Segmenting Remote Sensing Anomalies at Instance Level via Anomaly Map-Guided Adaptation

Jingtao Li, Yanfei Zhong, *Senior Member, IEEE*, Hengwei Zhao, *Member, IEEE*,
Zhi Gao, *Member, IEEE*, and Xinyu Wang, *Member, IEEE*

***Abstract*— Earth anomalies can locate valuable targets in an unsupervised manner for many defense and surveillance applications. Most models assign a continuous score at the pixel-level, resulting in object-agnostic results with higher false alarms than the instance level results. However, since the anomaly objects contain a variety of categories and have a large intraclass variance, the current state-of-the-art (SOTA) query-based models designed for certain categories perform unsatisfactorily when applied to the anomaly instances. The larger intraclass variance of anomalies makes the learning of general representation more difficult. To bridge this gap, we propose general adaptations guided by the pixel-level anomaly map for any query-based model, which adapts the model from learning certain category representation to learning anomaly-aware representation in different categories. The proposed adaptation first builds a separate branch to output the pixel-level anomaly map, where anomaly information is then extracted to guide the pixel embeddings and queries to focus on a variety of anomaly categories. Especially, the anomaly rank embeddings are devised to make the pixel embeddings aware of the anomaly rank order. The queries are dynamically selected from the anomaly candidates after aligning the anomaly map and pixel embeddings for better locating different anomalies. Finally, the selected queries dot-product the anomaly-aware pixel embeddings to output the anomaly instances. The proposed adaptations are simple, general, and additive, which bring the average improvements of +4.9 box AP and +5.1 mask AP in infrared, synthetic aperture radar (SAR), and hyperspectral modalities.**

***Index Terms*— Anomaly detection, instance segmentation, remote sensing.**

I. Introduction

CLASSICAL remote sensing anomalies are defined as the pixels that deviate from the background spectrally or spatially and are identified in an unsupervised manner [1]. Anomalies always have small sizes and occupy a low proportion, such as man-made targets, natural objects, and other interferers [2]. Monitoring the Earth anomalies using remote sensing technology holds significant meaning for the defense and surveillance applications [3], [4], where different modalities [e.g., infrared, synthetic aperture radar (SAR), hyperspectral (HSI)] encode unique information pertaining to the anomaly target [5].

Over the past decades, the community has been committed to building advanced anomaly segmentation models at the pixel-level, where each pixel is assigned to a continuous anomaly score [6], [7]. Since the real anomalies are difficult to acquire, most models focus on learning a high-quality background distribution with statistical-based [8], [9], representation-based [10], [11] or deep reconstruction-based methods [12], [13]. The anomaly score can then be measured by some distance metric between each pixel and obtained background distribution [7], [9]. Among recent studies, some follow this line and propose a more advanced background learning method with deep unfolding [14] or diffusion [15] techniques, achieving better detection accuracy. Some other studies focus on the transferring problem [5], [16], which converts the learning target from the varying background distribution to consistent relationship learning. The resulting transferring model can infer the unseen images directly with retraining.

Despite the improvement in accuracy and transferability, the detection results are limited in pixel-level and object-agnostic, bringing many noisy pixels [17] and inaccurate location information. With modern remote sensing sensors, the improvement of the spatial resolution brings fine-grained information and also noises [18], making the disadvantages of the pixel-level results more obvious. In contrast, the instance level results give a score for each instance rather than single pixel, producing object-aware, cleaner, and more accurate locations. Extending the scope of remote sensing anomaly segmentation from the pixel-level to the instance level constitutes the central objective of this study.

One natural solution to obtain the instance level anomalies is to train a current instance segmentation model with simulated anomalies. The effectiveness of simulated anomalies has been validated in pixel-level anomaly detection models [5], [16], [19]. However, we found although the query-based instance segmentation achieves state-of-the-art (SOTA) performance in computer vision communities [20], [21], the obtained results were poor, with many failed detections if we use the model directly (shown in Section IV-B). Diving

Manuscript received 14 April 2024; revised 12 June 2024; accepted 26 June 2024. Date of publication 7 August 2024; date of current version 30 August 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3903404 and in part by the National Natural Science Foundation of China under Grant 42192580, Grant 42192583, and Grant 42101327. *(Corresponding author: Yanfei Zhong.)*

Jingtao Li, Yanfei Zhong, and Hengwei Zhao are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing and Hubei Provincial Engineering Research Center of Natural Resources Remote Sensing Monitoring, Wuhan University, Wuhan 430072, China (e-mail: jingtaoli@whu.edu.cn; zhongyanfei@whu.edu.cn; whu_zhaohw@whu.edu.cn).

Zhi Gao and Xinyu Wang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China (e-mail: gaozhinus@whu.edu.cn; wangxinyu@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3439491

into the segmentation principle of query-based models, each instance is segmented by dot-producting the query with pixel embeddings [21]. Designed for the segmentation task with fixed categories, each query is related tightly with some certain category in pixel embeddings [22], [23], [24] while the anomalies may contain a variety of categories (vehicle, plane, ship, etc.)—even some unseen categories [2], [25]. This great difference in segmenting targets makes it difficult for the original query-based model to learn the anomaly representation in a variety of categories.

To tackle this, this article proposes a general and lightweight adaptation to prevent the query-based model from learning certain category anomalies. Given the meta-architecture of the query-based instance segmentation model [20], we leave the heavy backbone, encoder, and decoder unchanged and use the pixel-level anomaly map to guide the pixel embedding and query content to be anomaly-aware. Especially, the pixel-level anomaly detection branch is optimized jointly with the instance level anomalies, where the model is trained with both kinds of supervision signals. Based on the score ranking in pixel-level anomaly map, the anomaly rank embeddings are then further derived, which encodes the anomaly score rank of each pixel into the pixel embedding to make it anomaly-aware. Based on the spatial information of pixel-level anomaly map, the corresponding pixel embeddings with high anomaly scores can be selected out correspondingly, which acts as the content queries for more accurate anomaly localization. Besides, since different remote sensing images may have varying channels, a channel preprocessing strategy is also proposed to extract only three channels while keeping the anomaly information existing by computing the deviation distance explicitly. The above adaptations are simple, additive, and can adapt any query-based model for the better instance level anomaly segmentation. We validate its effectiveness on three modalities, including hyperspectral, infrared, and SAR. With increasing only 2M parameters and around 40G Flops, proposed adaptations can promote the query-based segmentation baselines with stable improvements of +4.9 box AP and +5.1 mask AP.

In brief, the main contributions of this article can be summarized as follows.

1) This work extends the anomaly segmentation from the pixel-level to the instance level for object-centric, cleaner, and counting support detection results. The meta-architecture of query-based models is adopted to output the object masks for the end-to-end advantage.
2) General adaptations are proposed, which extract the anomaly information from pixel-level anomaly map to guide the embedding refinement and query selection. The guidance injects the anomaly-aware representation to prevent the model from learning certain category anomalies.
3) A channel preprocessing strategy is designed to deal with the varying channels of input remote sensing images. The strategy extracts only three channels while keeping the anomaly information existing by computing the deviation distance explicitly.

The rest of this article is organized as follows. Section II introduces the related work in remote sensing anomaly detection and the query-based instance segmentation models. Section III provides a detailed description of the adaptations, including the motivation and the workflow. Section IV gives the experimental results and analysis. Finally, the article is concluded in Section V. The code would be available at https://github.com/Jingtao-Li-CVer/Adaptation_For_Instance_Anomalies.

II. Related Work

A. Earth Anomaly Detection

Earth anomalies can be defined as any object or phenomenon that does not conform to a well-defined notion of the expected pattern [26] in Earth observation data. The detection techniques vary according to the normal pattern definition, which mainly includes the following three categories according to the current studies.

1) Treating a collection of single-temporal images as a normal pattern, the related techniques have one-class classification (OCC) [3], [27], [28] and multiclass classification models [29], [30]. OCC models mostly learn the pattern of normal objects in an unsupervised manner and use some distance metric to compute the anomaly degree [31]. Representative models of OCC include one-class support vector machine (OC-SVM) for algal bloom detection [32], the anomaly segmentation model based on pixel descriptors (ASD) model for agriculture and landslide detection [3], and invasive tree species detection (ITreeDet) framework [28]. Different from OCC models, the multiclass classification models are trained with labeled samples for specific anomaly events [30]. Representative models include a refined DenseNet for fire recognition [30], a new deep multi-instance convolutional neural network (DMCNN) model for disaster classification [33], and a deep feed-forward neural network (DFNN) for drought monitoring [29].
2) Treating a collection of multitemporal images as a normal pattern, a common paradigm is to treat the changing locations between the pre-event and post-event images as the anomaly locations [34], [35]. Representative models include the kernel anomalous change detection model for natural disasters such as floods [36], an object-oriented change detection CNN (CDCNN) model with a fully connected conditional random field (CRF) for efficient landslide detection [37], and a deep object-based semantic change detection framework (ChangeOS) for large-scale building damage assessment [38].
3) Treating the background as a normal pattern, anomaly segmentation in remote sensing involves finding the rare pixels deviating from the background in an unsupervised manner [1], [6]. Both spectral and spatial deviation exist for the hyperspectral anomalies, while only spatial deviation exists for the spatial anomalies [39].

In this study, we focus on the third category of normal definition and review the above framework to make the position of our research clear within the entire field. The detailed review when treating the background as the normal pattern is given in Section II-B.

B. Anomaly Detection in Remote Sensing With the Background as Normal Pattern

Due to the high difficulty in acquiring the real anomaly samples, most methods focus on learning the background distribution first and then use some deviation metric to compute the anomaly score. For each pixel, the greater deviation from the background distribution equals to the higher anomaly score. Detection models learn the background distribution with different methods, which can be roughly divided into the statistical-based [9], [40] and representation-based [41], [42] and reconstruction-based methods [7], [43].

1) Statistical-based methods use parametric or nonparametric distribution forms to describe the background [8], where the likelihood is treated as the anomaly score. The classical Reed-Xiaoli detector (RXD) models the background as a multivariate Gaussian distribution and uses Mahalanobis distance to compute the anomaly score [9]. Inspired by RXD, many improved variants are proposed such as kernel RX-AD [44], the support vector data description (SVDD) [45], and the linear filter-based RX-AD [46]. Recently, Chang [47], [48] categorized these models into the generalized likelihood ratio test (GLRT)-based detectors and deviated them to design signal-to-noise ratio (SNR)-based detectors.
2) Representation-based methods exploit the low rank prior of background and the sparse prior of anomalies to obtain a clean background [49]. The remaining sparse components represent the anomalies. Low-rank and sparse-matrix decomposition (LSDM) [50] and the robust principal component analysis (RPCA) models [51] decompose the given image into a low-rank background and sparse anomaly components. By introducing the background dictionary, the low-rank and sparse representation (LRASR) model [52] extends the RPCA from a single subspace to multiple subspaces. Recently, some researchers combine the priors with the deep model. The deep low-rank (DeepLR) model [1] inserts the low rank prior into the background generating process, combining the data-driven and model-driven learning paradigm. The low-rank representation (LRRNet) model [14] solves the low-rank optimization problem with the deep unfolding technique, achieving better model interpretability.
3) Reconstruction-based methods assume the background can be more easily reconstructed by the detector than the anomalies [16], [53]. The classic collaborative representation detector (CRD) [41] reconstructs each pixel with the surrounding pixels constrained by a linear function. With the deep fully convolutional network, Wang al. [7] proposed the autonomous HSI anomaly detection network (Auto-AD) with an adaptive-weighted loss function to reconstruct the background. Both CRD [41] and Auto-AD [7] treat the reconstruction error as the anomaly score. Following similar ideas, a generative adversarial network (GAN) has also been used to reconstruct the background [54]. Recently, the diffusion model has been introduced for better background suppression [15].

