Few-Shot Segmentation of Remote Sensing Images Using Deep Metric Learning

Xufeng Jiang, Nan Zhou, and Xiang Li

***Abstract*— Current convolutional neural network (CNN)-based methods for remote sensing image segmentation require a large number of densely annotated images for model training and have limited generalization abilities for unseen object categories. In this letter, we propose a novel few-shot learning-based method for the semantic segmentation of remote sensing images. Our method can perform semantic labeling for unseen object categories with only a few annotated samples. More specifically, our model starts by using a deep CNN to extract high-level semantic features. The prototype representation of each class is then generated by using a masked average pooling on the feature embeddings of the support images with ground truth masks. Finally, our model performs semantic labeling over the query images by matching the feature embedding of each pixel to its nearest prototypes in the embedding space. Our model is optimized with a nonparametric metric learning-based loss function to maximize the intra-class similarity of learned prototypes while minimizing the inter-class similarity. Experiments on International Society for Photogrammetry and Remote Sensing (ISPRS) 2-D semantic labeling dataset demonstrate satisfying in-domain and cross-domain transferring abilities of our model.**

***Index Terms*— Few-shot learning, prototype representation, remote sensing, semantic segmentation.**

I. Introduction

THE semantic segmentation problem, which is often called image classification in the remote sensing field, is generally defined as determining the semantic class of each pixel in the input images. Recent years, deep learning, especially deep convolutional neural networks (CNNs), has achieved significant breakthroughs in many applications such as scene classification [1], [2], image classification [3], [4], and object detection [5], [6]. However, training a deep neural network model usually requires large-scale annotated data. While, data labeling is a time-consuming and labor-intensive task, especially for semantic segmentation which needs pixel-wise annotations. Semi- or weakly supervised learning methods, such as [7] and [8] are proposed to alleviate such requirements, but they still need a large amount of annotated data for training. Moreover, after training, these methods cannot generalize well to unseen classes.

Manuscript received 10 September 2021; revised 21 October 2021, 16 January 2022, and 12 February 2022; accepted 22 February 2022. Date of publication 24 February 2022; date of current version 21 March 2022. *(Corresponding author: Xiang Li.)*

Xufeng Jiang is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 10094, China, and also with the University of Chinese Academy of Sciences, Beijing 101408, China.

Nan Zhou is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 10094, China.

Xiang Li is with the Department of Electrical and Computer Engineering, New York University Abu Dhabi, Abu Dhabi, UAE (e-mail: xiangli92@ieee.org).

Digital Object Identifier 10.1109/LGRS.2022.3154402

In contrast to machine vision algorithms, humans can easily identify new object classes (e.g., in classification and semantic segmentation) after seeing only a few examples. To equip machine vision algorithms with this powerful learning ability, recent researches started the new research topic of few-shot learning that aims to learn a model that can generalize well to new classes with few annotated data. In recent years, few-shot learning has been actively explored in the computer vision field with applications to image classification, semantic segmentation, and object detection [9]–[11].

In remote sensing applications, a common situation is that one has trained a segmentation model for certain classes (e.g., building) but wants to perform segmentation on new classes (e.g., road). In existing methods, the model training and inference stage are operated on the same set of predefined classes, so one would always need a large amount of training data from novel classes to perform classification. In this letter, we propose the first few-shot learning-based method for the semantic segmentation of remote sensing images. More specifically, we aim to perform semantic segmentation of remote sensing images with only use a few annotated samples of new classes.

In the few-shot learning scenario, our model is designed to tackle new classes that have not been seen during model training. Instead of learning to extract feature representation for predefined classes, the model should have the ability to learn transferable knowledge, i.e., it should learn to extract feature patterns for both seen and unseen classes. To achieve this, we separate the training process as prototype extraction and nonparametric metric learning. The prototypes extraction part leverages the powerful feature learning ability of CNNs to extract a feature vector representation for each semantic class. While the nonparametric metric learning part conducts pixel-wise segmentation through nearest neighbor matching in the embedding space. After network training, each learned prototype is a compact and robust representation of the corresponding class and is at the same time sufficiently distinguishable from other classes. In the testing stage, our model performs semantic segmentation for each pixel of the query image by matching the feature embeddings to its closest prototype representation.

