frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 30 July 2024
DOI 10.3389/fenvs.2024.1395337

OPEN ACCESS

EDITED BY
Changchun Huang,
Nanjing Normal University, China

REVIEWED BY
Qiqi Zhu,
China University of Geosciences Wuhan, China
Dongping Ming,
China University of Geosciences, China

*CORRESPONDENCE
Tang Liu,
liut@lreis.ac.cn

RECEIVED 03 March 2024
ACCEPTED 12 July 2024
PUBLISHED 30 July 2024

CITATION
Lu Y, Wang J, Wang D and Liu T (2024), Efficient greenhouse segmentation with visual foundation models: achieving more with fewer samples.
Front. Environ. Sci. 12:1395337.
doi: 10.3389/fenvs.2024.1395337

Efficient greenhouse segmentation with visual foundation models: achieving more with fewer samples

Yuxiang Lu[1,2], Jiahe Wang[1,2], Dan Wang[3] and Tang Liu[1]*

[1]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]Provincial Geomatics Center of Jiangsu, Nanjing, China

Introduction: The Vision Transformer (ViT) model, which leverages self-supervised learning, has shown exceptional performance in natural image segmentation, suggesting its extensive potential in visual tasks. However, its effectiveness diminishes in remote sensing due to the varying perspectives of remote sensing images and unique optical properties of features like the translucency of greenhouses. Additionally, the high cost of training Visual Foundation Models (VFMs) from scratch for specific scenes limits their deployment.

Methods: This study investigates the feasibility of rapidly deploying VFMs on new tasks by using embedding vectors generated by VFMs as prior knowledge to enhance traditional segmentation models' performance. We implemented this approach to improve the accuracy and robustness of segmentation with the same number of trainable parameters. Comparative experiments were conducted to evaluate the efficiency and effectiveness of this method, especially in the context of greenhouse detection and management.

Results: Our findings indicate that the use of embedding vectors facilitates rapid convergence and significantly boosts segmentation accuracy and robustness. Notably, our method achieves or exceeds the performance of traditional segmentation models using only about 40% of the annotated samples. This reduction in the reliance on manual annotation has significant implications for remote sensing applications.

Discussion: The application of VFMs in remote sensing tasks, particularly for greenhouse detection and management, demonstrated enhanced segmentation accuracy and reduced dependence on annotated samples. This method adapts more swiftly to different lighting conditions, enabling more precise monitoring of agricultural resources. Our study underscores the potential of VFMs in remote sensing tasks and opens new avenues for the expansive application of these models in diverse downstream tasks.

KEYWORDS

visual foundation model, remote sensing downstream tasks, greenhouse, deep Learning, remote sensing foundation model

1 Introduction

The advent of the large language foundation model ChatGPT (Brown et al., 2020) has sparked a surge in research on big model technologies and their applications. The launch of GPTs has allowed users to create custom models for specific purposes, significantly expanding the scope and depth of model applications. Some studies have attempted to introduce the same learning paradigm into the field of computer vision, resulting in a series of self-supervised pre-training models (for example, Segment Anything Model (SAM) (Kirillov et al., 2023) and Dinov2 (Oquab et al., 2023), etc). They claim that those models are able to understand image content through contrastive learning (Chen et al., 2020; He et al., 2020) or Masked Image Modeling (MIM) (He et al., 2022; Xie et al., 2022), demonstrating superior performance in certain tasks based on natural images. For instance, SAM has shown its generalization capability on most natural images after being evaluated on various zero-shot tasks across more than 23 datasets.

However, when applied to certain specific remote sensing images or features, their generalization performances decreased significantly (Osco et al., 2023; Yan et al., 2023; Tolan et al., 2024). This situation can be intuitively attributed to the lack of remote sensing samples in the dataset, but the main reason lies in the differences between the features of natural and remote sensing images (Bioucas-Dias et al., 2013). This includes the unique perspective of satellite remote sensing imaging and the multi-scale nesting characteristics of features within image. Besides, features in remote sensing images displayed special optical and texture properties, which also cause poor performance of models trained from natural images. For example, a building has more blurred boundary and less feature in the remote sensing image than in the natural image.

To build a visual model used in the remote sensing field, an natural idea is to train a foundation model from scratch on a massive remote sensing dataset. Jamie Tolan (2024) trained a self-supervised remote sensing image model on 18 million remote sensing image data, which was applied to the estimation of tree heights with good results. Cong et al. (2022) developed a new network called SatMAE, which was trained on 700,000 remote sensing images and surpassed the performance of supervised training methods, achieving state-of-the-art (SOTA) results. Yuan et al. (2024) introduced ChatEarthNet, which includes 170,000 image-description pairs for training text-image foundational models. Guo et al. (2023) trained the largest multimodal remote sensing foundational model on 21.5 million images, setting a new benchmark for SOTA performance in multimodal remote sensing models. Although this approach indeed achieved some significant outcomes, the training process of the foundation model required a large amount of computational power (for example, training DINOV2 requires 96 A100 GPUs). In some current studies, this investment is nearly one-time and cannot serve well for other applications, leading to a certain waste of resources. Moreover, current remote sensing technology is more of an application-oriented method, with remote sensing data sources growing at a PB level daily, it is nearly impossible to traverse all regional remote sensing data. Also, the specific tasks faced by remote sensing are complex and varied, making it difficult to form a unified task description. Seen from this point of view, it limits the development of remote sensing foundation model to some extent. For example, with the same resolution, plastic greenhouses may present various shapes and tones, even combinations of multiple patterns; and at different spatial resolution scales, the same plastic greenhouses may show completely different features (such as the degree of light transmission caused by translucent materials) (Lin et al., 2021; Sun et al., 2021; Zhang et al., 2021; Feng et al., 2022).

As an important but relatively scattered agricultural infrastructure, it is obviously difficult to build a specialized visual model for categories like greenhouses. Ma et al. (2021) labeled a large number of greenhouses at multiple locations for training a dual-task neural network and achieved excellent extraction results. This method involved extensive labeling work. Zhang et al. (2023) used a method of automatically generating labels to train neural networks for greenhouse extraction, and the extraction performance lagged behind supervised learning using manually labeled data. Moreover, these methods require building complex models to achieve better results, and the cost of training these models from scratch is very high. Zhu et al. (2024)proposed a knowledge transfer framework that can use the knowledge of pre-trained models for the training of segmentation models, achieving label-free segmentation. Although this method used a large number of labels to train the pre-trained model, its knowledge transfer approach is highly insightful. With the advent of foundation models, utilizing the rich pre-training knowledge of unsupervised pre-trained foundation models is expected to reduce the dependence of models on the number of samples and lower the cost of training models. However, it is obviously difficult to build a specialized visual model for categories like greenhouses. The greenhouse has unique material, complex spectrum and blurred edge gradient features, which makes the application effects of visual foundation models (VFM) relatively poor. Therefore, how to effectively implement a rapid transfer from a general foundation model has become an inevitable issue in the development of remote sensing foundation models. Keyan et al. (2023) have proposed that foundation model can be seen as a repository for a large amount of general knowledge, which can be transferred and shared across domains to significantly reduce the need for specific task annotation data. In this context, most existing researches based on foundation models rely entirely on the feature encoding ability of the foundation model by constructing simple Head networks at the output end of the foundation model. Although some studies have proven that this method can indeed improve the accuracy and accelerate model convergence on specific tasks (Devlin et al., 2018; Clark et al., 2020; Raffel et al., 2020), they hardly transfer and test it totally on a new class or feature extraction task. In fact, the distribution of features and the observation method of remote sensing images mean that some features might not be explicitly encoded (i.e., self-supervised features are implicit in shallow information or combinations of multiple levels) (Kingma and Welling, 2013), so how to better utilize the encoding features of foundation model and how to transfer and share knowledge to new tasks will be a valuable research question. This idea has made some progress in the medical field (Hu et al., 2023; Mazurowski et al., 2023), but there are currently no good research examples in the remote sensing field.