Apart from the above models based on background learning, some recent models gradually pay attention to model transferability [5], [16]. They aim to build a model that can infer the unseen images directly rather than retraining. Li et al. [16] proposed the transferred direct detection (TDD) model to learn the consistent deviation relationship between the anomaly and background. TDD achieves detection transferability in hyperspectral modality. Li et al. [5] later extended it to achieve cross-modality transferability. Another transferring anomaly enhancement transformation network (AETNet) model [55] is trained on randomly masked images to learn the spatial context characteristics of anomalies. The masked images in AETNet act as the simulated anomalies in TDD, and these models have all demonstrated the zero-shot transferring ability.

Although great advances in accuracy and transferability, these methods are limited to pixel-level results, which are object-agnostic and may contain many noisy pixels. Our main difference with the previous models is that we first attempt to segment the instance level rather than pixel-level anomalies. The instance level anomaly map can provide object-aware and more interpretable results.

C. Instance Segmentation Models

Instance segmentation models can be roughly divided into box-based [56], [57], [58] and query-based [20], [21] categories. Most of the box-based models follow the architecture design of Mask R-CNN [56], where the instance features are localized by detected bounding boxes and then processed by the segmentation head to obtain the final mask. Based on the Mask R-CNN, many invariants are further proposed. Hybrid task cascade (HTC) [57] introduces the cascade spirit to interweave the detection and segmentation for a joint multistage processing. Mask Scoring R-CNN [58] focuses on the consistence between the mask score and IoU metric, which regresses the IoU score to make the model aware of the mask quality. QueryInst model [59] leverages the intrinsic one-to-one correspondence in object queries and parallel supervision for better performance. A recent dynamic mask selection method (DynaMask) [60] improves the fixed mask resolution in Mask R-CNN and adaptively assigns masks for each instance. However, the series of Mask R–CNN relies on many hand-crafted priors such as the anchor design or NMS post-processing [21].

To achieve end-to-end optimization, query-based models have gradually emerged and achieved the SOTA performance in many challenging datasets [21]. The detection transformer (DETR) [61] first converts the box-prediction problem in the series of Mask R–CNN to the set-prediction problem, removing the need for hand-crafted priors. Although DETR is limited in object detection task, instance segmentation task is closely correlated with it, and many researchers aim to unify them in a single query-based framework. MaskFormer [62] sets the meta-architecture of such unified models (backbone, pixel decoder, and transformer decoder) and treats all the segmentation tasks as the mask prediction task. To improve the instance segmentation ability of MaskFormer, Mask2Former [20] is

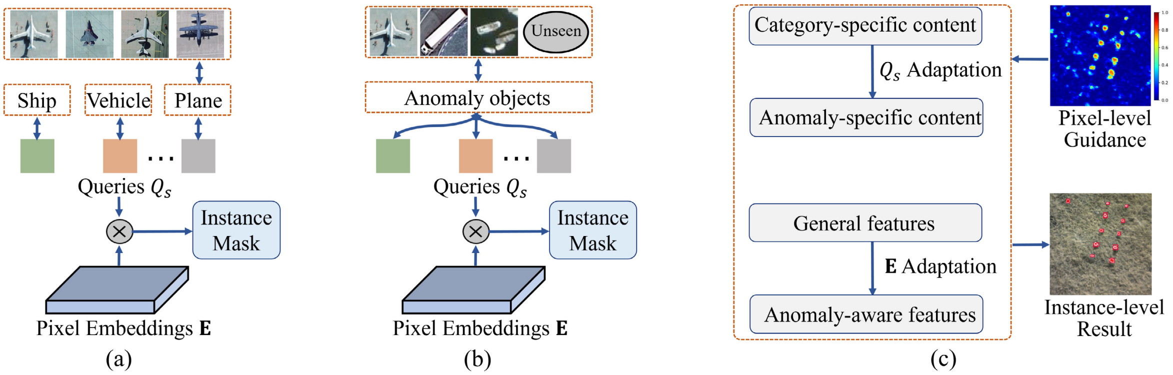


Fig. 1. Motivation of the proposed adaptation. (a) In the original query-based model, each query is correlated with one specific category (e.g., the ship, vehicle, and plane), where the objects in each category have similar spatial or spectral features. Each query dot-products the pixel embeddings to output the instance mask. (b) In the anomaly detection task, the anomaly category is very special, where the anomaly objects have a larger intraclass variance even with unseen objects. Each query is expected to be correlated with all the anomaly categories in embeddings. (c) To bridge this gap, our adaptations are designed to adapt the queries and pixel embeddings to learn anomaly-aware content and features guided by the pixel-level anomaly map. Proposed adaptations are general and can be used in any query-based model.

further proposed, where the core component (masked attention) is designed to extract the localized features. Recently, the Mask DINO (DETR with Improved Denoising Anchor Boxes) model [21] has been proposed, which adds a separate mask prediction branch to DINO and unifies all the segmentation tasks. Mask DINO surpasses many specialized architectures in different tasks.

Due to the end-to-end performance and similar meta-architectures, we choose the query-based models as our baselines for the instance level anomalies. However, the current query-based models are designed for a fixed number of categories [22], [23], [24], where each content query is correlated with a certain category in general pixel embeddings. The anomaly in our task is very special, which may contain a variety of different object categories [25] (vehicle, plane, ship, etc.) and even some unseen categories at the training stage [2]. In other words, the intraclass variance of the anomaly category in our task is much larger than the traditional closed set categories. To tackle this, a general adaptation method is proposed in this article to prevent the model from learning certain category anomalies.

III. General Adaptations for Instance Level Anomalies

In this section, we first clarify the motivation in Section III-A, where general adaptations are expected to inject the anomaly information into any query-based model and prevent the model from learning certain category anomalies (Fig. 1). We extract the anomaly information from pixel-level anomaly maps (Section III-B) to guide the embedding refinement (Section III-C) and the dynamic query selection (Section III-D). Besides, a channel preprocessing strategy is designed to deal with the varying channel numbers of the input image (Section III-E). Section III-F introduces the training process for the adapted model with the simulated anomalies.

A. Motivation: Any Query-Based Model Can Segment Anomaly Instances

The great advantages of instance level anomalies over pixel-level anomalies spawn this research, including object-centric results, fewer false alarm rates, and anomaly counting ability. Observing the SOTA performance of query-based models in instance segmentation community [21], one natural solution is to train some query-based model with simulated anomaly instances. Simulated anomalies have been successfully used to train a pixel-level anomaly detection model [5], [16], [19], and we hope the instance level anomaly segmentation can copy the paradigm directly.

Unfortunately, the obtained results are poor, with low average precision (as in Table II). Since we use the same data and loss function as pixel-level model training, the problem is obviously caused by the architecture difference. Different from the basic U-shape architecture of pixel-level segmenting models [1], [7], the meta-architecture of query-based model consists of the backbone, pixel decoder, and transformer encoder components [62]. Each instance mask is segmented by dot-producting the query from the transformer decoder with the pixel embeddings from the backbone and pixel decoder [21] [as shown in Fig. 1(a)]. As empirically proven in prior studies, each query is correlated with a certain category in pixel embeddings such as the plane or the vehicle [23], [24]. However, the remote sensing anomaly may contain a variety of categories and even unseen categories at the training stage [2]. This difference in segmenting targets [Fig. 1(a) and (b)] greatly increases the intraclass invariance and sets a barrier for the direct application.

To tackle this problem, this study chooses to design a general adaptation rather than a totally new architecture. On the one hand, since the meta-architecture of different query-based models is almost the same, the general adaptation can be used for any query-based model. On the other hand, the improvement from the query-based segmentation model can also enhance the better segmentation performance for instance level anomalies.

Our main idea to design the adaptation is to prevent the model from learning certain categories by making the query and pixel embeddings anomaly-aware, guided by the pixel-level anomaly map. As shown in Fig. 1(c), the query adaptation is designed to change it from category-specific to anomaly-specific, without limitation in fixed and certain categories. The pixel embedding adaptation is designed to change it from general features to anomaly-aware features.

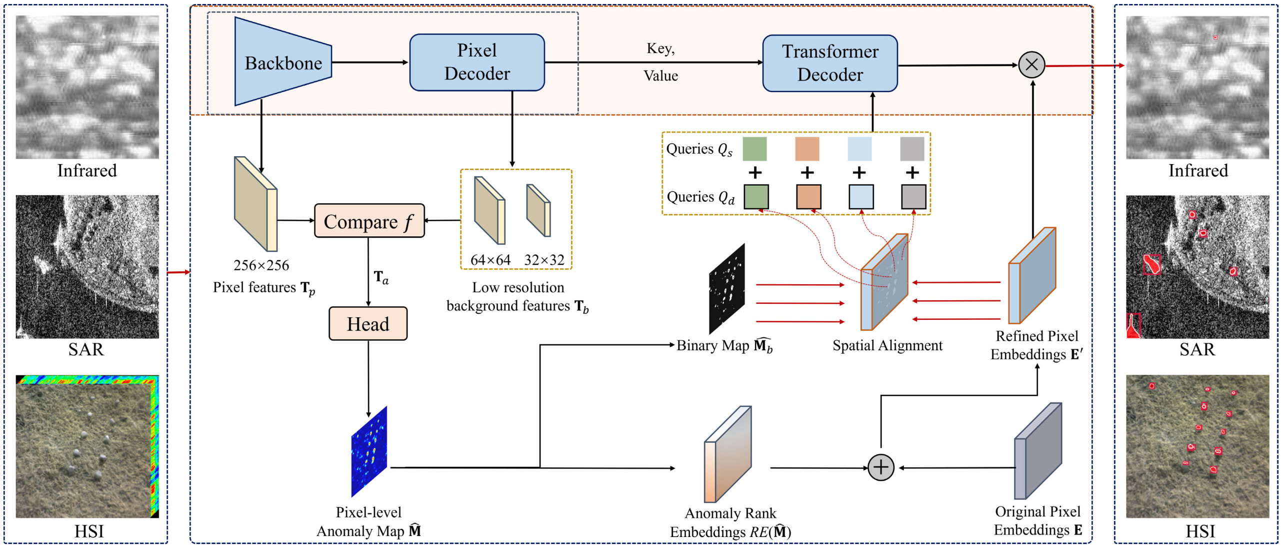

Infrared
SAR
HSI
Backbone
Pixel Decoder
Key,
Value
Transformer Decoder
256×256
Pixel features $\mathbf{T}_p$
Compare f
$\mathbf{T}_a$
Head
64×64 32×32
Low resolution background features $\mathbf{T}_b$
Queries Q_s
Queries Q_a
Binary Map $\widehat{\mathbf{M}}_b$
Spatial Alignment
Refined Pixel Embeddings $\mathbf{E}'$
Pixel-level Anomaly Map $\hat{\mathbf{M}}$
Anomaly Rank Embeddings $RE(\hat{\mathbf{M}})$
Original Pixel Embeddings $\mathbf{E}$
Infrared
SAR
HSI

Fig. 2. Proposed adaptation. To make the pixel embeddings and query anomaly-aware, we first build a sperate branch to output the pixel-level anomaly map $\hat{\mathbf{M}}$, which is further used to refine the embeddings and queries. Specifically, the anomaly rank embeddings are devised from $\hat{\mathbf{M}}$, which encodes the rank order in $\hat{\mathbf{M}}$ and can be added directly to original embeddings $\mathbf{E}$ without changing the feature size. For the queries, they are dynamically selected from refined embeddings $\mathbf{E}'$ according to the spatial location of candidate anomalies, which are obtained after converting the $\hat{\mathbf{M}}$ into a binary map and aligning with $\mathbf{E}'$. The selected queries have a high correlation with the anomaly objects, and the process is to randomly select a variety of anomaly categories. We leave the original architecture of any query-based model unchanged (in pink area) to make the adaptation general.

