Uncertainty-Guided Segmentation Network for Geospatial Object Segmentation

Hongyu Jia, Wenwu Yang, Lin Wang, and Haolin Li

Abstract—**Geospatial objects pose significant challenges, including dense distribution, substantial interclass variations, and minimal intraclass variations. These complexities make achieving precise foreground object segmentation in high-resolution remote sensing images highly challenging. Current segmentation approaches often rely on the standard encoder–decoder architecture to extract object-related information, but overlook the inherent uncertainty issues that arise during the process. In this article, we aim to enhance segmentation by introducing an uncertainty-guided decoding mechanism and propose the uncertainty-guided segmentation network (UGSNet). Specifically, building upon the conventional encoder–decoder architecture, we initially employ the pyramid vision transformer to extract multilevel features containing extensive long-range information. We then introduce an uncertainty-guided decoding mechanism, addressing both epistemic and aleatoric uncertainties, to progressively refine segmentation with higher certainty at each level. With this uncertainty-guided decoding mechanism, our UGSNet achieves accurate geospatial object segmentation. To validate the effectiveness of UGSNet, we conduct extensive experiments on the large-scale ISAID dataset, and the results unequivocally demonstrate the superiority of our method over other state-of-the-art segmentation methods.**

Index Terms—**Geospatial object segmentation, remote sensing (RS), semantic segmentation, uncertainty decoding mechanism.**

I. Introduction

GEOSPATIAL object segmentation aims to delineate foreground objects in high-resolution remote sensing (RS) images and assign a semantic label to each pixel. Given its capacity to extract diverse objects of interest, this technique finds extensive applications in various domains, including disaster assessment [1], building extraction [2], and urban monitoring [3].

In the early stages of geospatial object segmentation, a common approach involved combining handcrafted features such as texture [4], shape [5], color [6], etc., with machine learning algorithms like SVM [7], Random Forest [8], KNN [9], K-means [10], etc., to determine the location and classification of ground objects. However, manually designing discriminative features presented significant challenges, and the utilization of machine learning algorithms often resulted in substantial computational overhead and limited generalization performance.

In recent years, the rapid advancement of deep learning technology has triggered a revolution across various domains, including geospatial object segmentation. Deep learning methods, particularly convolutional neural networks (CNNs), have outperformed traditional approaches in terms of accuracy and speed, thanks to their potent feature representation capabilities and robust generalization performance. One prevalent paradigm in this domain is the fully convolutional network (FCN) [11], which divides the segmentation framework into two key components: the encoder and the decoder.

Typically, encoders are directly adapted from existing backbone networks that have been pretrained on large-scale datasets like ImageNet [12], examples being ResNet [13], VGG [14], among others. Consequently, the focus of current segmentation-related research primarily lies in the strategies implemented within the decoders.

Drawing from the multilevel features obtained by the encoder, general decoding strategies aim to aggregate foreground-related information through three key steps. These strategies initially construct a multibranch structure [15], [16] to gather comprehensive contextual information from encoded features. Subsequently, they incorporate attention modules [17], [18], [19] to establish global dependencies, and finally introduce various multilevel fusion techniques [20], [21] to merge the multilevel features and produce precise segmentation results.

In RS images, foreground objects often exhibit dense distribution, significant interclass variations, and minimal intraclass differences, introducing inherent ambiguity for general decoding strategies. In this article, we tackle these challenges by delving into the uncertainty issues present in RS segmentation and introduce a novel uncertainty-guided segmentation network, referred to as UGSNet.

To address these challenges, we begin with the conventional encoder–decoder architecture [11]. We employ the pyramid vision transformer (PVT) [22] to extract multilevel features rich in long-range information. Subsequently, we introduce an uncertainty-guided decoding mechanism, considering both epistemic and aleatoric uncertainty, to refine segmentation progressively. Within this uncertainty-guided decoding mechanism, we introduce two modules, the aleatoric uncertainty-quantify module (AUQM) and the epistemic uncertainty-quantify module (EUQM), to quantify these two types of uncertainty. These quantified uncertainties guide the model in minimizing ambiguity.

As illustrated in Fig. 1, we visualize the two types of uncertainty, demonstrating their complementary nature, which leads to finer segmentation results. With this uncertainty-guided

Manuscript received 8 October 2023; revised 27 December 2023; accepted 24 January 2024. Date of publication 2 February 2024; date of current version 8 March 2024. *(Corresponding author: Wenwu Yang.)*

The authors are with the School of Maritime Economics and Management, Dalian Maritime University, Dalian 116026, China (e-mail: jiahongyu@dlmu.edu.cn; yangwenwu@dlmu.edu.cn; wanglin@dlmu.edu.cn; lhl@dlmu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3361693

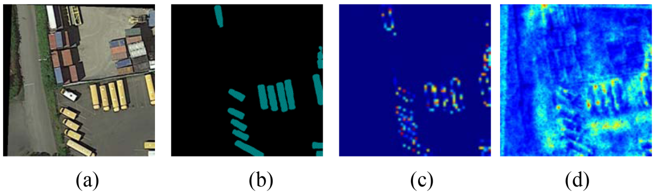

Fig. 1. Visualizations of the prediction of our proposed UGSNet, the epistemic uncertainty measured by AUQM, and the epistemic uncertainty output from EUQM. (a) Image. (b) Prediction. (c) Epistemic. (d) Aleatoric

decoding mechanism, our proposed UGSNet achieves precise geospatial object.

The main contributions of our UGSNet are as follows.

1) We address the uncertain segmentation problems resulted from dense distribution, large interclass variances, and small intraclass variances.
2) We establish an uncertainty-guided decoding mechanism from two perspectives, namely epistemic uncertainty and aleatoric uncertainty.
3) We conduct extensive experiments to prove that our UGSNet is fruitful and can achieve state-of-the-art (SOTA) performance over other segmentation networks.

The rest of this article is organized as follows. In Section II, we introduce the related work, and we give a detailed description of our proposed UGSNet in Section III. The experiments and corresponding ablation study are presented in Sections IV and V, respectively. Finally, Section VI concludes this article.

II. Related Work

In this section, we will review the related work and illustrate the differences between our proposed UGSNet and them.

A. *Early Deep Learning-Based Segmentation*

Since the inception of AlexNet [12] for image classification tasks, deep learning technology has witnessed significant and rapid evolution. Before the advent of the FCN [11] for semantic segmentation, deep learning algorithms often relied on relatively simple CNNs. For instance, in [23], a network comprising only three convolutional layers and two fully connected layers was used to assign semantic categories to all pixels. In [24], fully connected layers were replaced with global average pooling (GAP) to achieve patch-based segmentation. However, these early deep-learning-based segmentation networks imposed substantial computational demands and necessitated fixed input sizes during both training and inference stages.