To better utilize the encoding features of general visual foundation model and achieve rapid deployment and transfer of foundation model, this study proposes a joint inference framework combining basic visual models and lightweight convolutional

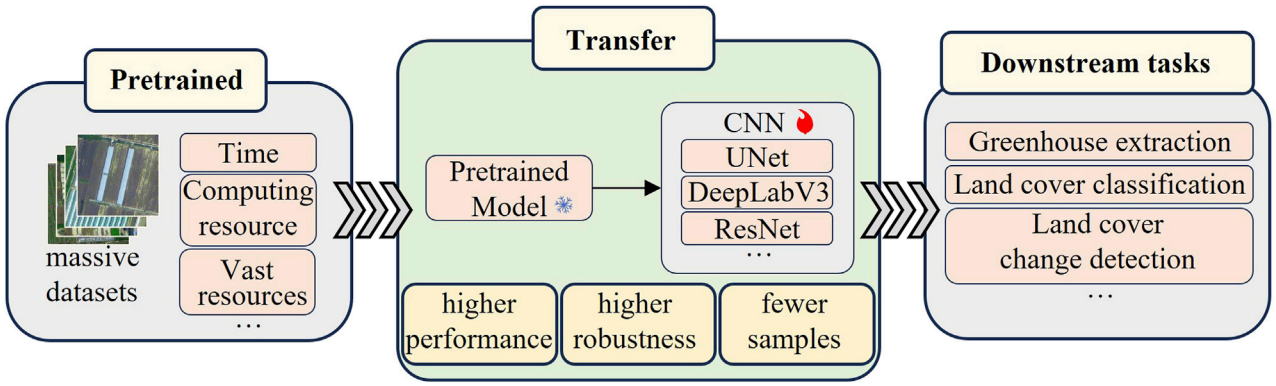

FIGURE 1
The above shows the entire process of utilizing the general knowledge of a foundation model. This study focuses on the transfer process, combining the trained foundation model with a lightweight convolutional network, enabling the model to achieve higher performance and greater robustness with fewer training samples.

models. This framework integrates the general encoding strategies of visual foundation models with geoscientific knowledge-driven model training schemes, focusing on agricultural greenhouse data to drive traditional lightweight convolutional networks with a small sample set, building a bridge between multi-level encoding features and inference results of general visual models. Extensive experiments show that our method can achieve swift transfer of application tasks for general visual models, improve land feature segmentation effects, enhance model generalization capabilities, and reduce model dependence on sample quantity. The remaining sections are organized as follows: Section 2 introduces the overall concept, specific implementation details, datasets, and experimental settings; Section 3 presents the experimental results and reflections prompted by the results, followed by a conclusion and outlook on the study.

2 Materials and methods

2.1 Overview

This study aims to explore how to obtain general knowledge during the model inference phase and transfer it to lightweight networks. As illustrated in Figure 1, the visual foundation model includes two parts: the training phase and the inference phase. During the training phase, researchers perform self-supervised pre-training on the foundation model to prepare for the subsequent inference process. This study focuses on the inference phase, specifically on directly extracting and utilizing the general knowledge from the baseline model during the inference process of applying it to downstream tasks. This general knowledge is integrated into a lightweight network, which is then trained with a few supervised samples to learn to use this knowledge. Here, UNet is used as a lightweight network and is combined with VIT, which is called VFM-UNet. The VFM-UNet network structure is displayed in Figure 2. Additionally, we designed a replaceable prediction head for this network structure, incorporating two mainstream architectures of current deep learning segmentation models: the encoder-decoder structure and the Atrous Spatial Pyramid Pooling (ASPP) (Chen et al., 2017) structure. Figure 3 shows the network with the prediction head replaced by the ASPP structure, which is called VFM-ASPP.

The core design of this study lies in training only the lightweight network to integrate the general knowledge from the foundation model, without training the foundation model weights. The purpose of this design strategy is to reduce the demand for computational resources during the model training process, simultaneously enhance the model performance on specific tasks by utilizing the pre-trained general features in the visual foundation model encoding blocks.

2.2 Knowledge transfer framework

As shown in Figure 2, this study proposes a Knowledge Transfer Framework that integrates a visual foundation model with a lightweight convolutional network. Within this framework, the encoding block employed by the visual foundation model is the Vision Transformer (ViT) (Dosovitskiy et al., 2020). The output features of ViT are combined with those from the encoding block of the convolutional network, followed by an up-sampling of the merged features through a decoding block, ultimately generating the segmented image. Throughout the entire VFM-UNet training process, the encoder of ViT and its outputs are frozen and do not participate in gradient calculations.

Given that ViT requires input shapes to be fixed, remote sensing images of varying sizes cannot be directly fed into ViT. Thus, image preprocessing is necessary to adapt these images to the size requirements of the ViT encoding block. This preprocessing involves resampling the image to match the long side to the required dimensions of the encoding block, followed by zero-padding to shape the image for ViT output. Taking Dinov2 as an example, an image from the training set sized 200 × 400 × 3 is resampled to 224 × 224 × 3. This allows for direct inference using ViT without modifying its structure and weights.

$$L_{BCE} = -\left(y\,log\,\left(p\left(x\right)\right) + \left(1-y\right)log\,\left(1-p\left(x\right)\right)\right) \quad (1)$$

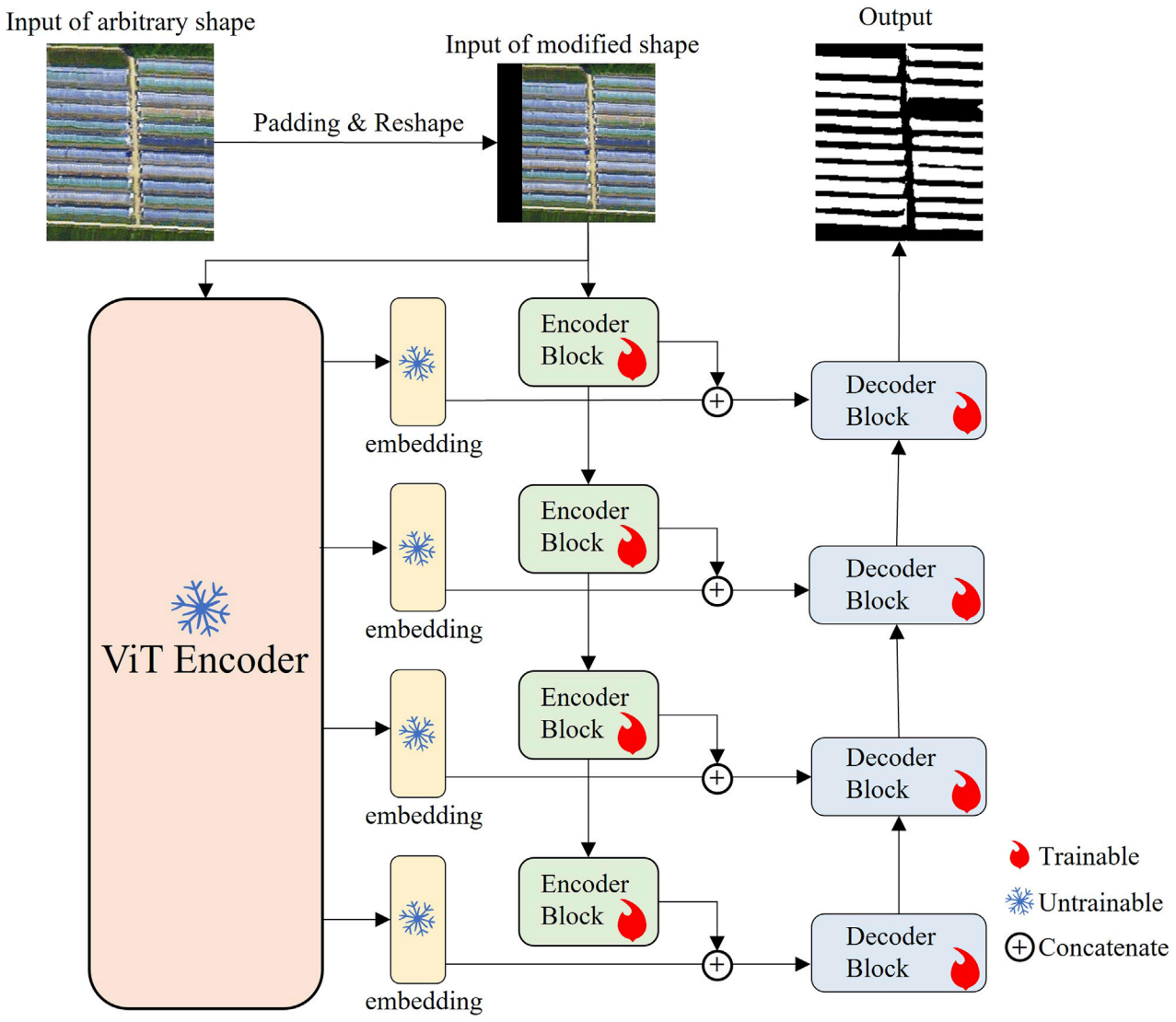

FIGURE 2
Overview of VFM-UNet. The ViT encoder serves as the feature extraction portion of the foundation model, and this component is frozen (untrainable). The self-attention blocks within the ViT encoder produce four embeddings, which are combined with the outputs from four trainable encoder blocks. These combined features are then input into decoder blocks to reconstruct the high-level features into segmentation maps.

