# Building Bridges across Spatial and Temporal Resolutions: Reference-Based Super-Resolution via Change Priors and Conditional Diffusion Model

Runmin Dong[1,5], Shuai Yuan[2], Bin Luo[1], Mengxuan Chen[1,5], Jinxiao Zhang[1,5], Lixian Zhang[4,5*], Weijia Li[3], Juepeng Zheng[3,5], Haohuan Fu[1,5*]
[1]Tsinghua University [2]The University of Hong Kong [3]Sun Yat-Sen University
[4]National Supercomputing Center in Shenzhen
[5]Tsinghua University - Xi'an Institute of Surveying and Mapping Joint Research Center
drm@mail.tsinghua.edu.cn, haohuan@tsinghua.edu.cn

## Abstract

*Reference-based super-resolution (RefSR) has the potential to build bridges across spatial and temporal resolutions of remote sensing images. However, existing RefSR methods are limited by the faithfulness of content reconstruction and the effectiveness of texture transfer in large scaling factors. Conditional diffusion models have opened up new opportunities for generating realistic high-resolution images, but effectively utilizing reference images within these models remains an area for further exploration. Furthermore, content fidelity is difficult to guarantee in areas without relevant reference information. To solve these issues, we propose a change-aware diffusion model named Ref-Diff for RefSR, using the land cover change priors to guide the denoising process explicitly. Specifically, we inject the priors into the denoising model to improve the utilization of reference information in unchanged areas and regulate the reconstruction of semantically relevant content in changed areas. With this powerful guidance, we decouple the semantics-guided denoising and reference texture-guided denoising processes to improve the model performance. Extensive experiments demonstrate the superior effectiveness and robustness of the proposed method compared with state-of-the-art RefSR methods in both quantitative and qualitative evaluations. The code and data are available at https://github.com/dongrunmin/RefDiff.*

## 1. Introduction

Spatiotemporal integrity of high-resolution remote sensing images is crucial for fine-grained urban management, long-time-series urban development study, disaster monitoring, and other remote sensing applications [6, 13, 49].

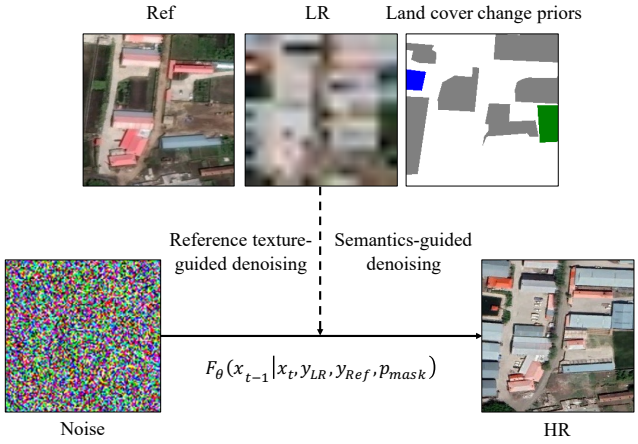


Figure 1. Illustration of the proposed change-aware diffusion model for RefSR. LR is a low-resolution image and HR is the corresponding high-resolution image. Ref represents a geographically matched reference high-resolution image acquired at another time.

However, due to limitations in remote sensing technologies and high hardware costs, we cannot simultaneously achieve high temporal resolution and high spatial resolution images on a large scale [24, 47]. To tackle this issue, reference-based super-resolution (RefSR) can leverage geography-paired high-resolution reference (Ref) images and low-resolution (LR) images to integrate fine spatial content and high revisit frequency from different sensors [7]. Although various RefSR methods achieve great progress, two major challenges remain to be solved for this scenario.

The first challenge is the land cover changes between Ref and LR images. Unlike the natural image domain, where Ref images are collected through image retrieval or captured from different viewpoints, Ref and LR images in remote sensing scenarios utilize geographic information to

*Corresponding authors

match the same location. Existing methods implicitly capture the land cover changes between LR and Ref images by adaptive learning or attention-based transformers [4, 26]. However, the underuse or misuse problems of Ref information still exist in these methods.

The second challenge is the large spatial resolution gaps between remote sensing sensors (e.g., 8× to 16×). Existing RefSR methods are usually based on the generative adversarial network (GAN) and designed for a 4× scaling factor [51, 54]. They can hardly reconstruct and transfer the details in the face of large-factor super-resolution. In recent years, conditional diffusion models have demonstrated greater effectiveness in image super-resolution and reconstruction than GAN [11, 41]. A straightforward way to boost RefSR is to use LR and Ref images as conditions for the diffusion model. To effectively utilize the reference information, some methods [18, 38] inject Ref information into the blocks of the denoising networks. However, they implicitly model the relationship between LR and Ref images for denoising, leading to ambiguous usage of Ref information and content fidelity limitation.

To alleviate the above issues, we introduce land cover change priors to improve the effectiveness of reference feature usage and the faithfulness of content reconstruction (as shown in Figure 1). Benefiting from the development of remote sensing change detection (CD), we can use off-the-shelf CD methods to effectively capture land cover changes between images of different spatial resolutions [22, 32, 52]. On the one hand, the land cover change priors enhance the utilization of reference information in unchanged areas. On the other hand, the changed land cover classes can guide the reconstruction of semantically relevant content in changed areas. Furthermore, according to the land cover change priors, we can decouple the semantics-guided denoising and reference texture-guided denoising in an iterative way to improve the model performance. To illustrate the effectiveness of the proposed method, we perform experiments on two datasets using two large scaling factors. Our method achieves state-of-the-art performance. In summary, our contributions are summarized as follows:

- We introduce the land cover change priors in RefSR to improve the content fidelity of reconstruction in changed areas and the effectiveness of texture transfer in unchanged areas, building bridges across spatial and temporal resolutions in remote sensing scenarios.
- We propose a novel RefSR method named Ref-Diff that injects the land cover change priors into the conditional diffusion model by the change-aware denoising model, enhancing the model's effectiveness in large-factor super-resolution.
- Experimental results demonstrate that the proposed method outperforms the existing SOTA RefSR methods in both quantitative and qualitative aspects.