B. *FCN-Based Segmentation*

Considering the limitations of early deep-learning methods, which lacked flexibility in segmentation tasks, Long et al. [11] introduced one of the pioneering deep learning solutions for semantic image segmentation: the FCN. The FCN comprises a sequence of convolutional layers, allowing it to process images of varying sizes and generate segmentation maps of equivalent dimensions. Notably, the FCN is structured around two fundamental components: the encoder and the decoder, establishing a foundational paradigm in the field of segmentation.

1) Encoder Network: In essence, the features extracted by the encoder play a pivotal role in determining the ultimate decoding outcome. As a downstream task of image classification, researchers traditionally opted for robust backbone networks like ResNet [13], VGG [14], GoogleNet [25], MobileNet [26], and others. These networks had been pretrained on large-scale classification datasets like ImageNet [27], with the final fully connected layers removed to obtain the encoded features. However, it is worth noting that CNNs inherently capture relatively local information due to the nature of convolution operations, which can result in incomplete feature interaction, particularly in RS images. To tackle these challenges, transformer-based backbone networks emerged as a solution. These networks, such as the PVT [22] and Swin Transformer [28], offer an effective alternative. Leveraging the self-attention mechanism in the spatial domain, transformer-based backbone networks excel at establishing long-range dependencies from a global perspective, addressing the limitations of traditional CNNs in capturing broader context information, especially for RS. Recently, some RS-related encoder networks trained on large-scale RS datasets have emerged. Among them, the typical encoders such as RingMo [29] and plain vision transformer [30] have achieved huge success, which can be directly applied to the downstream tasks.

In this article, because PVT can not only model the local continuity information, which is essential for geospatial object segmentation, but also reduce the high computational cost introduced by attention manipulation, we use it as the encoder of our proposed network.

2) Decoding Strategies: Following the processing of input images by the encoder network to obtain multilevel features, various decoding strategies have been developed. Initially, researchers introduced the feature pyramid network (FPN) [20], which employed up-sampling operations and concatenation to construct the final feature representation. Simultaneously, some scholars ventured into constructing multiscale structures to aggregate contextual information across different scales. For instance, Zhao et al. [16] introduced the pyramid scene parsing network (PSPNet), which aimed to enhance the global context representation of a scene. Zheng et al. [31] invented a foreground-aware relation network to capture the complex relationship between the foreground objects and the background in the RS images. Chen et al. [15] combined dilated convolutions with multiscale architectures to augment the contextual representation of high-level features. Wang et al. [32] adopted a multistage strategy to combine the texture and morphological features of images to guide feature learning, and Zhang et al. [33] utilized the spatial morphological differences to search for the boundary of fine-grained classes. Subsequently, researchers recognized the need to capture more global information, as dilated convolutions and multibranch approaches were still biased toward local context. In response, Fu et al. [17] designed a dual attention mechanism that models global dependencies through matrix operations in both spatial and channel dimensions. This approach inspired the development of numerous attention modules aimed

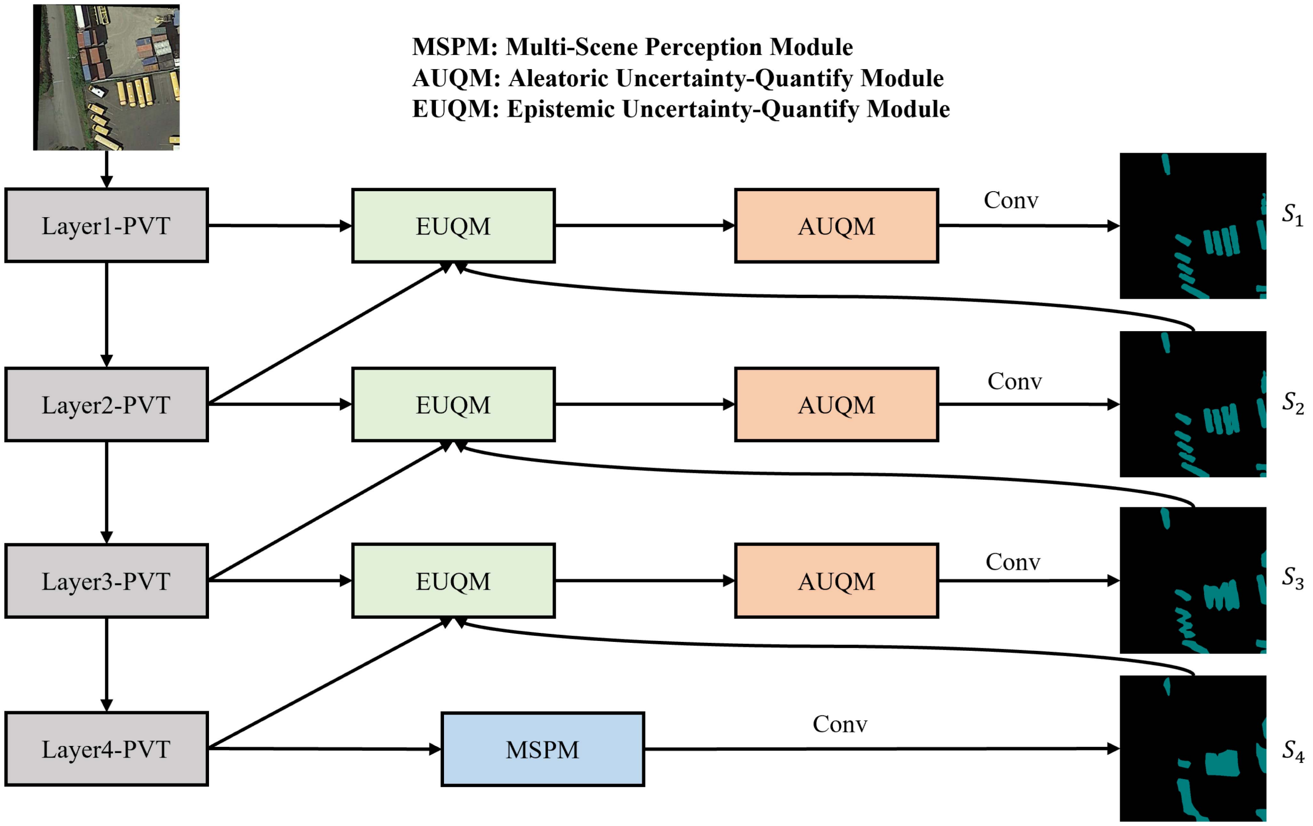
MSPM: Multi-Scene Perception Module
AUQM: Aleatoric Uncertainty-Quantify Module
EUQM: Epistemic Uncertainty-Quantify Module
Layer1-PVT
Layer2-PVT
Layer3-PVT
Layer4-PVT
EUQM
EUQM
EUQM
AUQM
AUQM
AUQM
MSPM
Conv
Conv
Conv
Conv
S_1
S_2
S_3
S_4

Fig. 2. Structure of our proposed UGSNet.

at capturing global information. For instance, Wang et al. [19] introduced the Non-local block, while He et al. [18] innovated the attention mechanism, using coarse segmentation results to guide the enhancement of higher level features. Furthermore, researchers explored coarse-to-fine decoding strategies, where an initial coarse segmentation result guided the refinement process to produce a more detailed segmentation. RefineNet [34] adopted this strategy to refine the details of the coarse segmentation, and DAD [21] mimicked the human eye's focusing process to achieve accurate segmentation from coarse to fine.