The two adaptations work together for better anomaly segmentation containing unlimited categories. Fig. 2 shows the overall workflow for any query-based model with our adaptations.

B. Pixel-Level Anomaly Map Guidance

To inject the anomaly information into queries and pixel embeddings, we first build a separate branch to output the pixel-level anomaly map and then use the anomaly map to guide the query content and embedding features to be anomaly-aware.

There are three main reasons why we use the guidance from pixel-level anomaly maps.

1) *Faster Convergence:* Compared to the instance segmentation, the semantic segmentation at pixel-level is a relatively easier task [63]. We can obtain the pixel-level anomaly map by simply adding a separate branch generating from the backbone and pixel decoder in query-based model. Without the need to be object-centric and simpler architecture design, the pixel-level anomaly map can converge faster than the instance level anomaly map. Faster convergence speed means the correct guidance from the pixel-level anomaly map can be obtained earlier during the training stage.
2) *Rich Anomaly Candidates:* Although more false alarm rates exist in the pixel-level anomaly map, it scores the anomaly degree for each pixel and most background pixels can be suppressed. A valuable anomaly candidate can be derived from the anomaly score and its corresponding spatial location.
3) *Multitask Joint Optimization:* Despite the difference exists between the pixel-level and instance level results, both aim to recognize the common deviation relationship between the anomaly and the background. This common point makes them possible to be optimized jointly with the same backbone and pixel decoder. The benefit of the multitask training has been verified in many previous works [64], [65].

We output the pixel-level anomaly map by conducting the "pixel-background" comparison in feature space. The pixel feature cube $\mathbf{T}_p$ has a high spatial resolution $H \times W$ to ensure the feature of each object exists, even for tiny objects. To preserve the low-level feature (e.g., texture) of anomalies, $\mathbf{T}_p$ is extracted from the shallow layers in backbone. In contrast, the background feature $\mathbf{T}_b$ consists of two low-resolution cubes, where the shape of $\mathbf{T}_b^1$ and $\mathbf{T}_b^2$ are 1/4 and 1/8 of the $\mathbf{T}_p$, respectively. With the large downsample ratio, $\mathbf{T}_b$ is expected to contain relatively clean background features without the contamination of the tiny anomalies. We extract $\mathbf{T}_b$ from the top two layers of the pixel decoder, where the deformable transformer [66] is used to exchange the information in different scales.

Given the pixel features $\mathbf{T}_p$ and background features $\mathbf{T}_b$, they are compared with the feature pyramid network (FPN) f [67]. After upsampling $\mathbf{T}_b$ to the same shape with $\mathbf{T}_p$, each pixel can be mapped to its surrounding background feature in $\mathbf{T}_b$ according to the spatial correspondence location. f implicitly judged the comparing feature distance, where the larger feature distance means a higher anomaly degree. The outputting anomaly feature $\mathbf{T}_a$ after pixel-background comparison is finally processed by the convolutional head H to output the pixel-level anomaly map $\hat{\mathbf{M}} \in R^{H\times W}$. H is instantiated by a single 1 × 1 convolutional layer, followed by the Sigmoid activation. The above process can be formulated by the following equation:

$$\hat{\mathbf{M}} = H\big(f\big(\mathbf{T}_p, \mathbf{T}_b\big)\big). \tag{1}$$

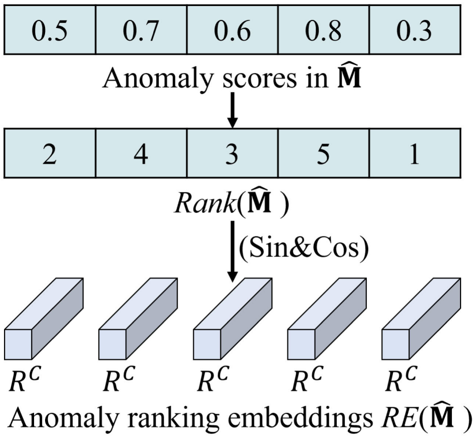

0.5
0.7
0.6
0.8
0.3
Anomaly scores in $\hat{\mathbf{M}}$
2
4
3
5
1
$Rank(\hat{\mathbf{M}})$
(Sin&Cos)
R^C
R^C
R^C
R^C
R^C
Anomaly ranking embeddings $RE(\hat{\mathbf{M}})$

Fig. 3. Designed anomaly rank embeddings. It is designed to inject the anomaly information from the anomaly map $\hat{\mathbf{M}}$ in dimension 2 to the original embeddings $\mathbf{E}$ in dimension 3. Since anomaly detection is a kind of rank task, the continuous anomaly scores are first converted into the discrete rank order in the consistent range $[1, H \times W]$. Then, each rank order is encoded with sine and cosine functions to be the corresponding embedding vector with shape R^C. Sine and cosine functions can ensure the uniqueness and normalization for each rank number. The finally obtained $\mathrm{RE}(\hat{\mathbf{M}})$ can be added into $\mathbf{E}$ directly with the same shape $R^{H\times W\times C}$.

To increase the separability between anomalies and the background, we supervise the pixel-level anomaly maps with simulated anomalies and the differentiable area under the curve (AUC) losses $L_{\mathrm{AUC}}(\hat{\mathbf{M}}, \mathbf{M})$. $\mathbf{M}$ is the simulated ground truth label. Anomalies are generated to simulate the deviation relationship between the background with the data argumentation techniques [5]. Simulated anomalies have been successfully used to help building the detection model with both high detection ability and the transferability in previous works [5], [16]. $L_{\mathrm{AUC}}(\hat{\mathbf{M}}, \mathbf{M})$ focuses on the anomaly-background separability to ensure correct score rank relationship in $\hat{\mathbf{M}}$ (about this point, we refer the readers to [5]). We adopt the similar training strategies to output the pixel-level anomaly end-to-end and can infer the unseen images directly to provide guidance for pixel embeddings and queries.

C. Embedding Refinement Guided by theAnomaly Map

In almost all the query-based models, pixel embeddings act as the general and high-resolution features, providing the low-level information for the queries to dot-product and generate the instance mask. The common manner is to fuse both the shallow features from the backbone and pixel decoder directly [20], [21], which contains little task-related semantic information. Set the original pixel embeddings $\mathbf{E} \in R^{H\times W\times C}$, where the spatial shape is the same as the pixel-level anomaly map $\hat{\mathbf{M}}$. Our target is to inject the anomaly information from $\hat{\mathbf{M}}$ into $\mathbf{E}$.

To interact the $\hat{\mathbf{M}}$ in dimension 2 and $\mathbf{E}$ in dimension 3, we devised the anomaly ranking embeddings to map $\hat{\mathbf{M}}$ from the shape $H \times W$ to shape $H \times W \times C$ without adding training paraments. Fig. 3 shows the detailed workflow to obtain the anomaly ranking embeddings. The original value in $\hat{\mathbf{M}}$ is continuous after the Sigmoid activation, which is intractable to be mapped into a unique vector in R^C. Since anomaly detection is a kind of ranking task [68], we sort the anomaly scores in $\hat{\mathbf{M}}$ and replace the original continuous anomaly score with the discrete rank order. In the ranked result $\mathrm{Rank}(\hat{\mathbf{M}})$, the rank value is in the range $[1, H \times W]$, and a larger value means a higher anomaly degree. With the consistent value range, this transformation can also prevent the influence of the score magnitude in different images. Then, the rank sequence is encoded with sine and cosine functions to ensure the uniqueness and normalized values for each rank number. Given some rank number N, the corresponding ranking embedding vector $\mathrm{RE}(N) \in R^C$ can be encoded as in the following equations:

$$\mathrm{RE}(N)_{2i} = \sin\left(\frac{N}{10\,000^{4i/C}}\right), \quad i \in Z \text{ and } 2i \le C \tag{2}$$

$$\mathrm{RE}(N)_{2i+1} = \cos\left(\frac{N}{10\,000^{4i/C}}\right), \quad i \in Z \text{ and } 2i+1 \le C \tag{3}$$

where $\mathrm{RE}(N)_{2i}$ represents the embedding value in the $2i$ position along the C dimension. Equations (2) and (3) compute the odd and even indexes, respectively. The hyperparameter 10 000 is chosen empirically, following the positional encoding technique in [69] to decide the embedding distance for difference rank order. Applying the operation RE for each rank value in $\mathrm{Rank}(\hat{\mathbf{M}})$ can obtain the final anomaly rank embeddings $\mathrm{RE}(\hat{\mathbf{M}}) \in R^{H\times W\times C}$, which is the same size with $\mathbf{E}$.

With the computed anomaly ranking embeddings $\mathrm{RE}(\hat{\mathbf{M}})$, we can inject its anomaly information into $\mathbf{E}$ through a simple adding operation as in (4). Beneficial from the same shape between $\mathbf{E}$ and $\mathbf{E}'$, the proposed embedding refinement can be used in any query-based model

$$\mathbf{E}' = \mathbf{E} + \mathrm{RE}\big(\hat{\mathbf{M}}\big). \tag{4}$$

D. Query Selection Guided by the Anomaly Map

Query selection is a widely focused improvement from static to dynamic initialization strategies in the query-based model. The static strategy treats the queries as the learned embeddings and is fixed for all the images [61], [70]. To make the query dynamic according to the input and accelerate convergence, dynamic selection strategies have been proposed, where the box prediction head is used to filter out the queries with high classification scores [24], [71].

The current strategies for dynamic query selection have two serve limitations, preventing the selection performance for our instance anomalies. 1) The strategies are limited in the frameworks where the detection and segmentation tasks are unified. For some query-based model without explicitly an object detection branch (e.g., MaskFormer [62] and Mask2Former [20]), the box classification scores do not exist. 2) The classification score differs from the desired anomaly score. Since the query-based model is designed for the fixed number of categories [23], the classification score can represent the model performance well. In contrast, our anomaly may have a variety of categories, and even some unknown category [25], and the anomaly detection task is a kind of ranking problem rather than a simple binary classification.

In this case, the classification score cannot represent model performance correctly and thus limits the selection of queries for segmenting the instance mask.

