# Map-Guided Cross-Training for Building Detection

Anna Zawadzka, Mateusz Żarski, Kamil Drejer, Przemysław Głomb, Michał Romaszewski, Michał Cholewa

***Abstract*—Extraction of building footprints from freely available, high temporal resolution satellite imagery is important for applications ranging from disaster response to monitoring informal settlements. While Sentinel-2 is globally available, limited spatial resolution restricts the reliable detection of small or densely packed structures. We propose a novel framework that re-targets a super-resolution backbone (ESRGAN) to generate cartographic map-style outputs directly and fuses it with a semantic segmentation network through joint training. Unlike pipelines treating super-resolution and detection as sequential tasks, our approach leverages map-centric supervision, unifying enhancement and building delineation. For evaluation, we introduce the S2-BDOT-PL dataset, integrating Sentinel-2 imagery with OpenStreetMap and BDOT10k ground truth. Results demonstrate improved precision and robustness in small-building detection compared to native resolution baselines.**

***Index Terms*—remote sensing; satellite mapping of buildings; Sentinel-2; super-resolution; semantic segmentation;**

## I. Introduction

AUTOMATIC extraction of building footprints is increasingly described as a crucial functionality for urban planning, infrastructure management, and land-use policy [1]. For example, during disasters and conflicts, rapid detection of damaged or newly built structures guides relief and reconstruction; e.g. satellite analysis of the war in Ukraine allowed to estimate damage to over 236 000 homes and US $170 billion in infrastructure losses [2]. At the same time, over 1 billion people live in informal settlements whose roofs are often less than a pixel wide on a satellite image, making accurate building detection essential for humanitarian action and for monitoring illegal encroachment into protected ecosystems [3]. While there are many high spatial resolution building datasets and pretrained models, they are often proprietary, expensive and updated infrequently [4].

Reliable practical detection of buildings demands imagery that is accessible, globally available and frequently updated. Among the freely accessible sensors now in orbit, Sentinel-2 (S2) is particularly well suited, providing worldwide coverage and 5-day revisit time [5] which allows for near real-time monitoring and change detection [6]. S2 imagery has already been shown to reliably detect and delineate buildings within densely packed informal settlements, despite its moderate 10 m spatial resolution [7]. However, its native-resolution imagery still misses many small roofs and merges dense building clusters, so a combination of super-resolution and detection is essential for achieving desired reliability [5], [8].

The Enhanced Super Resolution Generative Adversarial Network (ESRGAN), first introduced by Wang et al. [9] as an upgrade to SRGAN [10], marked a significant milestone in visual fidelity of super-resolution through its use of Residual-in-Residual Dense Blocks and relativistic loss, winning the PIRM2018-SR Challenge. Real-ESRGAN has further improved robustness to real-world degradations by introducing a U-Net-based discriminator [11]. Since then, the ESRGAN and its variants have been widely recognized as efficient solutions for various applications, including remote sensing [12]. Comparative studies have shown the superiority of the ESRGAN and its variants for the task of building segmentation, particularly in terms of contour detail and building extraction [13], as well as in building count study [14].

Among the many segmentation models applicable to small-building detection, DeepLab stands out as a widely adopted benchmark model. DeepLabv3+ is a semantic segmentation architecture that combines atrous spatial pyramid pooling (ASPP) with a decoder module, originally evaluated on the Cityscapes dataset [15]. Since then, it has been widely applied to remote sensing tasks and it consistently demonstrates superior performance; e.g. in a recent evaluation DeepLabv3+ outperformed multiple competing architectures in the context of building extraction on Inria Aerial Image Labeling dataset [16]. For accurate building detection, combining ESRGAN with DeepLab into a processing pipeline aligns with the recent survey findings. For example, [17] provides a comprehensive review of building extraction advances and challenges, pointing to Deep Learning segmentation frameworks such as DeepLab as notable breakthroughs, and highlighting the critical role of GAN-based super-resolution techniques for efficient building analysis.

