CTMANet: A CNN-Transformer Hybrid Semantic Segmentation Network for Fine-Grained Airport Extraction in Complex SAR Scenes

Keyu Wu, *Student Member, IEEE*, Feng Cai, *Student Member, IEEE*, and Haipeng Wang, *Senior Member, IEEE*

***Abstract*—Airports represent essential infrastructure, offering substantial research and application potential. However, extracting airports from complex synthetic aperture radar (SAR) scenes is challenging due to the cluttered background and fine structure of airports. This necessitates the integration of global and local information for fine-grained extraction. To tackle this issue, this article introduces a novel framework for fine-grained extraction of airports from large-scale SAR images. First, a convolutional neural networks (CNN) transformer hybrid semantic segmentation network with multiscale contextual fusion is proposed, named CNN-transformer network (CTMANet). In this network, the encoder combines CNNs and transformers to capture local and global information, while the multiscale context aggregation block fuses multiscale contextual information. Skip connections between the encoder and decoder are established to minimize the loss of detailed information and fuse low-level features with high-level semantic features. Moreover, a category balance block is designed to address class imbalance. Experimental results on the GF-3 dataset demonstrate that CTMANet outperforms state-of-the-art methods, proving its superior suitability for fine-grained airport extraction in large-scale scenarios.**

***Index Terms*—Convolutional neural networks (CNN) transformer hybrid, fine-grained airport extraction, multiscale context fusion, semantic segmentation.**

I. Introduction

AIRPORTS are significant infrastructure and transportation hubs. Therefore, airport extraction is widely employed in civil and military applications like airport navigation and aerial reconnaissance [1], [2], [3]. Synthetic aperture radar (SAR) has the advantage of all-day, all-weather imaging for earth observation, which can overcome the impact of inclement weather on airport observation [4], [5], [6]. In SAR image, rough ground surfaces exhibit higher backscatter while flat ground surfaces appear as dark areas. According to the scattering properties of SAR images, as shown in Fig. 1(a), the airport region typically presents distinctive visual features in SAR images [7]. In summary, SAR image is ideal for researching airport extraction. Therefore, this article aims to precisely extract airport regions from high-resolution large-scene SAR images, focusing on three crucial aspects.

1) The complex scenes often contain areas with visual characteristics similar to the airport such as rivers and roads, which can lead to false alarms.
2) A significant class imbalance exists between the airport class and the background class in large-scale SAR images. An example is shown in Fig. 1, where the airport occupies only a fraction of the total image area [see Fig. 1(a)], with the minimum bounding box encompassing merely 6% of the entire image [see Fig. 1(b) and (c)].
3) Airports exhibit diverse structures with rich details, and their runways are interconnected with long spans.

Therefore, it is essential for the airport extraction network to effectively capture and leverage both global and local information. Specifically, airport runways, rivers, and roads exhibit properties of long-span and connectivity. Therefore, the network requires a large receptive field and the capacity to extract and distinguish these global pieces of information. Conversely, the intricate structure of airports, including smaller-scale targets like aprons and the rich detail in the background, demands the network that focuses on minimizing resolution loss and retaining fine-grained features to capture these local pieces of information effectively. Additionally, multiscale feature fusion is crucial for this task.

In recent years, plentiful airport extraction studies have been proposed and we summarize these studies in two ways: 1) traditional methods and 2) deep learning-based methods. In traditional approaches, researchers mainly utilize priori knowledge and hand-designed features to extract airport. There are three primary categories of traditional approaches, but each has some clear flaws. 1) Extract the airport through the runway's edges [8], [9], [10]. Such methods merely use the most obvious low-level features of airport and are only effective for small scene images. However, this line segment-based method is prone to interference from other objects with line features in complex scenes, resulting in false alarms. 2) Using image segmentation to extract airports by leveraging high-level features such as texture and region [11], [12]. These approaches demonstrate superior accuracy compared to the line segment-based method. Despite this, its implementation is intricate and time-consuming due to

Manuscript received 4 October 2023; revised 10 December 2023 and 8 January 2024; accepted 30 January 2024. Date of publication 5 February 2024; date of current version 16 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62271153 and in part by the Natural Science Foundation of Shanghai under Grant 22ZR1406700. *(Corresponding author: Haipeng Wang.)*

The authors are with the Key Laboratory of Information Science of Electromagnetic Waves, Fudan University, Shanghai 200433, China (e-mail: kywu21@m.fudan.edu.cn; fcai21@m.fudan.edu.cn; hpwang@fudan.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3361927

