remote sensing

MDPI

Article

SAM-GFNet: Generalized Feature Fusion with Hierarchical Network for Hyperspectral Image Semantic Segmentation

Yinggang Guo *, Xu Liu, Jinhong Rong, Fengmiao Shi and Chao Tang

Northwest Institute of Nuclear Technology, Xi'an 710024, China; lxfeng_2005@163.com (X.L.); 18196440470@163.com (J.R.); 13139725795@163.com (F.S.); tangchao10@163.com (C.T.)
* Correspondence: fariel_gyg@163.com

Highlights

What are the main findings?

- The proposed SAM-GFNet integrates SAM-derived generalized visual features with hierarchical spatial–spectral characteristics for hyperspectral image semantic segmentation.
- We developed an efficient feature fusion module that adaptively integrates encoded features with decoder-generated characteristics through attention mechanisms.

What are the implications of the main findings?

- The hierarchical fusion of visual foundation model features with task-specific spectral–spatial characteristics significantly improves land cover classification accuracy.
- The approach demonstrates superior performance on large-scale hyperspectral datasets, bridging the domain gap between natural imagery and remote sensing data.

Abstract

Visual foundation models exhibit significant potential for extracting generalized characteristics from imagery, which is crucial for remote sensing image interpretation tasks. However, current approaches underutilize these capabilities for hyperspectral image semantic segmentation. To bridge this gap, this study establishes SAM-derived generalized feature fusion with a hierarchical network for hyperspectral image semantic segmentation (SAM-GFNet), which integrates generalized features with hierarchical task-specific spatial–spectral characteristics for the segmentation task. Specifically, hyperspectral images are partitioned into multiple spectral groups, and SAM-induced generalized features are extracted within each group's corresponding spectral range. To leverage multi-scale contextual information, the hierarchical hyperspectral image semantic segmentation architecture fuses multi-scale spectral–spatial features with SAM-induced generalized visual characteristics through a hierarchical fusion pattern. Furthermore, an efficient feature fusion module is further developed to adaptively integrate encoded features with decoder-generated characteristics, further enhancing semantic segmentation performance. The proposed semantic segmentation scheme is evaluated on WHU-OHS and WHU-H2SR benchmark datasets, achieving overall accuracies of 79.60% and 86.92% on two datasets, surpassing representative baseline methods by over 2.94% and 1.56%, respectively.

Keywords: hyperspectral image; semantic segmentation; visual generalized features; hierarchical network; feature fusion

check for updates

Academic Editor: Akira Iwasaki

Received: 13 October 2025
Revised: 4 November 2025
Accepted: 5 November 2025
Published: 7 November 2025

Citation: Guo, Y.; Liu, X.; Rong, J.; Shi, F.; Tang, C. SAM-GFNet: Generalized Feature Fusion with Hierarchical Network for Hyperspectral Image Semantic Segmentation. *Remote Sens.* **2025**, *17*, 3662. https://doi.org/10.3390/rs17223662

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The usage of remote sensing (RS) technology for Earth observation is an essential means for human understanding of the Earth's environment. As an important remote sensing tool, the imaging spectrometer captures scene reflectance information within the visible to near-infrared spectral range, and the resulting hyperspectral image (HSI) contains hundreds of diagnostic spectral bands, which provides significant advantages for precisely identifying ground object attributes. Therefore, it is a promising paradigm to utilize rich spectral and spatial information contained in hyperspectral data for land cover classification, which plays a crucial role in many application fields such as cartography [1], mineral exploration [2] and precision agriculture [3].

Thanks to rich spectral information contained in hyperspectral data, pixel-wise classification approaches are the prevailing pattern for land cover classification. The spectral-only classification approaches leverage the spectral signature of individual pixels, and typical methods include support vector machine (SVM) [4], one-dimensional convolutional neural network (1D-CNN) [5] and deep recurrent neural network (RNN) [6]. Because local spatial features surrounding the target pixel can significantly enhance classification performance, this leads to extensive research into patch-based approaches for hyperspectral image classification. In order to simultaneously utilize local spatial features and spectral information for classification, spatial feature extractors such as morphological attribute profiles [7] and extinction profiles [8] are employed to capture structural attributes, while two-dimensional convolutional neural networks (2D-CNN) directly leverage spatial characteristics for end-to-end classification [9]. Several studies directly process local hyperspectral cubes as input, employing three-dimensional convolutional neural networks (3D-CNN) [10], capsule networks (CapsNets) [11], transformers [12] and Mamba [13] for supervised classification.

Nevertheless, patch-based approaches essentially constitute a pixel-by-pixel classification paradigm, which leverages spatial information solely within individual patches, neglecting global spatial relationships contained in hyperspectral data. Patch-based classification methods require identifying optimal patch sizes, which significantly increases computational complexity, which is also vulnerable to data leakage between training and test sets. The end-to-end semantic segmentation of hyperspectral imagery has emerged as a highly promising research direction in land cover classification, demonstrating substantial advances in recent years. This classification paradigm leverages both global spatial context and spectral information while maintaining computational efficiency. Early hyperspectral segmentation approaches directly adopt computer vision-based semantic segmentation models. The fast patch-free global learning (FPGA) framework employs an encoder–decoder-based fully convolutional network (FCN) to capture global spatial contexts [14]. Subsequent work introduces spatial–spectral FCNs for end-to-end HSI classification, augmented with dense conditional random fields (CRFs) to optimize local–global information balance [15]. A scale attention-based fully contextual network adaptively aggregates multilevel features, complemented by a pyramid contextual module for comprehensive spatial–spectral context extraction [16]. In [17], a cross-domain classification framework is proposed to synergistically integrate multiple pre-trained models, achieving superior target-domain accuracy through adaptive feature fusion.

Transformers like ViT could effectively capture global spatial dependencies, but they demand substantially more labeled data and computational resources than CNN-based approaches. To expand the receptive field, the multilevel spectral–spatial transformer is investigated for HSI land cover classification [18]. A 3D Swin transformer-based hierarchical approach is proposed to effectively extract rich spatial–spectral characteristics in hyperspectral data, while contrastive learning is also employed to further enhance classification performance [19]. Conventional hyperspectral image segmentation approaches

often employ limited labeled data (e.g., Pavia University, Indian Pine, Houston 2018), resulting in poor model training and suboptimal semantic segmentation accuracy. Models like FPGA still follow patch-based limitations, relying on global random stratified sampling for batch updates, which raises training computational complexity. To overcome the data scarcity problem, large-scale hyperspectral segmentation datasets have been developed, enabling effective training of semantic segmentation models [20,21]. The segmentation of large-scale hyperspectral images is particularly challenging due to their rich spectral signatures, complex spatial patterns, multi-scale variations, and diverse fine-grained land cover features. Several hyperspectral-specific semantic segmentation approaches have recently been developed to address large-scale land cover classification challenges [22,23]. To further enhance segmentation performance, the self-supervised feature learning strategy has been explored for large-scale hyperspectral image semantic segmentation tasks [24]. Although dominant hyperspectral segmentation approaches are designed for large-scale datasets, their effectiveness in leveraging generalized features from visual foundation models remains substantially constrained.

Visual foundation models inherently encode transferable generalized representations, which can significantly enhance remote sensing image (RSI) interpretation tasks. Recently, several research efforts have attempted to leverage visual foundation models to enhance remote sensing image interpretation tasks, particularly in complex scenarios requiring cross-domain generalization. Pre-trained on a massive-scale general-purpose dataset, the segment-anything model (SAM) can perform fine-grained segmentation of images and videos based on user prompts, without requiring task-specific training data. This demonstrates its powerful zero-shot learning and generalization capabilities, and these capabilities can be effectively applied to remote sensing image interpretation tasks, such as RSI change detection [25], semantic segmentation [26], instance segmentation [27] and land cover classification [28]. These approaches primarily utilize prompts to extract generalized characteristics via encoder architectures for interpretation tasks, effectively bridging foundation models with domain-specific requirements. Another research direction focuses on constructing remote sensing foundation models using extensive remote sensing data, which aims to address the domain gap between computer vision and geospatial applications. These methods primarily build self-supervised learning models through masked learning as well as contrastive learning paradigms. RingMo employs large-scale unlabeled data for training the generative self-supervised learning architecture, and the trained network is then utilized for downstream tasks [29]. In hyperspectral data processing, self-supervised frameworks are established through masking and reconstruction techniques, subsequently applied to hyperspectral data interpretation [30]. Similarly, contrastive learning as well as masked learning are utilized for multimodal remote sensing data land cover classification [31,32], and the SAM is also utilized for assisting the training process [33].