Recent research shows that joint architectures with fused training–i.e. joint training of models targeting separate objectives–often outperform the solutions with staged and separated pipeline, also in context of remote sensing tasks. For example, DeepSent [18] achieves superior results in super-resolving S2 images by jointly fusing multitemporal and multispectral data within a single model, rather than processing them sequentially. Although not previously explored, we argue that super-resolution and semantical segmentation for building detection should not be treated as isolated stages and that performing joint training of the two aforementioned state-of-art models instead can yield superior results.

While freely available sensors such as S2 deliver comprehensive and extensive global coverage, they remain too coarse to detail fine landscape structure, motivating task-oriented super-resolution that blends auxiliary geospatial data

A. Zawadzka, M. Żarski, P. Głomb, M. Romaszewski and M. Cholewa are with Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Gliwice, Poland. Corresponding author: A. Zawadzka (e-mail: azawadzka@iitis.pl.

K. Drejer is with Space Research Center of the Polish Academy of Sciences, Warsaw, Poland.

Manuscript received April 19, 2005; revised September 17, 2014.

to optimize imagery for specific analyses [19]. In one study, ESRGAN was enhanced by incorporating elevation data, improving realism in the reconstructed scenes [20]. Likewise, SEG-ESRGAN jointly upsamples S2 imagery to 2 m while performing semantic segmentation, showing how multi-task ESRGAN variants can yield high-resolution land-cover segmentation from coarse inputs [21]. For building detection, vector maps can serve as such auxiliary cue for enhancing their performance. Satellite-to-map translation itself has already proved viable even without an explicit SR step: HPix employs a hierarchical GAN to turn satellite scenes directly into high fidelity vector-tile maps [22], while MapGen-Diff uses a diffusion-bridge model to generate online maps with crisper boundaries and lower RMSE/SSIM than earlier approaches [23]. Recognizing the benefit of coupling super-resolution with cartographic guidance, the Semantic Guided Diffusion Model (SGDM) [24] injects vector maps as semantic priors into a diffusion-based pipeline to sharpen large-factor SR outputs. SGDM, however, still strives to reconstruct photo-realistic high-resolution imagery, treating the maps only as auxiliary cues. In contrast to the above approaches, we argue that the map itself should be the target. We therefore fine-tune ESRGAN for direct image-to-map translation, and fuse it with building detector model, all driven with detection loss. This turns cartographic data from a side-channel into the core supervision signal and thus moving the method beyond map-assisted SR to map-centric, task-driven enhancement.

In this article:

1) We propose a novel approach for adapting the super-resolution model for accurate map rendering. To the best of our knowledge, this is the first framework that retargets a super-resolution backbone to output full map-style representations directly, unifying enhancement and cartographic abstraction in one end-to-end pipeline.
2) Furthermore, we optimize the map-generation framework towards accurate generation of building footprints by fused training of the generative model jointly with segmentation model. Instead of optimizing the general visual fidelity, and thus risking suboptimal performance and resource utilization, we make our framework inherently task-oriented.
3) We build a dataset, S2-BDOT-PL, which includes data from Sentinel-2 (input), OpenStreetMap (OSM, training), and BDOT10k, Poland's official topographic vector database (training and verification). Such selection provides versatility, accessibility, and reliability: the OSM has global availability, while the BDOT10k, being sourced from the national Land and Building Register, ensures geodetic accuracy and enables rigorous testing.

## II. Map-guided cross-training combining super-resolution and detection

Our objective is to recover urban structures, building footprints in particular, from a low-resolution satellite data. To achieve that, we use building detection via semantic segmentation. To achieve robust detection on low-resolution images, we fuse together super-resolution and segmentation models.

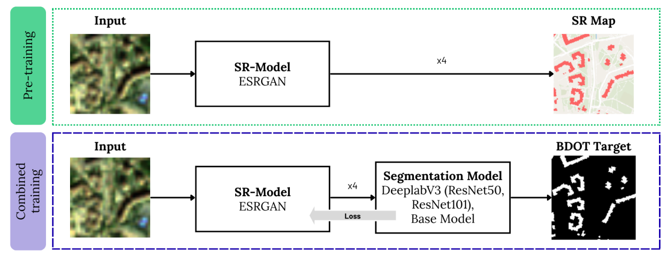



Fig. 1: Overview of the proposed joint training process for the combined model. In the pre-training phase, a super-resolution (SR) model is trained to upscale S2 imagery and produce cartographic map-like outputs. In the combined-training phase, the SR model is integrated with a segmentation model (e.g. DeepLabV3 with ResNet backbone), allowing gradients to flow through both networks to optimize segmentation performance on BDOT targets.

