remote sensing

MDPI

Article

Image Segmentation Method Based on a Dual-Domain Coupled 2-Additive Fuzzy-Measure Capacity Network

Ran Zhang and Kaihong Guo *

Faculty of Information, Liaoning University, Shenyang 110036, China; 4022310183@smail.lnu.edu.cn
* Correspondence: guokaihong@lnu.edu.cn

Highlights

What are the main findings?

- A dual-domain framework jointly models region membership and boundary separation for weakly supervised remote sensing image segmentation.
- A 2-additive fuzzy-measure capacity captures individual evidence contributions and pairwise interactions for boundary discrimination.
- DC2CNet improves segmentation performance on Potsdam, LoveDA, and DeepGlobe, particularly for small objects, elongated structures, and complex boundaries.

What are the implications of the main findings?

- Dual-domain coupling provides explicit structural constraints between region continuity and boundary separation under sparse annotations.
- Non-additive evidence fusion strengthens local structural discrimination beyond conventional concatenation or weighted feature fusion.
- The proposed framework improves segmentation accuracy in complex high-resolution remote sensing scenes.

Abstract

High-resolution remote sensing images exhibit large variations in object scale, ambiguous boundary transitions, and complex interactions among adjacent land-cover regions. Under sparse pixel supervision, these factors may cause object fragmentation, erroneous merging, and boundary displacement. To address these issues, a Dual-Domain Coupled 2-Additive Fuzzy Measure Network (DC2CNet) is proposed for sparse pixel-supervised remote sensing image segmentation. The proposed method extends pixel-level class prediction by jointly modeling regional consistency and boundary discrimination. A region-domain branch and a boundary-domain branch are constructed to characterize intra-region coherence and class-transition information, respectively. A dual-domain coupling mechanism is further introduced to integrate regional and boundary evidence. In addition, a 2-additive fuzzy measure decoding strategy is developed to model individual evidence contributions and pairwise evidence interactions through Möbius representation. During training, sparse pixel annotations are used to generate boundary discrimination probabilities, and neighborhood consistency constraints are introduced to optimize the boundary branch. Experimental results demonstrate that DC2CNet effectively reduces boundary ambiguity and region confusion, while maintaining accurate segmentation of small objects, elongated structures, and complex interface regions.

Academic Editors: Xinghua Li and Le Sun

Received: 30 August 2026
Revised: 11 October 2026
Accepted: 27 October 2026
Published: 30 October 2026

Copyright: © 2026 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.

Keywords: remote sensing image segmentation; 2-additive fuzzy measure; dual-domain coupling; boundary discrimination; sparse pixel supervision

1. Introduction

Remote sensing image segmentation is a fundamental task in fine-grained land-cover recognition, supporting urban spatial analysis, land-cover mapping, and ecological environment monitoring [1–3]. Compared with natural images, high-resolution remote sensing images contain more small-scale objects, elongated structures, and complex interface regions. Considerable variations in texture, appearance, and scale may occur within the same class, while different land-cover categories may exhibit similar spectral responses. Therefore, accurate segmentation requires not only pixel-level semantic discrimination but also the preservation of spatial connectivity and precise boundary locations.

Early remote sensing image segmentation methods mainly relied on handcrafted spectral, texture, and shape features combined with shallow classifiers for region partitioning. However, their performance was strongly dependent on feature design and was limited in complex land-cover scenes. With the development of deep learning, convolutional neural networks have formulated segmentation as pixel-level prediction and achieved improved feature representation through encoder–decoder architectures, skip connections, and multiscale context modeling [4–6]. Nevertheless, repeated downsampling may weaken local spatial details, resulting in inaccurate boundaries and fragmented predictions. Recently, Transformer-based architectures and hierarchical attention mechanisms have been introduced to capture long-range semantic relationships [7]. Although these methods improve global context modeling, subtle texture variations, narrow structures, and local transitions at class interfaces remain challenging in high-resolution remote sensing images.

Existing methods mainly improve segmentation performance through scale modeling or boundary enhancement. Multiscale feature pyramids, attention mechanisms, and feature interaction strategies have been developed to alleviate object-scale variations [8–10]. Boundary-aware methods introduce additional contour constraints through boundary branches or boundary losses [11,12]. However, scale-based methods mainly focus on semantic representation, while boundary information is often treated as auxiliary supervision rather than being integrated into feature learning or decoding. Consequently, the structural relationship between regional consistency and boundary transitions remains insufficiently explored, which may lead to region confusion, boundary displacement, and fragmented predictions in complex scenes [13,14].

These challenges become more severe under sparse pixel supervision. Pixel-level annotation of high-resolution remote sensing images is expensive, and only a limited number of labeled pixels may be available during training. Such sparse annotations provide insufficient constraints for intra-region consistency and class-transition boundaries, making it difficult to preserve complete objects and accurate interfaces. Therefore, effectively exploiting structural information beyond labeled pixels is important for improving segmentation reliability under limited annotation conditions.

Based on the above observations, a Dual-Domain Coupled 2-Additive Fuzzy Measure Network (DC2CNet) is proposed for sparse pixel-supervised remote sensing image segmentation. The proposed framework introduces a dual-domain representation to jointly model regional consistency and boundary discrimination. A dual-domain coupling mechanism is further designed to integrate complementary structural information between the two domains. In addition, a 2-additive fuzzy-measure decoding

strategy is introduced to model individual evidence contributions and pairwise interactions among local evidence, providing non-additive structural information for segmentation prediction. A sparse annotation-driven boundary optimization strategy is also developed to enhance structural consistency under limited pixel annotations. The main contributions of this study are as follows:

(1) A dual-domain coupled segmentation framework is proposed for sparse pixel-supervised remote sensing image segmentation. The framework models regional consistency and boundary transitions simultaneously and establishes interactions between them to improve structural representation under limited annotations.

(2) A 2-additive fuzzy measure decoding strategy is developed to model individual evidence contributions and pairwise evidence interactions through Möbius representation, enabling adaptive integration of local discriminative information.

(3) A sparse annotation-driven boundary optimization strategy is introduced, where limited pixel annotations are used to generate boundary supervision and improve structural consistency during segmentation.

2. Related Work

Remote sensing image segmentation research mainly focuses on semantic representation, spatial detail preservation, and structural modeling. Existing methods can be generally categorized into convolutional networks, Transformers, boundary-aware methods, multiscale fusion approaches, and fuzzy evidence fusion strategies. These methods improve segmentation performance from different perspectives, including local feature extraction, global dependency modeling, scale adaptation, and structural constraint learning.

2.1. Convolutional Network-Based Segmentation Methods

Convolutional neural networks extract local textures and neighborhood patterns through convolution operations and have been widely applied to remote sensing image segmentation [15]. FCN transforms classification networks into dense prediction frameworks to generate pixel-level outputs [16]. U-Net introduces an encoder–decoder architecture with skip connections to combine semantic features and spatial details [17,18]. The DeepLab series enlarges the receptive field through dilated convolutions and spatial pyramid pooling, improving multiscale contextual representation [19,20].

Although convolutional networks provide effective local feature extraction, their representation capability is constrained by limited receptive fields and repeated downsampling. Fine spatial details may be weakened during feature extraction, resulting in inaccurate boundaries and fragmented predictions for small objects and narrow structures. These limitations motivate the development of methods with stronger global context modeling and structural representation capabilities.

2.2. Transformer-Based Segmentation Methods

Transformers model long-range dependencies through self-attention mechanisms and provide a different feature representation paradigm from convolutional networks in remote sensing image segmentation [21–23]. Vision Transformer divides images into patches and represents them as tokens, enabling global interaction among different regions [24]. Hierarchical Transformers, such as PVT, introduce pyramid structures to progressively aggregate multiscale semantic information and improve dense prediction performance [25–28].

Despite their advantages in global semantic modeling, Transformer-based methods may weaken local boundary details due to patch-based representation and global feature aggregation. Small objects, narrow structures, and regions with subtle spectral variations

may remain challenging, especially when fine spatial information is required for accurate delineation.

2.3. Boundary-Aware and Structure-Constrained Methods

Boundary-aware methods improve segmentation accuracy by introducing contour supervision, boundary detection branches, or structural constraints. Some approaches adopt multitask learning strategies to jointly optimize semantic segmentation and boundary prediction [29–31]. Other methods incorporate boundary information into decoders through boundary attention or boundary-guided feature fusion to refine interface regions [32,33].

Although these methods improve boundary localization, boundary information is commonly introduced as an auxiliary constraint or additional prediction target. The interaction between boundary features and semantic representations still depends on the specific network design. Moreover, boundary prediction can be affected by texture variations and complex local patterns in high-resolution remote sensing images.

2.4. Multiscale Feature Fusion and Alignment Methods

Multiscale feature fusion methods aim to address object-scale variations in remote sensing images. FPN combines high-level semantic information with low-level spatial details through a top-down pathway and has been widely used for segmentation of roads, buildings, vehicles, and land-cover categories [34]. To improve cross-scale consistency, deformable convolution, attention-based alignment, and dynamic fusion strategies have been explored [35–37]. ASFF adaptively selects features from different scales through pixel-level weighting mechanisms [38–40].

However, feature fusion across different scales remains affected by differences in resolution, receptive fields, and semantic levels. Existing alignment strategies mainly focus on scale adaptation and feature selection, while the structural relationships between local regions and class transitions are less explicitly considered.

2.5. Fuzzy Measures and Non-Additive Evidence Fusion Methods

Fuzzy measures have been widely used in multisource information fusion and uncertainty modeling to represent the contributions of different evidence subsets. Unlike linear weighting strategies, fuzzy measures can model non-additive relationships among evidence elements. The Choquet integral further enables aggregation of ordered evidence according to fuzzy measure values, allowing both individual contributions and interactions among multiple evidence sources to be represented.

A 2-additive fuzzy measure retains first-order and second-order Möbius terms, which reduces model complexity while preserving pairwise interactions among evidence elements [41]. Such characteristics make it suitable for modeling situations where multiple types of evidence jointly influence the decision process.

Existing segmentation methods mainly rely on feature concatenation, weighted summation, or attention-based recalibration for information fusion. These strategies adjust feature responses but have limited capability in explicitly describing nonlinear interactions among different evidence sources. Therefore, this study introduces a 2-additive fuzzy measure into the decoding process to model local evidence interactions for sparse pixel-supervised remote sensing image segmentation.

2.6. Foundation Model-Based Segmentation Methods

Large-scale pretrained foundation models have recently attracted attention due to their transferable visual representations [42,43]. DINOv2 learns general visual features through large-scale self-supervised pretraining and has been applied to downstream

dense prediction tasks [44]. Segment Anything Model (SAM) introduces a prompt-driven segmentation paradigm and provides a general segmentation framework with strong transferability [45].

Despite their powerful representation capability, foundation models are mainly developed for general visual understanding and may encounter domain gaps when applied to remote sensing images. The diverse object scales, complex land-cover patterns, and fine spatial structures in remote sensing scenes still require task-specific adaptation.

2.7. Sparse Annotation-Based Segmentation Methods

Dense pixel-level annotation is expensive for high-resolution remote sensing images, motivating the development of sparse annotation-based segmentation methods [46,47]. SparseFormer utilizes limited labeled pixels for semantic segmentation and demonstrates the feasibility of reducing annotation requirements [48].

However, sparse annotations provide incomplete spatial supervision, especially for unlabeled regions and object boundaries. Effectively exploiting limited annotations while maintaining accurate segmentation remains a key challenge for sparse pixel-supervised remote sensing image segmentation.

3. Method

Image segmentation methods usually map an input image I to a pixel-level class probability map P, which can be expressed as: $P = f(I;\theta)$, where θ denotes the network parameters. Under the sparse pixel supervision setting, the commonly used segmentation loss is defined on the retained labeled pixels Ω_l:

$$L_{seg} = -\sum_{p\in\Omega_l} log\, P_{\tilde{Y}(p)}(p) \tag{1}$$

where p denotes a pixel or feature-map position, and $\tilde{Y}(p)$ represents the annotation of the retained labeled pixel. This loss constrains class prediction at labeled positions but does not explicitly model the structural relationships among neighboring pixels, such as regional consistency and class-transition boundaries.