However, current hyperspectral image segmentation approaches face two major limitations. The first one is the inability of effectively extracting generalized features from hyperspectral data with visual foundation models. Visual foundation models are typically trained on large-scale natural images, which exhibit significant structural and semantic discrepancies compared with remote sensing imagery. Multi-scale feature learning is capable of effectively mitigating the domain gap between natural and remote sensing imagery, particularly in applications requiring fine-grained land use classification [34]. FastSAM delivers SAM-comparable zero-shot segmentation performance at 2% computational cost while demonstrating superior multi-scale feature learning, especially advantageous for RSI interpretation [35]. And another is the ineffective information fusion of extracted visual generalized features with multi-scale spatial–spectral features contained in hyperspectral data. Visual generalization features can significantly enhance the performance of remote

sensing image interpretation; however, substantial challenges remain in effectively integrating these features with task-specific spatial–spectral characteristics [36]. And simplistic feature aggregation techniques, such as addition or concatenation, often result in significant information overlap and redundancy, severely hindering the network's ability for effective semantic segmentation. This article proposes leveraging a visual foundation model for extracting generalized features, which are then fused with task-specific spatial–spectral characteristics within the hierarchical semantic segmentation architecture. The principal contributions of this work are concisely outlined as follows:

(1) A visual foundation model is utilized to extract generalized features from spectrally partitioned HSI subgroups, and these features are subsequently transformed into multi-scale representations to bridge the domain gap between natural images and remote sensing data.

(2) An attention-enhanced segmentation architecture is developed to extract task-specific spatial–spectral features, and dedicated adaption blocks systematically integrate multi-scale generalized features with the task-specific characteristics at corresponding network stages.

(3) An efficient feature fusion module is explored within the decoder. This module dynamically and adaptively fuses the encoder's output (which now contains fused features) with the decoder's upsampled features to refine the segmentation output, moving beyond a simple feature fusion strategy.

The remainder of this paper is organized as follows. Section 2 presents the proposed methodology, including the overall framework, U-shape segmentation architecture, visual generalized feature extraction structure, and efficient feature fusion module, and datasets and evaluation metrics are also given in this section. Section 3 provides implementation details and result analysis with both qualitative and quantitative comparisons. Section 4 gives the analysis of learning rates, visual generalized features, the feature fusion module and the spectral grouping strategy, as well as a visualization of the feature fusion approach; future research directions are also discussed. Finally, Section 5 summarizes the main findings and contributions of this study.

2. Materials and Methods

In this section, the proposed hierarchical network with generalized feature fusion for hyperspectral image semantic segmentation is elaborated on, and experimental datasets as well as quantitative evaluation metrics are also introduced in this section. The overall architecture of SAM-GFNet is depicted in Figure 1.

2.1. Overall Framework

The overarching architecture of the proposed SAM-GFNet is designed to synergistically integrate visually generalized features from a foundation model with task-specific spectral–spatial features extracted from the hyperspectral image (HSI) for end-to-end semantic segmentation. The model follows a hierarchical encoder–decoder structure with skip connections, and its data flow can be systematically described in the following stages.

HSI partitioning and generalized feature extraction: The input HSI $X \in \mathbb{R}^{H \times W \times B}$ is first partitioned into non-overlapping subgroups along the spectral dimension. In each subgroup, the PCA operation is conducted and the first three principal components are employed by the pre-trained FastSAM encoder (kept frozen) to generate a set of multi-scale generalized feature maps $\{F_1, F_2, F_3, F_4\}$, corresponding to spatial scales of $1/4$, $1/8$, $1/16$, and $1/32$ of the original input, respectively. Each subgroup, corresponding to a distinct spectral range, yields generalized features at four different scales. Features of the same scale

across all subgroups are then stacked, forming a composite representation that encapsulates the multi-scale generalized characteristics of the hyperspectral data.

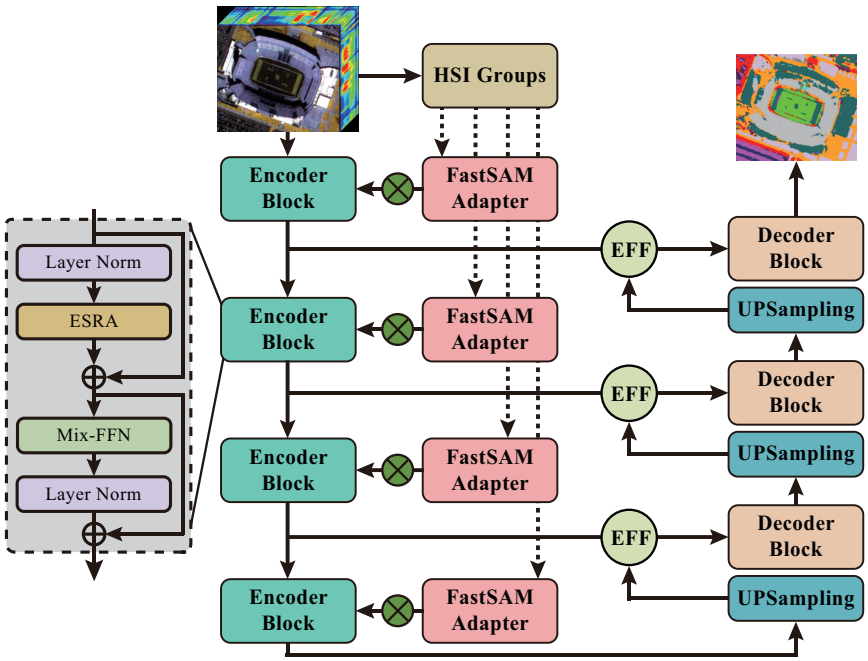


Figure 1. Overall architecture of SAM-GFNet, showing the workflow from hyperspectral input to segmentation output. The diagram illustrates the key components: spectral grouping and generalized feature extraction, task-specific encoder with hierarchical fusion, and decoder with efficient feature fusion modules.

Task-specific feature encoding and hierarchical fusion: The original HSI is simultaneously fed into a task-specific encoder, which is composed of four stages. Each stage contains two efficient spatial reduction attention (ESRA) blocks (detailed in Section 2.3) for progressive spectral–spatial feature extraction, generating a pyramid of feature maps $\{E_1, E_2, E_3, E_4\}$. A core design of the proposed architecture is the hierarchical fusion of generalized and task-specific features. Specifically, at each stage, the generalized feature map E_i is aligned and fused with the encoder feature map F_i via an adaptation and fusion block, producing a set of enriched feature maps $\{E_1', E_2', E_3', E_4'\}$ that carry both domain-general and HSI-specific information.

Decoder with efficient feature fusion strategy: The pyramid feature maps from the encoder are then processed by the decoder to recover spatial resolution and refine the segmentation map. The decoder progressively upsamples these features. At each upsampling stage, the decoder incorporates skip connections from the correspondingly scaled encoder features. The efficient feature fusion (EFF) module (detailed in Section 2.4) is critically employed to adaptively merge the spatially coarse, semantically strong decoder features with the spatially fine, semantically weak encoder features from the skip connections, ensuring a balanced flow of information.

Segmentation head: The final decoder output is passed through a segmentation head, consisting of a convolution layer and a softmax activation, to produce the per-pixel classification map $Y \in \mathbb{R}^{H \times W \times C}$, where C is the number of land cover classes.

This structured workflow ensures that multi-scale generalized features are efficiently integrated at multiple levels of the encoding process, while the decoder, equipped with the EFF modules, effectively leverages these enriched features for precise semantic segmentation. The subsequent subsections will elaborate on the key components mentioned in this overview.

2.2. U-Shaped Segmentation Architecture

The proposed SAM-GFNet adopts encoder–decoder architecture as well as skip connections for transmitting multilevel semantic characteristics. To alleviate the issue of high computation complexity in the original transformer, the efficient spatial reduction attention (ESRA) is exploited for spectral–spatial feature learning of hyperspectral data. The efficient self-attention in SegFormer mainly utilizes the sequence reduction process to reduce the length of the sequence, mainly including reshape and linear operations [37]. While the efficient spatial reduction attention primarily employs convolutional operations and the DropKey empress key and value, thereby reducing model parameters and computational cost. The ESRA block is composed of ESRA structure as well as a mix feed-forward network (Mix-FFN), and the Mix-FFN adopts a feed-forward network architecture similar to that employed in SegFormer.