## 2. Related Works

### 2.1. Reference-Based Super-Resolution Methods

Compared to single-image super-resolution (SISR), RefSR shows great potential in alleviating ill-posed problems and recovering realistic textures [1, 23]. Specifically, Jiang et al. [14] propose a contrastive correspondence network and a teacher-student correlation distillation method to address the misalignment issues in the texture transfer and resolution gaps between LR and Ref images. RRSR [48] and AMSA [40] contribute to high-quality correspondence matching. Besides, Huang et al. [12] decouple super-resolution and texture transfer tasks to alleviate the issues of the underuse and misuse of Ref images.

Owing to the pre-matching of LR and Ref images through geo-locations, existing RefSR methods for remote sensing images [4, 46] aim to transform relevant textures and suppress the irrelevant information fusion. However, their results contain apparent internal resolution inconsistencies between changed and unchanged regions in large-factor super-resolution. Because the details of changed regions can hardly be reconstructed using GAN-based methods. Therefore, recent works adopt the diffusion model to generate more realistic results [18]. For example, HSR-Diff [38] applies the conditional diffusion model and utilizes cross-attention as the conditioning mechanism to incorporate LR and Ref features into the denoising process, improving the perceptual quality. However, limited by implicit relationship modeling between LR and Ref images in the denoising process, the difficulty of denoising and the uncertainty of results are increased. In this work, we introduce the land cover change priors and explicitly use them to guide the denoising process.

### 2.2. Conditional Diffusion Model for Super-Resolution

Benefiting from diffusion models, recent image super-resolution techniques have witnessed significant progress in terms of visual appeal and high-quality output. The initial works [17, 30] utilize LR images as the condition for the diffusion processes to deal with the large-factor super-resolution. To further improve the effectiveness of image super-resolution, some works explore enhanced conditions to guide the denoising process. For example, ResDiff [31] and ACDMSR [27] use the CNN-enhanced LR prediction as a condition to accelerate the generation process and acquire superior sample quality. BlindSRSNF [39] and Dual-Diffusion [42] combine degradation representations to the condition of the diffusion model to achieve satisfactory results in real-world scenarios.

Except for simply combining those priors with the input of the conditional diffusion model, recent works integrate them into the denoising models [8, 34]. Wang et al. [35]

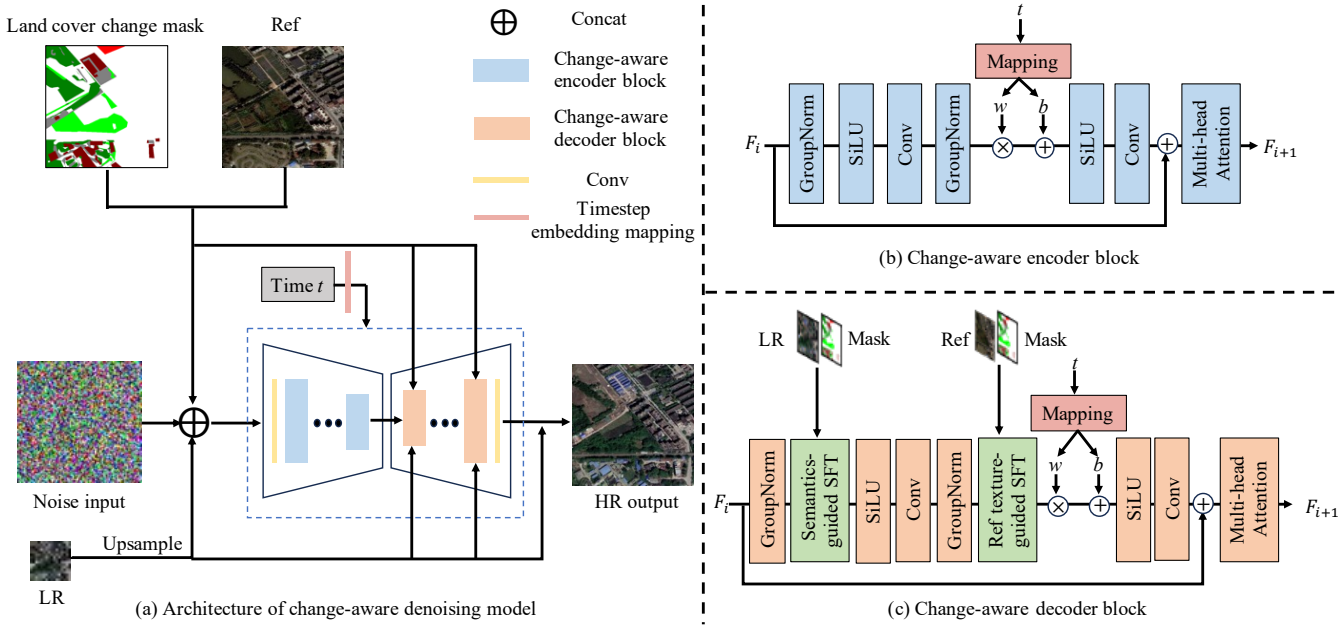


Figure 2. The architecture of the proposed change-aware denoising model. It consists of change-aware encoder and decoder blocks. The LR, Ref, and land cover change mask are combined with the noise input and are also injected into change-aware decoder blocks.

introduce the semantic layout to the decoder by multi-layer spatially-adaptive normalization operators for semantic image synthesis. PASD [45] and DiffBIR [20] adopt ControlNet to introduce priors like the high-level information extracted from CLIP or the enhanced LR representation by degradation removal. In this work, we explore the utilization of the land cover change priors in RefSR. We inject the priors into the denoising model, which decouples the semantics-guided and reference texture-guided denoising.

