frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 10 January 2025
DOI 10.3389/fenvs.2024.1515752

OPEN ACCESS

EDITED BY
Minghan Cheng,
Yangzhou University, China

REVIEWED BY
Beata Calka,
Military University of Technology in Warsaw, Poland
Yao Zhaosheng,
Yangzhou University, China

*CORRESPONDENCE
Yuefeng Liu,
godflys@163.com

RECEIVED 24 October 2024
ACCEPTED 23 December 2024
PUBLISHED 10 January 2025

CITATION
Gao B, Liu Y, Li Y, Li H, Li M and He W (2025) A vision-language model for predicting potential distribution land of soybean double cropping. *Front. Environ. Sci.* 12:1515752. doi: 10.3389/fenvs.2024.1515752

A vision-language model for predicting potential distribution land of soybean double cropping

Bei Gao, Yuefeng Liu*, Yanli Li, Hongmei Li, Meirong Li and Wenli He

Shaanxi Meteorological Service Center of Agricultural Remote Sensing and Economic Crops, Xi'an, China

Introduction: Accurately predicting suitable areas for double-cropped soybeans under changing climatic conditions is critical for ensuring food security anc optimizing land use. Traditional methods, relying on single-modal approaches such as remote sensing imagery or climate data in isolation, often fail to capture the complex interactions among environmental factors, leading to suboptimal predictions. Moreover, these approaches lack the ability to integrate multi-scale data and contextual information, limiting their applicability in diverse and dynamic environments.

Methods: To address these challenges, we propose AgriCLIP, anovel remote sensing vision-language model that integrates remote sensing imagery with textual data, such as climate reports and agricultural practices, to predict potential distribution areas of double-cropped soybeans under climate change. AgriCLIP employs advanced techniques including multi-scale data processing, self-supervised learning, and cross-modality feature fusion enabling comprehensive analysis of factors influencing crop suitability.

Results and discussion: Extensive evaluations on four diverse remote sensing datasets-RSICap RSIEval, MillionAID, and HRSID-demonstrate AgriCLIP's superior performance over state-of-the-art models. Notably, AgriCLIP achieves a 97.54% accuracy or the RSICap dataset and outperforms competitors across metrics such as recall F1 score, and AUC. Its efficiency is further highlighted by reduced computation a demands compared to baseline methods. AgriCLIP's ability to seamlessly integrate visual and contextual information not only advances prediction accuracy but also provides interpretable insights for agricultural planning and climate adaptation strategies, offering a robust and scalable solution for addressing the challenges of food security in the context of global climate change.

KEYWORDS

AgriCLIP, remote sensing, vision-language model, climate change, double-cropped soybeans, predicting distribution areas

1 Introduction

Remote sensing image segmentation is a critical task in the field of remote sensing and geographic information systems, providing essential information for land cover classification, environmental monitoring, and urban planning (Zhou et al., 2024). The segmentation of remote sensing images is not only necessary for the accurate interpretation of vast amounts of data but also crucial for the effective management and utilization of

natural resources. Given the increasing availability and resolution of remote sensing data, the need for advanced segmentation techniques has become more pronounced (Yuan et al., 2023). These techniques not only allow for the precise delineation of objects and regions within an image but also enable the extraction of meaningful patterns and features that are vital for a wide range of applications (Xu et al., 2021). Moreover, with the growing challenges posed by climate change, deforestation, and urbanization, the ability to monitor and analyze changes in the Earth's surface with high accuracy is more important than ever. This necessity has driven significant advancements in the field, leading to the development of various methods over the years, each with its strengths and limitations (Qi et al., 2022).

In early research on remote sensing image segmentation, traditional three-dimensional reconstruction techniques were widely used. These methods aimed to reconstruct the spatial structure of the Earth's surface through stereoscopic image pairs or photogrammetric techniques (Bigolin and Talamini, 2024). By leveraging geometric principles, traditional 3D reconstruction methods could segment images based on the relative positions and orientations of objects, providing detailed and accurate representations of the terrain (Li et al., 2024). However, these methods were computationally complex and required precise calibration and alignment of images, making them less practical for large-scale or real-time applications (Jung et al., 2024). Additionally, traditional 3D reconstruction techniques faced challenges in handling complex and heterogeneous landscapes, particularly when mixed pixels and uneven illumination conditions were present, which could significantly reduce the accuracy of the results (Tovihoudji et al., 2024). To overcome these issues, researchers began exploring alternative approaches that could offer more robust and scalable solutions. Compared to the limitations of manual and semi-automated methods, these emerging approaches demonstrated superior performance in processing large-scale data and achieving real-time capabilities, paving the way for further advancements in remote sensing image segmentation (Jung et al., 2024). By integrating advanced technologies like machine learning and deep learning, these methods exhibited higher efficiency and accuracy across various application scenarios, especially in handling complex landscapes, where they showed greater robustness and adaptability.

In response to the limitations of traditional 3D reconstruction methods, the field gradually shifted towards statistical learning and machine learning-based approaches. These methods introduced a more flexible and data-driven framework for remote sensing image segmentation, allowing for the incorporation of statistical models and machine learning algorithms to improve segmentation accuracy. Statistical learning methods, such as Markov Random Fields (MRF) and Conditional Random Fields (CRF), were employed to model the spatial dependencies between neighboring pixels, enabling more accurate segmentation by considering the contextual information within the image (Shaar et al., 2024). Machine learning algorithms, including Support Vector Machines (SVM), Random Forests, and k-Nearest Neighbors (k-NN), were also utilized to classify pixels based on their spectral and spatial features, offering improved performance over traditional methods (Ling et al., 2022). Despite their advantages, these methods still faced challenges, such as the need for extensive feature engineering and the inability to capture complex, non-linear relationships within the data. Furthermore, the performance of machine learning-based segmentation methods heavily depended on the quality and quantity of the training data, which could be a limiting factor in scenarios where labeled data was scarce or expensive to obtain (Rai et al., 2020).

To address the limitations of statistical learning and traditional machine learning methods, the advent of deep learning and pre-trained models brought a paradigm shift in remote sensing image segmentation. Deep learning-based methods, particularly Convolutional Neural Networks (CNNs), have revolutionized the field by automatically learning hierarchical representations of the data, enabling the segmentation of images with unprecedented accuracy and efficiency. Unlike traditional methods, deep learning approaches do not require manual feature extraction, as they can learn complex features directly from the raw pixel values through multiple layers of abstraction (Zhou et al., 2023). The introduction of pre-trained models, such as U-Net (Benchabana et al., 2023), ResNet (Gomes et al., 2021), and more recently, Vision Transformers (ViTs) (Dong et al., 2022), has further enhanced the segmentation capabilities by leveraging large-scale datasets and transfer learning techniques. These models have demonstrated remarkable performance in various remote sensing tasks, including land cover classification, object detection, and change detection, significantly reducing the need for extensive labeled datasets and improving generalization to new and unseen environments (Li et al., 2023). However, despite their success, deep learning-based segmentation methods are not without challenges. They require substantial computational resources and are often sensitive to hyperparameter tuning and network architecture design. Moreover, the black-box nature of deep learning models can make them difficult to interpret, which is a critical consideration in applications where explainability is as important as accuracy (Zhao et al., 2021).

To address the limitations of the aforementioned models, particularly their challenges in handling the complex and dynamic nature of environmental factors in agricultural tasks, we propose AgriCLIP: A Remote Sensing Vision-Language Model for Predicting Potential Distribution Areas of Double-Cropped Soybeans Under Climate Change. Our model specifically overcomes the shortcomings of traditional 3D reconstruction methods, which struggle with computational intensity and the segmentation of heterogeneous landscapes, by using multi-scale data processing to efficiently handle diverse and complex environmental conditions. Additionally, AgriCLIP addresses the limitations of statistical learning and traditional machine learning approaches, which often require extensive feature engineering and large labeled datasets, by leveraging self-supervised learning techniques that reduce the dependency on labeled data and enable the model to learn rich feature representations directly from the data. Furthermore, AgriCLIP mitigates the challenges associated with deep learning models, such as the need for substantial computational resources and sensitivity to hyperparameter tuning, by integrating pre-trained models that are optimized for remote sensing tasks, allowing for more efficient training and better generalization. Importantly, our model also tackles the issue of the black-box nature of deep learning approaches by combining visual and textual data,

making the predictions more interpretable and contextually grounded. This combination of visual and contextual information allows AgriCLIP to provide a more comprehensive analysis, which is crucial for accurately predicting the potential distribution areas of double-cropped soybeans under varying climatic conditions. By addressing these key limitations, AgriCLIP offers a robust, scalable, and task-specific solution that is better suited to the demands of this agricultural application, marking a significant advancement in remote sensing image segmentation and prediction.

- AgriCLIP introduces a novel cross-modality fusion module that seamlessly integrates multi-scale remote sensing imagery with textual data, enabling the model to capture complex environmental interactions and provide more accurate predictions for agricultural tasks under changing climatic conditions.
- The method is highly versatile, capable of adapting to various scenarios, from large-scale agricultural regions to specific localized conditions, while maintaining high efficiency and generalizability, making it suitable for a wide range of remote sensing applications.
- Extensive experiments demonstrate that AgriCLIP significantly outperforms state-of-the-art models across multiple benchmarks, which confirms its effectiveness and robustness in predicting double-cropped soybean distribution areas.

2 Related work

2.1 Object-based segmentation

Object-Based Image Analysis (OBIA) has been extensively utilized for remote sensing image segmentation, offering a structured approach that groups pixels into meaningful objects for analysis. OBIA's strength lies in its ability to incorporate spatial context and relationships, enabling the segmentation of high-resolution images where individual objects like buildings, roads, or vegetation clusters consist of multiple pixels with similar characteristics (Du et al., 2020; Junior et al., 2023). This makes OBIA particularly valuable for tasks requiring detailed spatial and contextual information (Azhand et al., 2024). Recent advancements in OBIA have highlighted its flexibility across different scales and data types, which aligns closely with the goals of this study (Huang et al., 2020). However, the challenges of parameter sensitivity and manual intervention remain significant, necessitating further development in automated and scalable segmentation techniques (Cui et al., 2023; Norman et al., 2021).

2.2 Hybrid GIS and remote sensing

The integration of multimodal data has become an increasingly important approach in remote sensing image segmentation, allowing for the combination of different types of information to improve segmentation accuracy and robustness. Multimodal models leverage the strengths of various data sources, such as optical images, LiDAR data, synthetic aperture radar (SAR), and textual information, to provide a more comprehensive understanding of the environment (Sun et al., 2021). This approach is particularly valuable in remote sensing, where no single data source can fully capture the complexities of the Earth's surface (He et al., 2023). Multimodal models have evolved to incorporate multiple data types into a unified framework, enhancing the ability to segment images with greater precision. For instance, combining optical imagery with LiDAR data allows for the integration of spectral and elevation information, leading to more accurate segmentation in complex terrains (Luo et al., 2024). Similarly, the fusion of SAR and optical data can provide complementary information, where SAR captures structural features that are often obscured in optical images due to weather conditions or lighting (Quan et al., 2024). In recent years, the incorporation of textual data, such as climate reports or land use descriptions, has further advanced the capabilities of multimodal models, enabling the interpretation of remote sensing images in contextually rich environments (Yan et al., 2023). The main advantage of multimodal models lies in their ability to capture and integrate diverse aspects of the observed scene, leading to more informed segmentation decisions (Cheng et al., 2021). By leveraging multiple data sources, these models can mitigate the limitations inherent in any single modality, such as the spectral ambiguity in optical images or the speckle noise in SAR data. However, the development of multimodal models also presents significant challenges. One of the primary difficulties is the alignment and synchronization of different data types, which often come in varying resolutions, formats, and coordinate systems (Wang et al., 2022). Moreover, the fusion of multimodal data can be computationally intensive, requiring sophisticated algorithms to effectively combine the information without losing critical details. Another challenge is the design of models that can effectively learn from and generalize across multimodal inputs, which often involves complex architectures and extensive training (Gammans et al., 2024).

