

Article

Research on the Spatiotemporal-Coupled High-Resolution Remote Sensing Land Use Classification Method

Jiawang Yang ¹, Xiaodong Hu ^{1,*} , Weifeng Ma ¹, Jiancheng Luo ², Tianjun Wu ³ , Zhongbao Shi ¹, Hongfeng Yu ¹, Peijie Jin ¹, Qirui Tan ¹ and Yufei Xu ¹

¹ School of Computer Science and Technology, Zhejiang University of Science and Technology, Hangzhou 310023, China

² State Key Laboratory of Remote Sensing and Digital Earth, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China

³ School of Land Engineering, Chang'an University, Xi'an 710064, China

* Correspondence: huxd@zust.edu.cn; Tel.: +86-1510-150-8309

Highlights

What are the main findings?

- A “spatiotemporal coupling” classification paradigm is proposed, which effectively mitigates the performance degradation in high-resolution remote sensing image classification caused by temporal differences through temporal segmentation and dedicated feature extractor strategies.
- A parcel-oriented deep texture feature extraction and quantification method is designed and implemented, transforming pixel-level features into statistically descriptive vectors with clear geographical significance, significantly enhancing feature discriminability and interpretability.

What are the implications of the main findings?

- This framework provides a scalable solution for fine-grained land cover recognition using multi-temporal, high-resolution remote sensing imagery, particularly suitable for agricultural areas with significant seasonal surface cover variations.
- The proposed technical workflow of “feature extraction–entity constraint–statistical quantification” promotes the effective integration of deep learning models with Geographic Information System (GIS) management needs, offering practical reference value for real-world applications.

Abstract

High-spatial-resolution remote sensing imagery provides a data foundation for fine-grained land use classification. However, due to long revisit cycles and susceptibility to cloud cover, large-area imagery often suffers from temporal inconsistency, which severely limits the classification accuracy of traditional unified models. To address this issue, this study proposes a geographic entity-oriented, spatiotemporally coupled land use classification method for high-resolution remote sensing imagery, with agricultural land (including paddy fields, dry farmland and gardens) as an example for validation. In this method, the study area is first divided into multiple sub-regions based on image acquisition time, ensuring temporal consistency within each sub-region. A dedicated deep texture feature extraction model is then constructed for each sub-region. This model is adapted from the advanced CAPTN texture recognition network: its classification head is removed, and a multi-scale feature fusion module is introduced, transforming it into an encoder focused on extracting spatial texture feature maps. Additionally, a self-supervised loss



Academic Editor: Giorgos Mallinis

Received: 2 January 2026

Revised: 5 February 2026

Accepted: 7 February 2026

Published: 10 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

function combining masked feature reconstruction and cross-view consistency is designed to improve the quality of the learned texture features. During the prediction stage, the corresponding feature extractor is invoked based on the temporal phase of the imagery to generate a full-region texture feature map. This feature map is then cropped using land parcel vectors, and statistical feature vectors describing the texture attributes of each parcel are formed by calculating the mean and standard deviation of the features within each parcel. Finally, a Random Forest classifier is employed to determine the land parcel categories. This study uses the Jiangjin District of Chongqing City as the experimental area. The results show that, compared to training a unified deep learning model directly on full-region multi-temporal imagery or using traditional texture features, the proposed spatiotemporally coupled classification framework achieves significant improvements in overall accuracy and Kappa coefficient, reaching 92.3% and 0.89, respectively.

Keywords: spatiotemporal coupling; land use classification; high spatial resolution remote sensing; parcel; texture features

1. Introduction

Land Use (LU) and Land Cover (LC) are fundamental categories in the study of the Earth's surface system [1,2], intrinsically linked yet distinct in essence. Land cover primarily describes the natural physical covers and artificial constructions on the Earth's surface, whereas land use emphasizes the functional planning and management practices of human societies on land resources. As a critical indicator of human–land interaction [3], the dynamic monitoring and precise identification of Land Use and Land Cover Change (LUCC) hold paramount scientific value and practical significance for understanding the evolution of surface processes [4,5], supporting territorial spatial governance [6,7], and promoting ecological civilization construction [8].

Remote sensing Earth observation technology has become the primary technical pathway for acquiring large-scale land use and land cover information, owing to its advantages of macroscopic, rapid, and periodic monitoring [9]. With the flourishing development of high-spatial-resolution remote sensing satellites, Earth observation capabilities have achieved breakthroughs. Sub-meter resolution imagery can clearly present the fine structure and spatial texture of ground objects, providing unprecedented data support for acquiring detailed LU/LC information and driving the extraction of such information from regional macro-monitoring towards fine-grained management of geographic entities [10]. However, due to the revisit cycle limitations of high-resolution satellites and inevitable cloud interference, acquiring complete, cloud-free imagery of an entire region at a single, ideal phenological phase is extremely difficult [11,12]. In practical applications, it is often necessary to mosaic multiple scenes captured in different seasons or years to cover the entire study area. Consequently, the same ground object can exhibit vastly different spectral and textural characteristics in high-resolution imagery depending on its acquisition time, a phenomenon particularly pronounced in agricultural land. Ignoring this inherent heterogeneity introduced by the temporal dimension and forcibly applying a unified classification model across the entire region forces the model to learn contradictory feature patterns, leading to degraded classification accuracy.

The methodological evolution of remote sensing-based land use classification has witnessed a shift in the fundamental classification unit, progressing from pixels to objects, and further to geographic entities. Early pixel-based classification methods, while algorithmically simple and computationally efficient, often produced significant “salt-and-pepper”

noise when processing high-resolution imagery, failing to meet fine-grained classification needs [13]. Object-Based Image Analysis (OBIA) methods improved spatial coherence by generating geographically meaningful polygon objects via image segmentation, but their accuracy heavily depended on segmentation scale selection. Moreover, the automatically generated object boundaries often deviated from real-world parcel boundaries with clear administrative significance. In recent years, with the increasing availability of multi-source geospatial data, directly using current vector land parcel boundaries as classification units has emerged as a new research trend [14]. This geographic entity-based approach not only effectively avoids the uncertainties of image segmentation but also aligns closely with the practical management needs of land surveys [15].

The feature system for classification has also evolved. Spectral features [16,17], as the most fundamental descriptors of remote sensing imagery, are plagued by the common phenomena of “same object with different spectra” and “different objects with the same spectrum” in complex geographical scenes, making it difficult to accurately discriminate classes using only spectral information [18]. To overcome this limitation, traditional spatial features—such as texture features based on Gray-Level Co-occurrence Matrices and shape features based on geometric morphology—were incorporated into the classification system [19–21]. However, these handcrafted features often capture only shallow spatial patterns of objects, with limited capacity for representing complex spatial structures.

The rise of deep learning methods, represented by Convolutional Neural Networks (CNNs), has brought revolutionary breakthroughs in feature learning for remote sensing imagery. CNNs can automatically learn hierarchical feature representations from raw images through multiple layers of nonlinear transformations, demonstrating far superior representational power compared to handcrafted features. Extensive studies have confirmed that CNN-based methods achieve state-of-the-art performance in various land use classification tasks [22–24]. Simultaneously, researchers have begun exploring the deep integration of temporal dimension information, such as using time-series imagery to capture the unique phenological responses of ground objects, thereby enhancing capabilities for dynamic monitoring and fine-grained category recognition [25].

To address the classification accuracy degradation caused by temporal inconsistency, several methods combining temporal analysis with self-supervised learning have emerged in recent years. For example, Wang et al. [26] proposed a cross-temporal self-supervised learning method that aims to extract denoised homogeneous features from multi-temporal data using a superpixel masking strategy and a cross-temporal attention module. The SeCo-Eco method [27] builds a global, multi-seasonal pre-training dataset to extract universal features from large-scale unlabeled data through self-supervised learning for macro-scale ecological applications. For processing multi-temporal, high-resolution imagery, 3D Convolutional Neural Networks (3D-CNNs) [28] simultaneously model spatial and temporal dimensions using volumetric convolution kernels to capture the spatio-temporal evolution patterns of ground objects.

A careful analysis of existing methods reveals persisting inadequacies in handling the “feature conflict” problem, where the same object exhibits vastly different characteristics across seasons due to multi-temporal image mosaicking. The core limitation lies in the fact that most approaches—whether 3D-CNNs or unified models incorporating cross-temporal attention—still strive to train a single, complex unified model. This model is tasked with reconciling all potentially conflicting feature patterns from different phenological periods throughout the year. This approach imposes a high learning difficulty on the model and cannot guarantee optimal feature representation for each specific temporal phase. Furthermore, existing methods are predominantly based on pixels or coarse-grained objects, failing to organically integrate high-precision land parcel boundaries with clear

management significance and deep feature learning. How to effectively combine the powerful pixel-level representational capacity of deep models with the spatial integrity and managerial attributes of parcel entities, thereby constructing a technical workflow that achieves both high accuracy and meets practical application needs, remains a critical and underexplored research challenge.

To address the issues mentioned above, this study proposes a geographic entity-oriented, spatiotemporally coupled land-use classification method for very high-resolution remote sensing imagery. The main contributions of this research are as follows: (1) A “spatiotemporally coupled” classification framework is proposed. Adopting a “divide-and-conquer” strategy, it divides the study area into sub-regions based on image acquisition time and trains dedicated networks for each, thereby methodologically resolving the interference of temporal inconsistency in entity-based classification caused by image mosaicking. (2) An improved self-supervised Deep Texture Feature Extraction Network (DTFEN) is constructed, which can learn discriminative and robust texture representations from unlabeled imagery and outputs spatially aligned feature maps. (3) A complete technical workflow of “deep feature extraction–vector entity constraint–statistical quantification–parcel classification” is established. By overlaying the feature maps with prior parcel vectors and performing statistical quantification, deep integration between intelligent remote sensing interpretation and geospatial management needs is achieved. (4) Using agricultural land with drastic land-cover changes over time as a typical scenario (Jiangjin District, Chongqing City, focusing on paddy fields, dry farmland, and garden land), comprehensive experiments demonstrate the superiority of the proposed method compared to traditional global models, laying a foundation for its extension to other land-use classification tasks.

