Few-Shot Rotation-Invariant Aerial Image Semantic Segmentation

Qinglong Cao, *Graduate Student Member, IEEE*, Yuntian Chen, *Member, IEEE*,
Chao Ma, *Member, IEEE*, and Xiaokang Yang, *Fellow, IEEE*

***Abstract*— Few-shot aerial image semantic segmentation is a challenging task that requires precisely parsing unseen-category objects in query aerial images with limited annotated support aerial images. Formally, category prototypes would be extracted from support samples to segment query images in a pixel-to-pixel matching manner. However, aerial objects in aerial images are often distributed with arbitrary orientations, and varying orientations could cause a dramatic feature change. This unique property of aerial images renders conventional matching manner without consideration of orientations fails to activate same-category objects with different orientations. Furthermore, the oscillation of the confidence scores in existing rotation-insensitive algorithms, engendered by the striking changes of object orientations, often leads to false recognition of lower scored rotated semantic objects. To tackle these challenges, inspired by the intrinsic rotation invariance in aerial images, we propose a novel few-shot rotation-invariant aerial semantic segmentation network (FRINet) to efficiently segment aerial semantic objects with diverse orientations. Specifically, through extracting orientation-varying yet category-consistent support information, FRINet provides rotation-adaptive matching for each query feature in a feature-aggregation manner. Meanwhile, to encourage consistent predictions for aerial objects with arbitrary orientations, segmentation predictions from different orientations are supervised by the same label and further fused to obtain the final rotation-invariant prediction in a complementary manner. Moreover, aiming at providing a better solution searching space, the backbones are newly pretrained in the base category to basically boost the segmentation performance. Extensive experiments on the few-shot aerial image semantic segmentation benchmark demonstrate that the proposed FRINet achieves a new state-of-the-art performance. The code is available at https://github.com/caoql98/FRINet.**

***Index Terms*— Consistent prediction, few-shot aerial semantic segmentation, rotation invariance, rotation-adaptive matching.**

I. Introduction

Semantic segmentation of remote sensing images is an indispensable task in Earth vision with a broad range of applications, such as surveillance of urban areas [1], [2], [3], building/road detection [4], [5], [6], and traffic management [7], [8], [9]. With the rapid development of deep-learning technology, fully supervised segmentation algorithms have shown impressive performance on aerial image semantic segmentation [10], [11], [12], [13], [14]. Yet the success of deep-learning-based methods heavily depends on the availability of large-scale annotated datasets [15], [16], [17] that are time-consuming and laborious to construct, and it is often infeasible to obtain annotated data for all semantic categories. In addition, traditional semantic segmentation models are rigid and lack the adaptability required for practical applications that demand recognition of new, unforeseen objects or environments [18]. This is exactly where the concept of few-shot learning comes into play. Few-shot learning empowers models to learn from a minimal amount of labeled data, making them versatile and adaptable. This adaptability is invaluable in situations where collecting extensive training data is impractical or impossible. Inspired by this, the few-shot aerial image semantic segmentation task is proposed to handle more practical semantic segmentation scenarios, which requires accurately parsing query aerial images with only a few annotated support samples.

Few-shot semantic segmentation (FSS) algorithms [19], [20], [21], [22], [23], [24], [25], [26] typically begin by generating category prototypes from the extracted support features. These prototypes are then utilized to accurately segment query images in a pixel-to-pixel matching manner. Following this pipeline, many advanced FSS algorithms have been proposed to boost segmentation performance via stronger category prototypes [20], [27], better matching strategies [28], [29], and distracting information elimination [30], [31]. Recently, aiming at dealing with the large variance of aerial objects' appearances and scales, Yao et al. [18] proposed the first few-shot aerial image semantic segmentation algorithm to provide a scale-aware detailed matching.

Previous simple matching frameworks directly expanded category prototypes as matching features and further concatenated these matching features with query features to activate category-related semantic objects. However, many semantic objects with the same category in aerial images usually appear with arbitrary orientations, resulting in a dramatic class-agnostic feature change. The existing simple matching

Manuscript received 1 June 2023; revised 16 October 2023 and 10 November 2023; accepted 29 November 2023. Date of publication 1 December 2023; date of current version 13 December 2023. This work was supported in part by the National Science Foundation of China under Grant 62106116 and Grant 62071388 and in part by the China Meteorological Administration Climate Change Special Program (CMA-CCSP) under Grant QBZ202316. *(Corresponding author: Yuntian Chen.)*

Qinglong Cao is with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo 315200, China (e-mail: caoql2022@sjtu.edu.cn).

Yuntian Chen is with the Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo 315200, China, and also with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: ychen@eitech.edu.cn).

Chao Ma and Xiaokang Yang are with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China.

Digital Object Identifier 10.1109/TGRS.2023.3338699

manner clearly could not tackle such a dramatic class-agnostic feature change, and same-category objects with different orientations might not be activated. Similarly, due to the consistent negligence of the critical rotation invariance in aerial images, current rotation-insensitive FSS solutions [Fig. 1(a)] have been prone to unstable aerial image segmentation, where aerial semantic objects' confidence scores can see drastic oscillation with the change of orientation, and the rotated semantic objects with low confidence scores may be regarded as background.

To tackle these problems, in this article, we propose a fresh few-shot rotation-invariant aerial semantic segmentation network (FRINet) to efficiently segment aerial semantic objects with diverse orientations. The design of the proposed FRINet [Fig. 1(b)] is inspired by human knowledge, i.e., the category of objects in aerial images remains consistent after arbitrary rotation. This knowledge could work as an implicit constraint for rotation-invariance learning. Particularly, for the support images, the aerial objects distributed with different orientations could provide orientation-varying but category-consistent support information that enables the network to achieve orientation-adaptive matching. For the query images, the segmentation model should make consistent predictions for the aerial objects distributed with varying orientations, which could encourage the network to make a steady aerial image semantic segmentation.

More specifically, FRINet first extracts support features and query features, respectively, for support and query images with varying orientations. Then, through mask averaging pooling, the corresponding support masks are applied to the support features to generate category-consistent support prototypes with diverse orientations. To provide orientation-adaptive matching for query features, the relation scores between each element of query features and these orientation-varying support prototypes are computed. Utilizing these relation scores as the aggregating scores, each element of query features could obtain an orientation-adaptive prototype. By correspondingly concatenating these prototypes with the query features, aerial semantic objects with different orientations could be adaptively activated. Subsequently, aiming at encouraging the segmentation model to make consistent predictions from different orientations, the predictions from different orientation branches are supervised with the same rotated labels. In this way, the same-category objects with different rotations would be pulled closer in the embedding space, and the foreground could be searched in varying rotations. Finally, by aggregating these predictions from different orientations to generate the final prediction in a complementary manner, the iterative meta-training process could provide abundant complementary visual patterns to develop a powerful few-shot rotation-invariant segmentation network.

Moreover, existing FSS solutions tend to leverage backbones (e.g., VGG16 [32] and Resnet50 [33]) pretrained in the large-scale ImageNet [34] as the feature extractor for both support and query images. Yet these pretrained backbones working for classification could only provide a global perceptive feature space for support and query images, which clearly cannot satisfy the pixel-level judge requirement for the segmentation task. To provide more rational and detailed embedding space, the backbones shall first be pretrained in the base category as the feature extractor for the segmentation task.

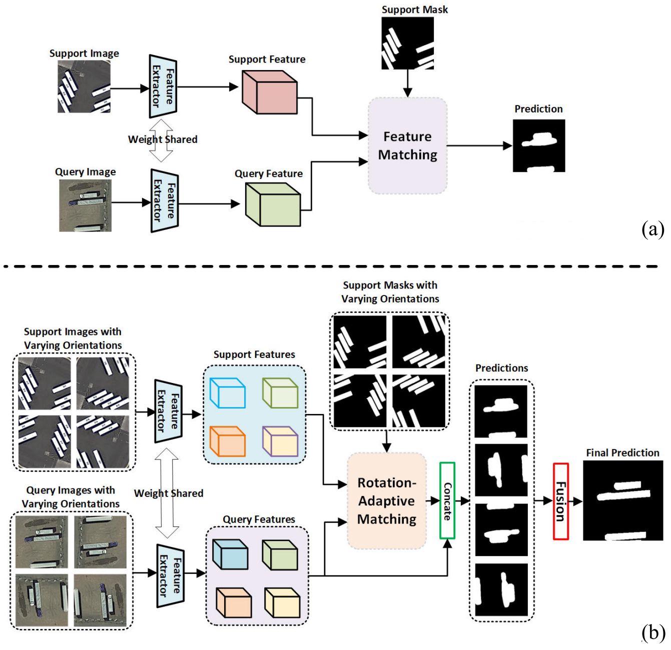


Fig. 1. Comparison between the existing FSS framework for aerial images and our proposed few-shot rotation invariant semantic segmentation framework.

