mathematics

MDPI

Article

Class-Aware Self- and Cross-Attention Network for Few-Shot Semantic Segmentation of Remote Sensing Images

Guozhen Liang [1,†], Fengxi Xie [1,†] and Ying-Ren Chien [2,*]

[1] Department of Electrical Engineering and Computer Science, Technische Universität Berlin, 10623 Berlin, Germany; guozhen.liang@campus.tu-berlin.de (G.L.); fengxi.xie@campus.tu-berlin.de (F.X.)
[2] Department of Electrical Engineering, National Ilan University, Yilan 260007, Taiwan
* Correspondence: yrchien@niu.edu.tw
† These authors contributed equally to this work.

Abstract: Few-Shot Semantic Segmentation (FSS) has drawn massive attention recently due to its remarkable ability to segment novel-class objects given only a handful of support samples. However, current FSS methods mainly focus on natural images and pay little attention to more practical and challenging scenarios, e.g., remote sensing image segmentation. In the field of remote sensing image analysis, the characteristics of remote sensing images, like complex backgrounds and tiny foreground objects, make novel-class segmentation challenging. To cope with these obstacles, we propose a Class-Aware Self- and Cross-Attention Network (CSCANet) for FSS in remote sensing imagery, consisting of a lightweight self-attention module and a supervised prior-guided cross-attention module. Concretely, the self-attention module abstracts robust unseen-class information from support features, while the cross-attention module generates a superior quality query attention map for directing the network to focus on novel objects. Experiments demonstrate that our CSCANet achieves outstanding performance on the standard remote sensing FSS benchmark iSAID-5^i, surpassing the existing state-of-the-art FSS models across all combinations of backbone networks and *K*-shot settings.

Keywords: few-shot learning; few-shot semantic segmentation; remote sensing; class-aware self- and cross-attention

MSC: 68U05; 68U10

check for updates

Citation: Liang, G.; Xie, F.; Chien, Y.-R. Class-Aware Self- and Cross-Attention Network for Few-Shot Semantic Segmentation of Remote Sensing Images. *Mathematics* **2024**, *12*, 2761. https://doi.org/10.3390/math12172761

Academic Editors: Volodymyr Ponomaryov, Vladimir Lukin, Bogdan Smolka and Beatriz P. García Salgado

Received: 7 August 2024
Revised: 29 August 2024
Accepted: 5 September 2024
Published: 6 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Remote sensing image analysis has greatly contributed to academic research, industrial development, and public affairs management, as remote sensing images are rich in geographical information [1–3]. In the context of remote sensing image analysis, semantic segmentation aims to assign predefined geospatial categories to the images at pixel level [4]. The emergence of convolutional neural networks (CNNs) has significantly advanced the development of semantic segmentation [5–8]. However, the remarkable performance of these CNN-based models relies heavily on large datasets. In addition, traditional semantic segmentation models struggle to generalize to classes that are absent from the training dataset.

To deal with these problems, Few-Shot Semantic Segmentation (FSS) has been developed. This technique enables the deep models to segment novel-class objects with scarce support examples, which has been proven effective in low-data scenarios [9]. The conceptualization of FSS was first defined by Shaban et al. [9]. Afterward, many researchers proposed their own insights and pushed the performance of FSS to a new limit. Zhang et al. [10] incorporated an attention module and an iterative optimization method into FSS, where the support information is successfully merged and the segmentation results are improved recursively. Lang et al. [11] proposed a base learner and an ensemble module to suppress the false-positive prediction caused by the similarities between base classes and novel

classes. Despite impressive results, these methods mainly focused on the segmentation of natural images, and few works investigated real-world scenarios [12–14]. The images of these application scenarios have special properties and pose great challenges to the segmentation task. For instance, remote sensing images, which are investigated in this paper, have greater foreground–background class similarity and more tiny objects compared with natural images. It can be observed in the first row of Figure 1, the target class ship, ground track field and harbor are greatly similar to the background class harbor, grassland and river bank, respectively. In addition, there is usually more than one target object to be segmented in an image, and in some circumstances, they are too tiny to identify (as shown in the second row of Figure 1). These unique characteristics would undoubtedly lead to unsatisfactory predictions in the existing FSS frameworks (e.g., false activation and coarse boundaries).

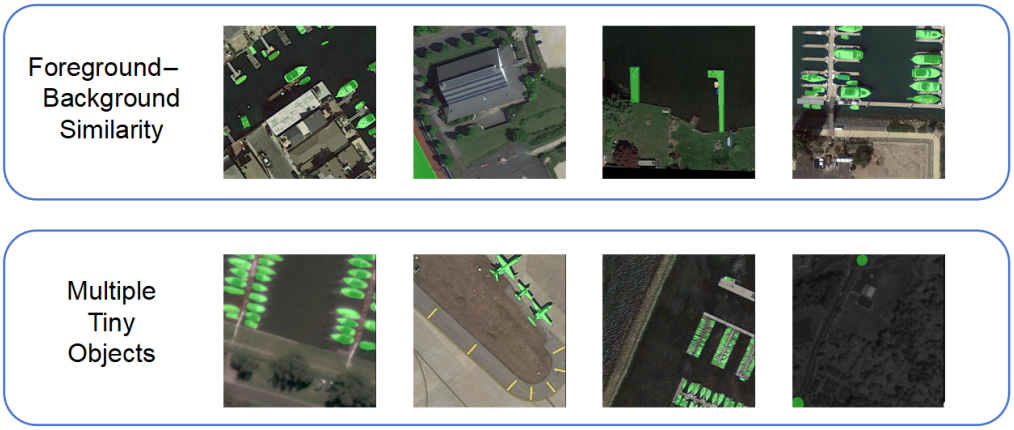

Figure 1. Characteristics of remote sensing images.

Furthermore, prevalent FSS approaches are mostly built on metric learning, which can be divided into affinity learning [15–17] and prototype learning [18–21]. Affinity-learning-based methods usually establish pixel-level support–query correspondences, which are then aggregated into query prediction. These methods, however, failed to utilize the semantic information from the extracted features and resulted in imperfect predictions.

In contrast, prototypical FSS approaches leverage one or two rich semantic class-wise prototypes to construct prototype–query connections for query segmentation. For instance, SG-One [10] applied masked average pooling (MAP) over support features to generate the class representative prototype vectors, against which the query feature is matched by the cosine similarity metric to yield query segmentation. More recently, researchers have striven to elevate the performance of the prototypical FSS paradigm by obtaining more guidance from class-wise prototypes such as PPNet [22], PFENet [20], ASGNet [21] and SD-AANet [17]. However, depending solely on scarce compressed prototypes is bound to incur information loss, making it difficult to deal with challenging scenarios in remote sensing image segmentation.

