A Multi-modal Geospatial-imagery Fusion Framework for Urban Agricultural Land Identification: A Case Study of Nanjing, China

Caiqiong Huang, *Yi* Qi*, *Haiwei* Yin, *Duoduo* Wang and *Zuqiang* Zhang

Nanjing University, School of Architecture and Urban Planning, 22 Hankou Road, Nanjing 210093, PR China

Abstract. Urban agriculture (UA) provides ecological, social, and economic co-benefits in compact cities, but accurately identifying UA land remains a challenge due to small patch sizes, fragmented patterns, and mixing with other urban land covers. We develop a multi-modal geospatial-imagery fusion framework that integrates high-resolution satellite imagery with multi-scale geospatial indicators, including points of interest (POIs), population, transportation, hydrology, and nighttime light data to improve UA land identification. Two imagery-only baselines are extended into fusion variants via skip-level and mid-layer integration. Using Nanjing, China, as the testbed, our experiments show that fusion significantly improves overall accuracy (OA) and mean Intersection-over-Union (mIoU) relative to imagery-only baselines. Channel-wise zero-out ablation further reveals the importance of transport-related POIs (especially at 1000 m), large-scale population distribution (1500 m), and neighborhood-scale public services (500 m). Applying the optimized fusion model to a 50 km strip across the urban-suburban-rural gradient uncovers a spatial transition: from fragmented, embedded UA patches in the core city to increasingly continuous and stable belts toward rural peripheries. The approach enhances both accuracy and interpretability, providing actionable evidence for UA spatial optimization in territorial planning and urban renewal.

1 Introduction

Urban agriculture (UA) has increasingly been recognized as a critical component of sustainable urban development, contributing to ecosystem resilience, food system sustainability, microclimate regulation, and social well-being [1]. In dense metropolitan contexts, however, UA is typically small-scale, fragmented, and embedded within complex urban fabrics, complicating accurate identification and management.

In recent years, the wide availability of high-resolution remote sensing imagery and large-scale geospatial data has enabled deep learning – particularly semantic segmentation networks – has been extensively applied to land use classification and urban functional zone identification [2-5]. Unlike coarse-grained land use types or functional zones delineated by

* Corresponding author: jnjn@163.com

road networks, urban agricultural land requires more fine-grained recognition. Advanced semantic segmentation models such as U-Net, DeepLabv3+, and transformers have demonstrated outstanding performance in fine-scale image recognition tasks. For example, Herlawati et al. utilized U-Net and DeepLabv3+ for land use classification in Karawang and Bekasi (Indonesia) and found that DeepLabv3+ outperformed U-Net in both speed and accuracy [6]. Likewise, Feng et al. improved DeepLabv3+ with multi-scale attention modules (ResNeSt and SENet) to better distinguish tailings ponds from other land covers, significantly enhancing recognition of these fine-grained features [7]. These studies underscore the effectiveness and potential of deep semantic segmentation models for detailed image-based land use identification.

Despite these advances, most current research has focused on models built exclusively on image features, with relatively little attention to the integration of ancillary geospatial features or the analysis of their role in the model's decision mechanism. Integrating multi-source data to improve model performance has only recently begun to receive attention. For instance, Bao et al. (2020) proposed a deep feature convolutional neural network (DFCNN) that combines high-resolution imagery with POI data, achieving a land use classification accuracy of 96.65% [8]. Xie et al. (2025) developed a multi-source scene feature fusion method based on a transformer, using graph convolutional networks to extract structural features from POI data and a ResNet-50 encoder for remote sensing features, then fusing them via multi-head attention [9]. This approach addressed the challenge of integrating diverse data sources and provided a new pathway for urban land use identification.

Another line of research highlighting the importance of scale is the multiscale geographically weighted regression (MGWR), an extension of traditional GWR that allows each explanatory variable to have a different spatial bandwidth [10]. MGWR has been used to reveal spatial heterogeneity in relationships such as the built environment's impact on urban vitality and the multiscale relationship between population distribution and land-use mix [11, 12]. These studies indicate that considering multi-scale geospatial characteristics enables different variables to exert influence at appropriate spatial extents, thereby capturing the drivers of urban spatial patterns more realistically. This concept forms the theoretical basis for our proposed fusion of image features with multi-scale spatial features.

Based on the above background, we propose a multi-modal fusion framework for urban agricultural land identification and spatial modeling that seeks to overcome the limitations of using only remote sensing imagery. The main contributions of this research are as follows:

I. Benchmark two imagery-only baselines: We employ two semantic segmentation models, U-Net and DeepLabv3-ResNet50 with ImageNet pretraining, as baseline models to evaluate their performance in urban agriculture land use identification. This comparison highlights the advantages of deeper networks and context modules for complex urban scenes.

II. Design two fusion variants that inject aggregated geospatial features: We develop improved frameworks that fuse remote sensing image features with multi-scale aggregated geospatial features (such as facility POIs, population, transportation, hydrology, and nighttime light). Two fusion strategies – multi-scale skip connections and mid-level feature fusion – are implemented for U-Net and DeepLabv3-ResNet50 respectively, effectively coupling imagery with spatial context to improve classification accuracy.

III. Quantify contributions of individual geospatial channels: We introduce a channel-wise zero-out ablation method to quantify the contribution of each multi-scale geospatial feature in the identification task. This approach provides insight into the model's decision mechanism by analysing performance changes when each feature channel is masked.

IV. Apply the optimized model to predict UA distribution: Using Nanjing as the study area, we apply the improved model to high-resolution imagery and multi-scale geospatial data at an urban–suburban–rural strip, revealing characteristic spatial gradients. The results provide methodological insights for multi-source data fusion in urban land classification, and the effectiveness and planning applicability of the proposed modeling approach are validated.