By elaborately leveraging category-invariance in rotation, the proposed FRINet provides a rotation-adaptive matching and further performs a steady segmentation process with the rotation-consistent constraint. FRINet successfully addresses previous challenges via rotation-invariant learning and achieves state-of-the-art segmentation performance. The main contributions of this article are summarized as follows.

1) To the best of our knowledge, we present the first attempt to build a rotation-invariant aerial semantic segmentation network under the few-shot setting, and the proposed network is able to provide a rotation-consistent prediction in a rotation-adaptive matching manner.
2) Aiming at constructing a more rational and detailed feature space for FSS in aerial images, we newly provide backbones pretrained on the base category to boost the segmentation performance to a higher level.
3) Experimental results on the FSS benchmark for aerial images demonstrate that the proposed FRINet outperforms existing state-of-the-art few-shot aerial segmentation methods by a large margin.

II. Related Work

Semantic segmentation has been extensively studied for the last few decades. In this section, we will first review the related works with regard to aerial image semantic segmentation and then introduce advanced FSS algorithms.

A. Aerial Image Semantic Segmentation

Aerial image semantic segmentation aims to precisely classify each pixel of aerial images. Recently, with the development of deep-learning technology, many advanced aerial image semantic segmentation algorithms have been proposed [12], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44]. For instance, Kampffmeyer et al. [35] first introduced the deep convolutional neural networks (CNNs) to parse the aerial images in an uncertainty quantifying manner. By introducing residual connections [33], atrous convolutions [45], pyramid scene parsing pooling [46], and multitasking inference into previous CNNs, Diakogiannis et al. [36] proposed a reliable segmentation framework for very high-resolution aerial images. Mining Rotation and Scale invariance is a critical issue for the semantic segmentation task. For instance, aiming to provide better class-activation-maps (CAMs) for the weakly supervised building segmentation, Wang et al. [47] designed a scale-invariant optimization module that cooperates with mutual learning and coarse-to-fine optimization to improve the completeness of CAMs. Directly from an architecture perspective, Mitton and Murray-Smith [48] designed the rotation equivariant CNN model to efficiently perform the deforestation segmentation. Moreover, through the designed densely connected feature network and the spatial feature recalibration module, the SaNet [49] successfully extracts the scale-aware feature representation to enhance the performance of aerial image semantic segmentation. Similarly, Sci-Net [50] leverages UNet hierarchical representation and dense atrous spatial pyramid pooling (ASPP) to achieve scale-invariant building segmentation from aerial imagery. Furthermore, Zhao et al. [51] integrated a refined rotated detector to extract rotation equivariant and invariant features to boost the rotation-aware building instance segmentation network. Simultaneously, by using rotating convolutions as building blocks, RotEqNet [52] treats rotated versions of the same object with the same filter bank, which efficiently mines the rotation-invariance and achieves advanced segmentation performance. Inspired by these advanced works, we also proposed the first rotation-invariant aerial semantic segmentation network to provide a rotation-consistent prediction in a rotation-adaptive matching manner.

B. Few-Shot Semantic Segmentation

The goal of FSS is to parse the unseen-category objects with only a few annotated support samples. The first attempt for FSS [22] is to leverage the category prototypes learned from the support samples to parse the category-related objects in a metric-learning manner. To provide strong category prototypes with stronger discriminative ability, prototype alignment regularization between support and query is introduced into PANet [20]. Simultaneously, aiming at refining the initial coarse predictions, CANet [21] utilizes multilevel feature comparison to iteratively refine the segmentation results. Previous works only provided a single prototype for each category. To provide more detailed category information, PMMs [28] introduce prototype mixture models to enforce the prototype-based semantic representation. Similarly, Liu et al. [53] adopt the super-pixels concept to generate part-aware prototypes and further design a graph neural network model to perform segmentation of query images. Though previous FSS methods have acquired good segmentation performance, these algorithms ignore the critical generalization problem between novel and base categories. Focusing on tackling this issue, PFEnet [24] leverages high-level semantic information as prior knowledge to tackle the spatial inconsistency between query and support targets. Accurately analyzing the relations between query and support features is the key factor of FSS. Inspired by this, HSNet [54] designs efficient 4-D convolutions to mine the e fine-grained correspondence relations between the query and the support images. Furthermore, efficiently removing distracting objects is also an appropriate method to boost the segmentation performance. BAM [55] and NERTNet [30] adopt this concept, and, respectively, eliminate the distracting objects with a base learner and a background mining loss. However, previous methods all focused on natural images and thus could not handle the unique property of aerial images. Meanwhile, though there are some works for few-shot classification [56], [57], [58] and object detection [59], [60], [61] on remote sensing images, there are still fairly few FSS works for aerial images. Recently, aiming at providing detailed support guidance for the aerial images, SDM [18] was first proposed to tackle few-shot aerial image semantic segmentation in a detailed matching manner. However, previous FSS solutions all ignore that the aerial objects in aerial images are often distributed with arbitrary orientations and varying orientations could cause a dramatic feature change. This property would lead to false recognition of orientation-varying aerial objects. However, the random rotation strategy could provide aerial objects with varying orientations in long-term training. Only a single-orientation view is leveraged in the query-support matching for an iteration. The orientation-varying objects could still not be activated and be falsely recognized. To conquer this challenge, a novel FRINet is first proposed in this article.

III. Proposed Method

A. Problem Definition

Few-shot aerial image semantic segmentation aims to parse novel-category objects in query aerial images with only a few annotated support aerial samples. Normally, episode-based meta-learning is adopted to perform the model training. The categories of datasets would be first divided into two nonoverlapping subsets, namely, C_{base} and C_{novel}. Subsequently, the aerial images containing base-category objects would be collected to construct the training dataset D_{train}, and the aerial images containing novel-category objects would be accumulated to build the test dataset D_{test}. Formally, given K-shot setting, during the training phase, $K+1$ labeled aerial images $\{(I_s^1, M_s^1), (I_s^2, M_s^2), \cdots (I_s^k, M_s^k), (I_q, M_q)\}$ of targeted category would be sampled from the D_{train} episodically, where (I_s^i, M_s^i) denotes the support image-mask pair and (I_q, M_q) denotes the query image-mask pair. The goal of the few-shot aerial image segmentation model is to learn how to precisely parse the query image under the guidance of K-shot annotated support samples. The predicted mask $\widetilde{M}_q$ is supervised by the ground truth mask M_q. Similarly, during the test stage, for each novel category, K shot annotated support samples would