$$L_{Dice} = 1 - \frac{2p(x)y}{p(x)^2 + y^2} \tag{2}$$

$$L_{Seg} = L_{BCE} + L_{Dice} \tag{3}$$

where y represents the ground truth values, and $p(x)$ represents the values predicted by the model.

In ViT, to capture both low-level and high-level features of images, four different levels of embeddings are selected, extracted from the outputs of the 0th, 1/3rd, 2/3rd, and the final self-attention block, respectively. As illustrated in Figure 2, these four embeddings are concatenated with the output of the encoder block of the lightweight convolutional network along with the channel dimension. The framework in this study includes four trainable encoder blocks, each reducing the image size by half and doubling the number of channels. The general knowledge integrated into the lightweight convolutional network is then progressively restored to the image size using decoder blocks, with the segmented image resized back to 224 × 224. This segmented image is then resized back to the original image size through a process inverse to the preprocessing.

This study utilizes a combination of Binary Cross-Entropy (BCE) loss and DICE loss as the loss function (see Eq. (3)). BCE loss (see Eq. (1)), appropriated for binary classification tasks, aims to minimize the divergence between model predictions and actual labels by calculating the cross-entropy between them (Jadon, 2020; Tian et al., 2022). DICE loss (see Eq. (2)), designed to assess the similarity between two samples, seeks to improve the detection of positive samples (Milletari et al., 2016; Sudre et al., 2017). In the ground feature extraction task from remote sensing imagery, due to the pronounced imbalance between background and target pixel counts, this loss function combination assists in

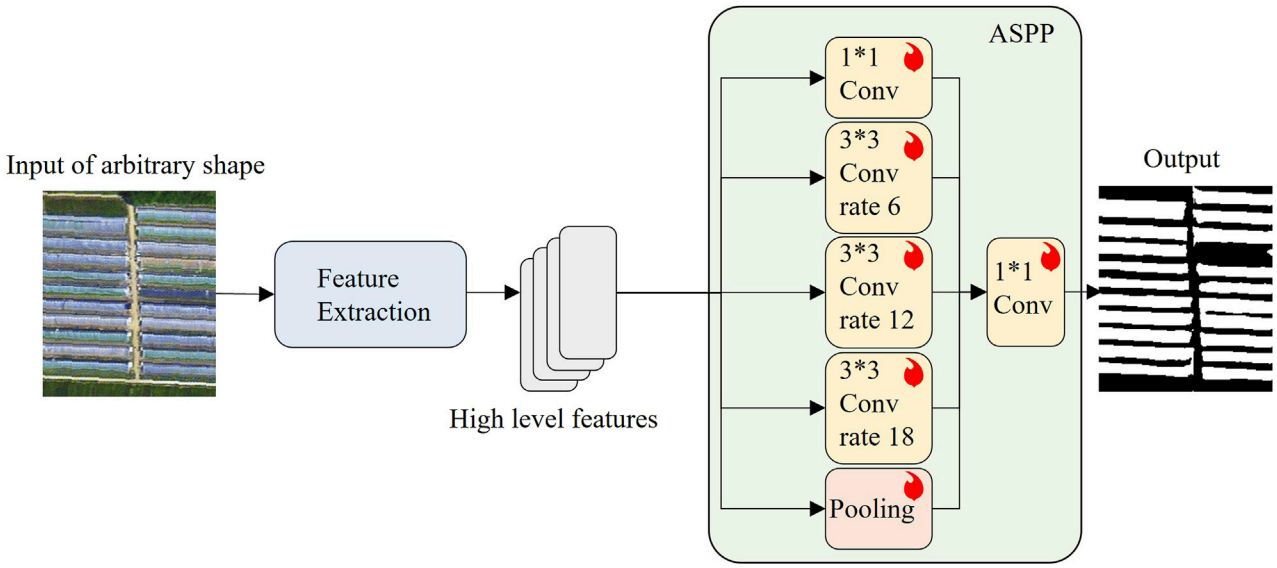

FIGURE 3
Overview of VFM-ASPP. Using the ASPP structure to process the high-level features obtained from the foundation model, the feature extraction here includes ViT and the encoder module in Figure 1. ASPP only requires high-level features and does not need low-level features.

alleviating data imbalance issues and further enhances the details in feature extraction (Wazir and Fraz, 2022).

2.3 Image decoder

After obtaining advanced features of the image, VFM-UNet predominantly uses a decoder to restore image features. Drawing on the classic encoder-decoder structure of UNet (Ronneberger et al., 2015), this study designed four decoder blocks to progressively restore image features. Each block increased the image size twofold while halving the number of channels. Each decoder block received the output from the corresponding encoder block as well as the output from ViT. Since the self-attention blocks in ViT do not alter the size of the features, bilinear interpolation for upsampling is necessary before entering the encoder block to adjust the feature size to meet the requirements of the decoder block.

Due to the use of multiple pooling operations, the encoder-decoder structure in VFM-UNet faces difficulties in extracting features from very small targets. Therefore, as shown in Figure 3, this framework also employ ASPPto process the advanced features acquired by the encoder, which is called VFM-ASPP. ASPP, which is the prediction head of DeepLabv3, uses atrous convolutions with different sampling rates to capture contextual information at various scales during convolution, enhancing the model's ability to acquire information about objects of different sizes. Unlike the decoder structure of UNet, ASPP includes 1 × 1 convolutions, preventing the complete loss of feature information for very small objects. Additionally, most of ASPP's structure is trainable, whereas the decoder's upsampling is based on a fixed algorithm and does not have learnable parameters. The ASPP block only needs to process the advanced features that are a fusion of the output from the fourth encoder block and embeddings. After undergoing multiple sizes of atrous convolution processing, the output features are directly upsampled to the image shape matching the input to ViT, and then restored to the original image size through operations that reverse the preprocessing steps.

2.4 Dataset and experiment

As of 2018, the area of greenhouse construction in Jiangsu Province, China, has reached 339,403.34 ha, ranking first among all provinces in China. In recent years, Jiangsu Province has been actively promoting the intelligent transformation of greenhouse facilities, involving a variety of greenhouse types, including glass, plastic, vegetable, and aquaculture greenhouses, etc.