To solve this, we propose a query-selection strategy guided by the pixel-anomaly map $\hat{\mathbf{M}}$ (as in Fig. 2). Compared to the classification scores, $\hat{\mathbf{M}}$ can better show the score ranking for the special "anomaly" category. Since the query-based model segments each instance mask by dot-producting the query vector with the pixel embeddings [20], the higher similarity between the query and the pixel embedding can output a higher response for the activation layer. Inspired by this observation, our strategy initializes the queries directly from the pixel embeddings, which can naturally meet the demand for high similarity. The $\hat{\mathbf{M}}$ is first converted to a binary map $\hat{\mathbf{M}}_b$ with some preset quantile, which can remove the influence from the absolute anomaly score and focus on the score ranking only. Considering the rarity and low probability properties of the remote sensing anomalies [2], the quantile can be set as a high value (0.9 in our study). The refined pixel embeddings $\mathbf{E}'$ and the $\hat{\mathbf{M}}_b$ have the same spatial shape $H \times W$, and finally a certain number of anomaly queries $\mathbf{Q}_d$ can be selected according to the locations in $\hat{\mathbf{M}}_b$ with value 1. The selection process can be formulated as follows:

$$\mathbf{Q}_d = f\left(\mathbf{E}'\left[\hat{\mathbf{M}}_b == 1\right]\right) \tag{5}$$

where f represents the random selection operation. f makes each selected query to be mapped into uncertain anomaly categories for our task. We do not use the top locations in $\hat{\mathbf{M}}_b$ because they may be gathered in the identity object. To increase the anomaly representation ability further, the classical static queries $\mathbf{Q}_s$ [21] and selected dynamic anomaly queries $\mathbf{Q}_d$ are added and mixed element-wise. $\mathbf{Q}_s$ is set with the same shape with $\mathbf{Q}_d$, and the final constructed anomaly-aware query $\mathbf{Q}$ for the latter dot production is obtained as follows:

$$\mathbf{Q} = \mathbf{Q}_d + \mathbf{Q}_s. \tag{6}$$

Similar to the embedding refinement in Section III-C, the designed anomaly-aware $\mathbf{Q}$ does not change the shape of original queries and can be used in any query-based model.

E. Preprocessing for Varying Remote Sensing Channels

Differing from the natural images, the channels can vary in different remote sensing images. For example, the number of channels can range from dozens to hundreds [72] in the hyperspectral modality and from a few to dozens in the SAR modality [12]. This contradicts the need for a fixed number of channels in the common query-based model. Although some band selection methods (e.g., principal component analysis (PCA) and minimum noise fraction (MNF) can be used to extract exactly three bands, they always aim to keep as much principal information as possible and ignore the rare signals such as the anomalies. Previous anomaly segmentation studies at the pixel level have fixed the model input channel as a large number and interpolated the channels of the test image to reduce the spectral information loss [5], [16]. However, this brings an unnecessary memory burden, especially for spectral images with dozens of channels only. Besides, the changing of the input channels prevents the usage of the large-scale pretrained parameters, which have been proven useful in different downstream tasks but only support the input with three channels.

To decrease the memory burden and utilize the pretrained parameters, we propose a channel processing method, which can extract only three channels while ensuring that the anomaly information is not lost. We aim to compute the anomaly map effectively first and treat each anomaly map as one channel. Statistical-based and representation-based models are both efficient in many hyperspectral datasets. However, statistical-based models (e.g., RXD [9]) are difficult to support parallel computing and have low speed for the large-scale images. Our method is based on the representation-based models [42], where we construct a background dictionary to provide a clean reference for the background categories. Different from the traditional representation models, we compute the anomaly with some deviation metrics rather than with the time-consuming alternating direction method of multiplier (ADMM) optimization. The reason lies in the fact that our preprocessing is used for instance level segmentation; the ultimate pursuit of a single pixel with ADMM is not necessary.

Specifically, the background dictionary is constructed with randomly selected pixels for efficiency, and the anomaly map under one deviation metric constitutes one channel (as shown in Fig. 4). As anomalies always occupy a low ratio in the image (lower than 1% commonly) [1], the anomaly-free property of randomly selected pixels can be guaranteed probabilistically. Set the input image $\mathbf{X}$, and j background pixels denoted by $\mathbf{b}_1$–$\mathbf{b}_j$ are randomly selected from $\mathbf{X}$ to constitute the background dictionary. For a certain deviation metric D_1, the corresponding anomaly map $\mathbf{A}_1$ is obtained by computing the minimum distance between each pixel $\mathbf{x}_k \in \mathbf{X}$ and the $\mathbf{b}_1$–$\mathbf{b}_j$. Formally, the anomaly score for pixel $\mathbf{x}_k$ under D_1 is defined as follows:

$$\min\left\{D_1(\mathbf{x}_k, \mathbf{b}_1), \ldots, D_1\left(\mathbf{x}_k, \mathbf{b}_j\right)\right\}. \tag{7}$$

With three different metrics D_1–D_3, the preprocessed image $\mathbf{X}'$ finally contains only three channels as in (8) after the concatenate operation. Since $\mathbf{A}_1$–$\mathbf{A}_3$ are already the anomaly map, the information of tiny objects can be preserved and also strengthened, even with three channels (as in Fig. 4). With different distance metrics, $\mathbf{A}_1$–$\mathbf{A}_3$ can provide the anomaly information from different aspects

$$\mathbf{X}' = \text{Concat}(\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3). \tag{8}$$

Based on Sections III-B–III-E, we have provided the pseudo code in Algorithm 1 to make the workflow clear.

F. Unsupervised Training Strategy

The anomaly segmentation task is unsupervised, and no real samples can be obtained during the training stage. To train the instance segmentation model, we follow Li et al. [5] to simulate deviating anomaly samples, where the spectral and spatial anomalies are simulated, respectively, with different data augmentation techniques. Considering the task difference

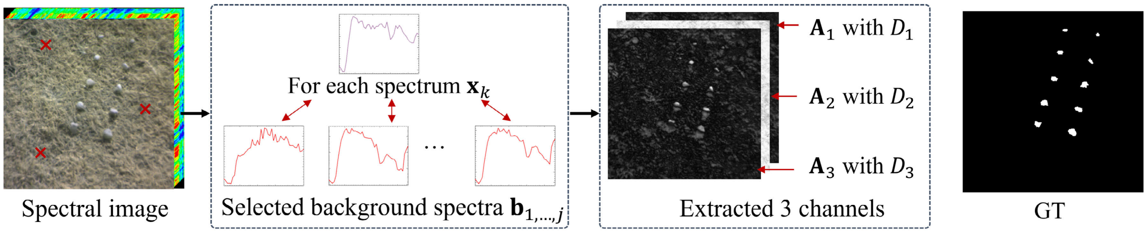

Spectral image
For each spectrum $\mathbf{x}_k$
...
Selected background spectra $\mathbf{b}_{1,\dots,j}$
$\mathbf{A}_1$ with D_1
$\mathbf{A}_2$ with D_2
$\mathbf{A}_3$ with D_3
Extracted 3 channels
GT

Fig. 4. Example to show the designed channel preprocessing strategy. To deal with the varying channels of input images and use the pretrained parameters from natural images, the designed strategy can extract only three channels from any input image while keeping anomaly information exist. We first select j spectra to constitute the background spectra $\mathbf{b}_1 - \mathbf{b}_j$ and use some deviation metric to compute the anomaly map. With metrics $D_1 - D_3$, three corresponding anomaly maps $\mathbf{A}_1 - \mathbf{A}_3$ can be obtained, respectively, and act as the extracted channels. Compared to the GT in the given example, it can be found the anomaly information is strengthened compared to the original image.

Algorithm 1 Workflow of Proposed Adaptation

```
Input: Remote sensing image $\mathbf{X}$
Compute $\mathbf{X}'$ to process the varying channels of $\mathbf{X}$.
Forward propagation with backbone and pixel decoder.
Extract features $\mathbf{T}_p$, $\mathbf{T}_b$ and $\mathbf{E}$ for $\mathbf{X}'$.
Compute $\hat{\mathbf{M}}$ as Eq. (1).
Refine $\mathbf{E}$ and get $\mathbf{E}'$ with rank embeddings as Eq. (4).
Compute $\mathbf{Q}_d$ with aligned $\mathbf{E}'$ and binarized $\hat{\mathbf{M}}$.
Compute $\mathbf{Q}$ based on $\mathbf{Q}_d$ as Eq. (6).
Dot product $\mathbf{E}'$ and processed $\mathbf{Q}$ from decoder.
Output: Segmented masks for each instance
```

between segmentation at the pixel-level with segmentation at the instance level, we make two modifications. 1) The number of simulated anomalies is increased from a few to dozens. The original few and low ratio setting would enlarge the data imbalance problem even with a modern balanced sampler and focal loss optimization. 2) The simulation process of large normal objects is skipped. Differing from the pixel-level segmentation task, the instance segmentation model already has the area-aware ability and would not misclassify the large normal objects. The final simulated training dataset had 5000 images in total, with 16 638 spectral anomalies and 94 365 spatial anomalies.

The simulated anomaly labels have two formats: the standard binary mask to supervise the pixel-level anomaly map and the coco format [73] to supervise the instance level segmentation results. Set the instance segmentation loss L_o. We keep the L_o unchanged as the original query-based model and only add the pixel-level supervising loss $L_{\text{AUC}}(\hat{\mathbf{M}}, \mathbf{M})$ (detailed in Section III-B). The overall loss L for the adapted query-based model is formally given in (9), where the λ controls the balance

$$L = L_o + \lambda L_{\text{AUC}}(\hat{\mathbf{M}}, \mathbf{M}). \tag{9}$$

IV. Experimental Results

A. Experimental Settings

1) Instantiated Models: The proposed adaptations are general and can be used for any query-based model. In this study, we instantiated the adaptations on two representative models: Mask2Former [20] and Mask DINO [21]. Mask DINO is the current SOTA instance segmentation model. The adaptations can be inserted easily, with the original model setting unchanged. We denote the adapted models as Mask2Former$^{\text{Ada}}$ and Mask DINO$^{\text{Ada}}$, respectively.

2) Comparing Methods, Datasets, and Evaluation Metrics: The comparing models cover both the box-based and the query-based models. The former includes Mask R-CNN [56] and its successful variants: Mask Scoring [58], HTC [57], QueryInst [59], and the recent cut-and-learn (CutLER) models [74]. The original query-based model Mask2Former [20] and Mask DINO [21] are also evaluated to show the promotion of proposed adaptations.

After training with simulated anomaly samples (as described in Section III-F), we tested models on three modalities, including HSI, SAR, and infrared. The 82 HSI scenes are collected from the classical Cri dataset and two large-scale UAV-borne WHU-Hi-Park and WHU-Hi-Station datasets [1]. Cri dataset covers the spectral range of 650–1000 nm with 46 spectral bands, and WHU-Hi datasets covers 400–1000 nm with 270 bands. Hyperspectral anomalies include plastic plane, metal object, and so on. The SAR dataset has 100 scenes collected from Gaofen-3 and Sentinel-1 satellites [53], [75], where various ships are considered as anomalies deviating from the sea background. We use the "Single" dataset in [76] to provide the infrared modality with various tiny objects such as cars and drones.

Table I shows the detailed information about the datasets. The three datasets are challenging, with the tiny anomaly size and various scene distributions. Many hyperspectral anomalies only have dozens of pixels and occupy less than 0.1% of the whole image, making the model struggle to learn the discriminating feature [1]. The infrared dataset is a similar situation, and intuitive samples are given in Figs. 5 and 6. Different from most anomaly detection datasets with several certain scenes, the three used datasets have 80–100 scenes with various background distributions. Since the trained model infers the unseen scenes directly, the setting has high demands on the model's robustness.