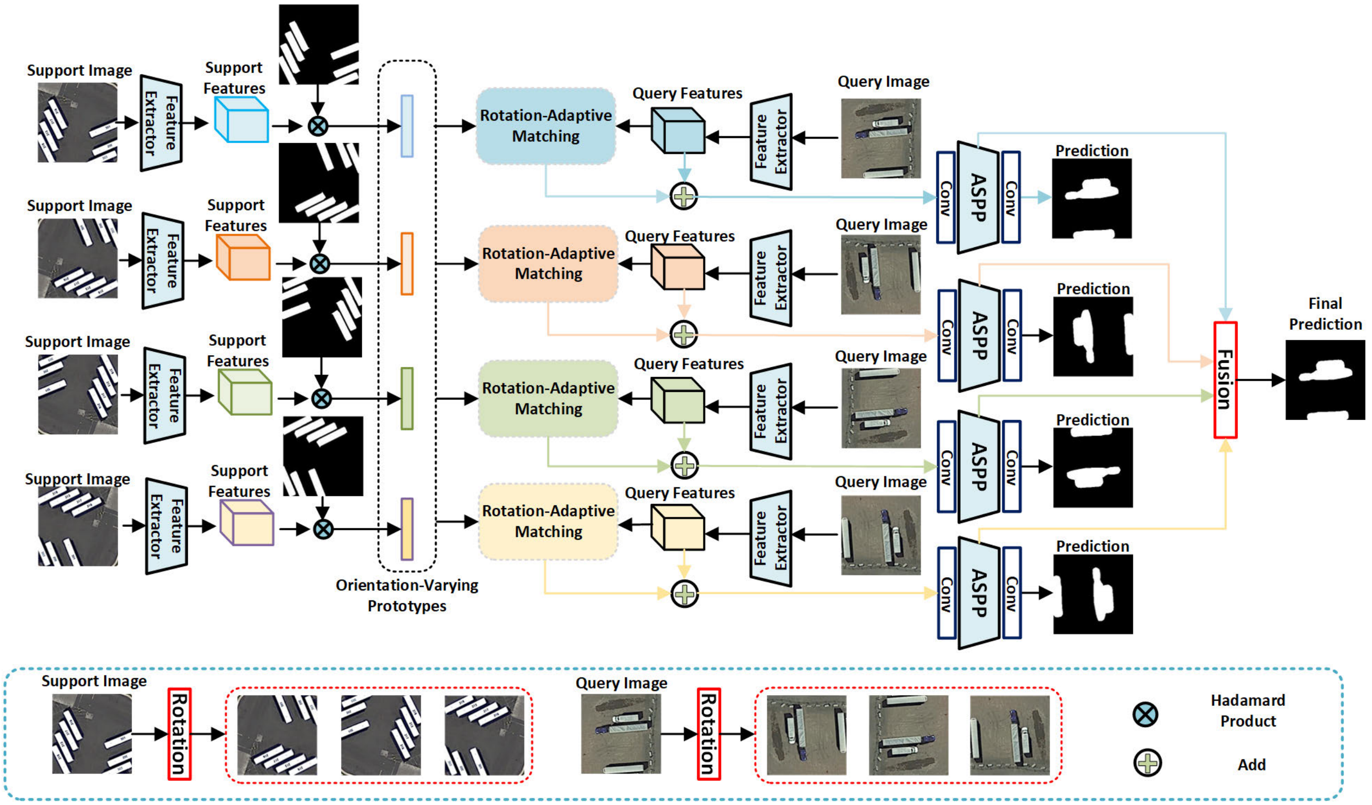


Fig. 2. Overall framework of the proposed FRINet. Original support and query images are first rotated to obtain support and query images with varying orientations. Then, the segmentation-pretrained backbones are leveraged as the feature extractor to obtain the corresponding support and query features. Then, with the category-consistent but orientation-varying prototypes, the network provides a rotation-adaptive matching for the orientation-varying query features. To encourage the network to make orientation-consistent predictions, the segmentation results from diverse orientations are supervised by the same ground truth. Finally, the predictions from different orientations are complementarily fused to obtain the final rotation-invariant prediction.

be sampled from the C_{novel} to inference the semantic objects of the query images.

B. Method Overview

As shown in Fig. 2, inspired by the intrinsic rotation invariance in aerial images, we propose a novel FRINet to efficiently segment aerial semantic objects with diverse orientations. First, different from previous FSS solutions that utilized backbones (VGG16 or Resnet50) pretrained in the ImageNet as feature extractors, the backbones are pretrained in the base category for segmentation. Then, the pretrained backbones (VGG16 or Resnet50) are leveraged as the feature extractor to obtain support and query features, respectively, for support and query images with varying orientations. In this manner, the newly pretrained backbones could provide a better solution for FSS segmentation. Subsequently, with given support masks and extracted support features, the rotation-varying but category-consistent support prototypes are collected through the mask average pooling operation. To provide a rotation-adaptive matching for aerial objects with varying orientations, the relation scores between orientation-varying query features and the generated support prototypes are computed. Leveraging the relation scores as guidance, each element of query features could obtain the newly generated orientation-adaptive prototype. By correspondingly utilizing these prototypes to activate the query features, the aerial objects with varying prototypes would be adaptively activated. We then propagate the activated features into the designed convolutional layers with the ASPP [45] to predict the final parsing result. To encourage the segmentation model to make consistent predictions on orientation-varying objects, the segmentation results for query images with varying orientations are supervised by the same ground truth mask. Finally, the predictions from different orientations are complementarily fused to generate the final rotation-invariant prediction.

C. Feature Extractor

Since the core idea of the FSS setting is learning how to transfer meta-knowledge from the base category to the novel category, weights frozen backbones are normally leveraged as the feature extractor to improve the generalization ability of the segmentation model. Previous FSS methods tend to use the backbones pretrained in the large-scale ImageNet, which works as the feature extractor for classification. However, the supervision from classification could only help the feature extractor learn a global visual pattern, this pattern clearly has limited ability to support the segmentation model in performing pixel-level classification. Thus, to provide a more detailed visual pattern for the feature extractor, a pretrained feature extractor for segmentation is a better solution. Thus, with the abundant annotated data for the base category, we pretrain these backbones as the feature extractor for segmentation. In this manner, the newly provided backbones

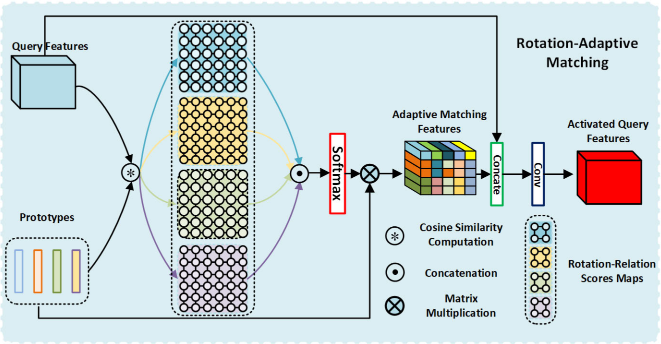

Query Features
Prototypes
Rotation-Adaptive Matching
Adaptive Matching Features
Softmax
Concate
Conv
Activated Query Features
Cosine Similarity Computation
Concatenation
Matrix Multiplication
Rotation-Relation Scores Maps

Fig. 3. Illustration of the rotation-adaptive matching. Utilizing relation scores as guidance, the prototypes are accumulated to construct the rotation-adaptive matching features. Through concatenation and convolutional layers, rotation-adaptive matching produces the activated query features.

could provide a better feature space for the few-shot aerial image semantic segmentation and basically boost the segmentation performance.

D. Rotation-Adaptive Matching

Existing FSS methods typically focus on how to mine better category prototypes from support samples to match the query images in a pixel-to-pixel manner. However, the category-consistent semantic objects in aerial images always appear in arbitrary orientations. The previous matching paradigm that does not take orientation into account would fail to activate all aerial objects. To conquer this challenge, as shown in Fig. 3, we propose rotation-adaptive matching to activate the orientation-varying aerial objects.

Assume we have the support image I_S and the query image I_Q, the support image and query image would be first rotated in different angles (90°, 180°, and 270°) to acquire rotated support images $\{I_{S1}, I_{S2}, I_{S3}\}$ and rotated query images $\{I_{Q1}, I_{Q2}, I_{Q3}\}$. Then, by propagating support images and query images with varying orientations, i.e., $\{I_S, I_{S1}, I_{S2}, I_{S3}\}$ and $\{I_Q, I_{Q1}, I_{Q2}, I_{Q3}\}$, into the pretrained feature extractor to obtain the corresponding support features $\{F_s, F_{s1}, F_{s2}, I_{s3}\}$ and query features $\{F_q, F_{q1}, F_{q2}, F_{q3}\}$, where the size of these features equals $C \times H \times W$.

Given the support mask M_s, we first rotate the original mask with the same angles to obtain the corresponding masks $\{M_s, M_{s1}, M_{s2}, M_{s3}\}$. Then, we, respectively, conduct masked average pooling operations on the orientation-varying support features to obtain the orientation-varying prototypes $\{P_s, P_{s1}, P_{s2}, P_{s3}\}$

$$\underset{i\in\{s,s1,s2,s3\}}{P_i} = \text{Avgpool}(F_i \odot M_i) \in \mathbb{R}^{C\times 1\times 1}. \tag{1}$$

With these category-consistent prototypes with varying orientations, we could leverage the similarities between prototypes and each pixel of the query to provide the rotation-adaptive matching features in an aggregating manner. The rotation-adaptive matching is applied to all query features $\{F_q, F_{q1}, F_{q2}, F_{q3}\}$. For better understanding, we then take F_q for a more detailed explanation.

Query feature F_q is first flattened in spatial dimension as query feature sets $\{l_1^q, l_2^q, \cdots, l_N^q\}$, where $l \in \mathbb{R}^{C\times 1\times 1}$ and $N = H \times W$. Subsequently, we choose the cosine similarity as the correlation function to compute the pixel-level relation scores between query features and orientation-varying prototypes

$$\cos\left(P_i, l_j^q\right) = \frac{P_i l_j^q}{\left\|P_i\right\|\left\|l_j^q\right\|}, \quad i \in \{s, s1, s2, s3\}, \quad j \in \{1, 2, \ldots, N\}. \tag{2}$$

By collecting relation cores for each element of query features, we could obtain the rotation-relation score maps $\{m_s, m_{s1}, m_{s2}, m_{s3}\} \in \mathbb{R}^{1\times H\times W}$. Then, by concatenating these score maps in channel dimension, and applying softmax operation in these maps, we could acquire the rotation-relation matrix $m_p \in \mathbb{R}^{4\times H\times W}$ between query features and orientation-varying prototypes

$$m_p = \text{Softmax}(\text{Cat}(m_s, m_{s1}, m_{s2}, m_{s3})). \tag{3}$$

It is notable that each score in the relation matrix could denote the orientation distance of each element of query features. Thus, according to the distance, the rotation-adaptive prototypes P_r could be computed for each element, by collecting these prototypes, we could obtain the rotation-adaptive matching features $F_r \in \mathbb{R}^{C\times H\times W}$. This process could be achieved by the matrix multiplication between the orientation-varying prototypes and rotation-relation matrix m_p

$$F_r = \text{Cat}(P_s, P_{s1}, P_{s2}, P_{s3}) \otimes M_p \tag{4}$$

where ⊗ denotes matrix multiplication.