2 Materials and methods

2.1 Study area and data

Nanjing is a major city in eastern China located in the lower Yangtze River Basin. The municipal territory covers approximately 6,587 km², with an officially delineated urban development boundary (UDB) of 1,492.53 km² (accounting for approx. 22.7%). In recent years, as urbanization enters a later stage, Nanjing has seen the emergence of urban agricultural land parcels. It appears as scattered or belt-like patches near residential, industrial, transport corridors, vacant construction land, and waterfronts. To validate model applicability across urbanization gradients, we further construct an approximately 50 km long strip (width ≈ 250 m) traversing core-urban, suburban, and rural zones. This setting provides a robust testbed for the multimodal model under different degrees of spatial heterogeneity.
The data used in this study include three main components: remote sensing imagery, geospatial features, and labeled urban agriculture samples (Fig. 1).

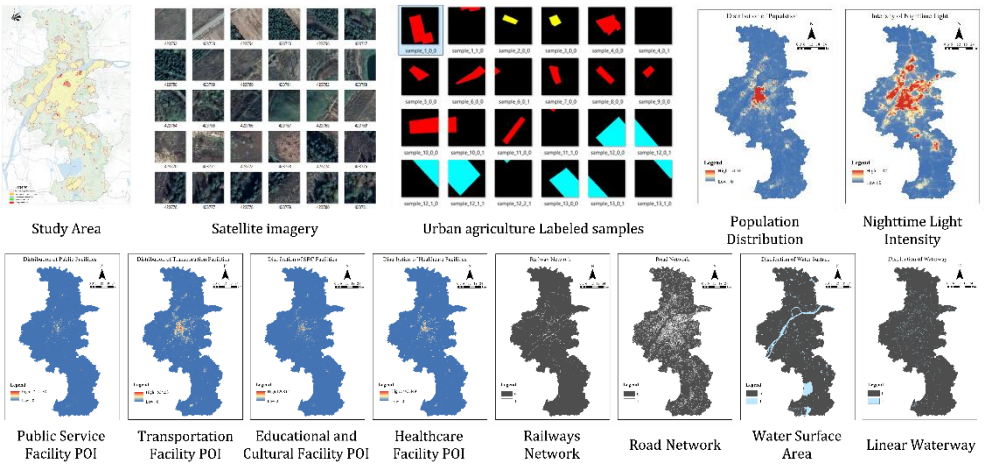

Fig. 1. Experimental data

2.1.1 Satellite imagery

We obtained high-resolution satellite images (Level-20 tiles from Google Maps, <1 m resolution) covering the entirety of Nanjing, approximately 11.84 million tiles. These tiles were mosaicked, registered, and clipped as needed (128 × 128). The imagery (RGB bands) serves as the primary input, providing fine-grained spectral and textural information of land cover.

2.1.2 Geospatial indicators

To capture the spatial heterogeneity influencing urban agriculture distribution, we gathered a variety of ancillary geospatial datasets, all of which were rasterized and aggregated using a multiscale geospatial aggregation tool. Specifically, spatial features were aggregated at four spatial extents: (1) the within-sample extent, where features were aggregated only within the spatial extent corresponding to each labeled sample unit, without additional spatial buffering; (2) a small-scale neighborhood using a 500 m circular buffer; (3) a medium-scale neighborhood using a 1000 m buffer; and (4) a large-scale neighborhood using a 1500 m buffer, thereby producing multi-scale feature rasters. For example, this process yielded features such as the count of each POI category, population sum, total road length, total river length, water area, and average nighttime light intensity within the sample extent, surrounding neighborhoods of 500 m, 1000 m, and 1500 m, centered on each sample location. These raster layers were spatially aligned with the 128 × 128 pixel image patches through resampling or clipping and normalized to unit scale, forming the geospatial feature channels used as model inputs.

I. Facilities POI data: Points of interest in four categories – healthcare (e.g., hospitals), transportation facilities, educational & cultural (e.g., schools), and public service facilities (e.g., nursing homes) – were acquired for Nanjing (2023) from the AMAP API. These facilities are included because they often contain idle vacant land that can be repurposed; for instance, the open spaces surrounding nursing homes are frequently utilized by the elderly to cultivate vegetables for daily consumption.

II. Population: A gridded population dataset for 2023 at 1 km resolution was obtained from the LandScan database, representing the spatial distribution of population and thus potential human activity intensity.

III. Transportation networks: 2025 vector data for railways and roads in Nanjing were retrieved from OpenStreetMap, characterizing transportation accessibility and locational factors.

IV. Hydrology: Data on water bodies (surface water area) and river networks (linear waterway length) in 2025 were also obtained from OpenStreetMap, reflecting relationships between agricultural land and water resources.

V. Nighttime light: Nighttime light intensity data for 2024 (15 arc-second resolution) were acquired from the Earth Observation Group, serving as a proxy for economic activity and urbanization intensity. We hypothesize a "threshold effect" regarding nighttime light intensity: urban agriculture is typically located in "dimly lit areas" within or on the fringes of built-up zones. Consequently, NTL data may aid the model in excluding both high-density built-up areas (high intensity) and pure natural wilderness (zero intensity).

