AN INVESTIGATION OF SUPER-RESOLUTION FOR CROSS-DOMAIN BUILDING EXTRACTION USING TRANSFORMER

Weitao Yue, Xiaowei Zhao*

Intelligent Control & Smart Energy (ICSE) Research Group, School of Engineering,
University of Warwick, Coventry, CV4 7AL, U.K. - (weitao.yue, xiaowei.zhao)@warwick.ac.uk

KEY WORDS: super-resolution, building extraction, transformer, ViT, remote sensing.

ABSTRACT:

The density of buildings is an important index to reflect the productivity and prosperity of an economic entity. Automatically monitoring the change and development of buildings through satellite can not only benefit the assessment of the status of urban development but also contribute to suburban construction planning. Apparently, more accurate building extraction performance can be guaranteed with higher-resolution remote sensing images. However, the desired high-resolution images are not always available limited by the remote sensing imaging technology and the expensive cost of updating the sensors and equipment. Therefore, the super-resolution technology, which aims at restoring the high-resolution images from the given low-resolution images, is a promising solution to resolve the dilemma. Therefore, in this paper, we investigate the potential application of super-resolution technology for cross-domain building extraction. The experiment results demonstrate that super-resolution can indeed improve building extraction accuracy.

1. INTRODUCTION

As an essential carrier of human productive activities, buildings have become one of the most changeable land use types (Hu et al., 2023). The dynamic information of buildings is beneficial for urban planning (Guo et al., 2021), map production (Lafarge et al., 2008), population statistics (Ji et al., 2019), and disaster assessment (Gupta and Shah, 2021). With more and more satellites have been launched worldwide in recent years, automatic extracting and monitoring of the change and development of buildings through remote sensing images become a feasible and efficient option (Chen et al., 2023).

In recent years, deep learning methods especially the convolutional neural network (CNN) represented by fully convolutional networks (FCN) (Long et al., 2015) have become the mainstream approach for building extraction from remote sensing images benefiting from their flexibility and adaptability (Ji et al., 2018). Since the pioneer FCN structure, the encoder-decoder structure for segmentation such as UNet (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2017) aiming at addressing the coarse-resolution segmentation of FCN-based networks was also introduced and improved for building extraction (Shi et al., 2022; Qiu et al., 2023; Deng et al., 2023).

The performance of these CNN-based methods, although promising, encounters bottlenecks in building extraction (Wang et al., 2022a). Specifically, CNN naturally lacks the capability for capturing long-range and non-local dependencies as it is originally designed to extract local patterns (Li et al., 2021b). However, in remote sensing images, buildings are normally in diverse appearances and are surrounded by complex backgrounds. Therefore, with only the local context, pixels will sometimes be ambiguous for identification, while the global context or long-range dependency can then provide extra information to determine the category, as illustrated in Figure 1.

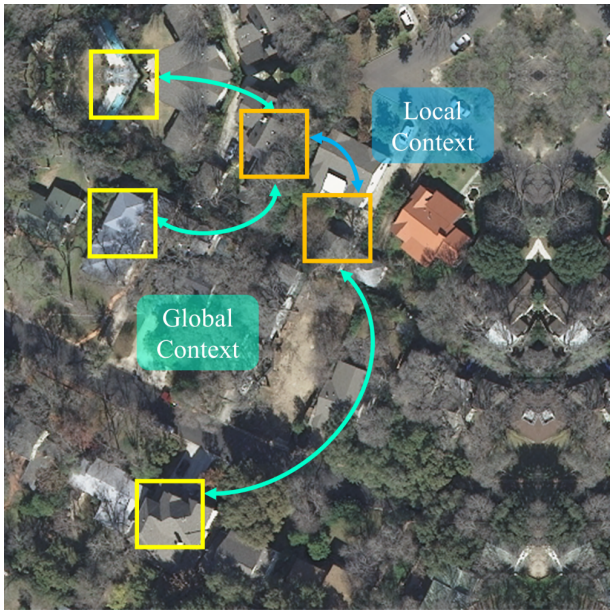

Figure 1. Illustration of the local context and global context. The squares represent the receptive field of the convolution operation. The orange regions represent the ambiguous building pixels where the local context is indistinguishable.