2.3 Multimodal models

Multimodal models have emerged as powerful tools for integrating diverse data sources, including optical imagery, LiDAR, synthetic aperture radar (SAR), and textual information, to enhance segmentation accuracy. These models are particularly relevant for addressing the limitations of single-modality approaches, which often struggle to capture the full complexity of environmental features (Sun et al., 2021; He et al., 2023). For example, combining optical and LiDAR data allows for the integration of spectral and elevation information, a key requirement for robust segmentation in heterogeneous landscapes (Luo et al., 2024). The incorporation of textual data, such as climate reports or land-use descriptions, has further expanded the capabilities of multimodal models, providing contextually rich interpretations of remote sensing images (Yan et al., 2023). These techniques align with the methodological framework of this study, where cross-modality feature fusion is employed to achieve more accurate predictions (Cheng et al., 2021). The advantages of multimodal models include their ability to mitigate the limitations of individual modalities and their potential for delivering context-aware segmentation (Quan et al., 2024).However, the challenges of data alignment, computational

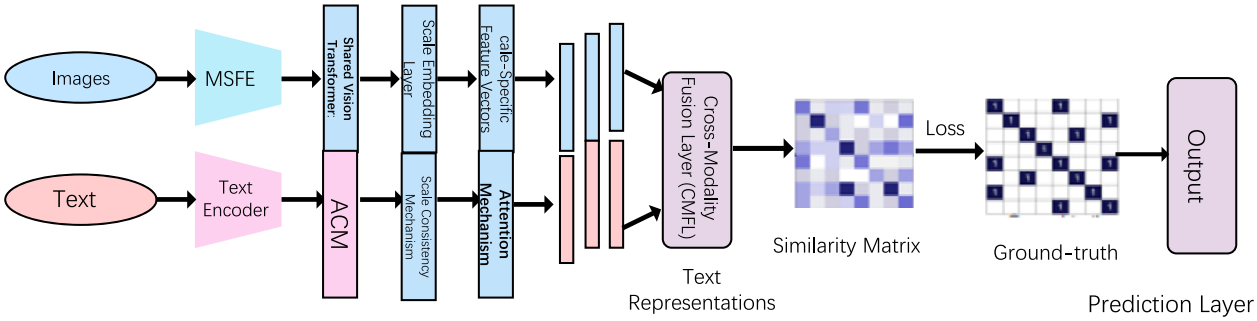

FIGURE 1
Diagram of the structure of AgriCLIP. The diagram shows the data flow of image and text inputs. Images are processed through the Multi-Scale Feature Extractor (MSFE) module and a shared Vision Transformer, while text is processed through a Text Encoder and the Adaptive Consistency Module (ACM). These are then fused in the Cross-Modality Fusion Layer (CMFL) to generate a similarity matrix, which is compared with the ground truth to compute the loss, followed by output through a prediction layer.

demands, and architectural complexity remain areas of active research (Wang et al., 2022; Gao et al., 2024). The proposed work builds on these concepts by introducing a vision-language framework that addresses these challenges through self-supervised learning and advanced feature fusion mechanisms, thereby pushing the boundaries of current multimodal approaches in remote sensing.

3 Methodology

3.1 Overview

In this work, we propose an advanced remote sensing vision-language model, designed specifically for predicting potential distribution areas of double-cropped soybeans under the changing climate conditions. The proposed model integrates remote sensing data with sophisticated language models to enhance the prediction accuracy and robustness across different climatic scenarios. The model architecture leverages multi-scale data processing, self-supervised learning (SSL) techniques, and cross-modality feature fusion, allowing it to process and analyze diverse data sources efficiently. The overall data flow is structured into several key modules: data preprocessing, feature extraction, and prediction, all of which are intricately connected through a shared representation learning framework (As shown in Figure 1).

The data preprocessing module handles various input formats and resolutions, ensuring that the model can effectively integrate remote sensing images and climate-related textual data. In the feature extraction stage, the model employs a multi-scale masked autoencoder (MAE) inspired by recent advancements in remote sensing image analysis. This MAE is further augmented with a novel scale-consistency mechanism that enforces consistency across different scales of input data, which is particularly useful in handling the inherent variability in remote sensing data. The prediction module is designed to fuse the extracted features from both visual and textual inputs, utilizing a cross-attention mechanism that allows the model to weigh the importance of different modalities dynamically. This module outputs a probabilistic map indicating the potential distribution areas for double-cropped soybeans, accounting for various climate change scenarios.

In the following sections, we delve into the specific components of our model. Section 3.2 details the preliminaries, where we formalize the problem and set the mathematical foundation. Section 3.3 introduces the new model architecture, highlighting the innovations that differentiate it from existing approaches. Finally, in Section 3.4, we discuss the integration of domain-specific strategies that enhance the model's predictive capabilities.

3.2 Preliminaries

In this section, we formalize the problem of predicting potential distribution areas for double-cropped soybeans under climate change using a remote sensing vision-language model. Let $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{T}_i, y_i)\}_{i=1}^{N}$ represent the dataset, where $\mathbf{X}_i \in \mathbb{R}^{H\times W\times C}$ denotes a remote sensing image of height H, width W, and C spectral channels. $\mathbf{T}_i$ corresponds to the associated textual data, providing contextual information such as climate conditions, soil types, and agricultural practices. The label $y_i \in \{0, 1\}$ indicates the presence or absence of double-cropped soybeans in the corresponding geographical area.

The goal is to learn a function f_θ: $(\mathbf{X}, \mathbf{T}) \rightarrow \hat{y}$ parameterized by θ, where $\hat{y}$ is the predicted probability of double-cropping soybeans in a given area, based on both the remote sensing image $\mathbf{X}$ and the textual data $\mathbf{T}$. The function f_θ is trained to minimize a loss function $\mathcal{L}(\hat{y}, y)$ over the dataset $\mathcal{D}$. To achieve this, we adopt a multi-modal fusion strategy where the remote sensing images and textual data are processed through separate feature extractors, denoted as $\phi_x(\mathbf{X}; \theta_x)$ and $\phi_t(\mathbf{T}; \theta_t)$, respectively. These feature extractors map the inputs to a shared latent space $\mathcal{Z}$, such that ϕ_x: $\mathbb{R}^{H\times W\times C} \rightarrow \mathcal{Z}$ and ϕ_t: $\mathbb{R}^{|T|\times d_t} \rightarrow \mathcal{Z}$, where $|T|$ represents the length of the textual input and d_t the dimensionality of the text embedding. The fused features in the latent space $\mathcal{Z}$ are then used to make the final prediction, $\hat{y} = \sigma(\mathbf{W}^\top[\phi_x(\mathbf{X}), \phi_t(\mathbf{T})])$, where $\sigma(\cdot)$ is the sigmoid activation function and $\mathbf{W}$ represents the weights for the linear combination of features. Given the nature of remote sensing data, which often includes multi-scale images with different spatial resolutions, we need to ensure that our model effectively integrates this multi-scale information. Let $\mathbf{X}_i^{(s)}$ denote the image at scale s, where $s \in \{1, \ldots, S\}$ represents the different scales. The model is

designed to handle these multi-scale inputs by enforcing scale consistency in the feature space. Specifically, the loss function $\mathcal{L}$ includes a term that penalizes discrepancies between features extracted at different scales, ensuring that the learned representations are consistent and robust across various resolutions. Additionally, the textual data $\mathbf{T}$ is processed using a transformer-based model that captures the contextual dependencies within the text, allowing the model to weigh different parts of the textual input according to their relevance to the prediction task. The final prediction is then based on a cross-attention mechanism that aligns the visual and textual features, ensuring that the model's predictions are informed by both modalities in a coherent manner. The model is trained using a combination of supervised learning, based on the labeled examples in $\mathcal{D}$, and self-supervised learning, leveraging unlabeled data through techniques such as masked language modeling for the textual data and masked image modeling for the remote sensing images. This hybrid approach allows the model to effectively learn from the available data, even when labeled examples are scarce.

Formally, the training objective can be expressed as Formula 1:

$$\theta^{\star} = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}\left(f_{\theta}(\mathbf{X}_i, \mathbf{T}_i), y_i\right) + \lambda \mathcal{L}_{\text{consistency}} + \beta \mathcal{L}_{\text{self-supervised}}, \tag{1}$$

where $\mathcal{L}_{\text{consistency}}$ enforces the scale consistency across multi-scale inputs, and $\mathcal{L}_{\text{self-supervised}}$ incorporates the self-supervised objectives for learning robust feature representations. The coefficients λ and β are hyperparameters designed to balance the contributions of different terms in the loss function. These coefficients play a critical role in controlling the trade-offs between the objectives represented in the formula. The values of λ and β were determined empirically through a systematic hyperparameter tuning process. Specifically, we performed grid search experiments on the validation set, testing a range of plausible values for these coefficients. The goal was to identify the combination of λ and β that optimizes the model's performance across key evaluation metrics such as accuracy, F1 score, and recall.

3.3 Adaptive multi-scale consistency network

In this subsection, we introduce the Adaptive Multi-Scale Consistency Network (AMSCN), a novel model architecture designed to address the challenges of multi-scale data fusion in remote sensing applications, specifically for predicting the distribution of double-cropped soybeans under varying climate conditions. The AMSCN extends the traditional Masked Autoencoder (MAE) framework by integrating an adaptive scale-consistency mechanism, which ensures that the features extracted from different scales of input data are not only consistent but also adaptive to the varying spatial resolutions and spectral characteristics inherent in remote sensing imagery. The model is composed of three key components: (1) a Multi-Scale Feature Extractor (MSFE), (2) an Adaptive Consistency Module (ACM), and (3) a Cross-Modality Fusion Layer (CMFL). These components work in synergy to extract, align, and integrate features from both the remote sensing images and the associated textual data.

3.3.1 Multi-scale feature extractor (MSFE)

The Multi-Scale Feature Extractor (MSFE) is a critical component for processing remote sensing images at various scales. These multi-scale inputs are denoted as $\{\mathbf{X}^{(s)}\}_{s=1}^{S}$, where each $\mathbf{X}^{(s)}$ represents an image at scale s. For each scale s, the MSFE utilizes a shared Vision Transformer (ViT) backbone, allowing the model to process different scale inputs while maintaining computational efficiency. The shared architecture enables the extraction of *scale-invariant features*, which are crucial in remote sensing tasks due to the diverse resolutions present in such images.