### 2.3. Change Detection

With the development of change detection (CD) methods, existing works can achieve up to 80% F1-score on different land cover categories [25]. For practical applications, recent CD models tend to be lightweight and can handle multi-temporal images with different resolutions [22]. For example, Zheng et al. [52] design a cross-resolution difference learning to bridge the resolution gap between two temporal images without resizing operations. Liu et al. [21] propose a SISR-based change detection network with a stacked attention module, achieving above 83% F1-score in 8× resolution difference on the building CD task. These CD methods show high confidence and plug-and-play abilities, which can directly provide high-quality land cover change prior information for this work.

## 3. Methodology

In this paper, we adopt the conditional diffusion model to boost the effectiveness of RefSR methods in large scaling factors. To enhance the fidelity of the generated content and improve the effectiveness of the Ref image transfer, we introduce land cover change priors. The architecture of the proposed method is shown in Figure 2. We propose a novel change-aware denoising model, injecting the Ref features and the land cover change priors into the denoising blocks. Leveraging the priors, we decouple the semantics-guided denoising and reference texture-guided denoising processes in the decoder, and cope with the two processes iteratively.

### 3.1. Preliminary

The conditional diffusion model extends the basic diffusion model by incorporating conditions, including forward and reverse diffusion processes. Karras et al. [15] unify different diffusion models into the EDM framework. The training objective of EDM is defined as:

$$\mathbb{E}_{\sigma,\mathbf{y},\mathbf{n}}[\lambda(\sigma)||D(\mathbf{y}+\mathbf{n};\sigma)-\mathbf{y}||_2^2], \quad (1)$$

where standard deviation $\sigma$ controls the noise level, $\mathbf{y}$ is a training image and $\mathbf{n}$ is noise. $D(\cdot)$ is a denoiser function. $\lambda(\sigma)$ is the loss weight.

To effectively train a neural network, the preconditioning of EDM is defined as:

$$D_\theta(\mathbf{x};\sigma) = c_{\text{skip}}(\sigma)\mathbf{x} + c_{\text{out}}(\sigma)F_\theta(c_{\text{in}}(\sigma)\mathbf{x}; c_{\text{noise}}(\sigma)), \quad (2)$$

where $\mathbf{x} = \mathbf{y}+\mathbf{n}$. $F_\theta(\cdot)$ represents the neural network undergoing training. $c_{\text{skip}}$ adjusts the skip connection. $c_{\text{in}}$ and $c_{\text{out}}$ scale the input and output magnitudes, respectively. $c_{\text{noise}}$ is used to map the noise level $\sigma$ into a conditioning input for

$F_\theta(\cdot)$. In this work, the diffusion architecture follows the formulations of the training objective, preconditioning, and other implementations in EDM.

### 3.2. Change-Aware Denoising Model

This work aims to exploit the land cover change priors to facilitate RefSR in large scaling factors for remote sensing images. The proposed change-aware denoising model is shown in Figure 2(a). Inspired by [28, 35], the land cover change prior can be regarded as the semantic layout of the changed areas between LR and Ref images for the semantics-guided denoising. Meanwhile, the texture details can be enhanced in the unchanged areas through reference texture-guided denoising. As a result, the semantics-guided denoising and reference texture-guided denoising processes can be decoupled in the change-aware decoder (see Figure 2(c)), further improving the denoising results.

**Land Cover Change Priors.** Land cover change priors used in this work are the pixel-level multi-category change detection mask for each image pair, including a no-change class and different land cover change classes. To fully unleash the potential of land cover priors in training, we use the ground truth of the land cover change mask as the condition. In real applications, change detection masks can be generated by the off-the-shelf end-to-end change detection methods or two-stage land cover classification methods. As shown in Figure 2(a), the land cover change mask is combined with the noise as input and also injected into the change-aware decoder.

**Change-Aware Encoder.** To improve the computational performance and avoid over-intervention of LR denoising, the LR image, Ref image, and land cover change mask are concatenated with the noisy image as the input to the encoder, instead of being injected into the encoder blocks as in [38]. The architecture of the encoder is based on the improved U-Net in [15] (see Figure 2(b)). Each change-aware encoder block consists of group normalization, convolution, SiLU, and a multi-head attention module. Since each timestep $t$ corresponds to a certain noise level, we map the timestep embedding into learnable weight $w(t)$ and bias $b(t)$ to regulate the features. Multi-head attention runs through the attention process multiple times in parallel, each with its own set of learnable parameters [33].

**Change-Aware Decoder.** As shown in Figure 2(c), we inject the features of land cover change masks and Ref images into the change-aware decoder blocks. With the land cover change priors, we decouple the semantics-guided denoising in the changed areas and reference texture-guided denoising in the unchanged areas. To tackle the mislabel problem in the land cover change priors, we combine the land cover change masks and LR images for the semantics-guided spatial feature transform (SFT) module. Considering the no-change class in land cover change mask can guide the utilization of Ref texture, we combine the land cover change masks and Ref images for the Ref texture-guided SFT module. In this way, the denoising of changed and unchanged areas can reinforce each other by an iterative solution. Considering the accuracy of predicting land cover change masks through change detection methods is usually between 60% to 80% in practical applications, we combine the guidance features and denoising features for the learning of spatially adaptive weight and bias, rather than only use the guidance features like the original SFT [36] and SPADE [28, 35]. The modified SFT module can be formulated as:

$$F_{i+1} = \gamma_i(F_e \oplus F_i) \cdot F_i + \beta_i(F_e \oplus F_i), \tag{3}$$

where $F_i$ and $F_{i+1}$ are the input and output features of the SFT module, respectively. $\gamma_i(\cdot)$, $\beta_i(\cdot)$ are the spatially-adaptive weight and bias learned from the combination of the guidance features $F_e$ obtained by the extractor and the input features $F_i$, respectively.