We list the main contributions of the proposed method as follows.

1) This letter introduces a novel few-shot learning-based method for the semantic segmentation of remote sensing images. Our proposed model can perform semantic labeling on new classes using only a few annotated samples.

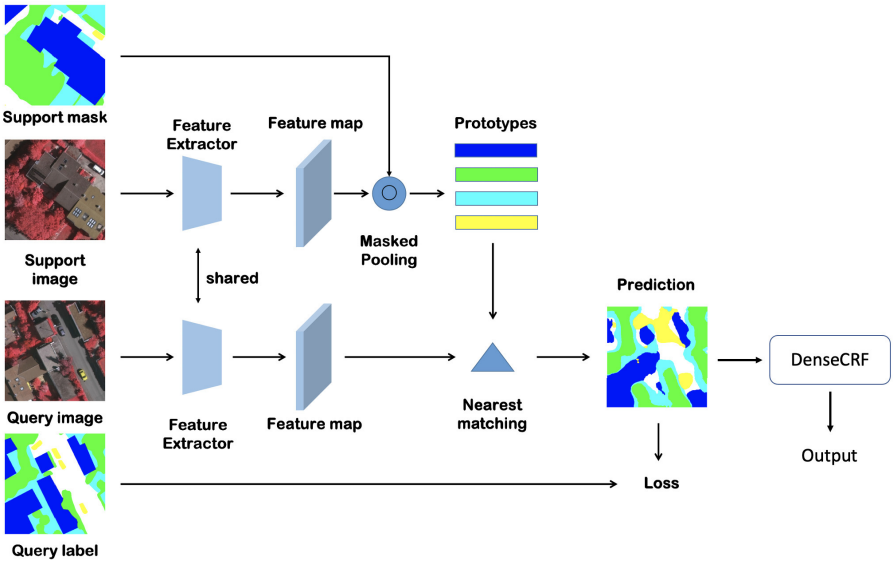

Fig. 1. Overview of the proposed method for few-shot segmentation of remote sensing images.

2) Our model is optimized by a nonparametric metric learning objective and learns to maximize in-class similarity while minimizing the inter-class similarity.
3) We investigate both in-domain and cross-domain transferring abilities of our model. We show that our model can learn transferable representations and perform semantic labeling on new classes in the same domain or from a different domain.

II. Methods

A. Problem Statement

The problem of few-shot segmentation aims at learning a segmentation model from the dataset of available classes that can perform semantic labeling for unseen classes using the information from only a few annotated images from the same unseen classes. To achieve this goal, a model is generally trained on a dataset D_{train} with the class set C_{seen} and learns to extract transferable representations that can be applied to the new dataset D_{test} with new classes C_{unseen}, where C_{unseen} and C_{seen} have no overlap.

To facilitate model training and evaluation, we construct numerous episodes from the training and testing set. Each episode E_i is constructed from a set of support images S_i (with annotations) and a set of query images Q_i. Given a C-way-K-shot segmentation task, each support set S_i consists of K annotated images per semantic class and there are in total C different classes. We denote the support set as, $S_i = \{(I_k, M_{ck})\}_{k=1}^{K}$ where I_k denotes the input image and M_{ck} denotes the binary mask for class c, $I_k \in \mathbb{R}^{H_i \times W_i \times 3}$, $M_{ck} \in \mathbb{R}^{H_i \times W_i}$, $k = 1, 2, \ldots, K$ and $c \in \mathcal{C}_i$ with $|\mathcal{C}_i| = C$. Note that we deal with the mask of each class independently. The query set Q_i contains N_q images from the same set of class $\mathcal{C}_i$ as the support set. The support images are used for prototype learning and our model performs segmentation for the query images by matching each pixel in the query images to these learned prototypes. During training, the predicted segmentation mask of query images is compared with the ground truth ones to guide network training. As each episode contains a different set of semantic classes, our model is, therefore, able to perform class-agnostic feature learning and thus can generalize well to unseen classes. After training on the training set D_{train} until convergence, we evaluate the segmentation performance on all episodes from the test set D_{test}.