2.1.3 Urban agriculture labeled samples

A ground-truth sample dataset was constructed by interpreting high-resolution images and conducting field surveys in Nanjing. Given the informal and fragmented nature of UA, relying solely on cadastral data is often insufficient. Therefore, we constructed the reference dataset through manual interpretation of high-resolution Google Earth imagery. To ensure data quality, we conducted targeted field surveys to verify the samples. Despite these efforts, we acknowledge potential biases; for instance, small-scale rooftop agriculture obscured by tree canopies may be underrepresented. To facilitate reproducibility and future benchmarking, the curated dataset is made publicly available at https://pan.dupetrc.qiyi.us:31583/nextcloud/index.php/s/UA_dataset . Each sample is a 128 × 128 pixel image patch labeled into one of five urban agriculture categories or a non-urban

agriculture category (total six classes). In total, 1,178 samples were collected. This labeled dataset was split into training, validation, and test sets for model training and evaluation.

I. Urban agriculture near transport land (80 samples),
II. Urban agriculture near residential and public service land (423 samples),
III. Urban agriculture near industrial land (239 samples),
IV. Rooftop urban agriculture (32 samples),
V. Urban agriculture on construction wasteland (297 samples),
VI. Non-urban agriculture / background areas (106 samples).

2.2 Model architecture and fusion framework

We selected two semantic segmentation architectures as baseline models: the classical U-Net and DeepLabv3 with a ResNet-50 backbone (DeepLabv3-ResNet50):

I. U-Net is a symmetric encoder–decoder network originally proposed by Ronneberger et al. for biomedical image segmentation. It features a contracting path (encoder) and an expanding path (decoder) with skip connections that concatenate feature maps at corresponding levels, which helps preserve spatial details and is advantageous for recognizing object boundaries and small targets in high-resolution imagery. In this implementation, the U-Net encoder consists of multiple DoubleConv blocks (each block has two convolutional layers with batch normalization and ReLU activation) and max-pooling layers for down-sampling, progressively reducing spatial resolution while increasing feature channels to capture higher-level semantic features. At the deepest layer (bottleneck, 512 channels), global features are aggregated before the decoder path begins. The decoder uses bilinear upsampling and concatenation with corresponding encoder feature maps at each level, followed by DoubleConv, to gradually recover spatial resolution and refine details. A 1×1 convolution output layer maps the final feature maps to the desired number of classes for pixel-wise classification.
II. DeepLabv3-ResNet50, on the other hand, is a deeper model that uses a ResNet-50 backbone for feature extraction and incorporates atrous (dilated) convolutions and an atrous spatial pyramid pooling (ASPP) module to capture multi-scale context. The ResNet-50 backbone provides robust multi-level feature representations through residual learning, and the ASPP module expands the receptive field to capture information at multiple scales. This architecture has proven effective in handling complex object boundaries and small fragmented patches, making it particularly suitable for recognizing urban agriculture in high-resolution imagery. Moreover, both prior studies and our own experimental trials indicate that, compared with other commonly used pre-trained models such as ResUNet, and EfficientNet, DeepLabv3 consistently demonstrates superior performance in urban agriculture identification tasks [13-15]. Therefore, the DeepLabv3-ResNet50 was selected as the baseline model for subsequent comparative experiments. We utilized a PyTorch-provided DeepLabv3-ResNet50 pre-trained on ImageNet, modifying the classifier head to output six classes (replacing the original fully convolutional layer with a 1×1 convolution with six channels).

Both baseline models were trained on the imagery data alone to serve as benchmarks. During training, we addressed class imbalance by using an Online Hard Example Mining (OHEM) cross-entropy loss. The OHEM strategy dynamically focuses training on hard-to-classify pixels (those with high loss), which improves the models' performance on under-represented classes and difficult boundary areas. For optimization, we used the Adam optimizer with an initial learning rate of 1×10^{-4} for both models. The models were trained using the PyTorch Lightning framework for efficient experiment management. We

monitored the mean IoU and loss during training to ensure stable learning, and saved model checkpoints based on validation performance. Visual outputs (comparisons of input image, ground truth, and prediction) on the validation set were also generated periodically to qualitatively assess model improvements.

Building upon the baseline models, we designed two multimodal fusion frameworks to integrate the multi-scale geospatial features with the image features. The design rationale is that urban agriculture distribution is driven by both fine-grained land cover characteristics and broader socio-environmental context. Remote sensing imagery provides detailed information on vegetation and land cover texture, while geospatial variables aggregated at multiple scales can reveal the influence intensity of various factors. By deeply coupling these two sources, the model can learn more discriminative representations for UA identification. We implemented fusion architectures for both U-Net and DeepLabv3-ResNet50, with different emphases (Fig. 2):

