

<https://doi.org/10.1038/s40494-025-02260-2>

Segmenting of historic landscape system along Jiangnan Canal based on deep learning and multimodal geodata

Check for updates

Li Ran¹, Keyu Chen², Shunhan Zhang¹, Qianting Gao¹, Yuqi Gao¹, Shangyu Tan³ & Qing Lin¹ ✉

As a living heritage, the Jiangnan Canal landscape features interwoven natural and artificial water networks. To accurately interpret its overall spatial composition and address heritage sustainability challenges, this study proposes Geo-SegFormer: a framework for the automated segmentation of historical landscapes by integrating deep learning with multimodal geospatial data. Using a self-constructed dataset, the proposed method achieves the first 1-meter-resolution reconstruction of the entire Jiangnan Canal landscape system across 35 categories, surpassing traditional manual methods in coverage, diversity, and precision. Quantitative analysis reveals the contribution weights to the segmentation: DEM 51.6%, hydrographic data 17.7%, and imagery data 30.7%, demonstrating that the canal landscape is structured upon natural terrain, with water networks as its spatial framework. This outcome establishes a critical data foundation for heritage conservation and interdisciplinary spatial quantitative research. Beyond this specific case, the developed methodology also possesses considerable potential for cross-regional application.

The Jiangnan Canal traverses the Taihu Basin. Since its dredging in the Sui dynasty (AD 610)¹, it has consistently served as a human settlement support system integrating drought-flood regulation, inland navigation, agricultural irrigation, and landscape recreation in Jiangnan region². The canals thereby led to the development of the landscape system along their routes, characterized by the interaction of natural and artificial water networks and the synergistic coexistence of towns, villages, and farmland. Its layout and essential functions persist today³. It stands not only as a living exemplar of ancient wisdom in human–water symbiosis and cultural memory, but also as a crucial theoretical and practical reference for contemporary territorial spatial planning, ecological management, and cultural inheritance. However, amid rapid urbanization and industrial modernization, these values of historical landscapes have not been systematically understood, leading to the accelerated disappearance of numerous canal branches and traditional agricultural and rural landscapes^{4,5}. Thus, deciphering the composition of the Jiangnan Canal historical landscapes and accurately reconstructing their complete spatial configuration have become pivotal tasks for revealing their formative mechanisms and comprehensive values, reshaping regional cultural identity, and addressing the ongoing crisis of heritage preservation^{6,7}. However, achieving this goal faces multiple challenges.

The concept of landscape type differs from land-use classification. The classification and spatial pattern recognition require professional interpretation that integrates geographical features, cultural semantics, and

historical records. The limitations in spatial coverage and precision of traditional documents make it difficult to fully depict historical landscapes. Some studies only redraw historical maps, seldom producing accurate spatial reconstructions^{8,9}. The recently declassified Keyhole (CORONA/KH) satellite imagery, with a resolution of 0.6–2.4 meters, provides a precise record of China's historical landscapes prior to its rapid urbanization, effectively filling gaps in traditional sources¹⁰. Integrating this imagery with historical records enables the accurate reconstruction of historical landscape systems across the entire region.

Some researchers have integrated both data types in historical landscape studies. The typical workflow involves: (1) selecting representative samples, (2) using historical satellite imagery as a base layer, (3) overlaying historical information, (4) manually inferring landscape composition, and (5) presenting the characteristics via schematic diagrams (Fig. 1). This approach, however, is reliant on expert knowledge and is labor-intensive, thus limiting research to small-scale, localized areas¹¹. Furthermore, studies often focus on narrowly defined heritage sites or their immediate contextual elements from a single perspective, which impedes the reconstruction of multi-category historical landscape systems and the comprehensive understanding across their full spatial extent.

Therefore, the development of automated identification and interpretation methods has become key to advancing research on historical landscapes. Although conventional shallow machine learning algorithms

¹School of Landscape Architecture, Beijing Forestry University, Beijing, China. ²Sichuan Institute of Computer Sciences, Chengdu, China. ³Chengdu Environment Investment Digital Intelligence Eco-Technology Co., Ltd., Chengdu, China. ✉e-mail: lindyla@126.com

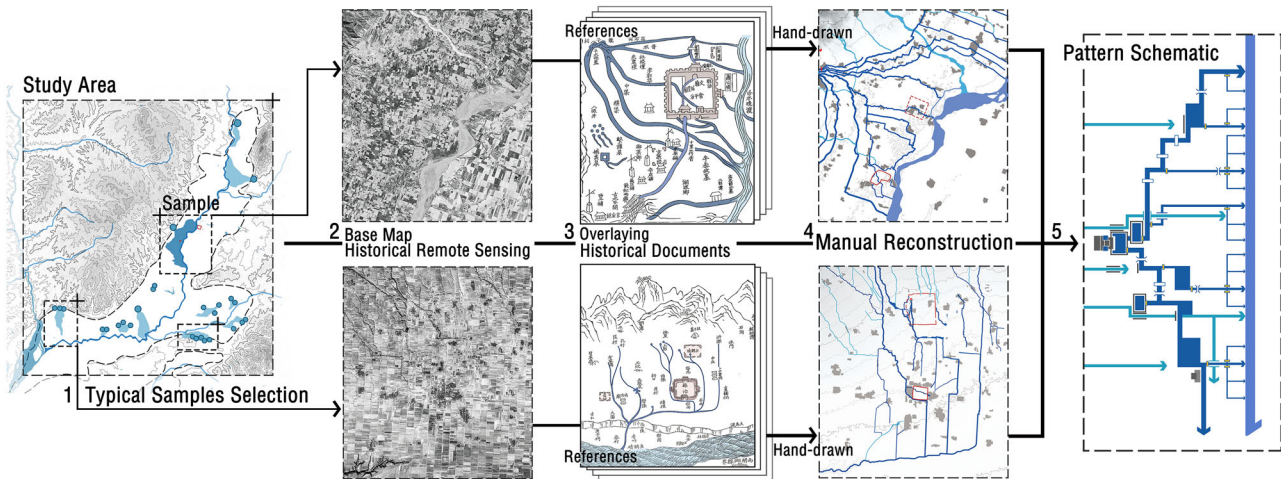


Fig. 1 | The typical workflow in historical landscape research. The figure illustrates a conventional workflow that uses both historical imagery and historical documents to manually reconstruct the historical landscape patterns of selected samples.

(e.g., maximum likelihood, random forest, self-organizing maps) have been widely applied in remote sensing interpretation, their reliance on manually designed feature extraction makes them inadequate for historical landscapes, which exhibit composite land-use, intricate cultural semantics, and absent multispectral data. Recent advances in Transformer-based deep learning models have significantly evolved their ability to capture both local textures and global contextual dependencies in remote sensing imagery, opening up new possibilities for this objective.

In deep learning applications, existing studies are predominantly confined to either generic land cover classification—utilizing RGB or multispectral data to categorize land use and settlement spatial elements—or to the analysis of specific individual targets, including landforms, water bodies, forests, buildings, and croplands. The deployment of panchromatic imagery remains in its infancy, with major advances primarily seen in the application of U-Net series and YOLO variants for reconstructing historical land use, identifying glacial features, and detecting archaeological sites¹². Consequently, there is a notable lack of methods and practical for the automated segmentation of historical landscape types. Overall, a significant gap persists in both the methods and practices for automated historical landscape segmentation.

To this end, this study aims to refine existing semantic segmentation methods to accurately reconstruct the Jiangnan Canal and its traditional landscape system across its entire extent. This will support cultural landscape and heritage research while overcoming current methodological limitations in spatial coverage, categorical diversity, and precision. Following a comprehensive evaluation, we adopt SegFormer as our fundamental architecture. Its layered Transformer encoder and positionless coding design enable precise and efficient extraction of both global and local information from large-scale imagery, while its lightweight MLP decoder ensures efficient semantic segmentation¹³. Focusing on the Taihu Basin, this research leverages KH-9 historical satellite imagery for pixel-level landscape classification, with the primary tasks and contributions outlined as follows:

- a. Provide a solution to the current lack of automated, region-wide recognition methods for multi-category historical landscapes. We develop Geo-SegFormer: a semantic segmentation framework for historical landscapes that integrates multimodal historical imagery, topographic and hydrographic data, with automatic adaptation to input channel variations.
- b. Transcend conventional land-use classifications, expand the research outcomes of the Jiangnan Canal landscape, surpassing prior work in coverage, accuracy, and classification granularity. This study pioneers a basin-wide, pixel-level spatial reconstruction of the Jiangnan Canal historical landscape, achieving 1-meter resolution modeling across 35 categories.

- c. Through ablation studies and XGBoost-Shap^{14,15}, interpretability analysis, we validate the method’s efficacy and quantify the contribution patterns of each modality to prediction outcomes.

The framework developed in this study enables the transition of historical landscape research from “singular-perspective, local-typicality, manual-interpretation, qualitative-analysis” to “systematic-research, full-spatial-coverage, automatic-interpretation, quantitative-analysis”. Furthermore, the spatial reconstruction of the Jiangnan Canal historical landscape serves not only as a historical record to directly inform the conservation and renewal of heritage landscapes but also provides a critical dataset for quantitatively analyzing its structural patterns and construction logic, thereby fostering the sustainable inheritance of canal culture and its ecological wisdom.