C. Uncertainty Strategy

The concept of uncertainty was initially introduced to aid deep learning model decision making, as outlined in [35], where it distinguishes two types of uncertainty: aleatoric and epistemic. Building upon this concept, some researchers have integrated uncertainty to enhance segmentation accuracy. For example, Czolbe et al. [36] introduced a probabilistic segmentation network to estimate the well-defined binary segmentation. DeVries et al. [37] utilized the maximum Softmax probability to estimate the uncertainty. However, previous approaches for estimating uncertainty often relied on Bayesian deep learning networks [35] or other tools [38], [39], which can be complex and challenging to implement.

In contrast to conventional decoding strategies, our approach is inspired by [35] and builds upon the concept introduced in [18]. He et al. [18] proposed a simple uncertainty rank algorithm to measure the uncertainty level of both the buildings and the background, respectively. However, such an uncertainty strategy was just designed for building extraction, which only modeled the epistemic uncertainty and cannot be applied to multiclass segmentation tasks. Different from it, we innovatively integrate an AUQM and an EUQM to address the uncertainties arising from both aleatoric and epistemic sources, as discussed in [35].

III. Methodology

A. Overview

As illustrated in Fig. 2, our innovative UGSNet utilizes the pretrained transformer-based backbone PVT-V2-B2 [22] as the encoder to obtain four levels of encoded features ($P_i, \{i = 1, 2, 3, 4\}$). These four-level encoded features possess varying resolutions and exhibit distinct semantic information. To maximize the utility of these features, we introduce a multiscale atrous spatial pyramid pooling (ASPP) mechanism, ensuring an ample receptive field for the highest level features, P_4. Subsequently, we employ a convolution operation to directly generate a segmentation result, S_4. Nevertheless, due to the lower resolution of the highest level features, P_4, S_4 represents a coarse segmentation lacking intricate details. Diverging from prior methods that sequentially fuse adjacent features without addressing the associated uncertainty generated during feature interactions, we introduce two pivotal modules: the AUQM and the EUQM. In the following sections, we will provide an

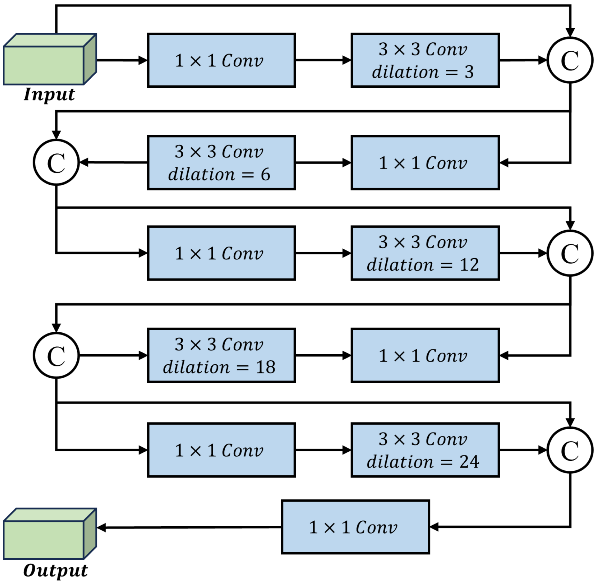

Fig. 3. Structure of the MSPM.

in-depth explanation of our proposed multiscale ASPP, AUQM, and EUQM, respectively.

B. Multiscene Perception Module (MSPM)

For RS images, the contained geographical information is exceedingly complex. To distinguish distinctive features of different objects within this intricate context, it is crucial to incorporate multiscale capabilities and a broad receptive field. To address this, we introduce an MSPM to achieve this objective. As previously emphasized, the capture of multiscale information is vital for acquiring detailed data, and dilated convolutions prove effective in enlarging the receptive field to encompass global information. Therefore, as depicted in Fig. 3, our approach initiates with a 1 × 1 convolution to reduce the channel dimensions of the input features, denoted as P_4. Subsequently, we employ a 3 × 3 dilated convolution (with a dilation rate of 3) to further enhance these features. Simultaneously, to retain the original information, we utilize a concatenation operation to merge these outcomes, resulting in the formation of P_4^3.

$$P_4^3 = \text{Concat}\left(P_4, \text{Conv}_{3\times3}^{d=3}(\text{Conv}_{1\times1(P_4)})\right) \tag{1}$$

where d represents the dilation rate of the 3 × 3 dilated convolution.

Furthermore, we follow such a process, adjust the dilation rates to 6, 12, 18, and 24, and get the corresponding features P_4^6, P_4^{12}, P_4^{18}, and P_4^{24} as follows:

$$\begin{aligned}
P_4^6 &= \text{Concat}\left(P_4^3, \text{Conv}_{3\times3}^{d=6}(\text{Conv}_{1\times1(P_4^3)})\right)\\
P_4^{12} &= \text{Concat}\left(P_4^6, \text{Conv}_{3\times3}^{d=12}(\text{Conv}_{1\times1(P_4^6)})\right)\\
P_4^{18} &= \text{Concat}\left(P_4^{12}, \text{Conv}_{3\times3}^{d=18}(\text{Conv}_{1\times1(P_4^{12})})\right)\\
P_4^{24} &= \text{Concat}\left(P_4^{18}, \text{Conv}_{3\times3}^{d=24}(\text{Conv}_{1\times1(P_4^{18})})\right).
\end{aligned} \tag{2}$$

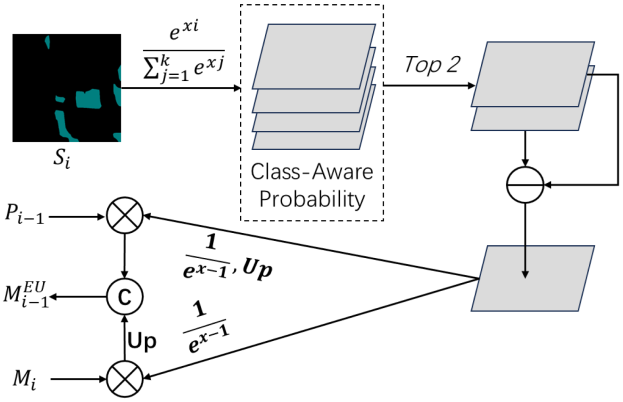

Fig. 4. Structure of the EUQM.

Finally, we use a 1 × 1 convolution to recover the number of channels of P_4.

