remote sensing

MDPI

Article

Weakly Supervised Semantic Segmentation of Remote Sensing Images Using Siamese Affinity Network

Zheng Chen [1,2], Yuheng Lian [1], Jing Bai [1,*], Jingsen Zhang [1], Zhu Xiao [3,4] and Biao Hou [1]

[1] The Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China; 22173110645@stu.xidian.edu.cn (Z.C.); yuheng_lian@stu.xidian.edu.cn (Y.L.); 22171110630@stu.xidian.edu.cn (J.Z.); houbiao@mail.xidian.edu.cn (B.H.)
[2] China Mobile Tietong Co., Ltd., Shanxi Branch, Taiyuan 030032, China
[3] The College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China; zhxiao@hnu.edu.cn
[4] The Shenzhen Research Institute, Hunan University, Shenzhen 518055, China
* Correspondence: baijing@mail.xidian.edu.cn

Abstract: In recent years, weakly supervised semantic segmentation (WSSS) has garnered significant attention in remote sensing image analysis due to its low annotation cost. To address the issues of inaccurate and incomplete seed areas and unreliable pseudo masks in WSSS, we propose a novel WSSS method for remote sensing images based on the Siamese Affinity Network (SAN) and the Segment Anything Model (SAM). First, we design a seed enhancement module for semantic affinity, which strengthens contextual relevance in the feature map by enforcing a unified constraint principle of cross-pixel similarity, thereby capturing semantically similar regions within the image. Second, leveraging the prior notion of cross-view consistency, we employ a Siamese network to regularize the consistency of CAMs from different affine-transformed images, providing additional supervision for weakly supervised learning. Finally, we utilize the SAM segmentation model to generate semantic superpixels, expanding the original CAM seeds to more completely and accurately extract target edges, thereby improving the quality of segmentation pseudo masks. Experimental results on the large-scale remote sensing datasets DRLSD and ISPRS Vaihingen demonstrate that our method achieves segmentation performance close to that of fully supervised semantic segmentation (FSSS) methods on both datasets. Ablation studies further verify the positive optimization effect of each module on segmentation pseudo labels. Our approach exhibits superior localization accuracy and precise visualization effects across different backbone networks, achieving state-of-the-art localization performance.

Keywords: Siamese network; SAM; remote sensing images; weakly supervised semantic segmentation (WSSS)

Academic Editor: Chiman Kwan

Received: 17 December 2024
Revised: 19 February 2025
Accepted: 20 February 2025
Published: 25 February 2025

Citation: Chen, Z.; Lian, Y.; Bai, J.; Zhang, J.; Xiao, Z.; Hou, B. Weakly Supervised Semantic Segmentation of Remote Sensing Images Using Siamese Affinity Network. *Remote Sens.* **2025**, *17*, 808. https://doi.org/10.3390/rs17050808

1. Introduction

Semantic segmentation is a fundamental task in remote sensing image processing, aiming to predict pixel-level classification results. With the rapid advancements in deep learning research in recent years, the performance of semantic segmentation models has seen significant improvement [1,2], facilitating various applications of remote sensing images such as building outline extraction and road detection [3]. However, compared to other tasks like image-level classification and box-level object detection, semantic segmentation requires pixel-level category labels, which are more time-consuming and labor-intensive to

collect than image and bounding box labels. Consequently, weakly supervised semantic segmentation (WSSS) has gained widespread attention in the segmentation field.

WSSS aims to train a multi-class semantic segmentation model using weak supervisory information such as image-level classification labels [4–6], scribbles [7] and bounding boxes [8,9]. Recently, image-level weakly supervised semantic segmentation methods have become a research hotspot. These methods primarily rely on image-level labels for model training, eliminating the need for fine-grained pixel-level annotations and significantly reducing annotation costs. Compared to other forms of weak supervision, such as scribbles and bounding boxes, image-level labels are more widely available and easier to obtain. However, there exists a significant gap between image-level supervision and pixel-level dense prediction [10]. The former provides global information on the target categories present in the image, while the latter involves fine-grained classification of each pixel in the image. This discrepancy in supervision signals poses a challenge for WSSS: how to effectively predict pixel-level classification results in the absence of detailed annotations.

Researchers commonly follow a two-stage pipeline [11], as illustrated in Figure 1. In the first stage, a classification model is trained using image-level supervision to generate pseudo masks (dense labels) for each training image. In the second stage, these pseudo masks are used to train the semantic segmentation model. High-quality pseudo mask generation is crucial for image-level WSSS. Most existing methods utilize intermediate feature maps from the classification model (e.g., the Class Activation Map, CAM [12]) to locate the approximate regions of the targets, serving as seeds for generating pseudo masks.

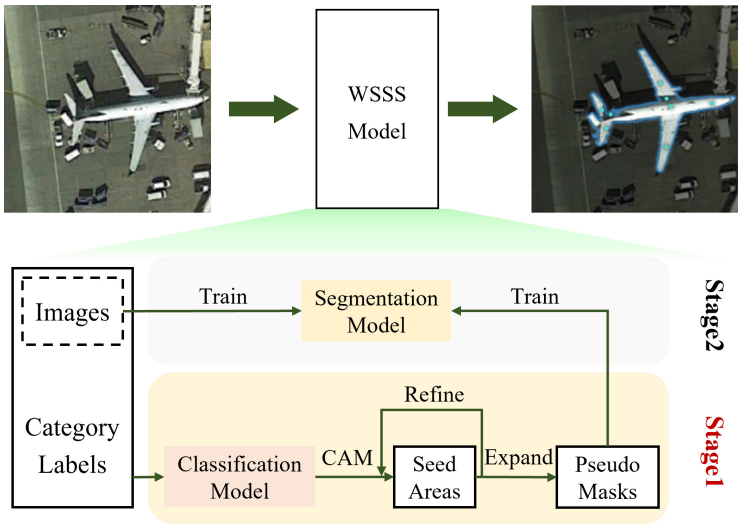

Figure 1. Image-level weakly supervised semantic segmentation (WSSS) pipeline.

Such a two-stage framework has achieved good results in the natural image domain. However, in the field of remote sensing images, WSSS research primarily focuses on single-label tasks such as building extraction [13,14], road detection [7], farm land [15] and water body segmentation [16], distinguishing only between the target and the background. In practice, the land cover and spatial distribution in remote sensing images are complex and diverse, requiring multi-class scene processing in applications like Land Use Land Cover (LULC), urban analysis, and agricultural monitoring [17]. In these scenarios, the boundaries between different land cover types are often unclear, necessitating more precise segmentation compared to natural images.

Due to the characteristics of the CAM method, the class activation maps obtained in the two-stage method usually only cover the most discriminative regions of the target, and errors in the activation of background parts of the image are inevitable, generating noise. This drawback is amplified in the WSSS task for remote sensing images, significantly

affecting the training of subsequent segmentation models [18]. Moreover, the low-level features (such as textures and patterns) in remote sensing images are overlooked in CAM, making it difficult to ensure the quality of pseudo labels. Therefore, the CAM is hard to use as an objective mask directly for supervision and usually requires additional supervisory information and constraints to bridge the gap between the two types of supervision.

To address the issues of missing supervisory information and discriminative region problems in WSSS, we draw on the idea of self-supervised learning to provide additional supervision for the training of the classification model. Self-supervised learning methods generate labels by designing pretext tasks to supervise the model in learning data distributions without extra manual annotations. There are many classic self-supervised pretext tasks, such as relative position prediction [19], spatial transformation prediction [20], image reconstruction [21], and image colorization [22]. The labels generated by pretext tasks provide self-supervision information to the network, enabling the model to learn more robust feature representations. We propose a WSSS method for remote sensing images based on a Siamese Affinity Network (SAN) and the Segment Anything Model (SAM) [23] following two constraints: cross-pixel similarity and cross-view consistency [10].

Based on the principle of cross-pixel similarity and drawing on the self-attention mechanism, we use a semantic affinity-based seed enhancement module to capture semantically similar regions in the image using contextual pixel information to enhance the CAM, alleviating the discriminative region problem. Leveraging self-supervised learning ideas, we propose a cross-view consistency Siamese network to provide additional supervision for the classification model training by supervising the CAM of images after affine transformations. Finally, we introduce the SAM [23], designing a semantic allocation strategy to further optimize the CAM seed regions. This provides high-quality labels for training a fully supervised segmentation model. Experiments conducted on two remote sensing image datasets show that our method achieves state-of-the-art segmentation performance using only image-level labels.

The main contributions of this paper are summarized as follows:

- We propose the Siamese Affinity Network (SAN), enhancing supervision quality using CAMs from affine-transformed images and a semantic affinity seed enhancement module, which improves the CAMs and mitigates the focus on discriminative regions.
- We introduce the SAM segmentation model to generate masks, which are then used to produce semantic superpixels. By designing a semantic allocation strategy, we further optimize the CAM seed regions, thereby enhancing the quality of the segmentation pseudo labels.
- We conduct extensive experiments to validate the positive optimization effects of the proposed modules on segmentation pseudo labels. The results demonstrate that our proposed method achieves superior localization accuracy and visualization effects, achieving state-of-the-art localization performance.