To improve the effectiveness of the process, we retask the super-resolution model (ESRGAN) to generate detailed map outputs of a given area based on a provided S2 image. The proposed method provides an effective approach to enhance spatial resolution and informativeness of S2 imagery.

### A. *Model setup and cross-training*

For the super-resolution task, we use an ESRGAN-based network, which incorporates a generator composed of Residual-in-Residual Dense Blocks (RRDB) without batch normalization, as introduced in [9]. The architecture further includes a U-Net discriminator, an enhancement proposed in [11]; the implementation is based on public repository [14].

The super-resolution model is pre-trained to generate 128×128 map tiles from corresponding 32×32 S2 inputs, using OSM patches as ground truth data (see Section II-B). Training is guided by a combination of L1 pixel loss, VGG19-based perceptual loss and vanilla adversarial loss. The model is optimized using Adam (learning rate $10^{-4}$, betas = 0.9, 0.99), with exponential moving average (EMA) applied to generator weights and USM sharpening applied to the ground truth images for L1 and perceptual losses.

For segmentation model, we experiment with two different architectures. First, we use the DeepLabV3 architecture [25], which combines a convolutional backbone with an Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale context. In our implementation, a ResNet backbone is used, and the model is trained to predict binary segmentation masks.

As a second model, we adopt a lightweight 'Base Model', to verify if the super-resolution backbone provides sufficiently rich features for binary segmentation; the simpler model also allows us to probe the backbone's robustness. Its architecture consisted of three one-layer convolution-deconvolution blocks with filter sizes $3\times3$ for maintaining the consistent feature map sizes, followed by batch normalization and ReLU activation function. The output layer was a single convolutional layer with a filter size $1 \times 1$ to obtain a finer pixel classification for small objects, and two output channels for each considered class, respectively.

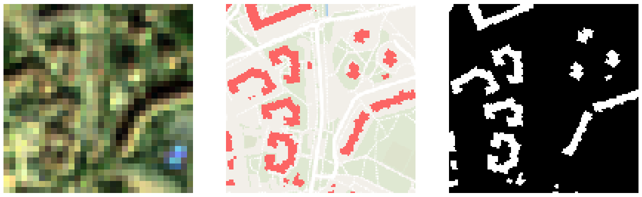


Fig. 2: S2-BDOT-PL dataset contents. Each data sample includes patches: Sentinel-2, OSM map and BDOT mask.

We first pretrained the super-resolution model (Fig. 1, 'Pre-training phase') for 1 850 000 iterations to obtain detailed map outputs. Then we combined it with the segmentation model in a way that allowed the gradient to flow through both segmentation and super-resolution model, thus rendering the super-resolution model as an extended feature extractor to the semantic segmentation model (Fig. 1, 'Combined-training phase'). The combined model was trained for 50 epochs and batch size of 32, and the target being the binary BDOT database. We used SGD optimizer along with initial learning rate of 0.001, weight decay of $10^{-4}$ and momentum of 0.9. We used a polynomial learning rate scheduler with exponent of 0.9. The loss function used for the training was weighted cross-entropy with weights proportional to class imbalance between background and building classes. Consequently, the model takes a S2 low-resolution $32 \times 32$ image as input and produces an upscaled $128 \times 128$ binary building mask.