2. Materials

2.1. Study Area

This study selects the Jiangjin District of Chongqing as the research area (Figure 1). Located in southwestern China, Jiangjin District is named after its position along the Yangtze River. Geographically, it lies between 105°49′–106°38′E and 28°28′–29°28′N. The region benefits from a humid subtropical monsoon climate, characterized by abundant rainfall and mild temperatures, with an average annual temperature of 19.5 °C and approximately 1141 h of annual sunshine. The fertile soil supports vigorous agricultural development, and the area exhibits the following characteristics: (1) Complex spatial distribution patterns: The terrain of Jiangjin District is rugged and undulating, featuring fold structures of varying sizes that form regular and irregular shapes, which are distributed in clustered or scattered patterns across the region. (2) Balanced development of food and economic crops: The uneven land limits large-scale mechanized farming but promotes the cultivation of specialized economic crops, particularly pepper and citrus. Differences in farming practices and cultivation methods result in unique textural features in high-resolution remote sensing imagery, which facilitates the analysis of current land use information.

As one of Chongqing’s major agricultural producers and an important grain-producing region, understanding the current land use information in Jiangjin District is crucial for promoting sustainable and high-quality agricultural development in the region.

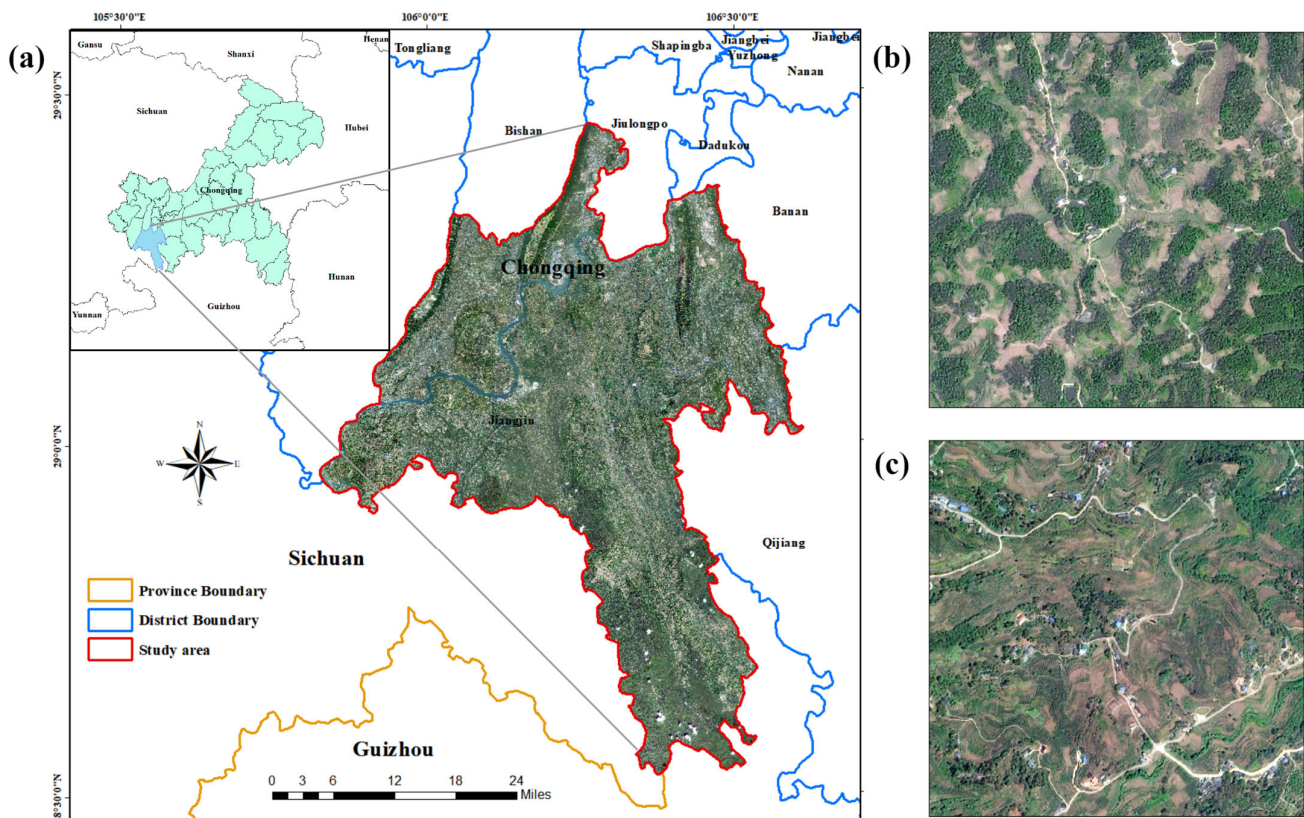


Figure 1. Location Map of the Study Area. (a) shows the location, and (b,c) show samples of different agricultural land use types from high-spatial-resolution remote sensing imagery.

2.2. Data Collection and Processing

The data foundation of this study is high-resolution remote sensing imagery covering the research area. The imagery data are sourced from China's Gaofen-2 (GF-2) satellite, which boasts a panchromatic band spatial resolution of 0.8 m and a multispectral band resolution of 3.2 m, providing an exceptional capability for presenting fine surface structures and spatial textures. The data were specifically acquired through the China Centre for Resources Satellite Data and Application (<https://data.cresda.cn/> (accessed on 12 May 2024)).

To validate the capability of the proposed spatiotemporally coupled classification method in handling multi-temporal inconsistency and to fully capture the key characteristics of agricultural land across different phenological stages, this study did not pursue coverage by a single “ideal” temporal image. Instead, we proactively constructed a multi-temporal image set encompassing several crucial phenological periods. GF-2 imagery from seven months in 2023—namely April, June, July, August, October, November, and December—was screened and acquired. This temporal span covers multiple typical phenological stages in the study area, from the initial growth and development phases to the maturity and post-harvest fallow periods of crops. Consequently, the imagery can reflect significant spectral and textural variations in the same agricultural land-use type across different seasons. During the data screening process, the primary principle was to ensure image quality, prioritizing clear, cloud-free images with minimal cloud cover (below 10%) to maximize the reduction in atmospheric interference and ensure the reliability of surface information extraction.

All acquired raw imagery underwent a systematic preprocessing pipeline. First, orthorectification was performed to eliminate geometric distortions caused by terrain relief and sensor attitude. Subsequently, the Gram-Schmidt (GS) pan-sharpening method was applied to fuse the panchromatic and multispectral bands, generating high-quality color

imagery with a unified spatial resolution of 0.8 m. Single-scene images from different time phases were then separately mosaicked to form complete base maps covering the entire study area for each month. Due to satellite revisit cycle limitations and inevitable cloud cover, localized data gaps still existed. To address this, qualified imagery from adjacent years was used to fill the missing areas. Ultimately, a multi-temporal high-resolution remote sensing image dataset was constructed, characterized by a large temporal span, complete coverage, and consistent geometric precision.

3. Methods and Experiment

3.1. Spatiotemporal Coupling Land Use Classification Method

This study proposes a spatiotemporally coupled land use classification method for high-resolution remote sensing imagery. It departs from the traditional approach of using pixels or image objects as classification units. Instead, it employs pre-extracted land parcel vectors—which hold practical geographical significance—as the fundamental, indivisible units of analysis. This ensures that the classification results are strictly aligned with real-world management entities. To address the issue of temporal inconsistency across the entire study area caused by crop phenology, seasonal changes, and varying imaging conditions, this study divides the mosaicked imagery covering the region into multiple temporal segments based on image acquisition dates. Each segment contains internally consistent temporal and phenological characteristics. Building upon this division, an independent deep texture feature extractor is trained for each temporal segment, thereby constructing a temporally adaptive feature extractor library. During parcel classification, the most discriminative deep texture features are extracted by adaptively selecting the corresponding feature extractor based on the dominant temporal phase of the imagery within each parcel. Subsequently, statistical methods are employed to aggregate and quantify the pixel-level texture features inside each parcel, forming a high-dimensional feature vector that characterizes the parcel's texture pattern. Finally, a machine learning classifier is applied to perform fine-grained land use type classification (such as paddy field, dry farmland, and gardens, as focused on in this study) at the parcel level. The overall technical workflow of the spatiotemporally coupled high-resolution remote sensing land use classification method is illustrated in Figure 2.

3.1.1. Farmland Parcel Extraction Based on Zoning and Layering

The strategy of zoning and layering has been widely applied in complex remote sensing image segmentation, showing significant results [29–31]. Based on the Digital Elevation Model (DEM), slope data, and the spectral characteristics of the high-resolution remote sensing imagery itself, Jiangjin District was partitioned into several sub-regions with relatively homogeneous internal terrain, landforms, and cropping patterns.

To validate the effectiveness of the zoning and its guarantee that “each zone is unique and has a reasonable topographic distribution,” we conducted two assessments: (1) Intra-zone Variance Analysis: The variance of all input features (standardized) within each zone was calculated. The results show that the variance of features within the sub-zones was reduced by approximately 65% on average compared to the overall variance of the entire study area. This quantitatively demonstrates that the zoning effectively enhanced the homogeneity within each unit. (2) Spatial Continuity Evaluation: We calculated the Spatial Autocorrelation Moran's Index to assess the smoothness of the zoning results. A higher positive Moran's Index value indicates that spatially adjacent units are more likely to be assigned to the same zone. This aligns with the practical pattern of spatial continuity of geographic entities, proving that the zone boundaries are not generated randomly but possess geographical rationality.