The remainder of the paper is organized as follows. In Section 2, we review the related work on image-level weakly supervised semantic segmentation and its current applications in remote sensing imagery. Section 3 introduces the overall framework of the proposed method, detailing its three main components. In Section 4, we conduct experiments on two types of remote sensing datasets and validate each module of the method through ablation studies. Finally, Section 5 provides the conclusion.

2. Related Works

2.1. WSSS for Natural Images

In natural images, the task of WSSS has been studied for many years. The SEC (Seed, Expand, and Constrain) [24] framework was the first to propose the basic process of image-

level WSSS, including seed generation, mask generation, and additional constraints. These methods have been adopted by many other research works.

Seed generation is usually achieved by computing the CAM [12,25] of a classification model, allowing the seed regions to cover each semantic region in the image. However, the initial seed regions obtained through the CAM are usually poorly discriminative. Existing methods either refine the seed regions or generate pseudo masks by incorporating prior constraints based on the seed regions.

Many methods have been designed to refine CAMs, including erasing overlapping regions [26], expanding incomplete regions [18,27], and optimizing CAM generation [28], among others. Reverse erasure is a popular CAM expansion method that erases the most discriminative parts of the CAM, guiding the network to learn classification features from other regions and expand the CAM. Wei et al. [27] introduced convolutional blocks with different dilated rates to generate dense and reliable object localization maps. IRNet [29] generates a transition matrix from boundary activation maps and extends this method to weakly supervised instance segmentation.

The generation of pseudo masks is achieved by propagating the semantic information of seed regions throughout the image. Common approaches include using random walk [30,31] to diffuse the boundaries of the seed regions. Ahn and Kolesnikov [32] trained a network to predict pixel affinity relationships by learning the similarity between neighboring pixels. They used the generated affinity matrix to multiply with the CAM, adjusting the coverage of its activation, and then performed a random walk on the CAM to obtain pseudo labels. Inspired by Kolesnikov, Huang et al. [33] adopted a Seeded Region Growing (SRG) strategy [29] to generate pseudo masks. Some studies also focus on integrating self-attention modules into the WSSS framework [30,34]. For instance, the work that designed CIAN [35] proposed a cross-image attention module that, guided by class activation maps, learns class activation maps from different images containing the same class of objects. In our method, we also utilize the self-attention mechanism, but unlike CIAN, we refine the CAM of a single image internally within the model.

2.2. WSSS for Remote Sensing Images

An increasing number of studies are applying weakly supervised semantic segmentation (WSSS) methods, originally used for natural images, to remote sensing imagery, focusing mainly on two areas: region extraction and multi-class semantic segmentation.

Research on region extraction primarily focuses on detecting and segmenting specific geographic objects in remote sensing images, such as roads, farm land, water bodies, and buildings. Early studies achieved segmentation by binarizing the CAM generated by classification networks [15]. To improve segmentation accuracy and mitigate boundary blurring issues, Cao et al. proposed a coarse-to-fine weakly supervised segmentation method (CFWS) [36]. Ali et al. [37] introduced an attention network to extract destructed patches under the constraint of sparsity loss and optimized the segmentation boundaries by combining Conditional Random Fields (CRFs) [38]. Some studies utilized superpixel methods to generate dense labels for training segmentation networks, which proved effective in the extraction of water bodies [39] and buildings [14]. For specific target extraction, ref. [40] proposed a global-to-local dynamic information enhancement (G2LDEE) framework, which improves pseudo-label accuracy by supplementing global information with local details and refining building boundaries using a Segment Anything Model (SAM) post-processing method. Similarly, ref. [41] introduced APSAM, a weakly supervised landslide extraction method that uses adaptive prompt engineering to generate fine-grained segmentation masks directly from the SAM, improving the boundary precision for the extraction of the target landslide without relying on high-quality CAMs.

Some work also focuses on change detection in target areas, which is applied in fields such as disaster management and urban management. For instance, ref. [42] proposed a novel approach called Knowledge Distillation and Multi-scale Sigmoid Inference (KD-MSI), which utilizes image-level labels and CAMs to enhance the accuracy of change detection while reducing the reliance on pixel-level annotations.Similarly, Zhao et al. [43] introduced a weakly supervised change detection framework (WSLCD) using a double-branch Siamese network. Their approach enforces cross-view consistency through mutual learning and equivariant regularization (MLER), combined with prototype-based contrastive learning (PCL), to improve pseudo-label quality and achieve state-of-the-art performance on multiple datasets.

Fewer WSSS studies have been applied to multi-class semantic segmentation. Zhou et al. [17] proposed a self-supervised twin network based on an explicit pixel-level constraint framework, significantly enhancing the quality and localization accuracy of class activation maps in multi-class remote sensing scenarios. Zeng et al. proposed a novel framework replacing the CAM with a saliency map generator (SMG) for more accurate localization and segmentation in remote sensing images [44]. Li et al. [45] introduced an integrated framework that improved segmentation performance through uncertainty-driven pixel-level weighted masks. Ref. [46] proposed a contrast token and foreground activation (CTFA) framework based on the ViT architecture, which improves CAM generation using contrast token learning module and a dual-branch decoder.

Weakly supervised methods have also been applied to change detection and other remote sensing tasks.

3. Proposed Method

The proposed framework for weakly supervised semantic segmentation of remote sensing images, based on a SAN and the SAM, is shown in Figure 2 and consists of three main components.

The first component (blue path in Figure 2) involves feature extraction from the input image using a classification network, generating a feature graph embedding. This embedding is used to produce a multi-label prediction via pooling in the GAP layer, followed by a soft-marginal classification loss calculation between the predicted and ground truth labels. Simultaneously, the CAM seed is obtained by weighting the feature map based on the GAP weights. To enhance the CAM seed, a self-attention mechanism computes semantic pixel affinity weights on the feature graph embedding, optimizing the CAM seed. The optimized CAM seed maps are upsampled to the original image size, yielding pseudo labels by selecting the maximum response across all CAM categories. Further refinement of the pseudo labels is performed using Conditional Random Fields (CRFs) to improve boundary accuracy.

The second component (gray path in Figure 2) introduces a cross-view consistency twin network to refine the CAM seed map using additional supervision. Inspired by self-supervised learning, the original image is affine transformed and passed through a twin network with shared parameters. The transformed CAM seed map is then inverse transformed, and an isovariant canonical loss is computed between the original and transformed CAM seeds to enforce consistency, providing additional supervision.

The third component utilizes the SAM to enhance the generated pseudo labels with external knowledge. The input image is processed by the trained SAM model, which produces segmentation masks based on grid point cues. Since the initial SAM masks are incomplete and may overlap, the pseudo labels are further refined by semantically assigning labels using mask filtering and category assignment.

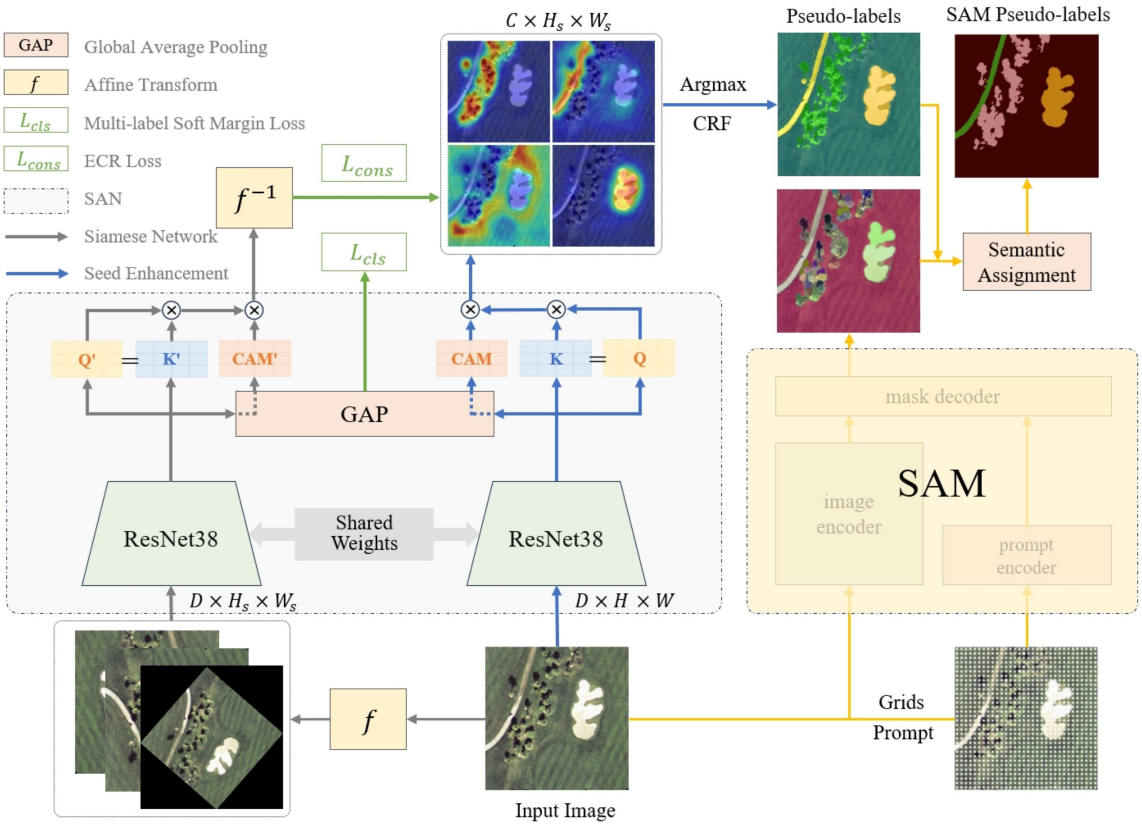

Figure 2. Overview of our Siamese affinity and SAM-based framework for WSSS of remote sensing images.