### B. Dataset S2-BDOT-PL

We construct S2-BDOT-PL, a combined dataset that aligns S2 imagery, OSM cartographic renderings, and BDOT10k topographic vector database government reference data over a 653 km$^2$ area in Poland spanning urban and natural landscapes. Each sample, as shown in Fig.2, is a triplet: (i) a 32×32 S2 Level-1C RGB patch (10 m GSD; acquisition on September; no atmospheric correction), (ii) a 128×128 OSM raster map (buildings, roads, green areas, water; building footprints highlighted in red), and (iii) a 128×128 BDOT10k raster mask of building footprints. We use S2 + OSM pairs to train and evaluate the super-resolution model; BDOT10k masks serve as authoritative labels to train and validate the downstream building-detection model. In total, the dataset contains 6 372 triplets, split 80/10/10 into train/validation/test.

BDOT10k is Poland's official topographic database, compiled primarily from the national Land and Building Register (EGiB) and maintained under uniform standards, yielding consistency and high positional and attribute accuracy. A comparative study in Piaseczno County reported higher attribute and areal completeness for BDOT10k than OSM, with OSM coverage dropping in sparsely populated areas while urban location accuracy remained strong [26].

### C. Evaluation

Segmentation performance is evaluated per-pixel, comparing the predicted class to the actual one for each pixel. In our case, this is a binary classification problem, classes being building and non-building, and we compare predicted building mask against BDOT mask. As the data is highly imbalanced, we consider balanced accuracy $\frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i+FN_i}$ and mean Intersection over Union (IoU) $\frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i+FP_i+FN_i}$ as most informative metrics. $TP_i$, $FN_i$ and $FP_i$ denote the number of true positives, false negatives and false positives for class $i$, respectively, and $C$ is the total number of classes.

During training, the best checkpoint is saved for each experiment based on balanced accuracy calculated on the validation set. The models are then evaluated based on their respective best checkpoints, using a held-out test split and per-pixel evaluation metrics as above.

## III. Results

TABLE I: Performance comparison of different models. All training variants include pre-training of ESRGAN on OSM data to output enhanced resolution map.

| Model training | Balanced Acc. | Mean IoU |
|---|---|---|
| DeeplabV3 (ResNet50) reference | 0.874 | 0.479 |
| DeeplabV3 (ResNet50) + ESRGAN | **0.883** | **0.540** |
| DeeplabV3 (ResNet101) reference | 0.877 | 0.485 |
| DeeplabV3 (ResNet101) + ESRGAN | **0.891** | **0.523** |
| Base model reference | 0.862 | 0.512 |
| Base model + ESRGAN | **0.886** | **0.641** |

All experiments followed the unified setup described above. We thus examined three model pairs: DeepLabV3–ResNet50, DeepLabV3–ResNet101, and a lightweight Base Model. Each was evaluated in a separate, single-training setting and in a joint-training variant coupled with the super-resolution model (denoted '+ ESRGAN' in Table I). To assess the impact of joint training, we held all settings constant within each pair (i.e., same hyperparameters), varying only the number of epochs to accommodate convergence: joint-training models ran up to 50 epochs, standalone DeepLabV3 models 70 epochs, and the standalone Base Model 350 epochs–indicating faster convergence with the ESRGAN-augmented joint training.

Across all three model pairs, the joint-training variants with ESRGAN outperformed their single-training baselines in both balanced accuracy and mIoU (Table I). The highest balanced accuracy was achieved by DeeplabV3 (ResNet101) + ESRGAN at 0.891, followed by the Base Model + ESRGAN at 0.886 and DeeplabV3 (ResNet50) + ESRGAN at 0.883; by comparison, the strongest single training variant—DeeplabV3 (ResNet101)—reached 0.877, i.e., below even the weakest ESRGAN variant. For mIoU, the Base Model + ESRGAN yielded the best score (0.641), and even the lowest ESRGAN mIoU (0.523) exceeded the best single training mIoU (0.512, Base Model). These improvements hold regardless of backbone capacity and reflect fewer false positives and sharper boundaries (e.g., avoiding merges of adjacent buildings), indicating that the GAN-augmented joint training contributes complementary detail that benefits downstream segmentation. We note that while the Base Model did not match the DeeplabV3 (Resnet101) architectures in balanced accuracy alone, its ESRGAN-augmented variant achieved the highest mIoU of all tested configurations, indicating the high informativeness of map-guided cross-training, that can be exploited with simple models.

