electronics

Article

STF-Net: A Robust Fine-Grained Land Classification Method Fusing Images and Spatiotemporal Metadata

Hengzhou Ye [1], Peikang Tang [1], Xiang Zhuang [2,3], Yongmei Tan [1], Gong Chen [1,2,4] and Haoxiang Wen [2,3,*]

[1] College of Computer Science and Engineering, Guilin University of Technology, Guilin 541006, China; 2002018@glut.edu.cn (H.Y.); 2120231256@glut.edu.cn (P.T.); 2120241296@glut.edu.cn (Y.T.); cg1900@glut.edu.cn (G.C.)
[2] Technology Innovation Center for Natural Resources Monitoring and Evaluation of Beibu Gulf Economic Zone, Ministry of Natural Resources, Nanning 530219, China; zhuangxiang@gxzrzydcjcy.cn
[3] Guangxi Institute of Natural Resources Survey and Monitoring, Nanning 530219, China
[4] Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541006, China
* Correspondence: wenhaoxiang@gxzrzydcjcy.cn

Abstract

Fine-grained land classification provides low-cost and efficient data support for land resource monitoring, agricultural assessment, and ecological protection. However, classification methods based on field-captured images often suffer from performance limitations due to challenges such as the complexity of land categories, variations in illumination and viewing angles, seasonal changes, and the absence of spatiotemporal metadata. Addressing the characteristics of significant seasonal variations in crops and strong correlations between adjacent land parcel categories, this paper proposes a robust Spatiotemporal Fusion Network (STF-Net) for fine-grained land classification by fusing images with spatiotemporal metadata. The main components of STF-Net include a visual backbone network, a spatiotemporal metadata encoder, a cross-modal multi-head attention fusion module, and a fallback branch designed for cases where metadata is missing. The model is robust to missing metadata and adaptable to different visual backbone networks such as the Swin Transformer and EfficientNet. Experiments on a dataset containing 91 categories of land use photos show that STF-Net achieves an overall accuracy of 93.54% and an F1-score of 0.92, significantly outperforming baseline models. Ablation studies further validate the necessity of fusing spatiotemporal metadata.

Keywords: land use classification; spatiotemporal metadata; feature fusion

Academic Editor: Toshifumi Moriyama

Received: 10 March 2026
Revised: 8 April 2026
Accepted: 9 April 2026
Published: 10 April 2026

1. Introduction

Against the backdrop of strengthening ecological protection and natural resource supervision, higher requirements are placed on the fine-grained monitoring of land use status. Particularly concerning the red-line policy for cultivated land protection and the supervision of illegal land use changes, efficient and accurate technical means are needed to assist management departments in identifying changes in land types [1]. Traditionally, remote sensing image classification has been widely used for macro-scale land cover monitoring [2–4]. However, in practical land change verification work, field survey personnel also capture a large number of on-site evidentiary photos as direct evidence for determining land type changes. Currently, the analysis of these evidentiary photos primarily relies on

manual inspection and comparison, which suffer from issues such as heavy workload and subjective influences. Therefore, developing an intelligent method to automatically identify land categories and cultivation status from photos is of significant practical importance. It can not only greatly improve the efficiency of land change verification but also objectively enhance the consistency and accuracy of determinations.

Unlike remote sensing images, ground-based photographs are characterized by diverse viewing angles, significant scale variations, and complex backgrounds. They are also typically accompanied by metadata such as capture time and GPS coordinates. This additional spatiotemporal information contains contextual cues for land categories: the longitude and latitude of the capture location imply the geographical environment, while the capture date reflects the crop phenological cycle [5]. Such multimodal data characteristics render the simple migration of remote sensing image classification methods insufficient for achieving ideal results. Therefore, it is necessary to explore innovative approaches that fuse visual models with spatiotemporal metadata. In recent years, deep learning has achieved remarkable progress in the field of image classification, with the continuous emergence of models from classic Convolutional Neural Network (CNN) to Vision Transformer (ViT). Among them, Transformer models, relying on the self-attention mechanism, excel at capturing global dependencies and have shown great potential in tasks such as image classification. Particularly, the Swin Transformer, with its hierarchical local self-attention design, has surpassed traditional CNNs in various visual tasks. By applying the Swin Transformer to land category classification and integrating temporal and spatial metadata, we aim to fully leverage the representational capacity of deep models to achieve fine-grained land type identification. This research differs from traditional remote sensing image classification, focusing more on fine-grained object recognition from a ground perspective and serving as an effective supplement and extension to remote sensing interpretation.

Building upon the aforementioned motivations, this paper proposes a land category recognition method based on the Swin Transformer and spatiotemporal information fusion. With the research objective of improving the classification accuracy of field evidentiary photos, we design a visual feature extraction network centered on the Swin Transformer. This network is combined with periodic temporal encoding and geographic location encoding to construct spatiotemporal metadata features, which are then effectively fused with image features through a multimodal attention mechanism. The main contributions of this paper include the following:

- Proposing a Robust Spatiotemporal Fusion Network (STF-Net). Its core modules include an image visual backbone network, temporal and geographic location metadata encoding, and a cross-modal multi-head attention fusion mechanism. The visual backbone network supports architectures such as the Swin Transformer and EfficientNet, with the Swin Transformer yielding the best performance.
- Addressing the issue of missing spatiotemporal metadata in some evidentiary photos by designing a lightweight fallback branch that enables the model to accommodate both complete and missing metadata inputs, thereby enhancing its robustness in practical applications.
- Systematically validating the effectiveness of STF-Net through extensive experiments, including its compatibility with various backbone networks like the Swin Transformer and EfficientNet, and demonstrating the necessity of fusing spatiotemporal metadata.

2. Related Work

2.1. Visual Applications of Transformer Architecture

The Transformer network architecture is widely applied in the field of Natural Language Processing (NLP). Due to its excellent performance, the Transformer network archi-

tecture has been introduced into the domain of natural images and is currently a research hotspot in visual recognition tasks. Both Transformer and Vision Transformer are based on the attention mechanism architecture, which is also extensively used in Convolutional Neural Networks. The attention-augmented convolutional network [6] concatenates convolutional feature maps with a set of feature maps generated through self-attention to enhance the model's ability to capture long-range information. Numerous experiments have shown that, with a comparable number of parameters, this model achieves higher classification accuracy than the Squeeze-and-Excitation Network (SENet), validating that the self-attention mechanism can enhance the classification and recognition capability of CNNs. In the object detection model Desktop Exploration of Remote Terrain (DERT) [7], CNNs are solely used to extract basic features, while the encoder and decoder parts fully leverage Transformers, forming a Transformer-based encoder–decoder sequence prediction framework capable of predicting all targets at once. Vision Transformer [8] abandons the CNN architecture and employs the Transformer architecture to address image-related problems. Input images are divided into patches, which are then fed as sequences into subsequent Transformer encoders. The Transformer in Transformer (TNT) [9] model improves upon ViT by dividing patches into sub-patches, representing the image at a finer granularity. Studies have shown that the TNT model exhibits stronger learning capacity and generalization ability. Swin Transformer [10] proposes a hierarchical Transformer, which divides patches into multiple windows and employs local attention mechanisms within these windows. Simultaneously, to address information interaction between windows across different layers, a shifted window strategy is adopted to facilitate cross-window connections between upper and lower layers. This hierarchical design provides substantial flexibility for modeling features at different scales. The Swin Transformer has also been widely applied in agricultural remote sensing and plant disease recognition tasks.

2.2. Deep Learning and Land Class Recognition

Land type image recognition aims to classify geographical scenes and surface cover types. Traditional methods primarily rely on remote sensing imagery and spectral information [11–14]. Convolutional Neural Networks have achieved significant results in this task due to their powerful local feature extraction capabilities. CNNs can automatically learn discriminative features from images, demonstrating high accuracy in remote sensing image scene classification.

