation form of CFG, the corrected velocity field $\mathbf{v}^{\theta^+}_t$ can be expressed as:

$$\mathbf{v}^{\theta^+}_t(\mathbf{x}_t \mid c_{\text{lr}}, c^+_{\text{ref}}) = (1-\omega)\, \mathbf{v}^{\theta^-}_t(\mathbf{x}_t \mid c_{\text{lr}}, c^-_{\text{ref}}) + \omega\, \mathbf{v}^\theta_t(\mathbf{x}_t \mid c_{\text{lr}}, c_{\text{ref}}), \tag{8}$$

where ω is the guidance coefficient. When $\omega > 1$, the prediction trajectory is extrapolated along the reference guidance direction, shifting toward more thorough exploitation of the texture information from Ref. It should be noted that, to preserve the structural fidelity of the reconstructed image, ω must be kept within a moderate range to prevent the prediction from departing from the structural constraints imposed by the LR image. A detailed analysis of the influence of ω is provided in the supplementary material.

4 Experiments

4.1 Experimental Settings

Datasets. SECOND [36] is a semantic change detection dataset comprising 4,662 pairs of aerial images. Each image has a size of 512×512 pixels with a spatial resolution ranging from 0.5 to 3 meters. The dataset covers six major land cover categories across diverse sensors and regions, encompassing various imaging styles and scene types. We follow the official split [36], using 2,968 image pairs for training and 1,694 pairs for testing.

FUSU [39] is a high-resolution aerial land cover change detection dataset sourced from Google Earth and annotated with 17 land cover categories. The cropped images are 512×512 in size with a spatial resolution of 0.2 to 0.5 meters. We select 7,436 image pairs for training and 2,192 pairs for testing.

CNAM-CD [47] is a remote sensing change detection dataset collected from Google Earth, with a spatial resolution of 0.5 m. In our experiments, we use 1,758 image pairs for training and 750 image pairs for testing.

Real-RefRSSRD [29] is a real-world remote sensing RefSR dataset. In contrast to synthetic-degradation settings, Real-RefRSSRD contains paired HR and LR satellite images collected from different sensors. Specifically, the HR images are from NAIP with a ground sampling distance (GSD) of 1 m, while the LR images are from Sentinel-2 with a GSD of 10 m. This cross-sensor setting provides a more realistic evaluation scenario.

Implementation Details. For SECOND, FUSU, and CNAM-CD, we conduct super-resolution experiments at two challenging scaling factors: $\times 8$ and $\times 16$. Low-resolution (LR) images are generated via bicubic downsampling. While we do not simulate complex real-world degradations, these large scaling factors themselves pose significant challenges for structural and textural recovery. For Real-RefRSSRD, we follow its original real-world setting and perform $\times 10$ super-resolution. DS-DiT is built upon the pretrained MM-DiT blocks of SD3 [8]. Specifically, we extract the denoising branch weights to initialize the dual siamese branches for stable training. For each dataset and scale setting, the model is trained at the target resolution, i.e., 512×512 for SECOND, FUSU, and CNAM-CD, and 480×480 for Real-RefRSSRD. LR images are upsampled to the corresponding target resolution using bicubic interpolation before being fed into the model. We utilize the AdamW [21] optimizer with a learning rate of 5×10^{-5} and a weight decay of 0.001. Each model is trained on 4 NVIDIA

Table 1: Quantitative comparison of RefSR methods on the SECOND and FUSU datasets at ×8 and ×16 scaling factors. Red indicates the best and blue indicates the second-best result. (Creati* denotes CreatiLayout*.)