Subsequently, to corresponding activate the aerial objects, the rotation-adaptive matching features are concatenated with the query feature and further propagated into the convolutional layers to obtain the activated query features F_a

$$F_a = \text{Conv}\left(\text{Cat}\left(F_r, F_q\right)\right). \tag{5}$$

E. Rotation-Invariant Segmentation

As shown in Fig. 4, aiming at encouraging the segmentation model to make consistent predictions, the segmentation results from orientation-varying query images are supervised with the same ground truth labels. Then, the segmentation predictions from different orientations are complementarily fused to obtain the final rotation-invariant result. Specifically, after applying rotation-adaptive matching to query features $\{F_q, F_{q1}, F_{q2}, F_{q3}\}$, we could obtain the correspondingly activated query features $\{F_a, F_{a1}, F_{a2}, F_{a3}\}$. By propagating these activated query features into the ASPP modules and designing convolutional layers, we could infer the segmentation results in $\{R_a, R_{a1}, R_{a2}, R_{a3}\}$

$$\underset{i\in\{a,a1,a2,a3\}}{R_i} = \text{Conv}(\text{ASPP}(\text{Conv}(F_i))) \tag{6}$$

$$\underset{i\in\{a,a1,a2,a3\}}{R_i} = \left\{R_i^f, R_i^b\right\} \tag{7}$$

where the $R_i \in \mathbb{R}^{2\times H\times W}$, R_i^f denotes the foreground probability map and R_i^b denotes the background probability map. In order to leverage the same ground truth M_{GT} as the

TABLE I

CATEGORIES IN THE iSAID-5^i DATASET

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
ship	storage tank	baseball diamond	tennis court	basketball court	ground track field	bridge	large vehicle	small vehicle	helicopter	swimming pool	roundabout	soccer ball field	plane	harbor

TABLE II

TESTING CLASSES AND TRAINING CLASSES FOR THE THREEFOLD CROSS-VALIDATION. THE TRAINING CLASSES OF iSAID-5^i, $i = 0, 1, 2$ ARE DISJOINT WITH THE TESTING CLASSES

Dataset	Training classes	Testing classes
iSAID-5^0	ground track field, bridge, large vehicle, small vehicle, helicopter, swimming pool, roundabout, soccer ball field, plane, harbor	ship, storage tank, baseball diamond, tennis court, basketball court
iSAID-5^1	ship, storage tank, baseball diamond, tennis court, basketball court, swimming pool, roundabout, soccer ball field, plane, harbor	ground track field, bridge, large vehicle, small vehicle, helicopter
iSAID-5^2	ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, bridge, large vehicle, small vehicle, helicopter	swimming pool, roundabout, soccerball field, plane, harbor

supervision, all these segmentation results would be rotated inversely in corresponding angles

$$\text{loss}_{\text{rotation}} = \sum_{i\in\{a,a1,a2,a3\}} \text{cross\_entropy}(M_{\text{GT}}, \text{Ro}(R_i)) \quad (8)$$

where Ro denotes that rotating the segmentation results in corresponding angles. Subsequently, these segmentation results are complementarily fused to obtain the final rotation-invariant segmentation results. Particularly, through convolutional neural layers, the foreground probability maps, and background probability maps are, respectively, fused to obtain the final foreground probability map and background probability map, which are further concatenated to obtain the final rotation-invariant segmentation result R_F

$$R_F^f = \text{Conv}\left(\text{Cat}\left(R_a^f, \text{Ro}\left(R_{a1}^f\right), \text{Ro}\left(R_{a1}^f\right), \text{Ro}\left(R_{a1}^f\right)\right)\right) \quad (9)$$

$$R_F^f = \text{Conv}\left(\text{Cat}\left(R_a^f, \text{Ro}\left(R_{a1}^f\right), \text{Ro}\left(R_{a1}^f\right), \text{Ro}\left(R_{a1}^f\right)\right)\right) \quad (10)$$

$$R_F = \left\{R_F^f, R_F^b\right\}. \quad (11)$$

The final rotation-invariant segmentation result is also supervised by the ground truth mask M_{GT}

$$\text{loss}_{\text{main}} = \text{cross\_entropy}(M_{\text{GT}}, R_F). \quad (12)$$

The overall supervision loss is computed as

$$\text{loss}_{\text{all}} = \text{loss}_{\text{main}} + \mu * \text{loss}_{\text{rotation}} \quad (13)$$

where μ is the adjusting factor, and the factor is empirically set as 0.25 in this article.

IV. EXPERIMENTS

To demonstrate the effectiveness of the proposed FRINet, extensive experiments are performed on the few-shot aerial segmentation benchmark, and the introduction of this section is illustrated as follows. We first describe the few-shot aerial segmentation benchmark and the evaluation metric. Then, the implementation details of FRINet are provided for better realization. Subsequently, the comparisons between the proposed FRINet and other advanced FSS solutions are present, and we analyze the segmentation results comprehensively. Finally, a series of ablation studies are conducted to illustrate the differences between FRINet and random flipping strategy and demonstrate the impact of each component in our proposed FRINet.

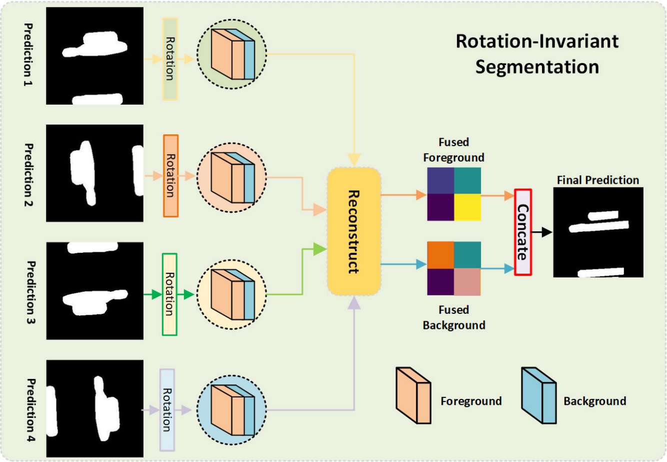


Fig. 4. Illustration of the rotation-invariant segmentation. The segmentation results from diverse orientations are complementarily fused to obtain the final rotation-invariant predictions.

A. Dataset and Evaluation Metric

We follow the experimental setting in SDM [18] to utilize the iSAID-5^i dataset as the benchmark. The iSAID-5^i dataset is constructed based on the iSAID dataset [17]. It mainly consists of 655 451 object instances across 15 categories, and the details of the categories are illustrated in Table I. Particularly, for the 15 object categories in the iSAID-5^i dataset, the cross-validation method is leveraged to evaluate the proposed model by sampling five classes as test categories D_{test} and leveraging the left ten classes as the categories of the training setD_{train}. The details of the class splits are shown in Table II, where i is the fold number. The iSAID-5^i dataset contains 18 076 images for training and 6363 images