To address this limitation, DC2CNet introduces dual-domain representation and non-additive evidence modeling. Region membership and boundary separation are characterized within local neighborhoods, and a 2-additive fuzzy measure is employed to model interactions among local evidence. The overall framework is illustrated in Figure 1.

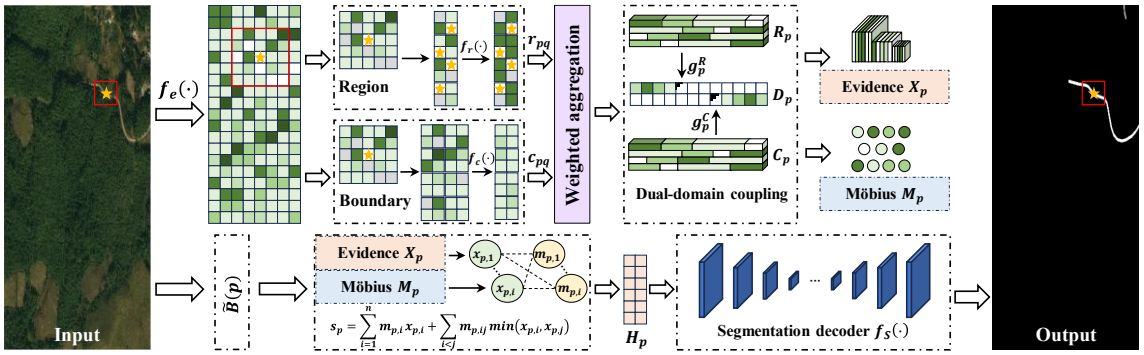


Figure 1. Overall architecture of DC2CNet. The red boxes highlight the corresponding local regions, while the yellow stars indicate the target pixels used to illustrate region and boundary modeling.

As shown in Figure 1, the input image $I \in \mathbb{R}^{H\times W\times 3}$ is first processed by a shared encoder $f_e(\cdot)$ to obtain the encoder feature:

$$A = f_e(I), A = \{a_p | p \in \Omega\} \tag{2}$$

Here, Ω is the set of spatial positions on the feature map, $a_p \in \mathbb{R}^d$ is the feature vector at position p, and d is the feature dimension. The encoder feature A contains color, texture, and semantic information, but it still uses an individual position as the representation unit. It cannot directly determine whether the position should remain connected to the neighborhood or form a boundary in the local structure. Therefore, a local neighborhood centered at position p is constructed as:

$$N_p = \{q | q \in \Omega, \|q - p\| \leq r\} \tag{3}$$

where q is a neighboring position and r is the neighborhood radius. To reduce the interference of irrelevant neighboring positions on local discrimination, normalized weights are calculated according to the similarity between the center feature and the neighboring features:

$$\omega_{pq} = \frac{exp\left(\eta(a_p, a_q)\right)}{\sum_{t \in N_p} exp\left(\eta(a_p, a_t)\right)} \tag{4}$$

Here, $\eta(\cdot)$ is a learnable similarity function, t is the neighborhood summation index, and ω_{pq} represents the influence weight of neighboring position q on center position p.

After the neighborhood structure is determined, DC2CNet describes local relationships through the region membership domain and the boundary separation domain. The region membership domain models semantic consistency between the center position and neighboring positions. Its relation value is defined as:

$$r_{pq} = \sigma\left(f_r([a_p, a_q, a_p - a_q])\right) \tag{5}$$

where r_{pq} represents the region membership relationship between positions p and q, $f_r(\cdot)$ is the region relation mapping function implemented by a two-layer multilayer perceptron (MLP) with nonlinear activation, which maps the concatenated local features into the region membership relationship, and $\sigma(\cdot)$ is the sigmoid function. The signed difference $a_p - a_q$ is retained in the region membership domain because the directional feature variation between neighboring positions provides additional information for modeling semantic connectivity. The boundary separation domain describes whether a structural boundary exists between the center position and the neighboring position. Its relation value is defined as:

$$c_{pq} = \sigma\left(f_c([a_p, a_q, |a_p - a_q|])\right) \tag{6}$$

where c_{pq} represents the boundary separation relationship between positions p and q, and $f_c(\cdot)$ is the boundary relation mapping function implemented by a two-layer MLP with nonlinear activation. $f_c(\cdot)$ adopts an independent parameter set from $f_r(\cdot)$ to model boundary separation relationships. Different from the region membership domain, the boundary separation domain focuses on the magnitude of local feature discontinuity rather than its direction. Therefore, the absolute difference $|a_p - a_q|$ is adopted to measure structural variation between neighboring positions. Based on the two relationships, neighborhood information is aggregated to obtain the region membership feature R_p and the boundary separation feature C_p:

$$R_p = \sum_{q \in N_p} \omega_{pq} r_{pq} f_R(a_q) \tag{7}$$

$$C_p = \sum_{q \in N_p} \omega_{pq}\, c_{pq} f_C(|a_p - a_q|) \tag{8}$$

where $f_R(\cdot)$ and $f_C(\cdot)$ are feature transformation functions implemented by linear projection layers, which transform neighborhood features into the region membership feature and boundary separation feature, respectively. R_p describes the semantic connectivity tendency within the neighborhood, while C_p describes the local structural variation tendency. Directly combining the two features may cause insufficient interaction between regional continuity and boundary discrimination. To handle the conflict between region connectivity and boundary separation, DC2CNet constructs a dual-domain coupling process. The boundary separation feature C_p produces a region suppression gate to restrict the propagation of region responses across boundaries. The region membership feature R_p produces a boundary modulation gate to adaptively modulate the boundary response according to regional semantic information:

$$g_p^R = \sigma(W_R C_p), g_p^C = \sigma(W_C R_p) \tag{9}$$

Here, W_R and W_C are learnable transformations, g_p^R is the region propagation suppression weight, and g_p^C is the boundary response modulation weight. The gated dual-domain features are:

$$\bar{R}_p = R_p \odot (1 - g_p^R), \bar{C}_p = C_p \odot (1 + g_p^C) \tag{10}$$

where $\odot$ denotes element-wise multiplication. When the local boundary response is strong, region propagation is restricted. When local differences lack regional semantic support, the boundary response is suppressed. The center feature, region membership feature, boundary separation feature, and their interaction terms are then jointly mapped into the dual-domain coupled feature D_p:

$$D_p = f_D([a_p, \bar{R}_p, \bar{C}_p, \bar{R}_p - \bar{C}_p, \bar{R}_p \odot \bar{C}_p]) \tag{11}$$

where $f_D(\cdot)$ is the coupling mapping function implemented by a two-layer MLP with nonlinear activation, which integrates the center feature, region membership feature, boundary separation feature, and their interaction terms. D_p contains center semantics, region consistency, boundary separability, and the differences between the two domains and is used as the input representation for subsequent capacity decoding. The dual-domain coupled feature D_p is further mapped into n local discriminative evidence components:

$$X_p = \sigma\left(f_X(D_p)\right) = [x_{p,1}, x_{p,2}, \ldots, x_{p,n}] \tag{12}$$

where $f_X(\cdot)$ is an evidence mapping layer implemented by a fully connected layer with sigmoid activation to generate n evidence components, $x_{p,i} \in [0,1]$ represents the i-th evidence component, and n is the number of evidence components. These evidence components correspond to local discriminative factors, including regional inconsistency, neighborhood inconsistency, directional variation, and boundary variation. Since individual evidence components may be affected by texture noise or local variations, DC2CNet introduces a 2-additive fuzzy measure to model interactions among evidence components. Position-dependent Möbius parameters are learned from D_p:

$$M_p = f_M(D_p) = \{m_{p,i}, m_{p,ij}\} \tag{13}$$

where $f_M(\cdot)$ is a capacity parameter mapping layer implemented by a fully connected layer to predict the first-order and second-order Möbius parameters, and $m_{p,i}$ represents the first-order contribution of the i-th evidence component, and $m_{p,ij}$ represents the second-order interaction between the i-th and j-th evidence components. Let $N =$

$\{1,2,\ldots,n\}$ be the evidence index set, and let $S \subseteq N$ be any evidence subset. The 2-additive capacity at position p is defined as:

$$\mu_p(S) = \sum_{i \in S} m_{p,i} + \sum_{\substack{i,j \in S \\ i<j}} m_{p,ij} \tag{14}$$

where $\mu_p(S)$ represents the degree to which the evidence subset S supports boundary discrimination. In DC2CNet, only first-order and second-order interaction terms are retained. This avoids modeling all higher-order evidence combinations while retaining the most direct synergistic or inhibitory relationships among evidence components:

$$m_p(S) = 0, |S| > 2 \tag{15}$$

where $m_p(S)$ is the Möbius term corresponding to evidence subset S. To reduce unnecessary complexity from higher-order interactions, while maintaining nonnegative first-order contributions and allowing second-order terms to represent positive or negative interactions, constrained parameterizations are applied to the first-order and second-order terms:

$$m_{p,i} = softplus(\alpha_{p,i}), m_{p,ij} = \lambda tanh(\beta_{p,ij}) \tag{16}$$

where $\alpha_{p,i}$ and $\beta_{p,ij}$ are output by $f_M(D_p)$, and λ controls the value range of the second-order Möbius terms. After nonnegative mapping, $m_{p,i}$ represents the basic contribution of an individual evidence component. $m_{p,ij}$ may take a positive or negative value. A positive value indicates that evidence components i and j jointly support boundary discrimination, whereas a negative value indicates that their combination represents suppressive interactions or redundant information. Negative second-order terms remain valid under the capacity constraint and indicate non-additive relationships among evidence components. After the evidence X_p and regularized capacity parameters M_p are determined, the 2-additive aggregation response corresponding to the Choquet formulation is calculated as:

$$s_p = \sum_{i=1}^{n} m_{p,i}\, x_{p,i} + \sum_{i<j} m_{p,ij}\, min(x_{p,i}, x_{p,j}) \tag{17}$$

Here, $min(x_{p,i}, x_{p,j})$ represents the degree of joint occurrence of evidence components i and j. Unlike concatenation, weighted summation, or attention recalibration, the fuzzy measure response in Equation (17) is determined by both individual evidence contributions and pairwise evidence interactions:

$$H_p = D_p + \sigma(s_p) f_H(D_p) \tag{18}$$

where H_p is the boundary-enhanced feature, and $f_H(\cdot)$ is the boundary mapping function. H_p retains the original representation of D_p and adds a boundary response term modulated by s_p to change the local feature distribution. Based on H_p, the boundary discrimination probability is defined as:

$$P_b(p) = \sigma(W_o H_p) \tag{19}$$

where W_o denotes the parameters of the boundary discrimination head. $P_b(p) \in [0,1]$ represents the probability that position p belongs to a boundary. This boundary discrimination probability is used only as an auxiliary constraint signal during training and is not output as an independent segmentation result during inference. After capacity decoding, the segmentation decoder jointly receives the semantic feature, dual-domain coupled feature, and fuzzy measure-guided boundary discrimination feature:

$$Z_p = f_S([a_p, D_p, H_p]) \tag{20}$$

where Z_p is the segmentation decoding feature, and $f_S(\cdot)$ is a lightweight convolutional decoding block that integrates the decoded features and generates pixel-level class predictions. This process places the basic semantic response a_p, dual-domain structural relationship D_p, and boundary-enhanced response H_p in the same decoding space and generates the feature representation for pixel-level classification through a unified mapping. The class probability is obtained through a linear mapping and softmax:

$$P(p) = softmax(W_s Z_p) \tag{21}$$

where W_s denotes the parameters of the classification head, and $P_k(p)$ represents the probability that position p belongs to class k. The final predicted class is:

$$Y(p) = arg\,max_k P_k(p) \tag{22}$$

where $Y(p)$ is the pixel-level class label. Because Z_p contains both D_p and H_p, the decision process is constrained by both regional structure and boundary variation.