As a promising alternative to CNN, transformer (Vaswani et al., 2017), a novel structure originally designed for natural language processing (NLP), adopts the self-attention mechanism to extract global interactions between contexts. Recently, the vision transformer (ViT), a variant of transformer for computer vision, has shown its huge potential in enhancing vision-related tasks (Dosovitskiy et al., 2021; Zhu et al., 2021; Liu et al., 2021, 2022; Wang et al., 2022b). Compared with content-independent convolutional operations, attention weights of the self-attention blocks in the ViT are generated according to the relationship between contexts (Conde et al., 2023). Meanwhile, the long-

* Corresponding author

range dependency modelling is enabled by the shifted window mechanism embedded within the ViT (Liu et al., 2021, 2022).

Apparently, a remote sensing image with a higher resolution can provide more accurate global context information than a lower one, thereby guaranteeing more reliable building extraction performance. Even though more and more high-resolution images are available with the rapid development of remote sensing technologies, a constellation of satellites launched several years ago still continuously provides low-resolution but high-quality images. To fully utilize those valuable resources for building extraction, the super-resolution (SR) technology, which aims at reconstructing the high-resolution (HR) images from the given low-resolution (LR) images, is an encouraging solution (Dong et al., 2022).

For super-resolution, the revolutionary deep-learning-based methods have replaced traditional solutions such as prediction-based methods, patch-based methods, and edge-based methods (Wang et al., 2020). Since the pioneering Super-Resolution Convolutional Neural Network (SRCNN) proposed by (Dong et al., 2015), a series of novel super-resolution models have been developed sequentially to improve both the performance and efficiency (Li et al., 2019; Ji et al., 2020). In the wake of the successful application of the transformer in vision-related tasks, the transformer-based models have already demonstrated their great potential to enhance the super-resolution performance (Liang et al., 2021; Conde et al., 2023; Lei et al., 2021).

Currently, super-resolution technology has been adopted to boost the performance of image classification (Pang et al., 2019), object detection (Li et al., 2021a) and semantic segmentation (Zhang et al., 2021b). For building extraction, super-resolution-based methods can be generally divided into two kinds: the end-to-end approach (Zhang et al., 2021c; Xu et al., 2021) and the two-stage approach (Zhang et al., 2021a; Chen et al., 2023). However, all their methods are based on CNN structure, it is still unknown if a transformer-based framework can perform well for super-resolution-based building extraction.

In this paper, we investigate the potential combination of transformer-based super-resolution and building extraction models. Specifically, a two-stage framework is developed using the LSwinSR (Li and Zhao, 2023) for super-resolution and BuildFormer (Wang et al., 2022a) for building extraction. To comprehensively examine the potential super-resolution-based building extraction, the experiments of $\times 4$ upsample scales are conducted and CNN-based networks are also included in the comparison.

2. METHODOLOGY

The flowchart of the super-resolution-based building extraction framework is demonstrated in Figure 2. The low-resolution images are first up-sampled by the super-resolution model, whereafter the up-sampled images are used to train, validate and test the building extraction model.

2.1 Super-Resolution Models

This paper evaluates the Bicubic up-sampling method and a deep learning super-resolution model LSwinSR (Li and Zhao, 2023). LSwinSR was designed based on Swin Transformer (Liu et al., 2021) and SwinIR (Liang et al., 2021) with a three-step structure i.e. shallow feature extraction, deep feature extraction and image reconstruction. The first step utilizes a convolutional layer to process early visual features and expand the feature space. The shallow features are then processed by several Residual Linear Swin Transformer Blocks (RLSTB) and a convolution layer with the shifted window mechanism to obtain deep features. Finally, the shallow features and deep features are aggregated and fed into the reconstruction module for generating high-resolution images.

One major improvement in LSwinSR is the reduction in memory usage and computation time for super-resolution of large images by replacing the self-attention module with the kernel attention module in Swin Transformer. The self-attention module has quadratic computation complexity and memory requirements with window size, which is not suitable for large images that require a larger window size. On the other hand, the kernel attention module, which linearizes the softmax operation in the self-attention module, has only a linear increase in computation complexity and memory requirements with window size. The above changes enhance the efficiency of LSwinSR while maintaining competitive performance.