TABLE III
PERFORMANCE COMPARISONS OF DIFFERENT METHODS ACROSS DIFFERENT SPLITS ON THE iSAID-5^i DATASET

Method	1-shot				5-shot			
	Split-0	Split-1	Split-2	mean	Split-0	Split-1	Split-2	mean
VGG16								
PANet [20]	17.43	11.43	15.95	14.94	17.70	14.58	20.70	17.66
CANet [21]	19.73	17.98	30.93	22.88	23.45	20.53	30.12	24.70
PMMs [28]	20.87	16.07	24.65	20.53	23.31	16.61	27.43	22.45
PFENet [24]	16.68	15.30	27.87	19.95	18.46	18.39	28.81	21.89
SDM [18]	29.24	20.80	34.73	28.26	36.33	27.98	42.39	35.57
DCP [62]	23.52	18.98	28.84	23.78	25.45	20.70	29.41	25.19
NERTNet [30]	9.15	13.57	9.81	10.84	6.95	20.06	11.75	12.92
HDMNet [63]	19.11	14.23	22.98	18.77	19.29	14.66	23.38	19.11
Ours	**42.86**	**36.73**	**42.51**	**40.70**	**43.80**	**37.18**	**45.41**	**42.13**
Resnet50								
PANet [20]	12.36	9.11	12.05	11.17	13.82	12.40	19.12	15.11
CANet [21]	18.80	15.62	25.79	20.07	23.86	18.54	32.00	24.80
PMMs [28]	19.02	18.51	28.42	21.98	20.89	20.87	31.23	24.33
PFENet [24]	18.75	17.24	22.09	19.36	19.57	18.43	26.14	21.38
SDM [18]	34.29	22.25	35.62	30.72	39.88	30.59	45.70	38.72
DCP [62]	26.66	24.56	21.68	23.34	28.81	24.21	31.73	28.25
NERTNet [30]	8.15	18.59	11.85	12.86	9.24	20.60	10.78	13.54
HDMNet [63]	36.04	32.12	34.87	34.34	39.95	36.01	40.68	38.88
Ours	**46.50**	**36.94**	**43.93**	**42.59**	**48.85**	**38.05**	**46.46**	**44.45**

for validation, where the size of images equals 3 × 256 × 256. Following the evaluation setting in [18], we adopt the class mean intersection over union (mIOU) as our evaluation metric, which could directly reflect the model performance. Formally, the mIOU could be defined as follows:

$$\mathrm{mIOU} = \frac{1}{C}\sum_{i=1}^{C}\mathrm{IOU}_i \tag{14}$$

where C is the number of categories in each split and IOU_i is the intersection over union of class i.

B. Implement Details

We leverage the VGG16 and Resnet50 pretrained in base categories as the feature extractors. Specifically, the VGG16 and Resnet50 are leveraged as the backbone of PSPNet [46], and the abundant base-category samples are utilized to train the PSPNet. After this training process, the pretrained backbones could provide a generalized segmentation feature space for the following few-shot segmentation phase. The overall network is built on Pytorch. The SGD optimizer with an initial learning rate of 5e-3 is used for updating the parameters, and the training batch size is set to 4. To facilitate the generalization ability of FRINet, the weights of the backbones are frozen, and only the weights of the rest network are learnable. Both Query and support images share the same-weight feature extractor and are propagated into the segmentation model with a size of 3 × 256 × 256. All the training images are augmented with random horizontal flipping. For the one-shot setting, the training phase lasts for 100 epochs, and the training process lasts for 50 epochs for the five-shot setting. To better demonstrate the superiority of our proposed segmentation model, a series of performance comparison experiments are performed in three splits of iSAID-5^i at one-shot and five-shot settings.

C. Performance Analysis

The comparisons between our proposed FRINet and other advanced FSS algorithms with the available source codes in the iSAID-5^i, dataset are shown in Table III, and the best performance in each comparison is bolded. Clearly, our proposed FRINet achieves the best segmentation performance and could acquire better performance with the reinforcement of backbones. With the VGG16 backbone, for the one-shot setting, FRINet achieves 40.70 mean mIOU performance, which brings 12.44% mIOU improvement. The best performance is achieved in the split0 with 42.86 mIOU performance. For the five-shot setting, FRINet achieves 42.13 mean mIOU performance, which brings a 6.56% mIOU improvement. The best performance is achieved in the split2 with a 45.41 mIOU performance. Particularly, with both one-shot and five-shot settings, all splits could achieve more than 3% mIOU improvement. This suggests the FRINet could truly bring a splendid performance boost. Simultaneously, with the Resnet50 backbone, FRINet could achieve 42.59 mean mIOU performance for a one-shot setting, which brings 11.87% mIOU improvement. Similar to the performance with VGG16, the best performance is obtained in the split0 with 46.50 mIOU performance. For the five-shot setting, FRINet acquires 44.45 mean mIOU performance, the best performance is also obtained in the spilt0 with 48.85 mIOU. We could see that through our proposed FRINet, all splits could acquire an impressive performance boost. This phenomenon powerfully illustrates the effectiveness of the proposed FRINet and proves the

TABLE IV

PERFORMANCE COMPARISONS OF DIVERSE CLASSES ON THE iSAID-5^i DATASET WITH ONE-SHOT SETTING

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
VGG16															
PANet [20]	13.85	17.78	19.78	17.80	17.94	22.41	11.75	11.75	6.36	4.88	12.96	19.11	23.49	11.87	12.33
CANet [21]	7.51	12.87	31.22	30.45	16.61	18.31	31.94	23.15	3.77	12.55	33.99	51.41	36.69	17.61	14.91
PMMs [28]	12.84	23.06	27.81	26.65	13.97	18.48	13.76	25.58	11.68	10.88	28.52	34.62	34.15	16.32	9.64
PFENet [24]	3.94	4.70	30.95	29.21	14.59	15.01	14.63	27.68	10.13	9.06	29.40	47.89	25.23	25.20	11.63
SDM [18]	26.55	30.57	33.01	35.06	21.03	20.54	28.89	32.45	9.62	12.59	**30.79**	43.52	49.47	27.51	**22.37**
DCP [62]	19.70	2.71	34.21	38.71	22.28	19.84	3.77	35.78	18.26	17.26	20.08	28.05	46.50	29.97	19.61
NERTNet [30]	0.30	**36.90**	5.12	2.14	1.59	2.21	24.23	14.70	13.45	13.29	9.31	7.49	14.48	4.89	12.85
HDMNet [63]	17.66	0.38	28.60	27.37	21.55	26.64	12.74	15.21	5.43	11.15	12.08	30.57	43.88	11.92	16.43
Ours	**37.26**	18.21	**43.32**	**60.56**	**54.96**	**26.86**	**39.84**	**52.26**	**31.29**	**33.39**	28.57	**65.20**	**51.08**	**48.27**	19.41
Resnet50															
PANet [20]	7.27	9.99	14.49	13.20	16.85	17.40	10.23	7.37	6.21	4.33	8.35	12.38	21.28	8.52	9.75
CANet [21]	7.25	27.25	43.32	6.17	9.99	6.16	17.81	28.58	8.84	16.73	17.73	56.97	13.82	27.41	13.06
PMMs [28]	12.93	17.92	41.97	7.43	14.87	13.24	32.67	26.08	4.17	16.37	12.26	61.49	38.37	17.92	12.02
PFENet [24]	2.98	6.30	33.29	34.49	16.71	8.68	8.67	34.62	18.19	16.08	28.58	24.47	7.37	39.76	10.31
SDM [18]	37.66	34.37	34.45	39.81	25.14	16.77	34.53	30.50	12.42	17.02	20.69	**56.83**	42.80	40.52	17.26
DCP [62]	21.14	0.50	35.15	46.52	30.51	**27.75**	19.87	39.44	13.37	22.37	28.55	14.79	24.97	33.20	7.77
NERTNet [30]	0.50	27.26	8.59	2.72	2.17	9.68	33.36	24.28	13.86	11.80	6.63	8.29	14.37	6.92	23.03
HDMNet [63]	**40.67**	3.58	40.74	**56.69**	38.53	24.16	40.05	53.05	24.42	18.92	**39.03**	42.44	28.73	51.24	12.92
Ours	35.80	**52.96**	**43.75**	53.36	**46.61**	24.99	**48.87**	**63.38**	**27.58**	**19.89**	34.30	53.44	**55.52**	**52.44**	**23.94**