During weakly supervised training, reliably labeled positions provide only sparse class information and cannot directly provide complete boundary annotations. To obtain an auxiliary boundary signal, a computable boundary candidate set is defined as: $\Omega_b = \{p | p \in \Omega_l, \exists q \in N_p \cap \Omega_l, q \neq p\}$. For $p \in \Omega_l$, the weak boundary candidate label $\tilde{B}(p)$ is defined as:

$$\tilde{B}(p) = \max_{q \in N_p \cap \Omega_l, q \neq p} \mathbf{1}[\tilde{Y}(p) \neq \tilde{Y}(q)] \tag{23}$$

where $\mathbf{1}[\cdot]$ is the indicator function. If a reliably labeled point with a different class exists in the neighborhood, position p is regarded as a potential boundary candidate rather than an exact boundary pixel. These candidates are only used as weak structural constraints to guide boundary response learning instead of pixel-level boundary supervision. Otherwise, the position is treated as an intra-region candidate. Because sparse annotations cannot provide continuous boundaries, $\tilde{B}(p)$ is used only to constrain the variation direction of the boundary response and is not required to align with the true boundary pixel by pixel. The boundary discrimination loss is:

$$L_b = -\frac{1}{|\Omega_b|} \sum_{p \in \Omega_b} \left[\tilde{B}(p) log\, P_b(p) + \left(1 - \tilde{B}(p)\right) log\left(1 - P_b(p)\right)\right] \tag{24}$$

If Ω_b is empty for a training sample, L_b is not calculated for that sample. To constrain prediction consistency within a region, a neighborhood consistency loss is further introduced. For center position p and neighboring position q, the region consistency weight is defined as:

$$k_{pq} = sg\left[\omega_{pq}\, r_{pq}\, \left(1 - P_b(p)\right)\left(1 - P_b(q)\right)\right] \tag{25}$$

where $sg(\cdot)$ denotes the stop-gradient operation. When two positions have a high region membership relationship and neither is located in a boundary region, k_{pq} takes a larger value. When the two positions are close to a boundary or their region relationship is weak, the constraint is weakened. The stop-gradient operation prevents the network from weakening the neighborhood consistency constraint by reducing the region relation weight or increasing the boundary probability. The neighborhood consistency loss is defined as:

$$L_n = \frac{1}{|\Omega|} \sum_{p \in \Omega} \frac{1}{|N_p|} \sum_{q \in N_p} k_{pq} \|P(p) - P(q)\|_2^2 \tag{26}$$

This loss constrains prediction distributions within the same region to remain consistent while avoiding forced smoothing across boundaries. Because the capacity parameters are generated by the network, normalization and monotonicity are encouraged through soft regularization rather than strictly enforced during optimization**.** The full-set capacity $\mu_p(N)$ is regularized to approach 1, and monotonicity is promoted through the constraint term in Equation (30):

$$\mu_p(S \cup \{i\}) - \mu_p(S) \geq 0 \tag{27}$$

According to Equation (14), this difference can be written as:

$$\mu_p(S \cup \{i\}) - \mu_p(S) = m_{p,i} + \sum_{j \in S} m_{p,ij} \tag{28}$$

To avoid enumerating all subsets, the following sufficient constraint is adopted:

$$\Delta_{p,i} = m_{p,i} + \sum_{j \neq i} min\left(0, m_{p,ij}\right) \tag{29}$$

$$L_\mu = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left|\mu_p(N) - 1\right| + \frac{1}{|\Omega| n} \sum_{p \in \Omega} \sum_{i \in N} max\left(0, -\Delta_{p,i}\right) \tag{30}$$

Equation (29) considers all negative second-order interactions associated with evidence component i. If $\Delta_{p,i} \geq 0$, adding evidence component i to any combination does not decrease the capacity. This constraint is stricter than checking only evidence pairs and avoids enumerating higher-order subsets. In practice, the constraint is optimized as a regularization term, and the degree of capacity-property violation is evaluated after convergence. The final training objective is:

$$L = L_{seg} + \rho L_b + \tau L_n + \gamma L_\mu \tag{31}$$

where ρ, τ, and γ are the weight coefficients of the boundary discrimination loss, neighborhood consistency loss, and capacity constraint loss, respectively. The satisfaction of normalization and monotonicity properties is further analyzed after model convergence in Section 4.7.1 to evaluate whether the learned parameters preserve the required fuzzy-measure characteristics.

During inference, forward computation is performed according to Equations (2)–(22), and the pixel-level class prediction $Y(p)$ is output. The boundary discrimination probability $P_b(p)$ is not output as an independent result but participates in the segmentation decoding process through the capacity-guided boundary discrimination feature H_p.

4. Experiments

4.1. Experimental Settings

4.1.1. Datasets

Experiments were conducted on three high-resolution remote sensing image segmentation datasets, including Potsdam, LoveDA, and DeepGlobe.

The Potsdam dataset released by ISPRS contains 38 aerial image tiles with a spatial size of 6000×6000 pixels. It includes six land-cover classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter. Following the commonly used ISPRS Potsdam protocol, 24 tiles were used for training, and 14 tiles were reserved for testing. A subset of the training tiles was randomly selected as the validation set. The original full-resolution ground-truth annotations without boundary erosion were adopted, and the clutter class was included in the mIoU calculation.

The LoveDA dataset contains 5987 remote sensing images collected from urban and rural regions, including 2522 training images, 1669 validation images, and 1796 test images. It contains seven land-cover classes: buildings, roads, water, barren land, forests, agricultural land, and background. Since the test labels are not publicly available, the official validation split was used for evaluation. Ignored pixels were excluded during metric computation.

The DeepGlobe Land Cover Classification dataset contains 1146 high-resolution satellite images with pixel-level land-cover annotations, including 803 training images, 171 validation images, and 172 test images. It includes seven land-cover categories: urban areas, agriculture, rangeland, forest, water, barren land, and unknown regions. The official training split was used for model optimization, and the official validation split was adopted for evaluation.

For all datasets, a sparse pixel-supervision setting was adopted. Specifically, 10% of pixels in each training image were randomly sampled from the complete annotations and retained as labeled pixels, while the remaining annotations were unavailable during training. For reproducible experiments, the same sparse pixel-supervision setting was adopted for model evaluation.

4.1.2. Comparison Methods

To comprehensively evaluate the effectiveness of DC2CNet, representative segmentation methods were selected for comparison based on their feature representation strategies and segmentation characteristics. The selected methods cover CNN-based, transformer-based, boundary-aware, multiscale fusion and alignment, Mamba-based, foundation model-based, and sparse annotation segmentation frameworks. These methods cover representative segmentation frameworks and enable comprehensive evaluation of DC2CNet under sparse pixel supervision. The detailed comparison methods are summarized in Table 1.

Since DC2CNet is designed for sparse pixel-supervised segmentation, comparison methods were evaluated using the same dataset splits and evaluation metrics. For reproducible methods, the sparse pixel-supervision setting described in Section 4.1.1 was adopted whenever applicable. For methods whose official implementations or pretrained models were unavailable, the reported results from the original publications were used as references.

Table 1. Comparison methods used in this study.

Category	Method
CNN-based methods	PSPNet [15], FCN [16], DeepLabV3+ [20]
Transformer-based methods	RSSFormer [18], SegFormer [25]
Boundary-aware methods	BoundaryFormer [32], Gated-SCNN [31]
Multiscale fusion and alignment methods	CATNet [34], UNetFormer [37], FaPN [39]
Mamba-based methods	VMamba [24]
Foundation model-based methods	DINOv2-based Segmentation(DINOv2) [44], SAM-based Segmentation(SAM) [45]
Sparse annotation-based methods	SparseFormer [48]

4.1.3. Evaluation Metrics

Model performance was evaluated using Intersection over Union (IoU), mean Intersection over Union (mIoU), Pixel Accuracy (PA), mean Pixel Accuracy (mPA), and Boundary Intersection over Union (BIoU). Let the confusion matrix be P, where P_{ij} represents the number of pixels whose ground-truth class is i and predicted class is j, and the total number of classes is $K+1$. The IoU of class i is defined as:

$$IoU_i = \frac{P_{ii}}{\sum_{j=0}^{K} P_{ij} + \sum_{j=0}^{K} P_{ji} - P_{ii}} \tag{32}$$

Here, P_{ii} represents the number of pixels of class i that are correctly predicted, and the denominator represents the union of the ground-truth and predicted regions of class i.

The mean IoU over all classes is:

$$mIoU = \frac{1}{K+1} \sum_{i=0}^{K} \frac{P_{ii}}{\sum_{j=0}^{K} P_{ij} + \sum_{j=0}^{K} P_{ji} - P_{ii}} \tag{33}$$

PA represents the proportion of correctly classified pixels among all pixels and is defined as:

$$PA = \frac{\sum_{i=0}^{K} P_{ii}}{\sum_{i=0}^{K} \sum_{j=0}^{K} P_{ij}} \tag{34}$$

mPA represents the mean pixel accuracy over all classes and is defined as:

$$mPA = \frac{1}{K+1} \sum_{i=0}^{K} \frac{P_{ii}}{\sum_{j=0}^{K} P_{ij}} \tag{35}$$

BIoU evaluates the agreement between predicted and ground-truth boundaries. For class i, let $B_d(G_i)$ and $B_d(S_i)$ denote the boundary regions derived from the ground-truth mask G_i and predicted mask S_i, respectively. BIoU is defined as:

$$BIoU_i = \frac{|B_d(G_i) \cap B_d(S_i)|}{|B_d(G_i) \cup B_d(S_i)|} \tag{36}$$

Here, $d = max\left(1, round\left(0.02\sqrt{H^2 + W^2}\right)\right)$, where H and W correspond to the inference crop size specified in Table 2. The ratio of 0.02 follows the original Boundary IoU setting [49] and is fixed for all methods without dataset-specific tuning. For multiclass segmentation, BIoU is averaged over all valid classes.

Table 2. Hyperparameter settings of DC2CNet.

Parameter	Value
Encoder backbone	ResNet-50
Pretraining	ImageNet
Neighborhood radius r	3
Evidence components n	4
Second-order interaction coefficient λ	0.5
Boundary loss weight ρ	0.5
Coupling loss weight τ	0.7
Capacity regularization weight γ	0.1
Training crop size	512 × 512
Inference crop size	512 × 512
Inference stride	256

4.2. Implementation Details

4.2.1. Experimental Environment and Evaluation Setup

All experiments were implemented using PyTorch 2.1.0 with CUDA 12.1 on Ubuntu 22.04. The experiments were conducted on a workstation equipped with an Intel Xeon Gold 6338 CPU, 128 GB RAM, and an NVIDIA RTX 4090 GPU with 24 GB memory.

The shared encoder of DC2CNet is based on ImageNet-pretrained ResNet-50. The remaining modules, including the region membership domain, boundary separation domain, dual-domain coupling module, and 2-additive fuzzy-measure fusion module,

were trained jointly with the encoder. All compared methods were implemented under the same experimental environment for fair comparison. The proposed framework is implemented using standard neural network operations without GPU-specific operators, enabling potential deployment on different hardware platforms through framework-level optimization.

For segmentation performance evaluation, all experiments were repeated three times with different random seeds, and the mean values with standard deviations were reported.

4.2.2. Hyperparameter Settings and Sensitivity Analysis

DC2CNet is designed for weakly supervised remote sensing segmentation, where only sparse pixel annotations are available during training. Following the experimental protocol, 10% of pixels in each training image were randomly selected as labeled pixels, while the remaining pixels were treated as unlabeled regions. The same sparse annotation strategy was applied to all datasets. During training, input images were cropped into 512×512 patches. For inference, overlapping sliding-window prediction was adopted with a crop size of 512×512 and a stride of 256 pixels.

The main hyperparameter settings of DC2CNet are summarized in Table 2. The values of r, n, λ, ρ, τ, and γ were determined through sensitivity analysis. Considering the large number of parameter combinations, the sensitivity analysis was conducted only on the Potsdam dataset. Potsdam was selected because it contains diverse land-cover categories and complex spatial structures, which are suitable for evaluating local neighborhood modeling, boundary-aware representation, and evidence interaction. After determining the optimal parameters, all experiments were conducted with the fixed configuration. The sensitivity analysis results are presented in Table 3.

Table 3. Sensitivity analysis of key hyperparameters on the Potsdam dataset. Bold values indicate the best results.