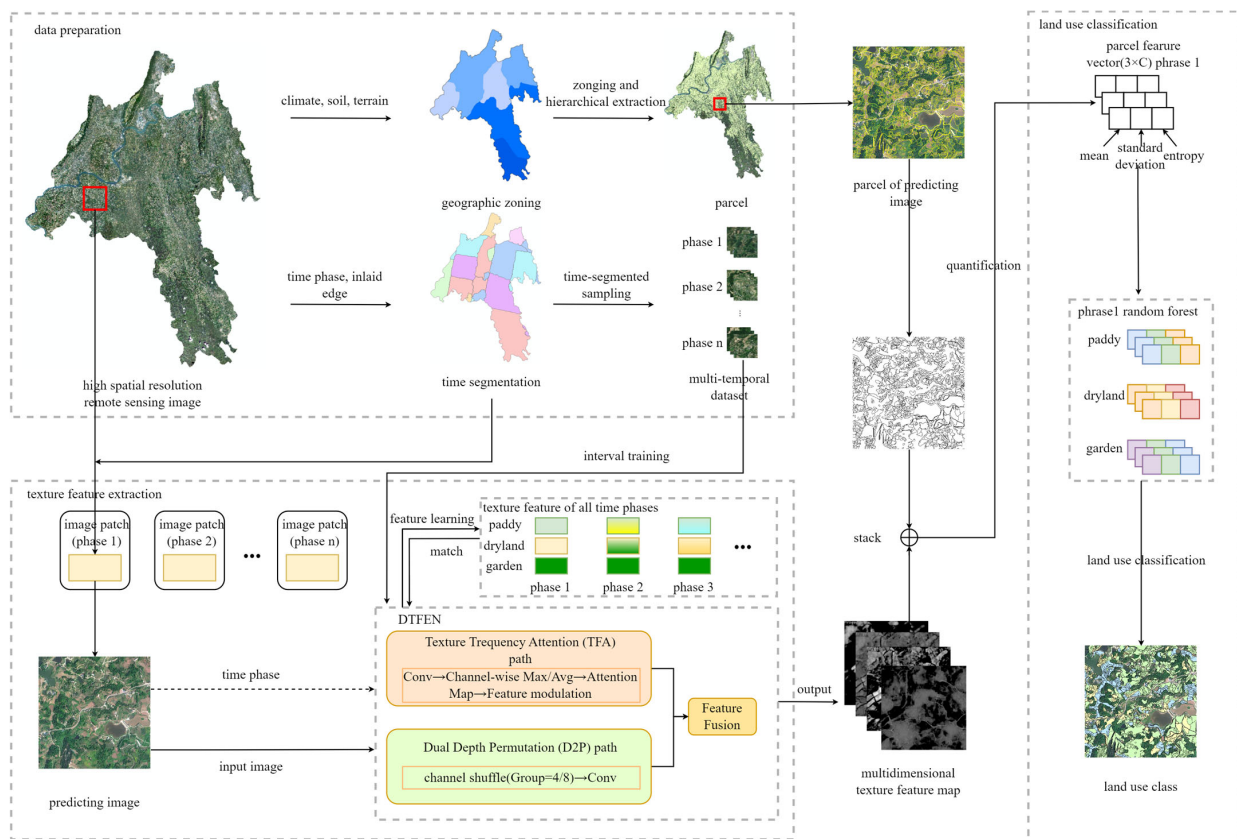


Figure 2. The technical workflow of the spatiotemporally coupled high-resolution remote sensing land use classification method.

Within each geographical zoning, a layered extraction was performed to obtain farmland parcels, as shown in Figure 3. This study uses these farmland parcels as analytical units, with each parcel defined as P_i (where $i = 1, 2, 3, \dots, N$, and N is the total number of parcels).

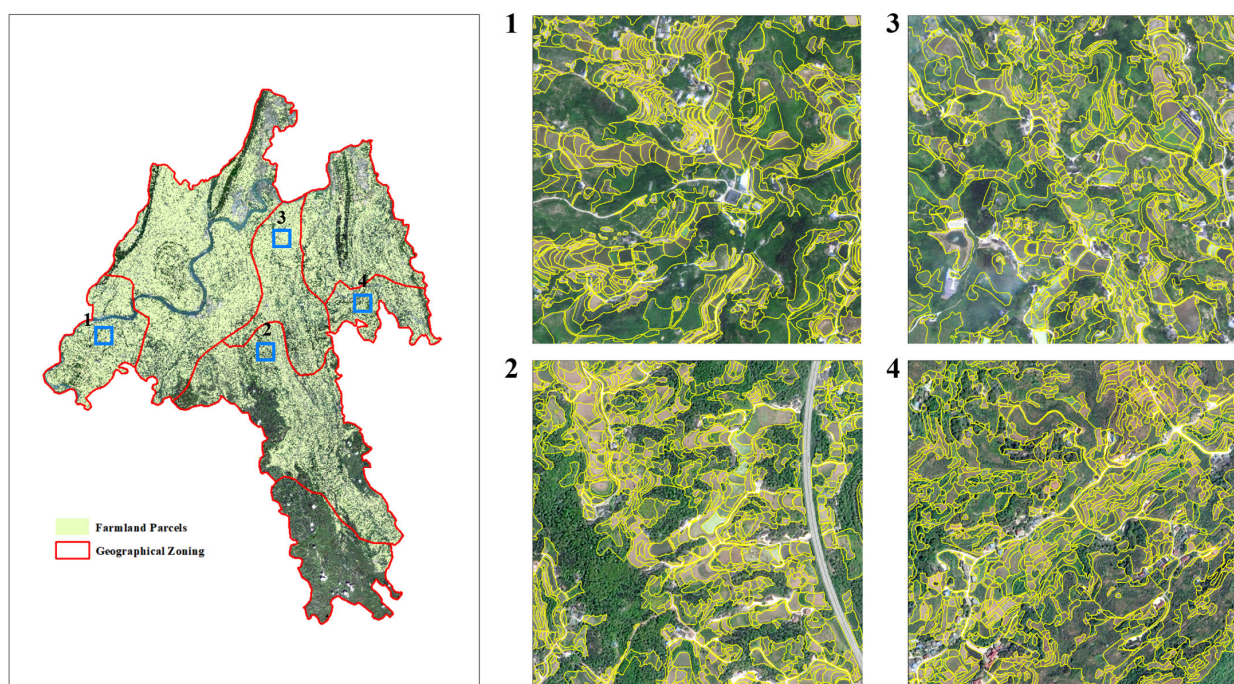


Figure 3. Geographical zoning and farmland parcels.

Although land parcels have been extracted through zonal and hierarchical methods with existing broad land-use category information, distinguishing more detailed land-cover classes is essential to meet the demands of refined land management. Furthermore, despite the availability of corresponding vector boundaries for these broad land-use categories, the actual land-use types within individual parcels may still be uncertain. For instance, confusion may exist between cropland and grassland, or between gardens and forest land. Therefore, it is necessary to conduct further analysis within the parcels to extract highly discriminative feature patterns.

3.1.2. Temporal Segments

High-spatial-resolution remote sensing images are inherently limited by their imaging mechanisms, resulting in a single scene covering only a small area. Therefore, a full-coverage image of a study area often needs to be constructed by mosaicking multiple images acquired at different times. This data characteristic leads to a situation where different parts of the final mosaic may originate from different acquisition dates. For agricultural land, there are significant differences in its textural features across various crop growth stages. Sampling and training a model directly on the entire mosaic would cause the model to learn conflicting textural cues, thereby degrading its classification performance. To address this issue, this study divides the full-coverage image, which is mosaicked from N scenes, into several temporal segments, as illustrated in Figure 4. Samples are then drawn, and models are trained independently for the images within each segment.

3.1.3. Deep Texture Feature Extractor Network (DTFEN)

To extract discriminative texture features from high-spatial-resolution remote sensing imagery for parcel-level classification, this study introduces improvements based on the Chebyshev Attention Depth Permutation Texture Network (CAPTN) [32], designing a Deep Texture Feature Extraction Network (DTFEN).

CAPTN was originally designed for general natural image texture classification. Its core innovation lies in modeling texture from two complementary perspectives through two key modules: the Texture Frequency Attention (TFA) module, which explicitly guides the model to focus on discriminative texture regions in the spatial domain, and the Dual Depth Permutation (D^2P) module, which implicitly encourages learning more robust and generalizable feature representations by randomly shuffling feature channels. These designs are universal at the level of fundamental texture analysis principles. In high-resolution remote sensing imagery, land-cover textures similarly manifest as visual patterns composed of primitives such as edges, lines, and spots with specific spatial distribution characteristics. For instance, the key to distinguishing regularly gridded paddy fields, directionally furrowed dry farmland, and dot-matrix distributed gardens lies precisely in capturing and differentiating these unique texture patterns. Therefore, the core philosophy and modules of CAPTN dedicated to texture modeling provide a powerful foundational tool for parsing the textural semantics of land cover in complex remote sensing scenes. However, as a classification network, CAPTN's final output is an image-level category label. This does not align with the core requirement of this study: to output a texture feature map that is spatially aligned with the input image and suitable for subsequent parcel-vector cropping and statistical quantification.

Therefore, this study refactors CAPTN, transforming it from a texture classifier into a self-supervised deep texture feature extractor. Improvements involve stripping away the original model's classification-related components, retaining and strengthening its core texture modeling capabilities, and designing self-supervised optimization objectives targeted at the quality of the feature map itself. The structure of the improved DTFEN

model is shown in Figure 5, and its workflow along with the core improvements are as follows.