To cope with the aforementioned problems, we proposed a Class-Aware Self- and Cross-Attention Network (CSCANet) for the FSS of remote sensing images. The proposed CSCANet consists of the self-attention module (SAM) and the prior-guided supervised cross-attention module (PG-CAM). Firstly, a CBAM [23]-like self-attention module is designed to exploit unseen-class information from support images. Specifically, we incorporate a weighted max pooling branch to extract robust discriminative novel-class features. Secondly, a prior-guided supervised cross-attention mechanism is proposed to direct our CSCANet to concentrate on the unseen classes in the query set. In detail, we first generate the prior similarity mask by measuring the cosine similarity between the intermediate-level support and query features. The prior similarity mask and support masks, along with

support and query features, are fed into the cross-attention module to yield a high-quality affinity attention map.

In summary, the contributions of our work include the following:

- We devise an efficient self-attention module, which makes use of support features and the corresponding ground-truth mask to mine the unseen-class information distinct from the background classes.
- We propose a prior-guided supervised cross-attention module to generate a high-quality query attention map. The query attention map can outline the tiny objects in images, which enhances the network's ability to segment tiny targets.
- The CSCANet outperforms the existing FSS methods across almost all the combinations of backbone networks (VGG-16, ResNet-50) and few-shot settings (one-shot and five-shot) on the standard remote sensing benchmark iSAID-5^i.

2. Related Work

2.1. Semantic Segmentation

Semantic segmentation stands as a foundational computer vision task with the primary goal of accomplishing pixel-level classification in images, categorizing each pixel into annotated semantic categories. Benefiting from the emergence of fully convolutional networks (FCNs) [5], significant advancements in this field have been achieved. For example, Unet [24] adopted an encoder–decoder-like architecture to generate the predicted mask in a symmetric manner. Later on, PSPNet [25] incorporated a pyramid pooling module to enhance the robustness of image features. In addition, an attention mechanism was also employed to direct the network to focus on the foreground regions [26]. Although traditional segmentation models have achieved impressive performance, they face a challenge in effectively adapting to novel-class objects as they heavily depend on a substantial number of annotated samples, hindering their practical applications to some extent.

2.2. Few-Shot Learning

Few-shot learning (FSL) aims to train models with scarce labeled examples, promoting the generalization ability of deep networks in scenarios with limited data. Most of the prevalent FSL approaches are implemented within the meta-learning paradigm [27], which has three subdivisions: metric-based [28–30], optimization-based [31–33] and augmentation-based [34]. Our work is built upon the metric-based approaches, where distance metrics (e.g., cosine distance, Euclidean distance) are leveraged to measure the support–query similarities.

2.3. Few-Shot Semantic Segmentation

Few-Shot Semantic Segmentation (FSS) has gained massive attention as an extension of FSL. FSS aims to adapt deep networks to predict pixel-to-pixel correspondence between support–query image pairs. This technique facilitates unseen-class segmentation, making it a promising solution for challenges in low-data regimes. The problem of FSS was initially formulated by Shaban et al. [9]. They proposed OSLSM to make query predictions using a classifier trained on the support branch. After that, Zhang et al. [10] proposed the first end-to-end prototypical FSS framework, which has become the paradigm in the field of FSS. ASGNet [21] adaptively extracted multiple prototypes according to the feature similarity and allocated them in the prototype–query matching based on an attention-like algorithm. Lang et al. [11] proposed a novel FSS paradigm where an auxiliary base learner was leveraged to explicitly identify confusing target regions that are similar to the base-class objects.

However, existing prevalent methods are mainly designed for natural image segmentation, which fails to consider the tricky properties of remote sensing images. Wang et al. [14] proposed a metametric-based FSS framework for few-shot geographical image segmentation, where the feature comparison sub-branch and affinity-based feature aggregation were introduced to improve the predictions. Lang et al. [35] designed a few-shot remote sensing image segmentation framework, in which the proposed global rectification and decouple

registration mechanism address the inter-class similarity and intra-class diversity to some extent. Nevertheless, these approaches did not thoroughly solve the aforementioned complicated cases in remote sensing image segmentation. Therefore, we propose a lightweight self-attention module and a supervised cross-attention module to solve these problems and push the performance to a new level.

3. Methodology

In this section, we first introduce the problem setting in Section 3.1. The overall architecture of our CSCANet is mentioned in Section 3.2. Then, in Sections 3.3 and 3.4, we describe our lightweight self-attention block and prior-guided supervised cross-attention block in detail, respectively. Section 3.5 is about the ASPP module and classifier. Finally, we briefly introduce the *K*-shot setting of our proposed method in Section 3.6.

3.1. Problem Definition

The goal of Few-Shot Semantic Segmentation is to segment novel-class targets with merely a few annotated exemplars. The training process of FSS models is usually performed within the meta-learning paradigm, also known as episodic training [36]. To ensure a reliable generalization ability, the model training and testing phases are separately performed on two subsets D_{train} (sufficient base classes) and D_{test} (scarce unseen classes) with no overlapped classes. Both image sets contain a series of episodes. Each episode includes a small number of support sets $S = \left\{\left(I_s^i, M_s^i\right)\right\}_{i=1}^{K}$ and a query set $Q = \left\{\left(I_q, M_q\right)\right\}$, where I_* denotes a raw image and M_* the corresponding ground-truth mask. In each episode of training, a support set S and a query image I_q are input to the model, with each query prediction supervised by its corresponding ground-truth mask. During each episode of the testing stage, the model is tested on D_{test} to assess the performance.

3.2. Overall Framework

Figure 2 depicts the overall architecture of our CSCANet under a 1-shot setting. Initially, a pre-trained backbone network is utilized to extract support and query features from input image sets. The support features F_s^2 of block2 and F_s^3 of block3 are concatenated and then processed by a 1×1 convolution to generate the intermediate-level support feature F_s^{23}:

$$F_s^{23} = Conv_{1\times1}\{F_s^2 \copyright F_s^3\}, \tag{1}$$

where © represents the concatenate operation. Thereafter, the support prototypes V_s can be calculated as follows:

$$F_{masked}^{23} = F_s^{23} \odot \zeta(M_s), \tag{2}$$

$$V_s = \mathcal{F}_{avg\_pool}(F_{masked}^{23}), \tag{3}$$

Here, $\odot$ denotes element-wise multiplication, ζ is the bi-linear interpolation function such that $\mathbb{R}^{H\times W} \rightarrow \mathbb{R}^{c\times h\times w}$. F_{avg_pool} represents the average pooling operation. In the self-attention module, the support feature F_s^{23} and its corresponding support mask are utilized to yield the support attention feature map A_s. Thereafter, the support and query features, as well as the prototype vector, are fed into the cross-attention module to yield a query attention map. Subsequently, the support attention feature map, query attention map and prototype vector, along with the query feature, are input to a dilated ASPP module for feature refinement. The enriched feature is processed by the classifier, where 3×3 and 1×1 convolution are applied to generate the query prediction.