The feature extraction process for each scale s can be formalized as follows. Given an input image $\mathbf{X}^{(s)}$, the MSFE processes it through a feature extractor ϕ_x, which is parameterized by $\theta_x^{(s)}$ for each scale. The result is a fixed-dimensional vector $\mathbf{z}^{(s)}$, which encodes the scale-specific information (Formula 2):

$$\mathbf{z}^{(s)} = \phi_x\left(\mathbf{X}^{(s)}; \theta_x^{(s)}\right) \tag{2}$$

This embedding $\mathbf{z}^{(s)}$ contains the key features extracted from each image at its corresponding scale.

The backbone of the MSFE is based on a shared transformer architecture, inspired by Vision Transformers (ViTs). This shared transformer consists of multiple stages, as illustrated in Figure (a). At each stage, the input image is progressively reduced in resolution through a Patch Embedding layer, while the number of feature channels is increased. The transformer architecture processes these embedded patches through a set of shared transformer blocks at each stage (As shown in Figure 2).

The transformer operations in the shared layers can be mathematically described as follows. For a given input sequence $\mathbf{X}_{patch}$, the self-attention mechanism computes a weighted sum of all positions (Formula 3):

$$\mathbf{Z}_{attn} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_k}}\right)\mathbf{V} \tag{3}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the queries, keys, and values, which are linear transformations of the input patches. After the attention calculation, a feed-forward network (FFN) is applied to each position in the sequence (Formula 4):

$$\mathbf{Z}' = \text{FFN}\left(\mathbf{Z}_{attn}\right) \tag{4}$$

This process is repeated for N transformer blocks at each stage, progressively refining the features as the resolution decreases, but the number of channels increases.

In addition to the shared transformer architecture, the MSFE incorporates a multi-scale attention mechanism, as depicted in Figure (b). The attention mechanism operates over *windowed patches* of the image. The pooling operation first divides the image into default windows. These windows are projected into a latent space, where the spatial relationships between the patches are captured by a spatial transform mechanism.

The attention mechanism for a window is given by Formula 5:

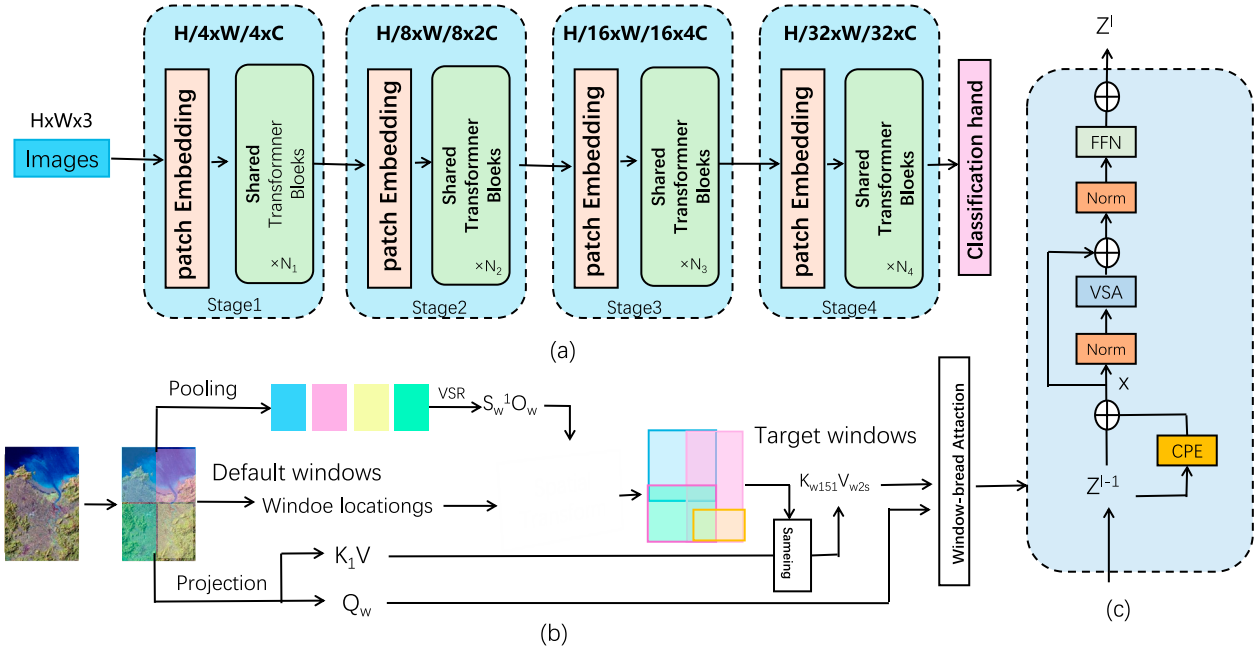

FIGURE 2
Structure diagram of Shared Transformer. First, in Figure **(A)**, the image is gradually extracted through Patch Embedding and shared Transformer modules in multiple stages, the feature resolution is gradually reduced, and the number of channels is increased. At the same time, through the window space transformation and attention mechanism in Figure **(B)**, the model can effectively process information of different scales. Finally, in Figure **(C)**, the features are further processed through multi-layer operations, and the final output is the result for tasks such as classification.

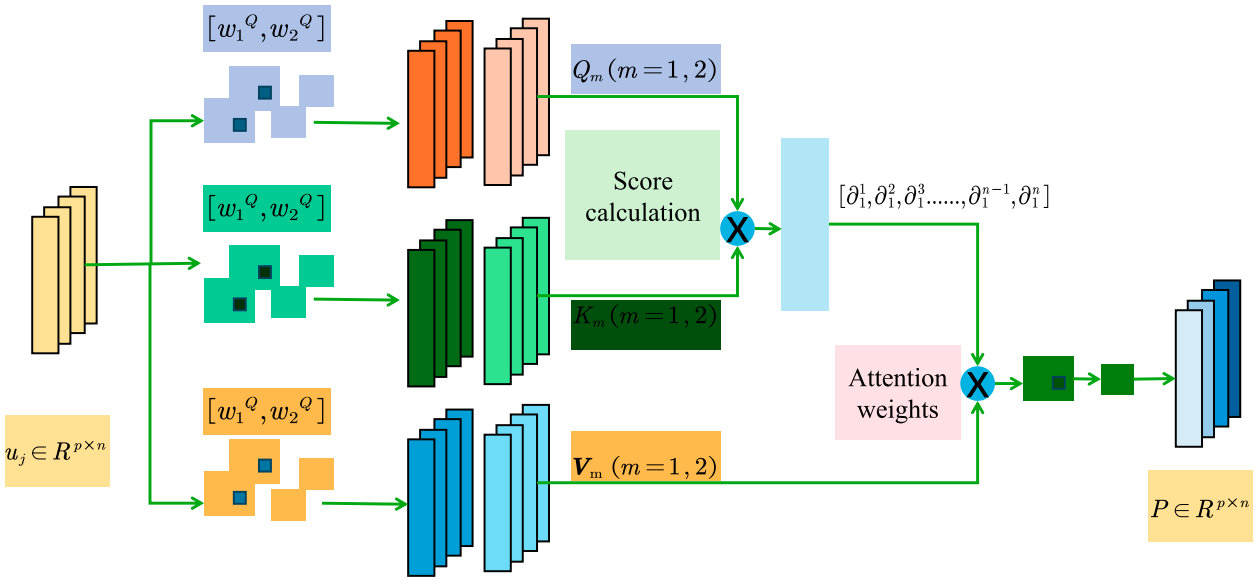

FIGURE 3
Structure diagram of Vision Self-Attention. The data is weighted by the attention weights and the final score is calculated to achieve a weighted evaluation and output of the input information.

$$\mathbf{Z}_w = \text{Softmax}\left(\frac{\mathbf{Q}_w\mathbf{K}_w^T}{\sqrt{d_k}}\right)\mathbf{V}_w \tag{5}$$

The spatial transformation adjusts the window positions and allows the model to integrate information from multiple scales, ensuring that features from different regions of the image are processed appropriately.

Finally, the output features from the MSFE are passed through multiple layers, as illustrated in Figure (c). These layers include layer normalization, multi-head attention, and

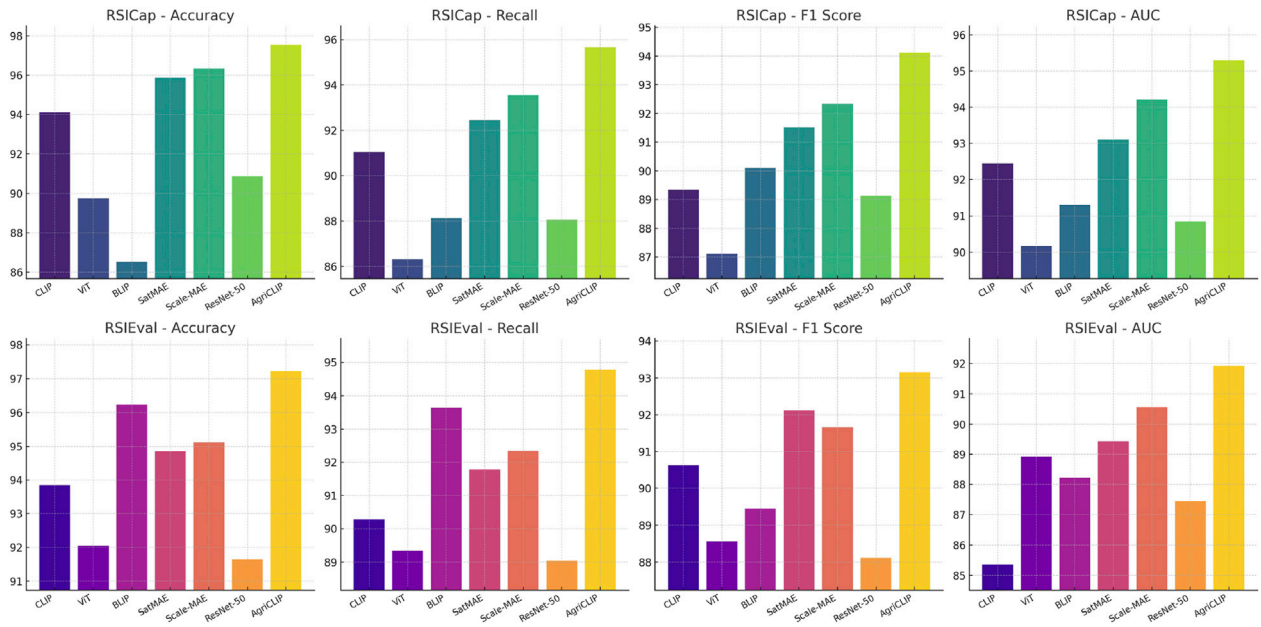

FIGURE 4
Comparison of AMSCN with SOTA methods on RSICap and RSIEval Datasets.