We evaluate the results with the standard (average precision) AP over varying IoU thresholds (from 0.5 to 0.95), AP$_{25}$ and AP$_{50}$. AP$_{25}$ is used following [77], considering that the tiny objects have low tolerance with the IoU metric. We reported both the box AP$^{\text{B}}$ and mask AP$^{\text{M}}$ evaluation results.

3) Implementation Details: $D_1 - D_3$ were implemented with the cosine distance, Euclidean distance, and the

TABLE I

Detailed Information of Constructed Multimodal Datasets for the ADRS Task

Modality	Source	Spatial resolution	Image size	Scene number	Anomalies
Hyperspectral	Nuance Cri; Nano-Hyperspec	4–8 cm/pixel	400×400; 200×200	82	Plastic plane, metal object, etc. [1]
SAR	Gaofen-3; Sentinel-1	3–10 m/pixel	256×256	100	Various ships [53], [75]
Infrared	\	\	173×98; 407×305	100	Cars, drones, etc. [76]

TABLE II

Quantitative Comparison Results Obtained on the Hyperspectral Dataset. All the Models Used the Proposed Channel Preprocessing Strategy to Deal With the Varying Channels, Through Which Its Effectiveness Is Widely Verified

Model	HSI dataset					
	AP^{B}	$\mathrm{AP}^{\mathrm{B}}_{25}$	$\mathrm{AP}^{\mathrm{B}}_{50}$	AP^{M}	$\mathrm{AP}^{\mathrm{M}}_{25}$	$\mathrm{AP}^{\mathrm{M}}_{50}$
Mask R-CNN	12.7	33.8	10.0	6.5	25.5	3.7
Mask Scoring	15.3	32.5	**24.6**	10.0	33.1	12.3
HTC	14.1	31.6	17.6	10.0	31.8	12.6
QueryInst	14.8	28.4	22.9	11.3	28.1	14.4
CutLER	14.1	32.3	20.8	10.7	28.8	14.4
Mask2Former	10.7	27.6	15.1	12.3	29.5	17.8
$\mathrm{Mask2Former}^{\mathrm{Ada}}$	15.8	36.1	21.2	**17.7**	39.7	**28.2**
vs. origin	+5.1	+8.5	+6.1	+5.4	+10.2	+10.4
Mask DINO	13.4	34.0	23.6	10.7	31.9	17.8
Mask $\mathrm{DINO}^{\mathrm{Ada}}$	**18.7**	**47.5**	24.0	15.7	**45.3**	24.4
vs. origin	+5.3	13.5	+0.4	+5.0	+13.4	+6.6

Manhattan distance, respectively. A larger backbone and longer training schedule may have further benefited the performance further, but this was not our focus. The box-based models were trained with the SGD optimizer (learning rate 0.002), and the query-based models were trained with the AdamW optimizer (learning rate 0.00003 for Mask DINO and 0.0001 for Mask2Former). λ was set to 1.0 and the batch size was set to 1. We use the standard data augmentations in the corresponding MMDetection v3.1.0 toolbox [78] for Mask2Former and the Detectron2 v0.6 toolbox for Mask DINO. The CPU was an Intel[1] Xeon[1] Gold 5218R CPU at 2.10 GHz with 251 GB memory, and the GPU was an NVIDIA GeForce RTX 4090 with 24 GB memory. For accurate reproduction, we would list the detailed information of all used packages in GitHub.

B. Comparison Results

1) Quantitative Comparison: The results are reported in Tables II and III. Note that the proposed channel preprocessing strategy was used in all the comparative models to process the varying channels. As a result of many unseen anomalies at the training stage and tiny object areas, most of the AP and AP_{50} results are only slightly higher than 10. Benefited from the multiround self-training, the pretrained CutLER shows a better performance than the other Mask R-CNN variants, especially in the SAR modality. The original query-based models show an overall better performance than the box-based models, which verifies the advantage of removing the hand-crafted priors. With the proposed adaptations further, $\mathrm{Mask2Former}^{\mathrm{Ada}}$ and Mask $\mathrm{DINO}^{\mathrm{Ada}}$ can outperform the corresponding original versions by around 4–7 points in the AP and AP_{50} metrics, and 10–17 points in the AP_{25} metric. These improvements are stable in both the query-based models, and Mask $\mathrm{DINO}^{\mathrm{Ada}}$ achieves the best overall performance. The average improvements of the two models over the three modalities are +4.9 AP^{B} and +5.1 AP^{M}.

2) Qualitative Comparison: The visualized examples in Fig. 5 shows the effect of adaptations. The original Mask2Former and Mask DINO have many false alarms in the SAR modality and fail to detect the anomalies in the hyperspectral modality. In contrast, the $\mathrm{Mask2Former}^{\mathrm{Ada}}$ and Mask $\mathrm{DINO}^{\mathrm{Ada}}$ show obvious improvements, and all the anomalies can be detected successfully. Since the test anomalies are unseen at the training stage, it is almost impossible to detect the anomalies without false alarms. The adapted models can decrease the false alarm rate to an acceptable level. The visual improvement intuitively validated the proposed adaptations.

C. Model Analysis

1) Ablation of the Adaptations: The proposed adaptation injects the anomaly information into the pixel embeddings and the query selection process. We conducted the related ablation analysis on all the three modalities with the SOTA Mask DINO as the baseline and reported the results in Table IV. Comparing the two adaptations (the second and third rows), it is difficult to analyze which adaptation has a better effect than the other, and there is no setting that is optimal in all the mask metrics. The increasement in AP_{25}^{M} always comes with the accuracy drop in AP_{50}^{M}. Compared to the baseline results (first row), both adaptations can improve the model obviously, and their improvements are additive. With both adaptations, we can obtain the optimal AP^{M} and overall performance in all three datasets.

2) Influence of the Channel Preprocessing Strategies: The proposed channel preprocessing strategy aims to process the anomalies with varying channels. We compare it with some other possible strategies (as listed in Table V). The pseudo-spectral strategy interpolates the channels of all the test images to a large number of 270, as in [5], to achieve the maximum spectral retention. The PCA transformation outputs the first three components as the input image. Uniform sampling selects the three bands at equal intervals along the spectral dimension. The structural similarity (SSIM) [79] is a recent band selection method, and we also used the first three components as the PCA.

The results show that the methods of the band selection (e.g., PCA and SSIM) perform worse than the pseudo-spectral

[1] Registered trademark.

TABLE III

QUANTITATIVE COMPARISON RESULTS OBTAINED ON THE INFRARED AND SAR DATASETS. THE ADAPTED MASK2FORMER$^{\text{Ada}}$ AND MASK DINO$^{\text{Ada}}$ MODELS HAVE SURPASSED THE ORIGINAL VERSIONS BY A LARGE MARGIN AND ACHIEVED THE BEST PERFORMANCE. THE ADAPTED MODELS HAVE THE BEST ABILITY TO DEAL WITH THE UNSEEN MODALITIES AND SCENES AT THE TESTING STAGE

Model	Infrared dataset						SAR dataset						Params	FLOPs
	AP^B	AP^B_{25}	AP^B_{50}	AP^M	AP^M_{25}	AP^M_{50}	AP^B	AP^B_{25}	AP^B_{50}	AP^M	AP^M_{25}	AP^M_{50}		
Mask R-CNN	7.4	22.0	9.4	7.7	22.7	11.4	13.5	38.9	4.4	5.3	22.5	0.8	44M	201G
Mask Scoring	8.8	24.9	12.3	9.6	27.1	14.0	12.6	35.4	4.7	8.8	33.3	0.6	60M	261G
HTC	5.4	15.9	6.4	6.0	17.7	7.8	8.3	20.8	3.4	4.8	17.6	0.4	77M	130G
QueryInst	6.9	21.2	9.4	6.9	21.4	9.3	13.3	32.0	7.5	9.2	29.6	1.6	172M	121G
CutLER	9.6	25.3	14.5	9.2	25.8	13.8	26.3	46.0	40.0	23.7	47.0	37.9	72M	162G
Mask2Former	11.6	27.6	16.2	10.6	27.8	14.1	9.3	27.8	2.9	8.4	30.4	1.3	44M	226G
Mask2Former$^{\text{Ada}}$	**16.8**	**38.8**	**24.7**	**14.9**	39.0	**21.5**	12.4	36.9	4.5	12.5	41.3	2.6	46M	267G
vs. origin	+5.2	+11.2	+8.5	+4.3	+11.2	+7.4	+3.1	+9.1	+1.6	+4.1	+10.9	+1.3	+2M	+41G
Mask DINO	9.6	26.5	15.2	7.7	24.2	10.6	32.5	57.7	50.1	30.7	57.7	53.1	52M	286G
Mask DINO$^{\text{Ada}}$	13.8	36.5	21.2	12.2	**41.7**	14.9	**39.0**	**64.5**	**60.2**	**38.0**	**64.8**	**62.0**	54M	319G
vs. origin	+4.2	+10.0	+6.0	+4.5	+17.5	+4.3	+6.5	+6.8	+10.1	+7.3	+7.1	+8.9	+2M	+33G

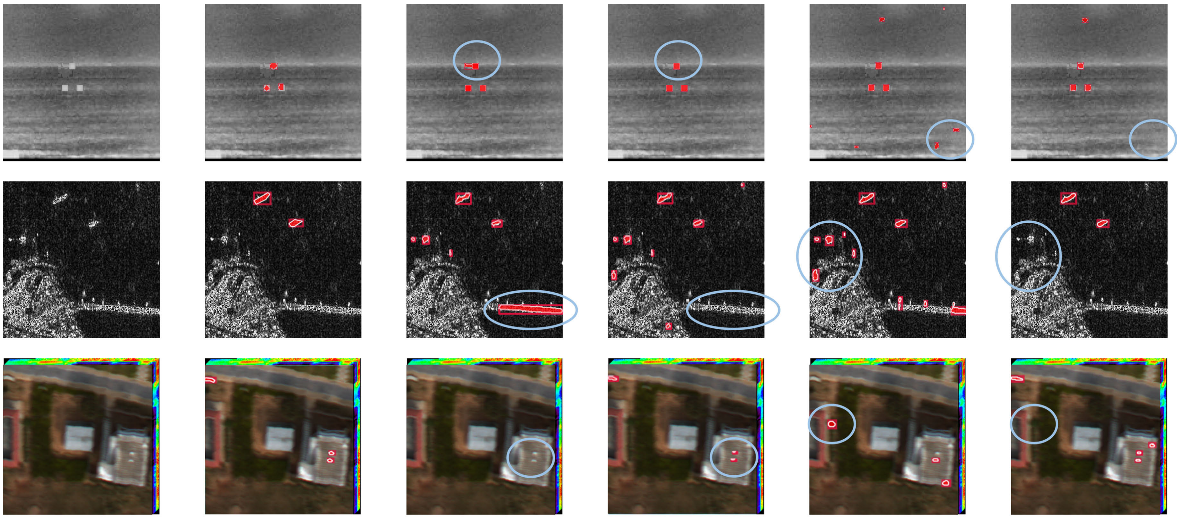


Image GT Mask2Former Mask2Former$^{\text{Ada}}$ Mask DINO Mask DINO$^{\text{Ada}}$

Fig. 5. Exemplified visualized results of the adapted Mask2Former$^{\text{Ada}}$ and Mask DINO$^{\text{Ada}}$ and their original versions. The adapted versions can detect all the anomalies successfully with the lowest false alarms. The models are compared at the same and default score thresholds.