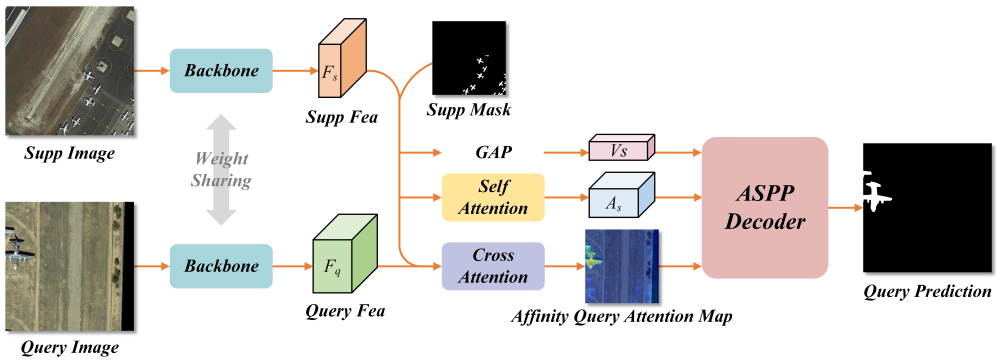

Figure 2. Meta learner of our proposed CSCANet.

3.3. Self-Attention Module

In the context of limited cues provided by the support prototypes, we proposed an efficient self-attention module to exploit novel-class cues from the scarce support images, which guides the network to concentrate on the unseen-class objects and avoid false activation. As shown in Figure 3, we first generate the pooling vector as follows:

$$V_{pool} = \mathcal{F}_{avg\_pool}(F^{23}_{masked}) \oplus \alpha * \mathcal{F}_{max\_pool}(F^{23}_{masked}), \tag{4}$$

Here, F_{max_pool} denotes the max pooling operation, and $\oplus$ represents the element-wise addition. The average pooling operation is employed to extract the global general features of the novel-class objects, while the max pooling operation is applied to abstract the local discriminative unseen-class features. However, we notice that directly incorporating the max pooling branch will result in a non-uniform feature representation of the novel classes. Therefore, we adopt a learnable parameter α to weight the max pooling branch and mitigate this side effect. We set the initial value of α to 1. Subsequently, the attention vector can be derived as follows:

$$V_a = \sigma(Conv_N(V_{pool})), \tag{5}$$

where $Conv_N$ refers to a series of convolutional layers, and σ denotes the activation function Sigmoid, respectively.

Finally, a foreground-focused support attention map is generated as follows:

$$A_s = F^{23}_s \odot V_a, \tag{6}$$

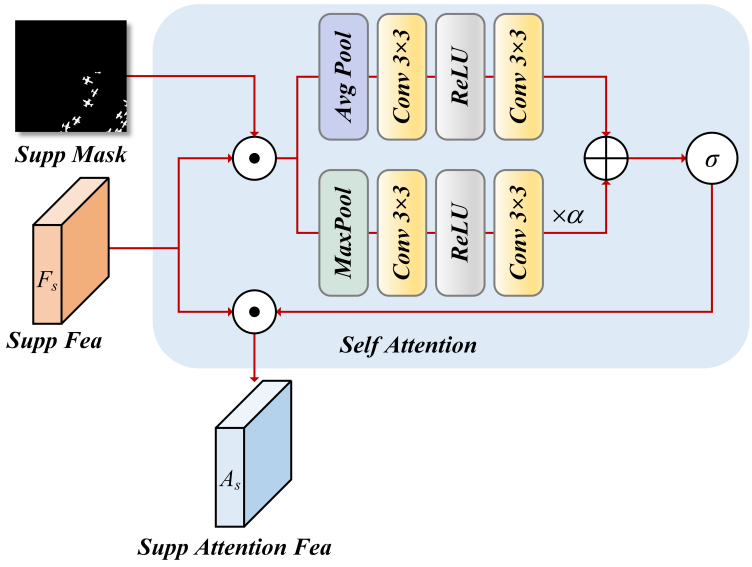

Figure 3. Architecture of the proposed SAM in 1-shot setting.

3.4. Prior-Guided Supervised Cross-Attention Module

A high-quality query attention map is an important hint for accurate novel-class segmentation. We proposed a prior-guided supervised cross-attention block to generate such an attention map, which is capable of accurately capturing the query targets regardless of their sizes. PFENet [20] introduced a similar attention mechanism, where the cosine similarity between the deepest support and query features is calculated to generate a query attention map. However, the backbone network adopted to extract the image features is pre-trained on ImageNet [37] for classification tasks, which would be ineffective for FSS. In contrast, we treat the cosine similarity map as a prior and adopt the pyramid pooling module (PPM) [25] as the feature extractor, which is trained in a standard supervised manner. The architecture of the proposed PG-CAM is visualized in Figure 4.

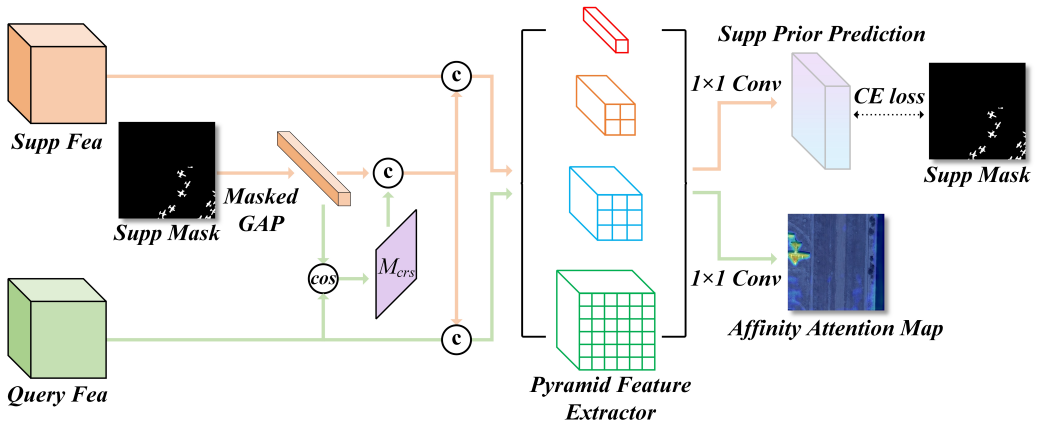

Figure 4. Architecture of the proposed PG-CAM in 1-shot setting.