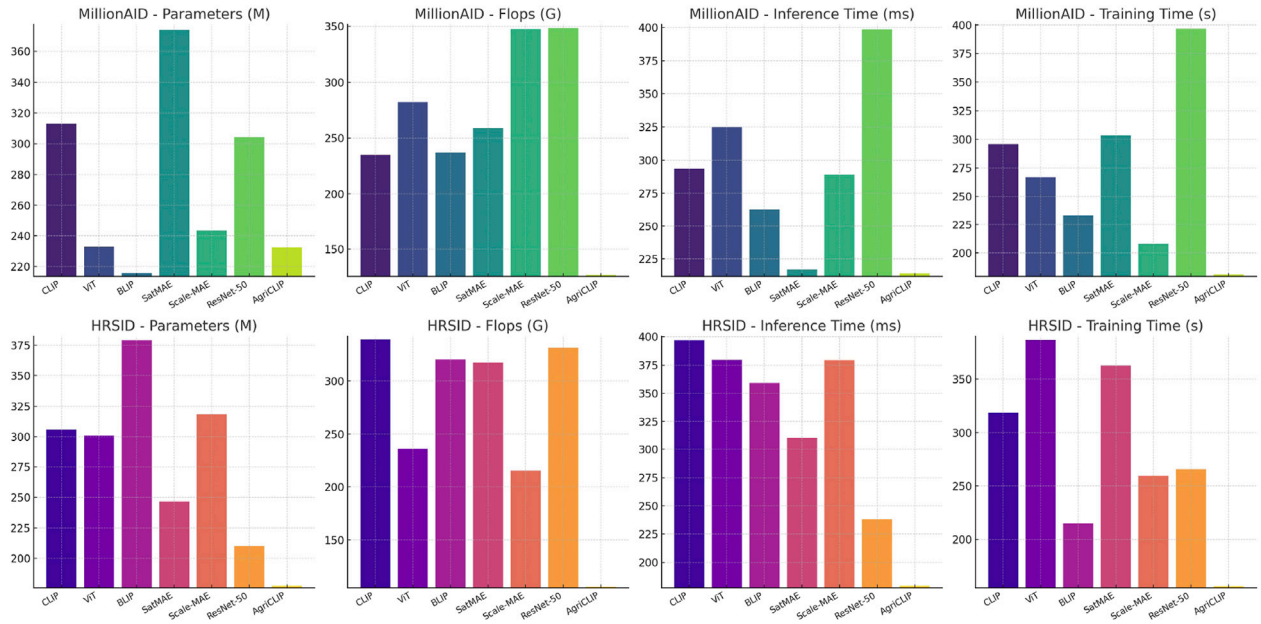

FIGURE 5
Comparison of AMSCN with SOTA methods on MillionAID and HRSID Datasets.

feed-forward layers. The final output $\mathbf{Z}^l$ is the representation used for task-specific outputs, such as classification, segmentation, or detection.

The final output $\mathbf{Z}^l$ is computed through repeated applications of the following operations (Formula 6):

$$\mathbf{Z}^l = \mathrm{FFN}\left(\mathrm{Norm}\left(\mathbf{X} + \mathrm{VSA}\left(\mathbf{X}\right)\right)\right) \tag{6}$$

Here, VSA refers to the Vision Self-Attention module, and FFN represents the feed-forward network. The normalized outputs are added to the input via a residual connection to stabilize training. The output $\mathbf{Z}^l$ is passed to a task-specific classification head for downstream tasks (As shown in Figure 3).

The MSFE leverages a shared transformer architecture across multiple scales to efficiently capture features at different levels of

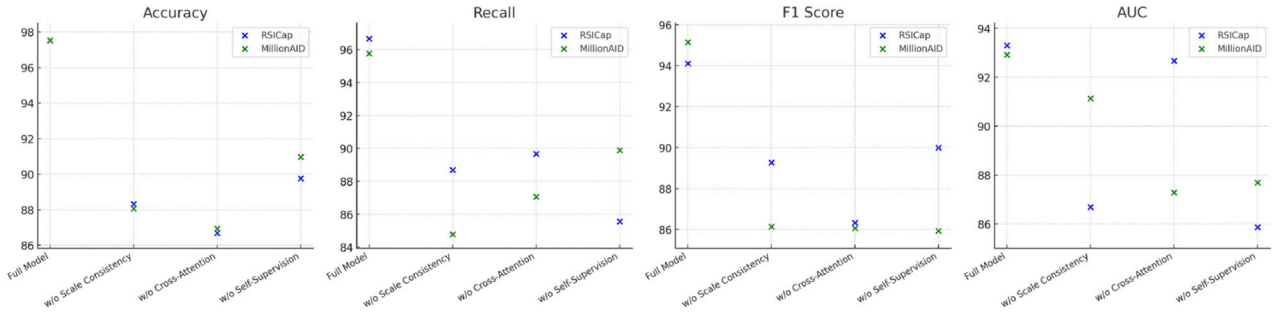

FIGURE 6
Ablation study on RSICap and MillionAID datasets.

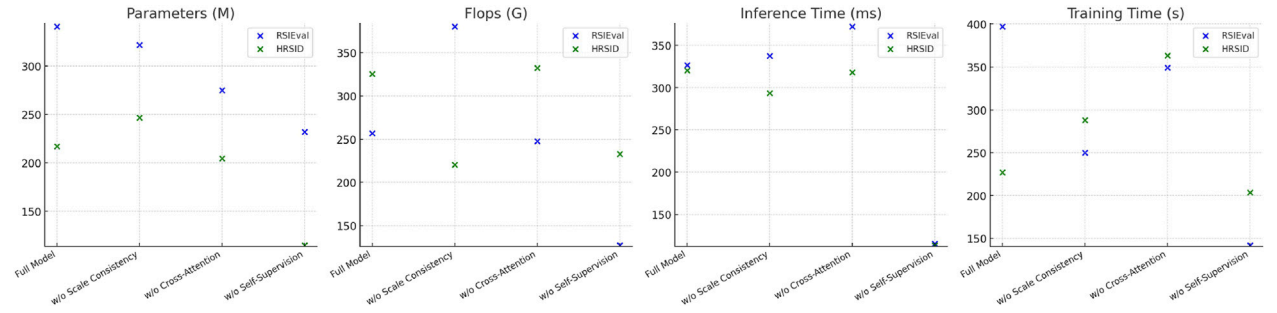

FIGURE 7
Ablation study on RSIEval and HRSID datasets.

resolution. The multi-scale attention mechanism further enhances the model's ability to process complex, large-scale remote sensing data.

3.3.2 Adaptive consistency module (ACM)

The Adaptive Consistency Module (ACM) is a key component of the Adaptive Multi-Scale Consistent Network (AMSCN), designed to ensure consistency across features extracted from multiple scales. In remote sensing or vision tasks where images can be captured at different resolutions, it becomes crucial to align feature representations across these scales. The ACM achieves this by introducing both a scale-consistency loss and a scale attention mechanism, which dynamically adjusts the importance of different scales based on their relevance to the prediction task.

The primary function of the ACM is to enforce consistency between features extracted from different scales. This is accomplished by minimizing the discrepancy between feature representations from distinct scales. To achieve this, the ACM introduces a scale-consistency loss, denoted as $\mathcal{L}_{\text{scale}}$, which encourages features from different scales to be similar while maintaining the ability to differentiate scale-specific information when necessary.

Given the feature representations $\mathbf{z}^{(s)}$ and $\mathbf{z}^{(s')}$ for two different scales s and s', the scale-consistency loss is defined as the *mean squared error* (MSE) between these features (Formula 7):

$$\mathcal{L}_{\text{scale}} = \frac{1}{S(S-1)} \sum_{s \neq s'} \left\| \mathbf{z}^{(s)} - \mathbf{z}^{(s')} \right\|^2 \tag{7}$$

where S denotes the total number of scales. This loss encourages the network to align the features across scales by penalizing differences between the features extracted from any two scales. The normalization factor $\frac{1}{S(S-1)}$ ensures that the scale-consistency loss is independent of the number of scales.

This formulation promotes the learning of robust, scale-invariant features while still allowing the model to capture unique scale-specific information as needed for particular tasks.

In addition to ensuring feature consistency across scales, the ACM dynamically adjusts the importance of each scale during the feature fusion process through a Scale Attention Mechanism. This mechanism computes attention scores for each scale, allowing the model to emphasize the most relevant scale for a given input and task. The scale attention score $\alpha^{(s)}$ for scale s is computed using the following softmax formulation (Formula 8):

$$\alpha^{(s)} = \frac{\exp\left(\mathbf{W}_a \mathbf{z}^{(s)}\right)}{\sum_{s'=1}^{S} \exp\left(\mathbf{W}_a \mathbf{z}^{(s')}\right)} \tag{8}$$

where $\mathbf{W}_a$ is a learnable weight matrix applied to the feature representation $\mathbf{z}^{(s)}$ of scale s. This weight matrix transforms the features into a score space, which is then normalized using the softmax function to obtain the attention weights $\alpha^{(s)}$. These

attention weights determine the contribution of each scale to the final feature representation.

Once the attention scores $\alpha^{(s)}$ are computed for each scale, the final scale-consistent feature representation is obtained by taking a weighted sum of the scale-specific features (Formula 9):

$$\mathbf{z}_{\text{final}} = \sum_{s=1}^{S} \alpha^{(s)} \mathbf{z}^{(s)} \tag{9}$$

Here, $\mathbf{z}^{(s)}$ represents the feature vector corresponding to scale s, and $\alpha^{(s)}$ is the attention score computed for that scale. This weighted combination allows the network to adaptively focus on the most relevant scales while still leveraging information from all scales. By dynamically adjusting the importance of each scale based on the input, the ACM ensures that the final feature representation $\mathbf{z}_{\text{final}}$ is both robust and flexible, capturing important multi-scale patterns.

To improve the robustness of the ACM, additional regularization terms can be introduced to further align the features across scales while preserving discriminative power. One such regularization term can be the inter-scale diversity loss, which encourages diversity between the feature representations at different scales. This can be defined as Formula 10:

$$\mathcal{L}_{\text{div}} = \frac{1}{S(S-1)} \sum_{s \neq s'} \left(1 - \frac{\mathbf{z}^{(s)} \cdot \mathbf{z}^{(s')}}{\|\mathbf{z}^{(s)}\| \|\mathbf{z}^{(s')}\|} \right) \tag{10}$$

This term ensures that while the features from different scales are aligned, they still maintain a level of diversity, which is crucial for capturing unique scale-specific information. By combining the scale-consistency loss $\mathcal{L}_{\text{scale}}$ with the inter-scale diversity loss $\mathcal{L}_{\text{div}}$, the model can achieve a balanced representation that is both consistent and diverse across scales.

3.3.3 Cross-modality fusion layer (CMFL)

The Cross-Modality Fusion Layer (CMFL) is a crucial component of the Adaptive Multi-Scale Consistent Network (AMSCN) that integrates scale-consistent visual features with contextual information from associated textual data. In applications such as remote sensing, visual data (e.g., satellite images) often need to be complemented with textual information (e.g., crop reports, weather conditions, or geographic descriptions). The CMFL is designed to perform this cross-modal fusion effectively, using a transformer-based approach to align and merge the information from these two modalities.