C. Epistemic Uncertainty-Quantify Module (EUQM)

In RS images, the objects are always presented in various scales and distinct appearances, which causes high degree of the uncertainty. Inspired by [18], we attributes epistemic uncertainty to the insufficient attention given to challenging-to-segment samples. Specifically, due to the extremely complex information of ground objects in RS images, it is extremely difficult to directly model the epistemic uncertainty, and we try to use the coarse segmentation to measure the epistemic uncertainty and achieve the quantification of degree of the epistemic uncertainty. Building upon these considerations, we introduce the EUQM.

As illustrated in Fig. 4, we employ the coarse segmentation result $S_4 \in \mathbb{R}^{C\times H\times W}$ generated by M_4 to quantify epistemic uncertainty, where C represents the total number of classes within the dataset. Initially, we apply the Softmax function to obtain the class-specific probabilities ($\text{CAP}_4 \in \mathbb{R}^{C\times H\times W}$) for all potential categories

$$\text{CAP}_4 = \text{Softmax}(S_4). \tag{3}$$

Subsequently, we introduce the top$k$ ($k = 2$) algorithm to calculate the two most likely classes and get the corresponding two probability maps $\text{Prob}_4^{\text{1st}}$ and $\text{Prob}_4^{\text{2nd}}$ as follows:

$$\text{Prob}_4^{\text{1st}}, \text{Prob}_4^{\text{2nd}} = \text{Top2}(\text{CAP}_4). \tag{4}$$

It is noteworthy that the difference in probability between $\text{Prob}_4^{\text{1st}}$ and $\text{Prob}_4^{\text{2nd}}$ signifies the extent of epistemic uncertainty. When this gap is not distinctly apparent, it suggests that the model struggles to differentiate between the two categories, indicating a high level of uncertainty. Consequently, it becomes essential to pinpoint uncertain samples and guide the model's focus toward these ambiguous pixels. In more detail, we compute the difference by subtracting $\text{Prob}_4^{\text{2nd}}$ from $\text{Prob}_4^{\text{1st}}$, and then, apply the function ψ to normalize the resulting subtraction, denoted as EU_4, into the range from 0 to 1, yielding $\text{EU}_4^{\text{norm}}$,

$$\begin{aligned}
\text{EU}_4 &= \text{Prob}_4^{\text{1st}} - \text{Prob}_4^{\text{2nd}}\\
\text{EU}_4^{\text{norm}} &= \frac{1}{e^{\text{EU}_4} - 1}.
\end{aligned} \tag{5}$$

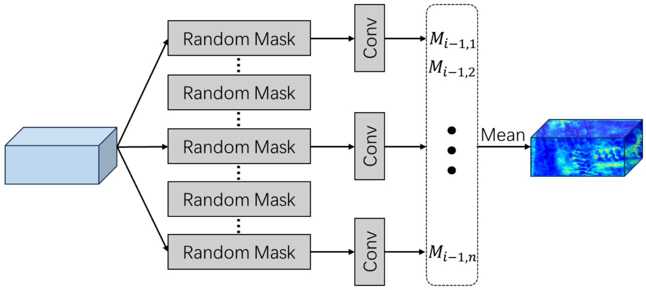

Fig. 5. Structure of the AUQM.

Following the quantification of epistemic uncertainty, we proceed to utilize the quantified uncertainty map, denoted as $\mathrm{EU}_4^{\mathrm{norm}}$, to emphasize uncertain pixels in the neighboring encoded layers. Specifically, we treat the uncertainty map as weights allocated to the encoded features. Consequently, we directly multiply the input features M_4 and P_3 by $\mathrm{EU}_4^{\mathrm{norm}}$ and the upsampled $\mathrm{EU}_4^{\mathrm{norm}}$, respectively. Finally, we concatenate these modified features together, resulting in the final features denoted as M_3^{EU} as follows:

$$M_3^{\mathrm{EU}} = \mathrm{Concat}(\mathrm{EU}_4^{\mathrm{norm}} \times M_4, \mathrm{Up}(\mathrm{EU}_4^{\mathrm{norm}}(\times P_3) \qquad (6)$$

where Up represents the Upsample operation.

D. Aleatoric Uncertainty-Quantify Module (AUQM)

In RS images, there will be differences in the RS imaging of the same geospatial object at different angles, at different heights, and at different times, which will result in huge uncertainty in feature interaction. However, previous decoding strategies tend to overlook this uncertainty generated during feature interaction, which can, to some extent, impact the final segmentation results. In this study, we realize that the disparity in feature information across different levels contributes to uncertainty, known as aleatoric uncertainty, in geospatial object segmentation. To address this, we introduce the AUQM for the measurement and elimination of aleatoric uncertainty.

Traditionally, prior research has often employed Bayesian neural networks to compute the posterior probability over the weights, denoted as $P(W|D)$, where W represents the learned parameters, and D signifies the corresponding dataset. However, as previously mentioned, this implementation can be quite intricate and challenging to integrate into the training process. Therefore, we opt to introduce the Monte Carlo dropout algorithm [40] to estimate the aleatoric uncertainty arising from feature interaction.

As depicted in Fig. 5, concerning the input features M_3^{EU}, we perform direct sampling of these features for a total of n iterations. During each iteration, we apply random channel-wise masking (with a mask ratio of q) to the sampled features. Following this process, we compute the variance among all the randomly masked sampled features. This variance estimation enables us to model the aleatoric uncertainty, resulting in M_3^{AU}, as follows:

$$M_3^{\mathrm{AU}} = \frac{1}{n}\sum_{t=1}^{n}(M_{i-1,t})^2 - \left(\frac{1}{n}\sum_{t=1}^{n}(M_{i-1,t})^2\right. . \qquad (7)$$

To address this aleatoric uncertainty, we simply choose the mean value among all the randomly masked sampled features, resulting in the creation of a more certain set of features denoted as

$$M_3 = \frac{1}{n}\sum_{t=1}^{n}(M_{i-1,t}). \qquad (8)$$

We employ a progressive strategy to systematically merge the features within the ith level, where i ranges from 2 to 4, ultimately yielding the refined features M_1. These refined features can be directly employed to generate the enhanced segmentation map S_1, as illustrated in Fig. 2.

E. Loss Function

In Fig. 2, we observe the presence of four output segmentation maps: S_1, S_2, S_3, and S_4. These maps serve as the foundation for constructing the loss function in conjunction with the ground truth (GT). In this context, we exclusively employ the cross-entropy loss as our loss function, denoted as L. Consequently, the comprehensive loss function $\mathrm{loss}_{\mathrm{all}}$ is as follows:

$$\mathrm{loss}_{\mathrm{all}} = \sum_{i=1}^{4} L(S_i, \mathrm{GT}). \qquad (9)$$

IV. Experiment

To evaluate the performance of our UGSNet, we conducted a comprehensive series of experiments. In this section, we will provide a rich set of experimental results along with in-depth analysis.