B. Method Overview

Different from classical semantic segmentation methods that learn to extract representative feature representations for predefined classes, our model is designed to perform class-agnostic feature learning and thus can generalize well to unseen classes. To achieve this goal, our model is designed to learn to extract robust and compact prototype representation for each semantic class. Then it predicts the semantic label for each pixel in the query images via nonparametric feature matching in the embedding space.

Fig. 1 gives an overview of the proposed method for few-shot segmentation of remote sensing images. Our model starts with a feature extraction network to learn deep feature representations for the support and query images. We use a shared backbone network to extract feature representations for both support and query images, details will be introduced in Section II-C. Then, a masked average pooling layer is used to extract prototypes from the support sets, as described in Section II-D. Finally, our model predicts the segmentation label for each pixel in the query images by matching its feature embedding to its nearest prototype representation. Our model is optimized with a nonparametric metric learning objective and learns to maximize in-class similarity while minimizing the inter-class similarity, as described in Section II-E.

C. Feature Extraction

The original ResNet-50 architecture consists of four residual blocks, with each block corresponding to a different level of representation. We choose the feature maps from block-2 and block-3 to formulate the feature embeddings for prototype learning. To maintain the spatial resolution, convolution layers in block-3 are replaced by dilated convolutions [12] with the dilation factor set to 2. In this way, all feature maps after

block-2 have a fixed spatial resolution of 1/8 of the input image. Moreover, in a dilated convolution, the convolutional layer has a larger receptive field which enables informative inputs from a larger area. We concatenate the output of block-2 and block-3 and use another 3×3 convolution layer to compress the feature dimension to 256. The feature extraction network is shared by both support images and query images for feature learning.

D. Prototype Learning

A prototype representation is generally regarded as a compressed feature vector that conveys the necessary information for distinguishing a specific class from others. In our method, the prototype representation is defined in the middle-level embedding space which ensures robustness toward different object categories and input variations, such as different structures, light change, occlusion, and truncation. To generate the class-wise prototype feature representation, we aggregate the deep features among different pixel locations of the same semantic class. Specifically, we exploit the ground truth annotation mask of the supported image and leverage a masked average pooling over the whole feature maps to generate the prototypes for each foreground class and background class independently. One may note that the deep feature maps and semantic masks are of different spatial resolutions. To tackle this issue, we leverage bilinear interpolation to upsample the feature maps to the same resolution as the semantic masks.

More specifically, given a support set $S_i = \{(I_k, M_{c,k})\}$ and the corresponding feature maps F_k after bilinear interpolation, where c and k denotes the class and sample index. The prototype representation of class c can be calculated through a masked average pooling following:

$$P_c = \frac{1}{K} \sum_k \frac{\sum_{(x,y)} F_k(x, y) \mathbb{I}\big(M_{c,k}(x, y) = 1\big)}{\sum_{(x,y)} \mathbb{I}\big(M_{c,k}(x, y) = 1\big)} \tag{1}$$

where (x, y) denotes the pixel coordinates on feature maps and on masked images and $\mathbb{I}(\cdot)$ represents an indicator function which outputs 1 if the input is true or 0 otherwise.

E. Nonparametric Metric Learning

Given the prototype representations, our model can predict the semantic segmentation maps for the query images by matching the deep feature vector of each pixel to its nearest prototype in the embedding space. To achieve this goal, we calculate the distance between the feature vector of each pixel location in query images to all prototypes in the embedding space, and then search the nearest prototype and assign the corresponding class to this pixel location. To facilitate back-propagation, we also generate the probability maps by applying a softmax function over the distance values for each pixel location. We achieve this goal by the following equation:

$$P_{c,j}(x, y) = \frac{e^{-\alpha \mathcal{D}\left(F_j(x,y), P_c\right)}}{\sum_c e^{-\alpha \mathcal{D}\left(F_j(x,y), P_c\right)}} \tag{2}$$