After these steps, the framework produces high-quality segmentation pseudo labels from multi-category image-level labels. These pseudo labels are then used to train the segmentation model, completing the weakly supervised semantic segmentation task.

3.1. Semantic Affinity-Based Seed Enhancement Module

The self-attention mechanism is a widely recognized attention model that effectively improves the representation of networks. The mechanism adjusts the representation of feature maps by capturing contextual correlations in the feature maps and weighting the features to follow the principle of uniform constraints on cross-pixel similarity, where pixel points of regions with the same semantics in the same image are characterized by a high degree of similarity, and thus, semantically similar regions can be captured by self-attention. This is in line with the idea that many WSSS methods utilize pixel similarity to improve the original CAM map. In general, the self-attention mechanism can be described as follows:

$$\begin{aligned} q_i = \theta_q(x_i), k_i = \theta_k(x_i), v_i = \theta_v(x_i) \\ y_i = \sum_j w_{q_i,k_j} v_j + q_i, w_{q_i,k_j} = \text{Softmax}\left(\frac{q_i^T k_j}{\sqrt{d_k}}\right) \end{aligned} \tag{1}$$

where x_i and y_i represent the input and output features at spatial position i. q_i, k_i, and v_i represent the query, key, and value vectors at position i. These three vectors are mapped using functions $\theta_q(\cdot)$, $\theta_k(\cdot)$, and $\theta_v(\cdot)$, which are typically implemented through a 1×1 convolutional layer. w_{q_i,k_j} is the attention weight between position q_i and k_j, which is calculated by taking the inner product of the query and key vectors and applying Softmax. Finally, y_i represents the output feature at position i, which is the sum of the weighted values and the original query vector.

In the framework of this paper, to improve the network's ability to predict pixel-level correlations, the classical self-attention module (given by Equation (1)) is integrated into the optimization of the CAM:

$$\hat{cam}_i = \frac{1}{\sum_{\forall j} w(q_i, k_j)} \sum_{\forall j} w(q_i, k_j)\theta_v(cam_i) + cam_i \tag{2}$$

where cam represents the original CAM, and $\hat{cam}$ represents the optimized CAM. In this structure, the original CAM is spatially mapped using the function θ_v. The three embedding functions θ_q, θ_k, and θ_v are still implemented through individual 1×1 convolutional layers.

The structural parameters used in this paper refer to the core parts of the attention mechanism and make some modifications to meet the task characteristics of WSSS. Here, we use the cosine distance to evaluate the feature similarity (affinity) between image elements:

$$w(q_i, k_j) = \frac{\theta_q(x_i)^T \theta_k(x_j)}{||\theta_q(x_i)|| \cdot ||\theta_k(x_j)||}, \theta_q = \theta_k \tag{3}$$

where the inner product in the normalized feature space is used to calculate the affinity between the current pixel i and other pixels. With some modifications to Equation (2), the final seed enhancement module based on semantic affinity is shown as follows:

$$\hat{cam}_i = \frac{1}{\sum_{\forall j} w(q_i, k_j)} \sum_{\forall j} \mathrm{ReLU}\big(w(q_i, k_j)\big) \cdot cam_i \tag{4}$$

Affinity is activated by a ReLU to suppress negative values, thereby screening out irrelevant pixel interference. The final CAM seed map is the weighted sum of the original CAM seed map and the normalized affinity. Compared to traditional self-attention, the residual connection is removed here to maintain the activation strength of the original CAM seed map. The embedded function θ_v is removed, and θ_k and θ_q are kept to reduce the parameters and computational interference caused by affinity calculation.

3.2. Cross-View Consistency Siamese Network

In self-supervised learning, the consistency of features embedded in the latent space through different augmented images can provide additional supervision for classification tasks. This clever approach compensates for the lack of additional supervision in weakly supervised semantic segmentation, echoing the principle of cross-view consistency.

In FSSS, the segmentation function $F_{ws}(\cdot)$ exhibits equivariance under spatial transformations $Aff(\cdot)$, meaning that applying a transformation to the input image and then segmenting it is equivalent to segmenting the original image and then applying the transformation to the output mask. This property can be expressed as follows: $F_{ws}(Aff(X)) = Aff(F_{ws}(X))$ However, in WSSS, where only image-level labels are available, the task relies on CAMs generated by a classification model $F_{wc}(\cdot)$. Unlike segmentation functions, classification models are designed to be invariant to spatial transformations, meaning that the predicted class label remains unchanged regardless of the transformation applied to the input image: $F_{wc}(Aff(X)) = F_{wc}(X) = Cls$ This invariance is achieved through pooling operations, which discard spatial information. As a result, the CAMs generated by F_{wc} lack the equivariance property needed for accurate segmentation. To bridge this gap between FSSS and WSSS, we introduce additional equivariance constraints by enforcing consistency between the CAMs of the original image and its transformed versions. This ensures that the model learns robust features that generalize well to different spatial transformations, improving the quality of the generated CAMs.

In the data augmentation phase of fully supervised semantic segmentation, pixel-level labels are subjected to the same transformations as the input images. This introduces implicit equivariance into the network. However, considering that WSSS only has image-level class labels, it cannot introduce such equivariance. Inspired by the self-supervised learning structure, in WSSS tasks, the corresponding CAM can be conveniently subjected to the same transformations to introduce equivariance. That is to say, the CAM of the image after transformation should correspond to the transformation of the CAM of the original image.

Therefore, we enforce consistency between the two CAMs to introduce an equivariance regularization for model learning:

$$\mathcal{L}_{\text{cons}} = \|Aff^{-1}(M(Aff(X))) - M(X)\|_1 \tag{5}$$

where $M(X)$ represents the CAM of X, and $Aff(\cdot)$ represents any arbitrary transformation such as scaling, rotation, or flipping. To integrate this regularization into the original network, similar to the self-supervised learning framework, the network is expanded to share the parameters of the backbone structure. Before forwarding through the backbone network, the image undergoes transformations, and the CAM points are then inversely transformed accordingly. Finally, the outputs of the original branch and the backbone branch are calculated, and the equivariance regularization loss is minimized to ensure the consistency of the CAMs.

At the same time, during the model training process, it is indispensable to have supervision from classification loss, considering the multi-class segmentation scenario. Here, a multi-label soft margin loss is used. For a foreground target with C classes, the multi-label soft margin loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{cls}} = -\frac{1}{C-1}\sum_{c=0}^{C}\Bigg[l_c \log\left(\frac{1}{1+e^{-p_c}}\right) + \\ (1-l_c)\log\left(\frac{e^{-p_c}}{1+e^{-p_c}}\right)\Bigg]\end{aligned} \tag{6}$$

where l_c and p_c represent the true one-hot labels and the class probability vectors output by the model after global average pooling, respectively. During the validation process, both the original branch and the backbone branch participate in the calculation of classification loss. Finally, the model loss is obtained by combining the classification loss and the equivariance regularization loss:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{cons}} \tag{7}$$

3.3. SAM-Based Pseudo-Label Expansion Module

The Segment Anything Model (SAM) [23] is a foundational image segmentation model recently released by Meta. Trained on over 1 billion masks across 11 million images, the SAM boasts impressive generalization capabilities and can be readily applied to a range of downstream vision tasks. Recently, the SAM has also been employed in some studies to enhance WSSS tasks. For instance, Chen et al. improved CAM pseudo labels by calculating overlap rates using SAM-generated segmentation results [47]. Jiang et al. used local maxima points on their CAM as prompts for the SAM to generate pseudo-segmentation labels [48]. Sun et al. utilized GroundedDINO [49] to generate bounding box prompts, which were then fed into the SAM to produce segmentation seeds [50]. In contrast, our approach introduces a seed expansion module based on the SAM, which sequentially filters masks covering the entire object, partial areas, and sub-regions. This module uses

SAM-generated masks to create semantic superpixels, thereby generating more complete and accurate segmentation seeds.

The seed expansion module based on the SAM takes the CAM as input to expand and produce seed maps. This is primarily achieved through the following three steps: semantic superpixel generation, semantic superpixel classification, and seed map generation.

3.3.1. SAM Superpixel Generation

When prompts use regular grid points, the SAM generates over 50 clear-boundary masks for each image. However, these masks lack semantic labels and many are overlapping, with some covering the entire object, while others only encompass parts or sub-regions of the object (regions with complete semantic concepts). To generate complete and accurate segmentation seed maps from these masks, it is essential to filter out a set of appropriate masks, which are referred to as semantic superpixels. Priority is given to semantic superpixels that cover the entire object. Additionally, similar to standard non-overlapping superpixels, the objective here is to minimize the overlap of semantic superpixels as much as possible.