Parameter	Value	mIoU (%)	mPA (%)	Best Value
r	1/2/3/4/5	82.91/83.54/**83.96**/83.72/83.18	91.96/92.54/**92.86**/92.71/92.32	3
n	2/3/4/5	83.21/83.67/**84.02**/83.74	92.28/92.67/**92.93**/92.72	4
λ	0/0.1/0.3/0.5/1.0	82.84/83.46/83.81/**84.06**/83.62	91.89/92.41/92.74/**92.97**/92.53	0.5
ρ	0.1/0.3/0.5/0.7/1.0	83.48/83.82/**83.93**/83.71/83.36	92.39/92.70/**92.84**/92.65/92.27	0.5
τ	0.3/0.5/0.7/0.9/1.0	83.52/83.79/**83.98**/83.75/83.44	92.43/92.69/**92.89**/92.73/92.35	0.7
γ	0.01/0.05/0.1/0.2/0.5	83.41/83.76/**84.00**/83.68/83.22	92.31/92.66/**92.91**/92.59/92.18	0.1

As shown in Table 3, the selected parameters achieve stable segmentation performance under different settings. Therefore, $r = 3$, $n = 4$, $\lambda = 0.5$, $\rho = 0.5$, $\tau = 0.7$, and $\gamma = 0.1$ were adopted in subsequent experiments.

4.2.3. Model Training

The model was optimized using the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The network was trained for 200 epochs with a batch size of 8, and the learning rate was adjusted using a cosine decay schedule.

During training, random cropping, horizontal flipping, and vertical flipping were applied for data augmentation. All input images were normalized using ImageNet statistics. The loss weights were set as $\rho = 0.5$, $\tau = 0.7$, and $\gamma = 0.1$, and the objective function defined in Section 3 was optimized during training. The training convergence behavior of DC2CNet is illustrated in Figure 2.

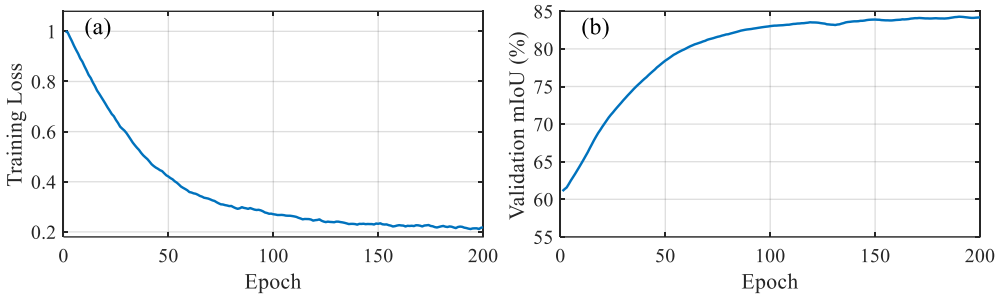


Figure 2. Training convergence analysis of DC2CNet: (**a**) training loss and (**b**) validation mIoU curves.

The training loss gradually decreases with increasing epochs, while the validation mIoU improves and converges after sufficient optimization epochs, indicating effective model convergence.

4.3. Performance Comparison

To quantitatively evaluate the effectiveness of DC2CNet, experiments were conducted on the three datasets introduced in Section 4.1.1, where DC2CNet was compared with the methods described in Section 4.1.2 under the same experimental settings. All methods adopted identical data splits, annotation ratios, and training configurations. The quantitative results on the Potsdam, LoveDA, and DeepGlobe datasets are reported in Tables 4–6, respectively. Bold values denote the best performance for each metric.

Table 4. Experimental results of different methods on the Potsdam dataset (%). Bold values indicate the best results.

Category	Method	IoU per Category						mIoU	PA	mPA	BIoU
		Imp. Surf.	Building	Low Veg.	Tree	Car	Clutter				
CNN-based methods	FCN	82.4	86.1	72.5	73.6	77.3	53.2	74.18 ± 0.35	88.36 ± 0.28	82.40 ± 0.31	68.42 ± 0.36
	PSPNet	86.2	90.0	77.7	78.7	81.0	59.8	78.90 ± 0.32	90.32 ± 0.25	85.03 ± 0.29	74.63 ± 0.33
	DeepLabv3+	88.3	91.6	80.2	80.9	84.6	62.3	81.32 ± 0.26	91.44 ± 0.22	86.60 ± 0.24	79.48 ± 0.29
Transformer-based methods	RSSFormer	89.1	92.5	81.0	82.4	85.7	65.8	82.75 ± 0.23	92.40 ± 0.19	87.60 ± 0.22	82.16 ± 0.26
	SegFormer	89.2	92.3	81.2	82.2	85.5	65.0	82.57 ± 0.25	92.24 ± 0.21	87.35 ± 0.24	81.72 ± 0.27
Boundary-aware methods	Gated-SCNN	88.7	91.9	80.6	81.5	85.2	66.2	82.35 ± 0.27	92.02 ± 0.23	87.10 ± 0.25	86.34 ± 0.23
	BoundaryFormer	89.8	93.1	81.4	82.5	87.8	69.5	84.02 ± 0.21	93.10 ± 0.18	88.60 ± 0.20	88.47 ± 0.20
Multiscale fusion and alignment methods	CATNet	83.5	95.9	78.0	70.9	85.0	52.8	82.67 ± 0.24	92.55 ± 0.20	87.85 ± 0.23	77.82 ± 0.30
	FaPN	83.9	96.5	78.7	71.5	83.7	53.1	83.05 ± 0.22	92.70 ± 0.19	88.10 ± 0.21	80.36 ± 0.27
	UNetFormer	84.1	96.9	79.9	72.3	86.5	53.6	83.67 ± 0.20	92.90 ± 0.17	88.40 ± 0.19	82.73 ± 0.24
Mamba-based methods	VMamba	88.8	92.8	82.1	83.1	86.9	68.3	84.00 ± 0.19	93.00 ± 0.16	88.50 ± 0.18	83.86 ± 0.22
	DINOv2-based Segmentation	89.9	92.8	82.0	83.4	87.0	68.6	84.12 ± 0.22	93.15 ± 0.17	88.70 ± 0.20	85.74 ± 0.21

Foundation model-based methods	SAM-based Segmentation	89.2	92.6	81.6	82.7	86.2	67.4	83.28 ± 0.24	92.70 ± 0.20	88.00 ± 0.22	85.16 ± 0.24
Sparse annotation-based methods	SparseFormer	90.0	92.7	82.4	83.2	87.1	68.8	84.03 ± 0.18	93.00 ± 0.15	88.60 ± 0.18	86.62 ± 0.19
Proposed method	DC2CNet	**90.1**	**92.9**	**82.5**	**83.4**	**87.6**	**69.2**	84.28 ± 0.14	93.05 ± 0.16	88.90 ± 0.15	91.28 ± 0.16

Table 5. Experimental results of different methods on the LoveDA dataset (%). Bold values indicate the best results.

Category	Method	IoU per Category							mIoU	PA	mPA	BIoU
		Background	Building	Road	Water	Barren	Forest	Agriculture				
CNN-based methods	FCN	36.0	47.4	39.2	63.1	31.6	45.1	45.4	43.97 ± 0.41	65.70 ± 0.35	55.80 ± 0.38	62.37 ± 0.42
	PSPNet	40.5	52.3	44.0	67.3	36.5	49.1	48.2	48.27 ± 0.36	68.60 ± 0.31	58.80 ± 0.34	66.84 ± 0.38
	DeepLabv3+	42.7	54.7	46.4	69.5	38.8	51.5	52.2	50.83 ± 0.33	70.20 ± 0.28	60.30 ± 0.30	71.63 ± 0.34
Transformer-based methods	RSSFormer	41.8	53.5	45	68.7	37.5	50.5	49.3	49.47 ± 0.30	69.30 ± 0.27	59.30 ± 0.29	74.82 ± 0.31
	SegFormer	43.7	56.2	48.4	71.3	40.1	53.0	56.0	52.67 ± 0.28	71.30 ± 0.24	61.60 ± 0.27	75.46 ± 0.30
Boundary-aware methods	Gated-SCNN	43.0	55.8	48.0	70.8	40.5	52.3	54.5	52.13 ± 0.31	70.80 ± 0.26	60.90 ± 0.28	81.35 ± 0.27
	CSBNet	43.2	56.0	48.2	71.0	40.7	52.6	55.0	52.39 ± 0.29	71.00 ± 0.25	61.20 ± 0.27	85.16 ± 0.24
Multiscale fusion and alignment methods	CATNet	43.5	55.9	48.0	70.9	39.8	52.8	55.8	52.39 ± 0.32	71.10 ± 0.28	61.30 ± 0.30	72.94 ± 0.34
	FaPN	43.9	56.5	48.7	71.5	40.7	53.1	57.2	53.09 ± 0.27	71.70 ± 0.23	61.90 ± 0.25	76.83 ± 0.30
	UNetFormer	44.1	56.9	49.0	72.3	41.0	53.6	58.2	53.59 ± 0.24	72.00 ± 0.21	62.30 ± 0.23	79.42 ± 0.28
Mamba-based methods	VMamba	44.0	57.1	49.3	71.5	41.5	53.5	59.8	54.11 ± 0.23	71.90 ± 0.20	62.40 ± 0.22	80.76 ± 0.26
Foundation model-based methods	DINOv2-based Segmentation	43.1	56.4	48.1	71.8	41.8	52.8	59.2	53.09 ± 0.26	71.20 ± 0.22	62.80 ± 0.25	83.68 ± 0.25
	SAM-based Segmentation	43.5	56.2	49.6	71.3	40.2	52.1	58.5	53.41 ± 0.28	70.60 ± 0.24	61.10 ± 0.26	82.91 ± 0.28
Sparse annotation-based methods	SparseFormer	43.0	56.8	49.7	71.6	41.6	53.7	57.9	52.93 ± 0.25	70.10 ± 0.23	61.60 ± 0.24	84.37 ± 0.23
Proposed method	DC2CNet	**44.3**	**57.4**	**50.2**	**72.1**	**42.0**	**54.0**	**60.1**	54.30 ± 0.18	72.50 ± 0.20	63.00 ± 0.17	88.24 ± 0.19

Table 6. Experimental results of different methods on the DeepGlobe dataset (%). Bold values indicate the best results.