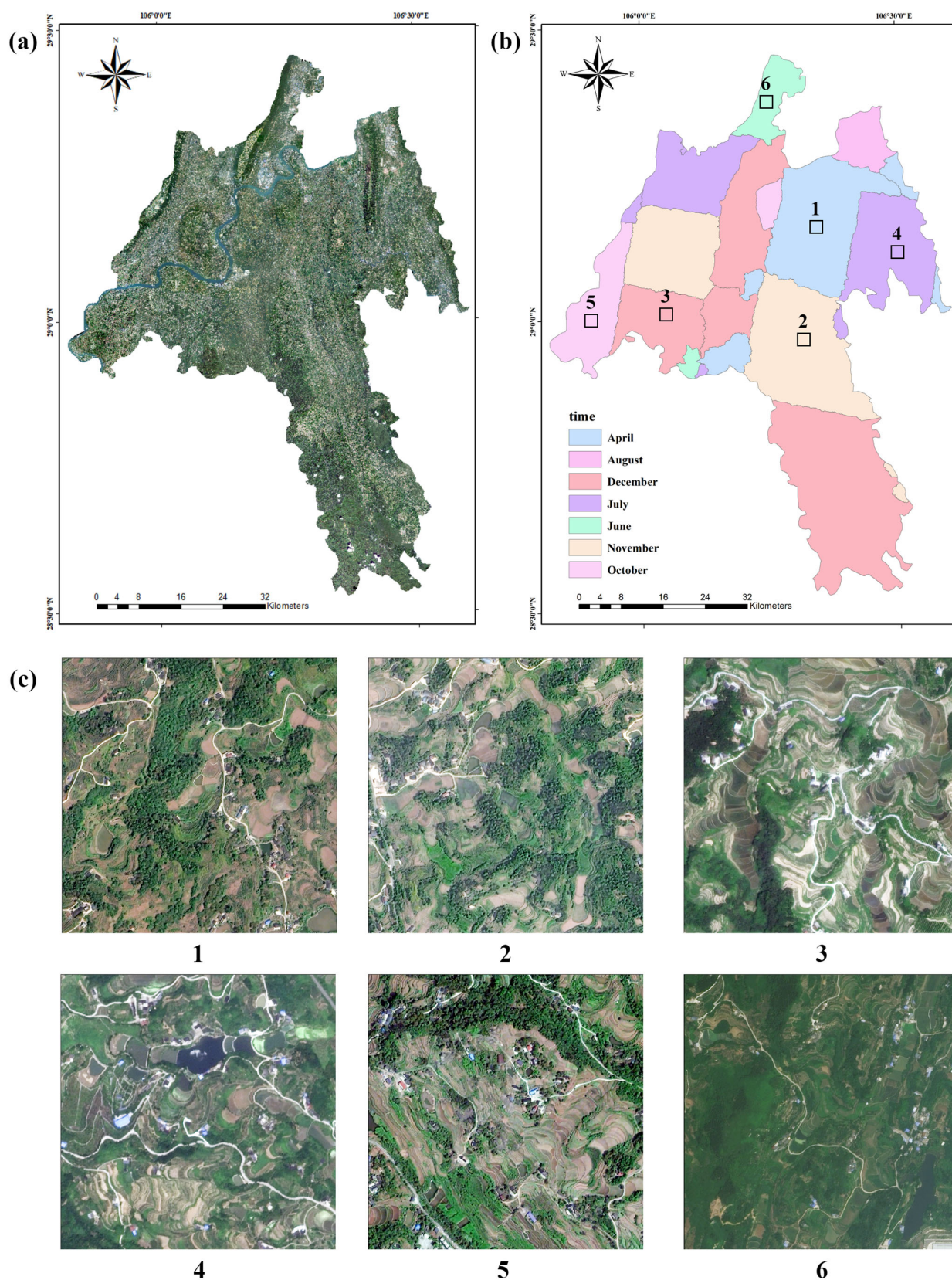


Figure 4. Temporal segments. (a) shows the study area, (b) shows temporal segments, and (c) shows images from different time segments.

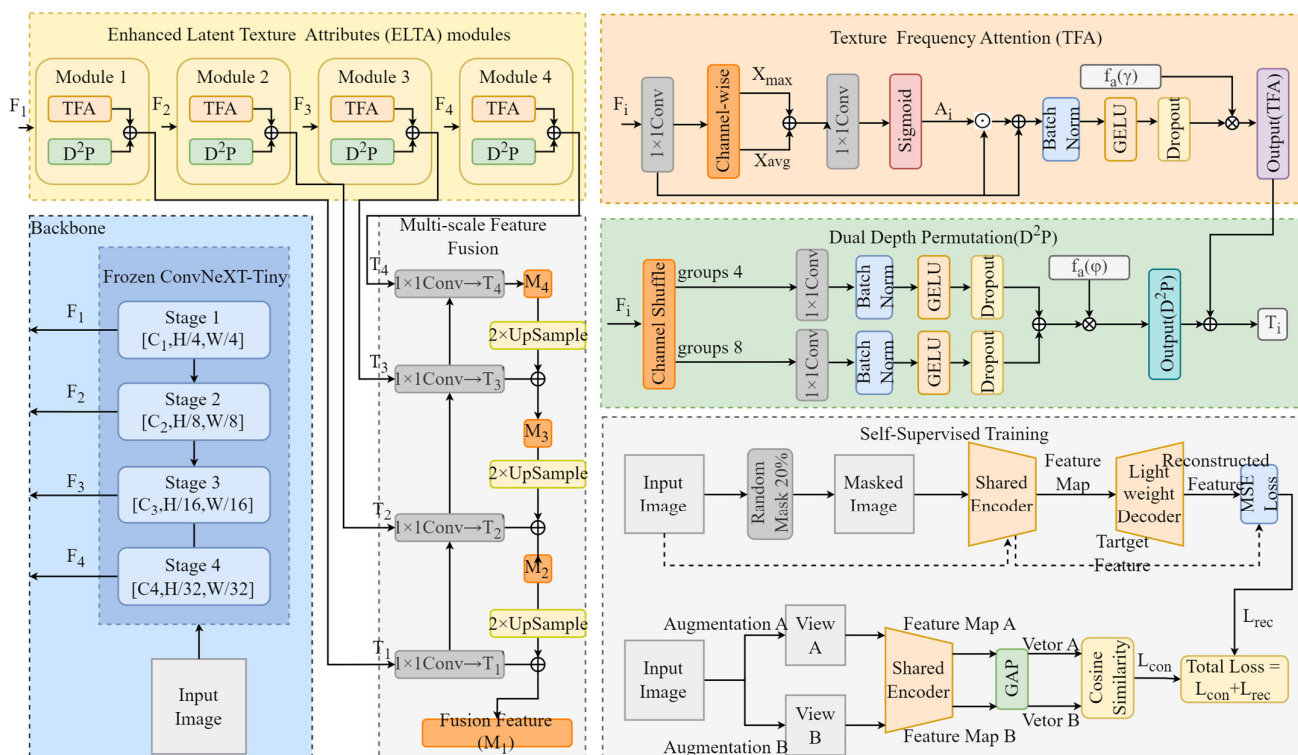


Figure 5. Architecture of the DTFEN Model.

This study removes the two final key components of the original CAPTN model designed for classification: the Learnable Chebyshev Polynomial (LCP) aggregation layer and the linear Classifier. The function of the LCP layer is to compress spatially distributed texture features into an orderless global vector—a process that discards the spatial location information crucial for parcel-level classification. By removing these components, the network’s output is preserved within the spatial domain, directly generating a high-dimensional deep texture feature map, $M \in \mathbb{R}^C \times H \times W$. Consequently, the network’s output is no longer a single category label but rather a “feature library” rich in texture information. This enables precise spatial overlay and feature cropping with parcel vector polygons from a Geographic Information System (GIS), establishing the foundation for subsequent statistical quantification.

To comprehensively utilize both the detailed texture information from shallow convolutional neural network features and the semantic context from deep features, this study designs and incorporates a lightweight feature pyramid fusion module. This module receives texture features from different stages of the backbone network, enhanced by both TFA and D²P. By introducing a top-down pathway and lateral connections, the rich semantic information from deep features is fused into the spatially higher-resolution shallow features via element-wise addition.

To drive the network to learn texture features beneficial for the classification task, this study designs two complementary self-supervised tasks: the Masked Texture Feature Reconstruction task and the Cross-View Texture Consistency task, along with their corresponding loss functions. These two tasks were applied simultaneously to train DTFEN from scratch.

The Masked Texture Feature Reconstruction task aims to force the model to understand the internal structure and spatial contextual relationships of textures. During training, random regions of the input image are masked, and the masked image is then fed into the network. The network must not only extract features but also, through an attached lightweight decoder, predict the representation of the masked regions in the feature space

based on the contextual features from the unmasked areas. Its loss function, $\mathcal{L}_{rec}$, calculates the mean squared error between the predicted features and the ground-truth features (obtained from a forward pass of the complete image) over the masked regions. This task guides the model to learn the holistic “gestalt” of textures, thereby extracting more semantically coherent features.

The Cross-View Texture Consistency task aims to enhance the invariance of features to variations in imaging conditions (e.g., illumination, color, minor viewpoint changes), focusing on the stable attributes of texture. For the same training image, we generate two views augmented with different random data augmentations (e.g., color jitter, small-scale rotation/cropping) and feed them separately into the network. The loss function $\mathcal{L}_{con}$ requires the feature vectors—obtained from the network’s processing followed by global average pooling of these two views—to be as similar as possible in cosine space. This drives the network to ignore appearance changes unrelated to the intrinsic nature of the objects, yielding more stable and robust feature representations.

The total training loss for the model is the weighted sum of the two:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{con}, \quad (1)$$

where λ is a balancing hyperparameter.

Through the improvements described above, we successfully integrate CAPTN’s texture modeling capabilities with the requirements of high-resolution remote sensing image processing, self-supervised learning, and geospatial entity analysis. Ultimately, the DTFEN becomes a feature extractor capable of autonomously learning high-quality, multi-scale texture features from unlabeled imagery.

3.1.4. Texture Feature Quantification

For any parcel P_i and its corresponding image patch $X_{i,t}$ at a given time phase, the texture feature map $F_{i,t}$ is extracted through a deep learning model. To transform these texture features into a feature vector with clear geographical significance, the following three statistical measures are computed for each channel’s feature map $F_{i,t}^{(c)}$:

Mean: Represents the average intensity of the texture feature within the parcel. The mean for

$$M_{i,t}^{(c)} = \frac{1}{H \cdot W} \sum_{h=1}^H \sum_{w=1}^W F_{i,t}^{(c)}(h, w). \quad (2)$$

each channel is expressed as:

Standard Deviation: Indicates the contrast of the texture feature within the parcel. The standard deviation for each channel is expressed as:

$$S_{i,t}^{(c)} = \sqrt{\frac{1}{H \cdot W} \sum_{h=1}^H \sum_{w=1}^W \left(F_{i,t}^{(c)}(h, w) - M_{i,t}^{(c)} \right)^2}. \quad (3)$$

Entropy: Represents the complexity or disorder of the texture feature distribution. The feature values are first discretized into L levels to form a probability distribution $p(l)$. The entropy for each channel is expressed as:

$$E_{i,t}^{(c)} = - \sum_{l=1}^L p(l) \cdot \log_2 p(l). \quad (4)$$

To clarify the theoretical correspondence between these statistics and the visual texture patterns exhibited by typical agricultural land types, and to provide prior knowledge for subsequent classification, we summarize the key relationships in Table 1, using the growing season as an example. The table explains, from a geographical perspective, that the Mean (M) reflects the average reflectance intensity of the texture, the Standard Deviation (S)

characterizes the internal contrast and heterogeneity within a parcel, and the Entropy (E) describes the complexity and disorder of the texture structure. Different land use practices shape distinct spatial patterns, which in turn manifest as separable numerical patterns through these three statistical measures.

Table 1. Texture patterns and statistical quantification characteristics of main agricultural land types.