This study utilized Gaofen-2 satellite imagery, with a panchromatic band spatial resolution of 0.8 m and a multispectral band resolution of 3.2 m. By fusing the panchromatic and multispectral bands, the spatial resolution was enhanced to 0.8 m. The study collected a total of 3,894 images. Using ArcGIS's vector drawing tool, the greenhouses in the images were manually annotated. Subsequently, the vector labels were rasterized through batch processing using code to obtain the corresponding binary image labels. Each image and its corresponding label together constitute a sample. During sample creation, the first step involved ensuring that all images and labels matched in terms of coordinate system, resolution, and size. Any null values generated during sample production were then removed. Finally, all samples were uniformly converted into TIFF format images. As illustrated in Figure 4, these samples are randomly distributed in areas of Jiangsu Province where greenhouses are either densely packed or dispersed. Additionally, some areas without greenhouses were randomly included to introduce noise into the dataset, enhancing the robustness of the trained model. Since the images required

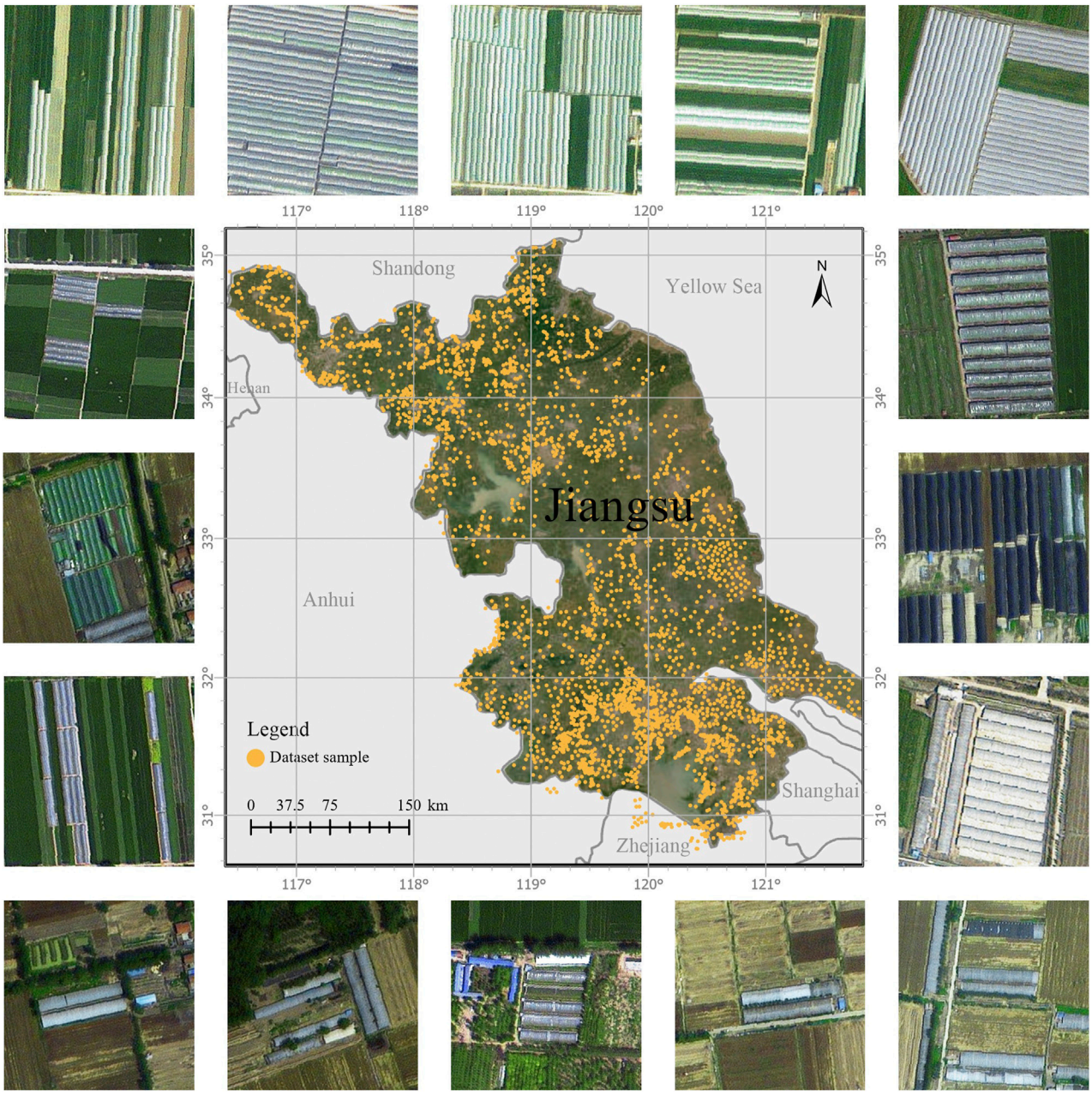

FIGURE 4
Distribution of all samples, along with the showcases of some images in the dataset. The spatial resolution of each image is 0.8 m.

TABLE 1 Comparison of the our method and benchmark network performance.

	IoU	Precision	OA	Recall	F1-score
UNet	0.7727	0.8098	0.8962	0.9436	0.8716
UNet (resnet50)	0.7532	0.7840	0.9068	0.8889	0.8332
VFM-UNet	0.7870	0.8230	0.9045	0.9472	0.8807
Deeplabv3 (resnet50)	0.5604	0.5943	0.7559	0.8533	0.7006
VFM-ASPP	0.6108	0.6669	0.7893	0.8261	0.7380

preprocessing to fit the input requirements of ViT, no uniform shape processing was applied when producing the dataset.

All samples were divided into training, validation, and test sets in a 6:2:2 ratio. To discover the impact of different numbers of samples on model performance, the number of samples in the validation and test sets was fixed at 778, ensuring consistency in evaluation across different training set sizes. By adjusting the number of training set samples, researchers could assess the trend of model performance with different sample quantities.

In this study, experiments were conducted on multiple foundation visual models and several lightweight convolutional networks. To ensure uniformity in evaluation, all models were

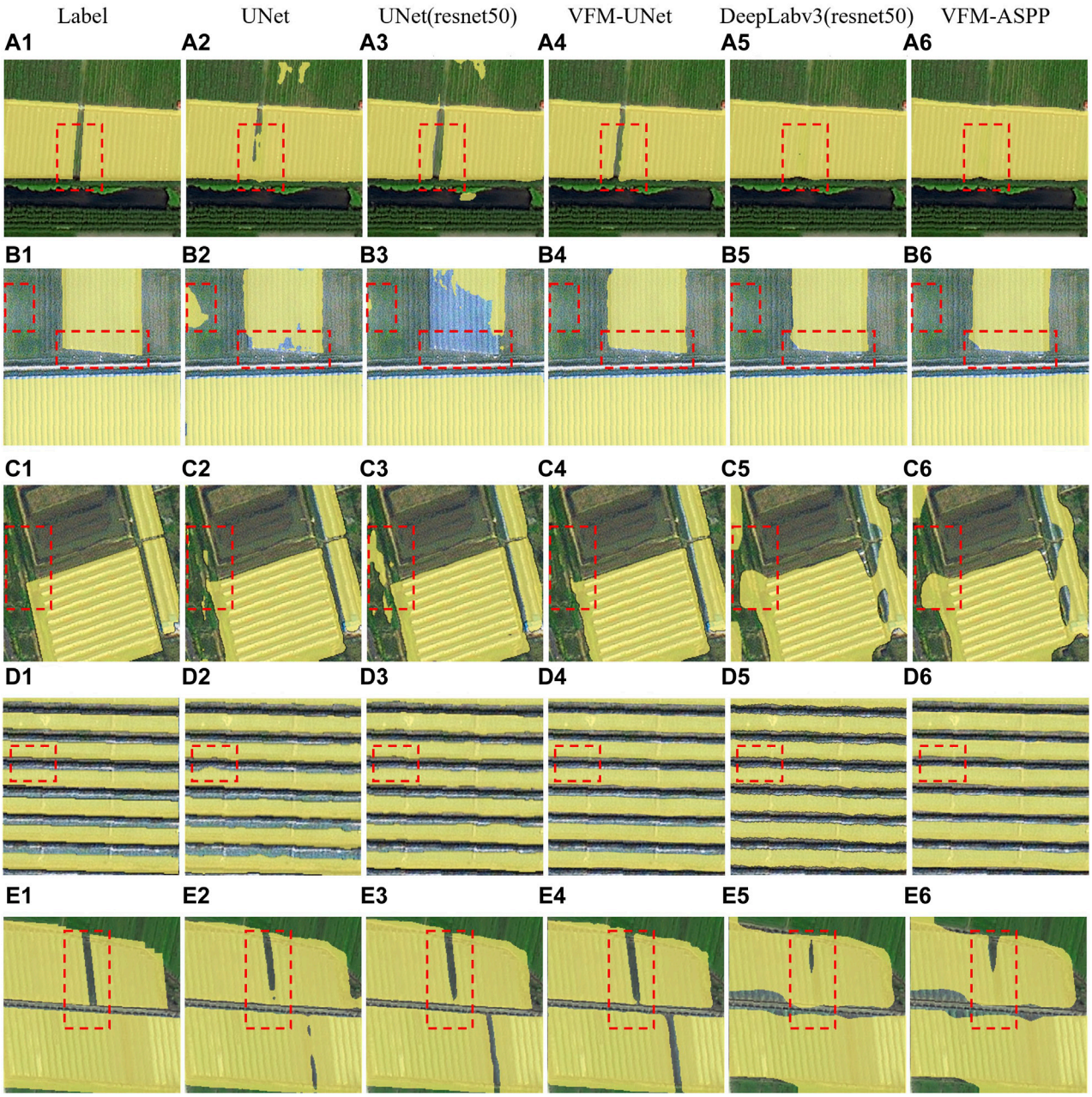

FIGURE 5
Comparing the performance of benchmark and our proposed methods. **(A1–E1)** are the manual annotations of greenhouses. **(A2–E2, A3–E3, A4-E4)** show the comparison of greenhouse extraction results between VFM-UNet and the benchmark. **(A5–E5, A6–E6)** show the comparison of results between VFM-ASPP and the benchmark.