To achieve this, the SAM mask generation process is modified as follows: (1) A lower confidence threshold $t_{m,whole}$ is set to retrieve more masks at the whole-object level. (2) In the Non-Maximum Suppression (NMS) process, masks at the whole-object level are prioritized to reduce the suppression of whole-object masks by masks covering partial or sub-regions. (3) A simple filter is designed to further filter masks post-NMS, reducing overlap. The filter processes masks in the order of whole-object, partial region, and sub-region. This means that masks covering the entire object are retained first, followed by masks of partial regions, and finally sub-region masks. A newly generated mask is retained only if its overlap rate with previous masks (i.e., the ratio of the overlapping area to the area of the new mask) is below a threshold t_r.

3.3.2. Semantic Superpixel Classification

The goal of semantic superpixel classification is to use the refined CAM to determine the semantic category of each semantic superpixel. The CAM for the foreground class c is defined as $M^{c*} \in \mathbb{R}^{H \times W}$, with each semantic superpixel Q^i represented by a binary mask of size $H \times W$. The probability Q^i belongs to the semantic category c and is calculated by averaging the activation weights within the semantic superpixel and then normalizing across all semantic superpixels. This can be expressed as follows:

$$S^c(Q^i) = Norm\left(\frac{1}{|Q^i|}\sum_{u=1}^{H}\sum_{v=1}^{W} Q^i_{u,v} \cdot M^{c*}_{u,v}\right) \tag{8}$$

Here, Norm$(\cdot)$ refers to min-max normalization, and the probability that a semantic superpixel is classified as background is calculated using the formula below:

$$S^{bg}(Q^i) = (1 - \max_{c \in P}(S^c(Q^i)))^{\alpha} \tag{9}$$

In this formula, α is an empirically determined hyperparameter. A larger α suppresses the background and produces more foreground. Then, the semantic category of semantic superpixel Q^i is determined by the following equation:

$$l_i = \underset{j \in P \cup bg}{\operatorname{argmax}}\left(S^j(Q^i)\right) \tag{10}$$

In specific experiments, the selected datasets do not include a background category. This formula is provided to enhance generality.

3.3.3. Seed Map Generation

Finally, the class label of each semantic superpixel is assigned to the pixels it contains, thus generating the seed map $D \in \mathbb{R}^{H \times W}$. If a pixel is covered by two or more semantic superpixels, it is assigned the class label of the superpixel with the higher score. If the scores are the same, it is assigned the class of the superpixel with the higher classification score. Pixels not covered by any superpixel are considered as the background. This seed generation method, based on preferential seeds, allows for more comprehensive detection of the target pixels and more accurate assignment of th semantic labels.

4. Experiment and Analysis

4.1. Datasets

In order to verify the effectiveness of the method proposed in this paper, experiments were conducted on two publicly available datasets of different scales.

The Dense Label Remote Sensing Dataset (DLRSD) [51] is a densely annotated dataset used for multi-label tasks such as remote sensing image retrieval (RSIR) and classification, as well as semantic segmentation tasks. The DLRSD dataset is an extension of the UC Merced dataset, sourced from the United States Geological Survey, and includes regions such as Fairbanks, Boston, and Rome. The DLRSD comprises images of 21 major categories, with 100 images per category. Each image is a remote sensing orthophoto with a resolution of 0.3 m and dimensions of 255 × 255 pixels. The DLRSD provides pixel-level annotations for each image in the UC Merced dataset, covering the following 17 categories: airplane, bare soil, buildings, cars, chaparral, court, dock, field, grass, mobile home, pavement, sand, sea, ship, tanks, trees, and water. In the experiments, the dataset was randomly split into training and testing sets in a 7:3 ratio. The training set contained 1470 images, while the testing set contained 630 images. Figure 3 shows some examples from the DLRSD with corresponding pixel-level annotations.

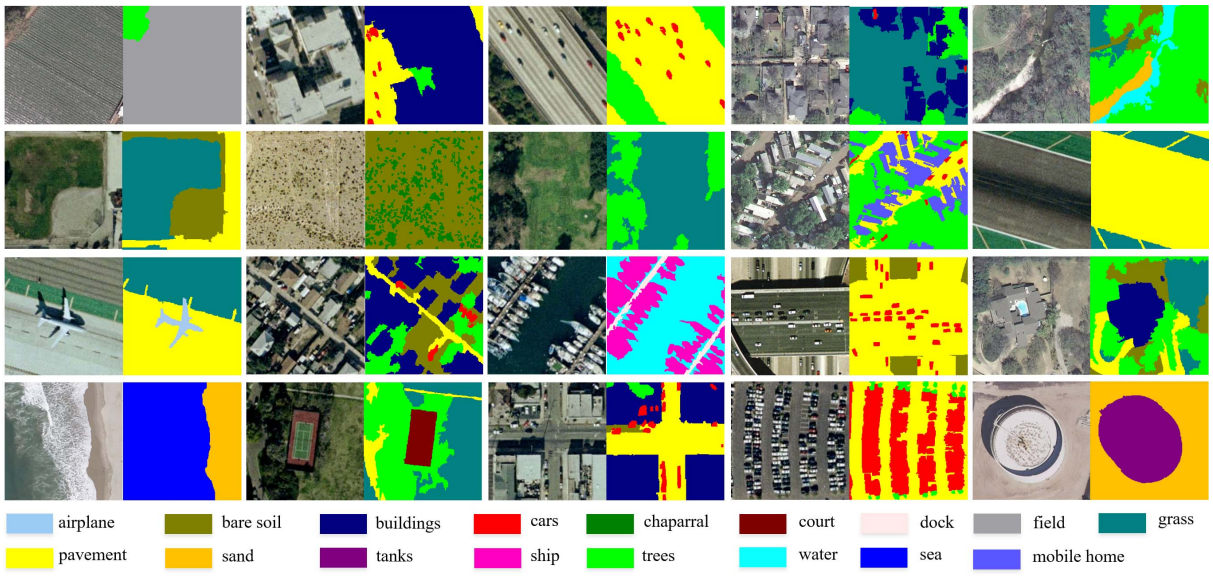

Figure 3. Examples of selected images of the DLRSD dataset and their annotations.

The ISPRS Vaihingen dataset is commonly used for semantic segmentation tasks of high-resolution remote sensing images, containing 33 images with an average size of 2494 × 2064 pixels in IRRG (infrared, red, and green) bands and a spatial resolution of 9 cm. The ISPRS provides not only high-resolution orthophoto but also standardized Digital Surface Models (DSMs) through corresponding dense image matching techniques. This dataset includes pixel-level annotations for five land cover types: impervious surface (imp.

surf), building, low vegetation (low veg.), tree, and car. According to the official data split, 16 images were used for training and 17 images for testing. During the training phase, these images were cropped to 128 × 128 pixel patches with a 64-pixel overlap between adjacent patches for data augmentation. In the testing phase, a non-overlapping patch approach was adopted. This resulted in a total of 17,336 training patches and 5074 testing patches. Figure 4 shows examples and annotation results from the Vaihingen dataset.

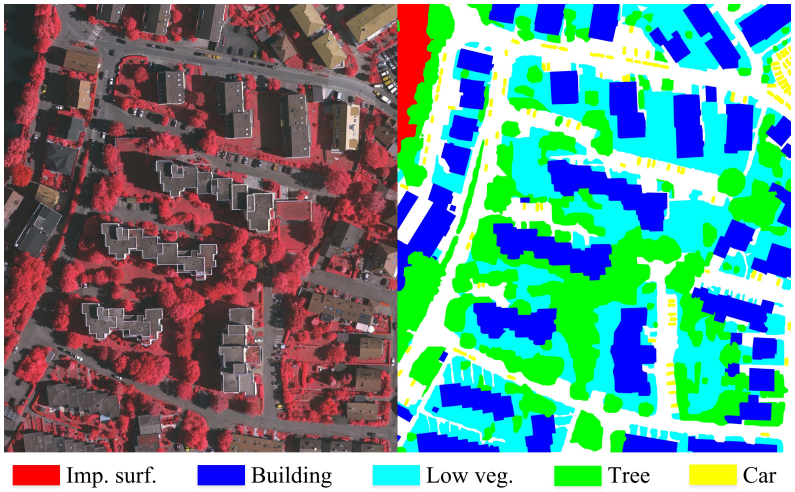

Figure 4. Examples of selected images from the ISPRS Vaihingen dataset (false color composite: near-infrared bands, red, green) and their annotations.

4.2. Metrics

Based on the aforementioned datasets, the mean Intersection over Union (mIoU), F1 score, and Overall Accuracy (OA) metrics were used to quantitatively evaluate the proposed model. In the domain of recognition, True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) were used to express these three metrics. TP refers to the number of samples that are correctly identified by the model as positive when they are indeed positive. FP refers to the number of samples that are incorrectly identified by the model as positive when they are actually negative, also known as the "false positive rate". TN refers to the number of samples that are correctly identified by the model as negative when they are indeed negative. FN refers to the number of samples that are incorrectly identified by the model as negative when they are actually positive, also known as the "false negative rate". These metrics are crucial in binary and multi-class problems for measuring model performance, especially when evaluating classifier accuracy.