Category	Dominant Texture Pattern	Statistical Quantification Features (Mean M /Std S /Entropy E) and Discriminative Implications	Representative Sample
Paddy	Regular grid-like uniform texture formed by field ridges.	- Mean M: Moderate, reflecting the average reflectance intensity of vegetation and ridges. - Standard Deviation S: Low. High internal homogeneity and low contrast, a key indicator for distinguishing paddy fields. - Entropy E: Low. Regular, ordered structure with low spatial complexity.	
Dryland	Directional striped texture due to crop rows and ridges.	- Mean M: Moderate, influenced by crop coverage. - Standard Deviation S: Medium. Periodic medium contrast arises from differences between inter-row soil and crop bands. - Entropy E: Medium. Directional structure exists, but with some management-induced variation.	
Garden	Granular dot-matrix texture formed by discrete tree canopies.	- Mean M: Can be relatively high or variable, depending on canopy closure. - Standard Deviation S: High. Significant brightness differences between canopy clusters and shaded gaps result in the highest internal contrast. - Entropy E: High. Random spatial distribution and the most complex, disordered structure.	

3.1.5. Spatiotemporal Coupling Feature Construction

Based on the three statistical measures mentioned above, each parcel obtains a $3 \times C$ -dimensional feature vector $V_{i,t}$ at its corresponding time phase:

$$V_{i,t} = [[M_{i,t}^{(1)}, S_{i,t}^{(1)}, E_{i,t}^{(1)}], \dots, [M_{i,t}^{(c)}, S_{i,t}^{(c)}, E_{i,t}^{(c)}]]. \quad (5)$$

3.1.6. Random Forest-Based Classification

The Random Forest (RF) algorithm was employed as the classifier due to its effectiveness in handling high-dimensional features and its capability to provide feature importance assessments. RF learns the complex discriminant rules described earlier by constructing a multitude of decision trees. Each tree identifies optimal split points using different subsets of features and temporal phases, and the final robust decision is made through ensemble voting.

A dataset was constructed by pairing the spatiotemporal coupling feature vectors $V_{i,t}$ with their corresponding ground truth land class labels. This dataset was then split into training and testing sets in a 7:3 ratio. A classifier was trained on this data to classify all farmland parcels across the entire Jiangjin District.

3.1.7. Multi-Temporal Feature Integration Strategy

The foundational workflow of the “spatiotemporally coupled” framework proposed in this study is to invoke the corresponding feature extractor for each parcel based on its dominant temporal phase, generating and utilizing a single-temporal feature vector for classification. This process effectively overcomes the model learning conflicts caused by the mosaicking of multi-temporal imagery across the entire region.

To further explore the potential of integrating multi-seasonal information to enhance classification accuracy, this paper designs and implements a multi-temporal feature integration strategy as a comparative experiment. The core premise of this strategy is that, where actual data permit—meaning the same parcel is covered by cloud-free imagery at multiple different temporal phases within a year—we can construct a more informative composite feature vector for that parcel.

The specific implementation steps are as follows: (1) Multi-Temporal Data Alignment: For a given parcel, identify which temporal segments, among all defined, have available, quality-qualified imagery coverage. (2) Multi-Temporal Feature Extraction: For each available temporal phase of the parcel, invoke its corresponding DTFEN model to generate the pixel-level texture feature map for that phase. This map is then cropped and statistically quantified based on the parcel’s vector boundary to obtain the feature vector for that phase. (3) Temporal Feature Concatenation: The feature vectors calculated for the parcel across all available temporal phases are concatenated in chronological order, forming a longer composite feature vector that integrates multi-temporal texture information. (4) Classifier Training and Prediction: A new Random Forest classifier is trained using the composite feature vectors of the parcels and their corresponding labels. During prediction, the composite feature vector for each parcel is similarly constructed and used for classification.

The primary distinction between this strategy and the basic single-temporal “spatiotemporally coupled” workflow lies in the fact that the latter uses features from only a single temporal phase per parcel, whereas the former integrates features from all available temporal phases for the parcel. This enables the capture of complementary texture patterns exhibited by the same parcel across different phenological stages (e.g., fallow period, growth stage, maturity stage).

3.2. Experiment

To clearly elucidate the experimental workflow and data usage in this study, this section provides a detailed explanation of the dataset composition, division methods, and specific purposes involved in model training and evaluation. The data usage in this research spans two core stages: (1) the self-supervised training of the Deep Texture Feature Extraction Network (DTFEN), and (2) the training and testing of the Random Forest (RF) classifier based on parcel feature vectors. These two stages utilize datasets of different natures and for distinct purposes.

The self-supervised training of DTFEN aims to enable the network to learn general and discriminative texture feature representations from high-resolution imagery without relying on extensive manual annotation. For this purpose, 200 unlabeled image patches of size 500×500 pixels were uniformly cropped as training samples from the multi-temporal GF-2 imagery covering the study area through manual visual interpretation. These samples encompass all major phenological stages and land cover conditions within the study area,

ensuring that the network is exposed to and learns diverse texture patterns throughout the annual cycle.

In the supervised training and testing stage of the Random Forest classifier, the features extracted by the aforementioned DTFEN are utilized to determine the category of each geographic entity. The classification unit in this study is the agricultural parcel, which possesses clear vector boundaries and practical management significance. A total of 901,263 agricultural parcels were identified within the study area, forming the target population for classification. To support model training and evaluation, a stratified random sampling approach was adopted to select 21,543 parcels as the labeled sample set. This sampling strategy was designed to ensure coverage of all land-use types (paddy fields, dry farmland, and garden land) and every temporal phase present across the study area, thereby maintaining representativeness of the overall parcel population. The corresponding land use type labels (paddy field, dry farmland, gardens) for each parcel were obtained through manual visual interpretation of high-spatial-resolution remote sensing imagery supplemented by field verification, thereby constituting the supervised learning dataset. This dataset was randomly split in a 7:3 ratio into a training set (15,080 parcel samples) and a test set (6463 parcel samples), which were used for training the Random Forest classifier and evaluating its final classification performance, respectively.

Regarding the specific settings for temporal segmentation, this study divides the mosaicked imagery covering the entire region into 7 temporal segments based on image acquisition dates, i.e., segmentation by month. Accordingly, we trained 7 independent DTFEN models, each specifically optimized for imagery from one temporal segment. Segmentation by month is based on considerations of data availability and experimental design precision. In practical applications, to balance accuracy and efficiency, months with similar spectral and texture characteristics (such as combining the vigorous growth months of June, July, and August) can be merged into a single phenological phase for processing, thereby reducing the number of models and lowering computational and storage costs.

In the specific operational procedure, DTFEN models are first trained separately for the imagery of each temporal segment. The balancing hyperparameter λ is set to 1. Training is performed using the Adam optimizer with a learning rate of 1×10^3 and a batch size of 4. Each DTFEN model is trained for approximately 300 epochs until the loss converges. Based on the temporal phase information of the imagery, the corresponding feature extractor is invoked to generate a pixel-level texture feature map. In order to balance texture feature extraction performance with computational cost, the dimensionality of the feature map output by the DTFEN model was set to 8 in this experiment. Subsequently, the vector boundaries of the parcels are used to crop the feature map, and statistical measures such as the mean, standard deviation, and entropy of the features within each parcel are calculated to generate a feature vector for each parcel. Finally, a Random Forest classifier is trained separately for the imagery of each temporal segment. Based on the temporal phase information of the imagery, the corresponding classifier is invoked to classify the parcels.

To ensure the rigor of the comparative experiments, the traditional CNN model used as a baseline in this study adopts a representative U-Net architecture. This model serves as a purely supervised feature extractor in the experiment. Its encoder comprises four downsampling stages, utilizing 3×3 convolutional kernels and ReLU activation functions to extract features, while the decoder restores spatial details through upsampling and skip connections. Unlike the self-supervised pre-training of DTFEN, this CNN model undergoes end-to-end supervised training directly on the same set of 200 labeled image patches (derived from the same source as DTFEN's training samples but with category labels assigned). The loss function used is cross-entropy loss. After training convergence, its classification head is removed, and the encoder along with the initial part of the decoder

are repurposed as a feature extractor to generate deep feature vectors for parcels in the same manner as DTFEN. The original CAPTN model strictly adheres to the architecture and classification objectives described in its source paper for training and testing, with its results used to quantify the performance gains achieved through the improvements introduced in this study.

The implementation of this study was based on the PyTorch 2.6 framework. The experiments were conducted on a hardware configuration comprising an NVIDIA GeForce RTX 4060 GPU (with 16 GB of VRAM) and an Intel i7-13700H CPU, with acceleration provided by the CUDA 12.4 library.

3.3. Evaluation Metrics

To comprehensively evaluate the performance of the land use classification method, this study employs four widely recognized evaluation metrics: Overall Accuracy (OA), User's Accuracy (UA), Producer's Accuracy (PA), and the Kappa coefficient. These metrics quantify the accuracy and reliability of the classification results from different perspectives.

Overall Accuracy (OA) measures the proportion of correctly classified samples, reflecting the overall performance of the classifier. It is calculated using the following formula:

$$OA = \frac{\sum_{i=1}^k TP_i}{N}, \quad (6)$$

where TP_i represents the number of correctly classified samples for the i -th category, k denotes the total number of categories, and N is the total number of samples.

User's Accuracy (UA) measures the reliability of classification results, representing the proportion of samples classified as a specific category that truly belong to that category. It is calculated as follows:

$$UA_i = \frac{TP_i}{TP_i + FP_i} \times 100\%, \quad (7)$$

where FP_i represents the number of samples from other categories that are falsely classified as the i -th category.

Producer's Accuracy (PA) measures the classifier's ability to correctly identify each category, representing the proportion of actual ground truth samples of a specific category that are correctly classified. It is calculated as follows:

$$PA_i = \frac{TP_i}{TP_i + FN_i} \times 100\%, \quad (8)$$

where FN_i denotes the number of samples from the i -th category that are misclassified into other categories.

The Kappa coefficient assesses the degree of agreement between classification results and ground truth, accounting for the effects of random classification. It is calculated as follows:

$$kappa = \frac{OA - P_e}{1 - P_e}, \quad (9)$$

where:

$$P_e = \sum_{i=1}^k \frac{(TP_i + FP_i) \times (TP_i + FN_i)}{N^2} \quad (10)$$

4. Results and Analysis

4.1. Land Use Classification Results

A total of 901,263 farmland parcels were identified with their current land use types in the study area, including 481,506 paddy field parcels, 390,998 dry farmland parcels, and

317,075 orchard parcels, accounting for 40.47%, 32.87%, and 26.66% of the total parcels respectively. The overall distribution of farmland parcels is shown in Figure 6.