In detail, the cosine similarity between query feature F_q^3 and support prototype V_s is calculated to generate the prior similarity mask M_{crs}, which serves as an important clue to locating the target regions:

$$M_{crs}(x,y) = \arg\max_k \frac{\exp(\gamma\phi(F_q^3(x,y), V_s^k))}{\sum_{V_s^k \in V_s^{all}} \exp(\gamma\phi(F_q^3(x,y), V_s^k))}, \tag{7}$$

where $x \in \{1,...,h\}, y \in \{1,...,w\}, k \in \{1,...,N\}$, and we set γ to 10 in all experiments.

For the support branch, we first concatenate the support prototype, the support feature F_s^{23} and the prior similarity mask M_{crs} and pass them through PPM. Subsequently, a 1×1 convolution is used to generate support prediction P_s with two output channels:

$$P_s = Conv_{1\times1}\Big(\mathcal{D}_e\Big(F_s^{23} \copyright V_s \copyright M_{crs}\Big)\Big), \tag{8}$$

Thereafter, the ground-truth support mask is applied to supervise the training of the proposed cross-attention module:

$$L_{ce,s} = -\sum_{x=1}^{h}\sum_{y=1}^{w}(M_s(x,y) \cdot log(P_s(x,y))), \tag{9}$$

where $L_{ce,s}$ represents the cross-entropy loss for the support prediction. $M_s(x,y)$ and $P_s(x,y)$ denote the (x,y) location of support ground truth and support prediction, respectively.

The same operation as in the support branch is applied for the affinity attention map prediction, except that the output of the 1×1 convolution is a binary mask:

$$M_{attn} = Conv_{1\times1}\Big(\mathcal{D}_e\Big(F_q^{23} \copyright V_s \copyright M_{crs}\Big)\Big), \tag{10}$$

3.5. Classifier

The obtained support attention feature map A_s and the query affinity attention map M_{attn} are concatenated along with the support prototype V_s and the query feature F_q^{23}. A dilated version of the ASPP module is introduced to merge and enrich these concatenated features. Finally, we obtain the mask prediction $P \in \mathbb{R}^{2\times h\times w}$ through

$$F_q^{23} = Conv_{1\times 1}\{F_q^2 \copyright F_q^3\}, \tag{11}$$

$$F_{merged} = \mathcal{F}_{guidance}\left(M_{attn}, A_s, V_s, F_q^{23}\right), \tag{12}$$

$$P = Softmax\left(\mathcal{D}_m\left(F_{merged}\right)\right), \tag{13}$$

where $\mathcal{F}_{guidance}$ denotes the combination of concatenate and expand operations. D_m consists of the ASPP module, convolutional operation and classifier.

Finally, binary cross-entropy (BCE) loss between $M_q(j)$ and $P(j)$ is employed to supervise the training of the meta learner:

$$L_m = \frac{1}{n_{ep}} \sum_{j=1}^{n_{ep}} BCE(M_q(j), P(j)), \tag{14}$$

where n_{ep} represents the number of training episodes in each batch.

3.6. K-Shot Setting

In K-shot ($K > 1$) segmentation, there are K support sets available. For the self-attention mechanism, we directly take the average of K and generate support attention maps. For the query affinity attention map prediction, K support features are fed into the cross-attention module separately, with each prediction supervised by its own label. Then, we average the K losses as follows:

$$L_{ce,s} = \sum_{i=1}^{K} L_{ce,s}^i, \tag{15}$$

where $L_{ce,s}^i$ denotes the cross-entropy loss of the i-th support image.

Finally, the K-times generated support attention feature map A_s and support prototypes V_s are averaged. Then, the averaged A_s and V_s concatenated with F_q^{23} and M_{attn} are passed through the ASPP module to obtain the predictions.

4. Experiments

4.1. Experimental Setup

Dataset. We assess the effectiveness of our approach on the standard remote sensing benchmark dataset iSAID-5^i [38], which is generated from 2806 high-resolution images. This publicly available aerial image dataset includes 655,451 object instances from 15 geospatial categories. We employ a cross-validation strategy for our experiments, dividing the dataset into three evenly distributed folds, where one fold is used for *meta testing* and the remaining folds are adopted for *meta training*. We randomly select 1000 support–query image pairs for validation in each training episode. As shown in Table 1, we select the unseen classes in each fold following the experimental settings of [13,35], in which the determination of the categories is based on the original sequence of the label dictionary [38].

Table 1. Selection of novel classes for each fold of iSAID-5^i dataset.

# Fold	Novel Classes				
0	Ship (C1)	Storage tank (C2)	Baseball diamond (C3)	Tennis court (C4)	Basketball court (C5)
1	Ground track field (C6)	Bridge (C7)	Large vehicle (C8)	Small vehicle (C9)	Helicopter (C10)
2	Swimming pool (C11)	Roundabout (C12)	Soccer ball field (C13)	Plane (C14)	Harbor (C15)

Evaluation Metrics. Consistent with previous studies [11,22,39], we employ the mean intersection over union (MIoU) for performance assessment. In addition, foreground–background IoU (FB-IoU) is also adopted as the evaluation metric.

Implementation Details. In order to enhance the network's generalization ability, most of the existing FSS approaches use a backbone network pre-trained on the large natural image dataset ImageNet [37], the parameters of which are frozen in the *meta training* phase. This backbone network cannot perfectly adapt to remote sensing image segmentation due to the unignorable domain shift. Hence, we train a more suitable backbone network on iSAID-5^i from scratch within the standard supervised learning paradigm. The backbone network is initialized with the parameters pre-trained on ImageNet [37]. We set the learning rate, training epoch and batch size to 1.25×10^{-3} , 50 and 16, respectively.

For the *meta training*, we adopt the episodic training strategy [11,36]. Specifically, we train the CSCANet using SGD optimizer for 12 epochs, with learning rate and batch size configured as 5×10^{-2} and 8, respectively. We adopt a similar data augmentation strategy to [35]. All experiments are conducted in PyTorch [40] on 4 NVIDIA Tesla T4s.

For a fair comparison, we run the source codes of the selected prevalent FSS approaches, except that we adopt the same retrained backbone network for training. Additionally, we use the same hyper-parameters for training as in our CSCANet.

4.2. Visualization Analysis

Visualization of segmentation results. We visualize some representative predicted masks generated by our CSCANet in Figure 5. The first two rows depict examples of support images (blue) and query images (green). The last two rows show the samples of baseline predictions and the results of CSCANet, respectively. It can be seen in all the examples that the proposed CSCANet is able to effectively reduce false activation. The last five columns show that the proposed method is capable of segmenting the multiple tiny query targets more precisely and completely than the baseline. The predicted masks are almost identical to the corresponding labels.