### 3.3. Degradation Model and Implementation Details

We adopt a comprehensive degradation to simulate LR images in real-world scenarios for training. According to off-the-shelf blind super-resolution methods [9, 37] and the characteristics of remote sensing sensors [5, 29], we adopt isotropic Gaussian blur, anisotropic Gaussian blur, motion blur, resize with different interpolation methods, additive Gaussian noise, and JPEG compression noise to synthesis LR images. The setting of degradation complexity is based on the scaling factor. In the experiments, the degradation model for $16\times$ datasets is simpler than that for $8\times$ datasets.

During training, each high-resolution (HR) image, Ref image, and land cover change mask are randomly cropped to a size of $256 \times 256$, and the size of the corresponding LR image is associated with the scaling factors. The implementation of the diffusion model is according to [15]. We utilize a dropout rate of 0.2. The batch size is set to 48. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initialized as $1 \times 10^{-4}$. The model is updated for 500k iterations using 4 NVIDIA A800 GPUs.

## 4. Experiments

### 4.1. Datasets and Evaluation

**SECOND Dataset.** SECOND [44] is a semantic change detection dataset with 7 land cover class annotations, including non-vegetated ground surface, tree, low vegetation, water, buildings, playgrounds, and unchanged areas. The images are collected from different sensors and areas with resolutions between 0.5 and 1 meters, guaranteeing style diversity and scene diversity. In this work, we use 2,668 image pairs with a size of $512 \times 512$ for training and 1,200 image pairs with a size of $256 \times 256$ for testing.

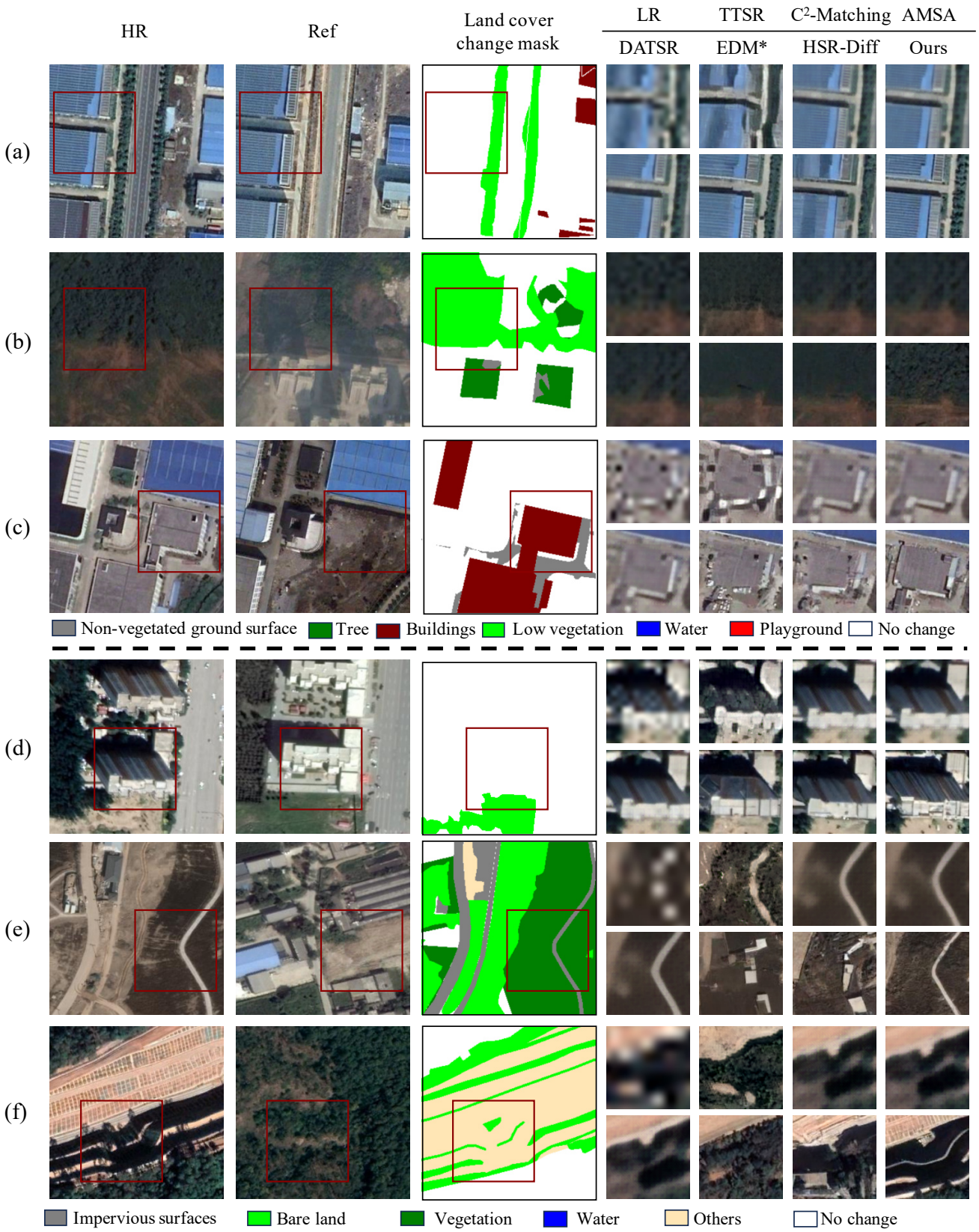


Figure 3. Comparison results on SECOND (a-c) and CNAM-CD (d-f) datasets with 8× and 16× scaling factors.