Statistical tests were performed to verify if the observed differences in balanced accuracy between the GAN-augmented setup and the baseline were significant. Statistical tests (Wilcoxon signed-rank, one-sided) confirmed that for each model pair the GAN-augmented variant significantly outperformed its baseline counterpart ($p < 0.0001$ in all cases).

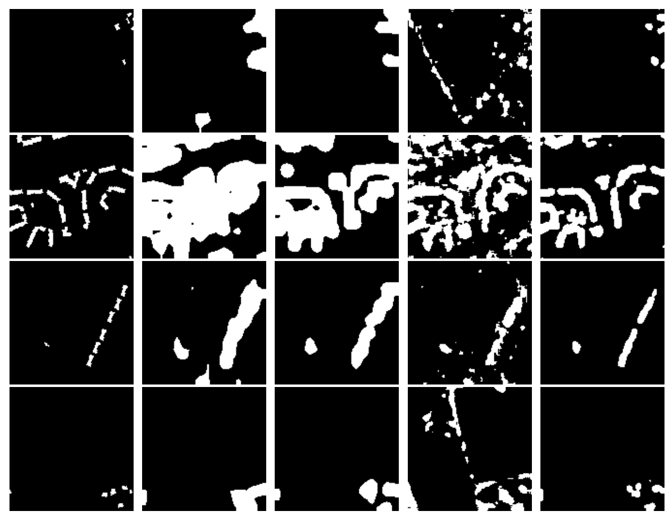
Inspecting output masks, we observe that DeepLabV3-based models tend to produce merged, spread, blob-like segmentation masks, particularly in dense built-up regions. These results highlight that a model can achieve high balanced accuracy even when building boundaries are poorly delineated, because false positives and over-segmentation are not strongly penalized by that metric (by contrast, mIoU aims to spatially align predictions with the ground-truth regions and, in our case, tracks the visual quality more closely). The inclusion of ESRGAN in joint training appears to improve this aspect. Predictions from this model in its current form seem to be better suited to indicating the presence and density of built-up areas, rather than counting individual buildings. Such setup is efficient for many tasks, i.e. change detection (assessment of disaster damages, urban expansion monitoring), land-use classification, risk assessment. The lightweight Base Model produces tighter, better-overlapping masks but occasionally misclassifies roads or tree canopies as buildings. Adding joint training with the ESRGAN improves this behavior: it reduces false positives, sharpens edges, and 'cleans up' merged blobs– both for the Base Model and for the DeepLabV3 backbones– yielding more conservative, well-localized delineations. This encourages further exploration of simpler architectures efficiency, especially for building count problems and with training set to optimize mIoU. Additionally, we notice that after joint training with the segmentation model, the ESRGAN output shifted from enhancing the entire scene, including roads and vegetation, to emphasizing building structures, effectively shifting its focus toward task-relevant details. This suggests that the ESRGAN's focus narrows over time, implicitly aligning with the final objective.

Overall, the ESRGAN-augmented variants shift predictions from blob-like sensitivity toward precise footprints, explaining stronger mIoU despite smaller (but consistent) difference in balanced accuracy; checkpoints were selected by balanced accuracy, yet the qualitative improvements and mIoU gains remain consistent (see Fig. 3 and 4). While the ESRGAN-augmentation introduced additional architectural component, the joint training for segmentation required much fewer epochs to converge compared versus the baseline one. While the ESRGAN pre-training for map generation can be time and resources-consuming, extending the architecture with ESRGAN improves the result significantly, justifying the additional computational cost.

(a) Sentinel-2 (the input) (b) BDOT (GT) (c) OSM map (d) ESRGAN (before joint training) (e) ESRGAN (after joint-training)

(f) DeeplabV3 (Plain) (g) DeeplabV3 + ESRGAN (h) Base Model (Plain) (i) Base Model + ESRGAN

Fig. 3: Qualitative comparison of outputs: Sentinel-2 satelllite image (a), BDOT mask (b), OSM map ground truth (c), ESRGAN output before joint training (d), ESRGAN output after joint training (e), predictions (the output) from: baseline DeeplabV3 (f) and its ESRGAN-augmented variant (g), Base model (h) and Base model with ESRGAN (i).