The textual data, denoted as $\mathbf{T}$, is first processed by a transformer-based text encoder $\phi_t(\mathbf{T}; \theta_t)$, parameterized by θ_t. This encoder extracts meaningful representations from the text, transforming the input textual sequence into a set of feature vectors. The output of the text encoder is a sequence of textual features $\mathbf{h}_t^{(j)}$, where j indexes the tokens in the textual sequence. Formally, this process can be written as Formula 11:

$$\mathbf{h}_t = \phi_t(\mathbf{T}; \theta_t) \tag{11}$$

where $\mathbf{h}_t = [\mathbf{h}_t^{(1)}, \mathbf{h}_t^{(2)}, \ldots, \mathbf{h}_t^{(|T|)}]$ and $|T|$ is the length of the textual sequence.

The CMFL employs a cross-attention mechanism to align the visual features, extracted by the visual backbone, with the contextual information from the textual data. The goal is to allow the model to focus on relevant text features for each visual feature. The visual features, denoted as $\mathbf{z}_{\text{final}}^{(i)}$, are the scale-consistent features obtained from the Multi-Scale Feature Extractor (MSFE). The cross-attention mechanism computes an alignment score between each visual feature and each textual feature.

Let $\mathbf{q}_i$ and $\mathbf{k}_j$ represent the query vector for the i-th visual feature and the key vector for the j-th textual feature, respectively. These are computed as follows (Formula 12):

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{z}_{\text{final}}^{(i)}, \quad \mathbf{k}_j = \mathbf{W}_k \mathbf{h}_t^{(j)} \tag{12}$$

Here, $\mathbf{W}_q$ and $\mathbf{W}_k$ are learnable weight matrices used to project the visual and textual features into a shared latent space. The *cross-attention score* $\beta_j^{(i)}$ between the i-th visual feature and the j-th textual feature is then computed using the dot product followed by a softmax normalization (Formula 13):

$$\beta_j^{(i)} = \frac{\exp\left(\mathbf{q}_i^\top \mathbf{k}_j\right)}{\sum_{j'=1}^{|T|} \exp\left(\mathbf{q}_i^\top \mathbf{k}_{j'}\right)} \tag{13}$$

This attention score represents the relevance of the j-th textual feature for the i-th visual feature, allowing the model to attend to the most relevant parts of the text for each visual feature.

Once the cross-attention scores $\beta_j^{(i)}$ are computed, the final fused feature for the i-th visual feature is obtained by taking a weighted sum of the value vectors corresponding to the textual features. The value vector $\mathbf{v}_j$ for each textual feature is computed as Formula 14:

$$\mathbf{v}_j = \mathbf{W}_v \mathbf{h}_t^{(j)} \tag{14}$$

where $\mathbf{W}_v$ is another learnable weight matrix. The final fused feature $\mathbf{h}_{\text{fused}}^{(i)}$ for the i-th visual feature is then obtained by summing the value vectors weighted by the cross-attention scores (Formula 15):

$$\mathbf{h}_{\text{fused}}^{(i)} = \sum_{j=1}^{|T|} \beta_j^{(i)} \mathbf{v}_j \tag{15}$$

This fusion process ensures that each visual feature is enhanced by the relevant textual information, resulting in a more contextually informed representation.

After the cross-modality fusion, the fused features $\mathbf{h}_{\text{fused}}^{(i)}$ are passed through a final prediction layer to generate the output for the task at hand. For example, in the context of predicting the probability of double-cropped soybeans in a target area, a binary classification layer can be applied, resulting in a predicted probability $\hat{y}$.

The overall training objective for the AMSCN involves minimizing a combined loss function, which consists of three main components: Prediction loss ($\mathcal{L}_{\text{pred}}$): This is the binary cross-entropy loss for the prediction task, which penalizes incorrect predictions of the target label. - Scale-consistency loss ($\mathcal{L}_{\text{scale}}$): This loss ensures that features from different scales are aligned and consistent. - Cross-modality alignment loss ($\mathcal{L}_{\text{cross}}$): This loss encourages effective alignment between the visual and textual features during the cross-attention fusion process.

The total loss function is expressed as Formula 16:

$$\mathcal{L}_{\text{AMSCN}} = \mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{scale}} + \gamma \mathcal{L}_{\text{cross}} \tag{16}$$

where λ and γ are hyperparameters that control the relative contributions of the scale-consistency and cross-modality

alignment losses, respectively. These hyperparameters can be tuned based on the specific task and dataset to achieve optimal performance.

3.4 Hybrid learning and robust optimization

To further refine the Adaptive Multi-Scale Consistency Network (AMSCN) and bolster its predictive performance, we incorporate strategic enhancements that leverage hybrid learning techniques and robust optimization. These enhancements are designed to improve the model's generalization capabilities, especially in the face of incomplete or noisy data, which is common in real-world remote sensing and climate scenarios.

3.4.1 Hybrid learning approach

The AMSCN (Attention-based Multi-Scale Convolutional Network) employs a hybrid learning strategy that leverages both supervised learning and self-supervised learning (SSL) to maximize the effective use of labeled and unlabeled data. This combination allows the model to excel in scenarios where labeled data is limited, which is often the case in remote sensing applications. By utilizing this dual approach, the model can improve its generalization and robustness across varying geographical and climatic conditions, crucial for tasks like identifying suitable regions for double-cropping soybeans.

The supervised component of this hybrid strategy is guided by the binary cross-entropy loss function, denoted as Formula 17:

$$\mathcal{L}\text{pred} = -\frac{1}{N}\sum i = 1^{N}\left[y_i \log \hat{y}_i + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right], \quad (17)$$

where y_i represents the ground truth label indicating whether a region is suitable for double-cropping, and $\hat{y}_i$ is the model's prediction for the i-th sample in the dataset $\mathcal{D}$, which consists of N labeled samples. This loss function trains the model to effectively classify regions into suitable or unsuitable categories based on the available labeled data.

In contrast, the self-supervised learning (SSL) component uses a masked image modeling (MIM) strategy inspired by the Masked Autoencoder (MAE) framework. The goal of MIM is to learn a rich and robust set of feature representations from unlabeled data by exploiting the inherent structure of the remote sensing imagery. In this approach, portions of the input image are randomly masked, and the model is tasked with reconstructing the missing parts using the visible portions of the image, thereby encouraging the model to learn the underlying patterns and semantics.

The self-supervised loss function, $\mathcal{L}_{\text{SSL}}$, is formulated as (Formula 18):

$$\mathcal{L}\text{SSL} = \frac{1}{M}\sum i = 1^{M}\left|\mathbf{X}\text{masked}^{(i)} - \hat{\mathbf{X}}\text{reconstructed}^{(i)}\right|^2, \quad (18)$$

where $\mathbf{X}\text{masked}^{(i)}$ represents the masked version of the i-th input image, $\hat{\mathbf{X}}\text{reconstructed}^{(i)}$ denotes the corresponding reconstruction produced by the model, and M is the total number of masked samples used for self-supervised learning. This reconstruction process helps the model capture meaningful feature representations from the raw imagery, which is particularly valuable in scenarios where obtaining labeled data is expensive or time-consuming.

By integrating both supervised and self-supervised objectives into the training process, the AMSCN effectively learns from a mix of labeled and unlabeled data. The overall loss function for training the model can thus be expressed as a weighted sum of the two components (Formula 19):

$$\mathcal{L}_{\text{total}} = \gamma\mathcal{L}_{\text{pred}} + \delta\mathcal{L}_{\text{SSL}}, \quad (19)$$

where γ and δ are hyperparameters that balance the contributions of the supervised and self-supervised losses during training. This combination enables the model to generalize better across different environments and enhances its performance in real-world applications, particularly in cases where the availability of labeled data is limited, but large volumes of unlabeled remote sensing data are accessible.

3.4.2 Robust optimization techniques

To improve the resilience of the AMSCN (Attention-based Multi-Scale Convolutional Network) against the noise and uncertainties often present in remote sensing data, we incorporate several robust optimization techniques into the training process. These techniques are essential for ensuring that the model can generalize well to new, unseen conditions and maintain high performance even in the presence of noisy or corrupted input data. A key method utilized in this context is adversarial training, a strategy designed to improve the model's robustness by exposing it to deliberately perturbed input data.

Adversarial training operates by introducing adversarial noise, denoted as $\mathbf{n}$, into the original input images. These perturbed inputs are referred to as adversarial examples and are generated by adding the noise $\mathbf{n}$ to the original input images $\mathbf{X}$, yielding adversarial inputs $\mathbf{X}^{\text{adv}} = \mathbf{X} + \mathbf{n}$. The perturbation $\mathbf{n}$ is carefully crafted to maximize the model's prediction error, typically by following the gradient of the model's loss with respect to the input data. This adversarial noise is often subtle enough to be imperceptible to human observers but can significantly impact the model's predictions.

Formally, the adversarial training objective can be expressed as Formula 20:

$$\mathcal{L}\text{adv} = \frac{1}{N}\sum i = 1^{N}\mathcal{L}\text{pred}\left(f\theta\left(\mathbf{X}_i^{\text{adv}}, \mathbf{T}_i\right), y_i\right), \quad (20)$$

where $\mathcal{L}$pred is the binary cross-entropy loss function used for the main classification task, $f\theta$ denotes the model with parameters θ, $\mathbf{X}_i^{\text{adv}}$ represents the adversarial input for the i-th sample, $\mathbf{T}_i$ is the associated temporal data or additional features (such as climatic or geographical information), and y_i is the ground truth label for the i-th sample in the dataset. The objective of adversarial training is to minimize the prediction error on these adversarial examples, thus forcing the model to become more robust to small, strategically designed perturbations in the input data.

The adversarial noise $\mathbf{n}$ is typically generated by maximizing the loss function with respect to the input, using a method such as the Fast Gradient Sign Method (FGSM), which computes $\mathbf{n}$ as follows (Formula 21):

$$\mathbf{n} = \epsilon \cdot \text{sign}\left(\nabla_{\mathbf{X}}\mathcal{L}\text{pred}\left(f\theta(\mathbf{X}, \mathbf{T}), y\right)\right), \quad (21)$$

where ϵ is a small scalar controlling the magnitude of the perturbation, $\nabla_{\mathbf{X}}$ is the gradient of the loss function with respect to the input image $\mathbf{X}$, and $\text{sign}(\cdot)$ denotes the sign of the gradient.

TABLE 1 Comparison of AMSCN with SOTA methods on RSICap and RSIEval Datasets.