Category	Method	IoU per Category						mIoU	PA	mPA	BIoU
		Urban	Agriculture	Rangeland	Forest	Water	Barren				
CNN-based methods	FCN	62.5	55.8	42.6	65.4	78.3	46.7	58.55 ± 0.38	79.60 ± 0.34	58.60 ± 0.36	66.85 ± 0.40

	PSPNet	65.8	59.6	45.1	68.7	81.2	50.3	61.78 ± 0.34	82.10 ± 0.30	61.80 ± 0.32	70.62 ± 0.36
	DeepLabv3+	68.9	62.7	48.5	71.6	83.5	53.8	64.83 ± 0.30	84.50 ± 0.27	64.80 ± 0.29	75.48 ± 0.32
Transformer-based methods	RSSFormer	70.2	64.5	50.4	73.1	84.6	55.2	66.33 ± 0.28	85.80 ± 0.24	66.30 ± 0.27	78.36 ± 0.29
	SegFormer	71.5	66.2	51.8	74.8	85.3	57.1	67.78 ± 0.26	86.90 ± 0.22	67.80 ± 0.25	79.14 ± 0.28
Boundary-aware methods	Gated-SCNN	71.8	66.0	52.1	74.5	85.6	57.8	67.97 ± 0.29	87.00 ± 0.24	68.00 ± 0.27	83.67 ± 0.26
	BoundaryFormer	72.1	66.5	52.7	75.0	85.9	58.2	68.40 ± 0.24	87.30 ± 0.21	68.40 ± 0.23	87.05 ± 0.22
Multiscale fusion and alignment methods	CATNet	71.6	65.8	51.9	74.6	85.1	57.5	67.75 ± 0.27	86.80 ± 0.23	67.80 ± 0.26	76.52 ± 0.32
	FaPN	72.0	66.3	52.4	75.2	85.5	58.0	68.23 ± 0.25	87.10 ± 0.22	68.20 ± 0.24	79.63 ± 0.29
	UNetFormer	72.6	67.0	53.1	75.8	86.0	58.7	68.87 ± 0.22	87.70 ± 0.19	68.90 ± 0.21	81.36 ± 0.26
Mamba-based methods	VMamba	73.1	67.8	53.8	76.5	86.4	59.4	69.50 ± 0.21	88.20 ± 0.18	69.50 ± 0.20	82.47 ± 0.25
Foundation model-based methods	DINOv2-based Segmentation	74.0	68.5	54.2	77.1	86.8	59.8	70.07 ± 0.23	88.80 ± 0.20	70.10 ± 0.22	85.86 ± 0.23
	SAM-based Segmentation	73.4	68.0	53.7	76.8	87.2	59.5	69.77 ± 0.25	88.50 ± 0.21	69.80 ± 0.24	84.73 ± 0.26
Sparse annotation-based methods	SparseFormer	73.6	68.2	54.0	76.7	86.6	60.1	69.87 ± 0.22	88.60 ± 0.19	69.90 ± 0.21	86.32 ± 0.21
Proposed method	DC2CNet	**74.2**	**68.6**	**54.3**	**77.2**	**87.3**	**60.2**	70.30 ± 0.16	89.40 ± 0.15	70.30 ± 0.17	89.76 ± 0.17

As shown in Tables 4–6, different segmentation architectures exhibit distinct performance characteristics under the sparse pixel supervision setting. CNN-based methods achieve relatively competitive results due to their effective local feature extraction capability, but their performance is limited when modeling long-range dependencies and complex structural transitions. Transformer-based and Mamba-based methods generally obtain improved segmentation performance by introducing global context modeling or more flexible feature interaction mechanisms. However, these methods are mainly designed for fully supervised segmentation and may suffer from insufficient structural constraints when only sparse pixel annotations are available.

Foundation model-based methods, such as DINOv2-based and SAM-based segmentation frameworks, provide strong representation capability from large-scale pretraining. Nevertheless, their performance gains are relatively limited under the current sparse supervision protocol, indicating that powerful feature extraction alone cannot fully compensate for the lack of pixel-level structural constraints. In comparison, DC2CNet achieves consistent improvements on all three datasets, with mIoU values of 84.28 ± 0.14%, 54.30 ± 0.18%, and 70.30 ± 0.16% on Potsdam, LoveDA, and DeepGlobe, respectively. The BIoU results further show a clearer advantage in boundary delineation. DC2CNet achieves BIoU values of 91.28%, 88.24%, and 89.76% on the three datasets, exceeding the strongest corresponding baselines by 2.81, 3.08, and 2.71 percentage points, respectively. Compared with the relatively small differences in mIoU, these larger BIoU gains indicate that the improvement of DC2CNet is more pronounced at object boundaries, consistent with the intended roles of the boundary separation and dual-domain coupling mechanisms. It should be noted that the performance of some compared

methods differs from their originally reported results. This difference mainly results from the unified sparse pixel supervision protocol adopted in this study, whereas many existing methods were originally evaluated with dense pixel annotations. Therefore, the results in Tables 4–6 reflect the capability of different segmentation methods under limited annotation conditions rather than their original fully supervised performance. The small standard deviations across repeated experiments further indicate the stable optimization behavior of the proposed DC2CNet.

4.4. Visual Analysis

To qualitatively analyze the segmentation characteristics of DC2CNet, visual comparisons were conducted on representative samples from the Potsdam, LoveDA, and DeepGlobe datasets introduced in Section 4.1.1. The compared methods followed the settings described in Section 4.1.2, and all visualization results were generated under identical experimental configurations. The qualitative results are presented in Figures 3–5. The black solid boxes mark the regions selected for detailed comparison.

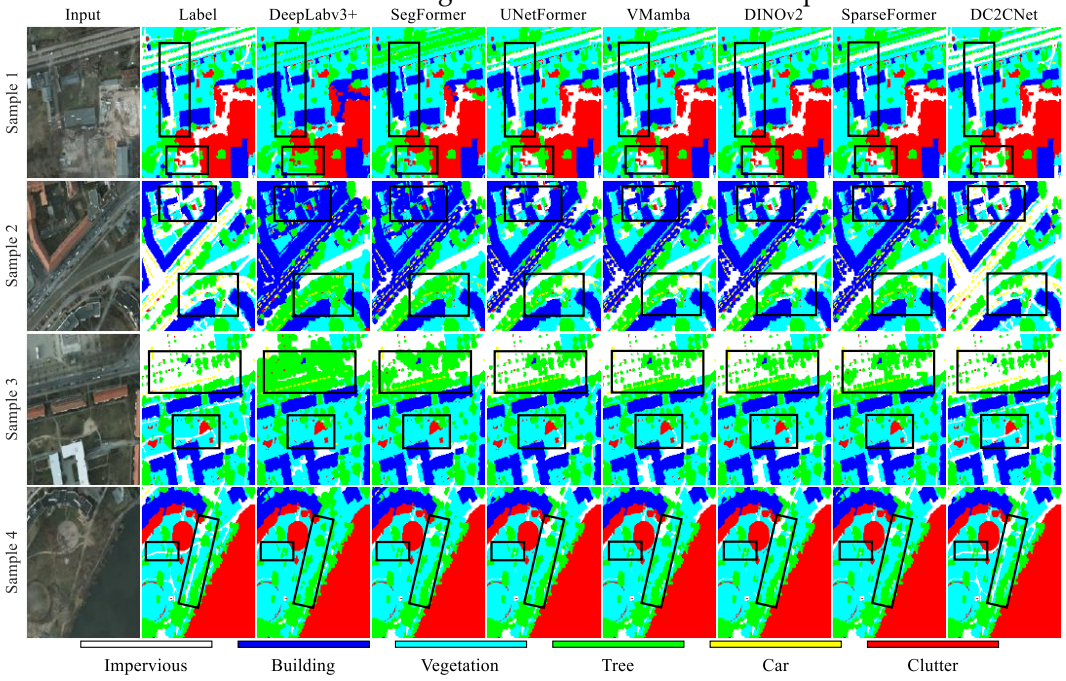


Figure 3. Visual comparison of segmentation results on the Potsdam dataset.

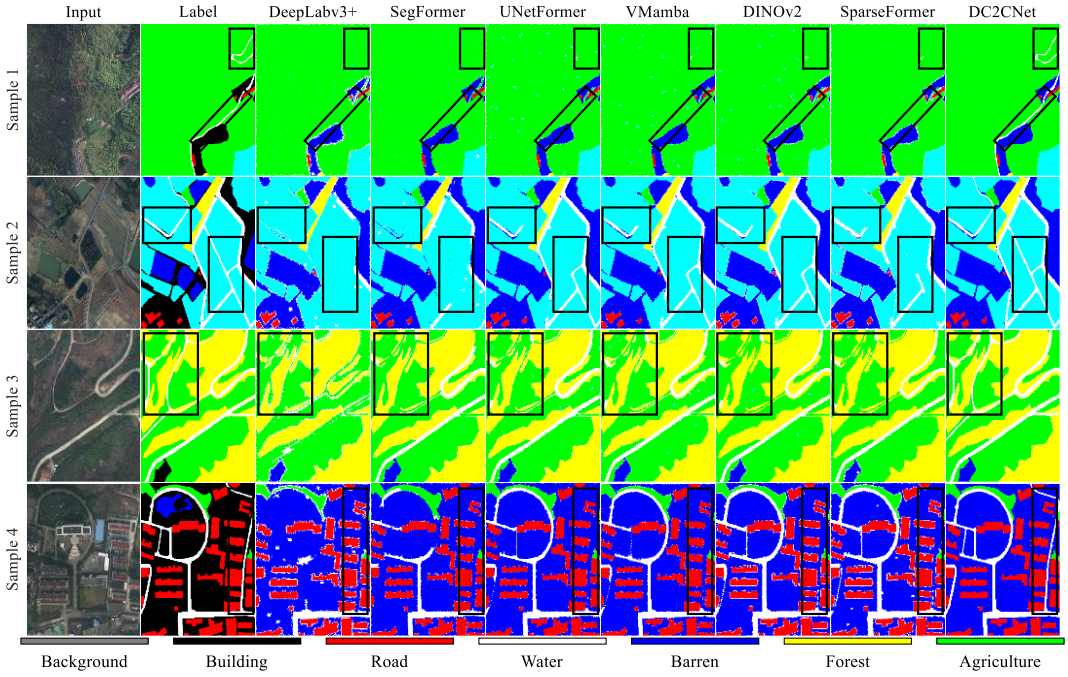


Figure 4. Visual comparison of segmentation results on the LoveDA dataset.

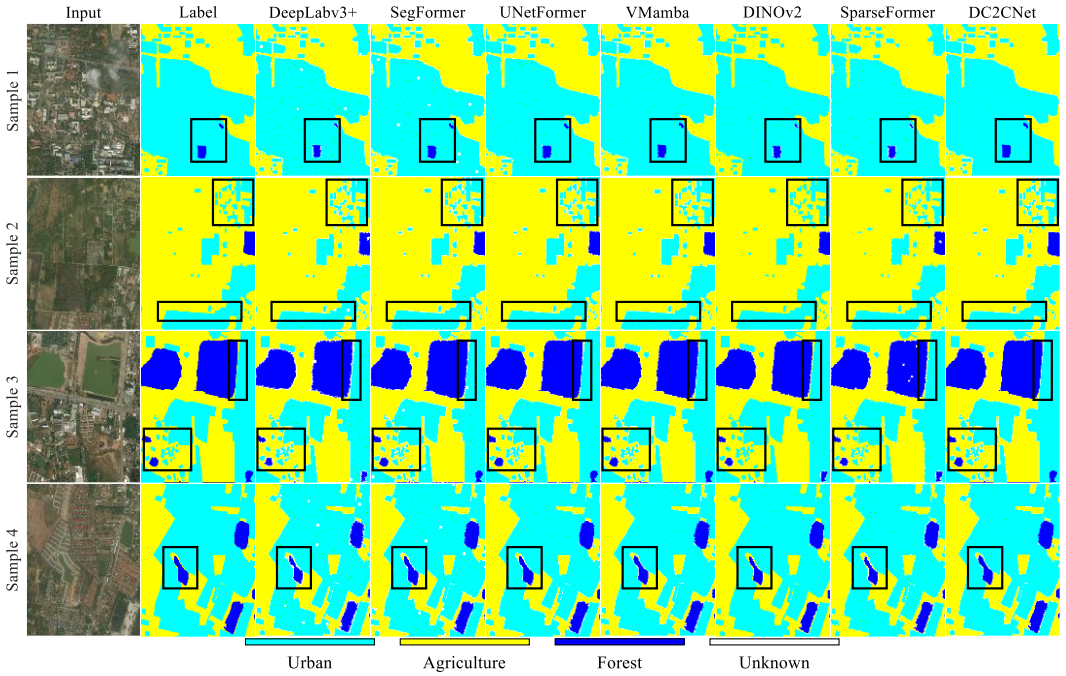


Figure 5. Visual comparison of segmentation results on the DeepGlobe dataset.

4.5. Ablation Study

To investigate the contribution of the proposed components and verify the effectiveness of the non-additive evidence fusion strategy, ablation experiments were conducted from two perspectives: component removal analysis and component replacement analysis. All experiments were performed on the Potsdam, LoveDA, and DeepGlobe datasets using identical training settings.

4.5.1. Component Removal Analysis

To evaluate the contribution of each proposed component, several variants were constructed by individually removing specific components from the complete DC2CNet

while keeping other modules unchanged. The results on different datasets are shown in Tables 7–9. The evaluated variants are defined as follows:

(1) w/o Region Membership Domain: The region membership domain is removed while all other components are kept unchanged. (2) w/o Boundary Separation Domain: The boundary separation domain and its corresponding boundary constraint are removed to evaluate whether boundary-related structural information provides complementary benefits beyond regional semantic representation**.** (3) w/o Dual-Domain Coupling: The coupling operation is removed, and the region and boundary domain features are directly fused without explicit interaction. (4) w/o Second-order Interaction: The second-order interaction term in the 2-additive fuzzy measure is removed by setting the coefficient of the second-order interaction term to zero ($\lambda = 0$), while the first-order fuzzy measure parameters and constraints remain unchanged. This variant evaluates whether pairwise evidence interactions provide additional benefits beyond first-order evidence aggregation. (5) Full DC2CNet: The complete DC2CNet with all proposed components is used as the reference model.