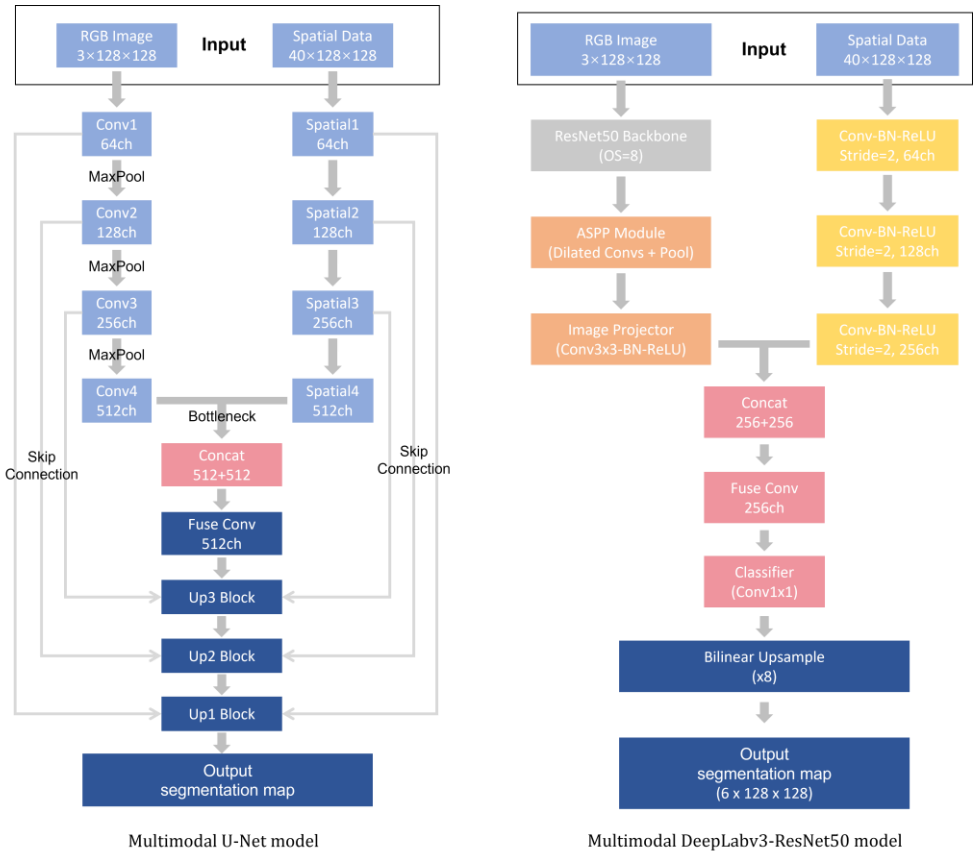

Fig. 2. Multimodal Fusion Framework

I. Multimodal U-Net model: In this framework, a parallel spatial feature encoder is constructed for the geospatial inputs, mirroring the four-level structure of the U-Net encoder. The geospatial encoder applies successive convolution and pooling operations to the multi-scale feature tensor (initially of size 128 × 128 × 40, where 40 is the number of spatial feature channels) to produce downsampled feature maps at resolutions 64 × 64, 32 × 32, 16 × 16 (with corresponding increasing channel depths). Each level of the geospatial encoder corresponds to one of the spatial scales (e.g., 128 px patch ≈ local scale, 64 px ≈ 500 m, 32 px ≈ 1000 m, 16 px ≈ 1500 m).

During decoding, the U-Net image decoder at each upsampling step concatenates the image feature map with the geospatial feature map of the same resolution. This implements a layer-wise fusion of "image details + spatial context," injecting the aggregated spatial features at the corresponding scale to inform the segmentation. Additionally, at the bottleneck (latent feature) level, the deepest image features (512 channels) are concatenated with the coarsest-scale spatial features (also 512 channels, corresponding to approx. 1500 m) to further enhance the model's understanding of global spatial context. This multimodal U-Net thus leverages skip connections to tightly integrate spatial variables with image features at multiple resolutions, which helps the model capture both fine boundary details and the broader spatial factors influencing urban agriculture presence.

II. Multimodal DeepLabv3-ResNet50 model: Here, we adopt a parallel fusion at the model output stage. The ResNet-50 + ASPP serves as the image branch, extracting image features (we use the output of the ASPP as image feature map). For the spatial branch, we process the multi-scale spatial feature tensor through a small convolutional network that downsamples it to a 16 × 16 feature map (to match the output stride of DeepLab, which is 16) and maps it to 256 channels. In practice, this can be a sequence of conv and pooling layers or an average pooling to 16 × 16 followed by 1 × 1 conv layers to achieve the desired channel dimensionality. The image feature map (also reduced to 16 × 16 with 256 channels by the ASPP) and the spatial feature map (16 × 16, 256 channels) are then concatenated, forming a 16 × 16 × 512 combined feature. This fused feature is passed through a fusion head consisting of a 1 × 1 convolution with batch normalization, a ReLU activation, a dropout layer, and a final 1 × 1 convolution to produce the six-class output. The multimodal DeepLab model thus captures global image context and spatial heterogeneity in parallel, allowing the network to model interactions between image-derived features and spatial variables.

In essence, the U-Net fusion emphasizes fine-scale, layer-by-layer coupling, whereas the DeepLab fusion emphasizes global context integration of spatial variables.

The input to the multimodal models includes two parts: (1) the remote sensing image patch (RGB bands, 128 × 128 pixels), and (2) the stack of co-registered multi-scale spatial feature raster corresponding to the same area (we have 40 spatial feature channels in total after aggregating 10 types of features across 4 spatial scales). Both inputs are normalized and aligned in resolution. The output is a 128 × 128 pixel map of predicted land use classes, distinguishing five types of urban agriculture and one non-UA category.
The training setting mirrored the baselines.

2.3 Model evaluation and feature interpretation

We evaluated model performance using several common segmentation metrics. A confusion matrix was constructed to analyse misclassification among all classes, which is useful for examining how the model confuses small urban agriculture classes with the non-UA background. From the confusion matrix, we derived:

I. Overall Accuracy (OA): the proportion of correctly classified pixels out of all pixels. OA indicates overall classification performance but can be biased if classes are imbalanced.

$$OA = \frac{\sum_{i=1}^{K} n_{ii}}{\sum_{i=1}^{K}\sum_{j=1}^{K} n_{ij}} \quad (1)$$

II. Intersection over Union (IoU) for each class, and mean IoU (mIoU): IoU is the ratio of the area of overlap between the predicted and true masks of a class to the area of

their union. The mIoU is the average IoU across all classes. This is a standard metric for segmentation that fairly accounts for each class's performance.