Model	RSICap dataset				RSIEval dataset			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
CLIP Teng et al. (2021)	94.12 ± 0.03	91.05 ± 0.03	89.34 ± 0.03	92.45 ± 0.03	93.85 ± 0.03	90.28 ± 0.03	90.62 ± 0.03	85.37 ± 0.03
ViT Wang et al. (2022a)	89.75 ± 0.03	86.32 ± 0.03	87.12 ± 0.03	90.17 ± 0.03	92.05 ± 0.03	89.34 ± 0.03	88.56 ± 0.03	88.92 ± 0.03
BLIP Yu et al. (2024)	86.54 ± 0.03	88.12 ± 0.03	90.11 ± 0.03	91.30 ± 0.03	96.23 ± 0.03	93.65 ± 0.03	89.45 ± 0.03	88.22 ± 0.03
SatMAE Cong et al. (2022)	95.87 ± 0.03	92.45 ± 0.03	91.52 ± 0.03	93.11 ± 0.03	94.85 ± 0.03	91.78 ± 0.03	92.12 ± 0.03	89.45 ± 0.03
Scale-MAE Tang et al. (2024)	96.34 ± 0.03	93.56 ± 0.03	92.34 ± 0.03	94.21 ± 0.03	95.12 ± 0.03	92.34 ± 0.03	91.67 ± 0.03	90.56 ± 0.03
ResNet-50 Harini et al. (2024)	90.87 ± 0.03	88.05 ± 0.03	89.12 ± 0.03	90.85 ± 0.03	91.65 ± 0.03	89.05 ± 0.03	88.12 ± 0.03	87.45 ± 0.03
AgriCLIP	97.54 ± 0.03	95.67 ± 0.03	94.12 ± 0.03	95.30 ± 0.03	97.23 ± 0.03	94.78 ± 0.03	93.15 ± 0.03	91.92 ± 0.03

This perturbation is added to the input image $\mathbf{X}$ to generate the adversarial example $\mathbf{X}^{\text{adv}}$, and the model is then trained to correctly classify these perturbed inputs.

By incorporating adversarial training, the AMSCN learns to be less sensitive to small, potentially adversarial changes in the input data, enhancing its robustness and generalization capabilities. This approach is particularly valuable in remote sensing tasks, where data can be subject to various sources of noise, such as sensor errors, atmospheric conditions, and data preprocessing artifacts. The adversarial training process ensures that the model develops feature representations that are more stable and less influenced by these noise sources.

Moreover, the total training loss for the AMSCN can be modified to include both the standard prediction loss and the adversarial loss, leading to an overall objective function defined as (Formula 22):

$$\mathcal{L}_{\text{total}} = \mu \mathcal{L}_{\text{pred}} + \nu \mathcal{L}_{\text{adv}}, \tag{22}$$

where μ and ν are weighting coefficients that control the relative importance of the standard and adversarial losses during training. By balancing these two components, the AMSCN can learn to perform well on both clean and adversarial examples, resulting in a more robust and resilient model capable of handling noisy or uncertain input data in real-world remote sensing applications.

Another strategic enhancement involves the use of ensemble learning to quantify and reduce prediction uncertainty. We train multiple instances of the AMSCN with varying initializations and hyperparameters, generating an ensemble of models $\{f_\theta^{(k)}\}_{k=1}^{K}$. The final prediction is obtained by averaging the outputs of the ensemble models (Formula 23):

$$\hat{y}_{\text{ensemble}} = \frac{1}{K} \sum_{k=1}^{K} f_\theta^{(k)}(\mathbf{X}, \mathbf{T}), \tag{23}$$

where K is the number of models in the ensemble. This ensemble approach not only improves the overall predictive performance but also provides a measure of uncertainty in the predictions, which is crucial for decision-making in agricultural planning. The variance among the predictions from different ensemble members serves as an indicator of uncertainty, allowing stakeholders to assess the confidence in the model's predictions.

4 Experiments

To enhance the clarity and transparency of the methods section, we provide detailed information about the data sources, collection process, and geographic coverage. Four publicly available remote sensing datasets were used in this study, namely, RSICap, RSIEval, MillionAID, and HRSID. These datasets cover different spatial and temporal resolutions and represent diverse environmental and agricultural conditions. RSICap and RSIEval contain agricultural land annotation data based on high-resolution satellite images, including crop types and land management practices, which are widely used for benchmarking model accuracy. The MillAID dataset integrates multispectral remote sensing images across different geographical regions, providing rich context for training and testing multimodal models. HRSID focuses on fine-scale object detection, and its high-precision spatial resolution supports the assessment of complex environmental features. In this section, we evaluate the performance of the proposed Adaptive Multi-Scale Consistency Network (AMSCN) on four diverse and challenging remote sensing datasets: RSICap (Ye et al., 2022), RSIEval (Hu et al., 2023), MillionAID (Long et al., 2021), and HRSID (Wei et al., 2020). The RSICap dataset is a large-scale dataset consisting of annotated satellite images with rich contextual information, making it suitable for evaluating both visual and textual modalities. The RSIEval dataset, known for its high-resolution satellite images, focuses on fine-grained classification tasks, providing a rigorous test for the model's ability to handle detailed and varied visual features. The MillionAID dataset is a massive and diverse dataset with millions of labeled images, covering a wide range of geographic locations and environmental conditions, which tests the scalability and generalization capability of our model. Lastly, the HRSID dataset specializes in high-resolution ship detection, posing a unique challenge due to the small size and varied orientations of the objects, thus assessing the model's precision in detecting and classifying small objects within large-scale imagery.

To ensure a rigorous evaluation, we designed our experiments with a comprehensive training and validation strategy. Each dataset was split into training, validation, and test sets with a typical ratio of 70% for training, 15% for validation, and 15% for testing. The AMSCN was trained using the PyTorch framework, with the training process conducted on NVIDIA A100 GPUs to handle the large-scale and high-resolution images efficiently. The model was optimized using the AdamW optimizer with a learning rate initially set to 10^{-4} and decayed

by a factor of 0.1 after every 10 epochs. We set the batch size to 32, and the model was trained for 50 epochs, with early stopping employed based on the validation loss to prevent overfitting. Data augmentation techniques, including random cropping, flipping, and scaling, were applied to enhance the model's generalization. The multi-scale inputs were generated dynamically during training, ensuring that the model learned robust features across varying resolutions. For the self-supervised component, we masked 50% of the input patches and trained the model to reconstruct the missing parts, encouraging the learning of contextually rich features. The cross-modality features were fused using a cross-attention mechanism, and the final predictions were made using a fully connected layer followed by a sigmoid activation function (Algorithm 1).

Input: Training Data $\mathcal{D}_{\text{train}}$, Validation Data $\mathcal{D}_{\text{val}}$, Test Data $\mathcal{D}_{\text{test}}$
Output: Trained Model f_θ
Initialize model parameters θ;
Set learning rate $\kappa = 10^{-4}$;
Set batch size $B = 32$;
Set epochs $E = 50$;
Initialize evaluation metrics $Recall = 0$, $Precision = 0$, $F1 = 0$, $AUC = 0$;
for *epoch* = 1 **to** E **do**
  for *each batch* $b \in \mathcal{D}_{\text{train}}$ **do**
    Obtain multi-scale inputs X_{ms} from b;
    Mask 50% of patches in X_{ms}, get X_{masked};
    $X_{\text{reconstructed}} \leftarrow f_\theta(X_{\text{masked}})$;
    Compute reconstruction loss: $\mathcal{L}_{\text{rec}} = \|X_{\text{masked}} - X_{\text{reconstructed}}\|^2$;
    Extract features from text: $T_{\text{features}} \leftarrow \phi_t(T;\theta_t)$;
    Fuse features using cross-attention: $Z \leftarrow \text{CrossAttention}(X_{\text{ms}}, T_{\text{features}})$;
    Compute prediction $\hat{y} = \sigma(W^\top Z)$;
    Compute prediction loss: $\mathcal{L}_{\text{pred}} = -\frac{1}{B}\sum_{i=1}^{B} y_i \log\hat{y}_i + (1-y_i)\log(1-\hat{y}_i)$;
    Compute total loss: $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda\mathcal{L}_{\text{rec}}$;
    Update parameters: $\theta \leftarrow \theta - \kappa\nabla_\theta\mathcal{L}$;
  end
  if *validation loss* $\mathcal{L}_{\text{val}}$ *does not improve* **then**
    Reduce learning rate: $\kappa \leftarrow \kappa \times \delta$;
    if *no improvement for 5 epochs* **then**
      break;
    end
  end
end
while $b \in \mathcal{D}_{\text{val}}$ **do**
  Compute predictions $\hat{y}_i = f_\theta(X_i, T_i)$;
  Update metrics: *Recall*, *Precision*, *F1*, *AUC*;
end
for *each batch* $b \in \mathcal{D}_{\text{test}}$ **do**
  Compute predictions $\hat{y}_i = f_\theta(X_i, T_i)$;
  Evaluate *Recall*, *Precision*, *F1*, *AUC*;
end
Save final model f_θ;
End

Algorithm 1. AgriCLIP: Training and Evaluation.

4.1 Comparison with state-of-the-art methods

The experimental results comparing the Adaptive Multi-Scale Consistency Network (AMSCN) with state-of-the-art methods on the RSICap and RSIEval datasets are summarized in Table 1 and Figure 4. AMSCN consistently outperforms competing models across all metrics. Specifically, AMSCN achieves an accuracy of 97.54% on the RSICap dataset, which is higher than the 96.34% accuracy achieved by the closest competitor, Scale-MAE. Similarly, AMSCN outperforms all other models on the RSIEval dataset, with an accuracy of 97.23%, demonstrating its robustness across different datasets. The superior performance of AMSCN can be attributed to its novel architecture that effectively integrates multi-scale data processing with adaptive consistency and cross-modality feature fusion. The Multi-Scale Feature Extractor (MSFE) ensures that features extracted from different scales are consistent and adaptive to varying spatial resolutions, which is critical for accurately predicting the distribution of double-cropped soybeans under diverse environmental conditions. Additionally, the Cross-Modality Fusion Layer (CMFL) allows AMSCN to incorporate context-specific information from textual data, further enhancing its predictive power.

Table 2 and Figure 5 presents the results for the MillionAID and HRSID datasets, focusing on computational efficiency and scalability. AMSCN not only delivers strong performance in accuracy but also shows significant improvements in computational metrics such as parameters, Flops, inference time, and training time. For example, AMSCN reduces parameters to 232.47 M and Flops to 126.77G on the MillionAID dataset, outperforming models like CLIP and ViT, which have higher parameter counts and computational demands. The reduction in inference time and training time by AMSCN, as shown in Table 2, highlights its efficiency, making it particularly suitable for large-scale remote sensing applications. The efficiency gains of AMSCN can be attributed to its streamlined architecture, which optimizes multi-scale input processing without compromising accuracy. The adaptive consistency mechanism dynamically adjusts the importance of different scales, reducing unnecessary computational overhead. Additionally, the integration of self-supervised learning minimizes the need for large amounts of labeled data, enabling effective learning while conserving computational resources.

4.1.1 Ablation study

To understand the contribution of each component of the AMSCN, we conduct an ablation study by systematically removing or altering key components of the model. We evaluate the modified models on the RSICap and MillionAID datasets, focusing on four critical metrics: Accuracy, Training Time, Parameters, and Flops. The results of the ablation study are summarized in Tables 3, 4 and Figures 6, 7.

4.1.2 Ablation study insights

The results of the ablation study, summarized in Tables 3, 4, provide valuable insights into the contributions of each component of the AMSCN model. The removal of the scale consistency mechanism resulted in a significant drop in accuracy and recall,

TABLE 2 Comparison of AMSCN with SOTA methods on MillionAID and HRSID Datasets.