Methods

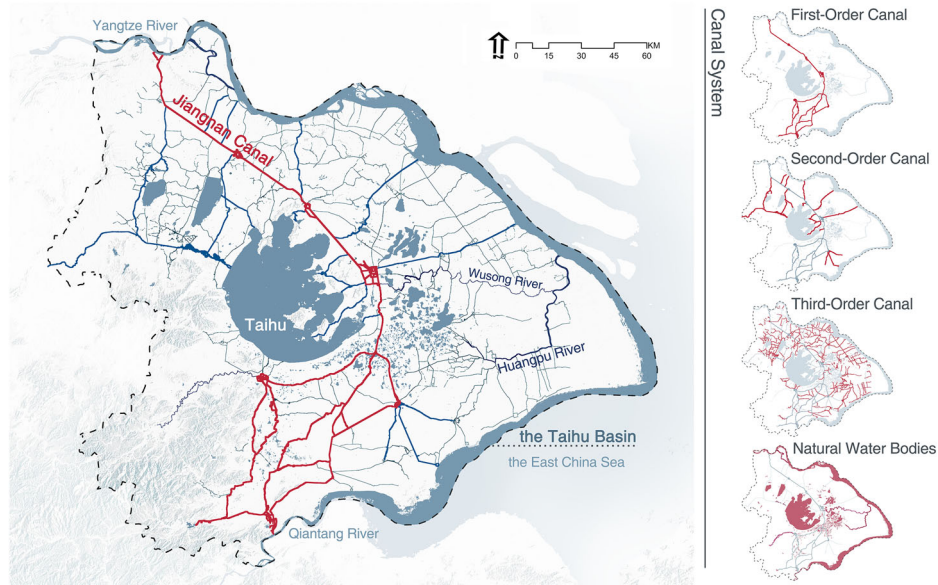
Study area and canal system

This study selects the Taihu Basin, traversed by the Jiangnan Canal, as the study area. While the region’s low-lying terrain and dense waterways weren’t naturally suited to the development of agricultural civilizations, they nonetheless provided a foundation for the construction of the canal¹⁶. To connect rivers, lakes, and the sea, various waterways were continuously extended, gradually forming a hierarchical canal system comprised of main canals supplemented by branch channels and natural water bodies (Fig. 2)¹⁷. The continued development of the canal also led to the rise of post stations along its route into prosperous commercial settlements. Simultaneously, the silt deposited along the riverbanks transformed into cultivated fields, which served as sources of food and handicraft raw materials for the urban settlements, continuously evolving under the irrigation and transportation guarantees of the canal system¹⁸. Ultimately, the Jiangnan Canal landscape system was formed, in which natural and artificial water networks interact, and towns, villages, and farmland coexist in a coordinated manner¹⁹.

Automatic interpretation of the landscape system

Current research on land cover interpretation with multifunctional attributes predominantly relies on morphological clustering analysis^{20–24}. Additionally, some studies utilize POI data to adjust the functional semantics of spatial objects through graph neural networks²⁵. However, these efforts focus solely on refining classifications of single surface types, rather than addressing landscape typology classification tasks that encompass cultural, functional, and other semantic dimensions. Studies on direct automatic identification of landscape types remain scarce. Li et al. conduct secondary clustering via SOM and K-means algorithms based on manually designed polder landscape morphological indices for landscape

Fig. 2 | Study area and canal system. The location of study area, Taihu Basin, China. Red lines indicate the Jiangnan Canal, and the right side shows the canal system with its main canal and lower-order branch channels.



interpretation²⁶. Deep-learning-based studies primarily employ convolutional neural networks. Fleischmann et al. categorized 16 composite landscape types centered on urban functions²⁷. Meng et al. combined water network and parcel division patterns, using ResNet50 to interpret five types of polder landscape²⁸. Shi et al., based on natural and cultural heritage characteristics, employed DeepLabV3+ to obtain five types of farmland landscapes²⁹. Additionally, Cui et al. used KH-5 imagery to interpret three types of sandy polder landscapes, but their approach still required substantial manual post-processing³⁰.

Automated interpretation of panchromatic remote sensing imagery

The keyhole series imagery, as the earliest accessible satellite imagery, is crucial aviation data for revealing past realities. However, due to its single-band panchromatic nature, combined with geometric distortions and noise from film data, apart from manual visual interpretation, there are currently few automated interpretation studies. Existing research predominantly employs conventional machine learning algorithms, including maximum likelihood (ML), random forest (RF), support vector machine (SVM), and self-organizing maps (SOM), focusing on historical land use applications. Limited studies have analyzed changes in urban built-up areas, forests, or agricultural land, generally employing binary or limited-category classifications. In light of the limitations of singular spectral data in panchromatic imagery, researchers have applied the gray level co-occurrence matrix (GLCM) to derive image texture features³¹. Building on this foundation, Rizayeva et al. further employed the SNIC algorithm to generate geometric features and integrated DEM data for multi-feature fusion, which enhanced model recognition accuracy³². Research has also been conducted on employing image texture features for pseudo-color synthesis³³ or utilizing U-Net generators to colorize panchromatic imagery within GAN frameworks^{34,35}. Despite issues such as spectral distortion, researchers have utilized SVM to extract land use from these colorized results³⁶.

In the field of deep learning, most studies deal with a limited number of classes. The interpretation themes primarily focused on employing U-Net series models for land cover interpretation^{37,38}, or reconstructing historical land use with multi-source data³⁹. To address the scarcity of training samples and high annotation costs for panchromatic imagery, Mboga et al. combined U-Net with domain adaptation networks to enhance cross-regional interpretation performance for historical imagery⁴⁰. Concurrently, Dahle et al. processed data into multi-scale representations for training and fused the outputs to integrate global and local features⁴¹.

Geo-SegFormer

Current research utilizing deep learning for remote sensing image interpretation predominantly employs samples with 512×512 pixel resolution or smaller, typically addressing a limited number of categories. Studies on automated interpretation of landscape types using historical imagery are even scarcer, with no publicly available models capable of segmenting such landscape types. The Jiangnan Canal and its landscape system are dominated by long-distance linear water networks and large-scale, cross-regional agricultural landscape patches, while also incorporating smaller-scale elements such as capillary polder networks and point-like settlement patches. Hence, landscape recognition requires models that effectively capture local feature details while contextually perceiving global structures. Accordingly, training sample selection requires multi-scale synergy. Large-scale samples enhance contextual relevance by covering broader scenes and more complete spatial structures, improving long-range dependency modeling while reducing boundary effects from local truncation, thus boosting the integrity and accuracy of long-distance coherent feature identification. In contrast, small-scale samples focus on high-resolution local features to compensate for detail dilution in large-scale training while strengthening modeling capabilities for capillary water networks and discrete patches (Fig. 3).

Therefore, to identify a suitable model architecture, we conducted a comparative performance analysis. U-Net, TransUNet, and SegFormer were selected as representative models of convolutional, hybrid convolutional-transformers, and pure transformer architectures, respectively. The training and validation data were prepared by replicating the original panchromatic imagery into three channels. As shown in Table 1, SegFormer outperformed the other models across all evaluation metrics on the test set, achieving an overall accuracy of 0.85, Kappa of 0.84, F1 score of 0.85, mIoU of 0.53, and mDice of 0.63. Simultaneously, SegFormer's hierarchical Transformer encoder eliminates positional encoding, enabling synchronous modeling of high-resolution local details and low-resolution global semantics while adapting to arbitrary input sizes without accuracy degradation. Paired with a lightweight MLP decoder that synergizes global and local information, this framework reduces computational complexity and meets the segmentation task's demands for the canal landscape system with multi-scale training. To this end, we select SegFormer as the foundational architecture.

However, canal landscape features exhibit high diversity with composite spatial and cultural semantics. Combined with the absence of multispectral data in historical panchromatic imagery and spectral confusion among features, these factors pose significant challenges for effective

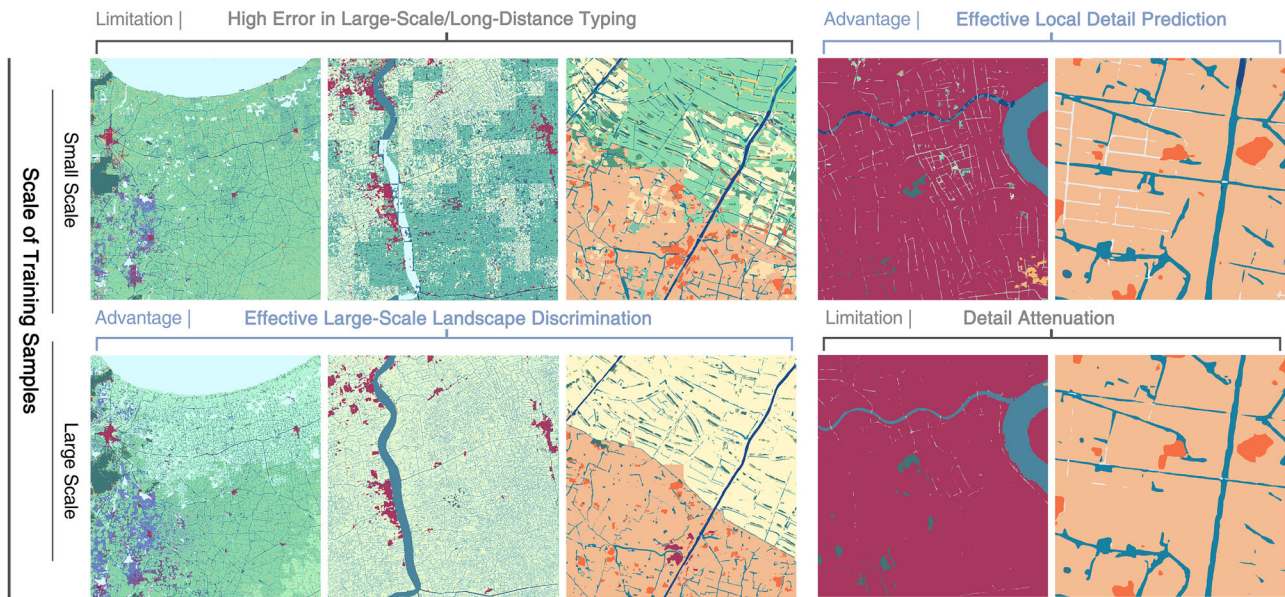


Fig. 3 | Comparison of training results using samples at different scales. The first column presents a macro-level identification performance comparison between small-scale and large-scale training samples, while the second column provides a micro-level comparison.