In recent years, Vision Transformer has been introduced into the fields of remote sensing image interpretation and ground-based scene classification, exhibiting characteristics distinct from CNN. By directly modeling long-range dependencies between image patches through self-attention mechanisms, ViT excels in capturing global contextual information [15]. However, notable differences exist between remote sensing images and ground-captured photographs in terms of data characteristics and classification objectives. Remote sensing images are typically captured from an overhead perspective, containing multi-scale and densely distributed targets within a single image, with complex scenes [16]. In contrast, ground-based photographs often have a prominent subject and relatively simpler backgrounds. For ground-based photographs, spatiotemporal metadata captured alongside the images can provide valuable prior knowledge. For instance, Tang et al. [17] utilized GPS coordinates to assist in internet image classification, confirming that incorporating seasonal and location context improves regional scene recognition accuracy. Research [18] has constructed spatiotemporal prior distributions for species, transforming observation data into probabilities of a category occurring at specific times and locations, and then fusing them with visual classification results, significantly enhancing fine-grained recognition performance. These studies indicate that for land feature categories with distinct

spatiotemporal distribution patterns, integrating metadata can effectively enhance the model's discriminative ability.

In the domain of multimodal information fusion, various methods have been proposed for images accompanied by auxiliary information. In tasks such as visual question answering and image caption generation, Transformer-based cross-modal attention mechanisms have achieved remarkable success. For the fusion of structured metadata (e.g., time, location) with images, common strategies include feature concatenation, weighted fusion, or attention-based interaction mechanisms. The dynamic MLP method proposed by Yang et al. [19] instantiates network weights based on the longitude, latitude, and time of each image to conditionally transform visual features. This approach achieved an accuracy of 91.39% on the iNaturalist fine-grained classification dataset, validating the effectiveness of deep cross-modal interaction. Minetto et al. [5], on the other hand, proposed the Hydra framework, which employs CNN ensembles combined with geographic metadata for land use classification. Overall, the fusion strategy is crucial for model performance. Simple feature concatenation often fails to fully exploit inter-modal relationships, while attention-based interaction mechanisms are more conducive to achieving efficient cross-modal representation.

Since the introduction of ViT by Dosovitskiy et al. [8], the Transformer architecture has been applied to vision tasks such as image classification. However, ViT produces a single-scale global feature map, making it difficult to directly adapt to the needs of multi-scale object analysis. Swin Transformer, proposed by Liu et al. [11], addresses this through hierarchical design and shifted window self-attention, establishing an efficient multi-scale feature extraction backbone. It has demonstrated excellent performance in ImageNet classification and various downstream tasks. Its core idea is to restrict self-attention computation to local windows to reduce complexity, while enabling cross-window information exchange through window shifting between layers, thereby balancing local details and global semantics. Swin Transformer has been proven to outperform traditional Convolutional Neural Networks in remote sensing image scene classification. For instance, Jannat et al. [20] applied it to land cover classification, achieving higher accuracy. Zhang et al. [21] developed the ED-Swin model based on an improved Swin Transformer for cassava disease classification in UAV imagery, significantly enhancing recognition performance in complex backgrounds. These advancements indicate the significant application value of the Swin Transformer in agricultural and geoscientific image analysis.

Additionally, Spatial Pyramid Pooling (SPP) [22] converts feature maps of variable sizes into fixed-length vectors through multi-scale pooling operations, enhancing the model's robustness to scale variations. It is commonly employed to improve the performance of classification and detection tasks.

In the field of fine-grained image classification, extracting discriminative features from limited samples has long been a research focus. Tang et al. [23] proposed a two-stage meta-learning framework for few-shot fine-grained recognition. By constructing a multi-scale feature pyramid and a multi-level attention pyramid, their method captures discriminative information at different granularities without requiring additional annotations, achieving leading results on standard fine-grained datasets such as CUB-200-2011 and Stanford Cars. Wang et al. [24] designed a depth-wise separable residual attention module and a global spatial attention mechanism based on Swin Transformer for fine-grained pest image recognition. By enhancing local feature representation and fusing multi-scale information, their method achieved strong performance on datasets such as IP102. Zhang et al. [25] introduced a multimodal fine-grained Transformer model that combines self-supervised learning with fine-grained recognition and leverages joint multimodal information from images and natural language descriptions, significantly improving classification accuracy

in few-shot pest recognition tasks. Ding et al. [26] proposed an attention pyramid convolutional network that uses a top-down feature pathway and a bottom-up attention pathway to exploit both high-level semantic and low-level detailed information, together with an ROI-guided refinement strategy, further advancing fine-grained classification accuracy.

Existing research has made important progress in both multimodal fusion and Transformer-based multi-scale representation. However, land classification scenarios have their own specific characteristics. First, phenological changes in crop categories cause visual features to fluctuate significantly over time. Second, geographical environment differences lead to variations in the growth status of the same crop. Third, some categories have scarce samples, and spatiotemporal metadata may be missing. Traditional fine-grained methods do not fully account for these characteristics. By drawing on the above ideas of feature mining and multimodal fusion, the proposed STF-Net achieves precise adaptation to land classification scenarios through spatiotemporal metadata encoding, cross-modal multi-head attention fusion, and an adaptive fallback branch for missing metadata. It constructs a fine-grained land classification model that effectively fuses visual and spatiotemporal metadata while possessing multi-scale modeling capabilities, thereby improving the accuracy and practicality of land classification and addressing the shortcomings of existing methods in this task.

3. Materials and Methods

3.1. Dataset Analysis

This study employs a field-captured land category photograph dataset from the Guangxi region of China for evaluation. The dataset contains a total of 38,580 land category photographs, covering 91 types of surface features (including 87 fine-grained cultivated land subcategories). The land category photograph dataset exhibits the following characteristics:

(1) Land categories exhibit spatiotemporal characteristics. The appearance of the same land category can vary significantly across different time periods and seasons, which increases the difficulty for models to discriminate between different land categories. Figure 1 displays land category photographs of the same parcel at different periods (a)–(d) and photographs of different parcels at different periods (e)–(h), presented in pairs. It illustrates the spatiotemporal variation features and growth environment characteristics of typical crops in Guangxi, China. The temporal evolution of the same parcel (a)–(d) clearly shows that crop morphology possesses distinct visual characteristics in different seasons. The spatial comparison of different parcels, (e)–(h) in Figure 1, reveals the differentiation in environmental adaptation, such as the stark contrast of ipomoea aquatica in paddy fields versus irrigated fields, and colocasia esculenta in cultivated land versus depressions. The visual similarity of rice paddies from the sowing stage to the maturity stage is low. These observations indicate that phenological cycles and geographical environments cause the visual features of the same land category to differ greatly. The discriminative ability of traditional visual models is insufficient, posing a severe challenge to the classification and recognition tasks of deep learning models.

(2) The land use classification dataset used in this study encompasses 5 major categories and 91 fine-grained subcategories, comprising a total of 38,474 samples. Some subcategories have extremely limited sample sizes, with certain classes containing as few as a dozen samples. This characteristic necessitates that the model, in achieving fine-grained land classification, must address challenges such as a wide variety of categories and imbalanced sample distribution.

Figure 1. Examples of spatiotemporal characteristics and growth environments of typical crops in Guangxi.

(3) There is a lack of spatiotemporal metadata. We have analyzed the spatiotemporal metadata of the dataset. 89.69% of the photographs possess capture time information, 89.56% have geographical location information, and the proportion with complete spatiotemporal information is 89.56%. This means that 10.44% of the photographs are missing spatiotemporal metadata. Spatiotemporal information holds significant auxiliary discriminative value in distinguishing between different land feature categories with similar visual characteristics. For samples lacking spatiotemporal information, no performance gain from spatiotemporal information fusion can be expected. Moreover, using invalid fill values for missing data may introduce noise, potentially interfering with the model's judgment.