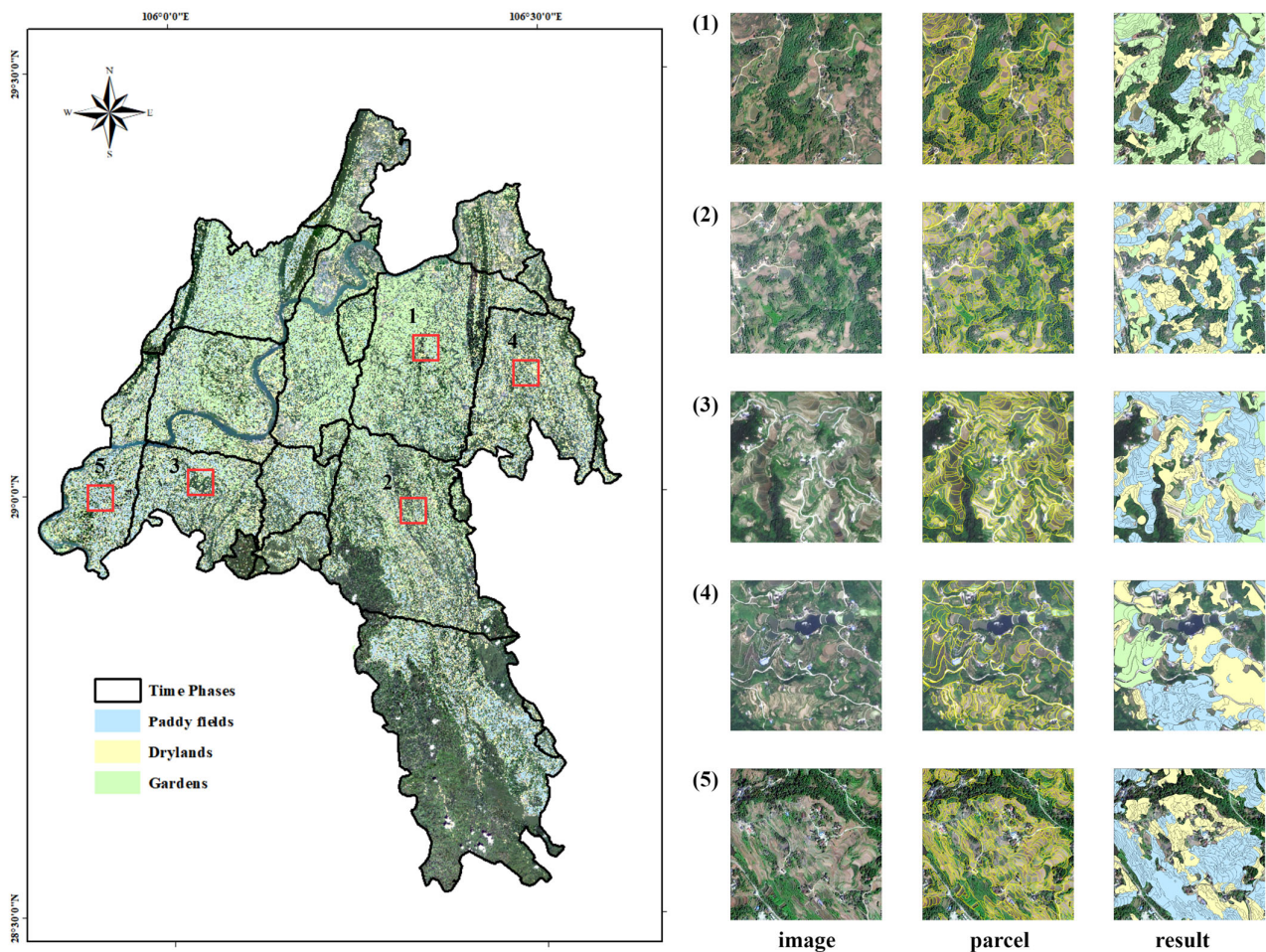


Figure 6. Farmland Parcel Distribution.

4.2. Comparison and Analysis

To evaluate the extraction results, this study designed the following set of experiments: (1) A comparison between the parcel-based method proposed in this study and traditional pixel-based and object-based methods is carried out. All methods utilized the same spectral + GLCM features and Random Forest classifier, with experiments conducted both with and without the spatiotemporal coupling approach. (2) A comparison between the deep texture feature method DTFEN used in this study and traditional methods using spectral features, traditional texture features, conventional convolutional neural networks, and the original CAPTN is conducted. In this set of experiments, all classification targets were parcels. (3) A comparison between classification results using single-temporal texture features and those using multi-temporal features is performed. (4) Ablation Experiments for Key Components in DTFEN are conducted.

Table 2 presents the performance of combinations involving different classification units and spatiotemporal coupling strategies. An in-depth analysis reveals that the parcel-based classification method achieved optimal performance when combined with the spatiotemporal coupling strategy (OA = 85.6%, Kappa = 0.812). This result validates the theoretical advantage of using geographic entities with clear management significance as classification units. Compared with the pixel-based method, the parcel-based method improved OA by 7.4 percentage points and the Kappa coefficient by 0.114, indicating that the parcel unit effectively overcomes the “salt-and-pepper noise” issue inherent in

pixel-based methods while avoiding the segmentation scale sensitivity associated with object-based methods.

Table 2. Comparative accuracy analysis of classification units and spatiotemporal coupling approaches.

Unit	Spatiotemporal Coupling	OA (%)	Kappa	Paddy PA (%)	Paddy UA (%)	Dry PA (%)	Dry UA (%)	Garden PA (%)	Garden UA (%)
Pixel	No	75.2	0.698	74.3	73.8	75.1	75.6	76.2	75.2
Pixel	Yes	78.1	0.726	77.2	76.9	78.3	78.7	78.8	78.1
Object	No	80.3	0.752	79.5	79.1	80.6	81.2	80.8	80.3
Object	Yes	83.5	0.791	82.7	82.3	83.8	84.1	84.0	83.5
Parcel	No	82.4	0.781	81.6	81.2	82.7	83.2	82.9	82.4
Parcel	Yes	85.6	0.812	84.9	84.8	85.3	85.6	86.1	85.7

The introduction of the spatiotemporal coupling strategy significantly enhanced classification accuracy across all methods. Specifically, the pixel-based method improved by 2.9 percentage points, the object-based method by 3.2 percentage points, and the parcel-based method by 3.2 percentage points. This demonstrates the universal applicability of the spatiotemporal coupling strategy in effectively integrating multi-temporal information and mitigating the uncertainty of single-temporal features.

Table 3 compares the performance of different feature extraction methods using parcel-based classification units. The proposed DTFEN method achieved the best classification results, with an OA of 92.3% and a Kappa coefficient of 0.892, significantly outperforming all other comparative methods. These results fully demonstrate the representational advantages of deep texture features in complex agricultural scenarios.

Table 3. Accuracy comparison of different feature extraction methods.

Feature	OA (%)	Kappa	Paddy PA (%)	Paddy UA (%)	Dry PA (%)	Dry UA (%)	Garden PA (%)	Garden UA (%)
Spectral	80.5	0.752	79.4	78.9	80.8	81.3	81.3	81.2
Traditional Texture	83.2	0.781	82.1	81.6	83.5	84.0	84.0	83.9
CNN	86.4	0.823	85.3	84.8	86.7	87.2	87.2	87.1
CAPTN	88.7	0.845	87.6	87.1	89.0	89.5	89.5	89.4
DTFEN	92.3	0.892	91.2	90.7	92.6	93.1	93.1	93.0

In-depth analysis reveals that deep learning methods substantially surpass traditional feature extraction approaches. DTFEN improved OA by 9.1 percentage points compared to traditional texture feature methods, proving the superiority of deep texture features in characterizing complex spatial patterns. Gardens achieved the highest User's Accuracy (UA) and Producer's Accuracy (PA) (both around 93%) under DTFEN features. This high accuracy can be attributed to the texture frequency attention mechanism, which enhances the perception of regular planting patterns—gardens (e.g., orchards, pepper gardens) typically exhibit a regular planting structure, appearing as periodic and directional texture dot-arrays in the imagery. In contrast, the identification accuracy for drylands (UA $\approx$ 93.1%, PA $\approx$ 92.6%), while still at a high level, is slightly lower than that of gardens. This may stem from the higher intra-class heterogeneity of drylands: their surface cover can vary rapidly from bare soil to sparse or dense crops. This inherent high intra-class variance poses a greater challenge to the model's feature generalization capability.

Compared to conventional CNN methods, DTFEN improved OA by 5.9 percentage points, indicating that pure convolutional feature extraction is insufficient for fully exploiting discriminative information in agricultural textures. Through the synergistic effects

of texture frequency attention, depth permutation, and Chebyshev transformation, DTFEN achieves comprehensive capture of multi-scale characteristics in agricultural textures. Particularly when dealing with complex terrain in hilly and mountainous areas, DTFEN demonstrates stronger robustness, effectively overcoming the influences of illumination and shadows caused by topography.

Table 4 presents the ablation experiment results for DTFEN. The experimental results show that after removing the self-supervised loss, the Overall Accuracy (OA) decreases by 4.2 percentage points and the Kappa coefficient drops by 0.061. This indicates that the masked reconstruction and cross-view consistency tasks effectively drive the network to learn more discriminative and robust intrinsic texture features rather than superficial appearance variations. When the TFA module is removed, OA decreases by 1.8 percentage points and Kappa drops by 0.025. The performance decline is particularly noticeable in the garden category, which aligns with the qualitative analysis. The TFA module explicitly guides the network to focus on discriminative texture regions in the spatial domain (e.g., regularly arranged plant-canopy dot-matrix patterns in gardens), thereby enhancing the perception of such ordered spatial patterns.

Table 4. Ablation study results of key components in DTFEN.

Model Variant	OA (%)	Kappa	Paddy PA (%)	Paddy UA (%)	Dry PA (%)	Dry UA (%)	Garden PA (%)	Garden UA (%)
DTFEN (Full)	92.3	0.892	91.2	90.7	92.6	93.1	93.1	93.0
DTFEN (w/o Self-Supervision)	88.1	0.831	88.2	87.9	88.5	88.3	88.1	87.4
DTFEN (w/o TFA)	90.5	0.867	89.3	88.9	90.2	91.3	90.8	91.2

The ablation experiments quantitatively confirm the effectiveness of each designed component in DTFEN. Self-supervised learning is the core contributor to improving feature quality, while the TFA module provides beneficial enhancement for specific regular texture patterns in agricultural scenes. The synergistic operation of these components enables DTFEN to extract highly discriminative and robust deep texture features from very high-spatial-resolution remote sensing imagery, laying a solid foundation for subsequent parcel-level classification.