The mIoU is a common metric for measuring the accuracy of each class's segmentation, particularly in image semantic segmentation tasks. It calculates the average value of the intersection over union for each class, representing the degree of overlap between the predicted result and the actual annotated region. The formula is defined as follows:

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c} \tag{11}$$

Among them, C is the total number of classes. The F1 score is a measure that considers both Precision and Recall. For a single class, the F1 score is defined as follows:

$$\text{F1}_c = 2 \times \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{12}$$

For multiple classes, the average F1 score (Macro F1 or Weighted F1) is the average of the F1 scores of all classes: $F1 = \frac{1}{C}\sum_{c=1}^{C} F1_c$. where

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{13}$$

The OA is the most straightforward metric of classification accuracy, reflecting the proportion of correctly classified samples out of the total samples.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

In multi-class tasks, the Precision, Recall, and F1 for each class need to be calculated independently, followed by computing the overall F1 score using either a macro average or a weighted average. The aforementioned metrics are the fundamental and general standards for evaluating the semantic segmentation process. By using these metrics, the model's ability to recognize land cover types on the DLRSD and Vaihingen datasets can be comprehensively evaluated.

4.3. Implementation Details

In the experiments in this paper, based on the usual experience in the field of WSSS, ResNet38 [32] was chosen as the backbone feature extraction network, and a global average pooling layer was used to replace the last fully connected layer in the model, while ImageNet was utilized for pre-training the weights in order to strengthen the model's feature extraction capability. Affine transformations such as random zoom rotation, random flip, and random crop were used in the Siamese network. Image enhancement additionally uses a color dithering strategy to randomly change the image brightness, contrast, saturation, and hue. During the training process of the multi-classification model, only the training set was used for extreme image-level labeling, the batch size was set to 16, the initial learning rate was set to 0.01, and the learning rate during the training process was controlled using a polynomial learning rate decay algorithm, with the momentum set to 0.9 and the weight decay coefficients set to 0.0005. Ultimately, the final round of classification model was used for the inference to compute the CAM seed maps. The comparison methods for both datasets are mainstream methods that have been published and validated in remotely sensed imagery, but the comparison methods are slightly different due to the different scales and scenarios of the two datasets. For the SAM model, the officially published sam_{vit_h} model was used to segment the images directly, and the prompt was chosen in the form of grid points.

For the segmentation network of stage 2 in Figure 1, the DeepLabV3+ [52] model was selected, ResNet50 [53] was used as the backbone network, and a random flip and random cropping strategy was used. The batch size was set to 4, with a total of 96,000 iterations, the model was saved every 8000 iterations, and the best validation set result among the 12 models was selected to record the metrics. The training of the segmentation model was based on the MMSegmentation repository in MMlab labs, and other configuration parameters were kept consistent with their given profiles.

All classification and segmentation models in this paper were trained in NVIDIA GeForce RTX 3060 (12 GB) and RTX 1080 (8 GB) GPUs, SAM inference was performed in the NVIDIA GeForce RTX3090 (24 GB) GPUs, and all experiments were based on the PyTorch implementation (version 2.2.1).

4.4. Experimental Results

In order to evaluate the performance of the proposed method in this paper, experiments were conducted here on two remote sensing image datasets of different sizes and resolutions. Firstly, the performance of the mainstream WSSS method and the proposed method in this paper were compared with the fully supervised algorithm under different metrics under the two datasets. Then, the effectiveness of each module of the proposed method for pseudo-label generation was demonstrated through ablation experiments.

4.4.1. DLRSD

Seven methods, MIL-FCN [54], Grad-CAM [25], SEC [24], DSRG [55], PSA [32], SEAM [56], and SMG [44], were compared under the DLRSD dataset, where SMG is the SOTA WSSS method published on this dataset. As shown in Table 1, in the case of using only image-level labels, MIL-FCN performed the worst among all the compared methods, with an mIoU of just under 10%. Three methods, Grad-CAM, SEC, and DSRG, all achieved an mIoU of around 33%, which is at the third tier level. The PSA and SEAM methods, on the other hand, with their affinity network design and regular constraints, both achieved more than 39% mIoU, which is at the second echelon water level. The published SOTA method SMG was significantly ahead of the previous WSSS method designed for natural images in all metrics, with an mIoU of 45.58%, exceeding the second tier level by 5.6%. With the Siamese network and affinity module, the proposed method in this paper reached the highest mIoU of 49.13%, which is also attributed to the finer segmentation edges brought by the SAM seed enhancement module. The five metrics of mIoU, OA, mPrecision, mRecall, and mFscore exceeded the SOTA method by 3.35%, 7.24%, 7.8%, 2.17%, and 0.84%, respectively. The method in this paper achieved the optimal weakly supervised segmentation effect at the image level.

Table 1. Overall segmentation performance of different WSSS methods on DLRSD dataset.

Supervision	Method	mIoU	OA	mPrecision	mRecall	mFscore
Image-level WSSS	MIL-FCN	3.23	13.49	9.57	7.01	11.08
	Grad-CAM	33.31	52.18	53.99	55.55	46.54
	SEC	33.05	54.59	48.03	51.52	48.43
	DSRG	33.94	58.96	56.07	50.24	48.44
	PSA	38.92	62.73	59.63	53.38	52.56
	SEAM	30.56	50.49	45.99	60.26	44.4
	SMG	45.58	64.28	58.71	62.43	63.07
	Ours	**49.13**	**71.52**	**66.51**	**64.60**	**63.91**
FSSS	DeepLabv3+	70.12	82.90	83.42	81.29	81.58

The bold values indicate the highest performance in each column among Image-level WSSS.

At the same time, this section also compares the differences between the proposed image-level weak supervision method and the full supervision method. Here, the DeepLabv3+ model [52], which was trained with all the training set dense labels, was used. The final model under full supervision achieved a 70.12% mIoU, which far exceeded the weak supervision methods. However, the method proposed in this article achieved a 70.07% mIoU and an 86.27% OA, which is the smallest gap compared to the full supervision method.

In addition, the IoU metrics for the 17 semantic classes in the DLRSD dataset are compared here, as shown in Table 2. It can be seen that the method in this article main-

tained leading performance in most categories, with the segmentation performance of the SOTA method SMG being the worst in some classes. However, in the airplane category, all methods performed poorly, with the best DSRG method only achieving a 15.31% mIoU. In further analysis in Section 4.5, it was found that in the airplane category images, there is a strong co-occurrence phenomenon between airplane and ground objects (airplane appear simultaneously with ground objects). This led to the model being unable to distinguish between the two classes, with attention equally distributed between airplane and ground objects, resulting in most airplane pixels being misclassified as ground objects. This labeling error caused all WSSS methods to fail in accurately segmenting the airplane category. In comparison, the full supervision method had relatively better segmentation performance for this category. Therefore, in the multi-class segmentation task, the co-occurrence phenomenon was found to significantly impact the attention of the segmentation model.

Table 2. IoU metrics for each category on the DLRSD dataset with different WSSS methods.

Category	Image-Level WSSS								FSSS
	MIL-FCN	Grad-CAM	SEC	DSRG	PSA	SEAM	SMG	Ours	DeepLabv3+
airplane	0	14.61	11.47	**15.31**	10.21	7.35	0.08	9.54	76.66
bare soil	6.59	3.19	30.25	24.03	28.05	12.26	**36.81**	33.88	42.38
buildings	2.89	6.33	38.25	33.82	48.13	40.54	40.56	**57.98**	75.09
cars	1.49	17.60	6.98	0.01	15.92	29.15	16.07	**45.46**	75.98
chaparral	0	42.73	21.68	24.47	40.82	1.91	**45.22**	38.91	47.78
court	0	54.38	0	58.12	40.88	30.26	**69.65**	53.46	78.08
dock	0	17.36	11.07	0	13.81	18.96	0	**19.61**	54.33
field	17.85	93.47	97.13	96.88	91.11	55.9	**98.88**	81.29	97.93
grass	4.93	37.58	35.33	39.81	39.98	32.28	50.63	**54.08**	66.02
mobile home	0	34.66	25.41	8.56	2.6	36.99	24.55	**40.91**	63.15
pavement	12.00	43.87	39.89	46.29	50.21	35.96	50.06	**64.59**	82.11
sand	0.85	26.55	26.75	28.95	37.64	23.46	40.98	**46.73**	57.33
sea	0	36.94	60.86	53.15	56.05	56.97	58.38	**65.74**	88.68
ship	0	8.77	15.46	0.16	22.16	11.55	**52.07**	46.55	79.05
tanks Tank	0	52.61	48.33	44.33	45.75	36.63	**51.32**	46.29	57.65
trees	6.15	44.45	56.63	55.06	63.66	47.36	59.01	**64.66**	74.06
water	2.24	31.12	36.44	48.06	54.56	42.06	**72.90**	65.58	77.69

The bold values indicate the highest IoU for each category among Image-level WSSS methods.