$$IoU_i = \frac{n_{ii}}{n_{ii}+\sum_{j=1,j\neq i}^{K} n_{ij}+\sum_{j=1,j\neq i}^{K} n_{ji}} \quad (2)$$

$$mIoU = \frac{1}{K}\sum_{i=1}^{K} IoU_i \quad (3)$$

III. F1-score (per class and mean F1): the harmonic mean of precision and recall for each class. The F1-score balances over- and under-prediction. We report the mean F1 (averaged over classes) as an overall measure.

$$F1_i = \frac{2\cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i} \quad (4)$$

$$Precision_i = \frac{n_{ii}}{\sum_{j=1}^{K} n_{ij}} \quad (5)$$

$$Recall_i = \frac{n_{ii}}{\sum_{j=1}^{K} n_{ji}} \quad (6)$$

To interpret the contribution of each geospatial feature to the model's predictions, we performed a channel-wise zero-out ablation analysis. This approach is analogous to permutation importance but more suitable for spatial features. We found that simply randomizing the values in a feature channel did not significantly affect the model due to the normalization and spatial structure (random shuffling largely preserved the value distribution and did not disrupt spatial continuity). Instead, for each spatial feature channel k, we set all its values to zero (effectively removing that feature entirely) while keeping all other inputs the same, and then recomputed the model's performance on the validation set. If the original model's mIoU is M and the mIoU with feature k zeroed out is $M^{(k)}$, we define the contribution of feature k as $\Delta\mathrm{mIoU}_k = M - M^{(k)}$. A larger $\Delta\mathrm{mIoU}_k$ indicates that the model performance drops more without that feature, meaning the feature is more important; conversely, a near-zero or negative $\Delta\mathrm{mIoU}_k$ implies the feature has little positive contribution or even a slight noise effect. This method avoids the issue of breaking the spatial continuity that permutation can cause and does not rely on inspecting internal model parameters, making it well-suited for interpreting "black-box" deep segmentation models. We conducted this ablation for all 40 spatial feature channels one by one and recorded the change in mIoU to identify the key features and their effective spatial scales. Additionally, we zeroed out all spatial feature channels simultaneously to quantify the overall contribution of the spatial features as a whole to the model's accuracy.

3 Results

3.1 Overall Comparison

Table 1. Model performance comparison

Model	OA	Mean IoU	Mean F1
Baseline U-Net	0.745	0.221	0.309
Fusion U-Net	0.817	0.414	0.537
Baseline DeepLabv3-ResNet50	0.874	0.575	0.711
Fusion DeepLabv3-ResNet50	0.896	0.615	0.745

We evaluated four models on the test set: the baseline U-Net, the fusion U-Net (with spatial features), the baseline DeepLabv3-ResNet50, and the fusion DeepLabv3-ResNet50. Table 1 summarizes their key performance metrics. Overall, the DeepLabv3-ResNet50

models significantly outperform the U-Net models under imagery-only settings, and incorporating multi-scale geospatial features further improves performance on both architectures.

3.2 Per-Class Behavior and Confusions

The baseline U-Net demonstrated limited discrimination of minority UA categories, particularly UA near residential and public-service land (class 2) and UA on construction wasteland (class 5)—with frequent misclassification into the background (class 0) (Fig. 3), suggesting that image-only features are insufficient for small or spatially ambiguous UA patches. Its overall accuracy primarily reflected the dominance of the non-UA background (IoU = 0.749) and moderate recognition of rooftop urban agriculture (class 4, IoU = 0.244; Table 2).

After introducing multi-scale geospatial indicators, the multimodal fusion U-Net achieved notable gains. IoUs for UA near residential & public-service, industrial, wasteland classes, and Rooftop UA all increased by over 0.2 (Table 2), indicating that contextual information – such as proximity to transport facility and population – significantly enhances detection of minor or fragmented UA areas and reduces boundary uncertainty.

The baseline DeepLabv3-ResNet50 already outperformed both U-Net variants. Its IoUs exceeded those of the U-Net models for every UA class (Table 2), demonstrating strong multi-scale contextual reasoning via the ASPP module and the benefit of pretrained deep representations. This confirms that deeper architectures capture complex spatial structures more effectively in heterogeneous urban landscapes.

The multimodal DeepLabv3-ResNet50 achieved the best overall performance (OA = 0.896, mIoU = 0.615, mean F1 = 0.745), improving by +0.022 OA, +0.040 mIoU, and +0.034 F1 over its image-only counterpart (Table 1). The most significant improvements occurred in previously challenging category: UA near wasteland (class 5, +46.2% in F1 score; Table 3).

Table 2. Comparison of IOU for each class

Model	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5
Baseline U-Net	0.749	0.103	0.038	0.127	0.244	0.064
Fusion U-Net	0.843	0.019	0.350	0.367	0.601	0.305
Baseline DeepLabv3-ResNet50	0.897	0.368	0.545	0.565	0.731	0.345
Fusion DeepLabv3-ResNet50	0.889	0.276	0.636	0.709	0.583	0.600

Table 3. Comparison of F1 score for each class

Model	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5
Baseline U-Net	0.857	0.186	0.074	0.225	0.392	0.120
Fusion U-Net	0.915	0.037	0.518	0.537	0.751	0.468
Baseline DeepLabv3-ResNet50	0.946	0.538	0.705	0.722	0.845	0.513
Fusion DeepLabv3-ResNet50	0.941	0.432	0.778	0.830	0.737	0.750