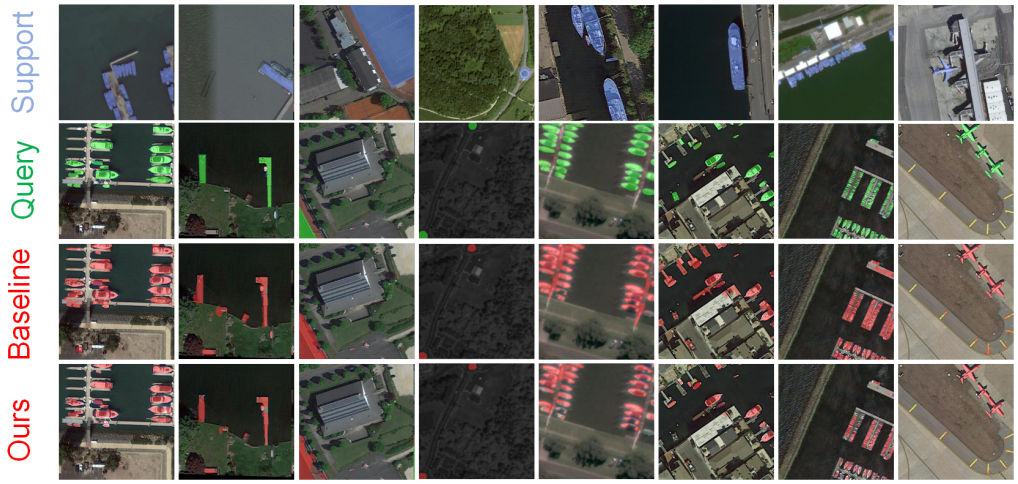

Figure 5. Qualitative examples of 1-shot prediction on the iSAID-5^i.

Visualization of query affinity attention map. To investigate the quality of query attention maps generated by PG-CAM, we plot some representative attention maps in Figure 6. Given the supported image(s) (the 1st row) and query image (the 2nd row), the cross-attention module is able to effectively capture the query targets regardless of their sizes and quantities.

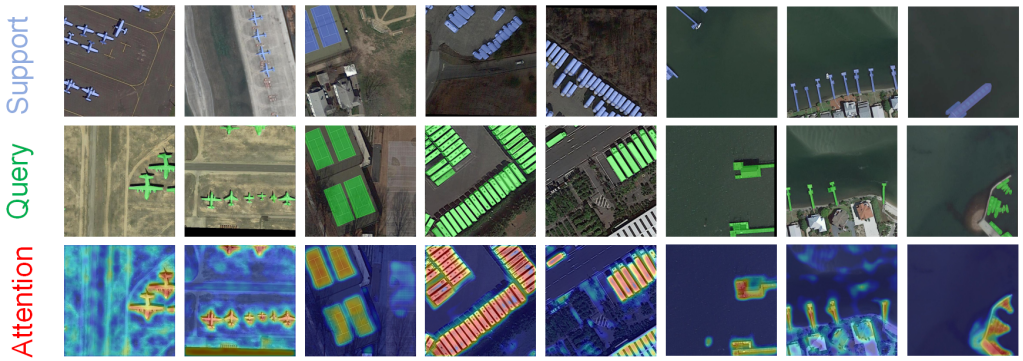

Figure 6. Visualization of the cross-attention maps generated by PG-CAM on the iSAID-5^i in the 1-shot setting.

4.3. Comparison with State of the Art

We compare the performance of CSCANet against other state-of-the-art FSS approaches. Table 2 demonstrates the performance of different approaches on iSAID-5^i in terms of MIoU and FB-IoU. The results indicate that our CSCANet outperforms all SOTA methods across almost all combinations of backbone network (VGG-16 and ResNet-50) and few-shot settings (1-shot and 5-shot), except in the case of backbone VGG-16 under the 1-shot setting. For backbone ResNet-50, we achieve 1.61%mIoU (1-shot) and 2.04%mIoU (5-shot) performance improvements over the best competitor R^2Net. Remarkably, CSCANet significantly surpasses the second-best approach under a 5-shot setting by 2.12%mIoU on average for both backbones. Additionally, we also list the model complexity and inference speed in Table 3. It can be observed that our proposed method reaches a superior balance between performance and efficiency.

Table 2. Comparison of the CSCANet with other FSS networks on iSAID-5^i under 1-shot and 5-shot settings. The results that are underlined denote the second-best performance, while the results that are **bold** show the best performance (the same applies to all the following tables).

Backbone	Method	1-Shot					5-Shot				
		Fold-0	Fold-1	Fold-2	MIoU%	FB-IoU%	Fold-0	Fold-1	Fold-2	MIoU%	FB-IoU%
VGG-16	PANet(ICCV-19) [18]	26.86	14.56	20.69	20.70	52.69	30.89	16.63	24.05	23.86	54.75
	CANet (CVPR-19) [19]	13.91	12.94	13.67	13.51	53.98	17.32	15.07	18.23	16.87	56.86
	SCL (CVPR-21) [41]	25.75	18.57	22.24	22.19	58.96	35.77	24.92	32.70	31.13	61.56
	PFENet (TPAMI-22) [20]	28.52	17.05	18.94	21.50	57.79	37.59	23.22	30.45	30.42	60.84
	NERTNet (CVPR-22) [42]	25.78	20.01	19.88	21.89	56.34	38.43	**24.21**	28.99	30.54	61.97
	DCP (arXiv-22) [43]	28.17	16.52	22.49	22.39	59.55	39.65	22.68	29.93	30.75	60.78
	BAM (CVPR-22) [11]	33.93	16.88	21.47	24.09	59.20	38.46	22.76	28.81	30.01	62.26
	DMML (TGRS-21) [14]	24.41	18.58	19.46	20.82	54.21	28.97	21.02	22.78	24.26	54.89
	SDM (TGRS-22) [13]	24.52	16.31	21.01	20.61	56.39	26.73	19.97	26.10	24.27	56.65
	DML (GRSL-22) [44]	30.99	14.60	19.05	21.55	55.98	34.03	16.38	26.32	25.48	56.26
	TBPN (IJON-23) [45]	27.86	12.32	18.16	19.45	54.26	32.79	16.28	24.27	24.45	55.79
	R^2Net (TGRS-23) [35]	**35.27**	19.93	24.63	**26.61**	**61.71**	**42.06**	23.52	30.06	31.88	63.55
	CSCANet (Ours)	33.26	**20.44**	**25.98**	26.56	61.45	40.08	24.15	**38.00**	**34.08**	**63.74**
ResNet-50	PANet(ICCV-19) [18]	27.56	17.23	24.60	23.13	56.56	36.54	16.05	26.22	26.27	57.37
	CANet (CVPR-19) [19]	25.51	13.50	24.45	21.15	56.64	29.32	21.85	26.91	26.03	59.46
	SCL (CVPR-21) [41]	34.78	22.77	31.20	29.58	61.30	41.29	25.73	37.70	34.91	64.13
	PFENet (TPAMI-22) [20]	35.84	23.35	27.20	28.80	60.09	42.42	25.34	33.00	33.59	63.25
	NERTNet (CVPR-22) [42]	34.93	23.95	28.56	29.15	59.97	44.83	26.73	37.19	36.25	64.45
	DCP (arXiv-22) [43]	37.83	22.86	28.92	29.87	62.36	41.52	28.18	33.43	34.38	63.37
	BAM (CVPR-22) [11]	39.43	21.69	28.64	29.92	62.04	43.29	27.92	38.62	36.61	65.00
	DMML (TGRS-21) [14]	28.45	21.02	23.46	24.31	57.78	30.61	23.85	24.08	26.18	58.26
	SDM (TGRS-22) [13]	27.96	21.99	27.82	25.92	59.58	28.50	25.23	31.07	28.27	59.90
	DML (GRSL-22) [44]	32.96	18.98	26.27	26.07	58.93	33.58	22.05	29.77	28.47	59.23
	TBPN (IJON-23) [45]	29.33	16.84	25.47	23.88	57.34	30.98	20.42	28.07	26.49	58.63
	R^2Net (TGRS-23) [35]	41.22	21.64	35.28	32.71	**63.82**	46.45	25.80	39.84	37.36	66.18
	CSCANet (Ours)	**42.30**	**24.17**	**36.50**	**34.32**	63.56	**47.85**	**30.04**	**40.32**	**39.40**	**66.32**