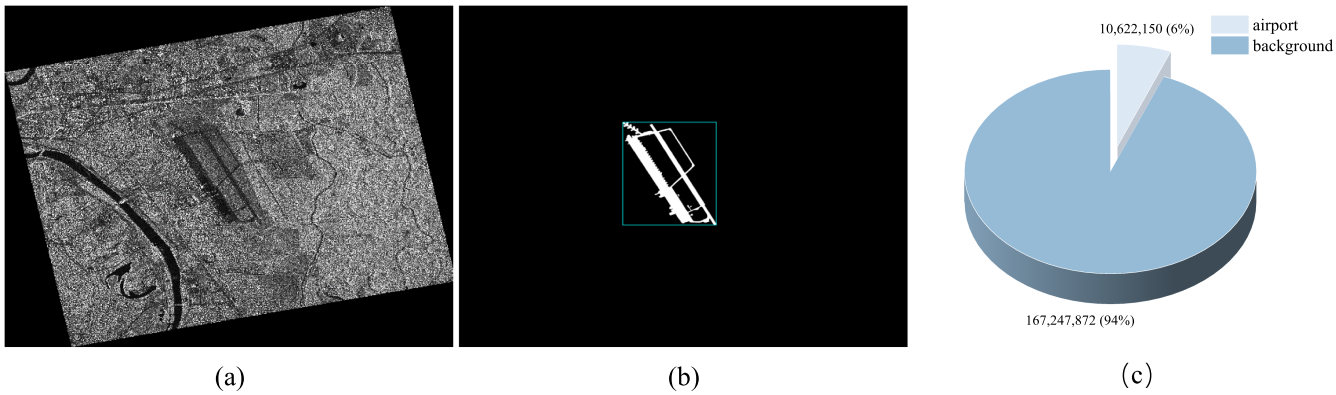

Fig. 1. Example of the airport in large-scale images. (a) SAR image. (b) Corresponding ground truth with the airport's minimum bounding box. (c) Ratio of the minimum bounding box to the area of the whole image.

the pixel-level analysis and multifaceted features. 3) Methods based on saliency analysis [13], [14], [15]. Such methods primarily utilize information on differences between the airport area and other areas to extract the airport. Nevertheless, areas such as rivers and roads generally present salient features similar to airports in SAR images. Furthermore, the background of large-scale SAR images is complicated, making it difficult for such methods to accurately and effectively extract airports. To sum up, most traditional methods rely on the priori knowledge and manually predesigned features, with limited parameters, making them challenging to apply in complex scenarios.

Deep neural networks (DNN) have the capability to learn feature representations from data, overcoming the limitations of traditional methods based on manual tuning of finite parameters and feature design [16]. In recent years, object detection and semantic segmentation based on DNN have been widely used for airport extraction. However, the airport extraction method based on object detection [2], [17], [18], only employs bounding boxes to represent the detected airport area, which cannot precisely identify the airport contour. The extraction of contours can be achieved by semantic segmentation method [7], [19], which employs neural networks to classify pixels into airport or background categories to obtain airport contours. Although the promising results have been obtained, they involve the following limitations. 1) The object detection [17], [18] or the semantic segmentation that only extracts runways [19] fails to finely extract the complete airport areas including runway, taxiway, and apron (hereinafter collectively referred to as airport areas). Extracting these components holds significant importance. First, these areas serve as distinguishing features that set airports apart from other regions. Second, extracting these areas can narrow the detection area of military targets in large-scale SAR images, thereby boosting the detection accuracy [20]. 2) Most approaches are based on convolutional neural networks (CNN), whose convolution can only model the relationship between neighboring pixels in an image, weakening the importance of contextual information [21], [22]. Transformer [23] can model relationships between all pixels in an image, thus addressing the limitation of CNN in obtaining global context [24]. But even so, in contrast to CNN, transformer lacks two inductive biases: 1) Locality: due to the sliding operation of convolutional kernels over the image, CNN assumes greater correlation among adjacent pixels, with closer proximity indicating stronger correlations. 2) Translation invariance. Consequently, transformer still falls short in capturing local details [25]. Therefore, combining the structures of CNN and transformer can complement each other well, taking into account both local features and global features. The combined CNN and transformer approaches have proven to be effective in both object detection task [26], [27] and semantic segmentation task [28], [29], [30]. CNN-transformer architectures also have widespread applications in the field of remote sensing [31], [32], [33], [34], [35], [36]. He et al. [31] embedded the Swin transformer into the CNN-based UNet, designing a transformer-CNN parallel structure with dual encoders. Zhang et al. [33] proposed a hybrid network based on a multilevel series-parallel combination of CNN and transformer. However, these methods faced challenges in effectively fusing features from the parallel CNN and transformer branches, potentially increasing computational demands. Li et al. [34] proposed a sequential structure combining transformer and UNet. However, it did not take into account the fusion of multiscale features.