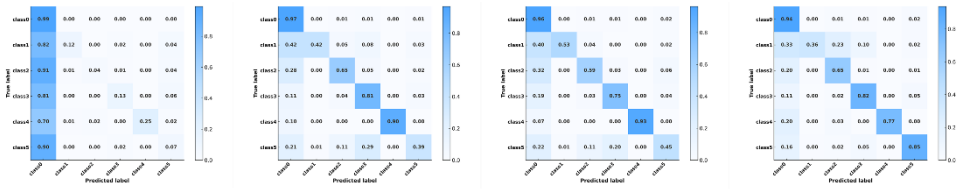

Fig. 3. Confusion matrices of the baseline U-Net, fusion U-Net, baseline DeepLabv3-ResNet50, and fusion DeepLabv3-ResNet50 models

3.3 Feature Interpretation

After training the fusion models, we conducted the channel-wise ablation analysis to analyze the contribution of each spatial feature. When all 40 spatial feature channels were set to zero, the multimodal DeepLabv3-ResNet50's mIoU dropped sharply from 0.6154 to 0.3908 (ΔmIoU = +0.2246). This severe performance degradation confirms that the geospatial features as a whole are indispensable for the model – without them, the accuracy falls well below even the image-only baseline (mIoU 0.575).

At the individual feature level, different types of spatial variables showed markedly different contributions (Table 4). The results indicate that transportation-related POIs at 1000 m have the largest ΔmIoU (+0.0461). Results show that urban agriculture presence is strongly correlated with transportation facility accessibility, especially at neighborhood to medium scales (500-1000m), possibly due to available vacant land or intentional planning along transportation corridors.

Population at 1500 m also contributes notably (+0.0400), suggesting UA distribution is shaped by broader human activity patterns. For instance, districts with larger population catchments might see more demand or more pressure for urban agriculture spaces. Small-scale population variation immediately around a site seems less important.

Features related to service facilities showed strong contributions at short ranges. Public-service POIs at 500 m (+0.0209) and healthcare POIs at 500 m (+0.0183) are beneficial at neighborhood scales. Such facilities might provide supportive environments or available land (for example, community gardens near hospitals or schools, or small farms near community centers), reflecting that UA often takes advantage of underutilized spaces in and around places of public service.

The nighttime light intensity feature exhibited a non-linear scale dependence. At the neighborhood scale (500 m), removing the feature caused a modest performance drop (ΔmIoU = +0.0152), indicating that local human activity signals (as proxied by lights) aid in identifying urban agriculture (UA). In contrast, negative impacts were observed at both the within-sample extent (ΔmIoU = -0.0326) and large-scale 1500 m buffer (ΔmIoU = -0.0105), although driven by different mechanisms. At the within-sample extent, nighttime light primarily captures built-up intensity within the labeled sample unit, rather than surrounding spatial context. In dense urban environments, small and embedded UA patches may exhibit nighttime light characteristics similar to adjacent built-up surfaces, biasing the model toward non-UA predictions. At the large-scale 1500 m buffer, nighttime light aggregates regional urbanization intensity, which is negatively associated with UA presence. In both cases, the nighttime light signal introduces semantic conflicts that increase the likelihood of false negatives. Removing these channels therefore reduces classification confusion and improves model performance.

On the other hand, linear infrastructure and natural features such as road networks, railways, rivers, and water surface showed negligible contributions (ΔmIoU ≈ 0) under

current resolutions. This indicates that these features were not key determinants in the model's discrimination of urban agriculture. There are a few possible reasons: (1) the spatial resolution or encoding of these features may not have been sufficient (many grid cells had zeros for these features, especially at small scales, due to sparse networks), (2) urban agriculture in Nanjing may not have a strong direct relationship with proximity to roads or water compared to other factors, or (3) these features had collinearity with other variables (e.g., transport POIs might implicitly capture effects of road proximity). In any case, the model did not rely on these linear or hydrological features to make its predictions. This could also be partly due to data limitations or the binary nature of these features (presence/absence rather than intensity), making their signal weak for a fine-scale land use like urban agriculture.

Table 4. Selected results of the channel-wise ablation experiment

Feature (spatial scale)	$M^{(k)}$	$\Delta\mathrm{mIoU}_k$
all 40 spatial feature channels were set to zero	0.3908	+0.2246
Transportation facilities POI (500 m)	0.5886	+0.0268
Transportation facilities POI (1000 m)	0.5693	+0.0461
Transportation facilities POI (1500 m)	0.6039	+0.0115
Population distribution (1500 m)	0.5754	+0.0400
Public service facilities POI (500 m)	0.5945	+0.0209
Healthcare facilities POI (500 m)	0.5971	+0.0183
Nighttime light intensity (within sample range)	0.6480	-0.0326
Nighttime light intensity (500 m)	0.6002	+0.0152
Nighttime light intensity (1500 m)	0.6259	-0.0105
Road length (any scale)	~0.6154	≈0.0000
Railway length (any scale)	~0.6154	≈0.0000
Water area (any scale)	~0.6154	≈0.0000
River network length (any scale)	~0.6154	≈0.0000

From these ablation results, we conclude that transport accessibility, population distribution, and proximity to service facilities are the dominant spatial factors influencing the accurate identification of urban agriculture in the model. Local indicators of human activity (e.g., lights) have auxiliary value at smaller scales, whereas major infrastructure and natural water features played a limited role in this study. These findings validate the importance of incorporating multi-scale socio-economic factors and also provide guidance for understanding the spatial distribution mechanisms of urban agriculture: UA tends to occur in areas that are moderately accessible and near communities and services, rather than in the densest urban cores or purely along natural features.