Table 7. Component removal analysis on the Potsdam dataset (%). Bold values indicate the best results.

	IoU per Category						mIoU	PA
	Imp. Surf.	**Building**	**Low Veg.**	**Tree**	**Car**	**Clutter**		
(1)	89.54	92.31	81.73	82.61	86.54	68.47	83.53 ± 0.11	92.62 ± 0.09
(2)	89.32	92.08	81.52	82.35	86.21	67.96	83.24 ± 0.13	92.43 ± 0.11
(3)	89.71	92.42	81.96	82.74	86.83	68.62	83.71 ± 0.10	92.78 ± 0.08
(4)	89.82	92.51	82.14	82.96	87.05	68.91	83.90 ± 0.09	92.86 ± 0.07
(5)	**90.10**	**92.90**	**82.50**	**83.40**	**87.60**	**69.20**	84.28 ± 0.14	93.05 ± 0.16

Table 8. Component removal analysis on the LoveDA dataset (%). Bold values indicate the best results.

	IoU per Category						mIoU	PA
	Building	**Road**	**Water**	**Barren**	**Forest**	**Agriculture**		
(1)	56.82	49.41	71.80	41.24	53.42	58.46	53.57 ± 0.14	71.81 ± 0.12
(2)	56.53	49.05	71.42	40.92	53.06	58.02	53.23 ± 0.16	71.55 ± 0.14
(3)	57.01	49.62	71.91	41.51	53.63	58.77	53.77 ± 0.13	71.97 ± 0.11
(4)	57.18	49.82	72.03	41.66	53.81	59.02	53.94 ± 0.12	72.12 ± 0.10
(5)	**57.40**	**50.20**	**72.10**	**42.00**	**54.00**	**60.10**	54.30 ± 0.18	72.50 ± 0.20

Table 9. Component removal analysis on the DeepGlobe dataset (%). Bold values indicate the best results.

	IoU per Category						mIoU	PA
	Urban	**Agriculture**	**Rangeland**	**Forest**	**Water**	**Barren**		
(1)	73.5	67.6	53.6	76.1	86.2	59.0	69.30 ± 0.18	88.47 ± 0.17
(2)	73.2	67.9	53.5	76.4	85.5	58.7	69.20 ± 0.19	88.20 ± 0.18
(3)	73.8	68.0	53.9	76.7	86.4	59.5	69.72 ± 0.17	88.75 ± 0.16
(4)	74.0	68.3	54.1	77.0	86.8	59.9	70.02 ± 0.15	89.08 ± 0.15
(5)	**74.2**	**68.6**	**54.3**	**77.2**	**87.3**	**60.2**	70.30 ± 0.16	89.40 ± 0.15

As shown in Tables 7–9, removing any individual component results in performance degradation across all datasets. Removing the region membership domain reduces semantic consistency within local regions, while removing the boundary separation domain mainly affects categories with complex structures, demonstrating the complementary role of boundary-aware information. The degradation caused by

removing dual-domain coupling indicates that explicit interaction between regional and boundary representations improves feature discrimination. In addition, removing second-order interactions consistently decreases performance, confirming the effectiveness of pairwise evidence modeling beyond first-order aggregation.

4.5.2. Component Replacement Analysis

To further verify the effectiveness of the proposed 2-additive fuzzy-measure fusion, the original capacity-based fusion strategy was replaced with conventional feature aggregation methods. The input evidence components and subsequent decoding process were kept unchanged. The results are presented in Table 10. The evaluated variants are defined as follows:

(1) Concatenation Fusion: All evidence components are directly concatenated and mapped to the prediction layer without explicit interaction modeling. (2) Weighted Summation Fusion: Evidence components are aggregated through weighted summation, where only independent contributions are considered. (3) Full DC2CNet: The complete DC2CNet with the proposed 2-additive fuzzy-measure fusion is used as the reference model.

Table 10. Component replacement analysis on different datasets (%). Bold values indicate the best results.

Dataset	(1)	(2)	(3)
Potsdam	83.47 ± 0.11	83.69 ± 0.10	**84.28 ± 0.14**
LoveDA	53.58 ± 0.17	53.76 ± 0.15	**54.30 ± 0.18**
DeepGlobe	69.48 ± 0.18	69.71 ± 0.17	**70.30 ± 0.16**

As shown in Table 10, DC2CNet consistently achieves better performance than conventional fusion strategies on all datasets. Although concatenation and weighted summation can combine multiple evidence components, they mainly model independent contributions and cannot explicitly represent interactions among evidence components. In contrast, the proposed 2-additive fuzzy-measure fusion considers both individual evidence contributions and pairwise interactions, enabling more effective evidence integration under weak supervision.

4.6. Analysis of Domain Discrepancy

To verify whether the region membership domain and boundary separation domain learn complementary rather than redundant representations, the correlation between region membership responses r_{pq} and boundary separation responses c_{pq} was measured after model training. Specifically, 10,000 neighboring pixel pairs were randomly sampled from each test dataset, and the Pearson correlation coefficient was calculated between the two domain responses. The Pearson correlation coefficient is defined as:

$$\rho_{r,c} = \frac{\sum_{i=1}^{N}(r_i - \bar{r})\,(c_i - \bar{c})}{\sqrt{\sum_{i=1}^{N}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{N}(c_i - \bar{c})^2}} \tag{37}$$

where N denotes the number of sampled pixel pairs, r_i and c_i represent the region membership response and boundary separation response of the i-th pixel pair, respectively. $\bar{r}$ and $\bar{c}$ denote the corresponding mean response values. The correlation results are reported in Table 11.

Table 11. Correlation between Region and Boundary Responses.

Dataset	Pearson Correlation
Potsdam	0.43

LoveDA	0.39
DeepGlobe	0.46

The obtained correlation values are moderate across all datasets, indicating that the two domains share related structural information but do not collapse into identical representations. The region membership domain mainly captures local semantic consistency, while the boundary separation domain focuses on structural transitions between adjacent regions.

Furthermore, a domain mapping exchange experiment was conducted to evaluate whether the two domain-specific mapping functions are interchangeable. Specifically, the mapping functions of the region membership domain and boundary separation domain were exchanged while keeping other components unchanged. The results are shown in Table 12.

Table 12. Effect of domain mapping exchange on segmentation performance.

Variant	Potsdam mIoU (%)	LoveDA mIoU (%)	DeepGlobe mIoU (%)
DC2CNet	84.28 ± 0.14	54.30 ± 0.18	70.30 ± 0.16
DC2CNet-Swap	83.82 ± 0.20	53.79 ± 0.21	69.74 ± 0.19

The performance degradation after exchanging the two mapping functions demonstrates that the two domains are not interchangeable. The domain-specific transformations enable the network to preserve different types of structural information, which contributes to the effectiveness of the proposed dual-domain representation.

4.7. Analysis of Model Properties

4.7.1. Capacity Constraint Analysis

In DC2CNet, the normalization and monotonicity properties introduced in Equations (27) and (29) are optimized through the regularization term in Equation (30). To evaluate whether the learned capacity parameters preserve these fuzzy-measure characteristics after optimization, the normalization deviation and monotonicity violation rate were analyzed on the Potsdam, LoveDA, and DeepGlobe datasets.

The normalization deviation is defined as:

$$E_{norm} = \frac{1}{|\Omega|}\sum_{p\in\Omega}\left|\mu_p(N) - 1\right| \tag{38}$$

where $\mu_p(N)$ denotes the full-set capacity at spatial position p, Ω represents the evaluated spatial positions, and $|\Omega|$ denotes their number. The monotonicity violation rate is defined as:

$$R_{mono} = \frac{\#\{\Delta_{p,i} < 0\}}{\#\{\Delta_{p,i}\}} \tag{39}$$

where $\Delta_{p,i}$ represents the capacity variation after adding the i-th evidence component at position p, and $\#(\cdot)$ denotes the number of elements satisfying the corresponding condition. A negative $\Delta_{p,i}$ indicates a monotonicity violation. The results are presented in Table 13.

Table 13. Capacity constraint analysis on different datasets.

Dataset	E_{norm}	R_{mono}
Potsdam	0.0031	0.42%
LoveDA	0.0045	0.57%
DeepGlobe	0.0038	0.49%

As shown in Table 13, the learned capacity parameters maintain small normalization deviations and low monotonicity violation rates on all datasets. The normalization deviation is below 0.005, and the monotonicity violation rate is lower than 0.6%, indicating that the regularization term in Equation (30) effectively maintains the desired fuzzy-measure characteristics.

4.7.2. Dual-Domain Coupling Analysis

The dual-domain coupling module is designed to adaptively integrate region membership information and boundary separation information. Unlike directly applying boundary features as a gate, the proposed coupling mechanism considers the interaction between the two domain responses and combines them through the proposed fusion strategy.

To analyze the activation behavior of the coupling module, the gate responses were investigated on different spatial regions. The results are reported in Table 14. For each pixel i, the coupling gate response is denoted as g_i. The average gate activation within a specific region Ω is defined as:

$$A_g(\Omega) = \frac{1}{|\Omega|}\sum_{i\in\Omega} g_i \tag{40}$$

where $|\Omega|$ represents the number of pixels in the region. The average gate activations for boundary regions Ω_b and non-boundary regions Ω_{nb} are calculated separately as:

$$A_g^b = \frac{1}{|\Omega_b|}\sum_{i\in\Omega_b} g_i \tag{41}$$

$$A_g^{nb} = \frac{1}{|\Omega_{nb}|}\sum_{i\in\Omega_{nb}} g_i \tag{42}$$

Table 14. Gate activation analysis in boundary and non-boundary regions.

Dataset	A_g^b	A_g^{nb}
Potsdam	0.78	0.34
LoveDA	0.74	0.31
DeepGlobe	0.76	0.36

The gate responses are consistently higher in boundary regions than in homogeneous regions, indicating that the coupling module mainly responds to local feature transitions rather than uniformly enhancing all regions.

Furthermore, the effect of dual-domain coupling on prediction discrimination was evaluated using pixel-level confidence. For each pixel, the prediction confidence is defined as:

$$C_i = \max_k\left(p_i^k\right) \tag{43}$$

where p_i^k denotes the predicted probability of class k for pixel i. The average confidence within a region is calculated as:

$$C(\Omega) = \frac{1}{|\Omega|}\sum_{i\in\Omega} C_i \tag{44}$$

The confidence values before and after applying the dual-domain coupling module are summarized in Table 15.

Table 15. Effect of dual-domain coupling on prediction confidence.

Dataset	Boundary Confidence w/o Coupling	Boundary Confidence with Coupling	Non-Boundary Confidence w/o Coupling	Non-Boundary Confidence with Coupling
Potsdam	0.71	0.78	0.91	0.93
LoveDA	0.68	0.75	0.88	0.91
DeepGlobe	0.70	0.77	0.89	0.92

The results show that the dual-domain coupling increases prediction confidence in boundary regions while maintaining stable responses in non-boundary regions. This indicates that the coupling module selectively enhances boundary-related information and suppresses unnecessary feature interference in homogeneous areas.

4.7.3. Small Object and Elongated Structure Analysis

To further evaluate the capability of DC2CNet in handling structure-sensitive categories, several representative classes related to small objects and elongated structures were selected for quantitative analysis. Specifically, Car and Clutter from the Potsdam dataset were used to evaluate small-object recognition and complex background discrimination, while Road and Barren from the LoveDA dataset were selected to analyze elongated structures and irregular regions. The experimental results are shown in Table 16, and the visualization results are shown in Figure 6. The black solid boxes mark the regions selected for detailed comparison.

Table 16. Segmentation performance comparison on small objects and elongated structures (%). Bold values indicate the best results.