A. Dataset

The iSAID dataset [41] comprises 2 806 high-resolution RS images sourced from various sensors and platforms, each with varying resolutions. This original dataset encompasses a diverse range of image sizes, spanning from 800 × 800 pixels to 4000 × 13000 pixels. It includes a total of 655 451 instance annotations across 16 categories, including the background class. The official iSAID dataset conveniently provides a predivided training dataset consisting of 1411 images and a validation dataset with 458 images.

Following the precedent set by previous research, we employed the training dataset to train our model and conducted our evaluation on the validation dataset. To facilitate our experiments, we uniformly cropped images from both the training and validation datasets into 512 × 512 pixels, ensuring no overlap between the cropped regions.

B. Implementation Details

During the training phase, we utilized the Adam optimizer with an initial learning rate set to 10^{-4}. In addition, we applied a learning rate decay, reducing it by a factor of 10 every 50 epochs.

TABLE I
PERFORMANCE COMPARISON WITH BASELINES ON BENCHMARK DATASET

Baseline	mIoU	IoU per category (%)														
		Ship	ST	BD	TC	BC	GTF	Bridge	LV	SV	HC	SP	RA	SBF	Plane	Harbor
FCN-8s	41.6	51.7	22.9	26.4	74.8	30.2	27.8	8.1	49.3	37.0	0	30.7	51.9	52.0	62.9	42.0
SPGNet	46.5	53.1	43.3	59.1	74.7	48.5	43.7	11.4	52.8	31.0	4.4	39.4	33.7	59.9	45.3	45.8
Dense ASPP	46.5	53.1	43.3	59.1	74.7	48.5	43.7	11.4	52.8	31.0	4.4	39.4	33.7	59.9	45.3	45.8
NonlocalNet	58.8	63.4	48.0	49.5	86.4	62.7	50.0	35.0	57.7	43.4	31.6	44.9	67.4	71.0	80.0	51.5
Semantic FPN	59.3	63.6	59.4	71.7	86.6	57.7	51.6	33.9	59.1	45.1	0	46.4	68.7	73.5	80.8	51.2
DANet	60.0	63.9	46.2	73.7	85.7	57.9	48.2	33.5	57.9	43.2	36.1	45.7	67.2	69.2	80.4	52.3
RefineNet	60.2	63.8	58.5	72.3	85.2	61.0	52.7	32.6	58.2	42.3	22.9	43.4	65.6	74.4	79.8	51.1
PSPNet	60.2	65.2	52.1	75.7	85.5	61.1	60.1	32.4	58.0	42.9	10.8	46.7	68.6	71.9	79.5	54.2
U-Net	60.4	63.7	52.5	67.1	87.1	57.6	49.5	33.9	59.2	47.8	29.9	42.2	70.2	69.5	82.0	54.6
CCNet	60.4	64.7	52.8	65.0	86.6	61.4	49.8	34.6	57.8	43.3	35.7	44.6	67.7	70.0	80.6	53.0
DNLNet	60.8	63.7	52.2	72.6	86.6	61.7	54.1	34.2	56.8	42.7	36.8	43.4	68.2	71.3	79.9	50.7
GCNet	60.9	64.9	49.8	72.4	85.8	59.3	51.1	34.1	58.3	43.5	34.9	46.7	68.8	72.6	80.8	53.2
OCNet	61.0	65.2	48.3	71.8	87.0	57.2	55.9	31.2	59.5	43.5	34.9	47.9	70.2	72.8	80.9	50.6
EMANet	61.2	65.3	52.8	72.2	86.0	62.8	49.0	34.9	57.6	43.1	38.6	46.0	69.2	69.3	80.7	52.8
Attention U-Net	61.4	65.6	50.2	73.8	87.7	63.1	54.8	33.7	59.8	47.7	19.5	45.8	70.6	74.3	82.9	54.6
Deeplab v3+	61.9	63.7	58.4	75.6	86.5	59.9	58.6	34.9	59.1	43.9	27.9	48.2	68.7	74.5	80.3	51.9
HRNet	62.7	67.2	64.4	78.2	87.6	60.9	57.5	34.8	59.9	47.7	15.9	48.9	68.2	74.5	82.3	57.0
UperNet	63.2	65.9	59.0	75.7	87.1	61.6	58.5	36.1	60.0	45.7	33.6	49.5	70.6	73.5	81.7	54.4
SFNet	63.3	69.2	**68.3**	77.5	87.5	59.4	55.1	29.7	60.3	46.8	29.3	50.8	71.0	72.7	82.9	53.4
FarSeg	63.7	65.3	61.8	**77.7**	86.3	62.0	56.7	36.7	60.5	46.3	35.8	51.2	71.3	72.5	82.0	53.9
UNetFormer	64.4	69.0	63.4	76.8	87.8	60.3	56.4	**39.9**	63.4	52.1	31.2	44.5	71.1	73.9	84.4	56.5
RSSFormer	65.2	69.5	64.0	76.0	88.2	62.1	59.1	37.6	62.9	51.2	36.1	51.3	71.9	73.0	85.0	55.0
MCCANet	66.9	**71.5**	65.9	77.0	89.7	63.4	**59.9**	38.9	63.9	52.6	**43.6**	52.8	73.1	**74.6**	**86.4**	58.7
UGSNet	**67.4**	70.8	63.6	74.0	**92.9**	**70.2**	55.9	39.4	**64.7**	**55.1**	43.1	**57.2**	**74.0**	73.5	84.5	**60.3**

The best score is highlighted in bold and the second score is underlined. The categories are defined as: ship (SHIP), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground field track (GTF), bridge (Bridge), large vehicle (LV), small vehicle (SV), helicopter (HC), swimming pool (SP), roundabout (RA), soccerball field (SBF), plane (Plane), harbor (Harbor).

TABLE II
PERFORMANCE COMPARISON WITH BASELINES ON BENCHMARK DATASET

Baseline	mIoU	IoU per category (%)														
		Ship	ST	BD	TC	BC	GTF	Bridge	LV	SV	HC	SP	RA	SBF	Plane	Harbor
Base	59.5	60.3	54.6	64.9	89.6	60.2	39.5	33.1	64.2	49.3	12.7	51.1	67.8	64.9	82.3	58.3
Base + MSPM	62.0	66.0	58.8	67.1	89.1	62.6	43.8	31.8	61.2	50.0	37.7	52.5	66.5	67.9	81.0	56.5
Base + MSPM + EUQM	62.4	65.2	48.3	71.8	87.0	61.7	53.0	34.2	56.8	42.7	36.8	43.4	70.2	69.5	82.0	54.6
Base + + MSPM + AUQM	64.6	67.6	61.8	70.9	90.0	67.1	50.9	36.1	**65.4**	51.2	38.1	54.0	71.2	70.1	83.3	57.1
Base + EUQM + AUQM + MSPM	**67.4**	**70.8**	**63.6**	**74.0**	**92.9**	**70.2**	**55.9**	**39.4**	64.7	**55.1**	**43.1**	**57.2**	**74.0**	**73.5**	**84.5**	**60.3**

The best score is highlighted in bold.