In the ESRA module, convolutional operations are utilized for reducing the dimensionality of key and value matrices, while dropkey operation is employed to automatically provide each attention head with an adaptive operator. This strategy constrains attention distribution by penalizing regions with high attention weights, thereby promoting smoother attention patterns. As a result, the model is encouraged to explore other task-relevant regions, enhancing its ability to capture robust global features. The overall structure of the ESRA module is formally formulated in Equation (1).

$$ESRA = Softmax(\frac{DropKey(\mathbf{Q} \times SR(\mathbf{K})^T)}{\sqrt{dim_{head}}} \times SR(\mathbf{V})) \quad (1)$$

The mathematical formulation of the $SR()$ operation is defined in Equation (2).

$$SR(\mathbf{x}) = Norm(Conv2d_R(\mathbf{x})) \quad (2)$$

in which $Conv2d_R$ denotes a 2D convolution operation with a kernel size of R. The DropKey operation is formally defined in Equation (3).

$$DropKey(\mathbf{x}) = \mathbf{x} + bernoulli(ones_like(\mathbf{x}) \times ratio) \times -e^{12} \quad (3)$$

in which $\mathbf{x}$ denotes attention weights. The $bernoulli()$ function samples from a Bernoulli distribution, while $ones_like()$ produces a matrix of ones matching the dimensions of $\mathbf{x}$ [38]. The Mix-FFN operation is formally defined in Equation (4).

$$Mix - FFN(\mathbf{x}) = FC(GELU(ConvDW(FC(\mathbf{x})))) \quad (4)$$

The encoder of the segmentation network contains four stages, and each stage combines two ESRA blocks. The channel dimensions in the four stages are defined as $C_1 = 64$, $C_2 = 128$, $C_3 = 320$ and $C_4 = 512$, respectively.

2.3. Visual Generalized Feature Extraction

Remote sensing data and natural images exhibit fundamental differences due to their distinct scene compositions and object categories; thus, multi-scale feature learning is essential for bridging the domain gap between these two types of images. The utilization of FastSAM for extracting visual features is driven by its dual advantages. Firstly, it enables prompt-free operation while effectively capturing multi-scale features, a critical requirement for hyperspectral data. Secondly, its lightweight design, characterized by fewer parameters, offers greater computational efficiency and ease of use. FastSAM utilizes a pyramid feature learning architecture for semantic segmentation, and through large-scale natural image pre-training, its encoder develops effective multi-scale feature extraction capabilities.

Hyperspectral images typically contain dozens to hundreds of spectral bands, covering a range from visible to infrared wavelengths. In contrast, visual foundation models are generally designed for three-band natural images. Due to the high redundancy between adjacent bands in hyperspectral data, directly using all bands for generalization feature extraction would lead to substantial computational cost and redundant information. To address this, the hyperspectral image is divided into multiple subgrouped sections by spectral bands. Within each subgroup, the first three principal components are extracted using PCA. These components are then processed by the visual foundation model to extract generalized features. This approach not only facilitates the extraction of visual features across different spectral ranges but also significantly reduces computational requirements. For each subgroup, FastSAM extracts multi-scale spatial features. Features of the same scale across all subgroups are then stacked, collectively representing the multi-scale generalized features of the hyperspectral data.

Adaptation has become a widely adopted strategy for fine-tuning pre-trained foundation models to downstream tasks, wherein supplementary trainable parameters are incorporated into the frozen backbone architecture to enhance domain-specific feature transfer. Given that FastSAM utilizes a CNN-based backbone, multi-scale feature maps are leveraged to capture hierarchical representations. Specifically, trainable convolutional operations are integrated into the frozen encoder of the pre-trained FastSAM to extract generalized visual features. Feature maps at spatial scales of $1/4$, $1/8$, $1/16$, and $1/32$ are utilized as hierarchical representations, denoted as F_1, F_2, F_3 and F_4, respectively. The adaption operations can be formally expressed through Equation (5).

$$f_i^* = \alpha(f_i) = \gamma(bn(conv_{1\times1}(f_i))) \qquad i = 1, 2, 3, 4 \tag{5}$$

where $conv_{1\times1}$ refers to the 1×1 convolutional transformation, bn indicates batch normalization applied to the feature maps, and γ corresponds to the rectified linear unit (ReLU) activation function. f_i and f_i^* represent characteristics before and after the adaptation operations, respectively. The adaptation-aware network components remain trainable, ensuring that the extracted features are aligned with the domain-specific characteristics of hyperspectral remote sensing imagery.

2.4. Efficient Feature Fusion

Conventional linear fusion approaches, such as summation or concatenation, lack the capacity for content-aware integration of features and are unable to adaptively assign weights to contributions based on informational relevance. In contrast, the human visual system utilizes attention mechanisms to dynamically assign weights according to target importance, selectively enhancing regions of interest. Attention-based fusion methods enable content-aware and nonlinear fusion by adaptively weighting features based on their semantic relevance.

The attention-driven efficient feature fusion (EFF) mechanism synergistically combines local and global representations while dynamically generating channel-wise weighting coefficients conditioned on the semantic and structural characteristics of the feature maps. Given input feature maps $\mathbf{X}$ and $\mathbf{Y}$, the initial fused representation $\mathbf{F}$ is obtained through element-wise summation across corresponding spatial and channel dimensions. During feature fusion, pointwise convolution operations are employed to integrate cross-channel information at each spatial location, while distinct computational branches separately extract local and global contextual features along the channel dimension. The global channel characteristics are derived through Equation (6).

$$G(\mathbf{F}) = \gamma(bn(PWConv(g(\mathbf{F})))) \tag{6}$$

The local channel features are obtained as Equation (7).

$$L(\mathbf{F}) = \gamma(bn(PWConv(\mathbf{F}))) \tag{7}$$

where $g()$ corresponds to the global average pooling operation, γ signifies a nonlinear activation transformation, and bn denotes batch normalization. Local channel features retain the spatial resolution of the original feature maps, thereby preserving and refining fine-grained local patterns. In comparison, global channel features aggregate comprehensive channel-wise contextual information. The attention weights are derived through Equation (8).

$$M(\mathbf{F}) = sigmoid(L(\mathbf{F}) \oplus G(\mathbf{F})) \tag{8}$$

The attention-guided fusion mechanism is formally defined by Equation (9).

$$\mathbf{Z} = M(\mathbf{F}) \otimes \mathbf{X} + (1 - M(\mathbf{F})) \otimes \mathbf{Y} \tag{9}$$

where $\mathbf{Z}$ denotes the fused output feature tensor, with $\oplus$ and $\otimes$ indicating broadcast addition and element-wise multiplication, respectively. Within this attention-driven fusion framework, the coefficients $M(\mathbf{F})$ and $1 - M(\mathbf{F})$, where $M(\mathbf{F}) \in [0, 1]$, constitute spatially adaptive weight matrices that produce the fused representation through a weighted combination of the input feature maps $\mathbf{X}$ and $\mathbf{Y}$. These fusion weights are dynamically optimized during training to facilitate effective integration of both feature representations. Consequently, the model achieves content-aware blending of the two feature types based on their semantic relevance, producing highly discriminative and task-adaptive representations.

During training, the model optimizes the predicted segmentation maps by minimizing a cross-entropy loss function that reduces the divergence between the predicted class probability distribution and the ground truth label distribution at each pixel location. The cross-entropy loss is formally defined in Equation (10).

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_i^{(j)} \cdot log(\hat{y}_i^{(j)}) \tag{10}$$

where N denotes the total number of pixels and C represents the number of semantic classes. Here, $y_i^{(j)}$ corresponds to the ground truth label, and $\hat{y}_i^{(j)}$ indicates the predicted probability that the i-th pixel belongs to the j-th class. During training, the model computes and minimizes the divergence between these distributions, thereby optimizing per-pixel classification accuracy across the image.

2.5. Benchmark Datasets

The WHU-OHS dataset is collected from the Orbita hyperspectral satellite (OHS) sensor, which contains 42 OHS satellite images acquired from more than 40 geographically diverse regions in China, annotated with 24 typical land cover categories. This dataset features a spatial resolution of 10 m and comprises 32 spectral channels, with an average spectral resolution of 15 nm. The annotated satellite imagery has been systematically partitioned into non-overlapping patches of dimensions $512 \times 512 \times 32$ to facilitate model training. These images are partitioned into training, validation and test sets, containing 4821, 513 and 2459 patches, respectively [20]. The land cover category distribution and visualization of the WHU-OHS dataset is shown in Figure 2.