Method	IoU per Category				Potsdam PA	LoveDA PA	Mean IoU
	Potsdam Car	Potsdam Clutter	LoveDA Road	LoveDA Barren			
DeepLabv3+	84.6	62.3	46.4	38.8	91.44 ± 0.22	70.2 ± 0.28	58.03
SegFormer	85.5	65.0	48.4	40.1	92.24 ± 0.21	71.3 ± 0.24	59.75
CATNet	85.0	52.8	48.0	39.8	92.55 ± 0.20	71.1 ± 0.28	56.40
UNetFormer	86.5	53.6	49.0	41.0	92.90 ± 0.17	72.0 ± 0.21	57.53
VMamba	86.9	68.3	49.3	41.5	93.00 ± 0.16	71.9 ± 0.20	61.50
Sparseformer	87.1	68.8	49.7	41.6	93.00 ± 0.15	70.1 ± 0.23	61.80
DC2CNet	**87.6**	**69.2**	**50.2**	**42.0**	93.05 ± 0.16	72.5 ± 0.20	62.25

As shown in Table 16, DC2CNet achieves the highest IoU on all four selected categories. Compared with the representative baseline methods, DC2CNet consistently improves the segmentation performance on Car, Clutter, Road, and Barren categories. These results indicate that the proposed method maintains better recognition ability for small-scale targets and structurally continuous regions under sparse pixel supervision.

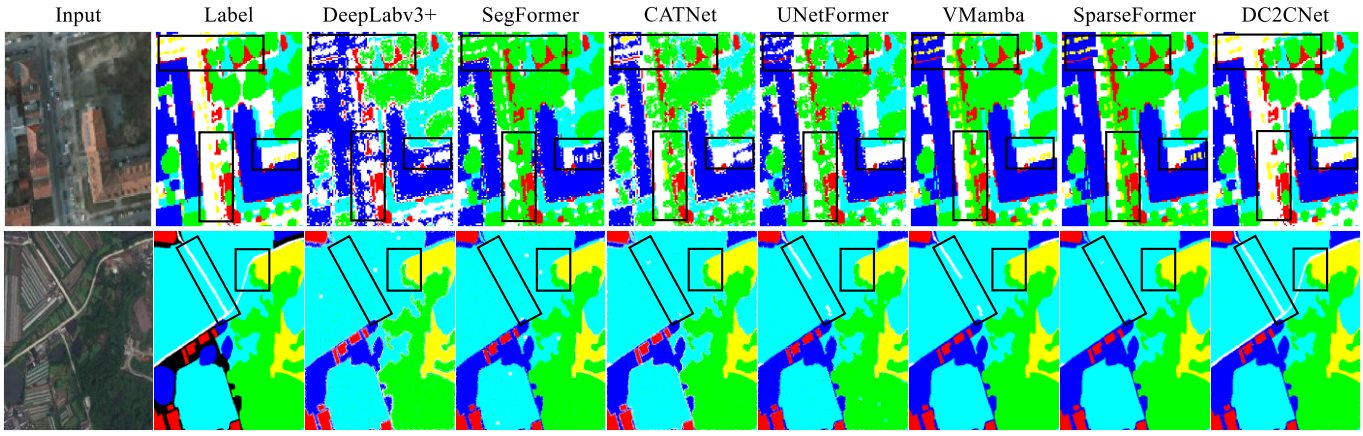


Figure 6. Visual comparison of segmentation results for small objects and elongated structures on Potsdam and LoveDA datasets. The color legends correspond to those in Figures 3 and 4, respectively.

The visualization results in Figure 6 are generally consistent with the quantitative results in Table 16. In the Potsdam scenes, the compared methods still exhibit missed regions, fragmented boundaries, or category confusion around small objects and complex backgrounds, whereas DC2CNet preserves more complete contours for Car and improves Clutter discrimination. In the LoveDA scenes, existing methods tend to produce fragmented predictions or inaccurate merging in Road and Barren regions, while DC2CNet maintains more continuous Road structures and clearer separation of Barren regions from surrounding classes.

4.7.4. Annotation Ratio and Training Efficiency Analysis

To analyze the performance of DC2CNet under different degrees of sparse supervision, experiments were conducted with annotation ratios of 5%, 10%, 20%, 50%, and full annotations. All settings adopted the same data splits, network architecture, training epochs, and optimization strategy, while only the proportion of labeled pixels involved in supervised loss calculation was changed. The results are reported in Table 17.

Table 17. Segmentation results of DC2CNet under different annotation ratios (%). Bold values indicate the best results.

Annotation Ratios	IoU per Category				Potsdam Train Time / h	LoveDA Train Time / h
	Potsdam mIoU	Potsdam PA	LoveDA mIoU	LoveDA PA		
1%	76.08	89.82	47.06	66.37	8.0	13.1
3%	78.79	90.84	49.32	68.41	8.0	13.2
5%	80.76	91.55	50.92	69.80	8.1	13.3
10%	84.28	93.05	54.30	72.50	8.3	13.6
20%	85.63	93.66	55.72	73.36	8.4	13.7
50%	86.52	94.12	56.84	74.18	8.5	13.8
Full	**87.20**	**94.45**	**57.48**	**74.62**	8.6	13.9

As shown in Table 17, the segmentation performance of DC2CNet on both datasets improves as the annotation ratio increases. Even under the more challenging 1% and 3% annotation settings, DC2CNet maintains reasonable segmentation performance, and the performance gradually improves with additional labeled pixels. The performance improvement from 5% to 10% annotations is relatively larger, while further increasing the annotation ratio leads to gradually smaller gains. This indicates that DC2CNet can effectively utilize limited pixel-level supervision and maintain stable segmentation performance under sparse annotation settings.

The training time remains relatively stable under different annotation ratios. Since the network architecture, input size, and optimization strategy are unchanged, the annotation ratio mainly affects the number of labeled pixels involved in loss calculation and has limited influence on the overall training cost.

4.7.5. Inference Crop Granularity Analysis

To further analyze the influence of inference crop granularity on small-object recognition, experiments were conducted using different image partition strategies during inference. The input crop size was kept consistent with the setting described in Section 4.2.2, while only the partition granularity during inference was modified. The training strategy and evaluation protocol remained unchanged. The third sample group

in Figure 3 was selected for analysis. The original image was divided into 16, 9, 4, and 1 subimages, respectively, and the predicted results were reconstructed into complete segmentation maps. The visualization results are shown in Figure 7.

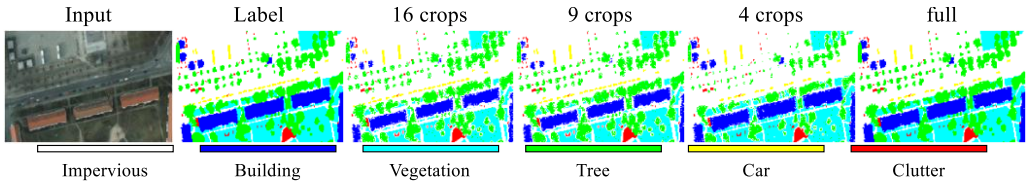


Figure 7. Visual comparison of DC2CNet segmentation results with different inference crop sizes.

As shown in Figure 7, the inference crop size affects Car recognition. With more subimages, small objects occupy a larger proportion of the local input region, which improves the preservation of vehicle responses. However, reduced contextual information may affect boundary completeness. In contrast, larger input regions provide richer contextual information but may weaken small-object responses due to the reduced relative object scale. These results indicate that small-object segmentation is influenced by both model representation and inference scale. Appropriate inference strategies are therefore important for improving the stability of small-object recognition in high-resolution remote sensing images.

4.8. Computational Efficiency Analysis

To further analyze the computational cost of DC2CNet, the number of parameters (Params), floating-point operations (GFLOPs), peak GPU memory usage, and inference latency were compared with representative baseline methods. All efficiency measurements followed the experimental environment described in Section 4.2.1. The experiments were conducted on the Potsdam dataset with the same input size and batch setting. The computational efficiency comparison is summarized in Table 18.

Table 18. Computational efficiency comparison of different segmentation methods. Bold values indicate the best results.

Method	Params (M)	GFLOPs	Memory (GB)	Latency (ms)
FCN [16]	15.3	35.6	2.1	18.7
PSPNet [15]	46.7	91.2	4.3	34.5
DeepLabV3+ [20]	41.2	78.6	3.9	31.8
SegFormer [25]	27.5	62.4	3.6	29.7
RSSFormer [18]	32.8	70.5	4.0	33.2
Gated-SCNN [31]	31.4	84.7	4.5	38.6
BoundaryFormer [32]	36.2	89.3	4.7	41.5
FaPN [39]	28.9	65.8	3.8	32.1
UNetFormer [37]	30.6	72.4	4.1	35.6
VMamba [24]	35.8	76.9	4.3	36.8
DINOv2 [44]	86.5	145.7	8.6	65.4
SparseFormer [48]	33.7	74.2	4.2	37.9
DC2CNet	**38.5**	**82.6**	**4.9**	**42.3**

As shown in Table 18, CNN-based methods generally achieve lower computational costs due to their efficient convolutional architectures, while Transformer-based and Mamba-based methods provide stronger representation capabilities with additional overhead. Boundary-aware and multiscale fusion methods introduce extra feature interactions, leading to increased complexity. DC2CNet introduces dual-domain feature

modeling and fuzzy-measure fusion, which increases computational complexity but maintains comparable efficiency while improving segmentation performance.

5. Conclusions and Future Work

A Dual-Domain Coupled 2-Additive Fuzzy Measure Network (DC2CNet) was proposed for weakly supervised remote sensing image segmentation. The proposed framework introduces dual-domain representation and fuzzy-measure-based feature interaction to enhance regional consistency and boundary discrimination under sparse pixel supervision.

Experimental results on the Potsdam, LoveDA, and DeepGlobe datasets demonstrate that DC2CNet achieves competitive segmentation performance compared with representative segmentation methods. The ablation studies verify the effectiveness of the proposed region membership domain, boundary separation domain, dual-domain coupling module, and 2-additive fuzzy-measure decoding mechanism. Additional experiments under different annotation ratios further demonstrate the effectiveness of DC2CNet under limited supervision.

Despite these advantages, the proposed method still has some limitations. The segmentation performance of small objects may be affected by the balance between local details and surrounding contextual information, especially in complex high-resolution scenes. Future work will investigate more adaptive multiscale feature interaction strategies, computationally efficient structural modeling methods, and the robustness of the proposed framework under degraded imaging conditions. In addition, the applicability of DC2CNet to broader remote sensing scenarios and different annotation settings will be further explored.

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

Funding: This research was funded by National Natural Science Foundation of China, grant number 71771110.

Data Availability Statement: The datasets used in this study are publicly available from their official releases. The Potsdam dataset was obtained from the ISPRS 2D Semantic Labeling Benchmark: https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx. The LoveDA dataset was obtained from the official LoveDA benchmark release: https://github.com/Junjue-Wang/LoveDA. The DeepGlobe Land Cover Classification dataset was obtained from the official DeepGlobe Satellite Image Understanding Challenge: https://deepglobe.org/. The dataset links and corresponding data preparation details are provided in the manuscript to facilitate reproduction. The source code of DC2CNet is not publicly available at this stage due to ongoing intellectual property protection procedures.

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

References

1. Li, X.; Qian, S.; Lyu, X.; Song, Y.; Liu, F.; Fang, Y.; Xu, Z.; Kaup, A. Frequency–spatial decoupled co-modeling transformer for fine-grained remote sensing image segmentation. *Inf. Sci.* **2026**, *747*, 123489. https://doi.org/10.1016/j.ins.2026.123489.
2. Feng, Z.; Yang, B.; Deng, B. SAM-assisted multimodal collaborative enhancement for remote sensing image segmentation. *Inf. Fusion* **2026**, *135*, 104432. https://doi.org/10.1016/j.inffus.2026.104432.