TABLE III
PERFORMANCE COMPARISON WITH BASELINES ON BENCHMARK DATASET

Baseline	mIoU	IoU per category (%)														
		Ship	ST	BD	TC	BC	GTF	Bridge	LV	SV	HC	SP	RA	SBF	Plane	Harbor
S_1	61.6	60.9	61.5	71.0	88.8	67.0	51.0	34.9	59.4	41.2	35.8	51.1	68.0	70.1	76.7	49.4
S_2	63.7	65.5	61.6	70.9	89.7	67.1	50.9	35.8	63.6	48.1	36.9	52.9	70.4	70.1	81.0	55.0
S_3	65.4	67.3	62.3	70.8	91.2	68.4	52.0	37.9	60.2	53.4	40.4	55.8	73.7	72.1	83.1	58.9
S_4	**67.4**	**70.8**	**63.6**	**74.0**	**92.9**	**70.2**	**55.9**	**39.4**	**64.7**	**55.1**	**43.1**	**57.2**	**74.0**	**73.5**	**84.5**	**60.3**

The best score is highlighted in bold.

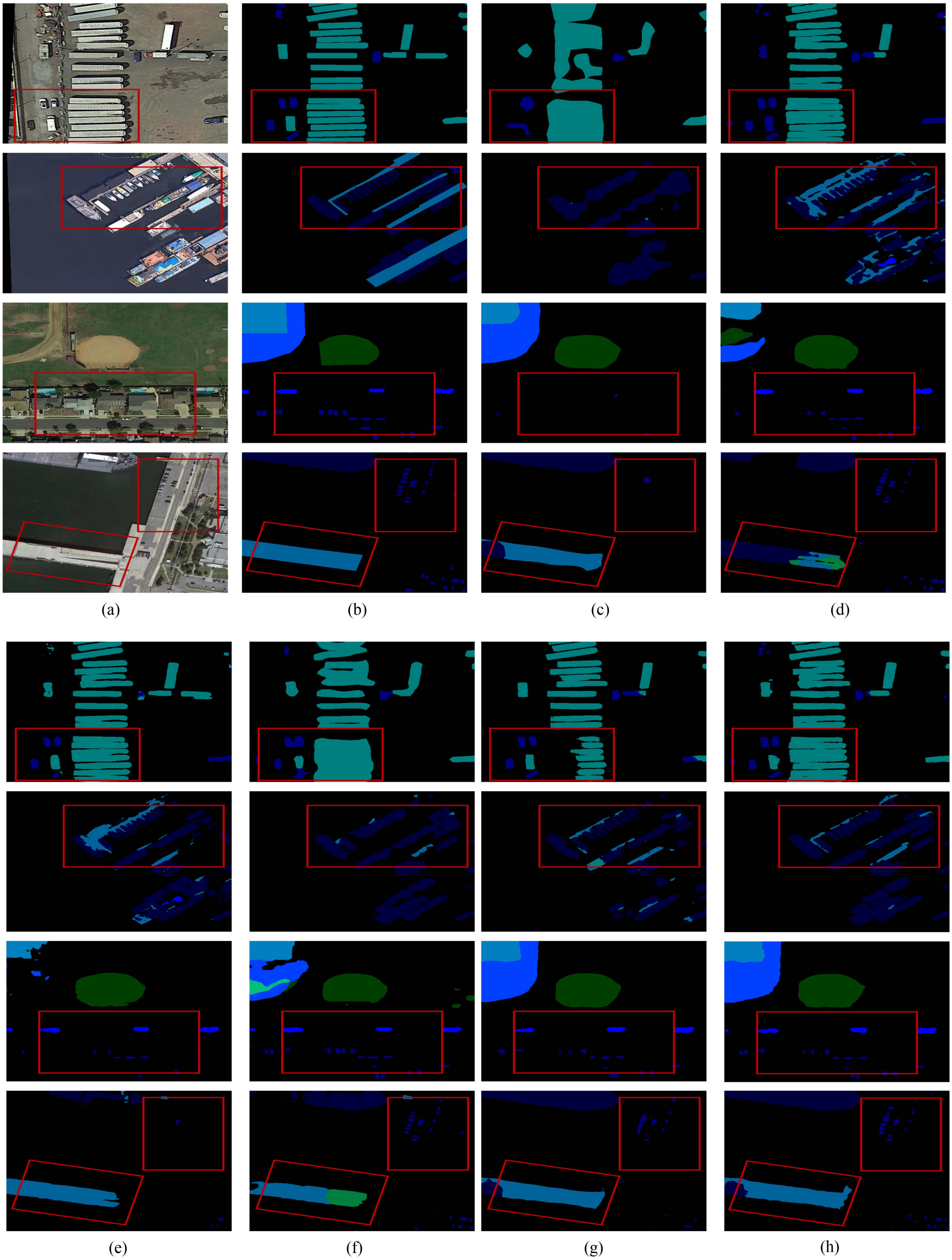

Fig. 6. Visual results of our method and the compared methods. (a) Image. (b) GT. (c) UNet. (d) Deeplab v3+. (e) HRNet. (f) SFNet. (g) Farseg. (h) Ours.

Our chosen batch size was 8, and we standardized the input image sizes to dimensions of 512 × 512 × 3. It is noteworthy that all experiments were executed on the NVIDIA GeForce RTX 3090Ti, which boasts a substantial 24 GB.

C. Evaluation Metrics

In line with established conventions [31], we selected the mean intersection over union (mIoU) as the primary metric for evaluating multiclass object segmentation. In addition, we