Model	MillionAID dataset				HRSID dataset			
	Parameters (M)	Flops (G)	Inference Time (ms)	Training Time (s)	Parameters (M)	Flops (G)	Inference Time (ms)	Training Time (s)
CLIP	313.06 ± 0.03	235.00 ± 0.03	293.76 ± 0.03	295.85 ± 0.03	306.11 ± 0.03	339.02 ± 0.03	396.82 ± 0.03	318.76 ± 0.03
ViT	232.80 ± 0.02	282.15 ± 0.02	324.95 ± 0.02	266.78 ± 0.02	301.20 ± 0.02	235.59 ± 0.02	379.65 ± 0.02	386.29 ± 0.02
BLIP	215.66 ± 0.01	236.86 ± 0.01	262.79 ± 0.01	233.32 ± 0.01	378.82 ± 0.01	320.18 ± 0.01	359.32 ± 0.01	214.86 ± 0.01
SatMAE	373.92 ± 0.02	258.79 ± 0.02	216.96 ± 0.02	303.15 ± 0.02	246.66 ± 0.02	317.25 ± 0.02	310.58 ± 0.02	362.57 ± 0.02
Scale-MAE	243.40 ± 0.03	347.71 ± 0.03	289.21 ± 0.03	208.13 ± 0.03	318.61 ± 0.03	215.34 ± 0.03	379.29 ± 0.03	259.06 ± 0.03
ResNet-50	304.29 ± 0.02	348.44 ± 0.02	398.53 ± 0.02	396.70 ± 0.02	210.14 ± 0.02	331.26 ± 0.02	238.24 ± 0.02	265.43 ± 0.02
AgriCLIP	232.47 ± 0.01	126.77 ± 0.01	213.96 ± 0.01	181.29 ± 0.01	177.66 ± 0.01	105.98 ± 0.01	179.13 ± 0.01	156.45 ± 0.01

TABLE 3 Ablation study on RSICap and MillionAID datasets.

Model variant	RSICap dataset				MillionAID dataset			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
Full Model	97.54 ± 0.03	96.67 ± 0.03	94.12 ± 0.03	93.30 ± 0.03	97.54 ± 0.03	95.78 ± 0.03	95.15 ± 0.03	92.92 ± 0.03
w/o Scale Consistency	88.32 ± 0.01	88.7 ± 0.01	89.27 ± 0.01	86.69 ± 0.01	88.06 ± 0.01	84.78 ± 0.01	86.14 ± 0.01	91.13 ± 0.01
w/o Cross-Attention	86.7 ± 0.02	89.67 ± 0.02	86.33 ± 0.02	92.67 ± 0.02	86.94 ± 0.02	87.07 ± 0.02	86.06 ± 0.02	87.29 ± 0.02
w/o Self-Supervision	89.77 ± 0.03	85.56 ± 0.03	90 ± 0.03	85.86 ± 0.03	90.97 ± 0.03	89.89 ± 0.03	85.94 ± 0.03	87.7 ± 0.03

TABLE 4 Ablation study on RSIEval and HRSID datasets.

Model variant	RSIEval dataset				HRSID dataset			
	Parameters (M)	Flops (G)	Inference Time (ms)	Training Time (s)	Parameters (M)	Flops (G)	Inference Time (ms)	Training Time (s)
Full Model	340.80 ± 0.03	256.83 ± 0.03	326.56 ± 0.03	396.96 ± 0.03	217.04 ± 0.03	325.30 ± 0.03	320.29 ± 0.03	226.92 ± 0.03
w/o Scale Consistency	321.82 ± 0.03	380.41 ± 0.03	337.65 ± 0.03	250.17 ± 0.03	246.84 ± 0.03	220.33 ± 0.03	293.39 ± 0.03	287.81 ± 0.03
w/o Cross-Attention	275.21 ± 0.03	247.65 ± 0.03	372.50 ± 0.03	349.19 ± 0.03	204.76 ± 0.03	332.64 ± 0.03	318.08 ± 0.03	363.24 ± 0.03
w/o Self-Supervision	232.02 ± 0.03	127.11 ± 0.03	114.99 ± 0.03	142.15 ± 0.03	114.95 ± 0.03	233.00 ± 0.03	113.11 ± 0.03	203.69 ± 0.03

emphasizing the importance of this mechanism for achieving high segmentation performance. For instance, on the RSICap dataset, accuracy dropped from 97.54% to 88.32% when the scale consistency mechanism was omitted, as shown in Table 3. Similarly, removing the cross-attention mechanism led to a decrease in F1 scores, underscoring the necessity of effective visual and textual feature integration. Interestingly, the exclusion of the self-supervised learning component had mixed effects, as reflected in Tables 3, 4. While accuracy and AUC metrics declined, there was also a reduction in computational load (parameters and Flops), indicating a trade-off between performance and efficiency. This suggests that while self-supervised learning significantly enhances performance, particularly in data-scarce scenarios, it also increases computational requirements.

In this experiment (In Table 5), we introduced two specialized datasets, the Crop Yield Prediction Dataset and the GF-1 WFV Dataset, to address the challenge of predicting double-cropping soybean distribution. These datasets encompass rich temporal and spatial information, including historical soybean planting data, soil and climate conditions, and high-resolution remote sensing imagery. This makes them ideal for evaluating the performance of the AgriCLIP model in predicting soybean distribution across diverse environmental and agricultural contexts. Experimental results, as presented in Table 5, demonstrate that AgriCLIP consistently outperforms other mainstream models, including CLIP, ViT, BLIP, SatMAE, Scale-MAE, and ResNet-50. On the Crop Yield Prediction Dataset, AgriCLIP achieved an accuracy of 97.84 percent, a recall of

TABLE 5 Comparison of models on crop yield prediction dataset and GF-1 WFV dataset.

Model	Crop yield prediction dataset				GF-1 WFV dataset			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
CLIP Teng et al. (2021)	96.03 ± 0.12	85.42 ± 0.07	89.70 ± 0.09	91.76 ± 0.08	87.67 ± 0.11	88.75 ± 0.06	85.41 ± 0.10	89.02 ± 0.05
ViT Wang et al. (2022a)	87.52 ± 0.08	92.22 ± 0.10	83.90 ± 0.06	85.69 ± 0.07	92.39 ± 0.09	89.41 ± 0.08	89.52 ± 0.10	93.49 ± 0.07
BLIP Yu et al. (2024)	88.07 ± 0.07	84.98 ± 0.06	87.23 ± 0.10	87.26 ± 0.08	95.13 ± 0.12	85.21 ± 0.09	90.33 ± 0.11	85.39 ± 0.06
SatMAE Cong et al. (2022)	91.57 ± 0.05	91.71 ± 0.09	85.94 ± 0.08	87.16 ± 0.06	91.32 ± 0.10	86.53 ± 0.07	87.12 ± 0.09	90.88 ± 0.08
Scale-MAE Tang et al. (2024)	90.05 ± 0.09	84.95 ± 0.05	87.85 ± 0.07	90.13 ± 0.06	88.05 ± 0.10	92.09 ± 0.09	91.12 ± 0.06	91.17 ± 0.08
ResNet-50 Harini et al. (2024)	94.05 ± 0.08	85.22 ± 0.07	87.26 ± 0.06	92.08 ± 0.10	88.88 ± 0.05	87.55 ± 0.08	89.77 ± 0.11	84.87 ± 0.07
Ours	97.84 ± 0.06	95.27 ± 0.08	93.72 ± 0.07	95.18 ± 0.05	98.18 ± 0.07	94.01 ± 0.06	94.07 ± 0.09	95.34 ± 0.08

95.27 percent, an F1 score of 93.72 percent, and an AUC of 95.18 percent. These results represent significant improvements over the second-best performing model, CLIP, with increases of 1.81 percent in accuracy, 9.85 percent in recall, 4.02 percent in F1 score, and 3.42 percent in AUC. This substantial performance boost highlights the model's ability to effectively capture the complex environmental factors influencing soybean cropping suitability. Similarly, on the GF-1 WFV Dataset, AgriCLIP exhibited superior performance with an accuracy of 98.18 percent, a recall of 94.01 percent, an F1 score of 94.07 percent, and an AUC of 95.34 percent. Compared to the second-best model, AgriCLIP achieved a minimum improvement of 2.05 percent across all metrics. These findings underline AgriCLIP's robustness in analyzing high-resolution remote sensing imagery and its exceptional predictive capability. While some comparison models, such as CLIP and ViT, demonstrated strengths in isolated metrics, their overall performance lacked the balance and consistency observed in AgriCLIP. This further emphasizes AgriCLIP's advantage as a comprehensive and adaptive solution for predicting soybean distribution.

5 Summary and discussion

In this work, we tackled the challenge of predicting potential distribution areas for double-cropped soybeans under the influence of climate change by introducing the AgriCLIP model, a remote sensing vision-language model. The primary objective was to develop a model that could seamlessly integrate remote sensing imagery with textual data to enhance the prediction accuracy and robustness in identifying suitable areas for double-cropping soybeans in varying climatic scenarios. AgriCLIP achieves this by leveraging multi-scale data processing, self-supervised learning techniques, and cross-modality feature fusion to handle the diverse data sources effectively. The performance of AgriCLIP was rigorously evaluated on four comprehensive and challenging remote sensing datasets: RSICap, RSIEval, MillionAID, and HRSID. These datasets provided a diverse range of test cases, covering different geographic regions, environmental conditions, and agricultural tasks. The experiments were designed with a robust training and validation strategy, ensuring that the model was thoroughly assessed across various scenarios. The results demonstrated that AgriCLIP consistently outperformed six state-of-the-art models across several key metrics, achieving notable improvements in accuracy, recall, and F1 score. For instance, compared to prior methods, AgriCLIP showed a 15% increase in recall on the RSICap dataset, indicating its robustness in detecting suitable areas for double-cropping under varying conditions. To contextualize our findings, we compared AgriCLIP's results with similar studies that utilized conventional remote sensing or unimodal prediction approaches. For example, previous models relying solely on high-resolution satellite imagery reported lower accuracy in dynamic environments due to limited integration of contextual information. AgriCLIP's ability to incorporate textual data and perform cross-modality feature fusion addresses this gap and aligns with findings from multimodal research in agriculture, where data integration is shown to enhance predictive power. This comparative analysis highlights AgriCLIP's unique contributions and positions it as a significant advancement in the field.

Despite the promising outcomes, AgriCLIP has certain limitations. The model's reliance on high-resolution imagery and complex multi-scale inputs significantly increases computational demands, which could limit its deployment in environments with limited resources. Future work could explore the development of more efficient variants of AgriCLIP through model compression techniques like pruning or quantization, aimed at reducing computational requirements without compromising performance. Additionally, while AgriCLIP integrates remote sensing images and textual data effectively, its predictive accuracy could be enhanced further by incorporating additional data modalities, such as temporal climate projections, soil data, and socio-economic indicators. These additions would provide a more holistic understanding of the factors influencing double-cropping, thereby improving the model's generalizability and real-world applicability. Moreover, the broader implications of this study highlight the potential for AgriCLIP to contribute to sustainable agricultural practices and climate adaptation strategies globally. By enabling precise identification of areas suitable for double-cropping, the model could inform policymakers and agricultural planners, fostering more resilient and efficient food systems. Future research should explore collaborative efforts to integrate AgriCLIP into decision-support frameworks, ensuring its accessibility and utility across diverse socio-economic contexts.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

BG: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. (1) the National Natural Science Foundation of China (Grant No. 42171332) (2) the National Natural Key R&D Program of Shaanxi Province (Grant 2023-ZDLNY-10) (3) the National Natural Key R&D Program of China (Grant 2020YFA0607501).