Table 1 | Comparison of Model Performance

Model configuration	Kappa	Overall accuracy	F1-score	mIoU	mDice
U-Net	0.7927	0.8081	0.8069	0.4806	0.6133
TransUNet	0.7341	0.7544	0.7335	0.3261	0.4100
SegFormer	0.8469	0.8583	0.8525	0.5314	0.6373

The bold values highlight the highest performance achieved in the test.

segmentation of historical canal landscapes. Tracing the formation of the Jiangnan Canal landscape system reveals that terrain and hydrographic elements constitute key geographical drivers. Therefore, building on the spatial-cultural attributes of historical landscapes and the characteristics of panchromatic data, while accounting for the aforementioned complexities, we propose Geo-SegFormer: a semantic segmentation method for historical landscapes that integrates deep learning with multimodal geospatial data (Fig. 4). The remainder of this section details the proposed method.

Differential three-channel CLAHE enhancement for panchromatic imagery

To address the single-band limitations of panchromatic remote sensing imagery and better adapt to pretrained model input structures, this study designs a differential three-channel CLAHE (Contrast Limited Adaptive Histogram Equalization) enhancement strategy.

Channel 1: Uses normalized orthoimagery data to preserve original surface texture information, serving as the baseline channel for multispectral synthesis.

Channel 2: Applies the standard CLAHE algorithm⁴² to normalized orthoimagery with a clip limit of 2.0. The output is:

$$C_2(x, y) = 255 \times \frac{\sum_{k=0}^{(x,y)} h_{clip}(k)}{\sum_{k=0}^{255} h_{clip}(k)} \tag{1}$$

Where $I(x, y)$ is the normalized pixel value, $h_{clip}(k)$ is the clipped histogram. The high-frequency texture of the imagery is enhanced, improving the local contrast and edge structural integrity of landscape elements, while strengthening visual separability between landscape

patches and discriminability of linear features such as canal systems and road networks.

Channel 3: First compresses the brightness of the normalized remote imagery, then applies CLAHE enhancement.

$$C_3(x, y) = \text{CLAHE}(0.8 \times I(x, y)) \tag{2}$$

This approach mitigates overexposure and noise risks from excessive bright-region histogram stretching during enhancement. It selectively enhances mid-low frequency information in shadow regions and other low-reflection, weak-texture areas such as mountains, farmlands, and lakes. These operations significantly improve discrimination among landscape types with similar textures but distinct cultural semantics.

At this stage, the panchromatic imagery is enhanced from a single-band to a three-band TIFF format. Given spectral heterogeneity in frequency domains among landscape types (e.g., high-frequency edges, mid-frequency patches, low-frequency homogeneous areas)^{43,44}, this processing enhances spectral differences between landscape types. After RGB pseudo-color mapping of the merged imagery, key landscape elements, including settlements, polders, dyke-ponds, lakes, marshes, and mountains, become distinguishable by color (Fig. 5).

Multimodal geographic data fusion

To enhance the accuracy of historical landscape identification, this study utilized terrain and hydrological data as additional inputs. This design is not a simple stacking of data, but is based on a deep understanding of the landscape evolution mechanism: in the Jiangnan Canal area, the terrain elevation and the distribution of the canal water system directly determine the spatial differentiation of landscape types. The height of the terrain dominates the distribution of water sources, forming different landform units, while the canal water conservancy further shapes various types of polder and settlement landscapes.

At the technical implementation level, the study normalized DEM and binarized water network data to eliminate sensor differences. These processed data were then stacked with the three-channel panchromatic image to form a five-channel input dataset, which was then fed into the Segformer model for high-dimensional fusion. This process simplifies data flow and improves fusion efficiency. Its encoder first extracts multi-scale feature embeddings with global semantics from multi-channel data, and then the MLP decoder performs nonlinear fusion of these embeddings in high-

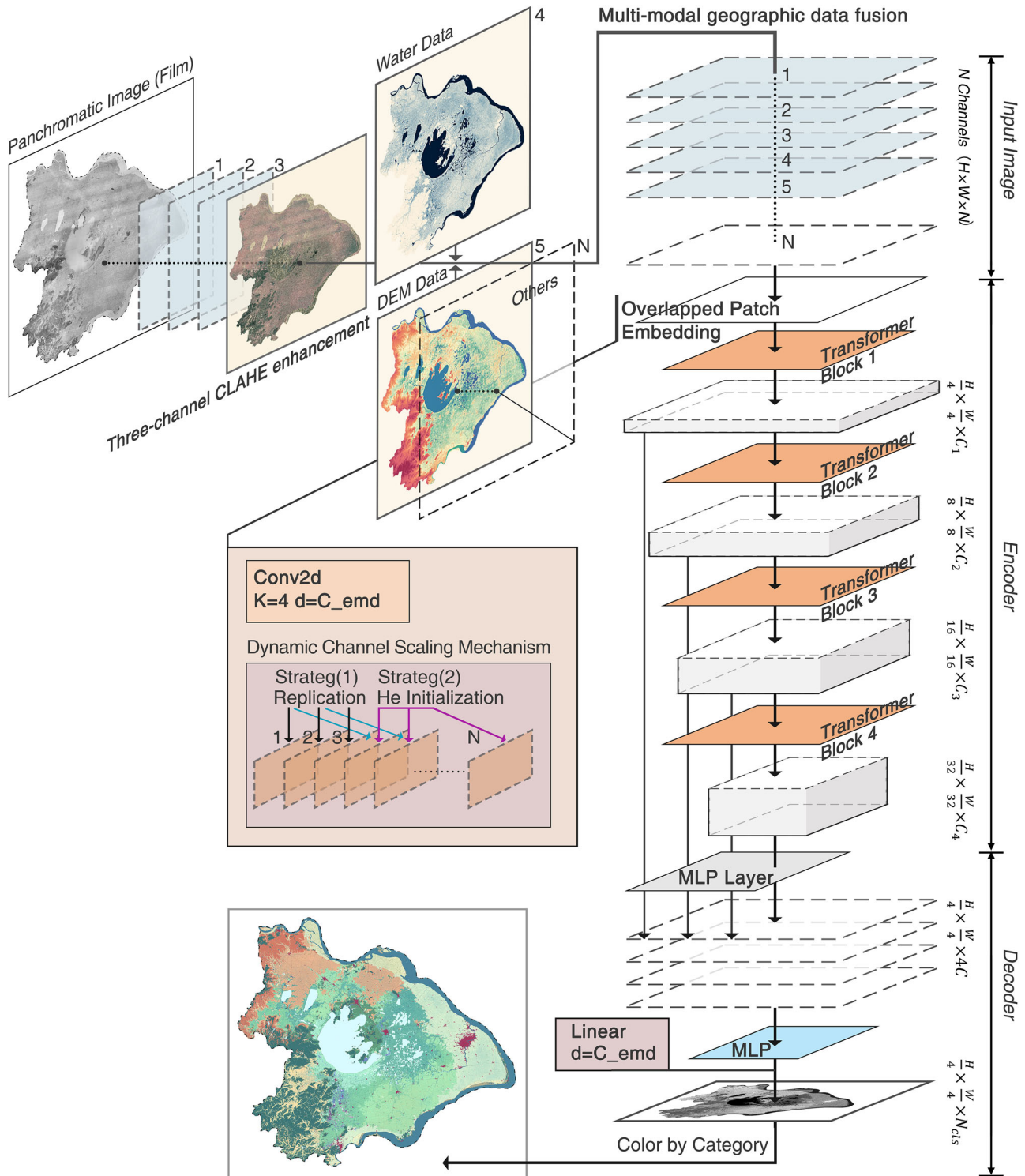


Fig. 4 | Overall structure of the proposed Geo-SegFormer framework. The framework comprises three components: differential three-channel CLAHE enhancement for panchromatic imagery, multimodal geographic data fusion, adaptation and modification of SegFormer.

dimensional space. This process achieves spatial, semantic alignment and aggregation of cross-scale features, rather than a shallow superposition of information from each channel. This design is simple and efficient, and is the key to achieving efficient semantic segmentation.

SegFormer architecture

SegFormer is a Transformer-based semantic segmentation model. As illustrated in Fig. 6, it comprises a hierarchical Transformer encoder omitting

positional encoding and a lightweight ALL-MLP decoder. This design enables the model to achieve a favorable trade-off between global semantic understanding and local detail preservation through multi-scale feature aggregation and computational efficiency optimization¹³.

In the hierarchical transformer encoder, the key advantage is its ability to simultaneously capture high-resolution fine details and low-resolution global contextual information, enabling the model to achieve both high accuracy and efficiency in segmentation at both macro and micro levels. It