(a) BDOT (GT) (b) DeeplabV3 (Plain) (c) DeeplabV3 + ESRGAN (d) Base Model (Plain) (e) Base Model + ESRGAN

Fig. 4: Four examples demonstrating the impact of ESRGAN on classifier outputs, with a visible reduction in false positives.

## IV. Discussion

Our results are consistent with the observations in [8] and [5], that dense clusters of small, individual buildings are sometimes misidentified as continuous regions. However, the IoU results in Table I, confirmed by visual inspection, indicate that this phenomenon is considerably less pronounced when using ESRGAN augmentation (See Fig. 4). Moreover, the small differences in IoU scores observed between the tested DeeplabV3 encoders are consistent with the 2% variations reported in [5]. In contrast, the sharp $\sim$10% improvement in the base model's accuracy may result from the DeeplabV3 encoders being pretrained on 10 m imagery, which makes them less adaptable to 2.5 m upsampled data.

Our results show significant improvement of performance (up to 10% IoU) introduced by applying super-resolution, the difference is more pronounced than e.g. in [8] (approx. 2%). We attribute this to the way super-resolution is applied in our approach: instead of simply upsampling the input, it generates a map-style representation, which appears to be critical for achieving improved results.

In addition to our existing comparisons, our results show map-fidelity after pretraining comparable to [27]. On balanced accuracy we are well above all baselines reported in [28] and match one of six cases, despite their much larger, multi-seasonal training dataset and significantly less complex ground truth structure. Our IoU is comparable to [29], exceeding some of their scenarios.

Although the BDOT/OSM ground truth is generally accurate, we identified a subset of images containing visible objects absent from the ground truth but classified as buildings. While this may affect model performance, it also highlights the potential of applying a similar pipeline for change detection [6].

To conclude, we show that pairing Sentinel-2 with map-guided super-resolution and joint training yields reliable small-building detection at scale; it improves both balanced accuracy and mIoU, with tighter boundaries and fewer false positives. Promising directions include instance-level evaluation and counting, boundary-aware objectives and checkpointing (optimizing for mIoU), temporal modeling for change detection, and cross-regional transfer (e.g., adding Sentinel-1). Together, these steps can turn our evaluation into an operational, low-cost pipeline for built-environment monitoring.

## Acknowledgments

The authors would like to thank Stanisław Lewiński and Marek Ruciński from the Space Research Center of the PAS for help in data preparation.

## References

[1] P. Dabove, M. Daud, and L. Olivotto, "Revolutionizing urban mapping: deep learning and data fusion strategies for accurate building footprint segmentation," *Scientific Reports*, vol. 14, no. 1, p. 13510, 2024.

[2] H. Chen, J. Song, O. Dietrich, C. Broni-Bediako, W. Xuan, J. Wang, X. Shao, Y. Wei, J. Xia, C. Lan, K. Schindler, and N. Yokoya, "BRIGHT: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response," *Earth System Science Data Discussions*, vol. 2025, pp. 1–51, 2025.

[3] N. Büttner, S. Stalder, M. Volpi, E. Suel, and K. Harttgen, "Large-scale slum mapping in sub-saharan africa's major cities: Remote sensing and deep learning reveal strong slum growth in the urban periphery between 2016 and 2022," *Habitat International*, vol. 161, p. 103403, 2025.

[4] M. Debella-Gilo, "Relative performance of super-resolved sentinel-2 and copernicus vhr images in mapping built-up areas and building footprints using deep learning," *European Journal of Remote Sensing*, vol. 58, no. 1, p. 2517381, 2025.

[5] S. Illarionova, D. Shadrin, I. Shukhratov, K. Evteeva, G. Popandopulo, N. Sotiriadi, I. Oseledets, and E. Burnaev, "Benchmark for building segmentation on up-scaled sentinel-2 imagery," *Remote Sensing*, vol. 15, no. 9, p. 2347, 2023.