To enhance the accuracy and robustness of land category recognition models, this study aims to design a multimodal fusion model that can effectively integrate image visual information with its corresponding spatiotemporal metadata to improve the classification accuracy of land category photographs. By fusing metadata such as capture time and geographic location with image features, the model can leverage these auxiliary cues to reduce ambiguity and more reliably distinguish between different land categories in complex environments. We employ image data augmentation methods to address the issue of imbalanced sample sizes. Additionally, a fallback mechanism is designed: when metadata is complete, the model fully utilizes the advantages of multimodal fusion; when metadata is missing, the model automatically reverts to relying solely on image visual features for classification. This ensures reliable operation of the model under various input conditions. This multimodal fusion strategy is expected to comprehensively improve both the accuracy and robustness of land category recognition.

3.2. Model Overview

Given field-captured photographs of ground objects and their associated metadata, the task is to determine the corresponding land category of each photograph. The categories include various fine-grained crop cultivation statuses under cultivated land, as well as forest land, garden land, roads, filled earth, etc., totaling 91 classes. Cultivated land is further subdivided into 87 subcategories. Some of these categories exhibit minimal visual differences, presenting challenges such as high inter-class feature similarity and significant

intra-class feature variation, making classification particularly difficult. The input for this task is an RGB photograph of a ground object I_n, along with its capture timestamp t_n and longitude–latitude coordinates l_n. The output is the predicted land category label y_n. In cases where metadata is missing, the model must predict the category based solely on the image.

The workflow of our land category recognition model, as illustrated in Figure 2, comprises four main components: an image visual backbone, a spatiotemporal encoding module, a multimodal fusion module, and a classification output layer. First, a pre-trained Swin Transformer is utilized as the visual backbone to extract hierarchical visual feature representations from the input photograph I_n. Concurrently, the photograph's capture time t_n and geographic coordinates l_n are fed into the spatiotemporal encoding module, which generates corresponding temporal and spatial features and combines them to form a spatiotemporal meta-feature vector. Then, within the fusion module, the visual features and spatiotemporal meta-features are interactively fused through a multimodal attention mechanism, resulting in a fused joint representation. Finally, the joint features are passed through a fully connected classification layer to output a probability distribution over the land categories. The entire model is trained end-to-end by minimizing the cross-entropy loss between the predicted values and the ground truth labels to learn the parameters of all modules. During the inference phase, the model directly predicts the land category based on the input photograph and its metadata. This design enables the model to leverage both the powerful image representation capability of the Swin Transformer and the effective integration of temporal and spatial prior knowledge, thereby meeting the requirements for fine-grained classification of land category photographs.

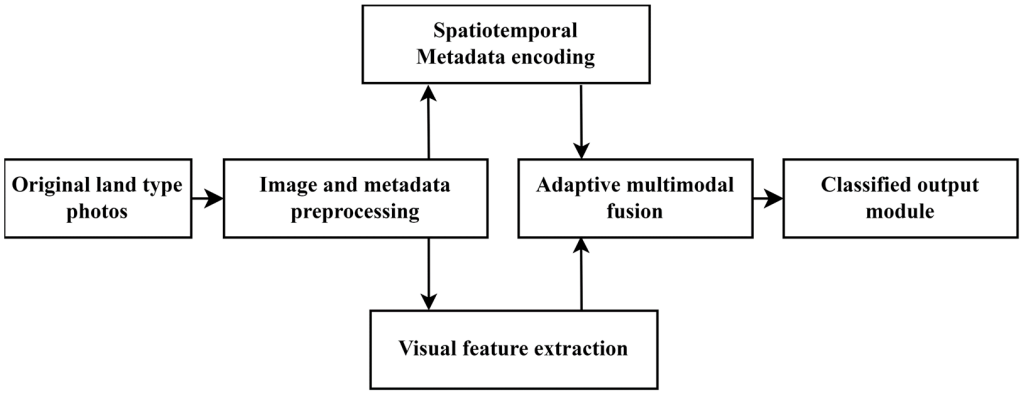


Figure 2. Flowchart of the land type classification model.

3.3. Methodology

This section systematically describes the core components and key techniques of the STF-Net model. The overall structure of the model is shown in Figure 3, and the data flow proceeds as follows. The input image first passes through the Swin Transformer visual backbone to extract multi-scale features, which are then aggregated into a fixed-dimensional image feature vector by the Spatial Pyramid Pooling module. Meanwhile, the capture time and GPS coordinates are encoded through temporal and spatial encoders, then mapped by a multilayer perceptron into a metadata feature vector. These two vectors are jointly fed into the multi-head attention fusion module for cross-modal interaction. If metadata is missing, the fallback branch is automatically activated and uses the image features directly. Finally, the fused features or the fallback features are sent to the classifier for land category prediction. The detailed design of each module is described in the following subsections.

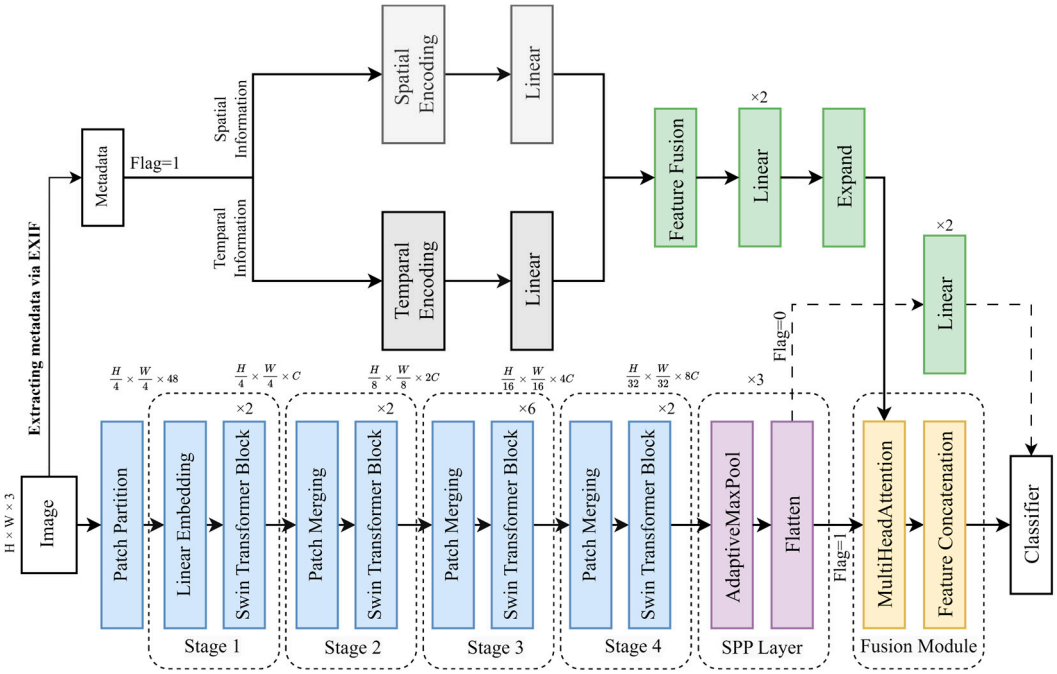


Figure 3. The STF-Net network structure diagram.

3.3.1. Visual Backbone Network

This paper adopts the Swin Transformer as the backbone network for visual feature extraction. The network divides the input image into fixed-size patches and converts them into a token sequence through linear projection. Subsequently, the sequence passes through multiple Transformer layers for feature transformation, and the spatial resolution is progressively reduced while the channel count is increased via Patch Merging operations, thereby constructing a pyramidal multi-scale feature representation. The Swin Transformer introduces the Shifted Window mechanism, which shifts the self-attention windows between adjacent layers to enable effective modeling of cross-region contextual information. This design preserves the advantages of the Transformer in capturing long-range dependencies while significantly reducing the computational cost of global self-attention. The Swin Transformer weights pre-trained on the ImageNet dataset are used for initialization and fine-tuned on the target dataset.

The final layer of the Swin Transformer outputs a low-resolution feature map. To convert it into a fixed-length feature vector for subsequent multimodal fusion, a Spatial Pyramid Pooling module is appended after the backbone network. This module performs multi-scale pooling operations on the input feature map using three pooling windows of sizes 1×1, 2×2, and 4×4, yielding 1, 4, and 16 pooled outputs, respectively. These outputs are flattened and concatenated to form a unified feature vector. Assuming the input feature map has C channels, the output dimension of the SPP module is $21 \times C$. This structure aggregates feature responses at different spatial scales, helping the model capture both broad scene context and local detail information simultaneously.