2.2 Building Extraction Models

In this paper, two deep-learning-based building extraction models are evaluated, i.e. ABCNet (Li et al., 2021b) and BuildFormer (Wang et al., 2022a). The success of building extraction algorithms depends on their ability to extract both local and global information. ABCNet utilizes an attentive bilateral network structure in its convolutional neural network to extract both types of features. This structure generates spatial paths to extract local features and contextual paths to extract global features, and an attention mechanism is used to capture rich contextual information. By integrating both low-level and detailed features from the spatial path and high-level and semantic features from the contextual path, ABCNet realizes a CNN-based building extraction model with competitive performance.

BuildFormer, on the other hand, is a dual-path variant of the Vision Transformer. It extracts spatial-detailed features through convolutional blocks and global context features through BuildFormer blocks with a window-based linear multi-head self-attention mechanism. Like ABCNet, it fuses the spatial-detailed and global context features to produce building extraction results. According to its claims, BuildFormer outperforms most CNN-based models for building extraction.

3. EXPERIMENTS AND DISCUSSIONS

3.1 Dataset

The Inria Aerial Image Labeling Dataset (Maggiori et al., 2017) is used to evaluate the different combinations of super-resolution and building extraction methods. The Inria dataset collects 360 high-resolution aerial images in 0.3m resolution from five cities (Austin, Chicago, Kitsap, Tyrol, and Vienna). In our experiment, the images of Austin, Chicago and Kitsap are taken to train super-resolution models, while the 1–5 tiles of each city remained to validate the performance and select the optimal model. After training and validation, the optimal model is then used to up-sample the images of Tyrol and Vienna, where the 1–5 and 6-10 tiles of each city are remained to test and validate models, respectively. For both super-resolution and building extraction, the original 5000×5000 images are first padded to 5120×5120 pixels and then cropped

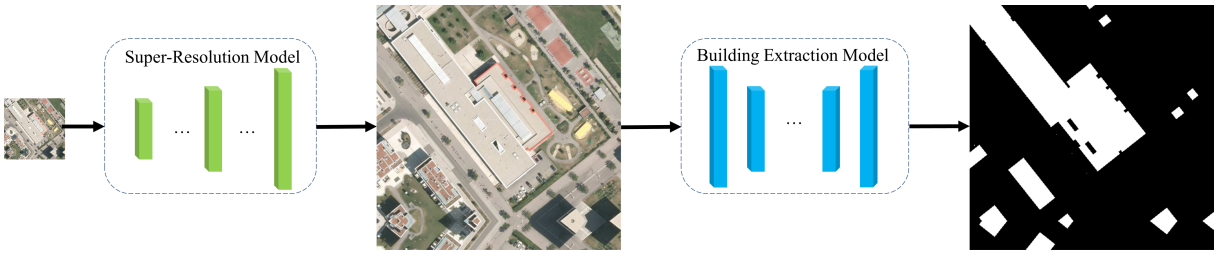

Figure 2. Flowchart of the two-stage super-resolution-based building extraction.

into 512 × 512 pixels image tiles. It is noteworthy that subsets for super-resolution and building extraction are collected in different cities. Therefore, the cross-domain capability of the super-resolution-based building extraction framework can be then verified.

3.2 Evaluation Metrics

Three frequently-used indexes are employed to evaluate the performance of the super-resolution result including the peak-signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM) and the mean absolute error (MAE). Thereafter, the building extraction model is trained, validated and tested based on the upsampled images generated from corresponding super-resolution algorithms, where the performance is measured by overall accuracy (OA), precision, recall, F1 score and intersection over union (IoU):

$$OA = \frac{TP}{TP + FP + TN + FN}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Recall}, \tag{4}$$

$$IoU = \frac{TP}{TP + FN + FP} \tag{5}$$

where TP, FP, and FN represent the true positive, the false positive, and the false negative, respectively.

3.3 Super-Resolution Experiments

Considering that the proposed model is a two-stage framework, the accuracy of the image super-resolution model in the first stage plays a crucial role in the performance of the entire model. In this paper, We evaluated two image super-resolution algorithms, Bicubic and LSwinSR, on the Ineria Aerial Image Labelling Dataset, with a super-resolution scale of ×4. The performance is evaluated based on the image quality evaluation metrics including PSNR, SSIM and MAE.