To tackle the aforementioned problems, this article proposes a novel framework for fine-grained extraction of airports in large-scale intricate scenes, in which a hybrid CNN-transformer network (CTMANet) with multiscale context aggregation (MCA) is presented. In CTMANet, an encoder is designed to capture low-level features and local information with CNN and then models long-range dependencies with transformer. The last component of the encoder is the MCA block, which consists of a sequence of cascaded dilated convolutions. This block enables the fusion of detailed and coarse information, enlarging the receptive field without compromising resolution. Moreover, skip connections are utilized between the encoder and decoder to restore the spatial information that is lost during downsampling. Additionally, a category balance (CB) block is introduced to split the SAR images and achieve category balance. The main contributions of this article can be summarized as follows.

1) A novel method for fine-grained airport extraction from high-resolution large-scene SAR images is proposed, effectively addressing the challenges posed by complex

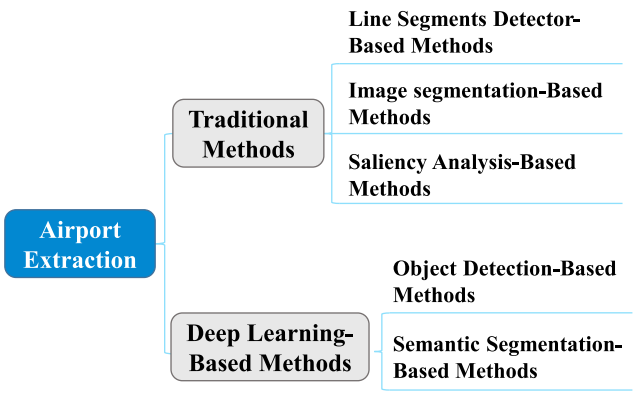

Fig. 2. Summary of the recent research on airport extraction.

backgrounds and intricate structures of airports. This method demonstrates notable efficacy, achieving a mIoU of 95.74% and an F_1 Score of 95.59% on the Gaofen-3 large-scale SAR image dataset.

2) We proposed the CTMANet, a semantic segmentation network that synergizes the strengths of CNN and transformer. The MCA block is introduced to capture and fuse multiscale features. This network can handle the complexity, long span, and scale variations of an airport, making it easier to distinguish airports from objects that have similar visual characteristics to airports.
3) We propose the CB block to process large-scale images, achieving class balance instead of direct cropping or scaling. This approach prevents the model from being biased toward the background and enables fair treatment of both positive and negative samples, subsequently enhancing the accuracy of segmentation.

The rest of this article is organized as follows. Section II reviews the current status of research on airport extraction. Section III describes in detail the airport extraction method proposed in this article. Section IV presents the experimental datasets and details, and showcases the experimental results of comparing CTMANet with state-of-the-art methods. In addition, partial ablation experiments are also shown for better analysis of the improved components of the network. Section V includes a discussion of the experimental results and implications. Finally, Section VI concludes this article.

II. Related Work

In this section, we summarize the recent research on airport extraction and categorize these works into two groups: traditional methods and deep learning-based methods, as depicted in Fig. 2.

A. Traditional Airport Extraction Methods

Previous researches focused on airport extraction using a priori knowledge and limited feature design. Xiong et al. [8] proposed an airport runway recognition method in SAR images based on Radon transform and hypothesis testing. Di et al. [9] extracted the airport using a combination of improved chain codes based edge tracking and the Hough transform. To address the drawbacks of the Hough transform such as long time consuming, Bai et al. [37] introduced an improved algorithm based on the Hough transform. Kou et al. [38] proposed a remote sensing image airport extraction method based on line segment detector. Wang et al. [39] introduced a fast line segment detection algorithm with a new filter for SAR images. This group of methods concentrates on extracting airport by detecting the edges of airport runways. Aytekin et al. [11] introduced a texture-based approach for airport runway detection. Liu et al. [40] combined texture segmentation and shape detection to extract airport. Tao et al. [12] proposed an airport extraction method based on the clustered scale-invariant feature transform key points and region information. Zhang et al. [41] used image segmentation, support vector machine classification, and shape analysis to extract airport. These methods mainly employ image segmentation to obtain the region of interest and extract airports using high-level features such as texture, shape, and region of the airports. Different from the pixel-by-pixel analysis method based on region segmentation, Wang et al. [13] proposed an airport extraction method based on saliency map. Zhang et al. [14] utilized a two-layer visual saliency analysis model and support vector machine to detect airport. To further enhance the efficiency, Liu et al. [42] combined line segment grouping and saliency analysis to extract airport. These saliency analysis-based methods primarily use the differences between airport areas and other areas to extract airport. However, traditional methods face limitations in adaptability, scalability, and model capacity to their reliance on prior knowledge and manually predesigned features. These approaches struggle with large-scale datasets and are sensitive to parameter tuning.