[6] A. M. Pacheco-Pascagaza, Y. Gou, V. Louis, J. F. Roberts, P. Rodríguez-Veiga, P. da Conceição Bispo, F. D. Espírito-Santo, C. Robb, C. Upton, G. Galindo *et al.*, "Near real-time change detection system using sentinel-2 and machine learning: A test for mexican and colombian forests," *Remote Sensing*, vol. 14, no. 3, p. 707, 2022.

[7] B. J. Gram-Hansen, P. Helber, I. Varatharajan, F. Azam, A. Coca-Castro, V. Kopackova, and P. Bilinski, "Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 361–368.

[8] T. Zhang, H. Tang, Y. Ding, P. Li, C. Ji, and P. Xu, "Fsrss-net: High-resolution mapping of buildings from middle-resolution satellite images using a super-resolution semantic segmentation network," *Remote Sensing*, vol. 13, no. 12, p. 2290, 2021.

[9] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 63–79.

[10] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1905–1914.

[12] Ş. B. Çetin, "Real-esrgan: A deep learning approach for general image restoration and its application to aerial images," *Advanced Remote Sensing*, vol. 3, no. 2, pp. 90–99, 2023.

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

[14] A. K. Piper Wolters, Favyen Bastani, "Zooming out on zooming in: Advancing super-resolution for remote sensing," *arXiv preprint arXiv:2311.18082*, 2023.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[16] S. Ekiz and U. Acar, "Improving building extraction from high-resolution aerial images: Error correction and performance enhancement using deep learning on the inria dataset," *Science Progress*, vol. 108, no. 1, p. 00368504251318202, 2025, pMID: 39943714.

[17] D. Yang, X. Gao, Y. Yang, K. Guo, K. Han, and L. Xu, "Advances and future prospects in building extraction from high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 6994–7016, 2025.

[18] T. Tarasiewicz, J. Nalepa, R. A. Farrugia, G. Valentino, M. Chen, J. A. Briffa, and M. Kawulok, "Multitemporal and multispectral data fusion for super-resolution of sentinel-2 images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–19, 2023.

[19] E. Alvarez-Vanhard, G. F. Garcia, and T. Corpetti, "Super-resolution by fusing multispectral and terrain models: Application to water level mapping," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[20] S. Karatsiolis, C. Padubidri, and A. Kamilaris, "Exploiting digital surface models for inferring super-resolution for remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[21] L. Salgueiro, J. Marcello, and V. Vilaplana, "SEG-ESRGAN: A multi-task network for super-resolution and semantic segmentation of remote sensing images," *Remote Sensing*, vol. 14, no. 22, 2022.

[22] A. Taparia and K. Nath, "HPix: Generating vector maps from satellite images," 2024, arXiv:2407.13680.

[23] J. Tian, J. Wu, H. Chen, and M. Ma, "MapGen-Diff: An end-to-end remote sensing image to map generator via denoising diffusion bridge model," *Remote Sensing*, vol. 16, no. 19, 2024.

[24] C. Wang and W. Sun, "Semantic guided large scale factor remote sensing image super-resolution with generative diffusion prior," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 220, pp. 125–138, 2025.

[25] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587v3*, 2017.

[26] S. Borkowska, E. Bielecka, and K. Pokonieczny, "Openstreetmap – building data completeness visualization in terms of "fitness for purpose"," *Advances in Geodesy and Geoinformation*, vol. vol. 72, no. No 1, p. article no. e35, 2023.

[27] C. Wang and W. Sun, "Semantic guided large scale factor remote sensing image super-resolution with generative diffusion prior," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 220, pp. 125–138, 2025.

[28] T. Hu, M. Han, and Z. Guo, "A seasonal-series LSTM network for irregular urban function zone recognition using Sentinel-2 images," *Applied Soft Computing*, vol. 172, p. 112876, 2025.

[29] Y. Wu, X. Li, Y. Gong, and D. Mwaniki, "A flexible framework for identifying urban villages using Sentinel-2 observations and deep learning," *International Journal of Applied Earth Observation and Geoinformation*, vol. 141, p. 104631, 2025.