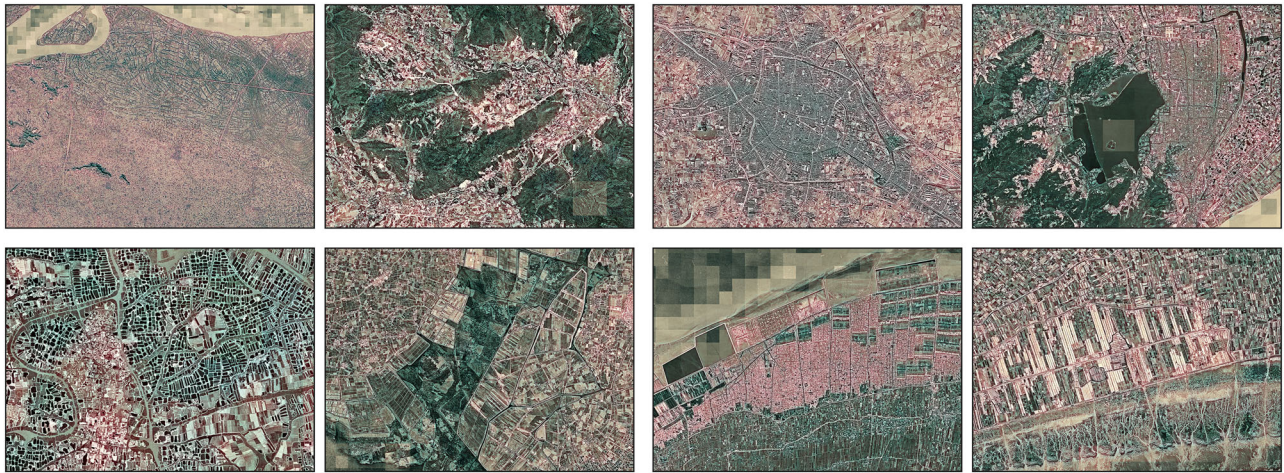
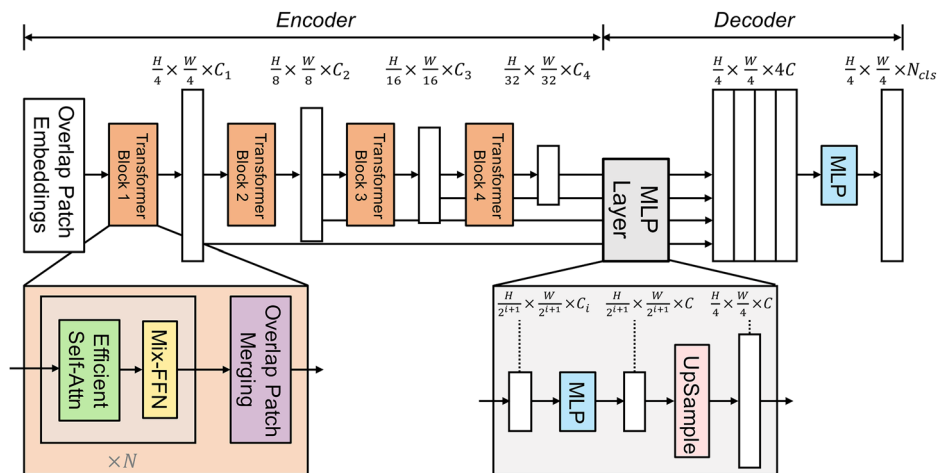


Fig. 5 | RGB mapping of three-channel CLAHE-enhanced images. The figure shows pseudo-colour composites generated from the enhanced panchromatic imagery, where different landscape types appear in contrasting hues.

Fig. 6 | Illustration of the SegFormer semantic segmentation framework architecture. The image is taken from ref. 13. The framework comprises a hierarchical Transformer encoder and a lightweight ALL-MLP decoder.



abandons the reliance on fixed positional encodings found in traditional Vision Transformers (ViTs) and instead constructs multi-scale feature representations through three core components—Overlapped Patch Embedding (OPE), Efficient Multi-head Self-Attention (EMSA), and Mixed Feed-Forward Network (Mix-FFN)—working in concert, with the design of the self-attention mechanism being particularly crucial.

The self-attention mechanism is capable of capturing both high-resolution local details and low-resolution global context. However, its quadratic computational complexity with respect to input size $\mathcal{O}(N^2)$ poses a major bottleneck for processing high-resolution images.

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V \quad (3)$$

$Q, K, V \in \mathbb{R}^{N \times C}$ refers to the Query, Key, and Value matrices. $N = H \times W$ denotes the total number of image patches; d_{head} represents the dimension of each attention head. Therefore, the EMSA module significantly reduces computational complexity to $\mathcal{O}\left(\frac{N^2}{R}\right)$ by applying a preset reduction ratio R (64, 16, 4, 1) across four stages, leveraging convolutional downsampling to compress the sequence length.

Moreover, the Overlapped Patch Embedding (OPE) in the encoder employs an overlapping convolutional downsampling strategy, where the kernel size K , stride S , and padding P are carefully configured to ensure that adjacent image patches overlap at their boundaries. This effectively

mitigates the loss of local continuity caused by non-overlapping patch partitioning in conventional models, thereby enhancing the joint representation of local details and global structure. Meanwhile, the Mixed Feed-Forward Network (Mix-FFN) completely discards positional encoding and instead explicitly captures local spatial information through depth-wise convolutions. This design avoids the performance degradation typically caused by interpolating positional encodings when the input resolution at test time differs from that during training, enabling the model to handle inputs of arbitrary sizes seamlessly.

The decoding process is handled by the lightweight ALL-MLP decoder. The MLP can perform nonlinear fusion of high-dimensional data from each channel, enabling thorough integration and effective utilization of information across different dimensions and channels, while the upsampling and downsampling design further enhances fusion efficiency. Owing to the larger effective receptive field (ERF) provided by the hierarchical Transformer encoder, the decoder of SegFormer can be simplified into a lightweight, fully MLP-based structure, eliminating the need for complex upsampling modules or fine-grained hyperparameter tuning. Its high-dimensional nonlinear fusion proceeds as follows: first, the four multi-scale feature maps output by the encoder are linearly projected into a common dimension; subsequently, all feature maps are upsampled via bilinear interpolation to one-quarter of the input resolution to achieve cross-scale spatial alignment; next, the aligned features are concatenated along the channel dimension and passed through a 1×1 convolution (i.e., a linear layer) to reduce channel dimensionality, thereby lowering computational

cost while simultaneously fusing semantic information across scales; finally, another MLP (or linear layer) nonlinearly integrates the fused high-dimensional features to directly produce the semantic segmentation prediction map. This design efficiently realizes nonlinear multi-scale feature fusion in a high-dimensional space, achieving a balance between performance and simplicity.

$$\hat{Y} = W_2\sigma(W_1[\uparrow(F_1), \uparrow(F_2), \uparrow(F_3), \uparrow(F_4))]) \quad (4)$$

F_i denotes the feature map from the i -th encoder stage, $\uparrow(\cdot)$ represents bilinear upsampling to one-quarter of the input resolution, $[\cdot]$ denotes channel-wise concatenation, W_1 and W_2 are learnable linear transformations (implemented as 1×1 convolutions), and σ is a nonlinear activation function.

Adaptation and modification of SegFormer

First, we introduce the Dynamic Channel Scaling mechanism. The original SegFormer architecture is designed for three-channel input, whereas in this study the input data has a variable number of channels; therefore, the model needs to be equipped with the capability to dynamically adapt to different channel counts. To effectively utilize the pre-trained model while accommodating these multi-channel inputs, this study proposes a systematic approach for input adaptation and pre-trained weight transfer. This mechanism preserves the representational capacity of the pre-trained parameters, while allowing the model architecture to be flexibly adapted to the specific task.

To begin, the input channels are dynamically allocated. In the newly constructed convolutional kernel matrix, the first C_0 channels can directly reuse the pre-trained weights from the original model:

$$W_{\text{new}}[:, C_0 :, :, :] = W_{\text{old}} \quad (5)$$

W_{old} is the original first-layer weight, and W_{new} is the new first-layer convolution kernel matrix. In cases where the input channel count is fewer than three (e.g., grayscale imagery), we employ the following adaptation strategy: the available channel weights are replicated to fill the first three dimensions, while the remaining dimensions are initialized with low-variance weights using He initialization⁴⁵, approximating the distribution of natural image channels.

For input modalities with more than three channels, we propose two complementary strategies to effectively incorporate the additional information while preserving the knowledge encoded in the pre-trained weights.

Channel Weight Replication: This strategy is particularly applicable when the additional input channels share semantic or spectral similarity with the original channels. For instance, in multimodal imagery where channels such as water bodies or vegetation exhibit spectral characteristics similar to the blue or green channels, the weights of the existing channels can be directly replicated to initialize the new ones:

$$W_{\text{new}}[:, C_0 :, :, :] = W_{\text{old}}[:, C - C_0 :, :, :] \quad (6)$$

Channel-wise He Initialization with Scaled Variance: When the additional channels represent new modalities (such as Digital Elevation Model (DEM) or enhanced feature maps) that lack a direct correspondence to the original input channels, simple weight replication may introduce undesirable bias. In such cases, the weights for the new channels are initialized using the Kaiming He initialization to ensure a statistically appropriate distribution. An empirical scaling factor $\gamma = \sqrt{\frac{2}{1+\pi}}$ is introduced to enhance the stability of the initial weight distribution. This factor is motivated by a modified form of He initialization, specifically adapted to maintain consistent variance across layer activations under multi-channel input conditions during forward propagation:

$$W_{\text{new}}[:, C_0 : C, :, :] \leftarrow W_{\text{new}}[:, C_0 : C, :, :] \cdot \gamma \quad (7)$$

Second, we implement dynamic weight loading and parameter migration in the model. The parameters of the SegFormer pre-trained model are only suitable for three-channel input, whereas the new model supports a variable number of channels; thus, the channel compatibility of the parameters needs to be optimized. In the context of transfer learning, aligning the model architecture with the pre-trained weights becomes a critical task. Particularly after modifications to key components such as the input channels, classification head, and decoder structure, the original pre-trained parameters can no longer be directly applied in their entirety. A naive loading of these weights would result in dimension mismatches, gradient anomalies, or even training failure. To address this challenge, a structural compatibility screening mechanism is introduced to selectively map and retain transferable parameters. To enable effective transfer learning under architectural modifications, we propose a structural compatibility screening mechanism. This approach identifies and maps the subset of pre-trained parameters that remain compatible with the adapted model structure, ensuring stable initialization and preserving the knowledge encoded in the original weights.

The structural compatibility mapping can be formulated as:

$$\theta_{\text{compatible}} = \left\{ (k, v) \in \theta_{\text{pre}} \mid k \in \theta_{\text{custom}} \wedge \text{shape}(v) = \text{shape}(\theta_{\text{custom}}[k]) \right\} \quad (8)$$

where θ_{pre} is the parameter set of the original pre-trained model, and θ_{custom} is the parameter set of the current custom model initialization. That is, only parameters with both identical key names and matching tensor dimensions are retained. Subsequently, the compatible parameters are transferred to the corresponding layers of the customized model:

$$\theta_{\text{custom}}[k] \leftarrow \theta_{\text{pre}}[k], \forall k \in \theta_{\text{compatible}} \quad (9)$$

Results

The experiment workflow encompasses landscape classification schema establishment, multi-source dataset construction, model training, and prediction. Performance was evaluated through ablation studies, and the contribution of each modality to the prediction outcomes was quantified using the XGBoost-SHAP explainable machine learning algorithm.