Table 3. Model complexity and average speed (FPS) comparisons between our approach (ResNet-50, 1-shot) and previous state-of-the-art methods.

	Ours	PANet [18]	CANet [19]	SCL [41]	PFENet [20]	DCP [43]
#Params.	5.2M	23.6M	22.3M	11.9M	10.8M	11.3M
FPS	40.36	58.1	32.7	39.2	45.7	37.9
	BAM [11]	DMML [14]	SDM [13]	DML [44]	TBPN [45]	R^2Net [35]
#Params	**4.9M**	23.6M	29.3M	23.6M	23.6M	5.0M
FPS	44.4	47.4	52.9	**59.5**	56.5	41.5

In addition, we also list the class-wise results in Table 4. It is noteworthy that our proposed CSCANet surpasses other prevalent FSS methods with the backbone ResNet-50 in class C12 (Roundabout) and C14 (Plane) by 13.32%mIoU and 4.73%, separately. The proposed method also obtained the second-best performances in class C1 (Ship), C2 (Storage tank), C3 (Baseball diamond) and C4 (Tennis court). The sizes of these categories are usually tiny and densely arranged in an image, indicating our proposed method is capable of accurately segmenting multiple tiny target objects.

Table 4. Class-wise comparison of CSCANet with other FSS networks on iSAID-5^i under 1-shot setting.

Method	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**	**MIoU%**
VGG-16																
PANet(ICCV-19) [18]	20.05	37.71	21.18	41.22	14.15	12.17	13.82	21.05	7.89	17.88	4.36	31.68	27.55	26.88	12.97	20.70
CANet (CVPR-19) [19]	24.13	6.73	13.83	16.32	8.54	14.12	3.24	21.04	3.35	22.96	9.57	14.91	17.83	16.11	9.92	13.51
SCL (CVPR-21) [41]	28.50	32.93	19.68	29.60	18.05	22.48	7.92	31.46	8.99	22.02	14.17	16.53	19.72	**39.40**	21.37	22.19
PFENet (TPAMI-22) [20]	34.32	31.81	24.20	35.43	16.86	13.98	6.01	31.68	6.76	26.85	8.15	17.75	20.56	33.34	14.87	21.50
NERTNet (CVPR-22) [42]	12.66	23.11	26.90	50.47	15.77	23.14	8.48	31.73	11.75	24.94	14.63	20.45	29.03	28.06	7.24	21.89
DCP (arXiv-22) [43]	27.69	38.45	25.92	33.20	15.57	17.62	12.36	26.79	8.05	17.80	**22.45**	18.29	18.03	37.57	16.10	22.39
BAM (CVPR-22) [11]	27.66	43.90	**31.48**	43.96	**22.66**	13.57	8.91	**31.76**	9.26	20.91	17.05	26.27	30.68	25.27	8.07	24.09
DMML (TGRS-21) [14]	34.75	37.36	15.15	22.85	11.94	21.41	13.85	23.92	10.24	23.50	8.17	16.32	21.08	29.63	**22.09**	20.82
SDM (TGRS-22) [13]	33.76	23.88	17.80	27.76	19.38	18.36	9.63	25.24	8.63	19.69	10.56	15.36	24.76	32.30	22.06	20.61
DML (GRSL-22) [44]	27.30	42.63	19.25	**50.63**	15.13	14.16	**15.94**	22.40	7.74	12.74	3.79	23.73	23.47	27.40	16.88	21.55
TBPN (IJON-23) [45]	22.03	39.75	20.80	42.80	13.94	10.41	6.87	16.54	4.38	23.41	5.68	23.66	22.13	24.63	14.72	19.45
R^2Net (TGRS-23) [35]	**37.82**	**45.16**	26.27	45.30	21.81	**24.11**	14.38	30.92	**12.21**	18.03	18.66	25.02	29.64	31.95	17.87	**26.61**
CSCANet (Ours)	36.21	43.88	26.01	43.39	16.81	21.80	15.84	26.65	10.58	**27.33**	9.05	**41.67**	**32.19**	31.01	15.97	**26.56**
ResNet-50																
PANet (ICCV-19) [18]	21.81	36.31	23.01	42.06	14.59	12.11	17.44	22.70	12.27	21.60	30.29	24.62	26.79	25.54	15.79	23.13
CANet (CVPR-19) [19]	39.57	18.54	18.46	33.63	17.34	9.78	5.49	22.15	5.17	24.89	9.96	36.50	19.12	38.82	17.85	21.15
SCL (CVPR-21) [41]	37.61	33.63	26.68	54.75	21.22	22.60	24.40	30.22	6.71	29.93	**33.00**	44.68	18.25	44.63	15.46	29.58
PFENet (TPAMI-22) [20]	39.02	45.63	20.86	49.96	23.72	21.00	24.76	31.59	6.98	32.42	13.34	47.64	30.65	32.82	11.54	28.80
NERTNet (CVPR-22) [42]	33.59	42.83	22.30	49.35	21.91	21.62	**28.82**	25.64	9.35	34.30	23.91	38.67	25.63	40.84	13.74	28.83
DCP (arXiv-22) [43]	37.42	42.44	35.16	56.55	17.58	21.66	19.57	32.97	10.60	29.50	24.02	35.34	28.44	39.80	17.02	29.87
BAM (CVPR-22) [11]	36.34	39.76	**38.23**	**58.13**	24.71	18.25	12.68	**35.91**	11.42	30.21	28.98	40.74	29.43	33.25	10.79	29.92
DMML (TGRS-21) [14]	40.14	40.18	21.31	27.02	13.60	15.56	15.19	26.05	13.84	**34.44**	11.26	17.57	23.27	39.11	**26.12**	24.31
SDM (TGRS-22) [13]	41.77	35.50	21.41	20.81	20.29	15.60	25.60	28.66	13.29	26.79	13.61	32.35	24.59	42.79	25.75	25.92
DML (GRSL-22) [44]	35.13	42.10	30.49	41.79	15.31	13.25	16.87	24.70	**14.62**	25.45	10.24	35.49	25.35	41.69	18.57	26.07
TBPN (IJON-23) [45]	25.36	41.28	30.67	32.88	16.48	13.48	9.74	27.88	12.52	20.56	11.12	34.31	23.57	40.36	17.98	23.88
R^2Net (TGRS-23) [35]	**46.87**	**49.06**	30.70	52.86	**26.62**	**24.31**	17.25	31.25	13.67	21.73	24.88	46.07	**42.29**	42.07	21.08	32.71
CSCANet (Ours)	45.96	47.83	36.62	57.99	23.10	21.27	23.45	29.87	11.98	34.28	18.69	**59.39**	37.45	**46.80**	20.17	**34.32**