TABLE IV

ABLATION OF THE PROPOSED ADAPTATION ON THE THREE MODALITIES WITH THE MASK DINO$^{\text{Ada}}$ MODEL

Embedding refinement	Query Selection	Infrared dataset			SAR dataset			HSI dataset		
		AP^M	AP^M_{25}	AP^M_{50}	AP^M	AP^M_{25}	AP^M_{50}	AP^M	AP^M_{25}	AP^M_{50}
×	×	7.7	24.2	10.6	30.7	57.7	53.1	10.7	31.9	17.8
×	√	11.4	28.9	**17.1**	32.1	60.0	55.2	13.5	40.2	24.7
√	×	10.0	32.3	12.0	34.8	**65.8**	58.5	12.7	33.3	**29.9**
√	√	**12.2**	**41.7**	14.9	**38.0**	64.8	**62.0**	**15.7**	**45.3**	24.4

strategy and the proposed one. It is easy to understand since their selection criterion is to sacrifice a little information and keep most of it, while the anomalies are exactly rare and to be sacrificed. Compared to the pseudo-spectral strategy, the processed preprocessing strategy can achieve a better performance without keeping the redundant information.

3) Influence of the Selected Background Pixels: The background pixels play an important role in the proposed channel preprocessing strategy, which constitutes the background

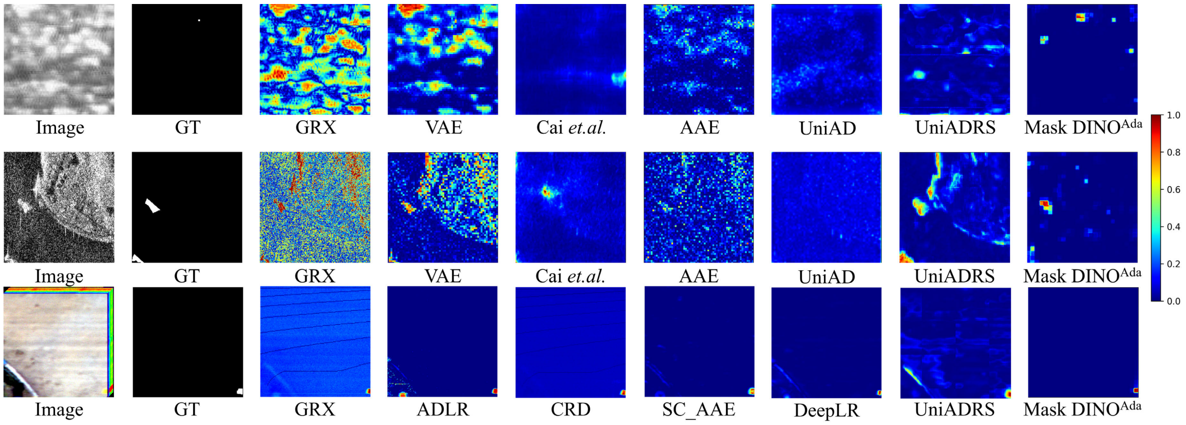


Fig. 6. (Top to bottom) Pixel-level segmentation maps comparison on the infrared, SAR, and HSI datasets. With the joint training of instance level and pixel-level anomaly segmentation, our output pixel-level anomaly maps have the most confidence and the lowest false alarms. High-quality anomaly maps can provide accurate guidance for the proposed adaptation.

TABLE V

Comparison of the Different Channel Preprocessing Strategies on the Hyperspectral Dataset With the Mask DINO$^{\text{Ada}}$ Model

Channel pre-processing	AP^B	AP^B_{25}	AP^B_{50}	AP^M	AP^M_{25}	AP^M_{50}
Pseudo spectra	18.5	39.8	**30.4**	15.3	38.5	**25.2**
PCA	13.5	30.4	19.6	12.8	31.0	14.5
Uniform sampling	12.8	34.8	16.4	9.8	30.9	12.8
SSIM	15.7	42.5	20.2	14.0	41.4	17.2
Ours	**18.7**	**47.5**	24.0	**15.7**	**45.3**	24.4

TABLE VI

Comparison of the Number of the Selected Background Spectra on the Hyperspectral Dataset With the Mask DINO$^{\text{Ada}}$ Model

Number	AP^B	AP^B_{25}	AP^B_{50}	AP^M	AP^M_{25}	AP^M_{50}	FPS
1	9.0	26.1	10.8	7.3	19.5	12.6	**3.15**
3	**18.7**	**47.5**	**24.0**	**15.7**	**45.3**	24.4	2.82
5	16.5	35.5	22.1	14.0	34.1	**24.5**	2.34

TABLE VII

Comparison of the Different Loss Hyperparameter Setting (i.e., λ) on the Three Datasets With the Mask DINO$^{\text{Ada}}$ Model

λ	Infrared dataset		SAR dataset		HSI dataset	
	AP^B	AP^M	AP^B	AP^M	AP^B	AP^M
0.1	11.1	11.2	31.7	30.1	17.1	19.0
0.5	13.5	11.8	32.5	31.5	16.9	12.9
1.0	13.8	12.2	39.0	38.0	18.7	15.7
5.0	13.6	11.8	40.9	39.3	19.4	13.8
10.0	11.6	11.5	35.5	34.3	19.6	12.9

dictionary for computing the anomaly map for each output channel. The number of selected pixels is a hyperparameter that needs to be tuned. We report the related analysis in Table VI with the hyperspectral modality. The selected number influences not only the precision but also the processing speed. The ideal background spectra can cover all the background categories but has nothing to do with the anomalies. Only one background spectrum is far from enough and causes a large precision drop compared to the three and five spectra. A slight precision drop is also observed when changing the number from 3 to 5, which may be caused by the contamination of the noisy pixels. Considering both the precision and speed, three spectra were selected for each hyperspectral image during the training and testing stages.

4) Influence of the Loss Hyperparameter λ: The hyperparameter λ is used to control the balance between the $L_{\text{AUC}}(\hat{\mathbf{M}}, \mathbf{M})$ for the designed anomaly map branch and the L_o for the instance level segmentation task. To set the parameter properly, we have conducted the related sensitivity analysis in Table VII, where we varied λ in different ratios. When the λ is set smaller than 1.0, we found the performance would decrease in positive correlation. This phenomenon is reasonable since the adaptation is designed based on the guidance from pixel-level anomaly map. Too small a λ can result in inaccurate anomaly maps and further damage the model performance. The value settings of 1.0 and 5.0 have a similar performance, implying the range [1.0, 5.0] is acceptable empirically. However, the value is not simply "the larger the better," and 10.0 leads to a severe decline in results, as shown in Table VII. Comparatively, the value range [1.0, 5.0] balances the anomaly map generation and instance segmentation well.

5) Accuracy of the Pixel-Level Anomaly Map: Our work extends the anomaly segmentation from the traditional pixel-level to the instance level. For the proposed adaptations, a branch was first built to output the pixel-level anomaly map to guide the instance level results. The accuracy of the anomaly map directly decides the quality of injected anomaly information for the query and embeddings. For this, we evaluate the pixel-level anomaly maps in three modalities to show their quality and report the results in Tables VIII and IX and Fig. 6. Due to the high spectral resolution and large spectral number, there are many specialized models for hyperspectral modality, and we report it separately in Table VIII. The used anomaly maps are output from the Mask DINO$^{\text{Ada}}$ model. The four used evaluation metrics come from the 3-D receiver operating characteristic (3-D ROC) curves [80], which are positively correlated with the detection performance.

TABLE VIII

Quantitative Comparison Results of the Pixel-Level Segmentation Performance on the Hyperspectral Modality

Method	**AUC(D,F)**	AUCTD	AUCBS	AUCODP
	HSI modality			
GRX [9]	0.9348	1.1304	0.8547	1.1146
ADLR [82]	0.9024	1.0861	0.8293	1.1414
CRD [41]	0.9334	0.9968	0.9109	1.0168
SC_AAE [43]	0.9379	1.0611	0.9596	1.0823
DeepLR [1]	0.9801	1.0914	0.9723	1.0999
UniADRS [5]	0.9859	**1.2608**	0.9530	**1.2353**
Mask DINO$^{\text{Ada}}$	**0.9923**	1.1231	**0.9914**	1.1299

From Tables VIII and IX, it can be observed that our anomaly maps have surpassed many modality-specialized models, where the UniADRS is the current SOTA model in pixel-level anomaly detection. The adapted Mask DINO$^{\text{Ada}}$ has surpassed the UniADRS by four points ($\text{AUC}_{(\text{D,F})}$) in infrared modality and around one point in hyperspectral modality. In SAR modality, Mask DINO$^{\text{Ada}}$ has a balanced performance with UniADRS.

We deduce that the overall promotion of Mask DINO$^{\text{Ada}}$ at the pixel-level results comes from the multitask training. Both pixel-level and instance level results have the same detecting target but only a different result format. As shown in Fig. 6, the visualized maps of Mask DINO$^{\text{Ada}}$ in three modalities have the highest confidence in anomalies and also have low false alarms or noisy pixels. The results are similar to the object-centric instance level anomalies, which shows that the supervision for the instance segmentation has also influenced the pixel-level results to be consistent. In reverse, the more accurate pixel-level results can provide better guidance for the anomaly instance segmentation.

6) Analysis of the Time and Space Complexity: The proposed adaptation can adapt any query-based model for the anomaly segmentation task and increase the performance obviously (as in Table III). However, the time and space complexity are also increased, and it is necessary to conduct related analysis. Specially, the adaptation refers to three procedures, including the anomaly map generation, embedding refinement, and query selection. The process of anomaly map generation has multiscale convolutional modules, which are all positively related to the image size. The input feature cube $\mathbf{T}_p$ has a size of $H \times W \times C$, and the processed multiscale features (i.e., $\mathbf{T}_b^1$ and $\mathbf{T}_b^2$) have sizes of $H/4 \times W/4 \times C$ and $H/8 \times W/8 \times C$, respectively, which are all proportional to the $H \times W$. Embedding refinement includes the ranking embedding generation and a dot addition operation, which are linearly proportional to image size $H \times W$. The situation for query selection is similar and mainly contains a dot addition operation. Therefore, the theoretical time and space complexity of the proposed adaptation are both $O(H \times W)$.

The increased computation cost was comparatively tested on the Mask2Former and Mask DINO models (as in Table III), where the param number reflects the space complexity and the FLOPs reflect the time complexity. The proposed adaptation would result in an increase of 2M params and around 30G FLOPs, which are lightweight compared to the original versions.

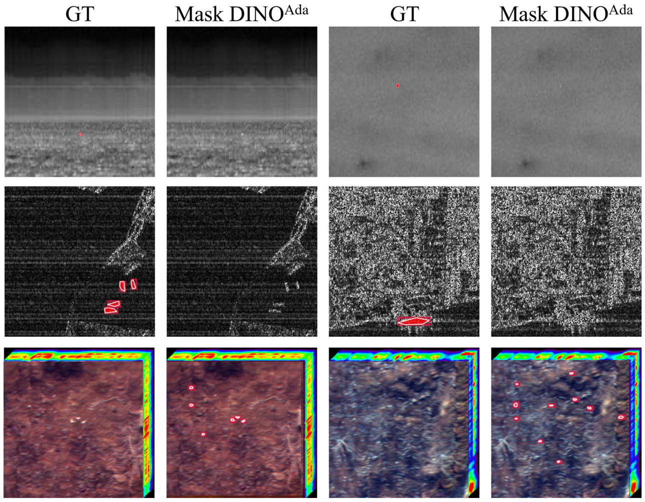


Fig. 7. (Top to bottom) Some failure examples on infrared, SAR, and HSI datasets, respectively.