3.3.2. Temporal and Spatial Metadata Encoding

The capture time and geographic location contain valuable prior information for land category recognition. Cultivated land images captured in different months typically correspond to distinct crop types or growth stages, while vegetation distribution also varies significantly across different geographical regions. Therefore, this paper encodes temporal and spatial metadata into structured feature vectors separately, which are then fused with visual features.

To capture the periodic patterns in capture time, we employ a sine–cosine periodic encoding method. Specifically, the timestamp is decomposed into three periodic components: season, month, and day of the year. For each component, its sine and cosine values are computed to form a dual-channel periodic encoding. Let m denote the month and d denote the day of the year. The encoding is calculated as follows:

Seasonal encoding:

$$S_{sin} = sin\left(\frac{2\pi \cdot \frac{m-1}{3}}{4}\right) \tag{1}$$

$$S_{cos} = cos\left(\frac{2\pi \cdot \frac{m-1}{3}}{4}\right) \tag{2}$$

Monthly encoding:

$$M_{sin} = sin\left(\frac{2\pi \cdot m}{12}\right) \tag{3}$$

$$M_{cos} = cos\left(\frac{2\pi \cdot m}{12}\right) \tag{4}$$

Daily encoding:

$$D_{sin} = sin\left(\frac{2\pi \cdot d}{365}\right) \tag{5}$$

$$D_{cos} = cos\left(\frac{2\pi \cdot d}{365}\right) \tag{6}$$

where m denotes the month of the capture time, d denotes the day of the year of the capture time, S_{sin} and S_{cos} represent the sine and cosine encoding values of the seasonal feature, M_{sin} and M_{cos} represent the sine and cosine encoding values of the monthly feature, and D_{sin} and D_{cos} represent the sine and cosine encoding values of the daily feature.

The seasonal, monthly, and daily features obtained from the above encoding are concatenated into F_T, which serves as the final temporal feature.

This encoding method effectively eliminates discontinuities at periodic boundaries, enabling the model to learn cyclic temporal patterns. By concatenating the encodings from the three periodic scales described above, a 6-dimensional temporal feature vector is obtained.

Geographic location is represented by GPS longitude and latitude coordinates. We first normalize the longitude and latitude and use the normalized continuous coordinates as basic features. To further capture regional semantics, GeoHash encoding is introduced to discretize the continuous coordinates into grid identifiers. In this paper, a 5-digit GeoHash is adopted, corresponding to a spatial resolution of approximately 4.9 km. The GeoHash output is a Base32 string, which is converted into an integer index via a lookup table and then mapped to a D_G-dimensional vector through an embedding layer. This discrete encoding enables adjacent locations to have similar representations in the feature space, thereby helping the model capture prior geographic distribution patterns. Finally, the normalized coordinate features and the GeoHash embedding vector are concatenated to form the spatial feature representation. The normalized longitude and latitude coordinates, together with the 5-digit GeoHash embedding vector are concatenated into an 18-dimensional raw spatial feature, which is then mapped to a 32-dimensional final spatial feature F_S through a two-layer MLP.

We concatenate the temporal and spatial encodings to obtain the initial metadata feature F_{meta}^{init}. A two-layer multilayer perceptron is then applied for nonlinear transformation and dimensionality reduction: the first layer expands the dimension to 128 with ReLU activation, and the second layer compresses the feature to 32 dimensions, yielding the final

metadata feature vector F_M. This vector concisely expresses the joint information of capture time and location and can learn the association patterns between spatiotemporal factors and land categories through training. If the input image lacks metadata, the system sets the flag bit to 0, which triggers the fallback mechanism in the subsequent fusion module.

3.3.3. Multi-Head Attention Modality Fusion

To fully exploit the complementary information between image features and metadata features, this paper designs a modality fusion module based on a multi-head self-attention mechanism. Traditional methods, such as directly concatenating features into $[F_I; F_M]$ before feeding them into a classifier, are simple to implement but fail to explicitly model the interaction between the two modalities. This module aims to adaptively adjust the representation of visual features based on metadata content. For example, seasonal information can guide the model to focus on visual cues related to crop maturity.

The specific implementation is as follows: First, the image feature F_I is reduced to 512 dimensions via a fully connected layer, while the metadata feature F_M is projected to the same 512-dimensional space through another fully connected layer. With both features now sharing the same dimensionality, they can serve as an input sequence for the Transformer self-attention mechanism. Treating F_I and F_M as two tokens, they are concatenated into a sequence of length 2: F_I, F_M, which is then fed into a multi-head attention layer. In this layer, the query (Q), key (K), and value (V) vectors are all derived from linear transformations of F_I and F_F. The multi-head mechanism allows the model to learn interaction patterns between image and metadata from multiple representation subspaces. The attention computation outputs a weighted fusion result for each modality with respect to itself and the other modality. This resultant vector serves as the final fused feature F_F, which encapsulates both visual patterns and spatiotemporal information, thereby possessing strong discriminative power.

Considering that, in practical applications, some images may lack temporal or location metadata, a fallback branch is additionally designed to ensure stable model performance when metadata is incomplete. When the flag bit $Flag = 0$, the model bypasses the attention fusion module and relies solely on image features for classification. Specifically, the image feature F_I' is passed through an expanded fully connected layer to 1024 dimensions, yielding feature F_F. This maintains dimensional consistency with the fused feature F_F, facilitating subsequent classifier processing. When $Flag = 1$, the fused feature F_F is used for prediction.

During training, the input to the classifier is dynamically selected as either F_F or F_I' based on the metadata flag of each sample. During inference, if an input image lacks metadata, the fallback path is automatically executed. This dynamic decision-making mechanism enables the model to fully leverage multimodal information when complete metadata is available while gracefully degrading to a robust visual-feature-based classifier when metadata is missing. Experimental results show that the fusion module significantly enhances classification performance, and the fallback mechanism maintains model usability without substantially compromising accuracy, thereby ensuring excellent performance across varying input conditions.

3.3.4. Classifier

The input to the classifier is the 1024-dimensional fused feature or the fallback feature. This feature is first mapped to 256 dimensions through a linear layer, followed by batch normalization to stabilize training and accelerate convergence. An SiLU activation function is then applied to introduce nonlinearity, and finally, another linear layer transforms the feature dimension to the number of classes. During training, the cross-entropy loss function

is adopted, with a small amount of label smoothing incorporated to improve the model's generalization ability.

Since the model processes both metadata-complete and metadata-missing samples during training, the classifier learns to adapt to two distinct feature distributions: when the input includes complete metadata, the classifier makes decisions based on the multi-modal fused features; when metadata is absent, it relies on the visual fallback features for classification. This design enhances the model's adaptability to varying input conditions.

During inference, the logits output by the model are transformed into a class probability distribution via the Softmax function, and the class with the highest probability is taken as the final recognition result. The classification loss is defined by the multi-class cross-entropy loss as follows:

Cross-entropy loss function:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} log(\hat{y}_{i,c}) \quad (7)$$

where N is the total number of samples, C is the number of classes, $y_{i,c}$ denotes the ground-truth probability of sample i belonging to class c, taking 1 if the true class of sample i is c and 0 otherwise, and $\hat{y}_{i,c}$ represents the predicted probability that sample i belongs to class c.

4. Results

4.1. Datasets and Evaluation Metrics

We employ a dataset of field-captured object photographs from the Guangxi region of China for evaluation. This dataset encompasses 91 types of surface features, including five major categories: cultivated land (subdivided into 87 fine-grained crop cultivation statuses), forest land, garden land, roads, and filled earth. Among these, the cultivated land category constitutes the majority and presents the highest difficulty in differentiation. The dataset is partitioned into training, validation, and test sets in a 7:1:2 ratio. To evaluate the model's ability to distinguish fine-grained cultivated land categories, we ensure that all 87 subcategories have sufficient samples in both the training and test sets. The entire dataset comprises 38,474 land category photographs, with an uneven distribution across categories. Therefore, during training, we adopt a weighted loss function to mitigate the impact of class imbalance.