Dataset	Scale	Metric	TTSR	DATSR	CoSeR*	Ref-Diff	Creati*	M^3-DiT	Ours
SECOND	×8	LPIPS ↓	0.2994	0.3835	0.2576	0.2819	0.2692	0.2417	0.2174
		FID ↓	34.62	40.87	23.06	28.48	24.05	22.66	15.87
		DISTS ↓	0.1518	0.1923	0.1289	0.1490	0.1412	0.1356	0.1090
		CLIPIQA ↑	0.2937	0.2076	0.2973	0.4361	0.4255	0.4434	0.4698
		MUSIQ ↑	43.86	27.36	42.11	51.36	43.18	44.92	49.04
	×16	LPIPS ↓	0.3927	0.5039	0.3426	0.3503	0.3556	0.3504	0.2905
		FID ↓	134.68	99.74	31.12	30.11	32.03	35.50	21.58
		DISTS ↓	0.2196	0.2555	0.1595	0.1555	0.1670	0.1715	0.1334
		CLIPIQA ↑	0.3391	0.1554	0.3134	0.3468	0.4530	0.4489	0.4875
		MUSIQ ↑	37.71	22.02	42.18	49.75	43.55	43.65	48.65
FUSU	×8	LPIPS ↓	0.2960	0.3295	0.2447	0.2809	0.2143	0.2167	0.1741
		FID ↓	54.73	35.06	24.69	26.21	16.88	17.95	13.01
		DISTS ↓	0.1660	0.1803	0.1346	0.1618	0.1282	0.1289	0.1054
		CLIPIQA ↑	0.4998	0.2623	0.4178	0.5495	0.5076	0.5095	0.5306
		MUSIQ ↑	51.33	32.82	49.54	54.43	51.45	51.45	52.84
	×16	LPIPS ↓	0.3398	0.4546	0.2954	0.3288	0.2766	0.2641	0.2301
		FID ↓	127.12	104.78	35.01	34.26	20.55	20.35	19.52
		DISTS ↓	0.1949	0.2493	0.1652	0.1714	0.1510	0.1456	0.1309
		CLIPIQA ↑	0.4836	0.2360	0.4036	0.5175	0.5066	0.5101	0.5339
		MUSIQ ↑	48.85	27.72	49.11	53.00	50.78	51.55	53.21

A100 GPUs with a total batch size of 16 for 80k steps, taking approximately 21 hours. For inference, we use the Euler ODE solver with 40 steps. The guidance coefficient ω is set to 1.2 for SECOND and 1.1 for the remaining datasets.

Evaluation Metrics. To quantitatively evaluate RefSR performance, we adopt three reference-based perceptual metrics. LPIPS [43] measures perceptual similarity between images using features extracted by a pretrained deep network. FID [10] computes the distributional distance between generated and real images in feature space. DISTS [3] evaluates structural and textural similarity. Compared to PSNR and SSIM, these metrics are more suitable for assessing generative super-resolution methods [37]. In addition, we introduce two no-reference metrics, CLIPIQA [30] and MUSIQ [16], to comprehensively evaluate the visual quality of reconstructed images.

4.2 Comparison Results

For SECOND and FUSU, we compare the proposed method with state-of-the-art GAN-based and diffusion-based RefSR methods. These include two GAN-based methods (TTSR [35] and DATSR [2]), two U-Net-based diffusion methods (CoSeR* [28] and Ref-Diff [5]), and two DiT-based methods (CreatiLayout* [40]

Table 2: Quantitative comparison on CNAM-CD and Real-RefRSSRD. Red indicates the best and blue indicates the second-best result. (Creati* denotes CreatiLayout*.)

Dataset	Scale	Metric	CoSeR*	Ref-Diff	Creati*	M^3-DiT	Ours
CNAM-CD	×8	LPIPS ↓	0.2593	0.2787	0.2560	0.2566	0.2101
		FID ↓	51.00	51.97	45.92	46.50	35.76
		DISTS ↓	0.1365	0.1465	0.1434	0.1441	0.1109
		CLIPIQA ↑	0.3322	0.3794	0.4058	0.3980	0.4446
		MUSIQ ↑	36.85	47.15	39.71	40.06	44.08
	×16	LPIPS ↓	0.3319	0.3405	0.3506	0.3486	0.2789
		FID ↓	62.99	60.76	72.05	64.84	45.79
		DISTS ↓	0.1716	0.1545	0.1915	0.1827	0.1361
		CLIPIQA ↑	0.3301	0.3284	0.4019	0.4052	0.4325
		MUSIQ ↑	34.79	46.25	36.82	38.97	43.35
Real-RefRSSRD	×10	LPIPS ↓	0.4459	0.3979	0.3825	0.3756	0.3261
		FID ↓	52.33	48.20	34.64	35.18	31.56
		DISTS ↓	0.2450	0.2067	0.2192	0.2072	0.1809
		CLIPIQA ↑	0.4818	0.5870	0.6281	0.6299	0.6456
		MUSIQ ↑	46.07	56.43	57.07	56.90	57.09

and M^3-DiT introduced in Sec. 3.2). For CNAM-CD and Real-RefRSSRD, we compare our method with the diffusion-based methods. Specifically, CoSeR* denotes an adapted version of the original method, in which the automatically generated reference image is replaced with the paired real reference image from the dataset. CreatiLayout* is likewise adapted for the RefSR task by replacing the original text and layout inputs with LR and Ref, and adopting a two-stage training strategy: a SISR model is first trained, after which most parameters are frozen and the Ref branch is introduced for continued training.