Meanwhile, in Figure 5, the segmentation results of the proposed model on the DLRSD dataset's 21 category scenes are visualized. It can be seen that the pixel segmentation results of the method in this article are basically consistent with the ground truth labels. On a large scale, the model's recognition effect was relatively good in categories with small target objects like cars, but it could not distinguish in many scenes, reflecting the situation of low classification in some categories, as shown in Table 2. Especially for the airplane category, due to the mixing with ground object classes, only a fuzzy segmentation of the airplane location could be achieved without identifying clear edges. This is a common problem faced by WSSS methods. Overall, in the DLRSD small scene dataset, the proposed method can accurately identify classes, most target boundaries are clear, complex scene segmentation boundaries are blurred, and the overall visual effect is clear.

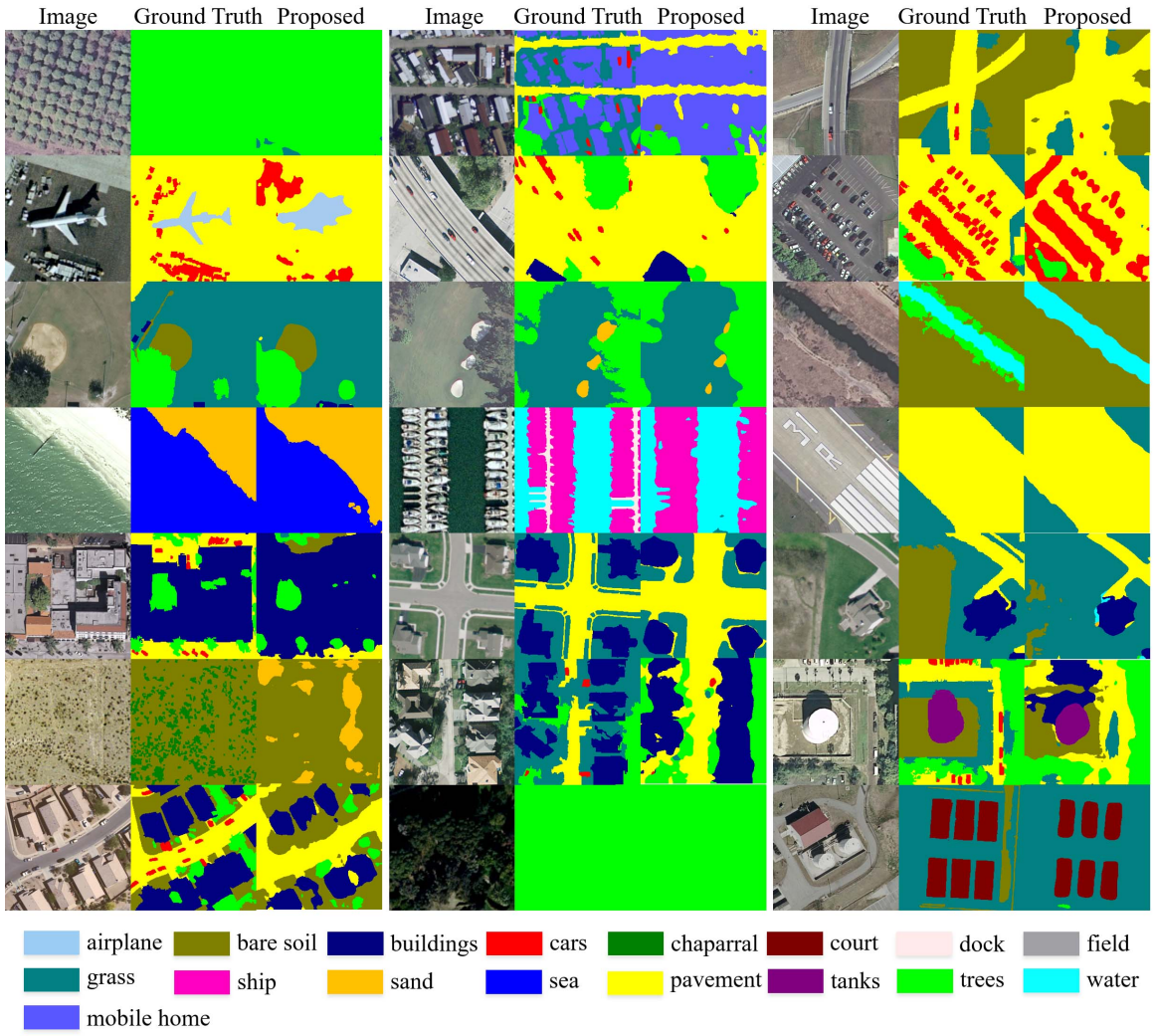

Figure 5. Segmentation visualization results of the proposed method on the DLRSD dataset.

In addition to quantitative metrics, we also visualized the segmentation results on this dataset, as shown in Figure 6. It should be noted that due to the high resolution of the images in this dataset, the results presented here were initially segmented with a 128 × 128 input size, and the final results were obtained by stitching all images together. It can be seen that the segmentation of small objects was still not very good, and some category misclassifications occurred. For larger objects, the edges of buildings and roads can be clearly seen. Overall, there is still a lot of room for improvement.

4.4.2. ISPRS Vaihingen

In addition to experiments conducted on small datasets such as DLRSD, we also compared the segmentation performance of our method with other mainstream WSSS methods on the large-scale ISPRS Vaihingen remote sensing dataset. To ensure a fair and comprehensive evaluation, we selected five representative baseline methods that are particularly suited for the unique characteristics of the Vaihingen dataset, which include complex land cover types such as dense vegetation and intricate urban structures. These methods include PSA [32], IRNet [29], SEAM [56], SC-CAM [57], and S2EPC [17]. The segmentation results of each category and the overall results on this dataset are shown in Table 3. It can be seen that, compared to mainstream methods, our method achieved the best results in terms of both the mIoU of individual categories and the overall segmentation metrics. Compared to the best-performing S2EPC method in the current dataset, our method improved the mIoU,

OA, and mFscore by 18.03%, 15.16%, and 20.74%, respectively. At the same time, compared to the fully supervised DeepLabv3+, our method achieved 85.74%, 93.09%, and 90.85% in the mIoU, OA, and mFscore, respectively. The weakly supervised localization performance was already very close to the fully supervised performance. In terms of the performance across various categories, the Vaihingen dataset was similar to the DLRSD dataset, where all WSSS methods performed poorly on small object localization, particularly for the car category, but performed well on large-scale scene segmentation.

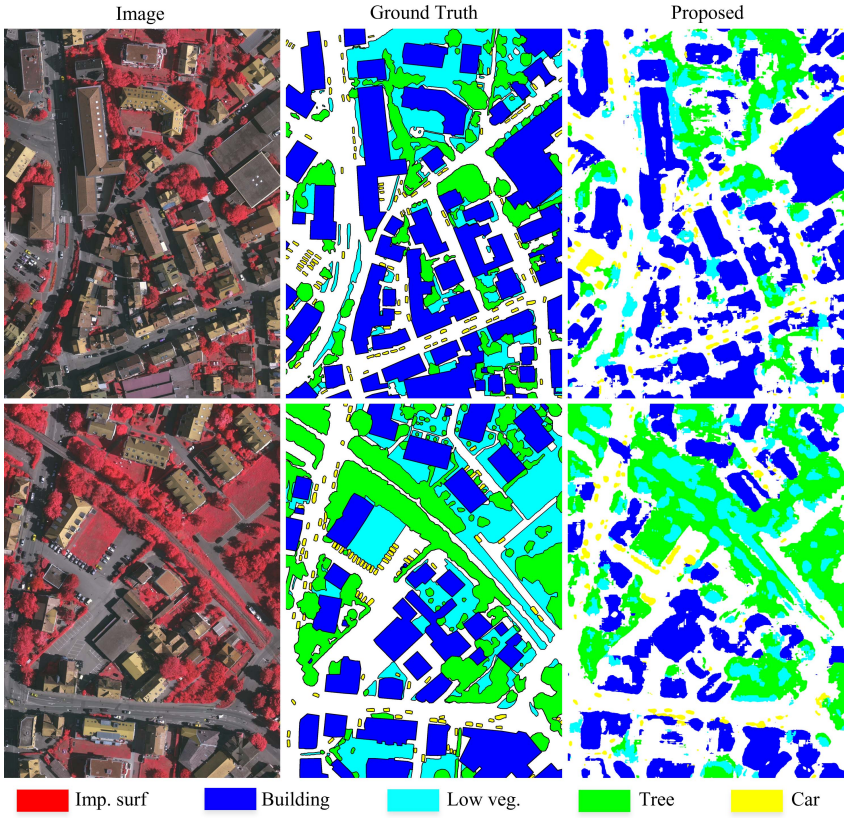

Figure 6. Segmentation visualization results of the proposed method on the Vaihingen dataset.

Table 3. IoU metrics and overall segmentation metrics for each category on Vaihingen dataset with different WSSS methods.