As shown in Table 1, LSwinSR, a transformer-based deep learning model, outperforms Bicubic significantly across all three evaluation metrics. For instance, on the training and validation sets, LSwinSR reaches 22.558 and 21.876 on PSNR and surpasses Bicubic by an obvious gap at about 2.73 and 2.81 respectively. Additionally, LSwinSR exceeds Bicubic on SSIM by more than 6.86% with lower MAE (at least 1.15% lower) on both training and validation sets. Higher PSNR and lower MAE indicate lower peak and mean error respectively while higher SSIM demonstrates a higher similarity between low-resolution images and super-resolution outputs. The outstanding performance of LSwinSR illustrates the effectiveness of both shallow and deep feature extraction through transformer-based deep-learning blocks for super-resolution.

Also, it is noted that compared with Bicubic, LSwinSR maintains obvious advances on test sets with respect to PSNR (2.44), SSIM (6.39%), MAE (1.13%). The similar gaps in the three evaluation metrics between training, validation and test sets imply the strong generalization and predictive ability of LSwinSR. The visual comparison in Figure 3 provides a more intuitive demonstration of the above analysis. As demonstrated in Figure 3, LSwinSR produces clear boundaries between buildings and other background structures while Bicubic only describes building with indistinguishable contours. These experiments validate the outstanding super-resolution performance of LSwinSR.

3.4 Building Extraction Experiments

This section evaluates the performance of building extraction based on the high-resolution images obtained by the super-resolution model in the section 3.3. In addition to using high-resolution images generated by Bicubic and LSwinSR, the original high-resolution images are also directly fed into the building extraction model as a comparison reference. Two building extraction models, namely BuildFormer and ABCNet, are evaluated in this section based on five evaluation metrics including Precision, Recall, F1, IoU and OA.

As listed in Table 2, the LSwinSR and BuildFormer combination achieves the highest IoU (81.13%), indicating the crucial role of super-resolution in the building extraction model and the success of the two-stage framework. Furthermore, LSwinSR super-resolution outperforms the Bicubic up-sampling method in improving building extraction performance. The advantage is obvious when ABCNet is utilized as the building extraction model. LSwinSR, compared with Bicubic, helps improve the performance of ABCNet in all five evaluation metrics, such as enhancing IoU from 76.14% to 77.24% and Recall from 83.71% to 84.70%. Also, replacing Bicubic with LSwinSR dramatically narrows the gap of evaluation metrics on ABCNet between super-resolution methods and the reference where original high-resolution images are directly input. For example, the F1 score gap is 1.59% between Bicubic and the reference while the number decreases to 0.88% when applying LSwinSR as the super-resolution method. Similarly, experiments with BuildFormer as the building extraction model have comprehensive performance improvements while the rises in evaluation metrics are relatively slight. The Precision error between LSwinSR and the reference remains at 0.63% although an improvement of 0.12% has been achieved.

Additionally, the experiments showed that transformer-based BuildFormer has comprehensive advantages over CNN-based

Model	Stage	PSNR	SSIM(%)	MAE(%)
Bicubic	Train	22.558	68.469	4.460
LSwinSR		25.293	75.330	3.345
Bicubic	Validation	21.876	67.383	4.590
LSwinSR		24.688	74.378	3.435
Bicubic	Test	22.316	67.230	4.907
LSwinSR		24.756	73.626	3.769

Table 1. Super-Resolution Comparison.

Figure 3. A visual comparison of super-resolution performance between low-resolution input (LR Input), Bicubic, LSwinSR, High-resolution images (Ground Truth) on training (top), validation (middle) and test (bottom) sets.

ABCNet in building extraction tasks. For example, BuildFormer achieved a leading IOU of 4.58%, 3.88%, and 4.22% than ABCNet respectively on images processed by Bicubic, LSwinSR, and the original high-resolution images. Similarly, when using the same source of high-resolution images, BuildFormer exceeded ABCNet in Precision, Recall, OA, and F1 by at least 1.88%, 2.90%, 1.19%, and 2.42%, respectively. This phenomenon is mainly due to two reasons. First, this suggests that Transformer-based methods have certain advantages over CNN-based methods. Second, ABCNet evaluated here has fewer parameter amounts than BuildFormer which may lead to worse performance of ABCNet. The visual comparison in Figure 4 further demonstrates the importance of the application of super-resolution. Compared with the original high-resolution image, building extraction with LSwinSR has less shape deformation. Also, BuildFormer performs better than ABCNet when processing the same input image. Among all results, the combination of LSwinSR and BuildFormer performs best, which validates the proposed two-stage framework using the transformer-based super-resolution model and building extraction model.