Quantitative Comparison. Tab. 1 presents the results on SECOND and FUSU at two large scaling factors (×8 and ×16). The proposed method achieves the best performance on all reference-based metrics, demonstrating that our results possess both high fidelity and perceptual quality. On the no-reference metrics, our method also shows strong competitiveness, falling slightly behind Ref-Diff only in certain settings. As shown in Fig. 2, Ref-Diff tends to generate visually rich yet less faithful textures in some regions. Since no-reference metrics such as CLIPIQA [30] and MUSIQ [16] primarily reflect perceptual quality rather than explicit fidelity to the ground truth, such rich but non-authentic details may still receive higher scores. In contrast, our method achieves a better balance between reconstruction fidelity and perceptual realism.

Tab. 2 reports the results on CNAM-CD and Real-RefRSSRD. On CNAM-CD, our method consistently achieves the best results on all reference-based metrics at both ×8 and ×16 scaling factors. On Real-RefRSSRD, our method achieves the best performance across all metrics under the real-world ×10 setting.

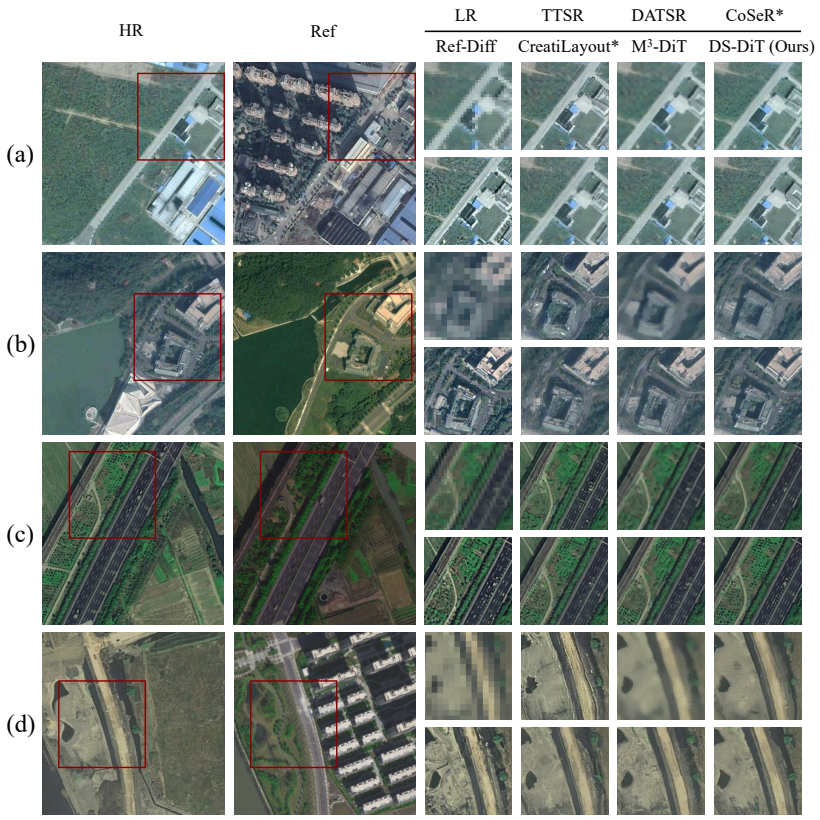

Fig. 2: Qualitative comparison of different RefSR methods on the SECOND and FUSU datasets. The visual results correspond to: (a) SECOND at $\times 8$ scaling factor, (b) SECOND at $\times 16$ scaling factor, (c) FUSU at $\times 8$ scaling factor, and (d) FUSU at $\times 16$ scaling factor. Please zoom in for better visual comparison.

These results show that the proposed method remains effective across different data distributions and realistic cross-sensor degradation scenarios.