Table 1. Quantitative comparison with different reference-based methods on two datasets, i.e., SECOND and CNAM-CD. Each dataset is evaluated at two large scaling factors, i.e., 8× and 16×. Lower LPIPS and FID values indicate better results. Bold indicates the best results.

| Methods | SECOND | | | | CNAM-CD | | | |
|---|---|---|---|---|---|---|---|---|
| | 8× | | 16× | | 8× | | 16× | |
| | LPIPS↓ | FID↓ | LPIPS↓ | FID↓ | LPIPS↓ | FID↓ | LPIPS↓ | FID↓ |
| TTSR [43] | 0.3799 | 142.4030 | 0.4743 | 232.0805 | 0.4041 | 163.8194 | 0.4700 | 191.8991 |
| WTRN [19] | 0.5081 | 110.4582 | 0.8426 | 260.4063 | 0.4915 | 142.1856 | 0.8121 | 283.1182 |
| $C^2$-Matching [14] | 0.3351 | 62.2991 | 0.4972 | 123.3611 | 0.3697 | 97.9389 | 0.5494 | 182.0406 |
| AMSA [40] | 0.3601 | 56.9689 | 0.5353 | 135.6814 | 0.3989 | 92.3733 | 0.5613 | 169.0448 |
| DATSR [3] | 0.3525 | 56.0531 | 0.5078 | 111.7564 | 0.3894 | 94.7989 | 0.5331 | 163.8926 |
| EDM* [15] | 0.2886 | 34.5802 | 0.3440 | 37.5573 | 0.3301 | 53.1511 | 0.3902 | 59.4250 |
| HSR-Diff [38] | 0.2689 | 45.4743 | 0.3437 | 51.1473 | 0.3045 | 67.2524 | 0.3771 | 68.1954 |
| Ref-Diff (ours) | **0.2642** | **32.5961** | **0.3433** | **33.9690** | **0.2791** | **43.0152** | **0.3519** | **45.6511** |

[1] EDM* represents that the original method is modified for the RefSR task, which combines LR and Ref images with the noise input.

**CNAM-CD Dataset.** CNAM-CD [53] is a multi-class change detection dataset with a resolution of 0.5 meter, including 6 land cover classes, i.e., bare land, vegetation, water, impervious surfaces (buildings, roads, parking lots, squares, etc.), others (clouds, hard shadows, clutter, etc.), and unchanged areas. The image pairs are collected from Google Earth from 2013 to 2022. We use 2,258 image pairs with a size of 512 × 512 for training and 1,000 image pairs with a size of 256 × 256 for testing.

**Evaluation.** The performance of the proposed approach and other competing methods on test datasets are assessed with the learned perceptual image patch similarity (LPIPS) and Fréchet inception distance (FID), which can better quantify both fidelity and perceptual quality than PSNR and SSIM [45]. LPIPS [50] is a full-reference metric designed to capture the perceptual quality of images. It measures the similarity between two images based on their perceptual features obtained by a pre-trained deep network which is the AlexNet model [16] in the experiments. FID [10] quantifies the similarity between the distributions of features extracted from a pre-trained Inception network for real and generated images. Lower LPIPS and FID values imply better results.

## 4.2. Comparison Results

The proposed method is compared with existing GAN-based and diffusion model-based RefSR methods on two datasets with two large scaling factors (i.e., 8× and 16×). The compared RefSR methods include five GAN-based methods (i.e., TTSR [43], WTRN [19], $C^2$-Matching [14], AMSA [40], and DATSR [3]), and two diffusion model-based RefSR methods (i.e., EDM* [15] and HSR-Diff [38]). EDM* represents that the original method is modified for the RefSR task, which combines LR and Ref images with the noise input. For a fair comparison, the LR images are synthesized by bicubic interpolation.

Table 1 shows the quantitative comparison results. Our method achieves the best LPIPS and FID performance in four sets of comparison experiments, demonstrating the advanced fidelity and perceptual quality of our results. According to the comparison results between GAN-based and diffusion model-based RefSR methods, the latter shows more powerful ability to bridge the resolution gap in large scaling factors. Owing to the utilization of the land cover changes priors, the proposed method performs consistently better than the other two diffusion model-based RefSR methods.

We further present the visual comparison in Figure 3. Figure 3(a) shows an example with slight changes on the SECOND 8× dataset. Our method can effectively transfer the relevant textures from the Ref image to the LR image in the unchanged areas, while the competing methods suffer from artifacts or blurred results. In the meantime, the vegetation reconstruction results are more realistic than comparison results in the changed areas. Figure 3(b) further demonstrates the effectiveness of vegetation reconstruction. Figure 3(c) shows an example with dramatic changes on the SECOND 16× dataset. The HR image contains a building where the non-vegetated ground surface is in the Ref image. The proposed method can guarantee the faithfulness of content reconstruction, while other methods cannot cope with this challenging scenario.