4. CONCLUSIONS

In this work, we investigated a two-stage building extraction framework based on LSwinSR and BuildFormer. The developed framework addressed the effectiveness of combining transformer-based super-resolution and building extraction models. Super-resolution and building extraction experiments were conducted separately on the Inria Aerial Labeling Dataset, which demonstrated superior building extraction performance with LSwinSR. Also, the super-resolution experiments demonstrated the dramatic enhancement of super-resolution by LSwinSR compared with the simple Bicubic interpolation. Furthermore, the building extraction experiments between ABCNet and BuildFormer indicated that transformer-based building extraction models always perform better than CNN-based methods.

In the future, we will further investigate the performance gaps on different super-resolution scales. In addition, the potential combination of other super-resolution models and building extraction models will be analyzed.

References

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2481–2495.

Chen, S., Ogawa, Y., Zhao, C., Sekimoto, Y., 2023. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195, 129–152.

Super-Resolution Model	Building Extraction Model	IOU	Precision	Recall	F1	OA
Bicubic	ABCNet	76.136	89.377	83.711	86.451	93.156
LSwinSR		77.243	89.768	84.700	87.161	93.491
High-Resolution		78.637	90.579	85.642	88.041	93.931
Bicubic	BuildFormer	80.725	91.531	87.241	89.335	94.566
LSwinSR		81.130	91.653	87.603	89.582	94.685
High-Resolution		82.866	92.280	89.039	90.630	95.198

Table 2. Building Extraction Comparison.

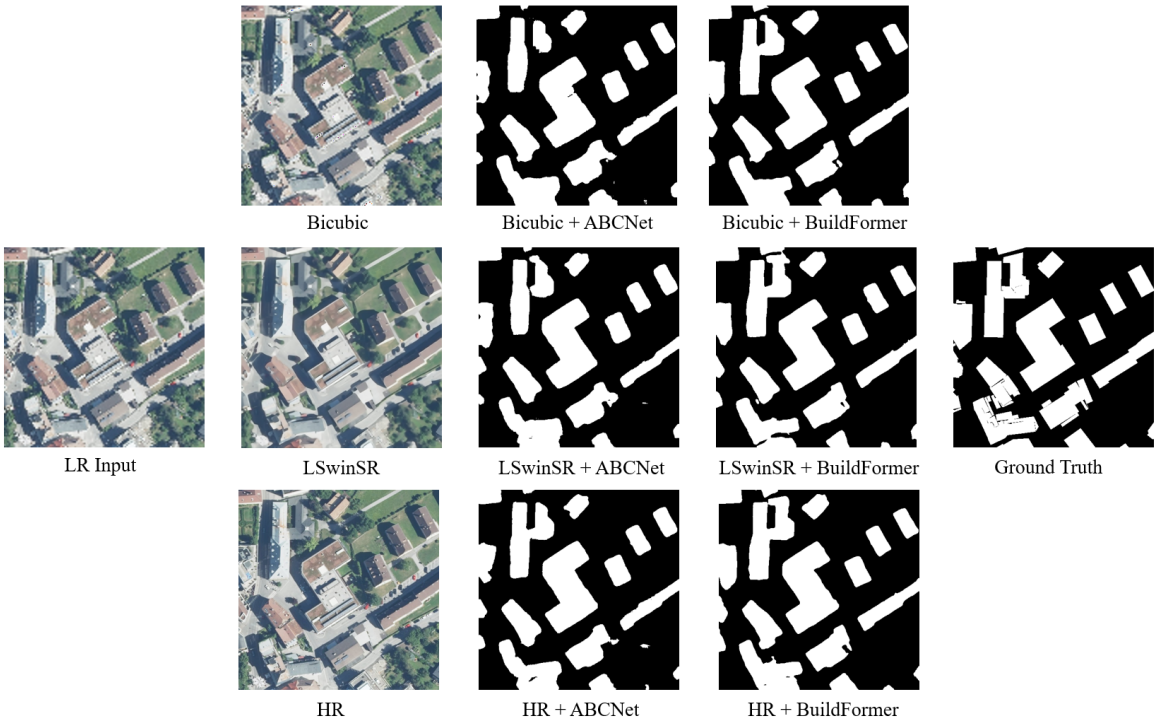

Figure 4. A visual comparison on building extraction results. Inputs include images from Bicubic, LSwinSR with $\times 4$ super-resolution and the original high-resolution image. The building extraction model includes ABCNet and BuildFormer.