TABLE V

PERFORMANCE COMPARISONS OF DIVERSE CLASSES ON THE iSAID-5^i DATASET WITH FIVE-SHOT SETTING

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
VGG16															
PANet [20]	13.40	17.79	19.04	18.11	20.15	22.97	16.18	15.65	9.37	8.73	31.65	21.63	25.36	12.70	12.15
CANet [21]	11.87	18.36	33.26	26.05	27.69	19.20	29.93	30.37	11.76	11.37	23.76	60.25	28.67	21.77	16.16
PMMs [28]	14.99	25.09	28.49	30.40	17.58	18.37	13.40	27.12	12.45	11.68	35.46	36.57	37.08	18.41	9.63
PFENet [24]	2.64	8.57	29.09	30.76	21.26	15.22	21.58	31.16	12.06	11.90	33.56	48.47	24.92	25.29	11.79
SDM [18]	30.81	**38.59**	37.34	42.58	32.35	**32.32**	39.71	37.23	17.73	12.93	39.20	**66.78**	39.06	46.17	20.72
DCP [62]	24.78,	2.26	38.27	36.99	24.93	25.87	4.65	36.98	14.31	21.69	23.05	24.06	38.88	42.72	18.34
NERTNet [30]	4.74	3.56	11.21	6.96	7.99	5.25	45.54	16.18	17.48	15.84	5.00	34.20	16.20	1.17	2.20
HDMNet [63]	17.39	0.44	28.16	27.19	23.28	26.00	15.18	14.52	5.52	12.06	13.10	25.24	46.56	13.44	18.56
Ours	**35.57**	8.01	**58.46**	**68.03**	**48.96**	24.73	**47.47**	**54.54**	**26.30**	**32.87**	**42.97**	56.07	**53.19**	**51.55**	**23.25**
Resnet50															
PANet [20]	8.45	11.68	17.81	14.22	16.93	18.86	15.70	9.40	7.90	10.13	21.41	24.94	28.41	9.80	11.04
CANet [21]	32.47	6.49	39.8	25.97	14.59	25.01	25.27	32.06	1.31	9.08	26.99	73.84	33.83	20.69	4.69
PMMs [28]	13.44	22.22	42.12	8.58	18.06	13.17	37.88	30.69	6.41	16.21	14.62	65.02	42.99	20.86	12.66
PFENet [24]	10.13	9.48	30.71	31.23	16.31	11.53	14.15	36.07	15.58	14.82	38.52	20.90	20.41	38.66	12.22
SDM [18]	38.76	49.06	50.06	39.25	22.26	30.68	45.34	41.49	20.21	15.21	32.61	**66.64**	**57.41**	49.12	**22.71**
DCP [62]	30.79	1.02	39.29	43.27	29.24	27.02	16.63	39.08	16.06	22.26	26.29	37.54	43.23	31.36	20.24
NERTNet [30]	0.50	40.75	3.16	1.48	0.79	13.49	37.21	19.25	15.99	17.09	8.53	9.56	10.52	10.70	14.61
HDMNet [63]	**42.39**	5.05	**53.34**	**67.61**	31.38	29.77	**50.90**	52.77	20.89	25.72	43.04	55.02	38.44	52.90	13.99
Ours	28.28	**49.50**	50.40	65.06	**51.01**	**33.84**	47.26	**57.01**	**22.40**	**29.74**	**48.87**	58.15	47.44	**56.67**	21.20

FRINet successfully encourages the improvement of few-shot aerial segmentation.

To further analyze the performance of diverse classes with the few-shot setting, the detailed segmentation results are collected, and the results are, respectively, illustrated in Tables IV and V. As shown in Table IV, the proposed FRINet could achieve the best parsing performance in almost all categories. Particularly, the best performance is 60.56 mIOU in the tennis court (C4) and 63.38 mIOU in the large vehicle (C8), respectively, for the VGG16 and Resnet50 backbones. Moreover, the best improvement is 32.68% mIOU improvement in the basketball court (C5) and 18.59% mIOU improvement in the storage tank (C2), respectively, for the VGG16 and Resnet50 backbones. These results suggest that the rotation invariance is very critical for these categories, and the proposed FRINet successfully mines the rotation invariance. However, our proposed FRINet still fails in some categories. For instance, a huge performance gap could be seen in the storage tank (C2). This phenomenon demonstrates that the proposed FRINet may still have limited ability to mine the rotation invariance of small-scale aerial objects like the storage tank.

Focusing on the performance in Table V, the proposed FRINet still obtains the best segmentation performance in almost all categories. Specifically, the best performance are 68.03 mIOU and 65.06 mIOU in the tennis court (C4),

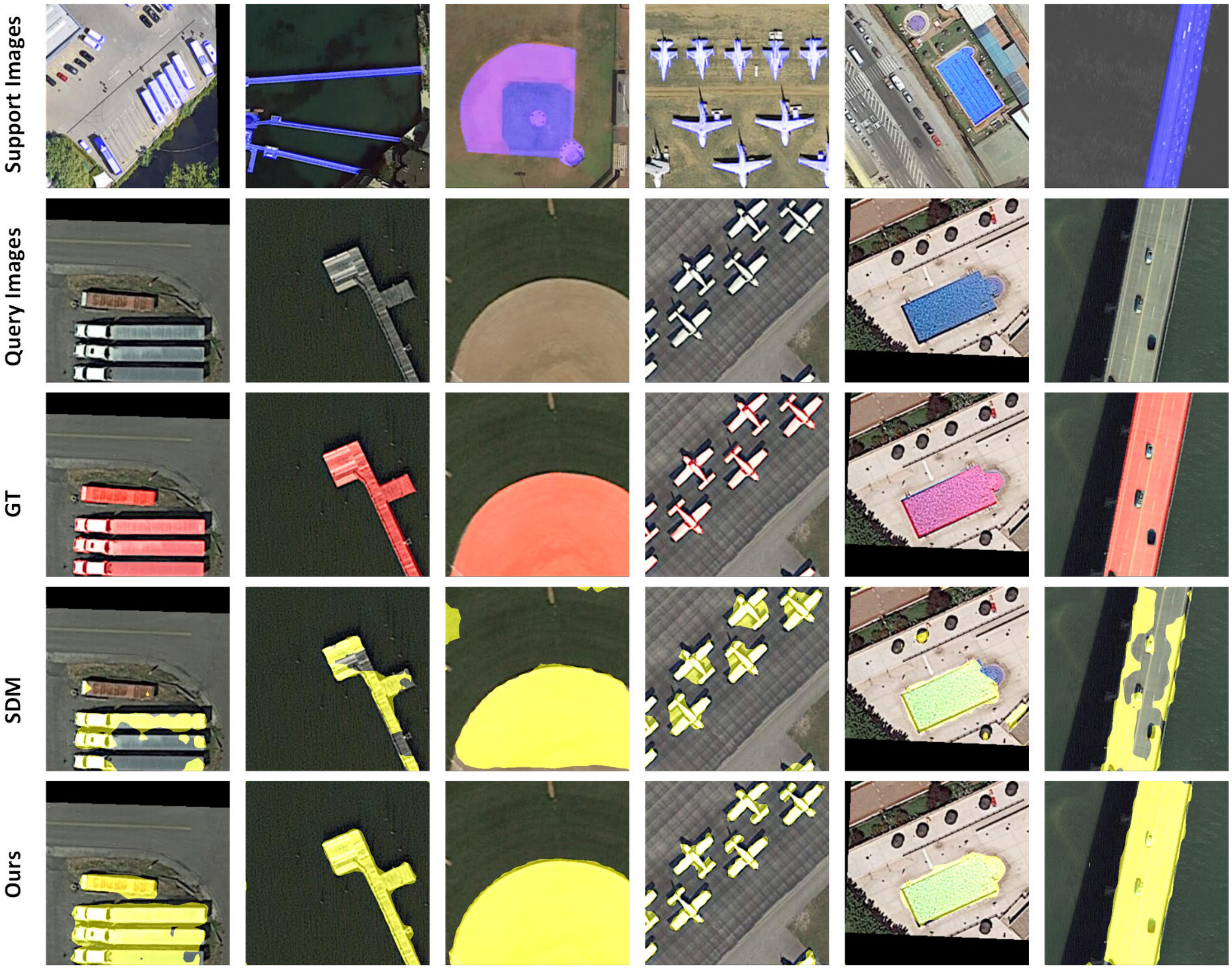


Fig. 5. Qualitative results of proposed FRINet. From top to bottom: support images, query images, the ground truth of query images, predictions of SDM network, and predictions of proposed FRINet.

respectively, for the VGG16 and Resnet50 backbones. Furthermore, the best performance boost is also achieved in the tennis court (C4) and swimming pool (C11), i.e., 25.45% and 5.83% mIOU improvement. This phenomenon again tells us the rotation invariance of the tennis court (C4) is well-mined by the FRINet. However, similarly, the performance in small-scale aerial objects like the storage tank is still very limited. We believe the main reason is that the resolution of the features of small-scale aerial objects is also very small, thus the change of rotation could not call a string-enough feature change to support the FRINet to mine the inner rotation invariance. This phenomenon could inspire further work to focus on designing an appropriate method to mine the rotation-invariance information-limited small-scale aerial objects.

If we scratch a little deeper into the performance, we can see that compared with the performance obtained through Resent50, our proposed method through VGG16 could achieve a more impressive performance improvement. The main reason could be explained as follows: In our method, we incorporate the rotation invariance as guidance to boost the few-shot aerial segmentation with different backbones. Meanwhile, it is known to us that Resnet50 could extract more valuable information than VGG16, which denotes method based on Restnet50 could be a better baseline than the method based on VGG16. Thus, the rotation invariance could offer more beneficial information for VGG16, and relatively less beneficial information for Resnet50. Correspondingly, with the VGG16, we could see more impressive performance improvements.