Similarly, Figure 3(d) shows an example with slight changes on the CNAM-CD 8× dataset. The texture of the building in our results is more realistic than other results, demonstrating the effectiveness of feature alignment and texture transfer in our method. Figure 3(e) shows an example with dramatic changes on the CNAM-CD 16× dataset. Our method correctly reconstructs the road and vegetation with the guidance of land cover change priors. The other two diffusion model-based RefSR methods produce illusory buildings and confusing layouts due to limited priors. Although remaining content fidelity, the GAN-based RefSR

Table 2. Ablation study of our method on SECOND 8× dataset. Lower LPIPS and FID values indicate better results.

| LR condition | Ref condition | Land cover change mask condition | Ref texture-guided SFT | Semantics-guided SFT | LPIPS↓ | FID↓ |
|---|---|---|---|---|---|---|
| ✓ | | | | | 0.3115 | 41.8340 |
| ✓ | ✓ | | | | 0.2886 | 34.5802 |
| ✓ | ✓ | ✓ | | | 0.2785 | 34.0638 |
| ✓ | ✓ | ✓ | ✓ | | 0.2709 | 33.7583 |
| ✓ | ✓ | ✓ | | ✓ | 0.2723 | 33.6805 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **0.2642** | **32.5961** |

Table 3. Results using land cover change predictions.

| Dataset | Land cover change mask | F1↑ | Precision↑ | Recall↑ | LPIPS↓ | FID↓ |
|---|---|---|---|---|---|---|
| SECOND 8X | GT | - | - | - | 0.2642 | 32.5961 |
| | Prediction | 87.72 | 86.41 | 86.30 | 0.2657 | 33.1453 |
| SECOND 16X | GT | - | - | - | 0.3433 | 33.9690 |
| | Prediction | 84.94 | 84.23 | 84.70 | 0.3404 | 35.0477 |
| CNAM-CD 8X | GT | - | - | - | 0.2791 | 43.0152 |
| | Prediction | 87.11 | 87.47 | 85.81 | 0.3159 | 48.3315 |
| CNAM-CD 16X | GT | - | - | - | 0.3519 | 45.6511 |
| | Prediction | 87.20 | 84.60 | 85.01 | 0.3889 | 56.7857 |

methods cannot reconstruct texture details, resulting in the spatial resolution gap between super-resolution results and HR images. Figure 3(f) also exhibits a dramatic change area on the CNAM-CD 16× dataset. The results of our method remain faithful to content reconstruction and can even remove the tree shadow in LR images.

In summary, the proposed method improves the content fidelity of reconstruction in changed areas and the effectiveness of texture transfer in unchanged areas, effectively building bridges across spatial and temporal resolutions.

### 4.3. Ablation Study

We perform ablation study on the SECOND 8× dataset to verify the effectiveness of the proposed method. In turn, we add the LR image, Ref image, and land cover change mask to the input of the conditional diffusion model. As shown in Table 2, using the Ref condition can largely improve the diffusion model ability in the super-resolution task, which is a promising way to narrow the gap between spatial resolutions in remote sensing scenarios. Taking the land cover change mask as a condition can further enhance the results. Still, the improvements are limited due to the simple combination between the conditions and the noise input.

We further conduct three experiments to verify the effectiveness of the semantics-guided SFT module, the Ref texture-guided SFT module, and the decoupled denoising strategy which uses both SFT modules. The results in Table 2 show that using either SFT module improves the denoising results because the guidance information further modulates the features in the decoder. Besides, building upon the two types of SFT modules, the decoupled denoising strategy with the iterative semantics-guided denoising and reference texture-guided denoising obtains the best results. Besides, we provide the ablation study of enhanced spatial feature transform module and results of real scenarios in the supplementary.

### 4.4. Experiments Using Land Cover Change Predictions

We conduct experiments to illustrate the impact of utilizing predicted land cover change masks. Specifically, we train change detection (CD) models based on ChangeFormer [2]. Combining a structured transformer encoder and an MLP decode, ChangeFormer is an ideal plug-and-play CD network for this work. Table 3 presents the quantitative comparison results using the prediction of land cover change masks. Although the performance is reduced compared to using ground truth (GT), our method still outperforms the comparison methods (refer to Table 1). This reinforces the effectiveness of using the proposed method in real scenarios.

### 4.5. Discussion of the Interaction between RefSR and CD Tasks

Ideally, accurate land cover change priors improve the confidence of Ref texture transfer in unchanged areas and content generation in changed areas. However, in practical scenarios, the utilization of land cover change priors may in-