We employ three evaluation metrics, namely Overall Accuracy (OA), Recall, and F1-score, to assess the fine-grained land classification performance of the model. OA reflects the overall classification accuracy across the 91 land categories, while Recall and F1-score focus more on the balanced recognition performance among different land categories. The specific calculation formulas for these metrics are as follows.

Overall Accuracy (OA): The probability that the model's predicted category matches the true category among all land type samples.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Recall: For a specific land category, the proportion of samples that actually belong to that category and are correctly predicted by the model.

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

F1: For a specific land category, the harmonic mean of its Precision and Recall, comprehensively reflecting the model's recognition accuracy and coverage for that category.

$$F1 = 2 \cdot \frac{Precision \cdot Receall}{Precision + Recall} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

TP: The number of samples that actually belong to a certain land category and are correctly predicted as that category by the model.

TN: The number of samples that do not belong to a certain land category and are correctly predicted as other categories by the model.

FP: The number of samples that do not belong to a certain land category but are incorrectly predicted as that category by the model.

FN: The number of samples that actually belong to a certain land category but are incorrectly predicted as other categories by the model.

4.2. Experimental Results and Analysis

We uniformly resized the input images to 224 × 224 pixels. For data augmentation, operations such as random cropping, horizontal flipping, and color jittering were applied to the training set photos, while the validation and test sets underwent only center cropping and normalization. The Swin Transformer backbone was initialized with ImageNet pre-trained weights. Model training employed the AdamW optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-3}. The loss function was cross-entropy with label smoothing set to 0.1. Training was conducted for 50 epochs.

This paper compares the proposed method with the following mainstream approaches, including ResNet [27], EfficientNet [28], Vit [8] and Swin Transformer [11].

We compared the overall performance of different models on the test set, and the experimental results are shown in Table 1. As a representative of traditional CNN methods, ResNet achieved the lowest performance, indicating its difficulty in handling the diverse viewpoints and complex backgrounds in fine-grained land cover recognition. The accuracy of ViT was slightly better than that of ResNet but remained relatively low compared to other methods, confirming that ViT underperforms when data scale is limited and lacks multi-scale feature extraction capability. Swin Transformer significantly outperformed both CNN and ViT, demonstrating that its hierarchical structure and shifted window mechanism can better capture multi-scale visual features. Dynamic MLP dynamically generates classifier weights using spatiotemporal metadata and has achieved excellent results on fine-grained datasets such as iNaturalist. However, on our land classification task, it achieved an accuracy of 88.10% and a recall of 81.83%, slightly lower than that of the Swin Transformer. The reason may be that some categories in our dataset have high spatiotemporal overlap and missing metadata, so the strategy of relying solely on metadata to dynamically generate weights does not fully exploit the discriminative power of visual features. The proposed STF-Net achieved an accuracy of 93.54%, which is 4.02% higher than the pure vision-based Swin Transformer and 5.44% higher than Dynamic MLP, with the F1-score increasing from 0.88 to 0.92. The experimental results indicate that incorporating spatiotemporal metadata combined with a multi-head attention fusion mechanism provides the model with powerful discriminative priors: temporal features help the model distinguish cultivated land crops with seasonal cycle characteristics, while spatial features assist in recognizing land cover categories with regional variations.

Table 1. Comparison results of land type recognition experiments.

Method	OA	Recall	F1
ResNet	81.62%	82.13%	0.81
Vision Transformer	82.94%	82.46%	0.82
EfficientNet	86.15%	85.78%	0.86
Swin Transformer	89.52%	88.96%	0.88
Dynamic MLP	88.10%	81.83%	0.81
STF-Net	93.54%	93.14%	0.92

Figure 4 presents the classification confusion matrix of STF-Net on 20 representative land categories, where rows and columns correspond to true and predicted categories, respectively. To balance visual readability and task completeness, 20 typical categories are selected for presentation, covering five major land types—cultivated land, forest land, garden land, road, and filled earth—and including fine-grained subcategories of cultivated land at different crop growth stages. The selection principle takes into account few-shot categories, easily confused fine-grained categories, and the distribution of major land types, fully reflecting the model's discriminative ability in the core land classification task.

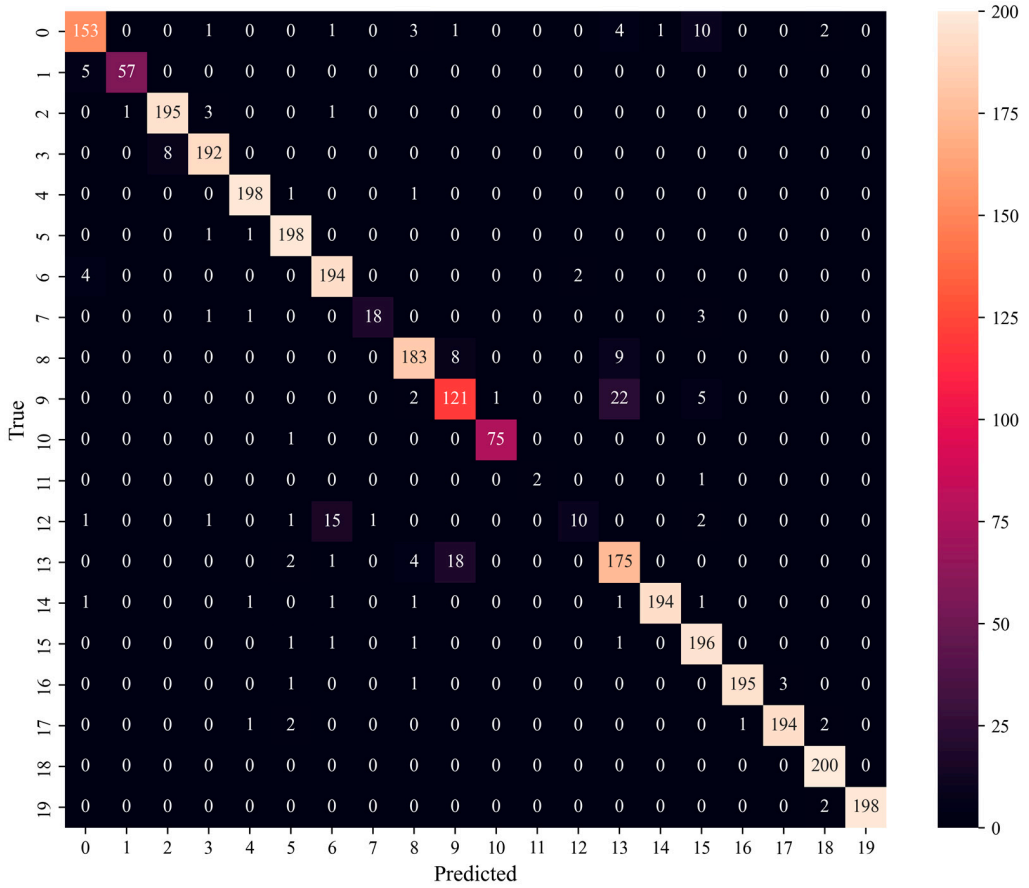


Figure 4. Confusion matrix of 20 representative land types.

In the matrix, the diagonal elements of the vast majority of categories are significantly higher than the off-diagonal elements, indicating that the model's predictions are consistent with the true labels for most samples, and the overall classification error rate is kept at a low level. For macro-level land categories such as forest land and road, the off-diagonal elements at the corresponding row and column positions are nearly zero, with recognition errors approaching zero. This reflects the model's strong robustness in capturing global

semantic features of prominent land objects, enabling it to easily distinguish land categories with clear differences.