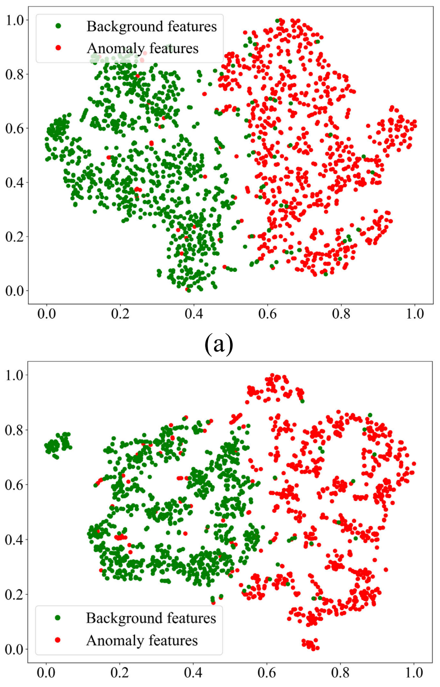
7) Analysis of the Failure Samples: To make the limitation of the proposed adaptation clear, we have also reported some failure examples of three modalities in Fig. 7. For the infrared modality, the anomalies in both examples have low contrast, and there are many normal objects with the same level of signal response, making the deviation degree not obvious and increasing the detection ability. For the SAR modality, the anomaly objects in the left example have extremely incomplete signals, where many internal signals of objects are missing due to the imaging characteristics of SAR. The right sample has a ship completely connected with the background, and it is difficult to distinguish even with human eyes. For the HSI modality, we show two special cases where the ground truth has no anomalies, but the results have detected many objects. It may not be appropriate to classify it as the failure example since the real anomalies are always varied and unknown, but ground truth only labels the known anomalies. Overall, the anomalies with low contrast or incomplete abjectness would make the detection more difficult. The challenging scenes also imply the points for further improvement.

8) Visualization of the Intermediate Features: To ensure the model has learned the correct features, we use the t-distributed stochastic neighbor embedding (t-SNE) [81] visualization technique to increase the model interpretability. Since the proposed adaptation is guided by the pixel-level anomaly map and the final instance mask is segmented with the pixel embeddings, the corresponding two key features are visualized in Fig. 8, including anomaly feature $\mathbf{T}_a$ and the refined pixel embeddings $\mathbf{E}'$. We select 2000 samples randomly for each feature of the trained Mask DINO$^{\text{Ada}}$ model and use t-SNE to downsample the features into the 2-D visualization map. Obviously, the background and anomaly features are separated effectively with a few indistinguishable cases. The linear separability proves that the adapted model has the ability to transform the data from the original non-linear space to a linearly separable space rather than learning noisy or meaningless features.

TABLE IX

Quantitative Comparison Results of the Pixel-Level Segmentation Performance on Infrared and SAR Modalities. The Adapted Mask DINOAda Model Can Achieve the Overall Best Performance, Even Though It Is Designed for the Instance Level Anomaly Segmentation

Method	**AUC$_{(D,F)}$**	AUC$_{TD}$	AUC$_{BS}$	AUC$_{ODP}$	**AUC$_{(D,F)}$**	AUC$_{TD}$	AUC$_{BS}$	AUC$_{ODP}$
	Infrared modality				SAR modality			
GRX [9]	0.6814	1.0899	0.4543	0.8629	0.8938	1.5250	0.7931	1.4243
CAE [83]	0.8291	0.9297	0.8180	0.9187	0.8281	0.9118	0.8210	0.9047
VAE [84]	0.7301	**1.2339**	0.4902	0.9941	0.8816	1.3315	0.8495	1.2995
Cai *et al* [85]	0.8853	1.2242	0.8415	1.1805	0.8610	1.0612	0.8347	1.0349
AAE [12]	0.7557	1.0686	0.6598	0.9727	0.8831	0.9699	0.8757	0.9626
UniAD [86]	0.8348	0.9145	0.8054	0.8850	0.9102	1.0678	0.8329	0.9905
UniADRS [5]	0.9437	0.9820	0.9394	0.9778	**0.9595**	0.9959	**0.9549**	0.9913
Mask DINO^{Ada}	**0.9806**	1.2159	**0.9761**	**1.2114**	0.9563	**1.5594**	0.9179	**1.5647**



(a)

(b)

Fig. 8. T-SNE visualization for (a) anomaly feature $\mathbf{T}_a$ and (b) refined pixel embeddings $\mathbf{E}'$.

9) Convergence on the Training and Validation Datasets: To show the model convergence, we have visualized the loss curve on the training dataset (red curve) and also the changing curve (green curve) of AP^{M} metric on the validation dataset in Fig. 9. The loss value is the sum of anomaly segmentation loss, multilevel box regression, and mask segmentation losses. The model loss drops rapidly during the first 100 iterations and then declines gradually between 100 and 1000 iterations, which exhibits an oscillating manner. Subsequent to the 850th iteration, the loss function exhibits convergence, maintaining a value proximate to 40. In parallel, we observe a monotonic rise in the AP^{M} metric of validation dataset, which contains 1100 unseen simulated images and anomalies. Similarly, the validation set accuracy also becomes stable around 850 iterations. Both the curves of training loss and validation accuracy show a healthy and convergent learning process.

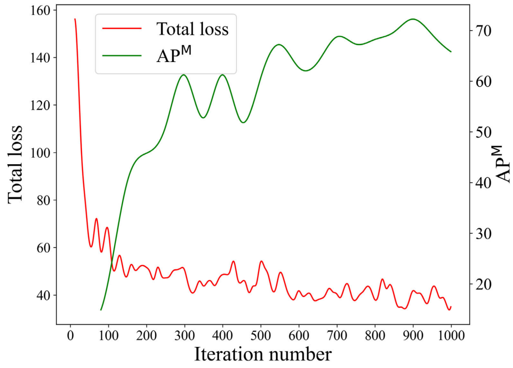


Fig. 9. Loss curve on the training dataset (red curve) and the changing curve of AP^{M} on the validation dataset (green curve) with theMask DINO^{Ada} model.

10) Potential for Real-Time Anomaly Detection: Real-time processing is the necessary demand for remote sensing processing [87], especially the Earth anomaly detection task. Traditional anomaly detection workflow has three main steps: on-board data generation, data downloading, and on-ground anomaly detection, where the data downloading step is time-consuming and increases the processing stresses. For our task, real-time processing is more necessary from two perspectives. First, most observations use hyperspectral data, which is always large in size and the size may be increased further with the improvement in spatial, spectral, and temporal resolution. This situation brings greater stress to the on-board data storage and downloading. Second, Earth anomaly detection always serves for disaster response [38] or defense safety [4], which are all high-timeliness applications.

This study can be seen as an extension of prior works [5] and [16], which all belong to the one-step paradigm without proxy tasks. In other words, the trained model can infer the unseen images directly without any fine-tuning. In addition to accuracy improvements, we believe that strong transferring performance is very meaningful to power the real-time anomaly detection.

V. Conclusion

In this work, we explored the anomaly segmentation task at the instance level for remote sensing imagery. Compared to pixel-level anomaly segmentation, the instance level results can reduce the noisy pixels significantly, support the anomaly counting, and be object-aware. Rather than designing a totally new architecture, we propose a general adaptation to adapt any query-based model from learning certain category objects to learning anomaly objects in a variety of categories. The adaptation extracts the anomaly information from pixel-level anomaly maps to guide the refinement of pixel embeddings and queries to be anomaly-aware. The proposed adaptation can be used on any query-based model and bring a stable promotion. Benefiting from the joint optimization of the pixel-level and instance level segmentation, we have also pushed the pixel-level anomaly detection accuracy to a higher level.

References

[1] S. Wang, X. Wang, L. Zhang, and Y. Zhong, "Deep low-rank prior for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3196409.

[2] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 25, no. 7, pp. 5–28, Jul. 2010.

[3] J. Li, X. Wang, H. Zhao, S. Wang, and Y. Zhong, "Anomaly segmentation for high-resolution remote sensing images based on pixel descriptors," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 4, pp. 4426–4434.

[4] S. Sun, J. Liu, X. Chen, W. Li, and H. Li, "Hyperspectral anomaly detection with tensor average rank and piecewise smoothness constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8679–8692, Nov. 2023.

[5] J. Li, X. Wang, H. Zhao, L. Zhang, and Y. Zhong, "A unified remote sensing anomaly detector across modalities and scenes via deviation relationship learning," 2023, *arXiv:2310.07511*.

[6] Y. Haitman, I. Berkovich, S. Havivi, S. Maman, D. G. Blumberg, and S. R. Rotman, "Machine learning for detecting anomalies in SAR data," in *Proc. IEEE Int. Conf. Microw., Antennas, Commun. Electron. Syst. (COMCAS)*, Nov. 2019, pp. 1–5.

[7] S. Wang, X. Wang, L. Zhang, and Y. Zhong, "Auto-AD: Autonomous hyperspectral anomaly detection network based on fully convolutional autoencoder," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5503314.

[8] J. Liu, Z. Hou, W. Li, R. Tao, D. Orlando, and H. Li, "Multipixel anomaly detection with unknown patterns for hyperspectral imagery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5557–5567, Oct. 2022.

[9] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct. 1990.

[10] L. Li, W. Li, Q. Du, and R. Tao, "Low-rank and sparse decomposition with mixture of Gaussian for hyperspectral anomaly detection," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4363–4372, Sep. 2021.

[11] Q. Ling, Y. Guo, Z. Lin, and W. An, "A constrained sparse representation model for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2358–2371, Apr. 2019.

[12] M. Muzeau, C. Ren, S. Angelliaume, M. Datcu, and J.-P. Ovarlez, "SAR anomalies detection based on deep learning," in *Proc. XXVIIIème Colloq. GRETSI*, 2022.

[13] H.-c. Shin and K. Na, "Anomaly detection using elevation and thermal map for security robot," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2020, pp. 1760–1762.

[14] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 10136197.

[15] J. Ma, W. Xie, Y. Li, and L. Fang, "BSDM: Background suppression diffusion model for hyperspectral anomaly detection," 2023, *arXiv:2307.09861*.

[16] J. Li, X. Wang, S. Wang, H. Zhao, L. Zhang, and Y. Zhong, "One-step detection paradigm for hyperspectral anomaly detection via spectral deviation relationship learning," 2023, *arXiv:2303.12342*.

[17] H. Cholakkal, G. Sun, F. S. Khan, and L. Shao, "Object counting and instance segmentation with image-level supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12389–12397.

[18] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "WHU-hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112012.

[19] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9659–9669.

[20] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.

[21] F. Li et al., "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3041–3050.

[22] W. Wang, M. Feiszli, H. Wang, J. Malik, and D. Tran, "Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4412–4422.

[23] C. Wang, G. Wang, Q. Zhang, P. Guo, W. Liu, and X. Wang, "OpenInst: A simple query-based method for open-world instance segmentation," 2023, *arXiv:2303.15859*.

[24] S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[25] H. Cevikalp, B. Uzun, Y. Salk, H. Saribas, and O. Köpüklü, "From anomaly detection to open set recognition: Bridging the gap," *Pattern Recognit.*, vol. 138, Jun. 2023, Art. no. 109385.