Conde, M. V., Choi, U.-J., Burchi, M., Timofte, R., 2023. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, Springer, 669–687.

Deng, S., Wu, S., Bian, A., Zhang, J., Di, B., Nienkötter, A., Deng, T., Feng, T., 2023. Scattered Mountainous Area Building Extraction From an Open Satellite Imagery Dataset. *IEEE Geoscience and Remote Sensing Letters*, 20, 1–5.

Dong, C., Loy, C. C., He, K., Tang, X., 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2), 295–307.

Dong, R., Mou, L., Zhang, L., Fu, H., Zhu, X. X., 2022. 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.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Guo, H., Shi, Q., Marinoni, A., Du, B., Zhang, L., 2021. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. *Remote Sensing of Environment*, 264, 112589.

Gupta, R., Shah, M., 2021. Rescuenet: Joint building segmentation and damage assessment from satellite imagery. *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 4405–4411.

Hu, Y., Wang, Z., Huang, Z., Liu, Y., 2023. PolyBuilding: Polygon transformer for building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199, 15–27.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 574–586.

Ji, S., Wei, S., Lu, M., 2019. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *International journal of remote sensing*, 40(9), 3308–3322.

Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F., 2020. Real-world super-resolution via kernel estimation and noise injection. *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 466–467.

Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M., 2008. Automatic building extraction from DEMs using an object approach and application to the 3D-city modeling. *ISPRS Journal of photogrammetry and remote sensing*, 63(3), 365–381.

Lei, S., Shi, Z., Mo, W., 2021. Transformer-based multistage enhancement for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–11.

Li, J., Zhang, Z., Tian, Y., Xu, Y., Wen, Y., Wang, S., 2021a. Target-guided feature super-resolution for vehicle detection in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5.

Li, R., Zhao, X., 2023. LSwinSR: UAV Imagery Super-Resolution based on Linear Swin Transformer. *arXiv preprint arXiv:2303.10232*.

Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L., Atkinson, P. M., 2021b. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181, 84–98.

Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W., 2019. Feedback network for image super-resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3867–3876.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. Swinir: Image restoration using swin transformer. *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L. et al., 2022. Swin transformer v2: Scaling up capacity and resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

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

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 3226–3229.

Pang, Y., Cao, J., Wang, J., Han, J., 2019. JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. *IEEE Transactions on Information Forensics and Security*, 14(12), 3322–3331.

Qiu, W., Gu, L., Gao, F., Jiang, T., 2023. Building Extraction From Very High-Resolution Remote Sensing Images Using Refine-UNet. *IEEE Geoscience and Remote Sensing Letters*, 20, 1–5.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 234–241.

Shi, X., Huang, H., Pu, C., Yang, Y., Xue, J., 2022. CSA-UNet: Channel-Spatial Attention-Based Encoder-Decoder Network for Rural Blue-Roofed Building Extraction from UAV Imagery. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, L., Fang, S., Meng, X., Li, R., 2022a. Building extraction with vision transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–11.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P. M., 2022b. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 196–214.

Wang, Z., Chen, J., Hoi, S. C., 2020. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10), 3365–3387.

Xu, P., Tang, H., Ge, J., Feng, L., 2021. ESPC_NASUnet: An end-to-end super-resolution semantic segmentation network for mapping buildings from remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 5421–5435.

Zhang, L., Dong, R., Yuan, S., Li, W., Zheng, J., Fu, H., 2021a. Making low-resolution satellite images reborn: a deep learning approach for super-resolution building extraction. *Remote Sensing*, 13(15), 2872.

Zhang, Q., Yang, G., Zhang, G., 2021b. Collaborative network for super-resolution and semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–12.

Zhang, T., Tang, H., Ding, Y., Li, P., Ji, C., Xu, P., 2021c. FSRSS-Net: High-resolution mapping of buildings from middle-resolution satellite images using a super-resolution semantic segmentation network. *Remote Sensing*, 13(12), 2290.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2021. Deformable detr: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations*.