where $F_{c,j}(x, y)$ denotes the query feature vector for the jth query image at location (x, y) and $\mathcal{D}(\cdot, \cdot)$ denotes the distance function defined in the embedding space, and α is a hyperparameter to control probability distribution with regard to distance values. In our experiments, we set α to 20 since different values lead to similar performance. For the distance function $\mathcal{D}(\cdot, \cdot)$, in this letter, we use cosine distance for $\mathcal{D}(\cdot, \cdot)$, calculated as

$$\mathcal{D}(a, b) = \frac{a \cdot b}{||a|| \cdot ||b||} \tag{3}$$

where a, b denote two feature vectors, and $|| \cdot ||$ calculates vector norm.

Given the predicted probability map $P_{c,j}$ and the corresponding ground truth segmentation label map $\tilde{M}_j$, we formulate the metric-based segmentation loss as

$$\mathcal{L}_{\text{seg}} = -\frac{1}{N_q * N} \sum_j \sum_{(x,y)} \sum_c \mathbb{I}\big(\tilde{M}_j(x, y) = c\big) \log\big(P_{c,j}(x, y)\big) \tag{4}$$

where N_q and N denote the number of query images and number of pixels in each query image, respectively.

III. Experiments and Results

A. Implementation Details

We conduct experiments on the International Society for Photogrammetry and Remote Sensing (ISPRS) 2-D Semantic Labeling Challenge dataset. This dataset contains very high-resolution aerial images collected in two cities of Germany: Vaihingen and Potsdam. And for each aerial image, the ground truth labels are provided on six classes: buildings, impervious surfaces (Imp. surf.), low vegetation (Low veg.), trees, cars, and clutter. We follow [3] to prepare experimental datasets, and we use overall accuracy and F1 score of each class to evaluate the classification performance.

In our method, the classification label of each pixel in query images is assigned independently according to their distance to the prototypes. This will unavoidably cause unwanted label inconsistency in the predicted segmentation maps. To address this issue, we leverage the widely adopted postprocessing method DenseCRF [13] to refine the segmentation maps.

B. Cross-Domain Segmentation

We first demonstrate the segmentation abilities of our proposed model under different domains. To achieve this, we train our model on PASCAL VOC 2012 dataset and evaluate the performance on ISPRS 2-D semantic labeling dataset. We follow the experimental settings used in [10] and train our model on the PASCAL-5i [14] dataset. In the evaluation stage, we divide the testing images into 417×417 patches. For each patch, we set it as the query/test image and randomly select five patches with ground truth labels from the training split as support images to formulate an episode.

1) Results on Vaihingen: We list the performance of all comparing methods in Table I. Note that all comparing methods use the full training set for model training, while our results are directly generated without training on the ISPRS dataset. As shown in Table I, our model gets a satisfying overall accuracy of 67.7%, which is significantly better than PANet [10]. The performance is reasonable considering our

TABLE I

RESULTS ON THE VAIHINGEN DATASET. THE NUMBERS INDICATE THE F1 SCORE ON EACH CATEGORY EXCEPT THE LAST COLUMN, WHICH INDICATES THE OVERALL ACCURACY

Method	Imp. surf.	Buildings	Low veg.	Trees	Cars	Overall
FCN [15]	90.5	93.7	83.4	89.2	72.6	89.1
SegNet (IRRG) [3]	91.5	94.3	82.7	89.3	85.7	89.4
PANet [10]	51.4	71.1	39.6	75.7	23.5	62.1
Ours	62.6	73.1	38.7	80.5	41.1	67.7

TABLE II

RESULTS ON THE POTSDAM DATASET. THE NUMBERS INDICATE THE F1 SCORE ON EACH CATEGORY EXCEPT THE LAST COLUMN, WHICH INDICATES THE OVERALL ACCURACY

Method	Imp. surf.	Buildings	Low veg.	Trees	Cars	Overall
FCN [15]	92.5	96.4	86.7	88.0	94.7	90.3
SegNet (IRRG) [3]	92.4	95.8	86.7	87.4	95.1	90.0
PANet [10]	49.7	66.7	40.9	60.1	20.0	52.5
Ours	50.6	67.7	40.5	60.3	24.9	53.2

results are generated from only five small patches with annotations. Also, our model gets an F1 score of above 70% on impervious surface and low vegetation classes.