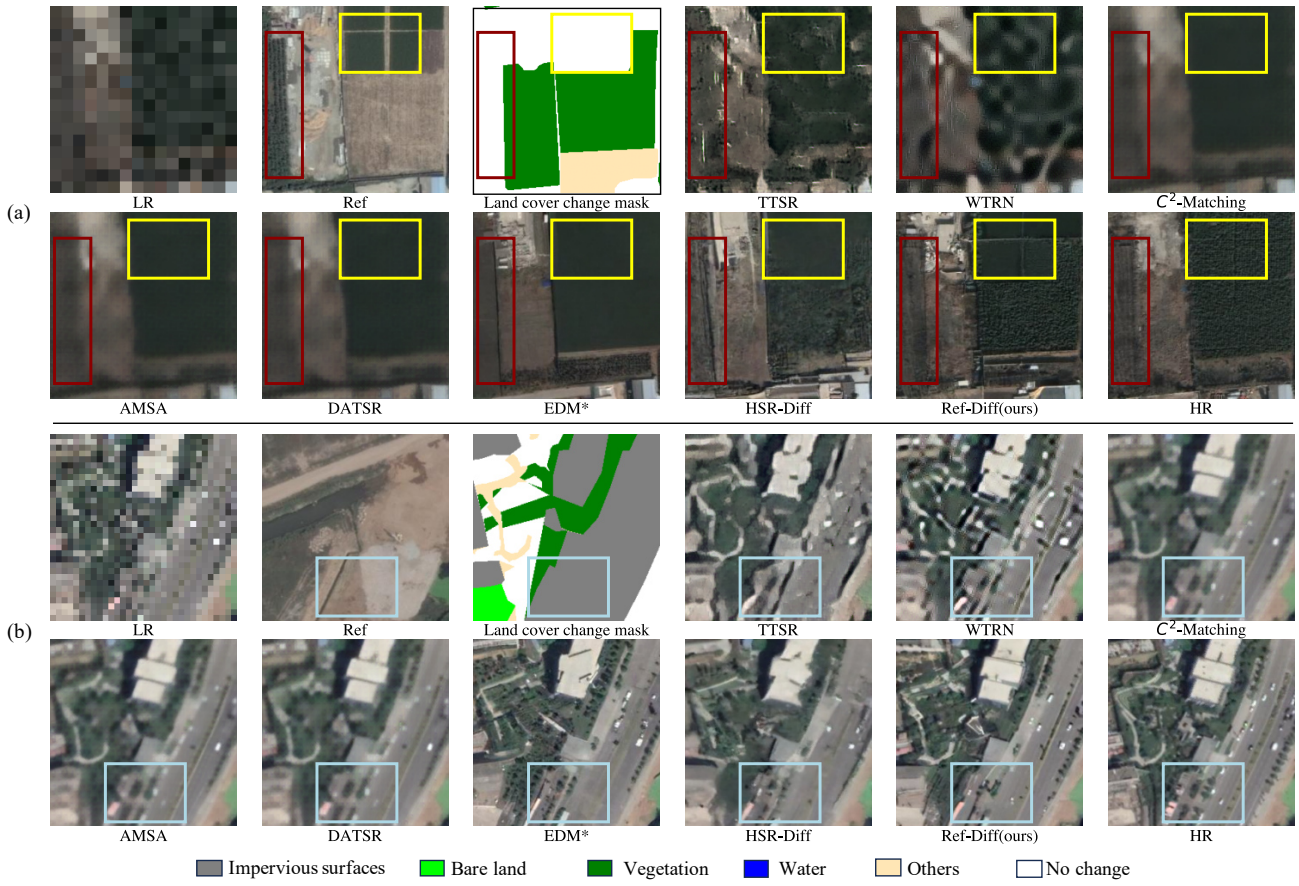


Figure 4. The results for two examples with mislabeled land cover change masks on CNAM-CD 8× and 16× datasets. (a) shows an example with false negative detection, and (b) shows an example with false positive detection.

troduce misleading information due to mislabeling issues or prediction errors in CD tasks. Figure 4 illustrates two common CD errors, i.e., false negatives (FN) and false positives (FP). The FN issue may lead to the introduction of false textures into the results, as depicted in Figure 4(a). Figure 4(b) presents an example of the FP issue, where vegetated areas are incorrectly labeled as impervious surfaces. The FP issue also undermines our results. Therefore, the improvement of CD accuracy will enhance the change-aware RefSR results. Moreover, a fine-grained classification system of land cover change will further facilitate the fidelity of content reconstruction in RefSR.

On the other hand, this work demonstrates the potential of change-aware RefSR in synthesizing well-labeled change detection data by semantic layout control and LR image collection. Consequently, RefSR and CD tasks can mutually reinforcement each other.

## 5. Conclusion

In this work, we propose a change-aware diffusion model for reference-based remote sensing image super-resolution to improve the faithfulness of content reconstruction and the effectiveness of texture transfer in large scaling factors. We inject the land cover change priors into the conditional diffusion model to explicitly guide denoising. With this powerful guidance, we decouple the semantic-guided denoising process in changed areas and the reference texture-guided denoising process in unchanged areas. We achieve the best quantitative and qualitative over state of the arts. This work also demonstrates the potential for mutual reinforcement between RefSR and change detection tasks. In future work, we will integrate change detection methods into the RefSR framework to enhance practicability.

## Acknowledgements

This research was supported in part by the National Natural Science Foundation of China (Grant No. T2125006 and No. 42301390), Jiangsu Innovation Capacity Building Program (Project No. BM2022028), China Postdoctoral Science Foundation (Grant No. 2023M731871), and Shuimu Tsinghua Scholar Project.

## References


[1] Masoomeh Aslahishahri, Jordan Ubbens, and Ian Stavness. Darts: Double attention reference-based transformer for super-resolution. *arXiv preprint arXiv:2307.08837*, 2023. 2

[2] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022. 7

[3] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European Conference on Computer Vision*, pages 325–342. Springer, 2022. 6

[4] Runmin Dong, Lixian Zhang, and Haohuan Fu. Rrsgan: Reference-based super-resolution for remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–17, 2021. 2

[5] Runmin Dong, Lichao Mou, Lixian Zhang, Haohuan Fu, and Xiao Xiang Zhu. Real-world remote sensing image super-resolution via a practical degradation model and a kernel-aware network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 191:155–170, 2022. 4

[6] Runmin Dong, Lichao Mou, Mengxuan Chen, Weijia Li, Xin-Yi Tong, Shuai Yuan, Lixian Zhang, Juepeng Zheng, Xiaoxiang Zhu, and Haohuan Fu. 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*, pages 16783–16793, 2023. 1

[7] Runmin Dong, Lixian Zhang, Weijia Li, Shuai Yuan, Lin Gan, Juepeng Zheng, Haohuan Fu, Lichao Mou, and Xiao Xiang Zhu. An adaptive image fusion method for sentinel-2 images and high-resolution images with long-time intervals. *International Journal of Applied Earth Observation and Geoinformation*, 121:103381, 2023. 1

[8] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2023. 2

[9] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 4

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[11] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 2

[12] Yixuan Huang, Xiaoyun Zhang, Yu Fu, Siheng Chen, Ya Zhang, Yan-Feng Wang, and Dazhi He. Task decoupled framework for reference-based super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5931–5940, 2022. 2

[13] Jiale Jiang, Qiaofeng Zhang, Xia Yao, Yongchao Tian, Yan Zhu, Weixing Cao, and Tao Cheng. 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. 1