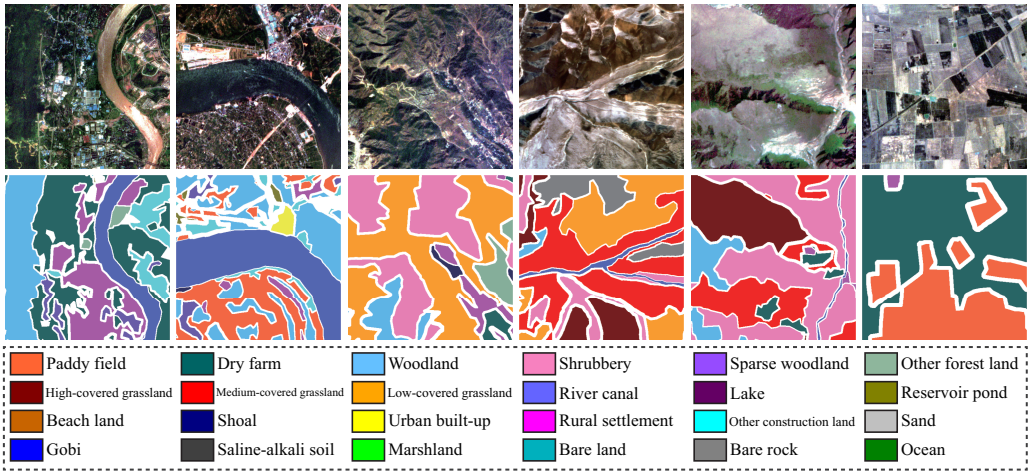


Figure 2. Land cover category distribution and visualization for the WHU-OHS dataset.

The WHU-H2SR dataset is collected by an airborne hyperspectral imaging spectrometer which constitutes a large-scale hyperspectral imaging resource characterized by a spatial resolution of 1 m, covering an extensive geographic area of 227.79 km^2 in southern Shenyang, Liaoning Province, China. The dataset comprises 249 spectral bands covering the wavelength range from 391 to 984 nm, with a spectral resolution of 5 nm. Each hyperspectral patch has dimensions of $300 \times 300 \times 249$, and the dataset encompasses eight distinct land cover categories. The data is partitioned into 1516, 253, and 762 patches for the training, validation, and test subsets, respectively [21]. The land cover category distribution and visualization of the WHU-H2SR dataset is shown in Figure 3.

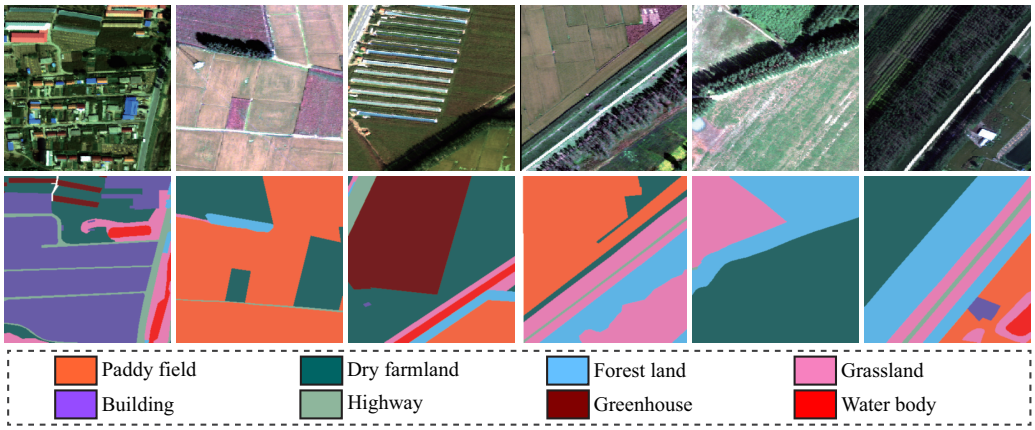


Figure 3. Land cover category distribution and visualization for the WHU-H2SR dataset.

2.6. Evaluation Metrics

To quantitatively evaluate the efficacy and generalization capacity of SAM-GFNet, widely adopted semantic segmentation metrics are utilized, including overall accuracy (OA), mean intersection over union (mIoU) and mean F1-score (mF1). More precisely, overall accuracy refers to the proportion of correctly classified pixels out of the total number of pixels evaluated, which is formally defined as Equation (11).

$$OA = \frac{\sum_{i=1}^{l} p_{ii}}{\sum_{i=1}^{l} \sum_{j=1}^{l} p_{ij}} \tag{11}$$

where i, j, and l correspond to the ground truth class, predicted class and total number of categories, respectively, and p_{ij} denotes the number of pixels with ground truth class i that are classified as class j. mIoU and mF1 are commonly employed to evaluate the remote

sensing image segmentation task. The mIoU metric computes the ratio of the intersection to the union of predicted and ground truth segments, formally expressed as Equation (12).

$$mIoU = \frac{1}{l}\sum_{i=1}^{l}\frac{p_{ii}}{p_{ij} + \sum_{j=1}^{l} p_{ji} - p_{ii}} \tag{12}$$

The mF1 is defined as the harmonic mean of precision and recall, expressed mathematically as shown in Equations (13)–(15).

$$mF1 = \frac{1}{l}\frac{2 \times P \times R}{P + R} \tag{13}$$

$$P = \sum_{i=1}^{l}\frac{p_{ii}}{\sum_{j=1}^{l} p_{ij} + p_{ii}} \tag{14}$$

$$R = \sum_{i=1}^{l}\frac{p_{ii}}{\sum_{j=1}^{l} p_{ji} + p_{ii}} \tag{15}$$

3. Results

3.1. Implementation Details

Land cover classification comparative experiments are carried out to assess the efficacy and advantage of the proposed methodology. All experiments are conducted on a workstation with an Intel(R) Xeon(R) processor and NVIDIA GeForce RTX 3090 24 GB GPU. The SAM-GFNet is implemented using PyTorch 2.5.1, while baseline approaches are reproduced in their original experimental configurations. For the training of SAM-GFNet, the Adaptive Moment Estimation (Adam) optimizer is chosen for model optimization, and the optimal learning rate is selected from a candidate set. The model is trained for 50 epochs on both datasets with batch size of 4. No data augmentation techniques are applied in the training procedure. The SAM-GFNet is trained from scratch, and the frozen FastSAM is employed to extract generalized features.

This section presents a comprehensive evaluation of the classification performance, employing both qualitative visual comparisons and quantitative metrics. For the WHU-OHS dataset with 32 spectral bands, it is divided into 4 groups, each containing 8 consecutive bands. For the WHU-H2SR dataset with 249 spectral bands, it is divided into 10 groups, with the first 9 groups containing 25 bands and the final group containing 24 bands. Within each subgroup, PCA is used to extract the first three principal components, followed by the usage of FastSAM to extract multi-scale visual generalized features. The features corresponding to different subgroups are then stacked according to their scales, and these stacked features represent the multi-scale generalized characteristics of the hyperspectral dataset. The comparative framework incorporates spectral-only methods, patch-based pixel-wise classification techniques, and state-of-the-art semantic segmentation models for hyperspectral imagery. The spectral-only classification approaches utilize established models, including support vector machine (SVM) [4] and 1D convolutional neural network (1D-CNN) [5]. For patch-based classification, comparative methods comprise 3D-CNN [39], hierarchical residual network with attention mechanism (HResNet) [10] and adaptive spectral–spatial kernel improved residual network (A2S2K-ResNet) [40]. Semantic segmentation techniques for hyperspectral imagery incorporate fast patch-free global learning (FPGA) [14], deformable transformer with spectral U-net (DTSU) [22], semantic segmentation of hyperspectral images (SegHSI) [23] and spectral–spatial hierarchical masked modeling framework (S2HM2) [24].

3.2. Quantitative Evaluation

The classification accuracy for each category and main evaluation coefficients of different different land cover classification approaches on two hyperspectral image semantic segmentation datasets are reported in Tables 1 and 2. Based on a comparative analysis of the quantitative results, the following conclusions can be drawn.

Segmentation-based land cover classification methods consistently achieve superior accuracy compared to patch-based approaches. Since pixel-wise segmentation approaches are generally developed for land cover classification in small-scale hyperspectral images, these methods demonstrate high performance on conventional datasets with simple scenes, but they exhibit significantly reduced accuracy when applied to large-scale hyperspectral imagery. The pixel-wise classification method achieved its highest accuracy on the WHU-OHS and WHU-H2SR datasets (obtained by the A2S2K method) at 54.14% and 80.62%, respectively, which have a significant performance gap compared to segmentation-based approaches, particularly evident on the WHU-OHS dataset. For categories 'Beach land', 'Shoal', 'Saline-alkali soil', 'Marshland', and 'Bare land' in the WHU-OHS dataset and the category 'Grassland' in the WHU-H2SR dataset, the pixel-wise classification methods demonstrate poor performance, whereas the segmentation-based approaches achieved substantially improved performance on these categories. This performance limitation stems from the high complexity of large-scale hyperspectral scenes, wherein patch-based techniques are constrained to local spatial feature extraction and demonstrate limited capability in learning semantically rich, globally coherent representations. In comparison, semantic segmentation models process the entire image as input to holistically integrate global spatial and spectral characteristics, resulting in substantially improved classification performance for large-scale hyperspectral image analysis tasks.