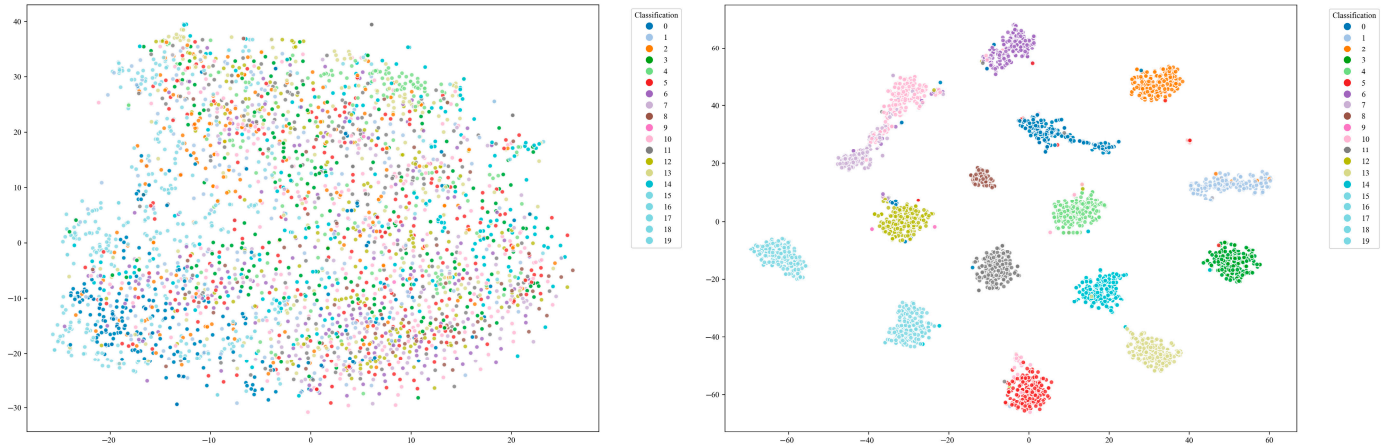
For fine-grained categories within cultivated land at different growth stages, such as categories 9 and 13, and categories 6 and 12, obvious misclassifications occur, with off-diagonal elements higher than the diagonal elements for the same categories. The causes of such confusion are closely related to the intrinsic characteristics of the land types and the nature of the task. Visual features are highly homogeneous; low-level visual features such as leaf morphology, plant height, and texture exhibit minimal differences, making it difficult to distinguish effectively using visual cues alone. The phenological periods of crops overlap, and they are concentrated in the same geographical region, resulting in insufficient discriminative power of spatiotemporal metadata, which further weakens the effectiveness of cross-modal fusion. Compared with macro-level land categories, the overlap in visual features and spatiotemporal priors for fine-grained land categories makes it difficult for the model to form clear decision boundaries in the fine-grained feature space, ultimately leading to confusion between categories.

Overall, the model performs excellently in recognizing major land categories, while fine-grained category confusion remains an inherent challenge in fine-grained land classification tasks. By integrating spatiotemporal information, the proposed model significantly reduces the degree of confusion among fine-grained categories. Compared with the model without metadata fusion, the error rate is effectively controlled, validating the critical role of spatiotemporal information in alleviating fine-grained category confusion.

Figure 5, the t-SNE dimensionality reduction visualization, displays the distribution of image features after dimensionality reduction, used to evaluate the effectiveness of the features learned by the model in distinguishing different categories. We projected the final fused features of the test set samples onto a two-dimensional plane using the t-SNE algorithm, with samples from different categories represented by distinct colors. As shown in Figure 5, data points from various categories form relatively separated clusters in the two-dimensional feature space. Samples from the same category tend to aggregate in close proximity, while samples from different categories are separated from one another. This indicates that the features extracted by the model already possess strong discriminative properties in the high-dimensional space, and the t-SNE projection visualizes this structure of intra-class aggregation and inter-class separation. For instance, categories with significant differences, such as forest land and roads, appear far apart in the figure, with almost no overlap between their clusters. In contrast, clusters of different crop categories within cultivated land are relatively closer in distance, yet it is still evident that the model has distinguished them based on subtle differences. Those crop categories that are visually extremely similar may lie close to each other in the feature space, suggesting that even with the fusion of spatiotemporal metadata, certain similarities remain between the features of these categories.

We also compared the computational load and parameter count of several models to assess whether the computational cost of the proposed method remains within an acceptable range while achieving performance improvements. Table 2 lists the computational load and parameter counts of the compared models and STF-Net. The models vary significantly in complexity. ViT has the highest computational load, reaching 11.29 GFLOPs, with a parameter size of 57.31 M, reflecting the substantial computational overhead of the self-attention mechanism, which leads to higher training and inference costs. ResNet has a computational load of 3.68 GFLOPs and 21.29 M parameters, offering moderate scale but lower classification accuracy. The Swin Transformer has a computational load of 5.76 GFLOPs and 33.07 M parameters, placing it between ViT and EfficientNet. Through local window self-attention, it achieves a good balance between accuracy and efficiency.

Dynamic MLP has a computational load of 8.91 GFLOPs and 53.74 M parameters, with computational costs close to those of ViT and a relatively large parameter size, indicating that the strategy of dynamically generating classifier weights, while improving fine-grained classification capability, increases computational burden. STF-Net has a computational load of 2.98 GFLOPs and 19.20 M parameters, lower than both the Swin Transformer and Dynamic MLP. The metadata branch introduces minimal overhead, and replacing large fully connected layers with several smaller MLP layers enables high performance at lower complexity. The prior knowledge provided by multimodal information allows the model to achieve excellent results without relying on an excessively large visual backbone, thereby controlling model size while maintaining accuracy. STF-Net holds certain advantages in practical deployment, achieving performance improvements without significantly increasing computational costs, and demonstrating good engineering application potential.



(**a**) (**b**)

Figure 5. t-SNE dimensionality reduction visualization. (**a**) Before model training. (**b**) After model training.

Table 2. Model calculation amount and parameter amount.

Method	Calculation Amount	Parameter Quantity
ResNet	3.68 G	21.29 M
Vision Transformer	11.29 G	57.31 M
EfficientNet	1.57 G	17.58 M
Swin Transformer	5.76 G	33.07 M
Dynamic MLP	8.91 G	53.74 M
STF-Net	2.98 G	19.20 M

5. Discussion

5.1. Ablation Experiments and Analysis

To gain a deeper understanding of the role of spatiotemporal information in land use type recognition, we conducted ablation experiments, with the results presented in Table 3. We summarized the performance of different information fusion models under the Swin Transformer backbone to analyze the contribution of each module. We compared the following models: a pure Swin Transformer visual baseline model without metadata fusion, a model fused only with temporal information, a model fused only with spatial information, and the complete STF-Net model that simultaneously fuses both temporal and spatial information. The table provides the accuracy, recall, and F1-score for each model.

Table 3. Ablation experiment results.

Method	OA	Recall	F1
Swin Transformer	89.52%	88.96%	0.88
Swin Transformer + Time information fusion	92.27%	91.51%	0.91
Swin Transformer + Spatial information fusion	91.58%	90.56%	0.9
STF-Net	93.54%	93.14%	0.92

The results show that the baseline Swin Transformer model achieved an accuracy of approximately 89.52%, representing the starting performance without any metadata. After fusing temporal information, the model's accuracy increased to 92.27%, with a recall of 91.51% and an F1-score of 0.91, marking a 2.8% improvement in accuracy over the baseline. This demonstrates that incorporating temporal features yields significant performance gains, validating the value of temporal dimension information in land cover classification. The model fused with spatial location information achieved an accuracy of 91.58%, a recall of 90.56%, and an F1-score of 0.90, representing a 2.1% improvement over the baseline—slightly lower than the gain from temporal features. This indicates that geographic spatial features also contribute to performance, though their independent effect is smaller than that of temporal features.

The STF-Net model, which simultaneously fuses both temporal and spatial information, performed best, achieving an accuracy of 93.54%, a recall of 93.14%, and an F1-score of 0.92. Compared to the baseline model, this represents a 4.02% improvement in accuracy, exceeding the gains from fusing either metadata type alone. From the perspective of recall, the temporal-only fusion achieved 91.51% and the spatial-only fusion achieved 90.56%. If the information provided by the two modalities were completely non-overlapping, this would indicate that the discriminative information from temporal and spatial features is both overlapping and complementary. The overlapping part is reflected in the macro-level patterns of crop distribution that both modalities capture, while the complementary part is shown in temporal features focusing on phenological cycles and spatial features focusing on geographical distribution. When used together, the model can cross-validate across both temporal and spatial dimensions, thereby more accurately determining category boundaries.