To further directly analyze the performance of the proposed FRINet, the qualitative results are shown in Fig. 5. Obviously, the pleasant segmentation performance successfully demonstrates the effectiveness of the proposed FRINet. Particularly, rotation-adaptive matching could adaptively activate aerial objects with varying orientations, and the rotation-invariant segmentation helps the network eliminate the background and recognize more missing foreground. For instance, in the first line, the orientation of large vehicles in the support image is totally different from the orientation of large vehicles in the query image. Thus, the previous method fails to recognize these orientation-changed objects. But our proposed FRINet

TABLE VI

ABLATION STUDY OF THE NEWLY PROVIDED BACKBONES. NB DENOTES NEWLY PROVIDED BACKBONES

Methods	Split-0	Split-1	Split-2	Mean
Baseline	29.24	20.80	34.73	28.26
Baseline+NB	**40.33**	**34.65**	**39.07**	**38.01**

TABLE VII

ABLATION STUDY OF THE ROTATION-ADAPTIVE MATCHING AND ROTATION-INVARIANT SEGMENTATION. RAM ILLUSTRATES THE ROTATION-ADAPTIVE MATCHING, RIS DENOTES THE ROTATION-INVARIANT SEGMENTATION

RAM	RIS	Split-0	Split-1	Split-2	Mean
		40.33	34.65	39.07	38.01
✓		41.44	35.04	41.91	39.46
	✓	40.75	35.07	41.72	39.18
✓	✓	**42.86**	**36.73**	**42.51**	**40.70**

TABLE VIII

DIFFERENCE BETWEEN THE PROPOSED FRINET AND THE TRADITIONAL ROTATION DATA AUGMENTATION. W/O DENOTES WITHOUT, W DENOTES WITH, AND AUG MEANS THE ROTATION DATA AUGMENTATION

Methods	Split-0	Split-1	Split-2	Mean
w/o Aug	39.33	33.36	38.33	37.01
w Aug	40.33	34.65	39.07	38.01
Ours	**42.86**	**36.73**	**42.51**	**40.70**

TABLE IX

ABLATION STUDY OF ORIENTATIONS

Orientations	Split-0	Split-1	Split-2	Mean
[0°]	41.44	35.04	41.91	39.46
[0°,90°]	41.98	35.99	42.11	40.03
[0°,90°,180°]	42.36	36.43	42.39	40.39
[0°,90°,180°,270°]	**42.86**	**36.73**	**42.51**	**40.70**

TABLE X

ABLATION STUDY OF DISCRETE ANGLES

Orientations	Split-0	Split-1	Split-2	Mean
[0°,90°,180°,270°]	**42.86**	**36.73**	**42.51**	**40.70**
[0°,45°,...,270°,315°]	42.73	36.45	42.41	40.53
[0°,30°,...,300°,330°]	42.19	36.12	41.98	40.10
[0°,20°,320°,340°]	41.32	35.76	41.17	39.42

successfully recognizes these orientation-varying objects. Furthermore, we could see a similar phenomenon in the second row (the harbor) and in the fifth row (the swimming pool).

Interestingly, for the fourth row (the airplane category), we found that our proposed FRINet could eliminate some very tiny but essential backgrounds. We believe these impressive results benefit from the rotation-invariant segmentation. The rotation-invariant segmentation encourages the network to make consistent predictions from diverse orientations and further fuses these predictions in a complementary manner. Apparently, this manner could help the network to accumulate the segmentation results in different views, which could produce a more precise parsing result from coarse results.

D. Ablation Study

To evaluate the effectiveness of our proposed FRINet, a series of experiments are performed to analyze the effect of the key components of the FRINet. All ablation experiments are conducted under the VGG16 backbone and one-shot setting.

First, the effect of the newly provided backbones is studied. As shown in Table VI, the newly provided backbone brings very impressive performance improvement. For split0 and split1, it provides more than 10% mIOU boost. For the split2, it also brings nearly 5% mIOU improvement. Overall, it gives 9.76% performance improvement. All these improvements prove the newly provided backbones truly provide a better feature space for the few-shot aerial segmentation task.

Then, based on the newly provided backbones, we further study the effect of the proposed rotation-adaptive matching and the rotation-invariant segmentation, i.e., the rotation-varying segmentation results supervised by the same ground truth are fused to obtain the final parsing result in a complementary manner. As shown in Table VII, RAM denotes the rotation-adaptive matching, and RIS denotes the rotation-invariant segmentation. The rotation-adaptive matching could bring a performance boost for splits. Particularly, rotation-adaptive matching leads to the best performance enhancement, namely, 2.84% mIOU improvement for split2. This result implies that more orientation-varying aerial objects are not activated for split2 in previous FSS solutions. For the rotation-invariant segmentation, the best performance improvement, namely, 2.64% mIOU improvement, is also obtained in the split2. This phenomenon tells us that some orientation-varying aerial objects are falsely recognized. By jointly leveraging RAM and RIS, the proposed FRINet successfully achieves 40.70 mIOU performance. These performance improvements powerfully illustrate that the proposed rotation-adaptive matching and the rotation-invariant segmentation could efficiently help the network parse the rotation-varying aerial objects.

To distinguish our proposed FRINet from the traditional rotation data augmentation, some ablation experiments are conducted, and the experimental results are shown in Table VIII. The traditional rotation data augmentation randomly rotates the images at a limited angle. Though this strategy could provide different orientation-views for training, only a single orientation-view is leveraged in the query-support matching. The orientation-varying objects could still not be activated and be falsely recognized. Moreover, as the experimental results imply, our FRINet could bring a more impressive performance boost. Thus, our FRINet is totally different from the traditional rotation data augmentation strategy.

In our FRINet, rotated images with three angles (90°, 180°, and 270°) and original images are adopted in our FRINet. To demonstrate the necessity of diverse angles, some ablation experiments are performed and results are shown in Table IX. Clearly, with the addition of orientations, all splits could an increasing of mIOU performance, and the best performance is obtained when all orientations are included in the network. This phenomenon illustrates that more orientations could help FRINet parse more orientation-varying objects.

To study the effect of discrete angles, several ablation study experiments are conducted and the experimental results are shown in Table X. The first row denotes our method with a 90° split, the second row denotes a 45° split, and

TABLE XI

Efficiency Comparison With ResNet50 on PASCAL-5i in One-Shot Setting. Params Denotes Learnable Parameters and FPS Denotes the Frame-per-Second

Methods	mIOU	FPS	Params
PANet [20]	12.36	9.11	12.05
CANet [21]	18.80	15.62	25.79
PMMs [28]	19.02	18.51	28.42
PFENet [24]	18.75	17.24	22.09
SDM [18]	34.29	22.25	35.62
DCP [62]	26.66	24.56	21.68
NERTNet [30]	8.15	18.59	11.85
HDMNet [63]	36.04	32.12	34.87
Ours	**46.50**	**36.94**	**43.93**

other rows follow this manner. As observed in Table X, when considering orientations more densely, there is a gradual decrease in performance. We found that using four orientations with a 90° split is sufficient to handle rotation invariance effectively in our proposed method. Beyond this point, additional orientations tend to introduce redundant information and hinder the segmentation process with the demoted performance.

E. Discussion

Mining rotation invariance in our approach does indeed come with an associated increase in computation time. Specifically, the forward process of the feature extractor and segmentation predictions requires four times the inference time compared to the nonrotated counterpart. In addition, providing simultaneous supervision for multiple predictions prolongs the backward process. To qualitatively analyze it, we have followed the setting in the ablation study to conduct an efficiency comparison, and comparison results are shown in Table XI. Obviously, our method has relatively lower FPS, yet it's crucial to emphasize that our method consistently achieves state-of-the-art performance by effectively harnessing rotation invariance for few-shot aerial image semantic segmentation. This improved performance validates the tradeoff in inference time. Looking ahead, we acknowledge the importance of striking a balance between exploiting rotation invariance and minimizing inference time. In our future work, we are dedicated to exploring more efficient techniques that will allow us to leverage rotation invariance with a lighter network and reduced inference time, ensuring practicality and efficiency in real-world applications.

V. Conclusion

In this article, we propose a novel FRINet. Particularly, aiming at activating aerial objects with varying orientations, orientation-varying yet category-consistent support information is leveraged to perform the rotation-adaptive matching. Meanwhile, rotation-varying segmentation results supervised by the same ground truth are fused to obtain the final rotation-invariant parsing result in a complementary manner. In addition, the backbones pretrained in the base categories are newly provided to offer a better feature space. The extensive experiments in the few-shot aerial semantic segmentation benchmark demonstrate our model achieves state-of-the-art performances.

References

[1] B. C. Forster, "An examination of some problems and solutions in monitoring urban areas from satellite platforms," *Int. J. Remote Sens.*, vol. 6, no. 1, pp. 139–151, Jan. 1985.

[2] M. K. Jat, P. K. Garg, and D. Khare, "Monitoring and modelling of urban sprawl using remote sensing and GIS techniques," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 10, no. 1, pp. 26–43, Feb. 2008.

[3] C. Ru et al., "Land Surface Temperature retrieval from Landsat 8 thermal infrared data over urban areas considering geometry effect: Method and application," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021, Art. no. 5000716.