[14] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2103–2112, 2021. 2, 6

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 3, 4, 6

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012. 6

[17] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2

[18] Shuangliang Li, Siwei Li, and Lihao Zhang. Hyperspectral and panchromatic images fusion based on the dual conditional diffusion models. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2

[19] Zhen Li, Zeng-Sheng Kuang, Zuo-Liang Zhu, Hong-Peng Wang, and Xiu-Li Shao. Wavelet-based texture reformation network for image super-resolution. *IEEE Transactions on Image Processing*, 31:2647–2660, 2022. 6

[20] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 3

[21] Mengxi Liu, Qian Shi, Andrea Marinoni, Da He, Xiaoping Liu, and Liangpei Zhang. Super-resolution-based change detection network with stacked attention module for images with different resolutions. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021. 3

[22] Mengxi Liu, Qian Shi, Jianlong Li, and Zhuoqun Chai. Learning token-aligned representations with multimodel transformers for different-resolution change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 2, 3

[23] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. 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*, pages 6368–6377, 2021. 2

[24] Yunan Luo, Kaiyu Guan, and Jian Peng. 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. 1

[25] ZhiYong Lv, HaiTao Huang, Xinghua Li, MingHua Zhao, Jon Atli Benediktsson, WeiWei Sun, and Nicola Falco. Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective. *Proceedings of the IEEE*, 2022. 3

[26] Xiaofeng Ma, Qunming Wang, Xiaohua Tong, and Peter M Atkinson. A deep learning model for incorporating temporal information in haze removal. *Remote Sensing of Environment*, 274:113012, 2022. 2

[27] Axi Niu, Pham Xuan Trung, Kang Zhang, Jinqiu Sun, Yu Zhu, In So Kweon, and Yanning Zhang. Acdmsr: Accelerated conditional diffusion models for single image super-resolution. *arXiv preprint arXiv:2307.00781*, 2023. 2

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 4

[29] Zhonghang Qiu, Huanfeng Shen, Linwei Yue, and Guizhou Zheng. Cross-sensor remote sensing imagery super-resolution via an edge-guided attention-based network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199: 226–241, 2023. 4

[30] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2

[31] Shuyao Shang, Zhengyang Shan, Guangxing Liu, and Jinglin Zhang. Resdiff: Combining cnn and diffusion model for image super-resolution. *arXiv preprint arXiv:2303.08714*, 2023. 2

[32] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[34] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 2

[35] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 2, 4

[36] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 4

[37] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 4

[38] Chanyue Wu, Dong Wang, Yunpeng Bai, Hanyu Mao, Ying Li, and Qiang Shen. Hsr-diff: hyperspectral image super-resolution via conditional diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2023. 2, 4, 6

[39] Hanlin Wu, Ning Ni, Shan Wang, and Libao Zhang. Blind super-resolution for remote sensing images via conditional stochastic normalizing flows. *arXiv preprint arXiv:2210.07751*, 2022. 2

[40] Bin Xia, Yapeng Tian, Yucheng Hang, Wenming Yang, Qingmin Liao, and Jie Zhou. Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2768–2776, 2022. 2, 6

[41] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. *arXiv preprint arXiv:2303.09472*, 2023. 2

[42] Mengze Xu, Jie Ma, and Yuanyuan Zhu. Dual-diffusion: Dual conditional denoising diffusion probabilistic models for blind super-resolution reconstruction in rsis. *arXiv preprint arXiv:2305.12170*, 2023. 2

[43] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020. 6

[44] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Semantic change detection with asymmetric siamese networks. *arXiv preprint arXiv:2010.05687*, 2020. 4

[45] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 3, 6

[46] Jiayang Zhang, Wanxu Zhang, Bo Jiang, Xiaodan Tong, Keya Chai, Yanchao Yin, Lin Wang, Junhao Jia, and Xiaoxuan Chen. Reference-based super-resolution method for remote sensing images with feature compression module. *Remote Sensing*, 15(4):1103, 2023. 2

[47] Lixian Zhang, Runmin Dong, Shuai Yuan, Weijia Li, Juepeng Zheng, and Haohuan Fu. Making low-resolution satellite images reborn: a deep learning approach for super-resolution building extraction. *Remote Sensing*, 13(15):2872, 2021. 1

[48] Lin Zhang, Xin Li, Dongliang He, Fu Li, Yili Wang, and Zhaoxiang Zhang. Rrsr: Reciprocal reference-based image super-resolution with progressive feature alignment and selection. In *European Conference on Computer Vision*, pages 648–664. Springer, 2022. 2

[49] Lixian Zhang, Shuai Yuan, Runmin Dong, Juepeng Zheng, Bin Gan, Dengmao Fang, Yang Liu, and Haohuan Fu. Swcare: Switchable learning and connectivity-aware refine-

ment method for multi-city and diverse-scenario road mapping using remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 127: 103665, 2024. 1

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[51] Jianwei Zheng, Yu Liu, Yuchao Feng, Honghui Xu, and Meiyu Zhang. Contrastive attention-guided multi-level feature registration for reference-based super-resolution. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(2):1–21, 2023. 2

[52] Xiangtao Zheng, Xiumei Chen, Xiaoqiang Lu, and Bangyong Sun. Unsupervised change detection by cross-resolution difference learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 2, 3

[53] Yanpeng Zhou, Jinjie Wang, Jianli Ding, Bohua Liu, Nan Weng, and Hongzhi Xiao. Signet: A siamese graph convolutional network for multi-class urban change detection. *Remote Sensing*, 15(9):2464, 2023. 6

[54] Han Zou, Liang Xu, and Takayuki Okatani. Geometry enhanced reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6123–6132, 2023. 2