Table 5 compares the performance differences between two feature utilization strategies based on the ‘spatiotemporally coupled’ framework: the single-temporal strategy (i.e., the basic workflow described in Sections 3.1.1–3.1.6, where each parcel uses only the features from its corresponding temporal phase) and the multi-temporal feature integration strategy (i.e., the method described in Section 3.1.7, which involves concatenating features from all available temporal phases for each parcel during training and classification). Building upon the DTFEN method, the multi-temporal strategy achieved an OA of 94.1%, representing an improvement of 1.8 percentage points over the single-temporal approach, with a Kappa coefficient increase of 0.032.

The advantage of the multi-temporal strategy is most evident in paddy field identification, where UA and PA increased by 1.9 and 1.6 percentage points respectively. This can be attributed to the highly complementary characteristic patterns of paddy fields across different phenological phases: the specular reflection features during spring plowing provide extremely high discriminative power but may diminish in other periods; the uniform vegetation coverage during the growth phase shows high similarity with other vegetation types; and the golden spike texture during the maturity phase exhibits unique spatial distribution patterns. The multi-temporal features effectively integrate this complementary

information through spatiotemporal coupling, whereas single-temporal methods can only capture localized characteristics from specific periods.

Table 5. Accuracy comparison of temporal feature utilization strategies.

Temporal	OA (%)	Kappa	Paddy PA (%)	Paddy UA (%)	Dry PA (%)	Dry UA (%)	Garden PA (%)	Garden UA (%)
Single-temporal	92.3	0.892	91.2	90.7	92.6	93.1	93.1	93.0
Multi-temporal	94.1	0.924	93.1	92.3	93.2	94.3	94.4	94.3

The multi-temporal strategy also delivered significant improvements for dry farmland and gardens. Dry farmland undergoes continuous transformation from plowed soil to dense vegetation across different growth stages, while gardens demonstrate periodic characteristics from flowering to fruiting phases. These temporal variation patterns are fully exploited through multi-temporal analysis, providing additional discriminative information for land class identification. Notably, the performance of the multi-temporal strategy during the fallow period proves particularly crucial, as the simplified surface coverage during this phase effectively reveals fundamental differences between perennial gardens and seasonal croplands.

5. Discussion

5.1. Advantages and Computational Trade-Offs of the Spatiotemporal-Coupled Framework

The spatiotemporally coupled classification framework proposed in this study effectively addresses the core challenge of temporal inconsistency caused by multi-temporal mosaicking in high-spatial-resolution remote sensing imagery for land use classification. This is achieved by proactively dividing the full-coverage multi-temporal mosaicked imagery into several internally temporally consistent sub-regions during the preprocessing stage and training dedicated feature extractors for each. It is important to note that a temporal segmentation strategy that is too fine-grained or fragmented could theoretically increase the complexity of model training and management. We argue that while this “divide-and-conquer” strategy increases the workload of parallel modeling in the early stages, its core advantage lies in significantly reducing the complexity of the sample distribution that a single model must fit. Each dedicated feature extractor needs only to focus on learning the relatively homogeneous surface texture patterns under a specific temporal phase, which is more efficient and stable than forcing a “universal” model to memorize and reconcile contradictory features from the entire year. By constructing a modular feature extractor library, the method can automatically invoke the corresponding module based on the image’s temporal phase during the inference stage, achieving “train once, reuse multiple times.” This automated workflow does not continuously increase the manual burden in practical applications; instead, by enhancing the focus and robustness of the single-temporal-phase models, it improves the overall repeatability and operational efficiency of the method.

The choice of the “temporal segment-specific model training” spatiotemporal coupling strategy is primarily motivated by methodological simplicity and effectiveness, rather than a disregard for computational cost. Compared to training a large unified model that needs to learn complex spatiotemporal invariance, our strategy decomposes the problem into several simpler sub-problems (single-temporal-phase feature modeling). Although this “divide-and-conquer” strategy requires training N independent DTFEN models for N temporal segments upfront, leading to approximately N -fold initial training time, each model converges faster due to its simpler task, and the overall training process is easily parallelizable. More importantly, during the inference and application stages, this strategy

avoids the complex fine-tuning often required by unified models to adapt to the temporal characteristics of new regions. Its modular design allows the model library to be pre-built and reused. For the Jiangjin District application, while the total training time for the 7 models was higher than for a single model, it remained within an acceptable range (approximately one week), and the resulting accuracy improvement (e.g., 94.1% OA achieved by the multi-temporal strategy) is significant. Therefore, this computational investment is reasonable and efficient for practical applications pursuing high accuracy, interpretability, and stability. Future optimizations could focus on developing more lightweight feature extraction networks or exploring meta-learning-based rapid model initialization methods to further reduce the marginal cost of the multi-model strategy.

5.2. Value and Future Directions of the Multi-Temporal Feature Integration Strategy

The experimental results of the multi-temporal feature integration strategy proposed in this study show that by concatenating deep texture features of the same parcel across different phenological stages, complementary information within the crop growth cycle—such as the specular reflection characteristics of paddy fields during flooding periods and the vegetation coverage features during the growth stage—can be effectively leveraged. This further improved the overall classification accuracy from 92.3% to 94.1%. This demonstrates that, within the spatiotemporally coupled framework, the evolution from ‘addressing temporal interference’ to ‘actively utilizing temporal information’ is both effective and necessary. The successful implementation of this strategy relies on the availability of multi-temporal, cloud-free imagery for the same parcel. In future work, exploring how to use time-series reconstruction techniques or fuse dense time-series data from medium-to-low resolution sources to compensate for the temporal coverage limitations of high-resolution imagery, thereby enabling reliable multi-temporal feature fusion across broader areas, represents a valuable research direction.

5.3. Description of Method Adaptability

The core idea of this method is to reduce model learning difficulty through temporal segmentation and combine it with feature statistics and classification based on geographic entities (parcels). Selecting the Jiangjin District, characterized by complex terrain and diverse planting structures, as the experimental area itself serves as a successful validation of the model’s application in a complex scenario. For plain areas, where large-scale, regularized farmland exhibits higher texture consistency, the temporal segmentation strategy in this method can more effectively isolate feature differences arising from different farming systems (e.g., single/double cropping rice). In plateau areas, strong illumination and topographic shadows are the main challenges. The cross-view consistency self-supervised task adopted in DTFEN is precisely designed to enhance feature robustness against unstable factors like illumination, and this mechanism can be directly transferred. For hilly areas, their terrain fragmentation is similar to that of the experimental area, and the method has demonstrated applicability. For special land types that may exist in coastal areas, such as saline-alkali land or aquaculture ponds, targeted training samples and features can be introduced in the “parcel classification” stage of this framework to enable extended recognition. Therefore, this framework provides an extensible solution paradigm, the applicability of which hinges on adjusting the temporal segmentation scheme and parcel sample library according to the characteristics of new regions. However, it must also be pointed out that the current research conclusions are based on a single region and three types of agricultural land. Their generalizability needs systematic validation across more diverse geographic units and a broader range of land use types (e.g., forests, grasslands, construction land). Especially for non-agricultural contexts like urban land use, the tempo-

ral change patterns of features may differ fundamentally from agricultural phenology (e.g., changes in built-up areas stem more from construction activities than natural rhythms). Nevertheless, the “temporal segmentation” concept of this framework remains valuable. For instance, temporal phases could be divided based on different development stages of urban areas or according to seasonal variations in illumination and vegetation (e.g., park green spaces) to address spectral-textural differences within cities arising from construction timing or seasons. The key is that this framework provides an extensible paradigm whose applicability is not a mechanical replication but requires flexible adjustment of the basis for temporal segmentation and the strategy for building parcel feature libraries according to the spatiotemporal variation patterns of new regions and categories.

5.4. Advantages over Existing Methods

Compared to previous models, the advantages of this study are mainly reflected in three aspects. First, compared to a unified deep learning model trained directly on mixed multi-temporal data, this method, through the spatiotemporally coupled framework, fundamentally avoids the performance degradation caused by the model learning contradictory features. This is reflected in the experimental results (Tables 2 and 3), especially in the significant improvement in the recognition accuracy of paddy fields. Second, compared to studies that focus solely on improving network architecture to enhance feature representation, this study innovatively and systematically integrates deep feature learning with geographic entities (parcels) that have clear management significance. Through the workflow of “deep feature extraction–entity constraint–statistical quantification,” it combines deep learning models adept at capturing pixel-level details with object-level analysis units (e.g., parcels) possessing clear geographic meaning, successfully transforming pixel-level feature representations into object-level geographic information products. This technical approach not only fully leverages the advantages of deep learning in extracting fine details from high-resolution imagery but also ensures that the final output directly corresponds to the geographic entities of concern in practical land management and agricultural survey operations, thereby significantly enhancing the practical value and interpretability of the research outcomes. Third, compared to supervised learning relying on large amounts of labeled samples, the self-supervised learning strategy introduced in the feature extraction stage reduces dependence on large-scale, finely annotated data, enabling the model to learn more generalizable texture representations from vast amounts of unlabeled imagery.

5.5. Limitations and Future Work

Of course, the method in this study also has certain limitations, which point the way for future work. First, the current temporal segmentation relies to some extent on prior phenological knowledge and the completeness of image acquisition dates. Future work could explore data-driven (e.g., cluster analysis) adaptive temporal segmentation methods to enhance the framework’s automatic adaptability to different regions and data acquisition conditions. Second, the method has relatively high requirements for input image quality (e.g., cloud cover, orthorectification accuracy) and may face challenges in areas with frequent cloud cover. Future research could consider integrating remote sensing data sources with strong penetration capabilities, such as radar, to build a multimodal spatiotemporally coupled analysis framework for improved robustness. Finally, although the method demonstrates superiority in the Jiangjin District, its generalizability across broader areas and more diverse geographic landscapes still needs further verification through cross-regional systematic comparative experiments. Developing lightweight model migration and adaptive fine-tuning strategies will be key to promoting the operational application of this method.