Generalized representations derived from visual foundation models enhance the hyperspectral image semantic segmentation performance. The self-supervised learning-based S2HM2 model, which is directly pre-trained on hyperspectral imagery before fine-tuning for supervised segmentation tasks, demonstrates improved overall segmentation performance. However, it still achieves low accuracy on specific categories, namely 'Sparse woodland', 'Other forest land', 'Beach land', and 'Bare land' in the WHU-OHS dataset and 'Building' in the WHU-H2SR dataset. Pre-trained on extensive natural image datasets, the foundation model feature extractor is capable of capturing transferable characteristics from remote sensing imagery through adaptation networks. By adaptively integrating these generalized features with domain-specific spectral–spatial information, the proposed method achieves state-of-the-art classification performance on both the WHU-OHS and WHU-H2SR datasets, particularly for the aforementioned categories.

A comparative assessment of classification performance across the two datasets indicates that the proposed method consistently attains higher accuracy on airborne hyperspectral imagery compared to spaceborne data. This discrepancy is primarily attributable to the enhanced spatial and spectral resolution characteristics of airborne systems. The WHU-H2SR airborne dataset, with its 249 spectral bands and high spectral resolution, enables pixel-wise classification methods to effectively leverage detailed spectral information, thereby achieving robust baseline accuracy. The proposed approach further augments performance by synergistically integrating spectral–spatial features from hyperspectral data with the generalized representational capacity of the vision foundation model, and the improvement in accuracy is more pronounced on the WHU-OHS dataset.

The number of model parameters is used to evaluate the architectural complexity of deep learning models, while floating point operations (FLOPs) serve as a metric for computational complexity. Analysis of these two metrics across different deep learning models reveals that pixel-wise deep learning models (such as 1DCNN and HResNet) typically

exhibit fewer parameters and lower computational complexity. In contrast, end-to-end image segmentation models generally possess greater model parameters and higher computational complexity, with the self-supervised S2HM2 model particularly demonstrating substantial parameter counts and computational demands. Since the proposed method employs pre-trained visual foundation models to extract generalized features, its parameter count remains comparable to other end-to-end models in terms of computational complexity, while achieving significantly superior performance in segmentation accuracy.

3.3. Classification Map Analysis

Land cover classification maps are employed for qualitative evaluation, with semantic segmentation results generated by various methods on both datasets presented in Figures 4 and 5. In these visualizations, distinct land cover categories are delineated using unique color mappings, and corresponding ground truth references are provided to facilitate comparative analysis.

Segmentation-based land cover classification methods yield visually superior results compared to pixel-wise approaches, particularly for large-scale polygonal features, such as "High-covered grassland" and "Beach land" categories in the WHU-OHS dataset and the "Grassland" category in WHU-H2SR. As evidenced in the classification maps, segmentation techniques achieve accurate delineation of extensive areal objects, whereas pixel-based methods display pronounced misclassification artifacts. Although patch-based classification strategies leverage spatial information within local receptive fields, their capacity remains constrained to learning features from limited contextual regions. In contrast, semantic segmentation models capture globally coherent spatial patterns, enabling more robust and precise classification of continuous geographical features.

By leveraging both visually generalized features and task-specific characteristics, the proposed SAM-GFNet achieves superior performance in large-area land cover classification and fine-grained object distinction. While spaceborne hyperspectral imagery often exhibits limited spectral discriminability, the proposed method effectively harnesses transferable representation from visual foundation models and global spatial–spectral features from hyperspectral data for segmentation. This integration enables accurate identification of extensive geographical regions, particularly on the WHU-OHS dataset with its diverse and complex land cover categories, demonstrating particularly pronounced performance on the 'Paddy field' and 'High-covered grassland' categories. Furthermore, the model incorporates detailed spectral information to support the discrimination of fine-grained spatial structures.

Table 1. Quantitative results of different approaches on the WHU-OHS dataset.

Category	Spectral-Only		Patch-Based			End-to-End				SAM-GFNet
	SVM	1DCNN	3DCNN	HResNet	A2S2K	FPGA	DTSU	SegHSI	S2HM2	
Paddy field	67.02	34.11	37.05	50.39	59.21	84.36	**87.76**	85.03	83.78	86.81
Dry farm	37.64	84.51	80.56	74.88	82.31	85.11	83.45	85.42	**86.60**	85.75
Woodland	66.78	54.29	63.62	65.78	69.79	79.17	82.26	80.13	**86.37**	83.80
Shrubbery	0.19	0.12	0.14	0.00	2.28	52.93	54.79	48.47	48.26	**57.03**
Sparse woodland	0.00	0.00	0.00	0.00	0.00	0.23	0.99	1.18	0.20	**18.79**
Other forest land	0.00	0.00	0.00	0.00	0.00	0.02	0.26	0.82	0.00	**10.56**
High-covered grassland	2.03	0.01	0.47	37.54	47.89	60.40	68.52	59.12	66.75	**69.46**
Medium-covered grassland	1.76	0.00	2.91	2.28	4.28	42.57	44.39	29.49	27.67	**49.54**
Low-covered grassland	16.50	7.06	18.84	15.33	13.79	45.25	57.59	63.50	55.94	**63.51**
River canal	17.11	1.56	39.08	9.68	28.50	68.14	76.28	66.88	71.51	**81.40**
Lake	78.32	35.79	74.63	10.19	32.22	90.28	91.41	93.81	92.26	**95.69**
Reservoir pond	4.70	4.49	19.75	22.58	24.42	58.88	67.25	60.41	67.45	**76.89**
Beach land	0.00	0.00	0.00	0.00	0.00	28.97	32.95	39.90	6.54	**55.22**
Shoal	0.00	0.00	2.00	0.00	0.00	54.32	56.06	52.47	54.34	**67.91**
Urban built-up	37.23	64.05	39.19	75.72	74.77	87.11	88.53	**89.91**	88.88	88.88
Rural settlement	0.26	0.01	6.73	27.44	25.79	58.25	61.83	61.38	58.62	**68.59**
Other construction land	0.00	1.01	0.46	0.05	3.52	38.19	39.70	27.57	31.51	**54.98**
Sand	31.76	39.00	46.71	0.66	45.77	78.01	82.29	82.46	71.22	**87.88**
Gobi	0.00	0.00	0.00	47.10	2.06	66.71	78.51	**80.72**	67.03	79.04
Saline–alkali soil	0.00	0.00	0.02	0.00	0.00	27.78	41.94	38.17	35.92	**55.44**
Marshland	0.00	0.00	0.00	0.00	1.29	35.46	47.21	23.20	43.09	**69.11**
Bare land	0.00	0.00	0.00	0.00	0.00	0.24	7.33	3.28	1.61	**25.83**
Bare rock	4.90	0.00	20.16	17.58	76.06	74.19	84.06	84.07	**87.41**	87.31
Ocean	84.86	98.30	89.49	96.99	96.45	99.13	99.13	99.75	99.56	**99.95**
OA	37.22	41.74	44.09	47.95	54.14	73.66	76.53	74.85	76.66	**79.60**
mF1	18.00	15.11	22.27	21.41	27.19	57.03	62.35	59.09	60.22	**69.64**
mIoU	12.30	9.95	15.36	14.14	18.55	44.47	49.93	46.75	48.38	**56.96**
Params (M)	-	0.03	0.03	0.02	0.02	2.55	8.65	2.57	26.45	3.42
FLOPs (G)	-	17.46	234.86	25.74	20.34	241.35	113.45	56.73	116.75	86.47

Table 2. Quantitative results of different approaches on the WHU-H2SR dataset.