B. Airport Extraction Methods Based on Deep Neural Networks (DNN)

DNN excel in automatically learning hierarchical features from data, exhibit superior performance across varied scenarios and extensive datasets. DNN overcome the limitations of traditional methods, which rely on manual tuning of limited parameters and feature design. Combining traditional methods with DNN is one of the airport extraction research directions. Zhang et al. [43] combined transfer learned CNN and region proposal with prior knowledge to extract airport. Zhu et al. [44] proposed a two-step object detection framework integrating saliency and ResNet. Xiao et al. [45] proposed an approach to detect airport by using a GoogleNet with a light feature module model and a support vector machine enhanced with hard negative mining. Although the above methods improve the efficiency of airport extraction, they still require tedious image preprocessing.

In recent years, object detection and semantic segmentation have emerged as the primary approaches for investigating airport extraction, which almost no longer relies solely on manual prior knowledge. Chen et al. [17] used an improved Faster RCNN algorithm to deal with the typical elongated linear geometric shape of an airport, reducing the image preprocessing to some extent. Yin et al. [18] applied multiscale training and hard sample mining strategy to Faster RCNN for airport extraction. Cai et al.

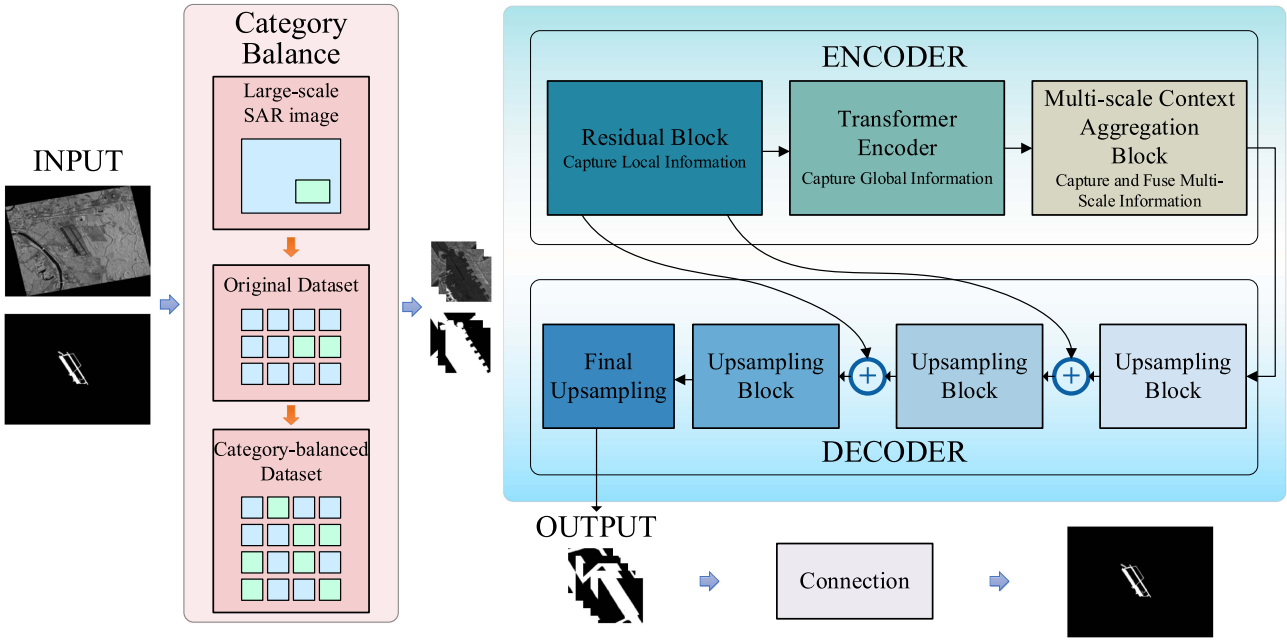

Fig. 3. Overview of the proposed airport extraction method.

