# Supplementary Material for
# 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

## 1. Ablation Study of Enhanced Spatial Feature Transform Module

Existing approaches, such as the spatial feature transform (SFT) [4] and SPADE [2, 3], leverage an ideal semantic map to regulate denoising features through its embedding. To enhance the robustness of utilizing land cover change priors, we introduce an enhanced SFT module for semantic guidance and reference (Ref) texture guidance. The enhanced SFT module integrates guidance features with denoising features to regulate the latter. Additionally, we incorporate the low-resolution (LR) image with the land cover change mask for semantics-guided SFT, going beyond the use of only semantic embedding. The results in Table 1 illustrate the effectiveness of the enhanced SFT module in alleviating the negative impact of imprecise priors.

## 2. Experiments in Real Scenarios

To illustrate the effectiveness of the proposed method, we evaluate its performance in real scenarios using two distinct datasets located in areas different from the training datasets. The first dataset is situated in Jiaxing, China, employing a classification system identical to the SECOND dataset. The second dataset is derived from the HRSCD dataset [1], covering two regions in France (i.e., Rennes and Caen). The classification system of this dataset can be roughly mapped to that in the CNAM-CD dataset.Real LR and high-resolution (HR) images are collected from Google Earth Engine based on the geographical information of reference images. Real LR and HR images include Google Earth images with different levels of RGB bands. The resolution of real HR and Ref images is 0.5 meter. Note that the real LR images and real HR images may be captured by different sensors.

Table 1. Results using the original SFT and enhanced SFT modules. Bold indicates the best results.

| Method | LPIPS↓ | FID↓ |
|---|---|---|
| With the original SFT | 0.2725 | 33.9478 |
| With the enhanced SFT | **0.2642** | **32.5961** |

As depicted in Figure 1, our method outperforms competing RefSR methods, showcasing superior visual results on the two real datasets. It demonstrates the advanced fidelity and perceptual quality achieved by our method.

## 3. Further Analysis of the Proposed Method

We summarize the shortcomings of our method in three aspects. Firstly, it is challenging for the proposed method to reconstruct small objects such as vehicles. Secondly, our method cannot handle the super-resolution in substantial scaling factors (e.g., 32×). Finally, diffusion model-based RefSR is time-consuming compared with GAN-based RefSR methods. But this issue is promising to be optimized by recent studies on diffusion model acceleration such as [5].

In our future work, we will further explore the end-to-end method, integrating the change detection methods into RefSR models, to improve the practicality of our method. As analyzed in Section 4.4 in the main text, the simple two-stage strategy of cascading a change detector and our RefSR model does not fully unleash the capabilities of the proposed method in real applications. An end-to-end approach is expected to bridge this gap. Additionally, we will also apply different treatments to different categories of land cover

[*] Corresponding authors

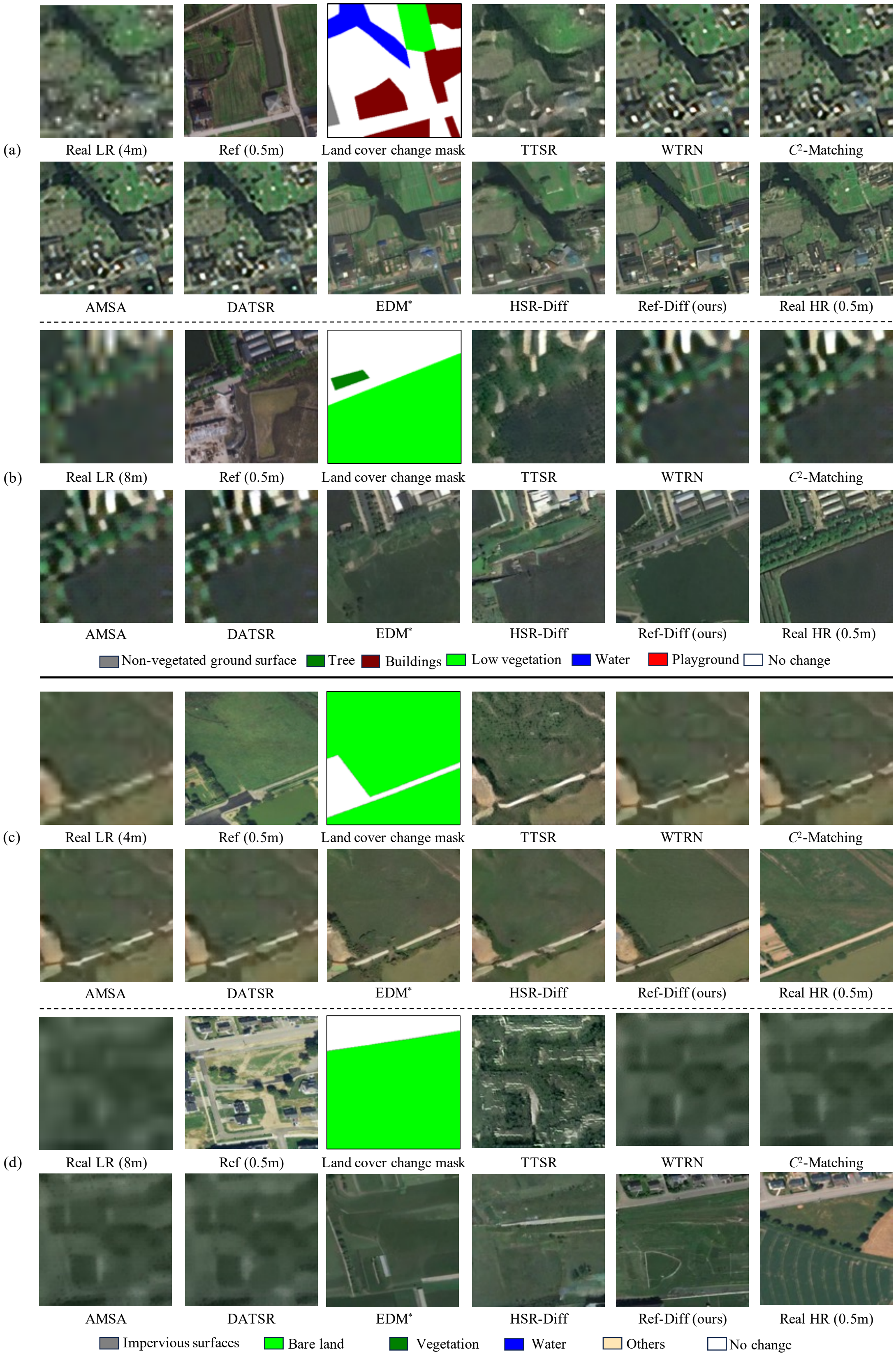

Figure 1. Comparison results on two real datasets. (a-b) are located in Jiaxing, China. (c-d) are located in Rennes and Caen, France. (a) and (c) are with 8× scaling factor. (b) and (d) are with 16× scaling factor.

change. Our method is designed to adaptively handle various types of changes. Given that the difficulty of reconstruction varies among land cover change types, introducing an explicit constraint or guidance could prove beneficial in further improving the results.

## References

[1] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019. 1

[2] 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. 1

[3] 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. 1

[4] 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. 1

[5] 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. 1