Thus, the two types of metadata exhibit a degree of complementarity, and their combined use further enhances accuracy. However, the performance gain from fusing spatiotemporal information is not simply additive; the 93.54% accuracy is slightly higher than that of temporal fusion alone but does not reach the sum of the improvements from temporal and spatial features. This indicates that the discriminative information provided by temporal and spatial features partially overlaps. Overall, temporal metadata contributes most significantly to model performance improvement, while spatial metadata also has a positive but relatively smaller impact. Utilizing both achieves the highest accuracy, demonstrating that the discriminative information they provide is both overlapping and complementary. These results fully validate the effectiveness of the spatiotemporal auxiliary features we introduced.

To validate the generalizability of multimodal fusion across different visual backbones, we conducted ablation experiments on the role of spatiotemporal information using the EfficientNet backbone. Table 4 presents the performance comparison of four configurations on the test set: EfficientNet with a single modality, EfficientNet fused with temporal information, EfficientNet fused with spatial information, and EfficientNet fused with both temporal and spatial information. As can be seen, the single-modality EfficientNet achieves an accuracy of 86.15%, serving as the baseline without metadata. After fusing temporal information, the accuracy increases to 90.39%, recall improves from 85.78% to

88.27%, and the F1-score rises to 0.89, representing a 4.24% accuracy improvement—a gain comparable to that observed with the Swin Transformer backbone when incorporating temporal features. This indicates that temporal metadata also provides significant benefits for CNN-based models. The accuracy for the model fused with spatial information is 87.28%, only a 1.13% improvement over the baseline, indicating a relatively limited gain. The model that simultaneously fuses both temporal and spatial information achieves the best performance, with an accuracy of 91.04%, a recall of 89.47%, and an F1-score of 0.90. The overall trend aligns with that observed for the Swin Transformer backbone: whether using CNN or Transformer architectures, incorporating temporal and geographic metadata consistently enhances classification accuracy. This demonstrates the robustness of our multimodal fusion strategy, which can be extended to other image classification models in the future to improve their ability to distinguish fine-grained categories.

Table 4. EfficientNet backbone plus spatiotemporal fusion.

Method	OA	Recall	F1
EfficientNet	86.15%	85.78%	0.86
EfficientNet + Temporal information fusion	90.39%	88.27%	0.89
EfficientNet + Spatial information fusion	87.28%	87.94%	0.88
EfficientNet + Spatiotemporal information fusion	91.04%	89.47%	0.9

Based on the validation of the effectiveness of spatiotemporal metadata, we further compared the impact of different cross-modal fusion strategies on classification performance. Experiments were conducted with the Swin Transformer as the visual backbone under the same training settings, employing four fusion strategies: simple concatenation, FiLM, cross-attention, and the multi-head attention fusion used in this paper. Simple concatenation directly concatenates image features and metadata features before feeding them into the classifier. FiLM modulates image features by generating affine transformation parameters from metadata features. Cross-attention treats image features as queries and metadata features as keys and values for interaction. The results are shown in Table 5.

Table 5. Comparison of different fusion strategies.

Fusion Strategy	OA	Recall	F1
No Metadata Fusion	89.52%	88.96%	0.88
Simple Concatenation	92.22%	92.87%	0.92
Cross-Attention	91.54%	92.04%	0.91
FiLM	91.64%	92.13%	0.91
Our	93.54%	93.14%	0.92

It can be observed from Table 5 that all four fusion strategies significantly outperform the baseline model without metadata, indicating that spatiotemporal information can effectively aid classification under different fusion approaches. Simple concatenation achieves an accuracy of 92.22%, already demonstrating the value of metadata. Cross-attention and FiLM achieve accuracies of 91.54% and 91.64%, respectively, slightly lower than simple concatenation. The proposed multi-head attention fusion outperforms the other three methods in terms of OA, Recall, and F1, achieving an accuracy of 93.54% and an F1-score of 0.92. This result indicates that treating images and metadata as different tokens for self-attention interaction captures the correlations between the two modalities better than simple concatenation or one-way modulation, thereby achieving superior classification performance.

5.2. Robustness Test Experiment

In real-world scenarios, photos sometimes lack time or location information. To address this, we designed a fallback branch that uses the pure visual path when metadata is unavailable, while the attention fusion branch fully utilizes spatiotemporal information to improve accuracy when metadata is available. As a comparison, the zero imputation strategy retains the fusion module when metadata is missing, simply setting the missing features to zero. We set seven missing ratios from 0% to 100% for metadata absence. Since the original dataset has only 89.56% complete spatiotemporal metadata, we manually removed metadata to obtain the 0% and 10% missing ratios.

It can be seen from Table 6 that when the metadata missing ratio is low (0–20%), the two strategies perform similarly, with the fallback branch slightly better. When the missing ratio reaches 50%, the fallback branch achieves an OA of 92.74% while zero imputation achieves 91.64%, a gap of 1.1 percentage points. At missing ratios above 80%, the difference becomes more pronounced: at 80% missing, the fallback branch OA is 91.82% versus 89.58% for zero imputation; at 90% missing, 91.27% versus 87.93%; when fully missing, 89.52% versus 86.45%.

Table 6. Robustness test of missing spatiotemporal information.

Missing Metadata Ratio	Fallback (Ours)			Zero Imputation		
	OA	**Recall**	**F1**	**OA**	**Recall**	**F1**
0%	93.52%	93.35%	0.92	93.52%	93.35%	0.92
10%	93.41%	93.10%	0.92	92.87%	92.53%	0.91
20%	93.29%	92.85%	0.91	92.87%	92.53%	0.91
50%	92.74%	92.31%	0.91	91.64%	91.28%	0.90
80%	91.82%	91.30%	0.90	89.58%	89.03%	0.88
90%	91.27%	90.53%	0.89	87.93%	87.26%	0.86
100%	89.52%	88.96%	0.88	86.45%	85.71%	0.84

The reason is that zero imputation sends zero vectors for missing metadata into the fusion module, effectively introducing invalid vectors unrelated to the true distribution. Although the multi-head attention mechanism can partially suppress such noise, when the missing ratio is high, the large amount of invalid information mixed into the fusion module still interferes with attention computation and degrades classification accuracy. The fallback branch, by contrast, skips the fusion module entirely, avoiding any disturbance from invalid information on image features and relying solely on visual features for classification, thus maintaining more stable performance under high missing ratios.

The above experiments demonstrate that the fallback branch has a clear performance advantage under high missing ratios, validating the necessity and rationality of this design. At a 50% missing ratio, the fallback branch still maintains an accuracy of 92.74%, significantly higher than the pure visual model's 89.52%. Under extreme missing scenarios of 90–100%, its performance is essentially on par with the pure visual model, introducing no additional instability. Overall, STF-Net maintains stable classification performance across various metadata missing ratios, showing good robustness when facing incomplete data in practical applications.

6. Conclusions

This paper addresses the task of fine-grained intelligent classification of land use type photographs by proposing a novel method that integrates multimodal spatiotemporal information fusion. By employing periodic temporal encoding and GeoHash spatial encoding to extract physically interpretable temporal and spatial features, which are deeply fused with

the image features from the Swin Transformer through a multi-head attention mechanism, we have achieved high-precision classification of land categories. Experimental results demonstrate that, compared to traditional unimodal models, this multimodal approach improves accuracy from 89.52% to 93.54% and achieves an F1-score of 0.92 on a dataset containing 91 fine-grained categories, representing a significant performance breakthrough and validating the effectiveness and potential of multi-dimensional information fusion.