3.4 Spatial Distribution Patterns of Urban Agriculture in Nanjing

Applying Fusion DeepLabv3-ResNet50 to the selected 50 km strip provides insight into the urban–rural distribution of urban agriculture (Fig. 4). The results reveal a clear urban-to-rural gradient: In the urban core areas, urban agriculture is found mostly in the form of very small

patches or point-like community gardens embedded within built-up areas. These appear as highly fragmented green spots amidst dense development, reflecting an "embedded" form of urban agriculture. Moving outwards to the peri-urban or suburban transitional zones, urban agriculture plots become larger and more linearly contiguous. They often extend along edges of industrial zones, transportation corridors, or idle construction land, forming strips of cultivation that take advantage of underutilized fringe spaces. This indicates a pattern of urban agriculture utilizing "edge spaces" along the urban fringe. Finally, in the outer rural areas at the far end of the strip, agricultural land becomes significantly larger in size and more continuous, resembling stable traditional agricultural landscapes. Overall, along this 50 km strip from the city center outward, there is a pronounced spatial pattern: highly fragmented and sporadic in the core, increasingly continuous in the suburbs, and predominantly continuous and extensive in the rural outskirts, representing a transition from fragmentation to semi-continuity to full continuity.

This gradient pattern suggests that the intensity of urbanization strongly influences the form of urban agriculture. In dense urban areas, only small pockets can exist (often in the form of rooftop gardens, community gardens, or tiny vacant lot farms), whereas in suburban areas, urban agriculture can take on larger, ribbon-like forms along undeveloped lands, and in rural edges it seamlessly connects with conventional agriculture. The analysis also reveals that the spatial arrangement of these urban agriculture sites is highly coupled with surrounding land uses: for instance, in the strip we observed many urban agriculture patches near living facilities, transportation land, and idle or vacant lands. This underscores that UA often emerges in interstitial spaces of the urban fabric, including near housing estates, alongside rail lines or highways, and on leftover land.

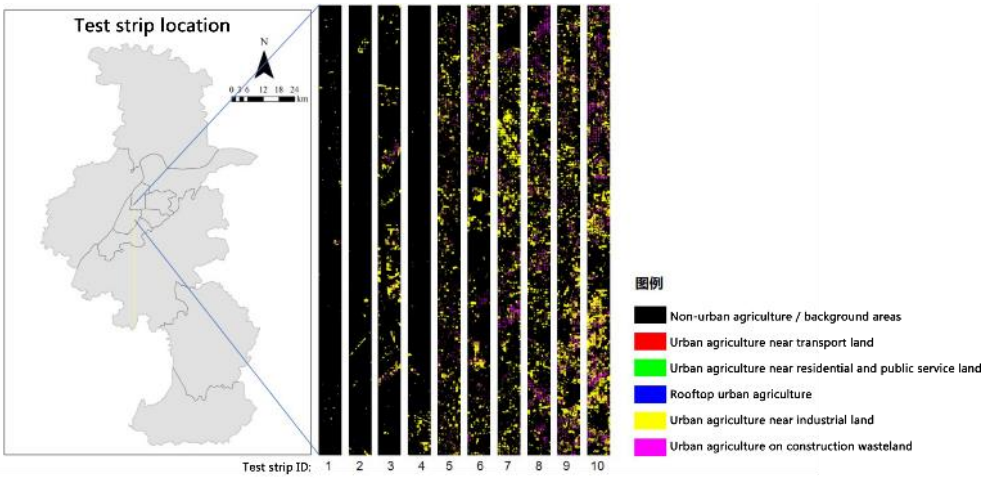

Fig. 4. The distribution of urban agriculture on the selected 50 km strip in Nanjing

4 Discussion

4.1 Why Multi-Modal Fusion Helps

The multimodal fusion of geospatial data with imagery further enhanced model discriminative power, especially for classes that were poorly captured by image-only models. This is because UA presence is jointly driven by biophysical contexts and socio-spatial factors (accessibility, service proximity, human activity). Imagery alone captures surface texture but misses these drivers. By encoding multi-scale geospatial indicators (like being

near certain facilities or within certain land use types), the fusion models inject contextual priors, improving minority class recognition and boundary disambiguation. This suggests that future urban land use studies should consider a multimodal approach, and additional data such as land use plans, soil data, or socio-demographics might further improve performance.

4.2 Scale Effects

The channel-wise ablation experiment also revealed the differing contributions across the 500, 1000, 1500 m scales indicate scale-dependent mechanisms: for example, population mattered at 1500 m but not at 500 m; nightlights exerted a positive impact at 500 m but a negative one at 1500 m. Selecting appropriate aggregation scales is thus critical. This aligns with the theoretical basis of MGWR and multiscale analysis in urban studies – it emphasizes that an urban phenomenon like UA cannot be fully understood at a single scale. Our approach of multi-scale feature aggregation is a practical way to let the model learn these scale effects. The ablation analysis, in turn, provides a level of interpretability (often lacking in deep learning models) by quantifying each feature's contribution. This contributes to bridging the gap between "black-box" model predictions and urban geographic interpretation, allowing urban planners and researchers to pinpoint which factors are most associated with urban agriculture presence.

4.3 Planning and Policy Implications

Findings suggest UA tends to cluster near accessible edges and service-supported neighborhoods. It is gradually becoming an important land use type to supplement ecological functions, regulate the environment, and meet residents' diverse needs within the built environment. In other words, rather than expanding outward, the city is finding opportunities to introduce or maintain agriculture within the urban and peri-urban footprint (for example, converting vacant lots into community gardens or preserving farming in urban fringes) as a strategy to enhance urban sustainability. This aligns with broader goals of urban resilience and carbon reduction. Planners can leverage the strip-gradient insight to designate UA priority zones within the UDB and buffer-zone corridors beyond it. The framework can support scenario testing for UA layout optimization in urban renewal and territorial spatial planning.