trained with the same parameter settings: a batch size of 14, an initial learning rate set at 2e-3, which decreased to 2e-5 over 200 training epochs using a cosine function. At the same time, both the training and validation sets underwent zero-mean normalization. All models employed a mixed loss function of binary cross-entropy loss (BCE loss) and Dice loss during training and validation phases. All metrics on the test set were calculated based on the model parameters at the end of the 200th training epoch. To evaluate model performance with a limited number of samples, no data augmentation techniques were used during the training of all models.

3 Result and discussion

3.1 The foundation model enhances the network's performance in segmentation tasks

In Table 1, "VFM-UNet" and "VFM-ASPP" represent our models that process high-level features from the foundation model through decoder and ASPP structures, respectively, as showed in Figures 2, 3. Separately, Table 1 presents a

TABLE 2 Comparison of the parameter quantity between our method and UNet with ResNet50.

Model	Million parameters
UNet (resnet50)	43.93
VFM-UNet	42.53

TABLE 3 Comparing the performance of models trained with training sets of different sample quantities.

	IoU	Precision	OA	Recall	F1-score
UNet	0.7727	0.8098	0.8962	0.9436	0.8716
VFM-UNet (100%)	0.787	0.823	0.9045	0.9472	0.8807
VFM-UNet (90%)	0.7786	0.8146	0.8995	0.946	0.8754
VFM-UNet (80%)	0.7841	0.8191	0.9024	0.9481	0.8789
VFM-UNet (70%)	0.7811	0.8177	0.9012	0.9454	0.8769
VFM-UNet (60%)	0.7766	0.8153	0.8988	0.9423	0.8742
VFM-UNet (50%)	0.7715	0.8119	0.8958	0.9393	0.8710
VFM-UNet (40%)	0.7771	0.8158	0.8989	0.9422	0.8745
VFM-UNet (30%)	0.7663	0.8091	0.8933	0.9352	0.8676
VFM-UNet (20%)	0.7637	0.808	0.892	0.9331	0.8661
VFM-UNet (10%)	0.7282	0.7829	0.8715	0.9123	0.8427

comparison of various networks' performance on the training set, including UNet, DeepLabv3, and a variant of UNet with ResNet50 as the encoder. The results demonstrate that after integrating foundation visual models, the method proposed in this study significantly outperforms all benchmark networks across multiple performance metrics, achieving the best results in terms of Intersection over Union (IoU), precision, recall, and F1-score. It slightly lags the UNet utilizing ResNet50 in Overall Accuracy (OA), which may be attributed to the network's incorporation of generic knowledge, leading to a preference for segmenting each category (background and target in this study). This results in a trade-off where the model sacrifices OA for substantial improvements in precision and recall. Furthermore, employing ResNet50 as UNet's encoder and the network structure used for the decoder in this study are similar to using UNet with different backbones. However, due to the parameters of the foundation visual models being frozen during training, only the encoder and decoder are trained. As shown in Table 2, compared to UNet with ResNet50 as the backbone, VFM-UNet not only achieves superior performance but also demands lower computational resources. Additionally, compared to some current Parameter-Efficient Fine-Tuning (PEFT) methods such as knowledge distillation and weight remapping, the approach in this study is structurally simpler. It directly places the smaller model after the feature output of the larger model, without the need to establish a connection between the foundation model and the smaller model.

As shown in Figure 5, the UNet network faces issues of under-segmentation and misclassification in the task of greenhouse segmentation. After integrating generic knowledge from foundation model, VFM-UNet significantly reduces such issues, leading to a notable improvement in segmentation outcomes. Figures 5B,C demonstrate that UNet and UNet with ResNet50 misidentify non-greenhouse objects as greenhouses in the imagery. Figures 5A,D,E show "adhesion" problems where two separate greenhouses located close to each other are incorrectly merged, with the road between them also being recognized as part of a greenhouse. VFM-UNet significantly lowers the likelihood of these two scenarios occurring. Figure 5D shows that in the segmentation of densely packed greenhouses, the boundaries of greenhouses segmented by UNet and UNet (resnet50) might appear distorted, which does not match the actual boundaries of the greenhouses. In contrast, VFM-UNet extracts smoother greenhouse boundaries compared to the benchmark. This might be because the foundation model contains knowledge of greenhouses, allowing it to extract straighter boundaries during feature extraction, resulting in smoother greenhouse boundaries in VFM-UNet's segmentation results.

In the task of extracting greenhouses from remote sensing images, the DeepLabv3 network uses ResNet50 as the feature extractor and combines ASPP to process high-level features. The atrous convolution settings in ASPP make it effective at capturing large-scale object features, but it may fail to capture enough detail when dealing with small, regular structures. As a result, its performance is worse than that of UNet and its variants. However, by integrating the general knowledge of the foundation model in the feature extraction part, this study's framework mitigates these issues to some extent. This additional information helps reduce misclassification and improves segmentation effectiveness (Figure 5A5–E5, A6–E6).

3.2 The foundation model decreases the requirement for the quantity of samples

Table 3 demonstrates the performance of VFM-UNet across training sets with varying sample quantities. During training, we only changed the number of training samples without altering any other hyperparameters. In Table 3, "VFM-UNet (90%)" represents the model trained with 90% of the randomly selected samples from the training set. Similarly, when training the model with different numbers of samples, the proportion of positive samples (those containing greenhouses) and negative samples (those without greenhouses) in the sample set may change. VFM-UNetrequire only about 40% of the training set sample volume to achieve performance levels comparable to those of benchmark networks without the use of visual foundation model on full training samples. This finding significantly substantiates the reduced necessity for training sample quantities in specific remote sensing object extraction tasks within our research framework. In fact, even when the quantity of training samples is reduced to 20%, the