Prior knowledge-based classification schema

The Taihu Basin, which the canal system traverses, has developed distinctive geomorphic zones extending inland from the coastline through fluvial-lacustrine sediment accretion mechanisms. These encompass the delta plain, limnetic plain, water-network plain, slightly high plain, and mountain landform. Through millennia of hydraulic engineering interventions, canal and polder water networks integrated lakes, marshes, and rivers. Concurrently, originally nature-adapted farmlands and settlements were sculpted by these water systems into diverse water-network-structured landscape subtypes. Synthesizing existing research on the formation and structure of natural and anthropogenic landform systems, this study establishes a 35-class canal-associated landscape schema. Detailed descriptions are omitted here due to space constraints.

Dataset construction

Experimental data include historical satellite imagery, DEM, and hydrographic data of the entire Taihu Basin. All data were resampled to 1-m resolution. The historical imagery for the Taihu Basin consisted of KH-9 film-based satellite data dating from 1971, sourced from USGS (<http://earthexplorer.usgs.gov>). After black margin removal and chromatic aberration correction, images were mosaicked and georectified in QGIS. The macro-topography has little difference between ancient and modern times, so the DEM data come from the 30-meter resolution Copernicus DEM of the European Space Agency (<https://doi.org/10.5270/ESA-c5d3d65>)⁴⁶. The

hydrographic data were derived via binary segmentation of the corrected historical imagery.

In the study area, since the actual distribution of various landscape types is uneven, the construction of training samples fully considers the

problem of category imbalance. Among these, agricultural landscapes such as terraced fields and stack fields, along with their derived settlement landscapes are the rarest. Some types are visually indistinguishable, and stack fields have almost disappeared in modern times. Therefore, we

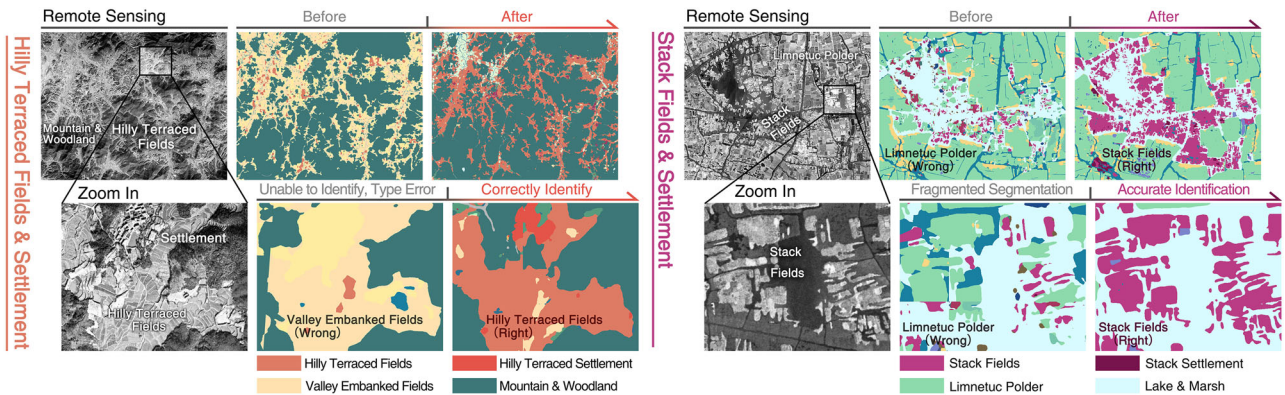


Fig. 7 | Performance of the class balancing strategy. This figure presents a before-and-after comparison of segmentation performance for rare categories at both macro and micro scales.

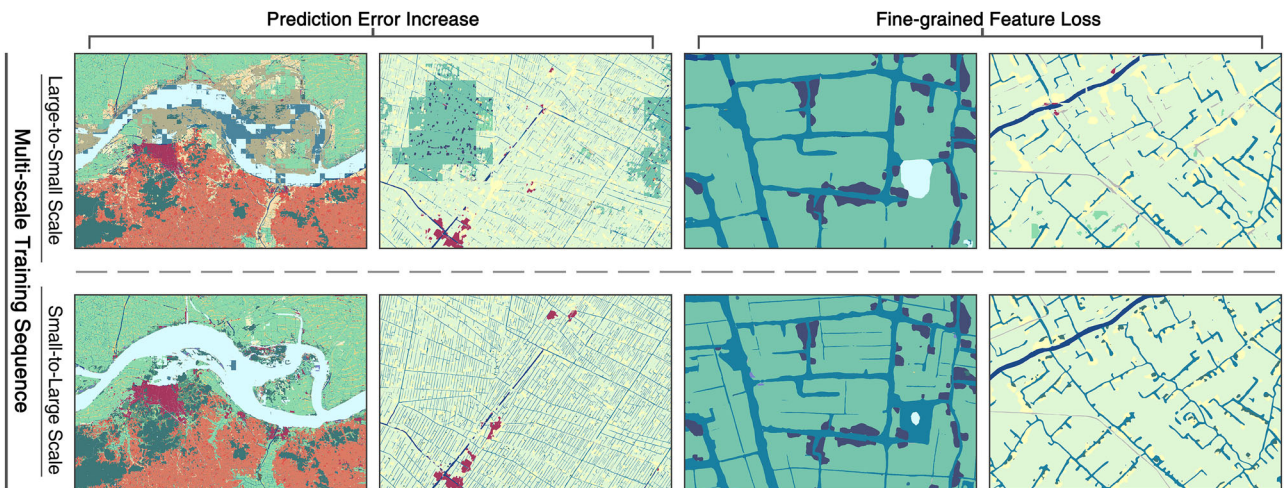
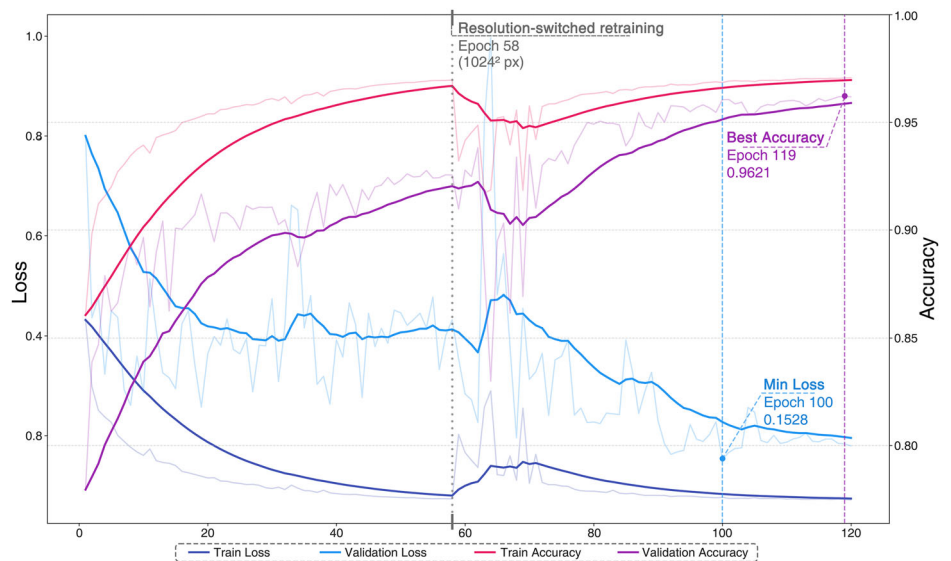


Fig. 8 | Comparison of multi-scale training performance under different input orders. This figure compares the prediction performance of two multi-scale training schemes—large-to-small (first row) and small-to-large (second row)—at both macro and micro landscape scales.

Fig. 9 | Loss and accuracy curves of training and validation sets. The dark blue and light blue lines denote the training and validation loss, respectively, while the red and purple lines represent the training and validation accuracy.



implemented multiple measures to balance the data volume between rare and common categories, including:

1. During sampling, approximate a balance in sample quantities between rare and common categories, guided by actual distribution.
2. For rare types: Expand the collection by supplementing corrected historical images from other years.

3. For existing data: Perform multi-scale cropping and rotation to increase the total sample size for rare types.

After data expansion for rare types, all data were preprocessed and labeled uniformly, including:

1. Manually label all samples to obtain label masks.

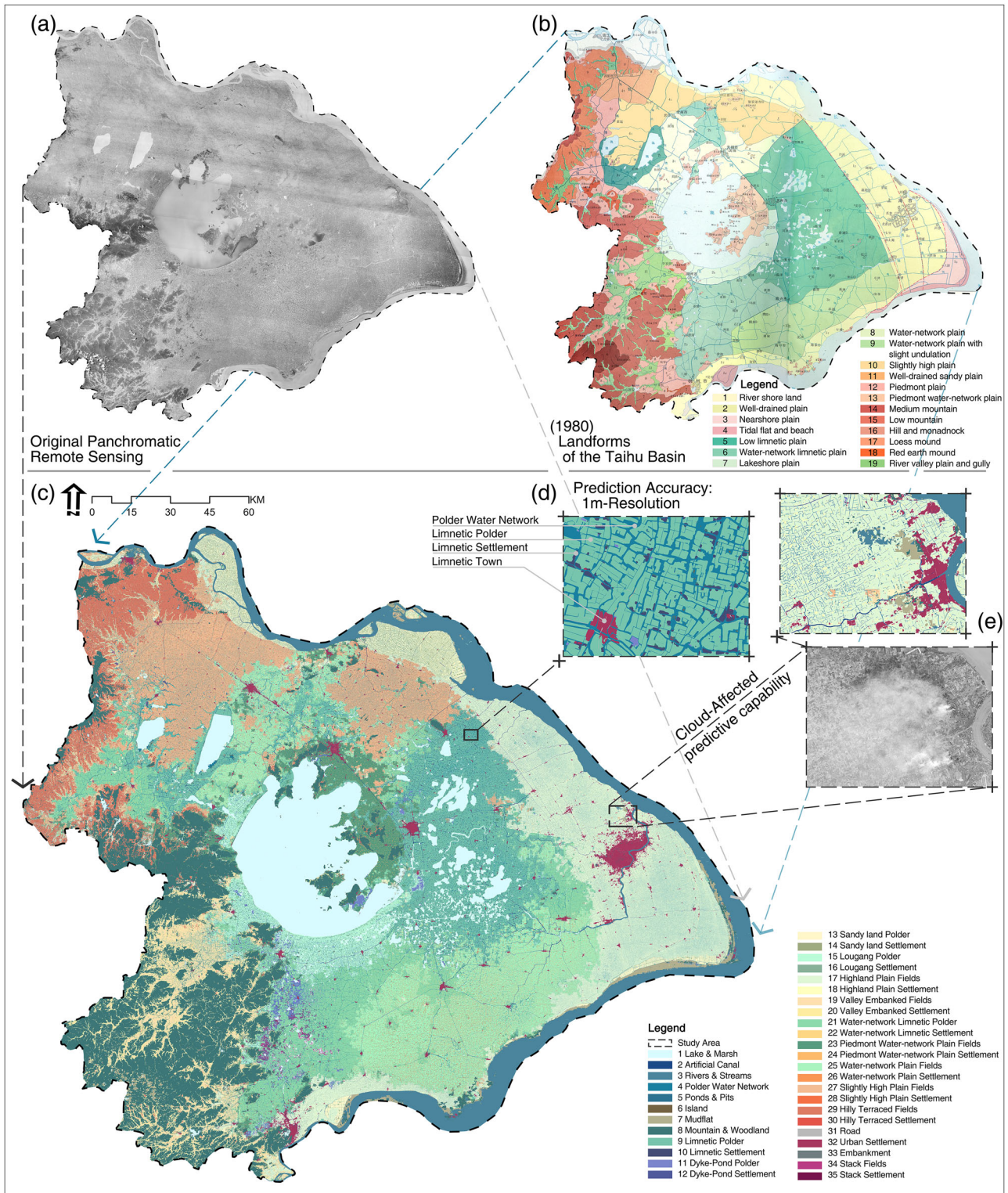


Fig. 10 | Prediction results. a Original panchromatic remote sensing. **b** Landforms of the Taihu Basin (1980). **c** Semantic segmentation result. **d** Demo of 1 m-resolution accuracy. **e** Cloud-affected sample prediction.

2. Apply three-channel CLAHE enhancement to the panchromatic imagery.
3. Fuse the three-channel panchromatic imagery with DEM and hydrographic data into a five-channel dataset.

The final multi-scale training dataset comprised approximately 300,000 samples at 1024×1024 and 256×256 resolutions, with 80% allocated to the training set and 20% to the validation set. Testing demonstrates that the class balancing strategy has markedly improved the interpretation

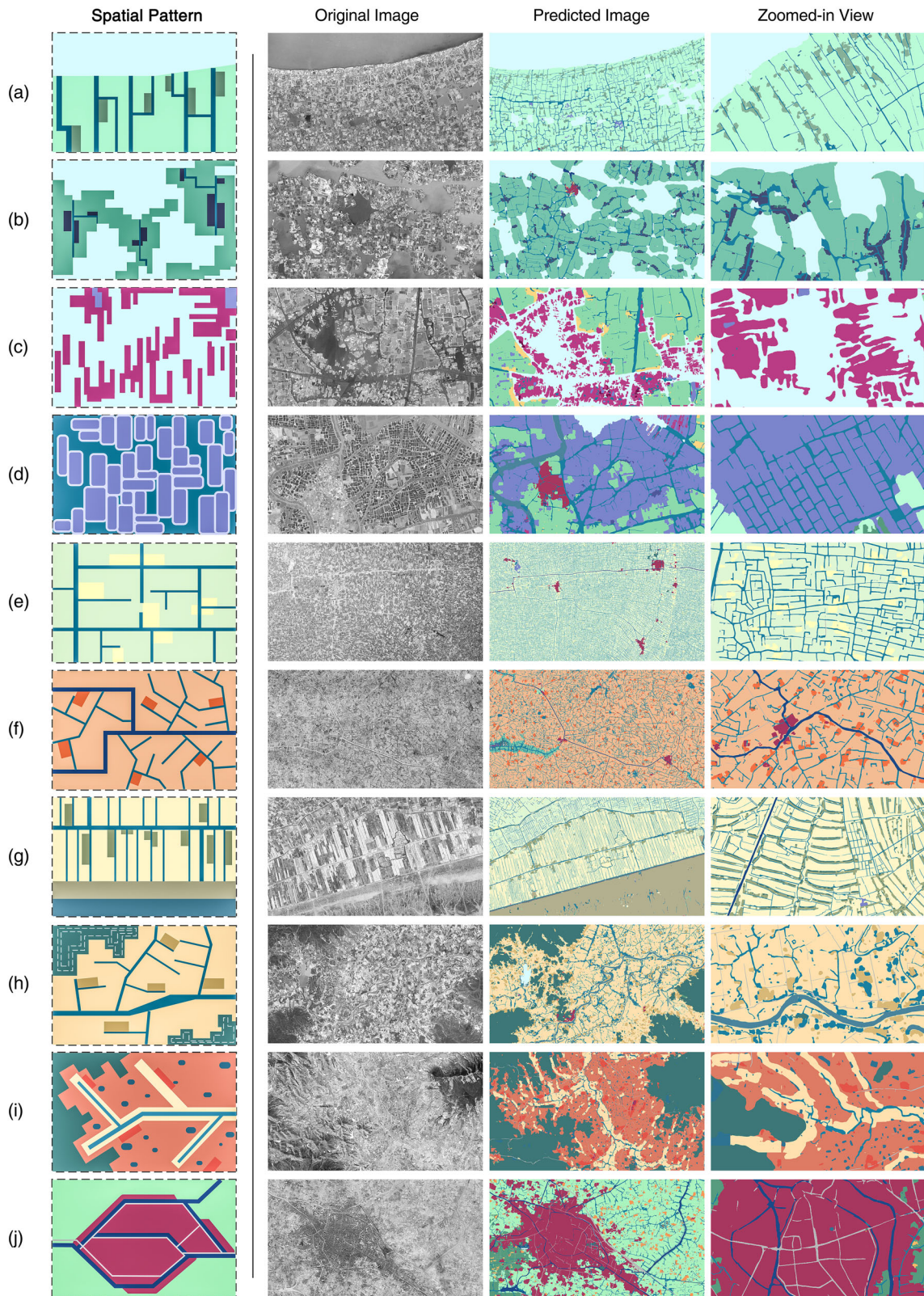


Fig. 11 | Prediction results for representative landscape samples. a Lougang Polder and Settlement. **b** Limnetic Polder and Settlement. **c** Stack Fields and Settlement. **d** Dyke-Pond Polder and Settlement; **e** Highland Plain and Settlement.

f Slightly High Plain Fields & Settlement. **g** Sandy land Polder and Settlement. **h** Valley Embanked Fields and Settlement. **i** Hilly Terraced Fields and Settlement. **j** Urban Settlement.

of rare categories, transforming it from widespread misidentification to precise segmentation with clear boundaries, confirming the efficacy of our approach (Fig. 7).

Model training

Multi-scale training efficacy was first tested with varying input resolution sequences. As Fig. 8 demonstrates, the small-to-large scale training sequence achieves enhanced detail preservation and global prediction capabilities

compared to the large-to-small approach. Accordingly, the small-to-large sequence was adopted as the formal training protocol.

The model was initialized with SegFormer-B4-Finetuned-Cityscapes-1024-1024 to balance accuracy and efficiency. The initial learning rate was set to 5e-5 with manual reduction upon training loss fluctuations. Figure 9 depicts the loss and accuracy of training and validation sets. Lighter curves represent raw values, while darker curves show smoothed values using Exponential Moving Average (EMA). After 58 epochs of small-scale

Table 2 | Comparison of ablation experiment results across different configurations

Model configuration	Kappa	OA	mIoU	mDice (F1)	mDice gain (vs Tri-Channel)
DOM-Water-DEM	0.9560	0.9601	0.8132	0.8778	+2.67%
DOM-DEM	0.9490	0.9539	0.7973	0.8639	+1.03%
DOM-Water	0.9437	0.9491	0.7903	0.8613	+0.73%
Tri-Channel DOM	0.9398	0.9455	0.7813	0.8551	-

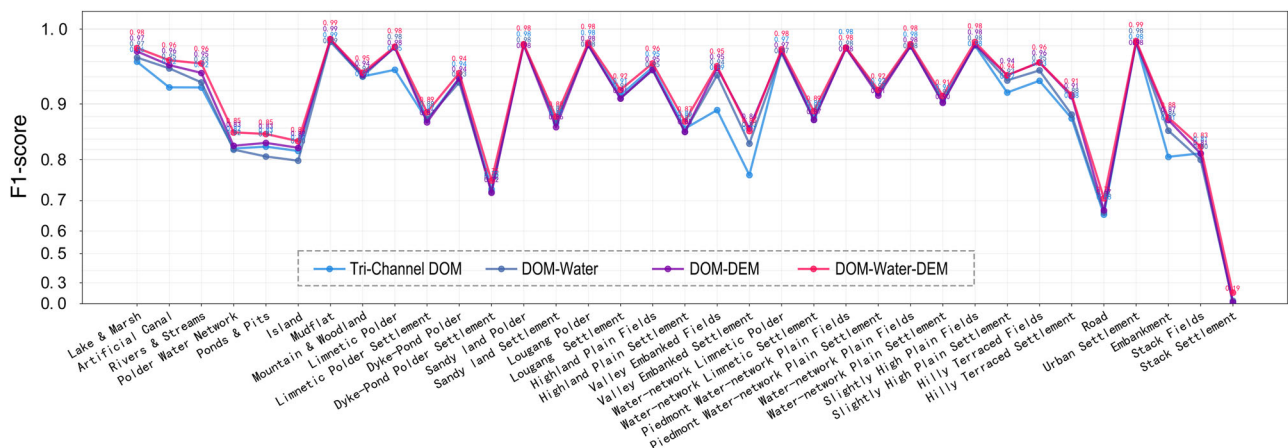


Fig. 12 | Segmentation accuracy in different model configurations. The F1-scores for each configuration are plotted as four colored polylines: Tri-Channel DOM (light blue), DOM-Water (blue), DOM-DEM (purple), and DOM-Water-DEM (pink).