6. Conclusions

This study addresses the issue of temporal inconsistency caused by multi-temporal mosaicking in high-resolution remote sensing imagery by proposing a land use classification method oriented toward geographic entities and based on spatiotemporal coupling. The method divides the study area into sub-regions based on image acquisition time and trains dedicated feature extractors for each sub-region, thereby reducing model learning complexity at its source. Furthermore, it integrates pixel-level texture features extracted via deep learning with vector boundaries of land parcels that carry clear management significance. By performing multi-dimensional statistical quantification—including mean, standard deviation, and entropy—it constructs parcel-level feature descriptors that are both highly discriminative and geographically meaningful. Finally, accurate classification of parcels is achieved using a Random Forest classifier. Empirical validation in the typical complex agricultural area of Jiangjin, Chongqing, demonstrates the effectiveness of the proposed spatiotemporal coupling framework, achieving an overall classification accuracy and Kappa coefficient of 92.3% and 0.89, respectively. The method shows particular improvement in the recognition accuracy of land types such as paddy fields, where texture features vary significantly with phenology. This provides a feasible technical solution for operational applications such as land surveying and agricultural monitoring.

Author Contributions: Conceptualization, X.H. and W.M.; methodology, J.Y., T.W. and X.H.; software, J.Y. and P.J.; formal analysis, X.H. and J.Y.; investigation, J.Y., Z.S. and H.Y.; resources, Q.T. and Y.X.; writing—original draft preparation, J.Y.; writing—review and editing, J.Y. and X.H.; visualization, J.Y.; supervision, W.M. and X.H.; project administration, P.J.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Independent Deployment Project of the Aerospace Information Research Institute, Chinese Academy of Sciences (No. E4Z202021F).

Data Availability Statement: The data used in this paper are available upon request from the corresponding author via email.

Conflicts of Interest: No potential conflicts of interest were reported by the author.

References

1. Tan, J.; Yu, D.; Li, Q.; Tan, X.; Zhou, W. Spatial relationship between land-use/land-cover change and land surface temperature in the Dongting Lake area, China. *Sci. Rep.* **2020**, *10*, 9245. [[CrossRef](#)] [[PubMed](#)]
2. Ramzan, M.; Saqib, Z.A.; Hussain, E.; Khan, J.A.; Nazir, A.; Dasti, M.Y.S.; Ali, S.; Niazi, N.K. Remote sensing-based prediction of temporal changes in land surface temperature and land use-land cover (LULC) in urban environments. *Land* **2022**, *11*, 1610. [[CrossRef](#)]
3. European Commission: Joint Research Centre; Soille, P.; Lumnitz, S.; Albani, S. From Foresight to Impact. In Proceedings of the 2023 Conference on Big Data from Space (BiDS'23), Vienna, Austria, 6–9 November 2023; Publications Office of the European Union: Luxembourg, 2023.
4. Chen, B.; Xu, B.; Gong, P. Mapping essential urban land use categories (EULUC) using geospatial big data: Progress, challenges, and opportunities. *Big Earth Data* **2021**, *5*, 410–441. [[CrossRef](#)]
5. Ye, J.; Hu, Y.; Zhen, L.; Wang, H.; Zhang, Y. Analysis on Land-Use Change and its driving mechanism in Xilingol, China, during 2000–2020 using the google earth engine. *Remote Sens.* **2021**, *13*, 5134. [[CrossRef](#)]
6. Bodhankar, S.; Gupta, K.; Kumar, P.; Srivastav, S.K. GIS-based multi-objective urban land allocation approach for optimal allocation of urban land uses. *J. Indian Soc. Remote Sens.* **2022**, *50*, 763–774. [[CrossRef](#)]
7. Mohammadyari, F.; Tavakoli, M.; Zarendian, A.; Abdollahi, S. Optimization land use based on multi-scenario simulation of ecosystem service for sustainable landscape planning in a mixed urban-Forest watershed. *Ecol. Model.* **2023**, *483*, 110440. [[CrossRef](#)]
8. MohanRajan, S.N.; Loganathan, A.; Manoharan, P. Survey on Land Use/Land Cover (LU/LC) change analysis in remote sensing and GIS environment: Techniques and Challenges. *Environ. Sci. Pollut. Res.* **2020**, *27*, 29900–29926. [[CrossRef](#)]

9. Schug, F.; Pfoch, K.A.; Pham, V.-D.; van der Linden, S.; Okujeni, A.; Frantz, D.; Radeloff, V.C. Land cover fraction mapping across global biomes with Landsat data, spatially generalized regression models and spectral-temporal metrics. *Remote Sens. Environ.* **2024**, *311*, 114260. [\[CrossRef\]](#)
10. Tong, X.-Y.; Xia, G.-S.; Zhu, X.X. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 178–196. [\[CrossRef\]](#)
11. Sun, X.; Wang, B.; Wang, Z.; Li, H.; Li, H.; Fu, K. Research progress on few-shot learning for remote sensing image interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2387–2402. [\[CrossRef\]](#)
12. Tian, L.; Cao, Y.; He, B.; Zhang, Y.; He, C.; Li, D. Image enhancement driven by object characteristics and dense feature reuse network for ship target detection in remote sensing imagery. *Remote Sens.* **2021**, *13*, 1327. [\[CrossRef\]](#)
13. Li, B.; Zhou, Z.; Wu, T.; Luo, J. Fine-Grained Land Use Remote Sensing Mapping in Karst Mountain Areas Using Deep Learning with Geographical Zoning and Stratified Object Extraction. *Remote Sens.* **2025**, *17*, 2368. [\[CrossRef\]](#)
14. Blaschke, T. Object Based Image Analysis for Remote Sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [\[CrossRef\]](#)
15. Wang, Z.; Yang, X.; Liu, Y.; Liu, B.; Zhang, J.; Liu, X.; Meng, D.; Gao, K.; Zeng, X.; Ding, Y. Geographical Principles of Remote Sensing Image Analysis and the Hierarchical Patch Model Based Analysis Framework. *Natl. Remote Sens. Bull.* **2024**, *28*, 1412–1424.
16. Xu, H.; Zheng, T.; Liu, Y.; Zhang, Z.; Xue, C.; Li, J. A joint convolutional cross ViT network for hyperspectral and light detection and ranging fusion classification. *Remote Sens.* **2024**, *16*, 489. [\[CrossRef\]](#)
17. Yin, Z.; Li, X.; Wu, P.; Lu, J.; Ling, F. CSSF: Collaborative spatial-spectral fusion for generating fine-resolution land cover maps from coarse-resolution multi-spectral remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2025**, *226*, 33–53. [\[CrossRef\]](#)
18. Zhang, C.; Wang, Q.; Atkinson, P.M. Unsupervised object-based spectral unmixing for subpixel mapping. *Remote Sens. Environ.* **2025**, *318*, 114514. [\[CrossRef\]](#)
19. Bantchina, B.B.; Gündoğdu, K.S.; Yazar, S.; Author, C. Crop type classification using Sentinel-2A-derived normalized difference red edge index (NDRE) and machine learning approach. *Bursa Uludağ Üniversitesi Ziraat Fakültesi Derg.* **2024**, *38*, 89–105. [\[CrossRef\]](#)
20. Moumni, A.; Lahrouni, A. Machine learning-based classification for crop-type mapping using the fusion of high-resolution satellite imagery in a semiarid area. *Scientifica* **2021**, *2021*, 6613372. [\[CrossRef\]](#)
21. Wang, X.; Zhang, J.; Xun, L.; Wang, J.; Wu, Z.; Henchiri, M.; Zhang, S.; Zhang, S.; Bai, Y.; Yang, S.; et al. Evaluating the effectiveness of machine learning and deep learning models combined time-series satellite data for multiple crop types classification over a large-scale region. *Remote Sens.* **2022**, *14*, 2341. [\[CrossRef\]](#)
22. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [\[CrossRef\]](#)
23. Aleissae, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.-S.; Khan, F.S. Transformers in Remote Sensing: A Survey. *Remote Sens.* **2023**, *15*, 1860. [\[CrossRef\]](#)
24. Yi, Y.; Zhang, Z.; Zhang, W.; Zhang, C.; Li, W.; Zhao, T. Semantic Segmentation of Urban Buildings from VHR Remote Sensing Imagery Using a Deep Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 1774. [\[CrossRef\]](#)
25. Zhu, Y.; Lu, L.; Li, Z.; Wang, S.; Yao, Y.; Wu, W.; Pandey, R.; Tariq, A.; Luo, K.; Li, Q. Monitoring Land Use Changes in the Yellow River Delta Using Multi-Temporal Remote Sensing Data and Machine Learning from 2000 to 2020. *Remote Sens.* **2024**, *16*, 1946. [\[CrossRef\]](#)
26. Wang, Q.; Qiu, Y.; Shen, S.; Shen, T. Cross-Temporal Self-Supervised Learning with Superpixel Mask for Multi-Temporal Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *18*, 2126–2136. [\[CrossRef\]](#)
27. Plekhanova, E.; Robert, D.; Dollinger, J.; Brun, P.; Wegner, J.D.; Zimmermann, N.E. SeCo-Eco: Global multiband seasonal pre-training dataset and self-supervised model for ecological applications. *EGU sphere* **2025**. [\[CrossRef\]](#)
28. Heidari, M.J.; Borges, J.G.; Najafi, A.; Alavi, S.J. A 3D convolutional neural network approach for monitoring land cover change using Sentinel-2 satellite imagery. *Front. For. Glob. Change* **2026**, *8*, 1672747. [\[CrossRef\]](#)
29. Zhang, J.; Wu, T.; Luo, J.; Hu, X.; Wang, L.; Li, M.; Lu, X.; Li, Z. Toward Agricultural Cultivation Parcels Extraction in the Complex Mountainous Areas Using Prior Information and Deep Learning. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4402414. [\[CrossRef\]](#)
30. Jiao, S.; Hu, D.; Shen, Z.; Wang, H.; Dong, W.; Guo, Y.; Li, S.; Lei, Y.; Kou, W.; Wang, J.; et al. Parcel-Level Mapping of Horticultural Crop Orchards in Complex Mountain Areas Using VHR and Time-Series Images. *Remote Sens.* **2022**, *14*, 2015. [\[CrossRef\]](#)

31. Liu, W.; Wang, J.; Luo, J.; Wu, Z.; Chen, J.; Zhou, Y.; Sun, Y.; Shen, Z.; Xu, N.; Yang, Y. Farmland Parcel Mapping in Mountain Areas Using Time-Series SAR Data and VHR Optical Images. *Remote Sens.* **2020**, *12*, 3733. [[CrossRef](#)]
32. Evani, R.; Rajan, D.; Mao, S. Chebyshev Attention Depth Permutation Texture Network with Latent Texture Attribute Loss. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 23423–23432. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.