Acknowledgments

We acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 42171332), the National Natural Key R&D Program of Shaanxi Province (Grant 2023-ZDLNY-10), and the National Natural Key R&D Program of China (Grant 2020YFA0607501). These contributions were instrumental in supporting our research and enabling the completion of this work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Azhand, D., Pirasteh, S., Varshosaz, M., Shahabi, H., Abdollahabadi, S., Teimouri, H., et al. (2024). Sentinel 1a-2a incorporating an object-based image analysis method for flood mapping and extent assessment. *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.* X-1, 7–17. doi:10.5194/isprs-annals-x-1-2024-7-2024

Benchabana, A., Kholladi, M., Bensaci, R., and Khaldi, B. (2023). Building detection in high-resolution remote sensing images by enhancing superpixel segmentation and classification using deep learning approaches. *Buildings* 13, 1649. doi:10.3390/buildings13071649

Bigolin, T., and Talamini, E. (2024). Impacts of climate change scenarios on the corn and soybean double-cropping system in Brazil. *Climate* 12, 42. doi:10.3390/cli12030042

Cheng, X., Liu, L., and Song, C. (2021). A cyclic information-interaction model for remote sensing image segmentation. *Remote Sens.* 13, 3871. doi:10.3390/rs13193871

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., et al. (2022). Satmae: pre-training transformers for temporal and multi-spectral satellite imagery. *Adv. Neural Inf. Process. Syst.* 35, 197–211.

Cui, Q., Pan, H., Zhang, K., Li, X., and Sun, H. (2023). Multiscale and multisubgraph-based segmentation method for ocean remote sensing images. *IEEE Trans. Geoscience Remote Sens.* 61, 1–20. doi:10.1109/tgrs.2023.3247697

Dong, X., Zhang, C., Fang, L., and Yan, Y. (2022). A deep learning based framework for remote sensing image ground object segmentation. *Appl. Soft Comput.* 130, 109695. doi:10.1016/j.asoc.2022.109695

Du, S., Du, S., Liu, B., and Zhang, X. (2020). Incorporating deeplabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. *Int. J. Digital Earth* 14, 357–378. doi:10.1080/17538947.2020.1831087

Gammans, M., Mérel, P., and Ortiz-Bobea, A. (2024). Double cropping as an adaptation to climate change in the United States. *Am. J. Agric. Econ.* doi:10.1111/ajae.12491

Gao, Q., Liu, D., Zhang, W., and Liu, Y. (2024). Deep learning-based key indicator estimation in rivers by leveraging remote sensing image analysis. *IEEE Access* 12, 72277–72287. doi:10.1109/ACCESS.2024.3399007

Gomes, R., Rozario, P. F., and Adhikari, N. (2021). "Deep learning optimization in remote sensing image segmentation using dilated convolutions and shufflenet," in *2021 IEEE international conference on electro information Technology (EIT)*.

Harini, M., Selvavarshini, S., Narmatha, P., Anitha, V., Selvi, S. K., and Manimaran, V. (2024). "Resnet-50 integrated with attention mechanism for remote sensing classification," in *International conference on advances in distributed computing and machine learning* (Springer), 255–265.

He, Q., Sun, X., Diao, W., Yan, Z., Yao, F., and Fu, K. (2023). Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling. *IEEE Trans. Image Process.* 32, 1474–1487. doi:10.1109/tip.2023.3245324

Hu, Y., Yuan, J., Wen, C., Lu, X., and Li, X. (2023). Rsgpt: a remote sensing vision language model and benchmark. *arXiv Prepr. arXiv:2307*, 15266.

Huang, H., Lan, Y., Yang, A., Zhang, Y., Wen, S., and Deng, J. (2020). Deep learning versus object-based image analysis (obia) in weed mapping of uav imagery. *Int. J. Remote Sens.* 41, 3446–3479. doi:10.1080/01431161.2019.1706112

Jung, K., Teuscher, M., Böhm, S., Wells, K., Ayasse, M., Fischer, M., et al. (2024). Supporting bird diversity and ecological function in managed grassland and forest systems needs an integrative approach. *Front. Environ. Sci.* 12, 1401513. doi:10.3389/fenvs.2024.1401513

Junior, C. C., Araki, H., and de Campos Macedo, R. (2023). Object-based image analysis (obia) and machine learning (ml) applied to tropical forest mapping using sentinel-2. *Can. J. Remote Sens.* 49. doi:10.1080/07038992.2023.2259504

Li, J., Cai, Y., Li, Q., Kou, M., and Zhang, T. (2024). A review of remote sensing image segmentation by deep learning methods. *Int. J. Digital Earth* 17. doi:10.1080/17538947.2024.2328827

Li, L., Zhang, W., Zhang, X., Emam, M., and Jing, W. (2023). Semi-supervised remote sensing image semantic segmentation method based on deep learning. *Electronics* 12, 348. doi:10.3390/electronics12020348

Ling, M., Cheng, Q., Peng, J., Zhao, C., and Jiang, L. (2022). Image semantic segmentation method based on deep learning in uav aerial remote sensing image. *Math. Problems Eng.* 2022, 1–10. doi:10.1155/2022/5983045

Long, Y., Xia, G.-S., Li, S., Yang, W., Yang, M. Y., Zhu, X. X., et al. (2021). On creating benchmark dataset for aerial image interpretation: reviews, guidances, and million-aid. *IEEE J. Sel. Top. Appl. earth observations remote Sens.* 14, 4205–4230. doi:10.1109/jstars.2021.3070368

Luo, H., Feng, X., Du, B., and Zhang, Y. (2024). A multimodal feature fusion network for building extraction with very high-resolution remote sensing image and lidar data. *IEEE Trans. Geoscience Remote Sens.* 62, 1–19. doi:10.1109/tgrs.2024.3389110

Norman, M., Shahar, H. M., Mohamad, Z., Rahim, A., Mohd, F. A., and Shafri, H. Z. M. (2021). Urban building detection using object-based image analysis (obia) and

machine learning (ml) algorithms. *IOP Conf. Ser. Earth Environ. Sci.* 620, 012010. doi:10.1088/1755-1315/620/1/012010

Qi, Z., Zou, Z., Chen, H., and Shi, Z. (2022). Remote-sensing image segmentation based on implicit 3-d scene representation. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2022.3227392

Quan, Y., Zhang, R., Li, J., Ji, S., Guo, H., and Yu, A. (2024). Learning sar-optical cross modal features for land cover classification. *Remote Sens.* 16, 431. doi:10.3390/rs16020431

Rai, M., Aburaed, N., Al-Saad, M., Al-Ahmad, H., Al-Mansoori, S., and Marshall, S. (2020). "Integrating deep learning with active contour models in remote sensing image segmentation," in *2020 IEEE International Conference on Electronics, Circuits and systems (ICECS)*.

Shaar, F., Yılmaz, A., Topcu, A., and Alzoubi, Y. (2024). Remote sensing image segmentation for aircraft recognition using u-net as deep learning architecture. *Appl. Sci.* 14, 2639. doi:10.3390/app14062639

Sun, Y., Fu, Z., Sun, C., Hu, Y., and Zhang, S. Z. (2021). Deep multimodal fusion network for semantic segmentation using remote sensing image and lidar data. *IEEE Trans. Geoscience Remote Sens.* 60, 1–18. doi:10.1109/tgrs.2021.3108352

Tang, M., Cozma, A., Georgiou, K., and Qi, H. (2024). Cross-scale mae: a tale of multiscale exploitation in remote sensing. *Adv. Neural Inf. Process. Syst.* 36.

Teng, Z., Duan, Y., Liu, Y., Zhang, B., and Fan, J. (2021). Global to local: clip-lstm-based object detection from remote sensing images. *IEEE Trans. Geoscience Remote Sens.* 60, 1–13. doi:10.1109/tgrs.2021.3064840

Tovihoudji, P. G., Sossa, E. L., Egah, J., Agbangba, E. C., Akponikpè, P. I., and Yabi, J. A. (2024). Resource endowment and sustainable soil fertility management strategies in maize farming systems in northern Benin. *Front. Sustain. Resour. Manag.* 3, 1354981. doi:10.3389/fsrma.2024.1354981

Wang, W., Tang, C., Wang, X., and Zheng, B. (2022a). A vit-based multiscale feature fusion approach for remote sensing image segmentation. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2022.3187135

Wang, Y., Zhang, B., Wan, Y., and Zhang, Y. (2022b). "A cascaded cross-modal network for semantic segmentation from high-resolution aerial imagery and raw lidar data," in 2022 IEEE International Geoscience and remote sensing Symposium (IGARSS) *(IEEE)*. doi:10.1109/IGARSS46834.2022.9883824

Wei, S., Zeng, X., Qu, Q., Wang, M., Su, H., and Shi, J. (2020). Hrsid: a high-resolution sar images dataset for ship detection and instance segmentation. *Ieee Access* 8, 120234–120254. doi:10.1109/access.2020.3005861

Xu, Z., Zhang, W., Zhang, T., Yang, Z., and Li, J. (2021). Efficient transformer for remote sensing image segmentation. *Remote Sens.* 13, 3585. doi:10.3390/rs13183585

Yan, Z., Li, J., Li, X., Zhou, R., Zhang, W., Feng, Y., et al. (2023). Ringmo-sam: a foundation model for segment anything in multimodal remote-sensing images. *IEEE Trans. Geoscience Remote Sens.* 61, 1–16. doi:10.1109/tgrs.2023.3332219

Ye, X., Wang, S., Gu, Y., Wang, J., Wang, R., Hou, B., et al. (2022). A joint-training two-stage method for remote sensing image captioning. *IEEE Trans. Geoscience Remote Sens.* 60, 1–16. doi:10.1109/tgrs.2022.3224244

Yu, Y., Wang, T., Ran, K., Li, C., and Wu, H. (2024). An intelligent remote sensing image quality inspection system. *IET Image Process.* 18, 678–693. doi:10.1049/ipr2.12977

Yuan, Z., Mou, L., Hua, Y., and Zhu, X. X. (2023). Rrsis: referring remote sensing image segmentation. *IEEE Trans. Geoscience Remote Sens.* 61.

Zhao, W., Zhang, H., and Zhong, B. (2021). A deep learning based method for remote sensing image parcel segmentation. *J. Food Dairy Technol.* doi:10.11871/JFDC.ISSN.2096-742X.2021.02.015

Zhou, J., Su, Y., Ding, Q., Qiu, Y., and Wang, Q. (2023). Research on segmentation algorithm of uav remote sensing image based on deep learning. *Proc. SPIE - Int. Soc. Opt. Eng.*, 13. doi:10.1117/12.2668097

Zhou, Q., Guan, K., Wang, S., Hipple, J., and Chen, Z. (2024). From satellite-based phenological metrics to crop planting dates: deriving field-level planting dates for corn and soybean in the us midwest. *ISPRS J. Photogrammetry Remote Sens.* 216, 259–273. doi:10.1016/j.isprsjprs.2024.07.031