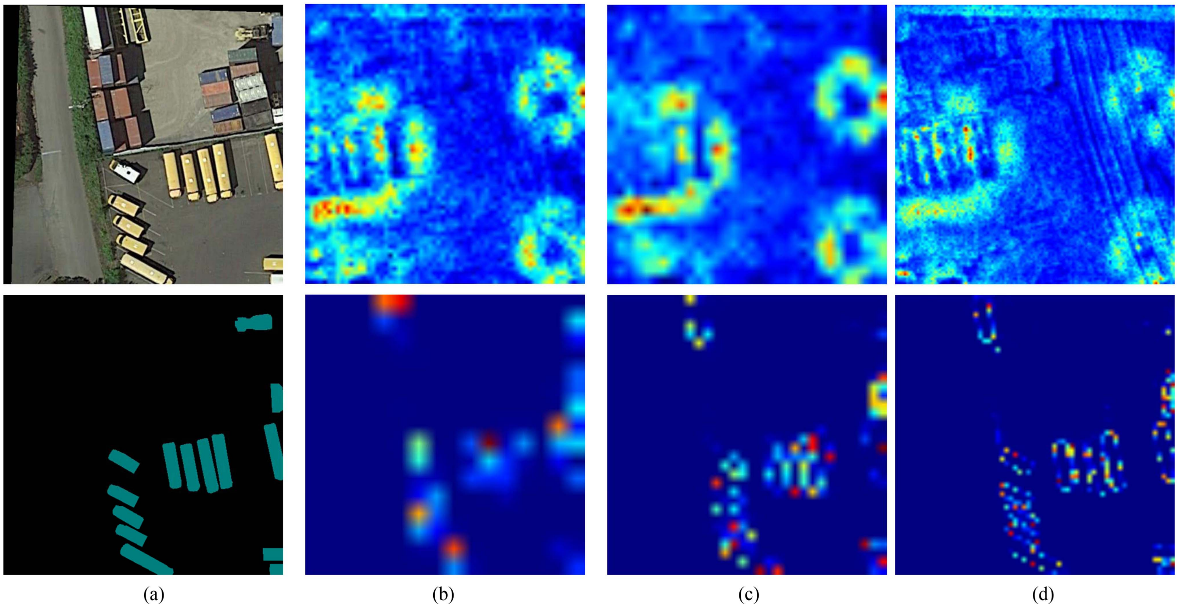

Fig. 7. Visualizations of the epistemic uncertainty and aleatoric uncertainty in the last three decoding layer. The upper represents the aleatoric uncertainty, and the lower represents the epistemic uncertainty. (a) Image/GT. (b) 2nd decoding layer. (c) 3rd decoding layer. (d) 4th decoding layer.

assessed the IoU for each individual class to gain a more detailed understanding of the model's performance.

D. Compared Methods

To verify our proposed USGNet, we select several SOTA semantic segmentation methods to compare with ours: FCN-8s [11], SPGNet [42], Dense Aspp [43], NonLocalNet [19], Semantic FPN [20], DANet [17], RefienNet [34], PSPNet [16], UNet [44], CCNet [45], DNLNet [46], GCNet [47], OC-Net [48], EMANet [49], Attention UNet [50], Deeplab v3+ [15], HRNet [51], UperNet [52], SFNet [53], FarSeg [31], UNetFormer [54], RSSFormer [55], and MCCANet [56].

E. Visual Results

As illustrated in Fig. 6, we present four illustrative examples to facilitate a comprehensive comparison between our pioneering UGSNet and the compared methodologies. In order to show certainty key areas in details, we highlight these areas in red rectangles. In the first scenario, we can discern the dense distribution of storage tanks, leading to less precise segmentations by UNet, SFNet, and FarSeg. Even though Deeplab v3+ and HRNet manage to capture the primary segments, they still lag behind our approach in terms of accuracy. In the second scenario, the challenge lies in accurately delineating small ships. Notably, among all the methods, the segmentation results produced by our UGSNet come closest to the GT, showcasing its prowess in handling intricate details. The third scenario poses a substantial challenge in detecting small vehicles at the image's bottom, where other methods exhibit prominent missed detections. In contrast, our UGSNet outperforms the competition, showcasing its robustness in challenging scenarios. The final scenario features the coexistence of harbors and small vehicles in the image, confounding competing methods in effectively distinguishing between the two. In addition, Deeplab v3+ and SFNet erroneously categorize certain harbor sections. Evidently, our UGSNet consistently delivers superior segmentation results. These visual outcomes unequivocally underscore the remarkable advantages of our strategies for mitigating uncertainty from dual perspectives, solidifying its position as a pioneering solution in the field.

F. Quantitative Comparison With SOTAs

As depicted in Table I, we conducted comprehensive experiments on the iSAID dataset. Notably, our UGSNet has demonstrated superior quantitative performance when compared to all SOTA methods. To delve into the specifics, UGSNet surpasses the leading SOTA method, MCCANet, by an impressive 0.5 percentage points in terms of the mIoU metric. Upon closer examination of individual categories, our proposed UGSNet excels in the segmentation of baseball diamond (BD), tennis court (TC), large vehicle (LV), small vehicle (SV), swimming pool (SP), roundabout (RA), and harbor (Harbor). In addition, it secures the second-best performance in segments such as ship, bridge, and helicopter (HC). This comparative analysis unequivocally underscores the significant effectiveness of our strategy for modeling uncertainty in both types, leading to highly accurate segmentation.

V. Discussion

In this section, we mainly focus on discussing the effectiveness of each module in our proposed USGNet on the iSAID benchmark dataset.

A. Effectiveness of MSPM

As previously discussed, the synergy between multiscale contextual and global information provides complementary cues, motivating us to introduce the MSPM technique for aggregating valuable information within the highest level features. As demonstrated in Table II, it becomes evident that the incorporation of the MSPM leads to a substantial enhancement in segmentation performance. Notably, the term "Base" refers to the UNet-shaped decoder, while "EUQM" and "AUQM" denote the utilization of our proposed EUQM and AUQM, respectively.

B. Uncertainty Strategies

In this article, our aim is to address the challenges posed by uncertainty stemming from dense distribution, significant interclass variations, and subtle intraclass differences, approaching them from two distinct angles: epistemic uncertainty and aleatoric uncertainty. As depicted in Table II, it becomes apparent that harnessing these two forms of uncertainty yields substantial benefits in terms of segmentation effectiveness. To investigate the stability of uncertainty reduction during feature interaction, we present visualizations of epistemic and aleatoric uncertainties in the last three decoding layers, as showcased in Fig. 7. Examining Fig. 7, a clear trend emerges—both epistemic and aleatoric uncertainties progressively diminish, conclusively validating the efficacy of our uncertainty mitigation strategy.

Furthermore, we extend our analysis to include the output of three additional levels and evaluate the results on the test dataset. As presented in Table III, a noticeable trend emerges—segmentation accuracy consistently improves with the gradual elimination of uncertainty. This tangible improvement underscores the significant advantages inherent in our proposed strategy. Meanwhile, it is easy to find that our proposed uncertainty strategy has different improvements for different types of objects. In detail, the use of the EUQM leads to a decrease in the accuracy of segmentation on categories such as ST, TC, BC, LV, SV, SP, and Harbor, but using AUQM basically improves the segmentation accuracy of each category. We think EUQM relies highly on the coarse segmentation result, when the deviation of the coarse segmentation is serious, EUQM cannot correctly measure the yielded epistemic uncertainty. However, the combination of EUQM and AUQM can constrain the coarse segmentation, and hugely improving some hard-to-segment categories, such as Ship, ST, BD, BC, GTF, HC, and SBF.