Category	Spectral-Only		Patch-Based				End-to-End			SAM-GFNet
	SVM	1DCNN	3DCNN	HResNet	A2S2K	FPGA	DTSU	SegHSI	S2HM2	
Paddy field	93.35	94.08	94.68	95.66	96.16	93.99	**96.53**	94.04	96.26	95.62
Dry farmland	85.38	7.95	89.13	90.33	86.93	90.23	86.53	**92.94**	92.86	89.26
Forest land	45.57	43.06	46.80	41.36	50.13	69.73	68.14	63.09	59.76	**71.79**
Grassland	0.00	4.67	5.60	0.10	0.31	27.29	41.96	18.93	23.58	**50.02**
Building	81.16	79.98	74.70	68.64	79.12	84.66	**88.64**	80.84	83.07	85.09
Highway	7.76	15.95	18.68	33.67	34.00	51.05	43.06	37.08	56.79	**58.33**
Greenhouse	0.02	1.78	16.11	25.39	28.94	55.04	54.81	32.24	**68.83**	61.14
Water body	60.09	54.96	56.52	60.77	68.38	68.31	77.12	67.62	43.60	**79.01**
OA	57.43	68.39	76.46	80.07	80.62	84.90	85.36	84.10	85.25	**86.92**
mF1	44.68	46.74	51.15	54.44	56.24	69.69	70.73	64.42	69.76	**74.47**
mIoU	32.25	36.61	39.73	42.33	43.98	56.15	57.22	51.04	56.59	**61.39**
Params (M)	-	0.05	0.06	0.04	0.06	3.86	9.35	4.38	44.27	4.69
FLOPs (G)	-	38.51	411.23	48.62	39.75	345.87	156.83	78.36	217.64	102.52

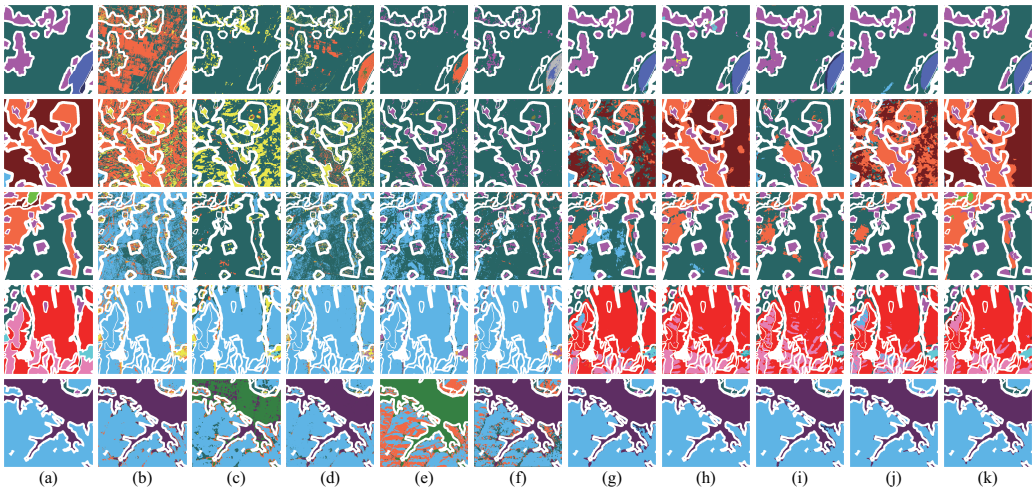


Figure 4. Ground truth and classification maps on the WHU-OHS dataset. (**a**) Ground truth, (**b**) SVM (37.22%), (**c**) 1DCNN (41.74%), (**d**) 3DCNN (44.09%), (**e**) HResNet (47.95%), (**f**) A2S2K (54.14%), (**g**) FPGA (73.66%), (**h**) DTSU (76.53%), (**i**) SegHSI (74.85%), (**j**) S2HM2 (76.66%), (**k**) SAM-GFNet (79.60%).

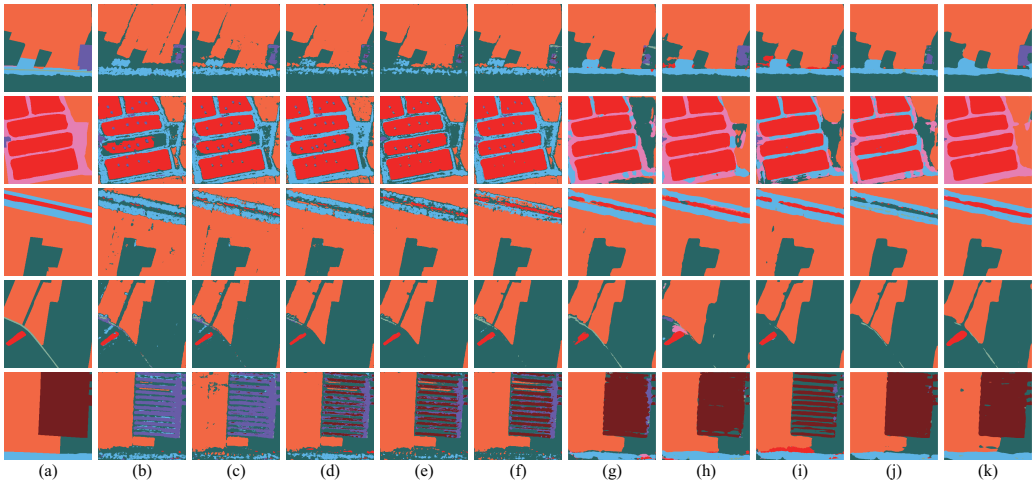


Figure 5. Ground truth and classification maps on the WHU-H2SR dataset. (**a**) Ground truth, (**b**) SVM (57.43%), (**c**) 1DCNN (68.39%), (**d**) 3DCNN (76.46%), (**e**) HResNet (80.07%), (**f**) A2S2K (80.62%), (**g**) FPGA (84.90%), (**h**) DTSU (85.36%), (**i**) SegHSI (84.10%), (**j**) S2HM2 (85.25%), (**k**) SAM-GFNet (86.92%).

4. Discussion

4.1. Learning Rate Analysis

In the training procedure, the learning rate affects the speed of model weight updates, thereby affecting the semantic segmentation accuracy of this model. This subsection evaluates the impact of this hyperparameter by presenting semantic segmentation results obtained using different learning rates on two datasets. The candidate set of learning rates is $\{0.0001, 0.001, 0.002, 0.005, 0.01\}$, and corresponding semantic segmentation accuracies are reported in Figure 6. The experimental results demonstrate that an excessively small learning rate prevents the model from converging to optimal solutions, while an overly large learning rate leads to training instability with significant oscillations. Based on empirical validation, the optimal learning rates for two experimental configurations were determined to be 0.001 and 0.002 respectively.

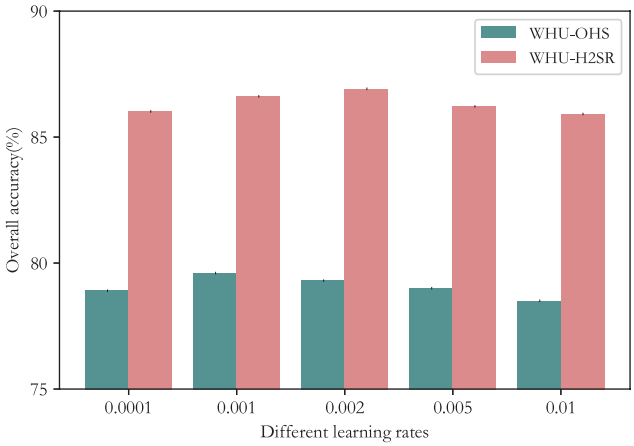


Figure 6. Sensitivity analysis of learning rate on two benchmark datasets.

4.2. Effect of Visual Generalized Features

To evaluate the contribution of visual generalization features to segmentation performance, the ablation study of visual generalized features is conducted in this subsection. In two comparative experiments, the quantitative results with and without generalization features are presented in Table 3. These experimental results clearly demonstrate that incorporating visual generalized features significantly improves land cover classification accuracy. This improvement is particularly notable for the WHU-OHS dataset, where the larger coverage area and more complex land cover types make the visual generalization features especially effective in enhancing object recognition precision.

Table 3. Ablation analysis of visual generalized features on two benchmark datasets.

Datasets	WHU-OHS			WHU-H2SR		
	OA	mF1	mIoU	OA	mF1	mIoU
Without Generalized Features	78.62	67.58	54.93	86.03	72.94	60.63
Using Generalized Features	79.60	69.64	56.96	86.92	74.47	61.39

4.3. Effect of Feature Fusion Module

This study introduces an adaptive feature fusion mechanism to integrate encoded features with decoder-generated characteristics. To assess the efficacy of this approach, ablation experiments are conducted comparing three traditional linear fusion strategies: feature summation, feature concatenation and the proposed adaptive fusion technique. To thoroughly validate the advancement of the proposed fusion method, the cross-modal spatio-channel attention (CSCA) module in [41] and multimodal feature self-attention

fusion (MFSAF) in [42] are modified as feature fusion modules for the proposed architecture, thereby enabling a more comprehensive comparative analysis. Results summarized in Table 4 confirm that adaptive fusion methods yield better classification performance across evaluated metrics. The superior performance stems from their ability to address the fundamental domain gap between encoded spatial–spectral and generalized features with decoder-generated characteristics, while simple summation or concatenation operations fail to reconcile these heterogeneous feature spaces. The proposed adaptive feature fusion method accounts for the feature disparities between the encoder and decoder, leveraging an adaptive weighting strategy for information fusion. This is particularly crucial for hyperspectral analysis, where the spatial–spectral specificity must be carefully preserved while incorporating generalized visual knowledge.