3. Wu, X.; Wang, L.; Wu, C.; Guo, C.; Yan, H.; Qiao, Z. Semantic Segmentation of Remote Sensing Images Using Multiway Fusion Network. *Signal Process.* **2024**, *215*, 109272. https://doi.org/10.1016/j.sigpro.2023.109272.
4. Zhou, Y.; Su, H.; Wang, T.; Hu, Q. Onet: Twin U-Net Architecture for Unsupervised Binary Semantic Segmentation in Radar and Remote Sensing Images. *IEEE Trans. Image Process.* **2025**, *34*, 2161–2172. https://doi.org/10.1109/TIP.2025.3530816.
5. Yang, Y.; Zheng, S.; Wang, X.; Ao, W.; Liu, Z. AMMUNet: Multiscale Attention Map Merging for Remote Sensing Image Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5. https://doi.org/10.1109/lgrs.2024.3506718.
6. Suresh, T.; Sundaralingam, H.; Akilan, T.; Ahmed, S.B. Ss-deepseg: An efficient deeplab with smart scaling for robust semantic segmentation. In *Proceedings of the 2025 IEEE 34th International Symposium on Industrial Electronics (ISIE)*; IEEE: New York, NY, USA, 2025; pp. 1–6.
7. Duan, Y.; Yang, D.; Qu, X.; Zhang, L.; Qu, J.; Chao, L.; Gan, P. LightCGS-Net: A novel lightweight road extraction method for remote sensing images combining global semantics and spatial details. *J. Netw. Comput. Appl.* **2025**, *242*, 104247. https://doi.org/10.1016/j.jnca.2025.104247.
8. Tang, P.; Wang, G.; Zhu, H. PVTv2-SegFormer: A Lightweight Transformer for Edge-Efficient Ship Segmentation in Remote Sensing Imagery. In *Proceedings of the 2025 China Automation Congress (CAC)*; IEEE: New York, NY, USA, 2025; pp. 3141–3146.
9. Sun, H.; He, X.; Li, H.; Kong, J.; Qiao, M.; Cheng, X.; Li, P.; Zhang, J.; Liu, R.; Shang, J. Adaptive sparse lightweight multi-scale hybrid network for remote sensing image semantic segmentation. *Expert Syst. Appl.* **2025**, *280*, 127347. https://doi.org/10.1016/j.eswa.2025.127347.
10. Lu, C.; Zhang, X.; Du, K.; Xu, H.; Liu, G. CTCFNet: CNN-Transformer Complementary and Fusion Network for High-Resolution Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. https://doi.org/10.1109/tgrs.2024.3458446.
11. Liang, B.; Luo, H. MEANet: An effective and lightweight solution for salient object detection in optical remote sensing images. *Expert Syst. Appl.* **2024**, *238*, 121778. https://doi.org/10.1016/j.eswa.2023.121778.
12. Sun, Y.; Liang, D.; Li, S.-Y.; Chen, S.; Huang, S.-J. Handling Noisy Annotation for Remote Sensing Semantic Segmentation via Boundary-Aware Knowledge Distillation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–20. https://doi.org/10.1109/tgrs.2025.3562073.
13. Wang, G.; Zhang, X.; Lai, J.; Peng, Z.; Zhang, T.; Wang, C.; Tang, X.; Jiao, L. Expanding then fusing: Weakly-supervised remote sensing semantic segmentation via progressive multi-modal fusion. *Pattern Recognit.* **2026**, *177*, 113392. https://doi.org/10.1016/j.patcog.2026.113392.
14. Li, Y.; Hu, C.; Ren, K.; Xi, H.; Fan, J. Multi-representation fusion learning for weakly supervised semantic segmentation. *Expert Syst. Appl.* **2025**, *277*, 127222. https://doi.org/10.1016/j.eswa.2025.127222.
15. Jiang, J.; Liu, H. Semantic segmentation algorithm for remote sensing images based on PSPNet. In *Proceedings of the Third International Conference on Environmental Remote Sensing and Geographic Information Technology (ERSGIT 2024)*; SPIE: Bellingham, WA, USA, 2025; pp. 91–96.
16. Hu, Q.; Wang, F.; Fang, J.; Li, Y. Semantic Labeling of High-Resolution Images Combining a Self-Cascaded Multimodal Fully Convolution Neural Network with Fully Conditional Random Field. *Remote Sens.* **2024**, *16*, 3300. https://doi.org/10.3390/rs16173300.
17. Dimitrovski, I.; Spasev, V.; Loshkovska, S.; Kitanovski, I. U-Net Ensemble for Enhanced Semantic Segmentation in Remote Sensing Imagery. *Remote Sens.* **2024**, *16*, 2077. https://doi.org/10.3390/rs16122077.
18. Xu, R.; Wang, C.; Zhang, J.; Xu, S.; Meng, W.; Zhang, X. RSSFormer: Foreground Saliency Enhancement for Remote Sensing Land-Cover Segmentation. *IEEE Trans. Image Process.* **2023**, *32*, 1052–1064. https://doi.org/10.1109/TIP.2023.3238648.
19. Feng, Y.; Fan, Z.; Yan, Y.; Jiang, Z.; Zhang, S. MFAFNet: Multi-Scale Feature Adaptive Fusion Network Based on DeepLab V3+ for Cloud and Cloud Shadow Segmentation. *Remote Sens.* **2025**, *17*, 1229. https://doi.org/10.3390/rs17071229.
20. Zeng, S.; Yao, Y. Research on an Improved Deeplabv3+-Based Algorithm for Urban Remote Sensing Image Segmentation. In *Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence*; ACM: New York, NY, USA, 2025; pp. 609–615.
21. Li, X.; Xu, F.; Li, J.; Su, Y.; Li, L.; Lyu, X.; Xu, Z.; Kaup, A. Frequency domain-enhanced spectral-spatial fusion transformer for semantic segmentation of remote sensing images. *Inf. Fusion* **2026**, *132*, 104248. https://doi.org/10.1016/j.inffus.2026.104248.
22. Du, J.; Peng, H.; Li, B.; Liu, Z. Local-Contextual Feature Fusion Network Based on Nonlinear Spiking Neural Model for Semantic Segmentation of Remote Sensing Images. *Int. J. Neural Syst.* **2026**, *36*, 2650029. https://doi.org/10.1142/S0129065726500292.

23. Wu, H.; Zhang, M.; Huang, P.; Tang, W. CMLFormer: CNN and Multiscale Local-Context Transformer Network for Remote Sensing Images Semantic Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 7233–7241. https://doi.org/10.1109/jstars.2024.3375313.
24. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. VMamba: Visual State Space Model. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2024.
25. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
26. Wang, C.; Duan, P.; Li, J. DFPF-Net: Dynamically Focused Progressive Fusion Network for Remote Sensing Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 5905–5918. https://doi.org/10.1109/jstars.2025.3531658.
27. Peng, S.; Xie, H.; Liu, N.; Zeng, Y. Semantic Segmentation of Multispectral Remote Sensing Imagery for Coastal Wetlands with SegFormer. *Remote Sens.* **2026**, *18*, 745. https://doi.org/10.3390/rs18050745.
28. Li, Z.; Zhang, Y.; Hu, J.; Wu, K.; Miao, J.; Wu, J.; Zhang, X.; Yang, B.; Xia, Z. HDC-Net:A multimodal remote sensing semantic segmentation network with hierarchical dual-stream fusion and cross-token interaction. *Knowl.-Based Syst.* **2026**, *337*, 115416. https://doi.org/10.1016/j.knosys.2026.115416.
29. Zhan, C.; Yang, K. WCMamba: Enhancing high-resolution remote sensing image semantic segmentation with pyramid wavelet convolution and SS2D. *Knowl.-Based Syst.* **2025**, *324*, 113877. https://doi.org/10.1016/j.knosys.2025.113877.
30. Li, Z.; Wang, Y.; Xu, D.; Gao, Y.; Zhao, T. TBNet: A texture and boundary-aware network for small weak object detection in remote-sensing imagery. *Pattern Recognit.* **2025**, *158*, 110976. https://doi.org/10.1016/j.patcog.2024.110976.
31. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 5228–5237.
32. Lazarow, J.; Xu, W.; Tu, Z. Instance Segmentation with Mask-supervised Polygonal Boundary Transformers. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18–24 June 2022*; IEEE: New York, NY, USA, 2022; pp. 4372–4381.
33. Qian, H.; Liu, X.; Li, Z.; Ma, Y.; Lu, Z. GS-USTNet: Global–Local Adaptive Convolution with Skip-Guided Attention for Remote Sensing Image Segmentation. *Remote Sens.* **2026**, *18*, 785. https://doi.org/10.3390/rs18050785.
34. Liu, Y.; Li, H.; Hu, C.; Luo, S.; Luo, Y.; Chen, C.W. Learning to Aggregate Multi-Scale Context for Instance Segmentation in Remote Sensing Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 595–609. https://doi.org/10.1109/TNNLS.2023.3336563.
35. Cao, Y.; Liu, C.; Wu, Z.; Zhang, L.; Yang, L. Remote Sensing Image Segmentation Using Vision Mamba and Multi-Scale Multi-Frequency Feature Fusion. *Remote Sens.* **2025**, *17*, 1390. https://doi.org/10.3390/rs17081390.
36. Wang, K.; He, D.; Sun, Q.; Yi, L.; Yuan, X.; Wang, Y. A novel network for semantic segmentation of landslide areas in remote sensing images with multi-branch and multi-scale fusion. *Appl. Soft Comput.* **2024**, *158*, 111542. https://doi.org/10.1016/j.asoc.2024.111542.
37. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. https://doi.org/10.1016/j.isprsjprs.2022.06.008.
38. Zhao, J.; Zhang, M.; Zhou, Z.; Wang, Z.; Lang, F.; Shi, H.; Zheng, N. CFFormer: A Cross-Fusion Transformer Framework for the Semantic Segmentation of Multisource Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–17. https://doi.org/10.1109/tgrs.2024.3507274.
39. Zhou, C.; Cai, B.; Xiong, C.; Liu, R. Edge-Guided and Multiscale Feature Optimization-Based Semantic Segmentation Network for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 1–16. https://doi.org/10.1109/tgrs.2026.3660697.
40. Bai, C.; Wu, K.; Bai, X.; Wu, X. Spectral–spatial collaborative perception and dynamic semantic reasoning for remote sensing image segmentation. *Eng. Appl. Artif. Intell.* **2026**, *181*, 115269. https://doi.org/10.1016/j.engappai.2026.115269.
41. Grabisch, M. K-order additive discrete fuzzy measures and their representation. *Fuzzy Sets Syst.* **1997**, *92*, 167–189. https://doi.org/10.1016/s0165-0114(97)00168-1.
42. 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. https://doi.org/10.1109/TPAMI.2024.3362475.
43. Li, J.; Wang, Z.; Xu, N.; You, Z.; Huang, D.-S. Vision Foundation Model-Driven Multiscale Expert Tuning for Multimodal Remote Sensing Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–17. https://doi.org/10.1109/tgrs.2025.3627904.

44. Fernandez, P.; Bojanowski, P.; Synnaeve, G.; Szafraniec, M.; Oquab, M.; Joulin, A.; Xu, H.; Galuba, W.; Sharma, V.; Darcet, T.; et al. DINOv2: Learning Robust Visual Features Without Supervision. In Proceedings of the ICLR, Singapore, 24–28 April 2025.
45. 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 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France 1–6 October 2023*; IEEE: New York, NY, USA, 2023; pp. 3992–4003.
46. Luo, X.; Xie, T.; Qin, W.; Duan, Z.; Tan, J.; Zhang, T. Combining hierarchical sparse representation with adaptive prompt for few-shot segmentation. *Expert Syst. Appl.* **2025**, *260*, 125377. https://doi.org/10.1016/j.eswa.2024.125377.
47. Yang, C.; Wang, J.; Meng, H.; Yang, S.; Feng, Z. Negative Class Guided Spatial Consistency Network for Sparsely Supervised Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 657–669. https://doi.org/10.1109/TCSVT.2024.3457622.
48. Gao, Z.; Tong, Z.; Wang, L.; Shou, M.Z. Sparseformer: Sparse visual recognition via limited latent tokens. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
49. Cheng, B.; Girshick, R.; Dollár, P.; Berg, A.C.; Kirillov, A. Boundary IoU: Improving Object-Centric Image Segmentation Evaluation. In *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 15329–15337.

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.