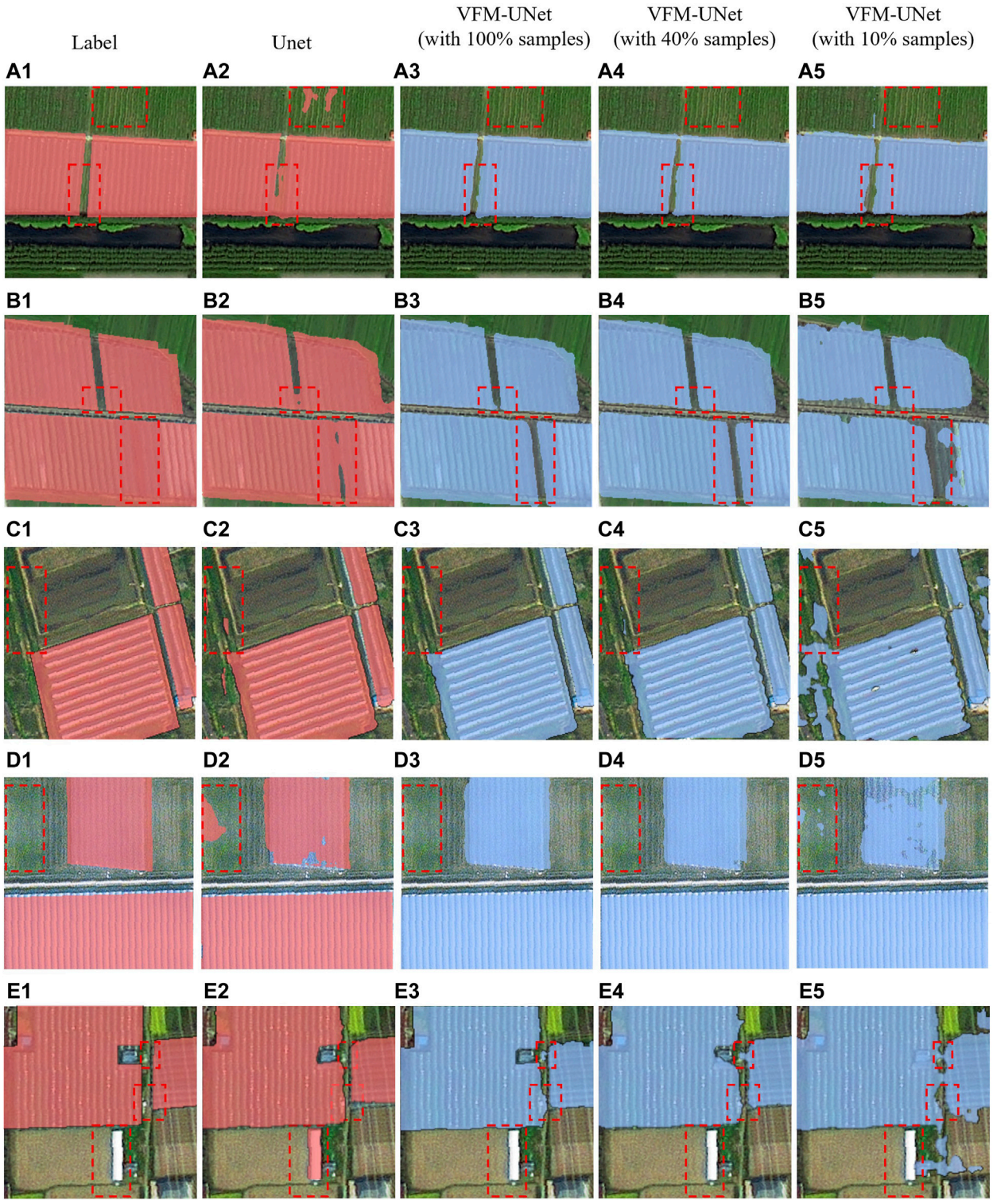

FIGURE 6
The performance comparison between UNet and VFM-UNet on training sets with varying numbers of samples. **(A1–E1)** are the manual annotations of greenhouses. **(A2–E2)** shows the greenhouse segmentation results of UNet, with extracted greenhouses marked in red. **(A3–E3, A4–E4, A5–E5)** show the greenhouse extraction results of VFM-UNet on training sets using different sampling rates, with extracted greenhouses marked in blue.

model's performance remains close to that of the benchmark networks. It is only when the sample quantity is further decreased to 10% that a notable decline in model performance is observed.

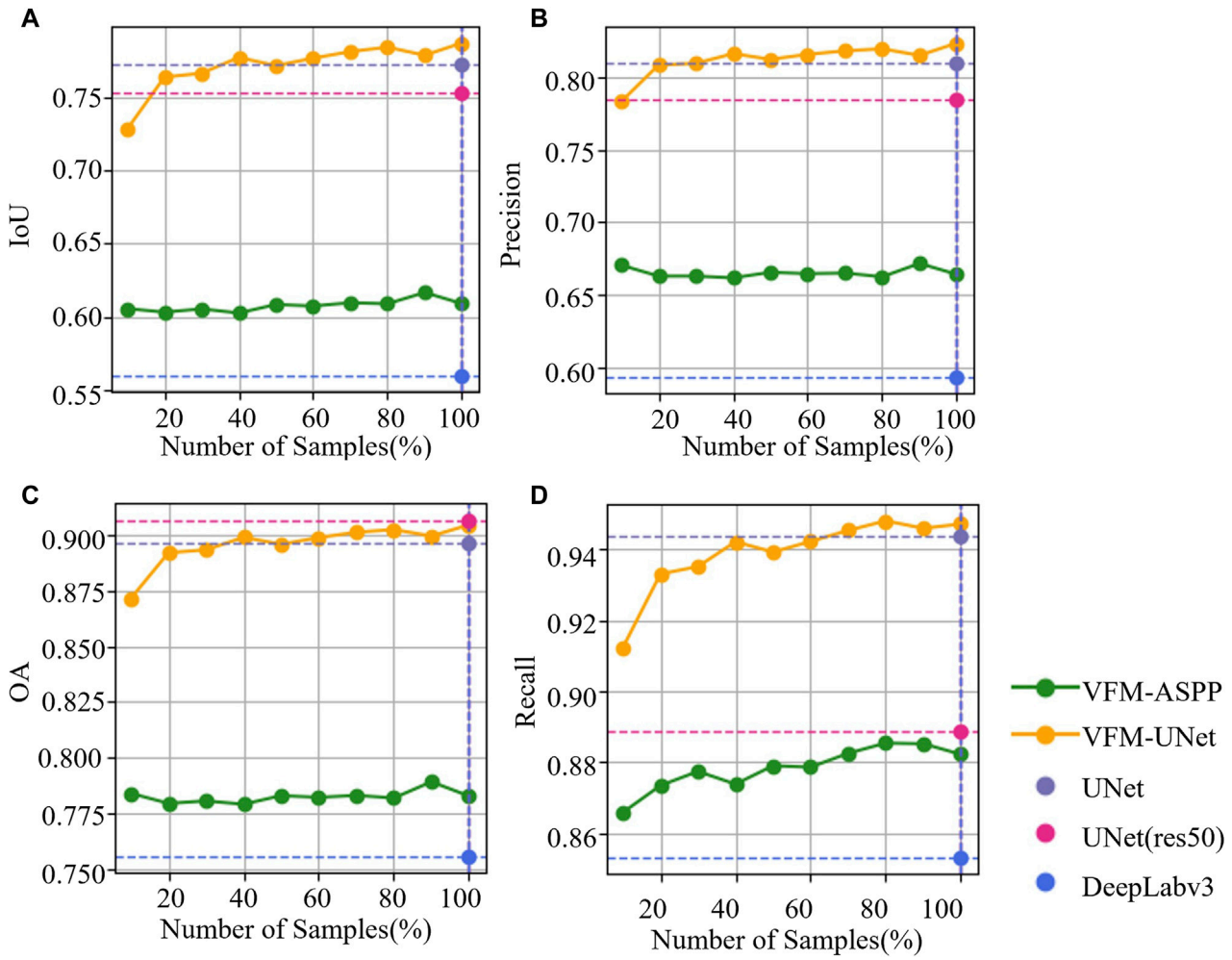

FIGURE 7
The metric comparison between the benchmark and our proposed methods on training sets with different numbers of samples is shown. **(A)**, **(B)**, **(C)**, and **(D)** compare the IoU, Precision, Overall Accuracy (OA), and Recall of these models, respectively.

Figure 6 demonstrates that VFM-UNet significantly reduces misjudgments on unseen data (Figure 6A, C, D, E) compared to the benchmark. After reducing the training data samples, VFM-UNet shows similar metrics to the benchmark (Table 3), but maintains good stability in areas where the UNet method misjudges, avoiding misclassification. It is only when the sample quantity is reduced to 10% that the model exhibits poorer prediction results. Our training data includes negative samples, and when the sample quantity is reduced to 10%, the proportion of negative samples may be too high, leading to a significant decrease in the model's ability to segment targets. Figure 6B shows that our method accurately judges the absence of greenhouses within the red box, whereas the UNet method tends towards identifying them as greenhouses. In reality, the area within the red box should not be considered a greenhouse, indicating a labeling error. VFM-UNet accurately determined the presence or absence of greenhouses at this location, demonstrating the model's accuracy on unseen data. Even in areas prone to misclassification during manual annotation, the model performed well.

UNet and DeepLabv3 represent two mainstream approaches in the field of deep learning for image segmentation. The UNet model employs a typical encoder-decoder structure, initially extracting advanced features of the image through the encoder and then using the decoder to upsample, gradually restoring the image size. In contrast, the DeepLabv3 model is based on the atrous convolution algorithm. It also refines semantic information step by step using an encoder but applies atrous convolution directly after the encoder to expands the receptive field and utilizes a pyramid structure to obtain multi-scale information, eventually upsampling directly to the original image size.