Table 4. Ablation analysis of feature fusion pattern on two benchmark datasets.

Datasets	WHU-OHS			WHU-H2SR		
	OA	**mF1**	**mIoU**	**OA**	**mF1**	**mIoU**
Summation	78.33	68.26	55.64	85.63	73.82	60.71
Concatenation	79.02	68.94	56.04	86.15	73.97	61.05
CSCA	79.26	69.14	56.23	86.56	74.02	60.87
MFSAF	78.88	68.89	55.86	85.87	73.89	60.64
EFF	79.60	69.64	56.96	86.92	74.47	61.39

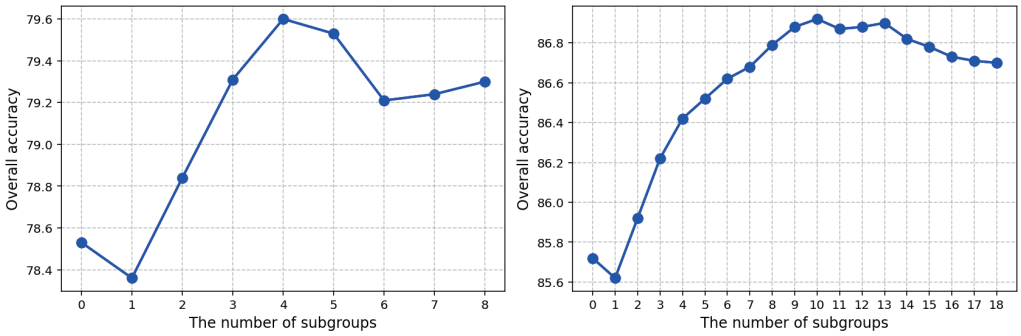
4.4. Effect of Spectral Grouping Strategy

In the proposed model, the hyperspectral image comprising multiple spectral bands is partitioned into several subgroups, within which visual generalized features corresponding to their respective spectral ranges are extracted. Given that the input to visual foundation models consists of three-band images, specific bands were selected for the experiments: bands 15, 9, and 5 for the WHU-OHS dataset, and bands 109, 63, and 34 for the WHU-H2SR dataset. This selection is primarily based on the spectral ranges of these bands being close to those of natural images. Additionally, the hyperspectral data are divided into multiple subgroups, with the number of bands in each subgroup kept as consistent as possible. The PCA operation is applied to each subgroup to extract the first three principal components. The corresponding overall classification accuracies are presented in Figure 7. The experimental results indicate that applying PCA to all bands of the hyperspectral image yields lower accuracy than using the three selected bands. This can be attributed to the fact that the selected bands are spectrally similar to natural images, making the extracted features more conducive to remote sensing image interpretation tasks. Furthermore, as the number of subgroups increases, the classification accuracy gradually improves until it plateaus and may even decline. This trend occurs because an increasing number of subgroups introduces growing redundancy among features from different subgroups. Considering both accuracy and computational cost, the number of subgroups is set to 4 for the WHU-OHS dataset and 10 for the WHU-H2SR dataset.

4.5. Visualization of Feature Fusion Approach

The proposed feature fusion approach utilizes an attention mechanism to integrate encoder and decoder features, achieving adaptive information enhancement based on feature content. To validate the effectiveness of this fusion mechanism, an ablation study of the attention-based feature fusion is conducted, with visual comparisons of multi-scale feature maps before and after feature fusion, as presented in Figure 8. In these visualizations, brighter colors indicate higher attention weights. The feature map analysis demonstrates that this fusion strategy adaptively enhances information for specific ground objects in

the feature representations, enabling the model to more effectively segment these target categories and consequently improve land cover recognition performance.



(a) WHU-OHS (b) WHU-H2SR

Figure 7. Sensitivity analysis of spectral grouping strategy on two benchmark datasets.

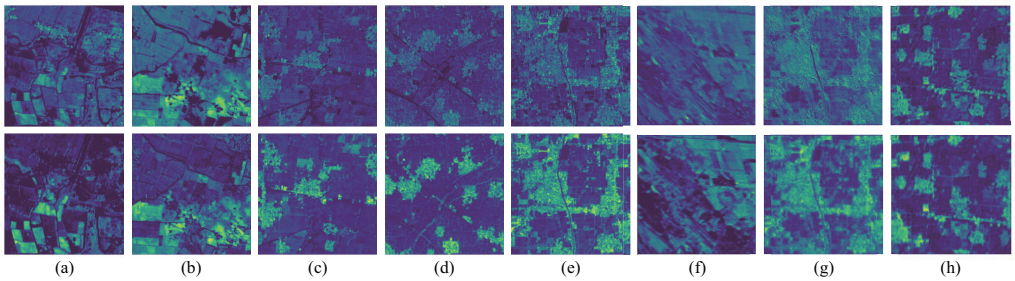


Figure 8. Visualization of feature fusion approach. Among them, (**a**–**h**) are some selected feature maps, with the top being the feature maps before fusion and the bottom being the feature maps after fusion.

4.6. Future Research Directions

Based on our findings, several promising directions for future research are identified to enhance the model's generalization capability and practical utility. First, investigating the framework's robustness across diverse hyperspectral data sources represents a critical direction. This includes validating the model on data from different sensors (e.g., AVIRIS, HyMap, PRISMA) with varying spatial–spectral resolutions and acquired under different atmospheric and illumination conditions. Such cross-domain evaluation would thoroughly assess the model's generalization capacity and guide the development of more universal hyperspectral segmentation solutions.

Second, conducting real-world field testing and validation is essential for translating laboratory performance to practical applications. Future work should involve collaborative studies with domain experts in areas such as precision agriculture, mineral exploration, and environmental monitoring. Comparing model predictions with in situ measurements and expert annotations would provide crucial insights into the method's operational strengths and limitations, particularly for time-sensitive applications like crop health assessment or disaster response.

Additionally, investigating model efficiency and deployment considerations represents another important direction. Developing lightweight versions of the architecture optimized for resource-constrained platforms, such as drones or satellite edge computing systems, would facilitate real-time analysis and decision-making. This could involve network pruning, knowledge distillation, or specialized hardware acceleration techniques tailored for hyperspectral processing.

4.7. Implications for Practical Applications

The proposed method has significant implications for practical remote sensing applications beyond general land cover classification. The enhanced capability to integrate detailed spectral information with spatial context enables more specialized applications that require fine-grained material discrimination. In urban environments, the method shows potential for precise material identification, including distinguishing between different roofing materials, pavement types, and building composites, which is crucial for urban heat island studies and infrastructure planning.

For vegetation monitoring, the model's sensitivity to subtle spectral variations can support species identification, stress detection, and biomass estimation with improved accuracy. In precision agriculture, this capability can be leveraged for crop health assessment, nutrient deficiency detection, and yield prediction. Furthermore, in geological applications, the method's robust feature extraction makes it particularly suitable for mineral mapping and exploration, where discriminating between mineral types with similar spectral signatures remains challenging.

Moreover, the hierarchical fusion architecture provides a flexible framework that can be adapted to other multimodal remote sensing tasks, such as combining hyperspectral data with LiDAR or SAR data. The attention-based fusion mechanism offers a principled approach for integrating heterogeneous data sources with varying characteristics and information content, extending the method's applicability to diverse environmental monitoring and resource management scenarios.

5. Conclusions

This study explores the integration of SAM-derived generalized features within a hierarchical network architecture for semantic segmentation of hyperspectral imagery. Since both visual generalized features and task-specific characteristics contribute to the semantic segmentation task, this study sought to learn spatial visual generalized features from different spectral ranges of hyperspectral data, which are adaptively integrated with multi-scale task-specific characteristics in the hierarchical architecture. Comprehensive evaluations on two large-scale benchmarks demonstrate the superiority and effectiveness of our approach. On the WHU-OHS dataset, SAM-GFNet achieves an OA of 79.60%, representing improvements of 2.94%, 3.07% and 25.46% over S2HM2, DTSU, and A2S2K respectively. Similarly, on the WHU-H2SR dataset, the proposed method obtains an OA of 86.92%, outperforming S2HM2, DTSU, and A2S2K by 1.67%, 1.56%, and 6.30% respectively. This study presents an approach integrating vision foundation models with end-to-end architectures fine-tuned on domain-specific datasets. While demonstrating effectiveness in closed-set scenarios, the method still shows limitations in open-category segmentation tasks.