Qualitative Comparison. Fig. 2 provides visual comparison results on SECOND and FUSU. Fig. 2(a) shows an example from the SECOND $\times 8$ dataset with substantial land cover changes. Even though only a few regions remain unchanged, our method successfully extracts the corresponding textures and reconstructs the fine details of the buildings on the right side of the image. Fig. 2(b) presents an example from the SECOND $\times 16$ dataset with minor changes. At this extreme scaling factor, other methods fail to transfer textures successfully, producing artifacts or blurry results despite the valuable information provided by Ref. In contrast, our method captures usable texture information and transfers it effectively, recovering the complex structures of the buildings. Fig. 2(c) illustrates an example from the FUSU $\times 8$ dataset. When the vegetation information in Ref

is untrustworthy due to seasonal changes, our method avoids blindly transferring erroneous textures; meanwhile, it effectively exploits reliable road textures to assist reconstruction—a selective transfer capability that other methods lack. Finally, Fig. 2(d) shows an example from the FUSU $\times 16$ dataset with drastic changes, where our method still successfully recovers the clear shapes of the lake and the river. In summary, the qualitative comparisons further validate the effectiveness of the proposed method, which selectively exploits Ref information rather than blindly copying it. Under the structural constraints imposed by the LR image, our approach generates reconstructed results with higher fidelity and superior perceptual quality. More visual comparison results are provided in the supplementary material.

4.3 Ablation Studies

We conduct ablation studies on the SECOND dataset at both $\times 8$ and $\times 16$ scaling factors to validate the effectiveness of each component. Detailed results are presented in Tab. 3.

Importance of Decoupled Siamese Attention. Compared to M^3-Attention, our decoupled siamese attention separates the interaction between LR and Ref conditions. Although this avoids their explicit interaction within the same attention layer, it achieves superior performance on nearly all metrics (see the comparison between M^3-DiT in Tab. 1 and the siamese attention baseline in Tab. 3). This confirms that by mitigating inter-source competition, the decoupled architecture improves information utilization and prevents feature dilution. As observed in Fig. 3(a) and Fig. 3(b), this architecture recovers more building and ground details compared to M^3-Attention.

Effect of Patch-Level Weighting Module. After introducing the PLW module, all three reference-based metrics show consistent improvement, indicating that this module effectively compensates for the deficiency of global attention in local modeling. By adaptively modulating the fusion weights based on the local context, the PLW module suppresses artifacts in changed regions while enhancing textures in unchanged areas. Fig. 3(c) further corroborates this observation, as the texture details of buildings are noticeably sharper and more refined.

Effect of Autoguidance. When autoguidance is applied to the baseline model, all metrics improve, with particularly significant gains in perceptual and no-reference metrics. This indicates that the corrected velocity field steers the sampling trajectory toward data distributions with richer high-frequency details, thereby producing more realistic reconstructions.

Complementarity of Components. When both modules are introduced simultaneously, all metrics reach their optimal values, showing larger gains than introducing either module alone. This suggests that the PLW module and autoguidance are highly complementary. Specifically, the PLW module facilitates the recovery of local structures, while autoguidance enhances global textural realism; together, they improve the overall reconstruction quality. As shown in Fig. 3(d), the reconstructed result maintains high fidelity while exhibiting stronger perceptual realism.

(a) M³-DiT (b) Siamese Attention (c) DS-DiT **w/o** Autoguidance (d) DS-DiT **w/** Autoguidance

Fig. 3: Visual ablation study on the SECOND dataset (×16 scaling factor).

Table 3: Ablation study on the SECOND dataset. The results are reported for ×8 (upper four rows) and ×16 (lower four rows) scaling factors. ✓indicates the component is included. The best result in each group is highlighted in **bold**.

Siamese Attention	PLW Module	Auto-guidance	LPIPS↓	FID↓	DISTS↓	CLIPIQA↑	MUSIQ↑
✓			0.2362	19.16	0.1251	0.4403	44.95
✓	✓		0.2291	18.93	0.1234	0.4399	45.07
✓		✓	0.2278	18.56	0.1159	0.4680	48.29
✓	✓	✓	**0.2174**	**15.87**	**0.1090**	**0.4698**	**49.04**
✓			0.3159	31.29	0.1571	0.4514	43.78
✓	✓		0.3085	30.32	0.1557	0.4604	43.88
✓		✓	0.3060	27.96	0.1449	0.4649	46.60
✓	✓	✓	**0.2905**	**21.58**	**0.1334**	**0.4875**	**48.65**

4.4 Discussion on LR and Ref Injection Strategy

It is worth noting that not all injection strategies lead to performance improvements. We initially designed two simpler injection approaches. Variant A keeps LR and Ref isolated from each other, avoiding their explicit interaction. The LR tokens $\mathbf{H}^l$ and Ref tokens $\mathbf{H}^r$ are injected into the noisy tokens $\mathbf{H}^z$ through independent zero-initialized linear layers:

$$\tilde{\mathbf{H}}^z = \mathbf{H}^z + \text{ZeroLinear}(\mathbf{H}^l) + \text{ZeroLinear}(\mathbf{H}^r). \tag{9}$$