4.4. Limitation Analysis

We observe that the proposed method has a poor performance in C9 (Small vehicle) with both backbone networks. We assume that this is due to the class similarity between C9 (Small vehicle) and other classes like C1 (Ship), C7 (Bridge), and C8 (Large vehicle) in the top-view conditions.

We also visualize some representative failure cases of our proposed method in Figure 7. Failure cases happen mainly due to different resolutions (row 1) and intra-class discrepancy (row 2 and row 3). These are also the major challenges faced by the current Few-Shot

Semantic Segmentation methods for remote sensing images. In the case of limited representativeness, our attention mechanism may concentrate on unrepresentative target information, leading to performance degradation.

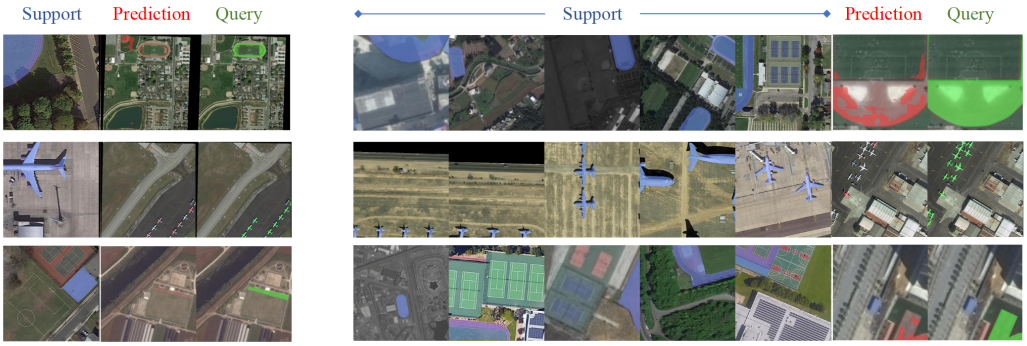

Figure 7. Visualization of the failure cases of the proposed CSCANet on iSAID-5^i (ResNet50, 1-shot setting).

4.5. Ablation Studies

The ablation study aims to examine the importance of each component of our CSCANet. We conducted a variety of ablation experiments on iSAID-5^i under a 1-shot setting, with ResNet-50 selected as the backbone network. The results are presented in Table 5.

Table 5. Ablation study of our CSCANet at module level. The first row represents the result of the baseline.

Self Attention	Cross Attention	Alpha	Prior	MIoU%	FB-IoU%
-	-	-	-	32.85	61.75
✓	-	-	-	33.01	61.81
✓	-	✓	-	33.18	62.13
-	✓	-	-	33.61	62.50
-	✓	-	✓	<u>34.08</u>	<u>62.92</u>
✓	✓	✓	✓	**34.32**	**63.56**

4.5.1. Effect of Self-Attention Module

Compared with the performance of the complete pipeline of CSCANet, the model without a self-attention module reduces it to 0.24% in terms of mIoU. Furthermore, the first two rows of Table 5 show that introducing the learnable parameter α in the SAM brings a further improvement of 0.17% mIoU, implying that α is important for abstracting a robust feature representation of novel classes. These results demonstrate our SAM can effectively extract robust class-relevant information and direct the model to concentrate on the novel class targets.

4.5.2. Effect of Cross-Attention Module

A high-quality query affinity attention map has a significant impact on the final prediction. Therefore, we conducted relevant ablation tests on PG-CAM, which is the core component of CSCANet. As shown in the second and fifth rows of Table 5, the model without PG-CAM decreases the performance to 1.14%. In particular, we also investigated the impact of the prior map on the proposed PG-CAM. Referring to the third and fourth rows, incorporating the prior similarity map achieved a 0.47% mIoU improvement, indicating that the prior information plays a crucial role in guiding the cross-attention module to focus on the unseen-class objects.

5. Conclusions

In this paper, we introduced a few-shot remote sensing image segmentation framework named CSCANet to address the problems of foreground–background similarity and

multiple tiny objects. The proposed CSCANet includes a simple yet effective self-attention module and a prior-guided cross-attention module. Specifically, the first module is able to extract robust unseen-class information from the support set and avoid undesired activation. The second module generates a high-quality query attention map, which can guide the network to concentrate on the tiny target regions. The proposed method demonstrates an outstanding ability to adapt to unseen classes, achieving state-of-the-art (SOTA) performance in both one-shot and five-shot settings.

The major factors in failure cases are different resolutions between support and query sets and the intra-class discrepancy. To address these issues, we will adopt stronger backbones (e.g., ResNet101, Swin-B) and incorporate transformer architecture to enhance the model's feature extraction ability in the future. Furthermore, we will validate the proposed method on more remote sensing benchmark datasets and try to create a new few-shot remote sensing image dataset. We will also explore the potential of extending the proposed framework to the zero-shot remote sensing image segmentation task.

Author Contributions: Conceptualization, G.L., F.X. and Y.-R.C.; Methodology, G.L., F.X. and Y.-R.C.; Experiments, G.L. and F.X.; Validation, G.L. and F.X.; Formal analysis, G.L., F.X. and Y.-R.C.; Investigation, G.L.; Data curation, F.X.; Writing—original draft, G.L., F.X. and Y.-R.C.; Writing—review and editing, Y.-R.C.; Visualization, G.L.; Project administration, Y.-R.C.; Funding acquisition, Y.-R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science and Technology Council, Taiwan (NSTC) under Grant 112-2221-E-197-022.