Supervision	Method	Imp. surf.	Building	Low veg.	Tree	Car	mIoU	OA	mFscore
Image-level WSSS	PSA	21.45	36.30	28.42	24.91	4.55	23.13	41.06	33.03
	IRNet	33.44	43.79	32.55	45.19	6.83	32.36	53.40	41.65
	SEAM	35.89	43.71	22.79	51.06	12.12	34.97	53.85	45.87
	SC-CAM	44.03	49.80	38.26	49.13	16.92	41.38	54.34	53.49
	S2EPC	53.86	53.43	44.40	55.79	19.66	45.43	65.57	56.33
	Ours	**70.93**	**76.88**	**53.25**	**69.44**	**47.05**	**63.51**	**80.73**	**77.07**
FSSS	DeepLabv3+	79.76	86.54	64.16	74.83	65.1	74.08	86.72	84.83

The bold values indicate the highest IoU for each category among Image-level WSSS methods.

4.5. Ablation Study and Pseudo-Label Quality Assessment

The method proposed in this paper is essentially a technique for extracting pseudo labels from multi-class models in segmentation. Therefore, in this section, we evaluate the pseudo labels generated by this method at each step on the DLRSD dataset to verify

the effectiveness of each module. As shown in Figure 7, we visualized the CAM and CRF pseudo labels generated at each stage of the model. Specifically, it includes the CAM seed map extracted by the model, the CAM seed map refined by the CRF, the initial CRF pseudo labels, the final pseudo labels after SAM processing, and the final pseudo labels merged with the SAM segmentation map.

It can be seen that the CAM obtained from the model have clear category discrimination. The activation regions of each class are distinct (except for UAV and ground class), and the boundaries of the segmentation regions are clear, providing a good seed region for subsequent steps. Taking the maximum activation index of CAM for each category as the initial pseudo masks and expanding them using the classical CRF method, the resulting initial CRF pseudo masks clearly segment the category boundaries. However, some target boundaries still contain mixed and noisy images. To further refine these noisy images on the boundaries, we infer using a large number of parameters of the SAM. Although the SAM can clearly segment most class boundaries, many overlapping regions exist where multiple classes may have been segmented into multiple fragments. Additionally, these segmentation results cannot be directly used for training a semantic segmentation model due to the noise and overlapping regions.

Finally, the method proposed in this paper combines the clear boundary segmentation of SAM and the accurate initial pseudo masks of the CRF. By merging the two, we obtain the final pseudo labels. The final pseudo labels have clear target boundaries, avoiding the issues of overlap and ambiguity, and provide high-quality pseudo labels for segmentation models.

For the effectiveness of the Siamese network and SAM expansion modules in the methodology of this paper, this subsection evaluates the quality of real labels and generated pseudo labels on the DLRSD training set, it uses these pseudo labels to train the segmentation model at the same time, and it evaluates the generalization performance of the corresponding segmentation model on the test set. The ablation results are shown in Table 4, where it is clear that the model performance was the worst before using Siamese networks and multi-scale inference, with pseudo-labeling and test set segmentation mIoUs of only 44.33% and 46.35%. After gradually adding the Siamese network, CRF, and SAM expansion module, the pseudo-labeling quality and segmentation quality of the model had a stepwise improvement. Among them, although the mIoU of pseudo labels only improved by 0.22% after adding the SAM expansion module, the segmentation mIoU of the model obtained a performance improvement of 1.13% after the corresponding training. This also proves the effectiveness of each module in the proposed method.

Table 4. Evaluation of pseudo-label quality and ablation study results on the DLRSD dataset.

Evaluation Methods	Siamese	CRF	SAM	mIoU	OA	mPrecision	mRecall	mFscore
Pseudo-mask Quality - Training Set	×	×	×	44.33	66.36	60.65	58.65	58.53
	✓	×	×	47.43	68.95	63.22	61.66	61.5
	✓	✓	×	49.87	72.2	66.28	62.72	63.1
	✓	✓	✓	**50.09**	**72.32**	**66.59**	**63.01**	**63.42**
Segmentation Quality - Test Set	×	×	×	46.35	66.36	60.82	66.19	61.25
	✓	×	×	47.07	69.46	64.11	61.77	61.71
	✓	✓	×	48.00	70.62	64.73	63.67	61.89
	✓	✓	✓	**49.13**	**71.52**	**66.51**	**64.60**	**63.91**

✓: Indicates module inclusion; ×: Indicates module exclusion. The bold values indicate the highest performance under each evaluation methods.

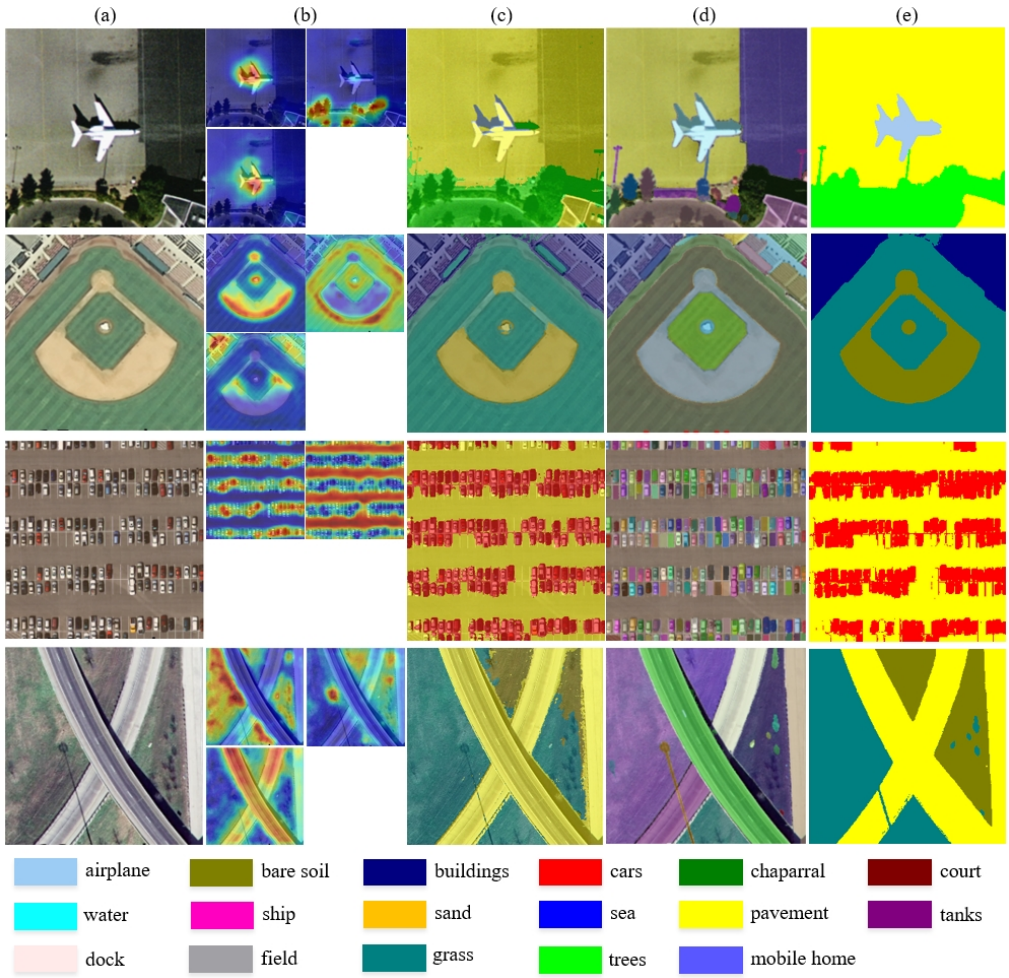

Figure 7. Visualization of the pseudo-label generation process on the DLRSD dataset. (**a**) Original image. (**b**) CAMs generated by the SAN. (**c**) Refined pseudo labels after CRF. (**d**) The segmentation results of SAM. (**e**) Refined pseudo labels after SAM-based pseudo-label expansion module. Legend applies to (**c**,**e**). (**b**,**c**,**d**) are superimposed on the original image.

5. Conclusions

We proposed a weakly supervised semantic segmentation method for remote sensing images leveraging the concepts of the Siamese Affinity Network in self-supervised learning and the SAM segmentation model. To accurately capture the attention of multi-class models, we introduced a semantic affinity-based seed enhancement module based on the unified constraint principle of cross-pixel similarity. This module enhances contextual relevance in feature maps, capturing semantically similar regions within the image. Additionally, to address the challenge of lacking extra supervision in weakly supervised semantic segmentation, we employed a Siamese network to achieve consistency constraints on the CAMs of differently affine-transformed images based on the prior of cross-view consistency, providing extra supervision for weakly supervised learning. Finally, by using the highly generalizable SAM segmentation model to generate semantic superpixels, we extended the original CAM seeds to more completely and precisely extract the target edges, thereby further improving the quality of the segmentation pseudo labels.

To validate the effectiveness of the proposed method, extensive experiments were conducted on the large-scale remote sensing datasets DRLSD and ISPRS Vaihingen. The results demonstrate that our method achieved segmentation performance close to that of FSSS methods on both datasets. Ablation experiments further confirm the positive optimization effect of each module on the segmentation pseudo labels. The proposed method exhibited

significantly superior localization accuracy and precise visualization effects across different backbone networks, achieving state-of-the-art localization performance.