[4] C. Ayala, R. Sesma, C. Aranda, and M. Galar, "A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery," *Remote Sens.*, vol. 13, no. 16, p. 3135, Aug. 2021.

[5] W. Qiao, L. Shen, J. Wang, X. Yang, and Z. Li, "A weakly supervised semantic segmentation approach for damaged building extraction from postearthquake high-resolution remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[6] C. Ding, L. Weng, M. Xia, and H. Lin, "Non-local feature search network for building and road segmentation of remote sensing image," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 4, p. 245, Apr. 2021.

[7] M. Zhao, S. Li, S. Xuan, L. Kou, S. Gong, and Z. Zhou, "SatSOT: A benchmark dataset for satellite video single object tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, Art. no. 5617611.

[8] E. Macioszek and A. Kurek, "Extracting road traffic volume in the city before and during COVID-19 through video remote sensing," *Remote Sens.*, vol. 13, no. 12, p. 2329, Jun. 2021.

[9] L. Jian, Z. Li, X. Yang, W. Wu, A. Ahmad, and G. Jeon, "Combining unmanned aerial vehicles with artificial-intelligence technology for traffic-congestion recognition: Electronic eyes in the skies to spot clogged roads," *IEEE Consum. Electron. Mag.*, vol. 8, no. 3, pp. 81–86, May 2019.

[10] D. Chai, S. Newsam, and J. Huang, "Aerial image semantic segmentation using DCNN predicted distance maps," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 309–322, Mar. 2020.

[11] G. Deng, Z. Wu, C. Wang, M. Xu, and Y. Zhong, "CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021, Art. no. 4401120.

[12] H. Luo, C. Chen, L. Fang, X. Zhu, and L. Lu, "High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3492–3507, Sep. 2019.

[13] A. Zheng, M. Wang, C. Li, J. Tang, and B. Luo, "Entropy guided adversarial domain adaptation for aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5405614.

[14] R. Zhou et al., "Weakly supervised semantic segmentation in aerial imagery via cross-image semantic mining," *Remote Sens.*, vol. 15, no. 4, p. 986, Feb. 2023.

[15] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[16] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020.

[17] S. W. Zamir et al., "Isaid: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 28–37.

[18] X. Yao, Q. Cao, X. Feng, G. Cheng, and J. Han, "Scale-aware detailed matching for few-shot aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, Art. no. 5611711.

[19] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "SG-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3855–3865, Sep. 2020.

[20] 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. 9196–9205.

[21] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5212–5221.

[22] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning," in *Brit. Mach. Vis. Conf.*, 2018, vol. 3, no. 4, pp. 1–13.

[23] G.-S. Xie, H. Xiong, J. Liu, Y. Yao, and L. Shao, "Few-shot semantic segmentation with cyclic memory network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7273–7282.

[24] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1050–1065, Feb. 2022.

[25] G. Cheng, C. Lang, and J. Han, "Holistic prototype activation for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4650–4666, Apr. 2023.

[26] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Base and meta: A new perspective on few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10669–10686, Sep. 2023.

[27] Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang, "Self-support few-shot semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 701–719.

[28] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 763–778.

[29] H. Wang, X. Zhang, Y. Hu, Y. Yang, X. Cao, and X. Zhen, "Few-shot semantic segmentation with democratic attention networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 730–746.

[30] Y. Liu, N. Liu, Q. Cao, X. Yao, J. Han, and L. Shao, "Learning non-target knowledge for few-shot semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11563–11572.

[31] Y. Liu, N. Liu, X. Yao, and J. Han, "Intermediate prototype mining transformer for few-shot semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 38020–38031.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[35] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 680–688.

[36] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[37] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12408–12417.

[38] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.

[39] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.

[40] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022, Art. no. 5603018.

[41] Z. Huang, Y. Liu, X. Yao, J. Ren, and J. Han, "Uncertainty exploration: Toward explainable SAR target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2023, Art. no. 4202314.

[42] Z. Qin, X. Lu, X. Nie, D. Liu, Y. Yin, and W. Wang, "Coarse-to-fine video instance segmentation with factorized conditional appearance flows," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 5, pp. 1192–1208, May 2023.

[43] Z. Qin, X. Lu, X. Nie, X. Zhen, and Y. Yin, "Learning hierarchical embedding for video instance segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1884–1892.

[44] X. Lu, W. Wang, J. Shen, D. J. Crandall, and L. Van Gool, "Segmenting objects from relational visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7885–7897, Nov. 2022.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[46] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[47] J. Wang, X. Yan, L. Shen, T. Lan, X. Gong, and Z. Li, "Scale-invariant multi-level context aggregation network for weakly supervised building extraction," *Remote Sens.*, vol. 15, no. 5, p. 1432, Mar. 2023.

[48] J. Mitton and R. Murray-Smith, "Rotation equivariant deforestation segmentation and driver classification," 2021, *arXiv:2110.13097*.

[49] L. Wang, C. Zhang, R. Li, C. Duan, X. Meng, and P. M. Atkinson, "Scale-aware neural network for semantic segmentation of multi-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 24, p. 5015, Dec. 2021.

[50] H. Nasrallah, M. Shukor, and A. J. Ghandour, "Sci-Net: Scale-invariant model for buildings segmentation from aerial imagery," *Signal, Image Video Process.*, vol. 17, no. 6, pp. 2999–3007, Sep. 2023.

[51] W. Zhao, J. Na, M. Li, and H. Ding, "Rotation-aware building instance segmentation from high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[52] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018.

[53] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 142–158.

[54] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmenation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6941–6952.

[55] C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8047–8057.

[56] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "DLA-MatchNet for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7844–7853, Sep. 2021.

[57] X. Li, D. Shi, X. Diao, and H. Xu, "SCL-MLNet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2021, Art. no. 5801112.

[58] G. Cheng et al., "SPNet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2021, Art. no. 5608011.

[59] L. Li, X. Yao, X. Wang, D. Hong, G. Cheng, and J. Han, "Robust few-shot aerial image object detection via unbiased proposals filtration," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023, Art. no. 5617011.

[60] 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, 2021, Art. no. 5601614.

[61] G. Cheng et al., "Prototype-CNN for few-shot object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2021, Art. no. 5604610.

[62] C. Lang, B. Tu, G. Cheng, and J. Han, "Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1–7.

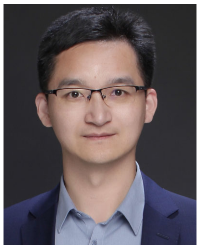
[63] B. Peng et al., "Hierarchical dense correlation distillation for few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23641–23651.

Qinglong Cao (Graduate Student Member, IEEE) received the B.E. and M.S. degrees from Northwestern Polytechnical University, Xi'an, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with Shanghai Jiao Tong University, Shanghai, China.

His research interests include computer vision and remote sensing image processing, especially on few-shot learning and semantic segmentation.

Yuntian Chen (Member, IEEE) received the B.S. degree from Tsinghua University, Beijing, China, in 2015, the B.S. degree from Peking University, Beijing, in 2015, and the Ph.D. degree (Hons.) from Peking University, in 2020.

He is currently an Assistant Professor at the Eastern Institute of Technology, Ningbo, China. His research interests include scientific machine learning, intelligent energy systems, and the integration of domain knowledge and data-driven models.



Chao Ma (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2016.

He was sponsored by the China Scholarship Council as a Visiting Ph.D. Student at the University of California at Merced, Merced, CA, USA, from Fall 2013 to Fall 2015. He was a Research Associate with the School of Computer Science, The University of Adelaide, Adelaide, SA, Australia, from 2016 to 2018. He is currently an Associate Professor at Shanghai Jiao Tong University. His research interests include computer vision and machine learning.

Xiaokang Yang (Fellow, IEEE) received the B.S. degree from Xiamen University, Xiamen, China, in 1994, the M.S. degree from the Chinese Academy of Sciences, Shanghai, China, in 1997, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, in 2000.

From September 2000 to March 2002, he worked as a Research Fellow with the Centre for Signal Processing, Nanyang Technological University, Singapore. From April 2002 to October 2004, he was a Research Scientist at the Institute for Infocomm Research (I2R), Singapore. From August 2007 to July 2008, he visited the Institute for Computer Science, University of Freiburg, Breisgau, Germany, as an Alexander von Humboldt Research Fellow. He is currently a Distinguished Professor at the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. He has published over 200 refereed articles and has filed 60 patents. His research interests include image processing and communication, computer vision, and machine learning.

Dr. Yang received the 2018 Best Paper Award of IEEE Transactions on Multimedia. He is an Associate Editor of IEEE Transactions on Multimedia and a Senior Associate Editor of IEEE Signal Processing Letters.