As illustrated in Figure 7, VFM-ASPP proposed in this study outperforms DeepLabv3 in terms of IoU, precision, and overall accuracy (OA). Even with only 10% of the training set's sampled data, the performance of VFM-ASPP surpassed that of DeepLabv3 trained on the entire training set. This significantly reduces the demand for labeled samples when training segmentation models. Figure 7 also shows some significant fluctuations, indicating that the model's performance does not always improve with an increase in the number of training samples. This is due to the presence of negative samples in the

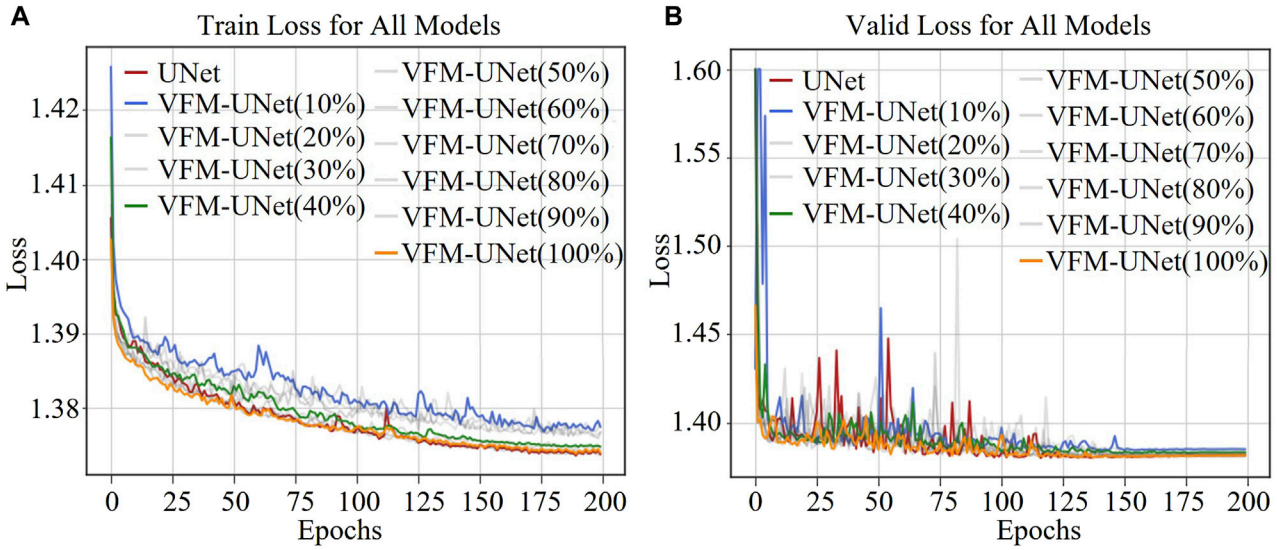

FIGURE 8
Training and validation loss curves for all models over 200 epochs. **(A)** Training loss for U-Net and VFM-U-Net with varying sample sizes (10%–100%). **(B)** Validation loss for the same models and sample sizes. The curves demonstrate how different sample sizes impact the convergence and performance of the models during training and validation.

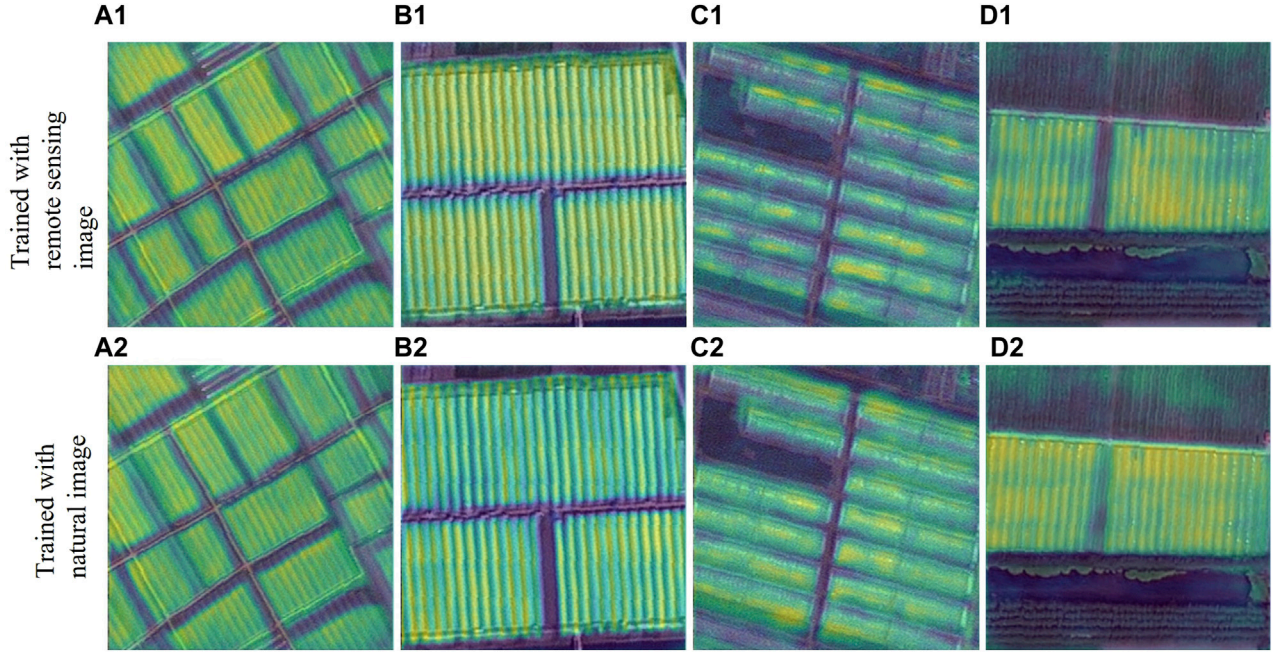

FIGURE 9
Comparison of feature maps generated by models trained with remote sensing images and natural images. Rows **(A1–D1)** display the feature maps from models trained with remote sensing images, while rows **(A2–D2)** show the feature maps from models trained with natural images.

training set (images without greenhouses, labeled entirely as zeros). A change of 10% in the number of samples could result in a higher proportion of negative samples, thereby reducing the model's ability to segment the target effectively. Compared to VFM-UNet, VFM-ASPP exhibits more pronounced performance fluctuations. This may be because VFM-ASPP has far fewer trainable parameters than VFM-UNet, causing the model to approach overfitting when trained on a small amount of data. The significant improvement in VFM-ASPP's recall with the increase in the number of training samples also supports this point. As VFM-ASPP overfits, the model tends to

predict more positive samples during inference, resulting in more false positives (FP), which leads to a decrease in precision and an increase in recall.

Although the improvements of VFM-UNet are more significant compared to VFM-ASPP, the encoder-decoder structure of VFM-UNet, exemplified by UNet, has its limitations, such as the potential loss of small object information due to the extensive use of pooling layers. ASPP was proposed to address this issue. Therefore, one future research direction is to explore how to further enhance the performance of visual foundation model on networks based on atrous convolution, such as DeepLabv3.

3.3 The foundation model enhances the network's generalization capability and stability

Figure 6 and Table 3 both display the model's results on a test set with the same data distribution. When the model is trained with 100% of the samples, it achieves a high accuracy of 90.45% on the test set, demonstrating its strong generalization ability. This indicates that the lightweight convolutional network has learned to utilize the general knowledge from the foundation model. As a result, on unseen data, the convolutional network using the foundation model can leverage general knowledge and high-level features of the imagery to segment specific objects, enhancing its generalization performance.

Figure 8 shows the train loss and validation loss of the model under different sample sizes as epochs progress. The results indicate that even with a very small number of samples, the model converges after 200 training epochs and exhibits excellent performance on the validation set. This demonstrates that the proposed framework effectively captures image features, enabling efficient model training and performance improvement. When using 100% of the data samples, the loss function curve on the validation set (yellow curve) for our method is smoother compared to the benchmark curve (red curve). This implies that VFM-UNet has stronger and more stable generalization capability on unseen data. As the sample size decreases, the network trained with 40% of the samples also shows a relatively smooth loss function curve on the validation set (green curve). In fact, even when the sample size is reduced to 30% and the loss function curve on the validation set begins to show significant fluctuations, it still exhibits less fluctuation compared to the benchmark. This proves the stability of our method when training with a small sample set.