Author Contributions: Conceptualization, J.Z.; Methodology, Z.C. and Y.L.; Software, Y.L.; Formal analysis, Y.L.; Investigation, Z.X.; Data curation, B.H.; Writing—original draft, Z.C.; Supervision, J.B., Z.X. and B.H.; Funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 62276206, Grant U24A20247, and Grant 62176196; by the Aeronautical Science Foundation of China under Grant 2023Z071081001; by the Key Research and Development Program of Hunan Province under Grant 2024AQ2032; by the Cultivation Project of Yuelu Mountain Industrial Innovation Center under Grant 2023YCII0123; by the Shenzhen Science and Technology (S&T) Program under Grant JCYJ20240813162405007; and by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011915.

Data Availability Statement: Data associated with this research are available online. The DLRSD dataset is available at https://sites.google.com/view/zhouwx/dataset (accessed on 11 January 2024), and the ISPRS Vaihingen dataset is available at https://www.isprs.org/education/benchmarks/UrbanSemLab (accessed on 9 January 2024).

Conflicts of Interest: Author Zheng Chen was employed by the company China Mobile Tietong Co., Ltd. The remaining 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.

References

1. Huang, L.; Jiang, B.; Lv, S.; Liu, Y.; Fu, Y. Deep Learning-Based Semantic Segmentation of Remote Sensing Images: A Survey. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2023**, *17*, 8370–8396. [CrossRef]
2. Wang, Y.; Bai, J.; Xiao, Z.; Zhou, H.; Jiao, L. MsmcNet: A modular few-shot learning framework for signal modulation classification. *IEEE Trans. Signal Process.* **2022**, *70*, 3789–3801. [CrossRef]
3. Liu, Y.; Bai, J.; Sun, F. Visual Localization Method for Unmanned Aerial Vehicles in Urban Scenes Based on Shape and Spatial Relationship Matching of Buildings. *Remote Sens.* **2024**, *16*, 3065. [CrossRef]
4. Bai, J.; Shi, W.; Xiao, Z.; Ali, T.A.A.; Ye, F.; Jiao, L. Achieving better category separability for hyperspectral image classification: A spatial–spectral approach. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 9621–9635. [CrossRef]
5. Bai, J.; Ding, B.; Xiao, Z.; Jiao, L.; Chen, H.; Regan, A.C. Hyperspectral image classification based on deep attention graph convolutional network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
6. Bai, J.; Yuan, A.; Xiao, Z.; Zhou, H.; Wang, D.; Jiang, H.; Jiao, L. Class incremental learning with few-shots based on linear programming for hyperspectral image classification. *IEEE Trans. Cybern.* **2020**, *52*, 5474–5485. [CrossRef]
7. Wei, Y.; Ji, S. Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]
8. Bai, J.; Ren, J.; Yang, Y.; Xiao, Z.; Yu, W.; Havyarimana, V.; Jiao, L. Object detection in large-scale remote-sensing images based on time-frequency analysis and feature optimization. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
9. Bai, J.; Ren, J.; Xiao, Z.; Chen, Z.; Gao, C.; Ali, T.A.A.; Jiao, L. Localizing from classification: Self-directed weakly supervised object localization for remote sensing images. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 17935–17949. [CrossRef]
10. Shen, W.; Peng, Z.; Wang, X.; Wang, H.; Cen, J.; Jiang, D.; Xie, L.; Yang, X.; Tian, Q. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9284–9305. [CrossRef]
11. Zhang, M.; Zhou, Y.; Zhao, J.; Man, Y.; Liu, B.; Yao, R. A survey of semi-and weakly supervised semantic segmentation of images. *Artif. Intell. Rev.* **2020**, *53*, 4259–4288. [CrossRef]
12. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
13. Chen, J.; Zhang, Y.; Ma, F.; Huang, K.; Tan, Z.; Qi, Y.; Li, J. Weakly-Supervised Semantic Segmentation of ALS Point Clouds Based on Auxiliary Line and Plane Point Prediction. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2024**, *17*, 18096–18111. [CrossRef]
14. Yan, X.; Shen, L.; Wang, J.; Deng, X.; Li, Z. MSG-SR-Net: A weakly supervised network integrating multiscale generation and superpixel refinement for building extraction from high-resolution remotely sensed imageries. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *15*, 1012–1023. [CrossRef]

15. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sens.* **2020**, *12*, 207. [CrossRef]
16. Lu, M.; Fang, L.; Li, M.; Zhang, B.; Zhang, Y.; Ghamisi, P. NFANet: A novel method for weakly supervised water extraction from high-resolution remote-sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
17. Zhou, R.; Zhang, W.; Yuan, Z.; Rong, X.; Liu, W.; Fu, K.; Sun, X. Weakly supervised semantic segmentation in aerial imagery via explicit pixel-level constraints. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
18. Lee, J.; Kim, E.; Lee, S.; Lee, J.; Yoon, S. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5267–5276.
19. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1422–1430.
20. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv* **2018**, arXiv:1803.07728.
21. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
22. Larsson, G.; Maire, M.; Shakhnarovich, G. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 577–593.
23. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 4015–4026.
24. Kolesnikov, A.; Lampert, C.H. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Computer Vision–ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 695–711.
25. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
26. Sun, W.; Zhang, J.; Barnes, N. Inferring the class conditional response map for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 2878–2887.
27. Yu, B.; Yang, L.; Chen, F. Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2018**, *11*, 3252–3261. [CrossRef]
28. Lee, J.; Choi, J.; Mok, J.; Yoon, S. Reducing information bottleneck for weakly supervised semantic segmentation. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2021**, *34*, 27408–27421.
29. Ahn, J.; Cho, S.; Kwak, S. Weakly supervised learning of instance segmentation with inter-pixel relations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2209–2218.
30. Gao, L.; Liu, H.; Yang, M.; Chen, L.; Wan, Y.; Xiao, Z.; Qian, Y. STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 10990–11003. [CrossRef]
31. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
32. Ahn, J.; Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4981–4990.
33. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
34. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
35. Fan, J.; Zhang, Z.; Tan, T.; Song, C.; Xiao, J. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10762–10769.
36. Cao, Y.; Huang, X. A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 157–176. [CrossRef]
37. Ali, M.U.; Sultani, W.; Ali, M. Destruction from sky: Weakly supervised approach for destruction detection in satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 115–124. [CrossRef]

38. Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; Volume 1, p. 3.
39. Fu, K.; Lu, W.; Diao, W.; Yan, M.; Sun, H.; Zhang, Y.; Sun, X. WSF-NET: Weakly supervised feature-fusion network for binary segmentation in remote sensing image. *Remote Sens.* **2018**, *10*, 1970. [CrossRef]
40. Sun, J.; He, W.; Zhang, H. G2LDIE: Global-to-local dynamic information enhancement framework for weakly supervised building extraction from remote sensing images. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5406714. [CrossRef]
41. Wang, J.; Zhang, X.; Ma, X.; Yu, W.; Ghamisi, P. Auto-Prompting SAM for Weakly Supervised Landslide Extraction. *arXiv* **2025**, arXiv:2501.13426.
42. Lu, B.; Ding, C.; Bi, J.; Song, D. Weakly Supervised Change Detection via Knowledge Distillation and Multiscale Sigmoid Inference. *arXiv* **2024**, arXiv:2403.05796.
43. Zhao, M.; Hu, X.; Zhang, L.; Meng, Q.; Chen, Y.; Bruzzone, L. Beyond Pixel-Level Annotation: Exploring Self-Supervised Learning for Change Detection with Image-Level Supervision. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5614916. [CrossRef]
44. Zeng, X.; Wang, T.; Dong, Z.; Zhang, X.; Gu, Y. Superpixel consistency saliency map generation for weakly supervised semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5606016. [CrossRef]
45. Li, Z.; Zhang, X.; Xiao, P. One model is enough: Toward multiclass weakly supervised remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4503513. [CrossRef]
46. Hu, Z.; Gao, J.; Yuan, Y.; Li, X. Contrastive Tokens and Label Activation for Remote Sensing Weakly Supervised Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5620211. [CrossRef]
47. Chen, T.; Mai, Z.; Li, R.; Chao, W.l. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv* **2023**, arXiv:2305.05803.
48. Jiang, P.T.; Yang, Y. Segment anything is a good pseudo-label generator for weakly supervised semantic segmentation. *arXiv* **2023**, arXiv:2305.01275.
49. Sun, W.; Liu, Z.; Zhang, Y.; Zhong, Y.; Barnes, N. An alternative to wsss? an empirical study of the segment anything model (sam) on weakly-supervised semantic segmentation problems. *arXiv* **2023**, arXiv:2305.01586.
50. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv* **2023**, arXiv:2303.05499.
51. Shao, Z.; Yang, K.; Zhou, W. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sens.* **2018**, *10*, 964. [CrossRef]
52. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
54. Pathak, D.; Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional multi-class multiple instance learning. *arXiv* **2014**, arXiv:1412.7144.
55. Huang, Z.; Wang, X.; Wang, J.; Liu, W.; Wang, J. Weakly-supervised semantic segmentation network with deep seeded region growing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7014–7023.
56. Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; Chen, X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12275–12284.
57. Chang, Y.T.; Wang, Q.; Hung, W.C.; Piramuthu, R.; Tsai, Y.H.; Yang, M.H. Weakly-supervised semantic segmentation via sub-category exploration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8991–9000.