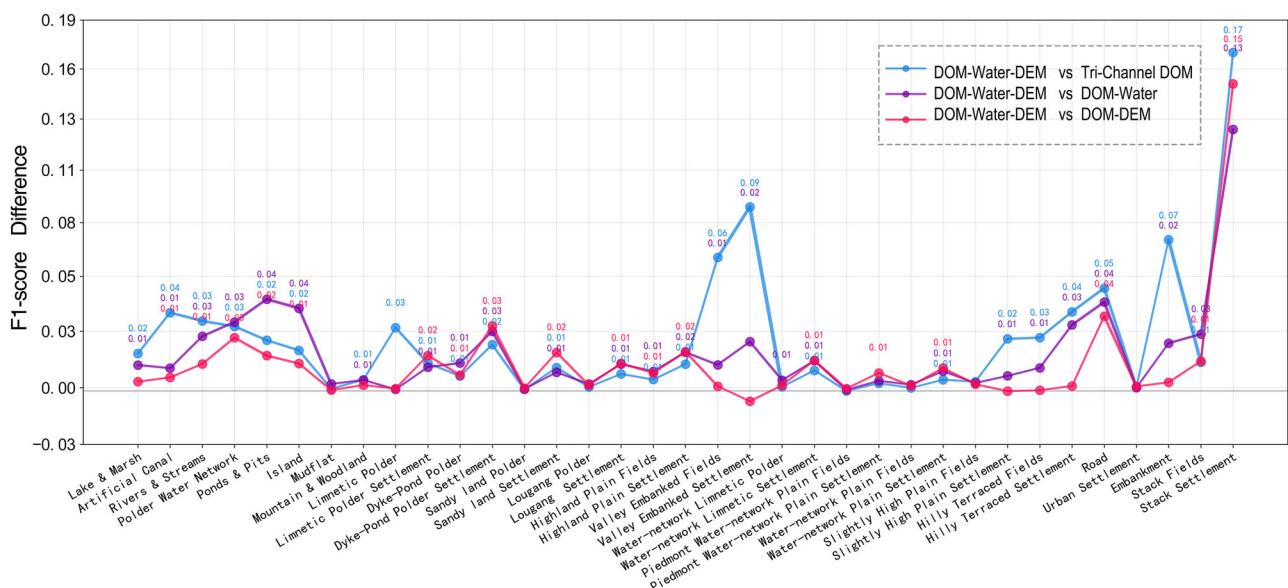


Fig. 13 | Performance differences between the complete model and alternative configurations. This figure compares the DOM-Water-DEM model with three other configurations by plotting their F1-score gaps as colored polylines: Tri-Channel DOM (blue), DOM-Water (purple), and DOM-DEM (red).

training, the training loss reached 0.073 with no overfitting indicated by validation. The training then advanced to 1024-resolution inputs for 62 epochs, achieving stable convergence with final metrics: 0.071 train loss, 0.179 val loss, 0.971 train accuracy, and 0.962 val accuracy, validating high model accuracy. All experiments were executed in PyTorch on an NVIDIA RTX 5090D GPU.

Prediction results

Employing a 50% overlap ratio, the basin-wide imagery was partitioned into 1 × 1 km tiles. Following five-channel preprocessing identical to the training phase, tile-wise segmentation and mosaic stitching generated a 1-m resolution semantic segmentation map of the historical landscape system along the Jiangnan Canal (Fig. 10c). Figure 10a presents the original panchromatic image, while Fig. 10b illustrates the landforms of the Taihu Basin in the 1980s¹⁶. The predictions exhibit strong alignment with the manually delineated landform map, while demonstrating the model’s partial reconstruction capability in cloud-affected regions (Fig. 10d).

Ten representative landscape types are demonstrated in Fig. 11. Each sample includes (left to right): spatial pattern schematic, original image, predicted image, and zoomed-in view. As evident from the chart, the model effectively identifies not only categories with distinctive features but also visually indistinguishable classes that challenge human perception, validating our approach’s efficacy.

Ablation study

To validate the effectiveness of the proposed Geo-SegFormer method for historical landscape semantic segmentation, an ablation study with five model configurations was conducted on the validation dataset. Given the imbalance in the real-world distribution scales of landscape types, four evaluation metrics were adopted: Kappa coefficient, Overall Accuracy (OA), Mean Intersection over Union (mIoU), and mDice (mean Dice/F1-score).

As evidenced in Table 2, the performance hierarchy across configurations is DOM-Water-DEM > DOM-DEM > DOM-Water > Tri-Channel DOM. Figure 12 details per-class segmentation accuracy for all 35 landscape categories, while Fig. 13 quantifies performance gaps between the full multimodal model and other configurations. The Tri-Channel DOM baseline model achieves superior results on texture-sensitive categories. The DOM-Water model significantly enhances segmentation of linear hydrographic features (e.g., canals, rivers), but does not comprehensively surpass the baseline model. The introduction of DEM data significantly improves

F1-scores for 32 categories relative to the baseline model (mean: +2.84%). The full model outperforms alternatives in 34 categories (mean F1-score gain: 2.27%), demonstrating the synergy of multimodal data integration.

The categories with low accuracy are those that are both extremely sparse in distribution and visually indistinct (e.g., dyke-pond settlements, stack-field settlements), yet are retained for taxonomic completeness. Roads, on the other hand, were not prioritized in labeling as they constitute integral components of cultural landscapes rather than standalone types.

The role of multimodal data in interpretation

After the ablation study verified the effectiveness of the proposed model, we introduced the XGBoost-SHAP interpretable machine learning algorithm to further dissect the contribution mechanism of each modality to the Jiangnan Canal landscape interpretation. Leveraging the algorithm’s model-agnostic, post-hoc interpretation capability, we quantified the importance ranking, direction of effect, and contribution intensity of each modality. Utilizing the explainable machine learning algorithm XGBoost-SHAP, this study provides an in-depth interpretation of the proposed model’s segmentation performance across the entire Jiangnan Canal dataset. Figure 14 reveals the ranking of influence importance from different data sources on prediction outcomes, along with positive/negative impacts of data values (feature values colored blue-to-red from low to high). Figure 15 quantifies each feature’s net importance after offsetting positive/negative effects and their cumulative contribution process, where Base Value represents the average prediction across all samples. Comprehensive analysis confirms that all data types contribute positively to predictions. DEM data delivers the highest positive contribution (51.6%), though sparse, low values on the left indicate its limited influence in low-elevation areas. Water data provides the second-highest contribution (17.7%) despite having the lowest global impact due to its binary nature, confirming its critical role in identifying specific landscape categories. Film-based imagery exhibits the widest SHAP value distribution, indicating it provides foundational semantic and frequency-enhanced features while collectively contributing 30.7% (high-frequency: 6.4%; mid-low: 7.1%; original: 17.2%). These results validate the fundamental premise formulated during the classification system’s establishment: the canal-associated landscape is structured upon anthropogenically modified landforms with water networks as spatial frameworks. These findings also align closely with the ablation study conclusions.

Discussion

Deciphering historical landscape composition and accurately reconstructing its full spatial layout are pivotal for addressing heritage sustainability crises and advancing human-land interaction research. However, current studies on the Jiangnan Canal landscapes exhibit significant gaps in both outcomes and methodologies. To bridge this gap, this paper proposes Geo-SegFormer: a framework for the automated semantic segmentation of historical landscapes by integrating deep learning with multimodal geospatial data. Leveraging this dataset, our approach achieves the first 1-m-resolution reconstruction of the entire Jiangnan Canal landscape system across 35 categories, surpassing traditional manual methods in coverage, diversity, and precision.

The canal landscape types identified in this study are not simple geographical classifications. Building upon natural conditions, each landscape type was created through the appropriate modification of nature by human construction, under the multifaceted influences of religion, culture, and

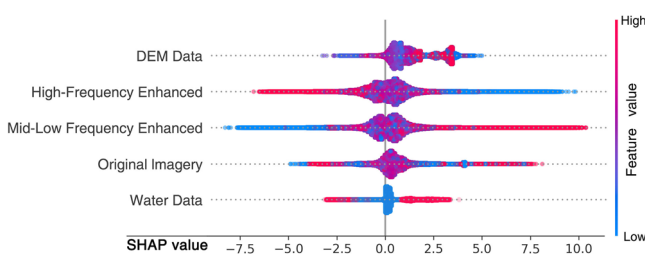


Fig. 14 | SHAP value (impact on Dom-Water-DEM model output). The left panel displays the contribution importance ranking of each modality, while the right panel shows their positive and negative contributions. The color in the right panel represents the magnitude of each modality’s raw values.

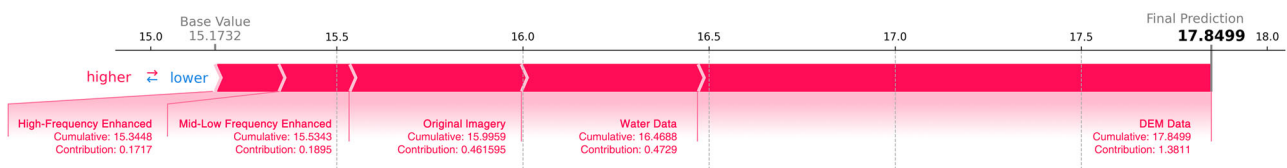


Fig. 15 | Mean SHAP value accumulation process. This figure visualizes the cumulative contribution process of each modality to the model’s predictions and their net importance after offsetting positive and negative effects.

institutions. This classification system reflects the characteristics of both geographical features and cultural landscapes, while also aligning with local traditional cognition. Consequently, the basin-wide interpretation results not only provide a scientific basis for the holistic conservation and perpetuation of the canal heritage but also establish a foundational dataset. This dataset can be integrated with intangible cultural research—including history, religion, folklore, institutions, and handicrafts—to quantify the structural characteristics of the historical canal landscape and reveal the interactive mechanisms between material and intangible heritage from ecological, economic, and cultural perspectives. Thereby, this study propels traditionally text-based historical research into a new framework that integrates qualitative and spatial analysis, enabling a fuller revelation of the construction wisdom and overall value of the canal landscape system.