Models integrated with general knowledge exhibit enhanced generalization capabilities and more stable performance, likely because foundation model can inherently capture the advanced features required for segmentation tasks. Figure 9 shows the class activation maps (CAM) generated by the dinov2 model when trained on different datasets, specifically a natural image dataset and a remote sensing imagery dataset (Tolan et al., 2024). In the maps, bright yellow indicates the areas the model focuses on during the task of segmentation, while blue-green signifies areas the model tends not to concentrate on. The baseline model trained on the remote sensing images accurately focuses on the locations of greenhouses in the images. Similarly, the baseline model trained on the natural image dataset also effectively focuses on the distribution of greenhouses, demonstrating that both the natural and remote sensing image-trained baseline models are capable of accurately targeting the correct locations of objects in images. This showcases the strong generalization ability of the baseline models. When a lightweight convolutional network learns to utilize the general knowledge of the baseline model, it can more quickly focus on "the areas it should focus on." This is likely why networks integrating baseline models converge faster and exhibit greater stability.

4 Conclusion

The foundation model is difficult to apply directly to various remote sensing image tasks. PEFT methods such as knowledge distillation and weight remapping require the construction of complex relationships between the foundation model and the smaller model. This study proposes a knowledge transfer framework for visual foundation models, which transfers the general knowledge of the visual foundation model to a lightweight convolutional network, allowing knowledge transfer without the need to construct a complex model structure. Compared to benchmark networks, the network constructed by this research framework significantly enhances accuracy, generalization ability, and the stability of predictions in remote sensing image segmentation tasks, also reducing the dependency on the quantity of training samples. This research not only demonstrates the potential of visual foundation models in knowledge transfer, but also showcases their capability on how to accelerate and optimize the network training process through the use of pretrained models in specific domain applications, especially under data-limited conditions. Furthermore, we discuss potential applications and directions for improving the framework in future work, including enhancing the efficiency and effectiveness of transfer learning and optimizing the framework structure further.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation. All the codes and dataset can be found in the https://github.com/Luyuxiang-Hi/LLM-Knowledge-Transfer.

Author contributions

YL: Conceptualization, Data curation, Methodology, Software, Validation, Writing–original draft. JW: Formal Analysis, Software, Validation, Visualization, Writing–review and editing. DW: Data curation, Resources, Validation, Writing–review and editing. TL: Conceptualization, Funding acquisition, Project administration, Supervision, 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. This study is

supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB 0740200).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bioucas-Dias, J. M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N., and Chanussot, J. (2013). Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience remote Sens. Mag.* 1, 6–36. doi:10.1109/mgrs.2013.2244672

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. neural Inf. Process. Syst.* 33, 1877–1901.

Chen, K., Liu, C., Li, W., Liu, Z., Chen, H., Zhang, H., et al. (2023). Time travelling pixels: bitemporal features integration with foundation model for remote sensing image change detection. arXiv preprint arXiv:2312.16202.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). "A simple framework for contrastive learning of visual representations," in International conference on machine learning, Vienna, Austria, July 12-18, 2020 (PMLR), 1597–1607.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Feng, J., Wang, D., Yang, F., Huang, J., Wang, M., Tao, M., et al. (2022). PODD: a dual-task detection for greenhouse extraction based on deep learning. *Remote Sens.* 14, 5064. doi:10.3390/rs14195064

Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., et al. (2023). Skysense: a multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. arXiv preprint arXiv:2312.10115.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). "Masked autoencoders are scalable vision learners," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA, 18-24 June 2022, 16000–16009.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). "Momentum contrast for unsupervised visual representation learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, 13-19 June 2020, 9729–9738.

Hu, X., Xu, X., and Shi, Y. (2023). How to efficiently adapt large segmentation model (SAM) to medical images. arXiv preprint arXiv:2306.13731.

Jadon, S. (2020). "A survey of loss functions for semantic segmentation," in *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)* (Piscataway, NJ: IEEE), 1–7.

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. arXiv preprint arXiv:2304.02643.

Lin, J., Jin, X., Ren, J., Liu, J., Liang, X., and Zhou, Y. (2021). Rapid mapping of large-scale greenhouse based on integrated learning algorithm and Google Earth engine. *Remote Sens.* 13, 1245. doi:10.3390/rs13071245

Ma, A., Chen, D., Zhong, Y., Zheng, Z., and Zhang, L. (2021). National-scale greenhouse mapping for high spatial resolution remote sensing imagery using a dense object dual-task deep learning framework: a case study of China. *ISPRS J. Photogrammetry Remote Sens.* 181, 279–294. doi:10.1016/j.isprsjprs.2021.08.024

Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N., and Zhang, Y. (2023). Segment anything model for medical image analysis: an experimental study. *Med. Image Anal.* 89, 102918. doi:10.1016/j.media.2023.102918

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in 2016 fourth international conference on 3D vision (3DV), Stanford, CA, USA, 25-28 Oct. 2016 (IEEE), 565–571.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., et al. (2023). Dinov2: learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Osco, L. P., Wu, Q., Lemos, E. L. d., Gonçalves, W. N., Ramos, A. P. M., Li, J., et al. (2023). The Segment Anything Model (SAM) for remote sensing applications: from zero to one shot. *Int. J. Appl. Earth Observation Geoinformation* 124, 103540. doi:10.1016/j.jag.2023.103540

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 5485–5551.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, Part III 18 (Springer), 234–241.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. (2017). "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support: Third international workshop, DLMIA 2017, and 7th international workshop, ML-CDS 2017, held in conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, proceedings 3* (Berlin: Springer), 240–248.

Sun, H., Wang, L., Lin, R., Zhang, Z., and Zhang, B. (2021). Mapping plastic greenhouses with two-temporal sentinel-2 images and 1d-cnn deep learning. *Remote Sens.* 13, 2820. doi:10.3390/rs13142820

Tian, Y., Su, D., Lauria, S., and Liu, X. (2022). Recent advances on loss functions in deep learning for computer vision. *Neurocomputing* 497, 129–158. doi:10.1016/j.neucom.2022.04.127

Tolan, J., Yang, H.-I., Nosarzewski, B., Couairon, G., Vo, H. V., Brandt, J., et al. (2024). Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sens. Environ.* 300, 113888. doi:10.1016/j.rse.2023.113888

Wazir, S., and Fraz, M. M. (2022). "HistoSeg: quick attention with multi-loss function for multi-structure segmentation in digital histology images," in 2022 12th international conference on pattern recognition systems (ICPRS), Saint-Etienne, France, 7-10 June 2022 (IEEE), 1–7.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., et al. (2022). "Simmim: a simple framework for masked image modeling," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA, 18-24 June 2022, 9653–9663.

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

Yuan, Z., Xiong, Z., Mou, L., and Zhu, X. X. (2024). *ChatEarthNet: a global-scale image-text dataset empowering vision-language geo-foundation models.*

Zhang, P., Guo, S., Zhang, W., Lin, C., Xia, Z., Zhang, X., et al. (2023). Pixel–scene–pixel–object sample transferring: a labor-free approach for high-resolution plastic greenhouse mapping. *IEEE Trans. Geoscience Remote Sens.* 61, 1–17. doi:10.1109/tgrs.2023.3257293

Zhang, X., Cheng, B., Chen, J., and Liang, C. (2021). High-resolution boundary refined convolutional neural network for automatic agricultural greenhouses extraction from gaofen-2 satellite imageries. *Remote Sens.* 13, 4237. doi:10.3390/rs13214237

Zhu, Q., Li, Z., Song, T., Yao, L., Guan, Q., and Zhang, L. (2024). Unrestricted region and scale: deep self-supervised building mapping framework across different cities from five continents. *ISPRS J. Photogrammetry Remote Sens.* 209, 344–367. doi:10.1016/j.isprsjprs.2024.01.021