Furthermore, the channel-wise ablation results (Table 4) offer a pathway for "lightweight deployment." Planners can implement a simplified version of the framework by focusing only on the top-three contributors—transportation POIs (1000 m), population (1500 m), and public services (500 m). This feature-contribution guided simplification significantly reduces data processing burdens while retaining the core spatial logic, making the tool more accessible for routine planning tasks.

4.4 Limitations and Future Work

Despite the progress made, several limitations of this study should be acknowledged, which also point to directions for future research.

I. Generalizability and Transferability: the current model is validated only in Nanjing, a high-density city with specific urban form and urban agriculture practice. While the core image-spatial fusion architecture is conceptually transferable, its performance in cities with different urban forms (e.g., sprawling low-density cities), climatic conditions (affecting vegetation phenology in imagery), or data availability (where geospatial indicators are scarce) remains untested and may require adaptive parameter tuning or generalization training. Future studies could address this issue

through cross-city validation. This includes developing a "lightweight" version that relies on globally available datasets and exploring adaptive spatial aggregation scales to suit different city sizes and densities.

II. Scope and Data Constraints: First, our definition of "urban agriculture" is tied to the urban development boundary; without explicit masking, the model may confuse urban-edge agriculture with rural farmland at the fringe. Second, the training dataset is relatively small and class-imbalanced, which limits the model's ability to recognize rare UA types perfectly. Thus, for practical applications, post-processing steps, such as clipping results to the official urban boundary, can be integrated to resolve edge-case confusion at the rural fringe. Expanding the training dataset through additional sampling, data augmentation, or semi-supervised learning could improve model robustness and generalization.

III. Model Complexity and Efficiency: While the multi-modal fusion improves accuracy, it introduces significant computational overhead compared to lightweight networks. Furthermore, despite using ablation studies for interpretation, the deep learning architecture remains a "black box," lacking the transparency of traditional regression models. Future models could employ more advanced Explainable AI (XAI) techniques to decode the decision logic, increasing trust for urban planning practitioners.

5 Conclusions

In conclusion, this study proposes and validates a multi-modal geospatial–imagery fusion framework for fine-grained identification of urban agricultural land. On Nanjing data, fusing multi-scale geospatial indicators with DeepLabv3-ResNet50 significantly improves OA, mIoU, and F1 over imagery-only baselines, particularly for minority and boundary-ambiguous classes. Channel-wise ablation highlights the leading roles of transport-related POIs, large-scale population, and neighborhood-scale public services. Strip-scale mapping confirms the model's utility for practical planning insights and reveals a transition from fragmented embedded UA in core urban areas to continuous belts toward rural peripheries.

Although we focused on one city, the methodology is generalizable and, with further enhancement in terms of dynamic data, broader testing, data volume, and model interpretability, it holds promise for wider application in various UA contexts. Ultimately, the framework offers robust technical support for the optimal allocation of urban agricultural spaces in territorial spatial planning and urban sustainability initiatives, helping cities integrate food production and green space into their development for greater resilience and livability.

We acknowledge the support the National Key Research and Development Program of China (2024YFE0197900): "Multiserviceability Evaluation and Spatial Optimization Technologies for Urban Agriculture". We thank the contributors of Amap, OpenStreetMap, LandScan, and the Earth Observation Group for open data support. We also acknowledge field survey assistance from colleagues at Nanjing University.

References

1. H. Yin, F. Kong, H. Liu, Z. Shen, C. Chen, P. Chen, T. Sun, Acta Ecol. Sin. **45**, 1 (2025) (in Chinese)
2. P. Cheng, G. Qi, Y. Zhong, Bull. Surv. Mapp. **5**, 90 (2024) (in Chinese)

3. W. Shi, *Urban functional zone identification method integrating high-resolution remote sensing and open geographic data*, M.S. thesis, Nanjing University of Posts and Telecommunications (2023) (in Chinese)
4. A. A. Gharahbagh, S. Siachalou, M. J. Valadan Zoej, A. Mohammadzadeh, Sensors **25**, 1988 (2025)
5. S. Sierra, R. Ramo, M. Padilla, A. Cobo, Environ. Monit. Assess. **197**, 1 (2025)
6. R. T. Handayanto, J. Ilm. Komput. Inf. **17**, 89 (2024)
7. X. Feng, Y. Zhang, T. Dong, M. Liu, Remote Sens. **17**, 482 (2025)
8. H. Bao, D. Ming, Y. Guo, K. Zhang, K. Zhou, S. Du, Remote Sens. **12**, 1088 (2020)
9. Z. Xie, L. Han, L. Sun, B. Peng, Remote Sens. Technol. Appl. **40**, 708 (2025) (in Chinese)
10. A. S. Fotheringham, W. Yang, W. Kang, Ann. Am. Assoc. Geogr. **107**, 1247 (2017)
11. W. Wu, W. Yang, T. Dai, Z. Xie, Y. Liu, Sci. Rep. **15**, 23459 (2025)
12. L. Liu, H. Huang, J. Yang, Land **13**, 2154 (2024)
13. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*, in Proceedings of the European Conference on Computer Vision (ECCV) (2018)
14. M. Tan, Q. V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, in Proceedings of the 36th International Conference on Machine Learning (ICML) (2019)
15. Z. Zhang, Q. Liu, Y. Wang, IEEE Geosci. Remote Sens. Lett. **15**, 749 (2018)