2) Results on Potsdam: For the Potsdam dataset, we downsample both the aerial images and digital surface model (DSM) images by a factor of 2 and we report our performance on the downsampled labels. From Table II, our model achieves a reasonable overall accuracy of 52.2% with only five annotated small patches, and the performance is better than PANet [10]. Moreover, our model gets an F1 score of above 50% on impervious surface, building, and tree classes.

C. In-Domain Category Transfer

In this section, we conduct experiments to verify the in-domain transferring abilities of our model. We train the model on selected classes from the ISPRS Vaihingen dataset and evaluate the segmentation performance on other classes with only a few annotated images. We conduct experiments with two settings: 1) we train our model on the classes including impervious surface, low vegetation, tree, and car and test the performance on building class and 2) we train our model on the classes including building, low vegetation, tree, and car and test the performance on impervious surface. We do not use DenseCRF to postprocess the prediction results in this section. Table III lists the performance of these two experiments.

As shown in Table III, with only five labeled patches, our proposed method gets an overall accuracy of 77.4% and 72.0% on building and impervious surface categories of the Vaihingen dataset. Moreover, after training our model on other classes, our model is able to get significantly better performance than the cross-domain setting (73.1% and 62.6% for building and impervious surface) on a specific class with only a few annotated label images. This finding further demonstrates a satisfying property of our model, i.e., the few-shot segmentation performance can be boosted by training on some other classes of the same domain.

TABLE III

IN-DOMAIN VERSUS CROSS-DOMAIN SEGMENTATION ACCURACY ON BUILDING AND IMPERVIOUS CLASSES OF VAIHINGEN DATASET

Category	Cross-domain	In-domain
building	73.1	77.4
imp. surface	62.6	72.0

TABLE IV

EFFECT OF NETWORK DEPTH AND DENSECRF

Method	Vaihingen		Potsdam	
	Acc.	Kappa	Acc.	Kappa
VGG-init	58.8	46.9	50.0	35.7
VGG-Net	60.0	47.6	52.5	39.4
ResNet50	62.1	49.6	48.7	35.2
Ours w DenseCRF	67.7	56.3	53.2	40.2

IV. DISCUSSION

A. Network Depth

We first explore the effect of different backbone networks for prototype learning. We tried two different backbone networks, i.e., VGG-Net and ResNet. As shown in this Table IV, our model with ResNet backbone achieves better performance on the ISPRS Vaihingen dataset, with an improvement of 2.1% on overall accuracy. While for the Potsdam dataset, using a deeper backbone network hurts the performance. The overall accuracy produced by VGG-Net decreases from 52.5% to 48.7% by using the ResNet backbone. In the following experiments, we use ResNet and VGG-Net for feature extraction and prototype learning on Vainhingen and Potsdam dataset, respectively.

B. Effect of DenseCRF

In this section, we investigate the effect of DenseCRF on segmentation refinement. Table IV lists the performance of our model with and without DenseCRF for segmentation refinement. As shown in Table IV, our model achieves significantly better performance on the ISPRS Vaihingen dataset by applying DenseCRF postprocessing. More specifically, our method with DenseCRF improves the overall accuracy by 6.6%. This is because by using DenseCRF, our model can produce more consistent and smooth segmentation maps.