Our method exhibits the following notable advantages: the decoupled spatiotemporal representation strategy successfully introduces an interpretable auxiliary feature space, enabling the model to learn seasonal variation patterns and regional distribution characteristics, thereby more accurately distinguishing categories that are challenging to differentiate based on visual information alone. This spatiotemporal feature space, constructed through sine encoding and GeoHash encoding, exhibits excellent periodic continuity and regional clustering properties. The adaptive fusion and dynamic decision-making mechanisms ensure the model's versatility. Through attention-weighted fusion based on metadata flags, our model fully leverages multimodal information when metadata is available and maintains high classification accuracy even when metadata is missing, effectively addressing the issue of incomplete data in practical applications.

Nevertheless, there remains room for further improvement in this study. In future work, we plan to incorporate additional metadata modalities, such as meteorological data and textual descriptions, to construct richer multimodal features. For categories that remain prone to confusion, we aim to design more advanced fusion networks, utilizing cross-modal Transformer encoder–decoder architectures to enable iterative interaction between images and metadata for extracting higher-level semantics. Subsequent efforts may involve using generative algorithms to mitigate the impact of class imbalance on classification accuracy and applying the proposed method to broader scenarios, such as integrating drone aerial imagery for land cover recognition and multi-temporal monitoring of land use changes. By further enhancing model performance and expanding its applicability, the approach presented in this paper holds the potential to offer more efficient land category recognition solutions for future land use monitoring.

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

Funding: This research was supported by the Technology Innovation Center for Natural Resources Monitoring and Evaluation of Beibu Gulf Economic Zone, Ministry of Natural Resources (No. BBW2025006), the Guangxi Major S&T Special Program (No. GuikeAA23062035-2), and the National Natural Science Foundation of China (No. 62262011).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study was provided by the Guangxi Institute of Natural Resources Survey and Monitoring. It is proprietary and not publicly available due to institutional data policies. Access to the dataset may be granted upon reasonable request to the corresponding author, subject to the formal approval of the Guangxi Institute of Natural Resources Survey and Monitoring.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun, Z.; Zhong, Y.; Wang, X.; Zhang, L. Identifying Cropland Non-Agriculturalization with High Representational Consistency from Bi-Temporal High-Resolution Remote Sensing Images: From Benchmark Datasets to Real-World Application. *ISPRS J. Photogramm. Remote Sens.* **2024**, *212*, 454–474. [CrossRef]
2. Mohammed, E.A.; Lakizadeh, A. A Swin Transformer-Based Method for Classification of Land Use and Land Cover Images. *Int. J. Adv. Soft Comp. Appl. (IJASCA)* **2024**, *16*, 328–347. [CrossRef]
3. Lu, T.; Wan, L.; Qi, S.; Gao, M. Land Cover Classification of UAV Remote Sensing Based on Transformer–CNN Hybrid Architecture. *Sensors* **2023**, *23*, 5288. [CrossRef] [PubMed]
4. Martinez-Sanchez, L.; See, L.; Yordanov, M.; Verhegghen, A.; Elvekjaer, N.; Muraro, D.; d'Andrimont, R.; Van Der Velde, M. Automatic Classification of Land Cover from LUCAS In-Situ Landscape Photos Using Semantic Segmentation and a Random Forest Model. *Environ. Model. Softw.* **2024**, *172*, 105931. [CrossRef]
5. Minetto, R.; Pamplona Segundo, M.; Sarkar, S. Hydra: An Ensemble of Convolutional Neural Networks for Geospatial Land Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6530–6541. [CrossRef]
6. Bello, I.; Zoph, B.; Le, Q.; Vaswani, A.; Shlens, J. Attention Augmented Convolutional Networks. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Seoul, Republic of Korea, 2019; pp. 3285–3294.
7. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12346, pp. 213–229.
8. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
9. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. In *Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 15908–15919.
10. Park, K.-B.; Lee, J.Y. SwinE-Net: Hybrid Deep Learning Approach to Novel Polyp Segmentation Using Convolutional Neural Network and Swin Transformer. *J. Comput. Des. Eng.* **2022**, *9*, 616–632. [CrossRef]
11. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Montreal, QC, Canada, 2021; pp. 9992–10002.
12. Li, Z.; Hu, J.; Wu, K.; Miao, J.; Zhao, Z.; Wu, J. Local Feature Acquisition and Global Context Understanding Network for Very High-Resolution Land Cover Classification. *Sci. Rep.* **2024**, *14*, 12597. [CrossRef]
13. Hao, M.; Dong, X.; Jiang, D.; Yu, X.; Ding, F.; Zhuo, J. Land-Use Classification Based on High-Resolution Remote Sensing Imagery and Deep Learning Models. *PLoS ONE* **2024**, *19*, e0300473. [CrossRef]
14. Mehmood, M.; Shahzad, A.; Zafar, B.; Shabbir, A.; Ali, N. Remote Sensing Image Classification: A Comprehensive Review and Applications. *Math. Probl. Eng.* **2022**, *2022*, 1–24. [CrossRef]
15. Thapa, A.; Horanont, T.; Neupane, B.; Aryal, J. Deep Learning for Remote Sensing Image Scene Classification: A Review and Meta-Analysis. *Remote Sens.* **2023**, *15*, 4804. [CrossRef]
16. Wang, X.; Xu, H.; Yuan, L.; Dai, W.; Wen, X. A Remote-Sensing Scene-Image Classification Method Based on Deep Multiple-Instance Learning with a Residual Dense Attention ConvNet. *Remote Sens.* **2022**, *14*, 5095. [CrossRef]
17. Tang, K.; Paluri, M.; Fei-Fei, L.; Fergus, R.; Bourdev, L. Improving Image Classification with Location Context. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*; IEEE: Santiago, Chile, 2015; pp. 1008–1016.
18. Sun, J.; Futahashi, R.; Yamanaka, T. Improving the Accuracy of Species Identification by Combining Deep Learning With Field Occurrence Records. *Front. Ecol. Evol.* **2021**, *9*, 762173. [CrossRef]
19. Yang, L.; Li, X.; Song, R.; Zhao, B.; Tao, J.; Zhou, S.; Liang, J.; Yang, J. Dynamic MLP for Fine-Grained Image Classification by Leveraging Geographical and Temporal Information. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New Orleans, LA, USA, 2022; pp. 10935–10944.
20. Jannat, F.-E.; Willis, A.R. Improving Classification of Remotely Sensed Images with the Swin Transformer. In *Proceedings of the SoutheastCon 2022*; IEEE: Mobile, AL, USA, 2022; pp. 611–618.
21. Zhang, J.; Zhou, H.; Liu, K.; Xu, Y. ED-Swin Transformer: A Cassava Disease Classification Model Integrated with UAV Images. *Sensors* **2025**, *25*, 2432. [CrossRef] [PubMed]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
23. Tang, H.; Yuan, C.; Li, Z.; Tang, J. Learning Attention-Guided Pyramidal Features for Few-Shot Fine-Grained Recognition. *Pattern Recognit.* **2022**, *130*, 108792. [CrossRef]
24. Wang, X.; Xiao, Z.; Deng, Z. Swin Attention Augmented Residual Network: A Fine-Grained Pest Image Recognition Method. *Front. Plant Sci.* **2025**, *16*, 1619551. [CrossRef] [PubMed]

25. Zhang, Y.; Chen, L.; Yuan, Y. Multimodal Fine-Grained Transformer Model for Pest Recognition. *Electronics* **2023**, *12*, 2620. [CrossRef]
26. Ding, Y.; Ma, Z.; Wen, S.; Xie, J.; Chang, D.; Si, Z.; Wu, M.; Ling, H. AP-CNN: Weakly Supervised Attention Pyramid Convolutional Neural Network for Fine-Grained Visual Classification. *IEEE Trans. Image Process.* **2021**, *30*, 2826–2836. [CrossRef] [PubMed]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Las Vegas, NV, USA, 2016; pp. 770–778.
28. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the Proceedings of the 36th International Conference on Machine Learning*; PMLR (Proceedings of Machine Learning Research): Long Beach, CA, USA, 2019; Volume 97, pp. 6105–6114.