VI. Conclusion

In this article, we recognize that foreground objects often exhibit characteristics such as dense distribution, significant interclass variability, and subtle intraclass distinctions, resulting in inherent ambiguity for conventional decoding strategies. To address these uncertainty challenges, we introduce the UGSNet. In this approach, we begin by leveraging the PVT to extract multilevel features, rich in long-range information, within the framework of a typical encoder–decoder architecture. Subsequently, we introduce an uncertainty-guided decoding mechanism, considering both epistemic and aleatoric uncertainty, to progressively refine the segmentation at each level. Employing this uncertainty-guided decoding strategy, our UGSNet excels in achieving precise geospatial object segmentation. To validate the effectiveness of our UGSNet, we conducted extensive experiments on the comprehensive ISAID dataset. The results unequivocally highlight the superiority of our method over other SOTA segmentation techniques.

Although our proposed UGSNet outperforms current SOTA RS segmentation methods, there are still some drawbacks in the segmentation results. We attribute the flawed segmentation to the complex environment presented in the RS images. In detail, an RS image consist of various ground objects, which inevitably results in complex contextual information. Our uncertainty-guided strategy can solve such a problem to some extent, but it still cannot deal with the very high-uncertainty situation. We think an ideal solution is to introduce other-modal image data, which can reflect more comprehensive class-related information. Therefore, in the future, we intend to extend our uncertainty strategies to diverse domains, including cross-modal segmentation, medical diagnosis, change detection, and so on.

References

[1] S. Dotel, A. Shrestha, A. Bhusal, R. Pathak, A. Shakya, and S. P. Panday, "Disaster assessment from satellite imagery by analysing topographical features using deep learning," in *Proc. 2nd Int. Conf. Image Video Signal Process.*, 2020, pp. 86–92.

[2] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 27, 2022, Art. no. 5625711.

[3] N. Audebert et al., "Deep learning for urban remote sensing," in *Proc. Joint Urban Remote Sens. Event*, 2017, pp. 1–4.

[4] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Improved building detection using texture information," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 38, pp. 143–148, 2011.

[5] Y. Song and J. Shan, "Building extraction from high resolution color imagery based on edge flow driven active contour and JSEG," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 37, pp. 185–190, 2008.

[6] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 313–328, Jan. 2013.

[7] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 1998.

[8] S. J. Rigatti, "Random forest," *J. Insurance Med.*, vol. 47, no. 1, pp. 31–39, 2017.

[9] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, 2009, Art. no. 1883.

[10] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 281–288.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

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

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

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[17] J. Fu et al., “Dual attention network for scene segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[18] J. Li, W. He, W. Cao, L. Zhang, and H. Zhang, “UANet: An uncertainty-Aw extraction from remote sensing images,” *IEEE Trans. Geosci.*, vol. 62, 2024, Art. no. 5608513, doi: 10.1109/TGRS.2024.3361211.

[19] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[21] J. Li, W. He, and H. Zhang, “Towards complex backgrounds: A unified difference-aware decoder for binary segmentation,” 2022, *arXiv:2210.15156*.

[22] W. Wang et al., “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[23] S. Saito, T. Yamashita, and Y. Aoki, “Multiple object extraction from aerial imagery with convolutional neural networks,” *Electron. Imag.*, vol. 2016, no. 10, pp. 1–9, 2016.

[24] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, “Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 139–149, 2017.

[25] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[26] A. G. Howard et al., “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, *arXiv:1704.04861*.

[27] 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.*, 2009, pp. 248–255.

[28] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[29] X. Sun et al., “RingMo: A remote sensing foundation model with masked image modeling,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 28, 2022, Art. no. 5612822.

[30] D. Wang et al., “Advancing plain vision transformer toward remote sensing foundation model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 21, 2022, Art. no. 5607315.

[31] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, “Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4096–4105.

[32] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, “Remote sensing scene classification via multi-stage self-guided separation network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 17, 2023, Art. no. 5615312.

[33] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, “Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 3, 2023, Art. no. 5501212.

[34] G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1925–1934.

[35] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5580–5590.

[36] S. Czolbe, K. Arnavaz, O. Krause, and A. Feragen, “Is segmentation uncertainty useful?,” in *Proc. 27th Int. Conf. Inf. Process. Med. Imag.*, Jun. 2021, pp. 715–726.

[37] T. DeVries and G. W. Taylor, “Leveraging uncertainty estimates for predicting segmentation quality,” 2018, *arXiv:1807.00502*.

[38] Y. Fang, H. Zhang, J. Yan, W. Jiang, and Y. Liu, “UDNet: Uncertainty-aware deep network for salient object detection,” *Pattern Recognit.*, vol. 134, 2023, Art. no. 109099.

[39] F. Yang et al., “Uncertainty-guided transformer reasoning for camouflaged object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4146–4155.

[40] F. James, “Monte Carlo theory and practice,” *Rep. Prog. Phys.*, vol. 43, no. 9, 1980, Art. no. 1145.

[41] 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*, 2019, pp. 28–37.

[42] B. Cheng et al., “SPGNet: Semantic prediction guidance for scene parsing,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5218–5228.

[43] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, “DenseASPP for semantic segmentation in street scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.

[44] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Munich, Germany, 2015, pp. 234–241.

[45] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[46] M. Yin et al., “Disentangled non-local neural networks,” in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 191–207.

[47] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Global context networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6881–6895, Jun. 2023.

[48] Y. Yuan, X. Chen, and J. Wang, “Object-contextual representations for semantic segmentation,” in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 173–190.

[49] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, “Expectation-maximization attention networks for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9167–9176.

[50] O. Oktay et al., “Attention U-Net: Learning where to look for the pancreas,” 2018, *arXiv:1804.03999*.

[51] J. Wang et al., “Deep high-resolution representation learning for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[52] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.

[53] X. Li et al., “Semantic flow for fast and accurate scene parsing,” in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 775–793.

[54] L. Wang et al., “UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, 2022.

[55] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, “RSSFormer: Foreground saliency enhancement for remote sensing land-cover segmentation,” *IEEE Trans. Image Process.*, vol. 32, pp. 1052–1064, Jan. 26, 2023.

[56] J. Zheng, A. Shao, Y. Yan, J. Wu, and M. Zhang, “Remote sensing semantic segmentation via boundary supervision aided multi-scale channel-wise cross attention network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 4, 2023, Art. no. 4405814.

Hongyu Jia received the graduated degree from Dalian Maritime University, Dalian, China, in 2008 with a PI engineering.

He is a Professor with the School of Maritime Economics and Management, Dalian Maritime University, Dalian, China. His main research interests include the application of deep learning machine vision to decision-making evaluation.

Wenwu Yang received the master’s degree in industrial engineering and management from the School of Maritime Economics and Management, Dalian Maritime University, Dalian, China.

His main research interests include deep learning machine vision.

Lin Wang received the master’s degree in management science and engineering from the School of Maritime Economics and Management, Dalian Maritime University, Dalian, China.

Her main research interests include deep learning natural language processing.

Haolin Li received the master’s degree in management science and engineering from the School of Maritime Economics and Management, Dalian Maritime University, Dalian, China.

His main research interests include convolutional neural network prediction.