By comparing our prediction with and without DenseCRF in Fig. 2, one can find that our original prediction (w/o DenseCRF) leads to inconsistent predictions, e.g., the pixels of the same building instance have different predicted labels. By enforcing spatial continuity using DenseCRF,

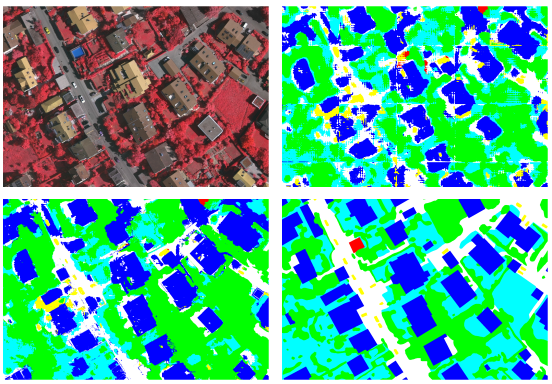

Fig. 2. Segmentation results: aerial image, baseline prediction, our model with DenseCRF, ground truth.

TABLE V
SEGMENTATION PERFORMANCE WITH DIFFERENT NUMBERS OF SUPPORT IMAGES

#Samples	Vaihingen		Potsdam	
	Acc.	Kappa	Acc.	Kappa
SegNet (1 shot)	49.2	30.3	34.0	10.8
SegNet (5 shots)	60.9	50.4	43.8	25.8
SegNet (10 shots)	71.3	61.1	50.6	36.9
Ours (1 shot)	56.9	42.2	45.0	28.5
Ours (5 shots)	67.7	56.3	53.2	40.2
Ours (10 shots)	72.4	62.9	55.4	40.8

our model successfully corrects inconsistent predictions and smoothness the segmentation maps.

C. Number of Samples

We conduct experiments on both ISPRS Vaihingen and Potsdam datasets with 1/5/10 support labels. Note that when selecting the support images, we make sure that the selected images cover all semantic classes. We list the performance under different configurations in Table V. From Table V, we can see that the segmentation performance of our model improves consistently as more annotated images are provided. Especially, our model achieves a satisfying segmentation accuracy of 72.4% on ISPRS Vaihingen with ten labeled patches. The results on the Potsdam dataset also enjoy a performance boost when using more support labels. Table V also shows that with only a few support labels, our model obtains significantly better performance than its supervised counterpart on both the Vaihingen and Potsdam datasets, especially when only one or five support labeled patches are provided. This finding is even more significant on the Potsdam dataset, where the SegNet model gets an overall accuracy 11% less than our model when only one support image is provided.

V. CONCLUSION

In this letter, we introduce a novel few-shot learning-based method for semantic segmentation of remote sensing images. Our method can perform segmentation for unseen object categories with only a few annotated samples. More specifically, our model starts with a deep CNN to extract high-level semantic features. Then, the prototype representation of each class is generated by a masked average pooling over feature maps by leveraging the ground truth masks of support images. Finally, our model performs semantic labeling over the query images by matching the feature embedding of each pixel to its closest prototype. Our model is optimized with a nonparametric metric learning-based loss function in order to maximize the intra-class similarity of learned prototypes while minimizing the inter-class similarity. We conduct both in-domain and cross-domain experiments to demonstrate the generalization abilities of our model on unseen categories. Experiments on ISPRS 2-D semantic labeling dataset demonstrate satisfying in-domain and cross-domain transferring abilities of our model. Specifically, with only five annotated image patches, our model gets satisfying performance on the ISPRS 2-D Semantic Labeling Challenge dataset, with an overall accuracy of 67.7% and 53.2% for Vaihingen and Potsdam dataset, respectively.

REFERENCES

[1] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
[2] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
[3] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.
[4] Y. Feng, X. Sun, W. Diao, J. Li, and X. Gao, "Double similarity distillation for semantic image segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 5363–5376, 2021.
[5] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.
[6] Y. Hu, X. Li, N. Zhou, L. Yang, L. Peng, and S. Xiao, "A sample update-based convolutional neural network framework for object detection in large-area remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 947–951, Jun. 2019.
[7] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1635–1643.
[8] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1742–1750.
[9] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.
[10] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9197–9206.
[11] X. Li, J. Deng, and Y. Fang, "Few-shot object detection on remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[13] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.
[14] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 167.1–167.13.
[15] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*.