Author Contributions: Conceptualization, X.L.; methodology, Y.G. and J.R.; data curation, C.T.; writing—original draft preparation, Y.G.; writing—review and editing, Y.G. and C.T.; supervision, F.S.; funding acquisition, J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available because of privacy restrictions.

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

References

1. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [CrossRef] [PubMed]
2. Hajaj, S.; El Harti, A.; Pour, A.B.; Jellouli, A.; Adiri, Z.; Hashim, M. A review on hyperspectral imagery application for lithological mapping and mineral prospecting: Machine learning techniques and future prospects. *Remote Sens. Appl. Soc. Environ.* **2024**, *35*, 101218. [CrossRef]
3. Zhang, B.; Zhao, L.; Zhang, X. Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images. *Remote Sens. Environ.* **2020**, *247*, 111938. [CrossRef]
4. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
5. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
6. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
7. Dalla Mura, M.; Benediktsson, J.A.; Waske, B.; Bruzzone, L. Morphological Attribute Profiles for the Analysis of Very High Resolution Images. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3747–3762. [CrossRef]
8. Ghamisi, P.; Souza, R.; Benediktsson, J.A.; Zhu, X.X.; Rittner, L.; Lotufo, R.A. Extinction Profiles for the Classification of Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5631–5645. [CrossRef]
9. Aptoula, E.; Ozdemir, M.C.; Yanikoglu, B. Deep Learning With Attribute Profiles for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1970–1974. [CrossRef]
10. Xue, Z.; Yu, X.; Liu, B.; Tan, X.; Wei, X. HResNetAM: Hierarchical Residual Network With Attention Mechanism for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3566–3580. [CrossRef]
11. Paoletti, M.E.; Haut, J.M.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.J.; Pla, F. Deep Pyramidal Residual Networks for Spectral–Spatial Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 740–754. [CrossRef]
12. Zhong, Z.; Li, Y.; Ma, L.; Li, J.; Zheng, W.S. Spectral–Spatial Transformer Network for Hyperspectral Image Classification: A Factorized Architecture Search Framework. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3115699. [CrossRef]
13. He, Y.; Tu, B.; Liu, B.; Li, J.; Plaza, A. 3DSS-Mamba: 3D-Spectral-Spatial Mamba for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3472091. [CrossRef]
14. Zheng, Z.; Zhong, Y.; Ma, A.; Zhang, L. FPGA: Fast Patch-Free Global Learning Framework for Fully End-to-End Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5612–5626. [CrossRef]
15. Xu, Y.; Du, B.; Zhang, L. Beyond the Patchwise Classification: Spectral-Spatial Fully Convolutional Networks for Hyperspectral Image Classification. *IEEE Trans. Big Data* **2020**, *6*, 492–506. [CrossRef]
16. Wang, D.; Du, B.; Zhang, L. Fully Contextual Network for Hyperspectral Scene Parsing. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3050491. [CrossRef]
17. Lee, H.; Eum, S.; Kwon, H. Exploring Cross-Domain Pretrained Model for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3165441. [CrossRef]
18. Yu, H.; Xu, Z.; Zheng, K.; Hong, D.; Yang, H.; Song, M. MSTNet: A Multilevel Spectral–Spatial Transformer Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3186400. [CrossRef]
19. Huang, X.; Dong, M.; Li, J.; Guo, X. A 3-D-Swin Transformer-Based Hierarchical Contrastive Learning Method for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3113721. [CrossRef]
20. Li, J.; Huang, X.; Tu, L. WHU-OHS: A benchmark dataset for large-scale Hersepctral Image classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 103022. [CrossRef]
21. Yu, B.; Li, J.; Huang, X. STSNet: A cross-spatial resolution multi-modal remote sensing deep fusion network for high resolution land-cover segmentation. *Inf. Fusion* **2025**, *114*, 102689. [CrossRef]
22. Zhang, T.; Xue, Z.; Su, H. Deformable Transformer and Spectral U-Net for Large-Scale Hyperspectral Image Semantic Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 20227–20244. [CrossRef]
23. Liu, H.; Li, W.; Xia, X.G.; Zhang, M.; Guo, Z.; Song, L. SegHSI: Semantic Segmentation of Hyperspectral Images With Limited Labeled Pixels. *IEEE Trans. Image Process.* **2024**, *33*, 6469–6482. [CrossRef]
24. Tu, L.; Li, J.; Huang, X.; Gong, J.; Xie, X.; Wang, L. S2HM2: A Spectral–Spatial Hierarchical Masked Modeling Framework for Self-Supervised Feature Learning and Classification of Large-Scale Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5517019. [CrossRef]
25. Zhang, D.; Wang, F.; Ning, L.; Zhao, Z.; Gao, J.; Li, X. Integrating SAM with Feature Interaction for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3483775. [CrossRef]
26. Ma, X.; Wu, Q.; Zhao, X.; Zhang, X.; Pun, M.O.; Huang, B. SAM-assisted Remote Sensing Imagery Semantic Segmentation with Object and Boundary Constraints. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3443420. [CrossRef]

27. Luo, M.; Zhang, T.; Wei, S.; Ji, S. SAM-RSIS: Progressively Adapting SAM with Box Prompting to Remote Sensing Image Instance Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3460085. [CrossRef]
28. Ma, K.; Yao, C.; Liu, B.; Hu, Q.; Li, S.; He, P.; Han, J. Segment Anything Model-Based Hyperspectral Image Classification for Small Samples. *Remote Sens.* **2025**, *17*, 1349. [CrossRef]
29. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A Remote Sensing Foundation Model with Masked Image Modeling. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3194732. [CrossRef]
30. Wang, D.; Hu, M.; Jin, Y.; Miao, Y.; Yang, J.; Xu, Y.; Qin, X.; Ma, J.; Sun, L.; Li, C.; et al. HyperSIGMA: Hyperspectral Intelligence Comprehension Foundation Model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 6427–6444. [CrossRef]
31. Xue, Z.; Liu, B.; Yu, A.; Yu, X.; Zhang, P.; Tan, X. Self-Supervised Feature Representation and Few-Shot Land Cover Classification of Multimodal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3217893. [CrossRef]
32. Xue, Z.; Yang, G.; Yu, X.; Yu, A.; Guo, Y.; Liu, B.; Zhou, J. Multimodal self-supervised learning for remote sensing data land cover classification. *Pattern Recognit.* **2025**, *157*, 110959. [CrossRef]
33. Yan, Z.; Li, J.; Li, X.; Zhou, R.; Zhang, W.; Feng, Y.; Diao, W.; Fu, K.; Sun, X. RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3332219. [CrossRef]
34. Jiao, L.; Wang, M.; Liu, X.; Li, L.; Liu, F.; Feng, Z.; Yang, S.; Hou, B. Multiscale Deep Learning for Detection and Recognition: A Comprehensive Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 5900–5920. [CrossRef]
35. Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; Wang, J. Fast Segment Anything. *arXiv* **2023**, arXiv:2306.12156. [CrossRef]
36. Ding, L.; Zhu, K.; Peng, D.; Tang, H.; Yang, K.; Bruzzone, L. Adapting Segment Anything Model for Change Detection in VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3368168. [CrossRef]
37. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Álvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; pp. 12077–12090.
38. Li, B.; Hu, Y.; Nie, X.; Han, C.; Jiang, X.; Guo, T.; Liu, L. DropKey for Vision Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 22700–22709.
39. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*, 258619. [CrossRef]
40. Roy, S.K.; Manna, S.; Song, T.; Bruzzone, L. Attention-Based Adaptive Spectral–Spatial Kernel ResNet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7831–7843. [CrossRef]
41. Zhang, Y.; Choi, S.; Hong, S. Memory-efficient cross-modal attention for RGB-X segmentation and crowd counting. *Pattern Recognit.* **2025**, *162*, 111376. [CrossRef]
42. Wang, T.; Chen, G.; Zhang, X.; Liu, C.; Wang, J.; Tan, X.; Zhou, W.; He, C. LMFNet: Lightweight Multimodal Fusion Network for high-resolution remote sensing image segmentation. *Pattern Recognit.* **2025**, *164*, 111579. [CrossRef]

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