In addition, Geo-SegFormer effectively provides a solution to the lack of automated identification methods in historical landscape research. At the application level, the framework also demonstrates versatility and scalability. It is applicable to landscape interpretation tasks using panchromatic aerial or satellite imagery across diverse regions and allows for flexible configuration with ancillary data based on specific regional characteristics and task requirements, which allows it to adapt to various analytical scenarios.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available, since they are used under license and will be further researched, so they are not publicly available. Once the study is complete and the results are disseminated, the data may be made available from the corresponding author upon reasonable request.

Received: 11 August 2025; Accepted: 7 December 2025;

Published online: 31 December 2025

References

- Pan, Y. *The Grand Canal and Grain Transport System of the Sui and Tang Dynasties* (San Qin Press, 1987).
- Lu, J. *The Chinese Canal Encyclopedia: Volume on River Engineering and Management* (Phoenix Publishing & Media Group, Nanjing, 2019).
- Jing, M. Development of irrigation and water conservancy and evolution of the relationship between canals and lakes: with the reclamation of Linping Lake during the Tang and Song dynasties as an example. *J. Zhejiang Univ.* **54**, 127–137 (2024).
- Wang, C. et al. The features of spatial distribution of hydraulic facilities along the Grand Canal based on database development. *China Cult. Herit.* **21**, 53–63 (2024).
- Tan, X., Wang, Y., Li, Y. & Deng, J. *Composition and Value Assessment of China's Grand Canal Heritage* (China Water & Power Press, 2012).
- Wan, Y. & Ye, Y. Perception of the value of the historical canal heritage in Jiangsu province under the background of territorial spatial planning. *Shanghai Urban Plann. Rev.* **3**, 78–83 (2023).
- Sang, W., Liu, M. & Shi, T. The grand canal cultural heritage preservation and renewal planning path based on cultural gene theory: the cultural corridor of Jiaying section. *Planners* **40**, 40–47 (2024).
- Tang, B., Zhan, X. & Zhang, J. The agricultural heritage value of the polder system in Gaochun, Nanjing. *Built Herit* **3**, 47–58 (2019).
- Wang, Y., Pendlebury, J. & Nolf, C. The water heritage of China: the polders of Tai Lake Basin as continuing landscape. *Plann. Perspect.* **38**, 949–974 (2022).
- Burnett, M. G. *Hexagon (KH-9) Mapping Camera Program and Evolution* (Center for the Study of National Reconnaissance, 2012).
- Sun, J., Wang, Q. & Guo, W. Origin and characteristics of the traditional dike–polder human settlement system of Xitiaoqi watershed. *Landsc. Archit.* **31**, 64–70 (2024).
- Bulawka, N., Orengo, H. A. & Berganzo-Besga, I. Deep learning-based detection of qanat underground water distribution systems using HEXAGON spy satellite imagery. *J. Archaeol. Sci.* **171**, 106053 (2024).
- Xie, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **34**, 12077–12090 (2021).
- Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794 (ACM, 2016).
- Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
- Nanjing Institute of Geography & Limnology, Chinese Academy of Sciences, Taihu Basin Management Bureau, Ministry of Water Resources. *Atlas of Natural Resources of The Taihu Basin* (Science Press, 1991).
- Ge, H. & Liu, Y. Distribution and attribution of cultural heritage in Jiangsu segment of the Grand Canal. *J. Econ. Water Resour.* **42**, 72–78 (2024).
- Liu, S. The pattern, history, and culture of Jiangnan canal city. *Soc. Sci. Nanjing.* **33**, 47–54 (2022).
- Yang, J. *Evolution of the ecological environment in typical areas along the Grand Canal* (Publishing House of Electronics Industry, 2014).
- Van Strien, M. J. & Grêt-Regamey, A. Unsupervised deep learning of landscape typologies from remote sensing images and other continuous spatial data. *Environ. Model. Softw.* **155**, 105462 (2022).
- Li, N. & Quan, S. J. Discovering urban block typologies in Seoul: combining planning knowledge and unsupervised machine learning. *Cities* **150**, 104988 (2024).
- Wang, J. et al. EO + morphometrics: understanding cities through urban morphology at large scale. *Landsc. Urban Plan.* **233**, 104691 (2023).
- Metzler, A. B. et al. Phenotyping urban built and natural environments with high-resolution satellite images and unsupervised deep learning. *Sci. Total Environ.* **893**, 164794 (2023).
- Henne, A. et al. Ground-truthing of a data-driven landform map in southwest Australia. *CATENA* **248**, 108619 (2025).
- Tao, Y. et al. A graph-based multimodal data fusion framework for identifying urban functional zones. *Int. J. Appl. Earth Obs. Geoinf.* **136**, 104353 (2025).
- Li, Z. et al. Quantitative morphology of polder landscape based on SOM identification model: case study of typical polders in the south of Yangtze River. *Comput. Intell. Neurosci.* **2022**, 1–12 (2022).
- Fleischmann, M. & Arribal-Bel, D. Decoding (urban) form and function using spatially explicit deep learning. *Comput. Environ. Urban Syst.* **112**, 102147 (2024).
- Meng, C. et al. Association between multilevel landscape characteristics and rural sustainability: a case study of the water-net region in the Yangtze River Delta, China. *Ecol. Inform.* **82**, 102677 (2024).
- Shi, Y., Yang, H., & Liu, Z. Large-scale cultural landscape recognition based on deep learning: taking the coastal areas of Jiangsu as an example. *Dev. Small Cities Towns* **42**, 95–102 (2024).
- Cui, Z. Q. & Guo, W. Type distribution identification and morphological characteristic analysis of the sand flat polders in the Yangtze River Delta. *J. Landsc. Archit.* **30**, 117–124 (2023).
- Chen, Y., Lin, W., Hsiao, L. & Cheng, K. A multidecadal change analysis for irrigation ponds in Taoyuan, Taiwan, using multisource data. *Paddy Water Environ* **18**, 1–14 (2020).
- Rizayeva, A., Nita, M. D. & Radeloff, V. C. Large-area, 1964 land cover classifications of Corona spy satellite imagery for the Caucasus Mountains. *Remote Sens. Environ.* **284**, 113343 (2023).
- Shahtahmassebi, A. R., Shahtahmassebi, G., Moore, N. & Atkinson, P. M. Identifying fine-scale archaeological features using KH-9 HEXAGON mapping and panoramic camera images: evidence from Liangzhu Ancient City. *Int. J. Remote Sens.* **45**, 5544–5576 (2024).
- Poterek, Q., Herrault, P., Skupinski, G. & Sheeren, D. Deep learning for automatic colorization of legacy gray scale aerial photographs. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **13**, 2899–2915 (2020).

35. Wang, Y. *Colouring and Interactive Visualization of Historical Earth Observation Data*. Master's Thesis, Technische Universität Dresden (2022).
36. Agapiou, A. Land cover mapping from colorized CORONA archived greyscale satellite data and feature extraction classification. *Land* **10**, 771 (2021).
37. Mboga, M. et al. Fully convolutional networks for land cover classification from historical panchromatic aerial photographs. *ISPRS J. Photogramm. Remote Sens.* **167**, 385–395 (2020).
38. Sertel, E. et al. HexaLCSeg: a historical benchmark dataset from hexagon satellite images for land cover segmentation [software and data sets]. *IEEE Geosci. Remote Sens. Mag.* **12**, 197–206 (2024).
39. Suh, J. W., Ouimet, W. B. & Dow, S. Reconstructing and identifying historic land use in northeastern United States using anthropogenic landforms and deep learning. *Appl. Geogr.* **161**, 103121 (2023).
40. Mboga, N. et al. Domain adaptation for semantic segmentation of historical panchromatic orthomosaics in Central Africa. *ISPRS Int. J. Geo-Inf.* **10**, 523 (2021).
41. Dahle, F., Lindenbergh, R. & Wouters, B. Revisiting the Past: a comparative study for semantic segmentation of historical images of Adelaide Island using U-nets. *ISPRS Open J. Photogramm. Remote Sens.* **11**, 100056 (2024).
42. Reza, A. M. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. *J. VLSI Signal. Process. Syst. Signal Image Video Technol.* **38**, 35–44 (2004).
43. Gonzalez, R. C. & Woods, R. E. *Digital Image Processing* 4th edn (Pearson Education, 2018).
44. Behjati, P. et al. Frequency-based enhancement network for efficient super-resolution. *IEEE Access* **10**, 57383–57397 (2022).
45. He, K., Zhang X., Ren, S., & Sun, J. Delving Deep into Rectifiers: surpassing human-level performance on imagenet classification. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 1026–1034 (IEEE, 2015).
46. Buchhorn, M. et al. Copernicus global land cover layers-collection 2. *Remote Sens* **12**, 1044 (2020).

Acknowledgements

This research was supported by the National Key Research and Development Program of China (Grant No. 2022YFC3800203). The funders played no role in study design, data collection, and writing of the manuscript.

Author contributions

L.R.: Conceptualization, Methodology, Dataset curation, Experiments, Writing, Visualization, Project Administration. K.C.: Methodology, Model development, Writing. S.Z.: Dataset collection, Dataset curation, Translation revision. Q.G.: Dataset curation, Translation revision. Y.G.: Dataset curation. S.T.: Code debugging. Q.L.: Research topic formulation, review, and funding acquisition. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Qing Lin.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025