Variant B directly sums LR and Ref before injecting them into the noisy tokens through a single zero-initialized linear layer:

$$\tilde{\mathbf{H}}^z = \mathbf{H}^z + \text{ZeroLinear}(\mathbf{H}^l + \mathbf{H}^r). \tag{10}$$

Our proposed Patch-Level Weighting (PLW) injection strategy is denoted as Variant C, with the corresponding formulation given in Eq. (6). The comparison results on the SECOND ×8 dataset are presented in Tab. 4. Variant A improves PSNR and SSIM but leads to a higher FID, suggesting that independently injecting LR and Ref information does not necessarily improve perceptual quality.

Table 4: Comparison of different injection strategies on the SECOND ×8 dataset. The best result is highlighted in **bold**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	DISTS ↓
Siamese Attention	24.50	0.5921	0.2362	19.16	0.1251
+ Variant A	24.63	0.5981	0.2327	23.07	0.1303
+ Variant B	24.63	0.5966	0.2308	19.81	0.1254
+ Variant C (PLW)	**24.70**	**0.6005**	**0.2291**	**18.93**	**0.1234**

Variant B reduces FID compared with Variant A, but its direct summation still provides limited gains over the baseline. In contrast, Variant C (PLW) achieves the best overall performance, indicating the benefit of adaptive per-patch fusion.

However, the design of the PLW module implicitly assumes that LR and Ref are roughly spatially aligned. This assumption generally holds in remote sensing scenarios, where both images originate from the same geographic location. For spatially misaligned reference images, *e.g.*, captured from different viewpoints or covering different regions, the per-patch weight allocation would lose its spatial correspondence, potentially leading to degraded performance. How to adaptively retrieve and exploit useful texture information from spatially misaligned reference images remains an important direction for our future exploration.

5 Conclusion

In this work, we propose the Decoupled Siamese Diffusion Transformer (DS-DiT), a framework designed to balance LR structural fidelity and Ref texture utilization in remote sensing RefSR. Through the decoupled siamese attention mechanism, DS-DiT separates the interaction between LR and Ref conditions, preserving structural fidelity while incorporating high-frequency details. To further address the limitations of global attention, we design a Patch-Level Weighting module for adaptive local feature fusion. Additionally, we introduce an autoguidance strategy that improves reconstruction quality during inference by exploiting the model's internal structural priors without requiring additional training. This work demonstrates the significant potential of DS-DiT for high-fidelity remote sensing reconstruction, providing a robust and flexible approach for practical applications where precise texture recovery must be balanced with strict structural authenticity.

Acknowledgements

This research was supported in part by the Guangdong S&T Program (Grant No. 2024B0101040005), the National Natural Science Foundation of China (Grant No. T2125006), and the Natural Science Foundation of Guangdong Province, China (Grant No. 2026A1515010676).

References

1. Black Forest Labs: Flux. `https://bfl.ai/blog/24-08-01-bfl` (2024), accessed: 2026-06-24
2. Cao, J., Liang, J., Zhang, K., Li, Y., Zhang, Y., Wang, W., Gool, L.V.: Reference-based image super-resolution with deformable attention transformer. In: European conference on computer vision. pp. 325–342. Springer (2022)
3. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
4. Dong, R., Mou, L., Chen, M., Li, W., Tong, X.Y., Yuan, S., Zhang, L., Zheng, J., Zhu, X., Fu, H.: Large-scale land cover mapping with fine-grained classes via class-aware semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16783–16793 (2023)
5. Dong, R., Yuan, S., Luo, B., Chen, M., Zhang, J., Zhang, L., Li, W., Zheng, J., Fu, H.: Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 27684–27694 (2024)
6. Dong, R., Zhang, L., Fu, H.: Rrsgan: Reference-based super-resolution for remote sensing image. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–17 (2021)
7. Duan, Z.P., Zhang, J., Jin, X., Zhang, Z., Xiong, Z., Zou, D., Ren, J.S., Guo, C., Li, C.: Dit4sr: Taming diffusion transformer for real-world image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18948–18958 (2025)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
9. Guo, H., Dai, T., Ouyang, Z., Zhang, T., Zha, Y., Chen, B., Xia, S.t.: Refir: Grounding large restoration models with retrieval augmentation. Advances in Neural Information Processing Systems **37**, 46593–46621 (2024)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
13. Jiang, J., Zhang, Q., Yao, X., Tian, Y., Zhu, Y., Cao, W., Cheng, T.: Histif: A new spatiotemporal image fusion method for high-resolution monitoring of crops at the subfield level. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 4607–4626 (2020)
14. Jiang, Y., Chan, K.C., Wang, X., Loy, C.C., Liu, Z.: Robust reference-based super-resolution via c2-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2103–2112 (2021)
15. Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S.: Guiding a diffusion model with a bad version of itself. Advances in Neural Information Processing Systems **37**, 52996–53021 (2024)

16. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
17. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
18. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: European conference on computer vision. pp. 430–448. Springer (2024)
19. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
20. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
22. Luo, Y., Guan, K., Peng, J.: Stair: A generic and fully-automated method to fuse multiple sources of optical satellite data to generate a high-resolution, daily and cloud-/gap-free surface reflectance product. Remote Sensing of Environment **214**, 87–99 (2018)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
24. Sadat, S., Kansy, M., Hilliges, O., Weber, R.: No training, no problem: Rethinking classifier-free guidance for diffusion models. In: International Conference on Learning Representations. vol. 2025, pp. 76833–76858 (2025)
25. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4713–4726 (2022)
26. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
27. Stability AI: Stable diffusion 3.5. `https://stability.ai/news/introducing-stable-diffusion-3-5` (2024), accessed: 2026-06-24
28. Sun, H., Li, W., Liu, J., Chen, H., Pei, R., Zou, X., Yan, Y., Yang, Y.: Coser: Bridging image and language for cognitive super-resolution. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 25868–25878 (2024)
29. Wang, C., Sun, W.: Controllable reference-guided diffusion with local-global fusion for real-world remote sensing super-resolution. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2026)
30. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)
31. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)
32. Wang, Y., Wan, Y., Zheng, S., Li, B., Hou, Q., Jiang, P.T.: Trust but verify: Adaptive conditioning for reference-based diffusion super-resolution via implicit reference correlation modeling. arXiv preprint arXiv:2602.01864 (2026)

33. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25456–25467 (2024)
34. Xia, B., Tian, Y., Hang, Y., Yang, W., Liao, Q., Zhou, J.: Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In: Proceedings of the aaai conference on artificial intelligence. vol. 36, pp. 2768–2776 (2022)
35. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5791–5800 (2020)
36. Yang, K., Xia, G.S., Liu, Z., Du, B., Yang, W., Pelillo, M., Zhang, L.: Semantic change detection with asymmetric siamese networks. arXiv preprint arXiv:2010.05687 (2020)
37. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: European conference on computer vision. pp. 74–91. Springer (2024)
38. Yang, Y., Zhang, Y., Zhang, K., Zhang, J., Chen, X., Fu, H., Dong, R.: Task-oriented data synthesis and control-rectify sampling for remote sensing semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42147–42157 (2026)
39. Yuan, S., Lin, G., Zhang, L., Dong, R., Zhang, J., Chen, S., Zheng, J., Wang, J., Fu, H.: Fusu: A multi-temporal-source land use change segmentation dataset for fine-grained urban semantic understanding. Advances in Neural Information Processing Systems **37**, 132417–132439 (2024)
40. Zhang, H., Hong, D., Wang, Y., Shao, J., Wu, X., Wu, Z., Jiang, Y.G.: Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18487–18497 (2025)
41. Zhang, J., Zhang, W., Jiang, B., Tong, X., Chai, K., Yin, Y., Wang, L., Jia, J., Chen, X.: Reference-based super-resolution method for remote sensing images with feature compression module. Remote Sensing **15**(4), 1103 (2023)
42. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
43. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
44. Zhang, Y., Zhang, Z., DiVerdi, S., Wang, Z., Echevarria, J., Fu, Y.: Texture hallucination for large-factor painting super-resolution. In: European Conference on Computer Vision. pp. 209–225. Springer (2020)
45. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7982–7991 (2019)
46. Zheng, H., Ji, M., Wang, H., Liu, Y., Fang, L.: Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In: Proceedings of the European conference on computer vision (ECCV). pp. 88–104 (2018)
47. Zhou, Y., Wang, J., Ding, J., Liu, B., Weng, N., Xiao, H.: Signet: A siamese graph convolutional network for multi-class urban change detection. Remote Sensing **15**(9), 2464 (2023)