Data Availability Statement: The original data presented in the study are openly available in iSAID at https://captain-whu.github.io/iSAID/ (accessed on 23 May 2024).

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

Abbreviations

The following abbreviations are used in this manuscript:

FSS	Few-Shot Semantic Segmentation
FSL	Few-Shot Learning
CNN	Convolutional Neural Network
FCN	Fully Convolutional Network
ASPP	Atrous Spatial Pyramid Pooling
PPM	Pyramid Pooling Module
MAP	Masked Average Pooling
SAM	Self Attention Module
PG-CAM	Prior-Guided Supervised Cross-Attention Module
BCE	Binary Cross Entropy
MIoU	Mean Intersection Over Union
FB-IoU	Foreground–Background Intersection Over-Union

References

1. Sun, W.; Du, Q. Graph-regularized fast and robust principal component analysis for hyperspectral band selection. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3185–3195. [CrossRef]
2. Peng, J.; Sun, W.; Ma, L.; Du, Q. Discriminative transfer joint matching for domain adaptation in hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 972–976. [CrossRef]
3. Sun, X.; Yin, D.; Qin, F.; Yu, H.; Lu, W.; Yao, F.; He, Q.; Huang, X.; Yan, Z.; Wang, P.; et al. Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery. *Nat. Commun.* **2023**, *14*, 1444. [CrossRef] [PubMed]
4. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

6. Lin, D.; Dai, J.; Jia, J.; He, K.; Sun, J. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3159–3167.
7. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
8. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7262–7272.
9. Shaban, A.; Bansal, S.; Liu, Z.; Essa, I.; Boots, B. One-shot learning for semantic segmentation. *arXiv* **2017**, arXiv:1709.03410.
10. Zhang, X.; Wei, Y.; Yang, Y.; Huang, T.S. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Trans. Cybern.* **2020**, *50*, 3855–3865. [CrossRef] [PubMed]
11. Lang, C.; Cheng, G.; Tu, B.; Han, J. Learning what not to segment: A new perspective on few-shot segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8057–8067.
12. Ouyang, C.; Biffi, C.; Chen, C.; Kart, T.; Qiu, H.; Rueckert, D. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXIX 16; Springer: Cham, Switzerland, 2020; pp. 762–780.
13. Yao, X.; Cao, Q.; Feng, X.; Cheng, G.; Han, J. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5611711. [CrossRef]
14. Wang, B.; Wang, Z.; Sun, X.; Wang, H.; Fu, K. Dmml-net: Deep metametric learning for few-shot geographic object segmentation in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5611118. [CrossRef]
15. Zhang, C.; Lin, G.; Liu, F.; Guo, J.; Wu, Q.; Yao, R. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9587–9595.
16. Wang, H.; Zhang, X.; Hu, Y.; Yang, Y.; Cao, X.; Zhen, X. Few-shot semantic segmentation with democratic attention networks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIII 16; Springer: Cham, Switzerland, 2020; pp. 730–746.
17. Zhao, Q.; Liu, B.; Lyu, S.; Chen, H. A self-distillation embedded supervised affinity attention model for few-shot segmentation. *IEEE Trans. Cogn. Dev. Syst.* **2023**, *16*, 177–189. [CrossRef]
18. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9197–9206.
19. Zhang, C.; Lin, G.; Liu, F.; Yao, R.; Shen, C. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5217–5226.
20. Tian, Z.; Zhao, H.; Shu, M.; Yang, Z.; Li, R.; Jia, J. Prior guided feature enrichment network for few-shot segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1050–1065. [CrossRef] [PubMed]
21. Li, G.; Jampani, V.; Sevilla-Lara, L.; Sun, D.; Kim, J.; Kim, J. Adaptive prototype learning and allocation for few-shot segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8334–8343.
22. Liu, Y.; Zhang, X.; Zhang, S.; He, X. Part-aware prototype network for few-shot semantic segmentation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IX 16; Springer: Cham, Switzerland, 2020; pp. 142–158.
23. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
24. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18*; Springer: Cham, Switzerland, 2015; pp. 234–241.
25. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
26. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–November 2019; pp. 603–612.
27. Jindal, S.; Manduchi, R. Contrastive representation learning for gaze estimation. In Proceedings of the Annual Conference on Neural Information Processing Systems, PMLR, New Orleans, LA, USA, 10–16 December 2023; pp. 37–49.
28. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015; volume 2.
29. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.

30. Li, H.; Eigen, D.; Dodge, S.; Zeiler, M.; Wang, X. Finding task-relevant features for few-shot learning by category traversal. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1–10.
31. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
32. Jamal, M.A.; Qi, G.-J. Task agnostic meta-learning for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11719–11727.
33. Ravi, S.; Larochelle, H. Optimization as a model for few-shot learning. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
34. Chen, Z.; Fu, Y.; Chen, K.; Jiang, Y.-G. Image block augmentation for one-shot learning. *AAAI Conf. Artif. Intell.* **2019**, *33*, 3379–3386. [CrossRef]
35. Lang, C.; Cheng, G.; Tu, B.; Han, J. Global rectification and decoupled registration for few-shot segmentation in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5617211. [CrossRef]
36. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1–9.
37. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
38. Zamir, S.W.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Khan, F.S.; Zhu, F.; Shao, L.; Xia, G.-S.; Bai, X. Isaid: A large-scale dataset for instance segmentation in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 28–37.
39. Yang, B.; Liu, C.; Li, B.; Jiao, J.; Ye, Q. Prototype mixture models for few-shot semantic segmentation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VIII 16; Springer: Cham, Switzerland, 2020; pp. 763–778.
40. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.P.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst* **1912**, *32*, 8026.
41. Zhang, B.; Xiao, J.; Qin, T. Self-guided and cross-guided learning for few-shot segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8312–8321.
42. Liu, Y.; Liu, N.; Cao, Q.; Yao, X.; Han, J.; Shao, L. Learning non-target knowledge for few-shot semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11573–11582.
43. Lang, C.; Tu, B.; Cheng, G.; Han, J. Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation. *arXiv* **2022**, arXiv:2204.09903.
44. Jiang, X.; Zhou, N.; Li, X. Few-shot segmentation of remote sensing images using deep metric learning. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6507405. [CrossRef]
45. Puthumanaillam, G.; Verma, U. Texture based prototypical network for few-shot semantic segmentation of forest cover: Generalizing for different geographical regions. *Neurocomputing* **2023**, *538*, 126201. [CrossRef]

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