[26] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009, doi: 10.1145/1541880.1541882.

[27] C. Hu, "Remote detection of marine debris using satellite observations in the visible and near infrared spectral range: Challenges and potentials," *Remote Sens. Environ.*, vol. 259, Jan. 2021, Art. no. 112414.

[28] H. Zhao et al., "Mapping the distribution of invasive tree species using deep one-class classification in the tropical montane landscape of Kenya," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 328–344, May 2022.

[29] R. Shen, A. Huang, B. Li, and J. Guo, "Construction of a drought monitoring model using deep learning based on multi-source remote sensing data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 79, pp. 48–57, Jul. 2019.

[30] H. Yar, W. Ullah, Z. A. Khan, and S. Wook Baik, "An effective attention-based CNN model for fire detection in adverse weather conditions," *ISPRS J. Photogramm. Remote Sens.*, vol. 206, pp. 335–346, Dec. 2023.

[31] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, 2021.

[32] P. H. M. Ananias and R. G. Negri, "Anomalous behaviour detection using one-class support vector machine and remote sensing images: A case study of algal Bloom occurrence in inland waters," *Int. J. Digit. Earth*, vol. 14, no. 7, pp. 921–942, Jul. 2021, doi: 10.1080/17538947.2021.1907462.

[33] C. Li, Z. Zhang, L. Liu, J. Yoon Kim, and A. Kumar Sangaiah, "A novel deep multi-instance convolutional neural network for disaster classification from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2098–2114, 2024, doi: 10.1109/JSTARS.2023.3340413.

[34] M. Noman, M. Fiaz, H. Cholakkal, S. Khan, and F. S. Khan, "ELGC-Net: Efficient local–global context aggregation for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4701611.

[35] X. Ning, H. Zhang, R. Zhang, and X. Huang, "Multi-stage progressive change detection on high resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 207, pp. 231–244, Jan. 2024.

[36] J. A. Padrón-Hidalgo, V. Laparra, N. Longbotham, and G. Camps-Valls, "Kernel anomalous change detection for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7743–7755, Oct. 2019, doi: 10.1109/TGRS.2019.2916212.

[37] W. Shi, M. Zhang, H. Ke, X. Fang, Z. Zhan, and S. Chen, "Landslide recognition by deep convolutional neural network and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4654–4672, Jun. 2021.

[38] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sens. Environ.*, vol. 265, Nov. 2021, Art. no. 112636.

[39] C.-I. Chang, S. Chen, S. Zhong, and Y. Shi, "Exploration of data scene characterization and 3D ROC evaluation for hyperspectral anomaly detection," *Remote Sens.*, vol. 16, no. 1, p. 135, Dec. 2023.

[40] A. Banerjee, P. Burlina, and C. Diehl, "A support vector method for anomaly detection in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2282–2291, Aug. 2006.

[41] W. Li and Q. Du, "Collaborative representation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1463–1474, Mar. 2015.

[42] P. K. Pokala, R. V. Hemadri, and C. S. Seelamantula, "Iteratively reweighted minimax-concave penalty minimization for accurate low-rank plus sparse matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8992–9010, Dec. 2022.

[43] W. Xie, J. Lei, B. Liu, Y. Li, and X. Jia, "Spectral constraint adversarial autoencoders approach to feature representation in hyperspectral anomaly detection," *Neural Netw.*, vol. 119, pp. 222–234, Nov. 2019.

[44] H. Kwon and N. M. Nasrabadi, "Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 2, pp. 388–397, Feb. 2005.

[45] P. Gurram and H. Kwon, "Support-vector-based hyperspectral anomaly detection using optimized kernel parameters," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 6, pp. 1060–1064, Nov. 2011.

[46] Q. Guo, B. Zhang, Q. Ran, L. Gao, J. Li, and A. Plaza, "Weighted-RXD and linear filter-based RXD: Improving background statistics estimation for anomaly detection in hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2351–2366, Jun. 2014.

[47] C.-I. Chang, "Hyperspectral target detection: Hypothesis testing, signal-to-noise ratio, and spectral angle theories," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5505223.

[48] C. I. Chang, "Hyperspectral anomaly detection: A dual theory of hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5511720.

[49] T. Guo, L. He, F. Luo, X. Gong, Y. Li, and L. Zhang, "Anomaly detection of hyperspectral image with hierarchical antinoise mutual-incoherence-induced low-rank representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510213.

[50] L. Du, Z. Wu, Y. Xu, W. Liu, and Z. Wei, "Kernel low-rank representation for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 477–480.

[51] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[52] Y. Xu, Z. Wu, J. Li, A. Plaza, and Z. Wei, "Anomaly detection in hyperspectral images based on low-rank and sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 1990–2000, Apr. 2016.

[53] N. Wang, B. Li, Q. Xu, and Y. Wang, "Automatic ship detection in optical remote sensing images based on anomaly detection and SPP-PCANet," *Remote Sens.*, vol. 11, no. 1, p. 47, Dec. 2018.

[54] Y. Li, T. Jiang, W. Xie, J. Lei, and Q. Du, "Sparse coding-inspired GAN for hyperspectral anomaly detection in weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5512811.

[55] Z. Li, Y. Wang, C. Xiao, Q. Ling, Z. Lin, and W. An, "You only train once: Learning a general anomaly enhancement network with random masks for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5506718.

[56] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[57] K. Chen et al., "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4969–4978.

[58] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6402–6411.

[59] Y. Fang et al., "Instances as queries," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6890–6899.

[60] R. Li, C. He, S. Li, Y. Zhang, and L. Zhang, "DynaMask: Dynamic mask selection for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11279–11288.

[61] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[62] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. NIPS*, Dec. 2021, pp. 17864–17875.

[63] B. Kim, Y. Yoo, C. E. Rhee, and J. Kim, "Beyond semantic to instance segmentation: Weakly-supervised instance segmentation via semantic knowledge transfer and self-refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4268–4277.

[64] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12737–12747.

[65] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 228–239, Jan. 2022.

[66] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[67] 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. (CVPR)*, Jul. 2017, pp. 936–944.

[68] A. Li, C. Qiu, M. Kloft, P. Smyth, S. Mandt, and M. Rudolph, "Deep anomaly detection under labeling budget constraints," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19882–19910.

[69] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[70] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query DeNoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13609–13617.

[71] H. Zhang et al., "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[72] H. Zhao, X. Wang, J. Li, and Y. Zhong, "Class prior-free positive-unlabeled learning with Taylor variational loss for hyperspectral remote sensing imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16827–16836.

[73] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[74] X. Wang, R. Girdhar, S. X. Yu, and I. Misra, "Cut and learn for unsupervised object detection and instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3124–3134.

[75] H. Guo, X. Yang, N. Wang, and X. Gao, "A CenterNet++ model for ship detection in SAR images," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107787.

[76] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. False alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8508–8517.

[77] Y. Kondo et al., "MVA2023 small object detection challenge for spotting birds: Dataset, methods, and results," in *Proc. 18th Int. Conf. Mach. Vis. Appl. (MVA)*, Jul. 2023, pp. 1–11.

[78] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.

[79] B. Xu, X. Li, W. Hou, Y. Wang, and Y. Wei, "A similarity-based ranking method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9585–9599, Nov. 2021.

[80] C.-I. Chang, "Multiparameter receiver operating characteristic analysis for signal detection and classification," *IEEE Sensors J.*, vol. 10, no. 3, pp. 423–442, Mar. 2010.

[81] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[82] Y. Qu et al., "Hyperspectral anomaly detection through spectral unmixing and dictionary-based low-rank decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4391–4405, Aug. 2018.

[83] S. Mabu, K. Fujita, and T. Kuremoto, "Disaster area detection from synthetic aperture radar images using convolutional autoencoder and one-class SVM," *J. Robot. Netw. Artif. Life*, vol. 6, no. 1, pp. 48–51, 2019.

[84] S. Sinha et al., "Variational autoencoder anomaly-detection of avalanche deposits in satellite SAR imagery," in *Proc. 10th Int. Conf. Climate Informat.*, Sep. 2020, pp. 113–119.

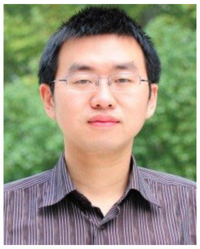
[85] C. Lile and L. Yiqun, "Anomaly detection in thermal images using deep neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2299–2303.

[86] Z. You et al., “A unified model for multi-class anomaly detection,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Jun. 2022, pp. 4571–4584.

[87] Z. Zhang, Z. Qu, S. Liu, D. Li, J. Cao, and G. Xie, “Expandable on-board real-time edge computing architecture for Luojia3 intelligent remote sensing satellite,” *Remote Sens.*, vol. 14, no. 15, p. 3596, Jul. 2022.

Jingtao Li received the B.S. degree from the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China, in 2021. He is currently pursuing the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan.

His major research interests include anomaly detection in remote sensing.



Yanfei Zhong (Senior Member, IEEE) received the B.S. degree in information engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

Since 2010, he has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. He has organized the Intelligent Data Extraction, Analysis and Applications of Remote Sensing (RSIDEA) Research Group. He has published more than 100 research articles in international journals, such as *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include hyperspectral remote sensing information processing, high-resolution remote sensing image understanding, and geoscience interpretation for multisource remote sensing data and applications.

Dr. Zhong is a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). He won the Second-Place Prize in the 2013 IEEE GRSS Data Fusion Contest and the Single-View Semantic 3-D Challenge of the 2019 IEEE GRSS Data Fusion Contest. He is currently serving as an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and the *International Journal of Remote Sensing*.

Hengwei Zhao (Member, IEEE) received the B.S. degree in surveying and mapping engineering from the School of Resources and Civil Engineering, Northeastern University, Shenyang, China, in 2019. He is currently pursuing the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.

Zhi Gao (Member, IEEE) received the B.Eng. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

In 2008, he joined the Interactive and Digital Media Institute, National University of Singapore (NUS), Singapore, as a Research Fellow (A) and the Project Manager. In 2014, he joined the Temasek Laboratories, NUS (TL@NUS), as the Research Scientist (A) and a Principal Investigator. He is currently working as a Full Professor with the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 70 research papers on top journals and conferences, such as *International Journal of Computer Vision* (IJCV), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS (TIE), IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), *ISPRS Journal of Photogrammetry and Remote Sensing* (JPRS), *Neurocomputing*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), Conference on Computer Vision and Pattern Recognition (CVPR), European Conference on Computer Vision (ECCV), Asian Conference on Computer Vision (ACCV), and British Machine Vision Conference (BMVC). Since 2019, he has been supported by the Distinguished Professor Program of Hubei Province and the National Young Talent Program, China. His research interests include computer vision, machine learning, and remote sensing and their applications. In particular, he has a strong interest in vision for intelligent systems and intelligent-system-based vision.

Dr. Gao serves as an Associate Editor for the *Unmanned Systems* journal.

Xinyu Wang (Member, IEEE) received the B.S. degree in photogrammetry and remote sensing and the Ph.D. degree in communication and information systems from Wuhan University, Wuhan, China, in 2014 and 2019, respectively.

Since 2019, he has been an Associate Research Fellow with the School of Remote Sensing and Information Engineering, Wuhan University. His major research interests include hyperspectral data processing and applications.