[46] proposed a hard examples mining based CNN to solve the airport-background class imbalance problem in training. Men et al. [19] utilized DeepLabv3 with Lovasz-Softmax loss to extract contours of airport runways from remote sensing images. Daltla et al. [47] introduced a multimodal semantic segmentation for extracting airport runway in panchromatic remote sensing images. The two above methods [19], [47] focus solely on pixelwise airport runway extraction, ignoring other important components of the airport such as taxiway and apron. Therefore, Chen et al. [7] proposed the multilevel dual attention mechanism network, which can extract the airport's runway area including runway, taxiway, and apron. However, they downsampled the SAR images by a factor of five before splitting, which class imbalance between airport and background was also ignored. To overcome these limitations, CB block is proposed to preserve more detailed information and achieve class balance.

III. Methodology

This section provides a comprehensive introduction to the proposed method. We first present an overview of the airport extraction process and the general framework of CTMANet. Next, we delve into two crucial components of CTMANet: the transformer encoder and the MCA block. Finally, we introduce the specific functionality of the CB block.

A. Overview

Fig. 3 illustrates the overall workflow of the proposed method. Given a large-scale SAR image, its size is too large to be directly processed by DNNs. As a preliminary step, the CB block is applied to slice images and equalize the class distribution within the training set. Specifically, the CB block first segments these images into 512 × 512-pixel nonoverlapping slices. It then focuses on slices containing airports (determined from the minimum bounding box coordinates of the airports), and performs overlapping slicing on these areas along with data augmentation. The CB block ensures the generation of a class-balanced dataset. This dataset is then fed to CTMANet, a hybrid network composed of an encoder and a decoder.

The encoder operates in three stages: it first employs the residual block for initial feature extraction. On the one hand, CNN can offer more prior knowledge to capture local information because of its inductive biases. On the other hand, downsampling the input feature map with CNN before applying transformer's self-attention can reduce computational load. Then, utilizes the transformer encoder to capture long-range contextual information, and finally incorporates the MCA block to fuse multiscale contextual information. In addition, skip connections between the encoder and decoder are added to preserve low-level feature details.

Finally, the decoder yields an output image of 512 × 512 pixels, generated through four up-sampling operations. During the prediction phase, the image slices are stitched together to generate the final large-size prediction results. To be specific, slices are systematically named and indexed row-by-row and column-by-column during the CB block slicing process. The network's predictions align with the naming scheme. Subsequently, the slices are indexed row-by-row and column-by-column to facilitate the connection of the slices to reconstruct the complete image.

B. CTMANet Architecture

The overall architecture of the proposed CTMANet is shown in Fig. 4, which follows an encoder–decoder structure. We design a three-stage encoder consisting of a residual block,

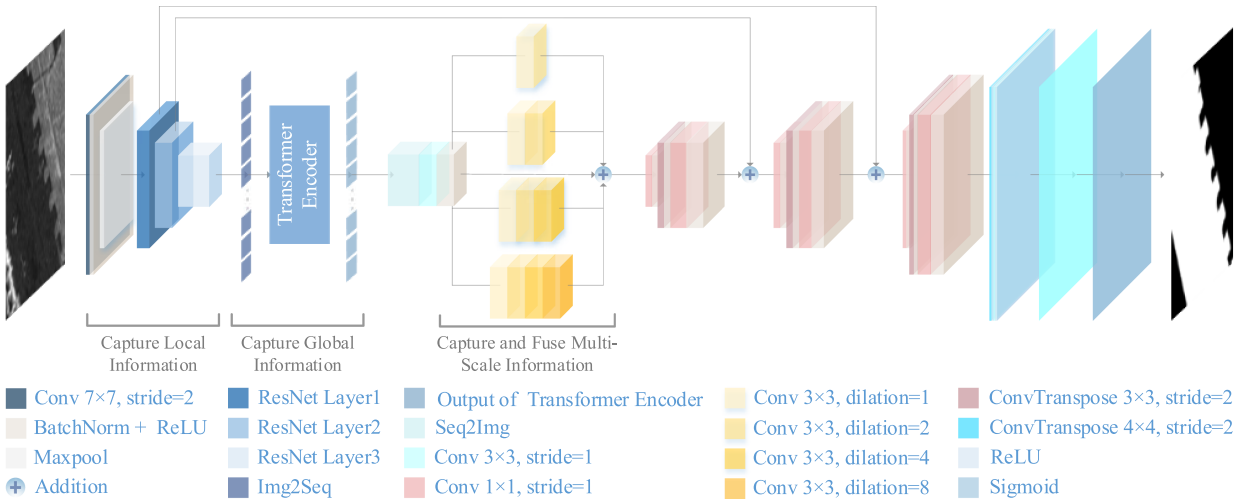

Fig. 4. Structure of the proposed CTMANet. It mainly consists of a CNN-transformer hybrid encoder with multiscale aggregation and a decoder with upsampling by transposed convolution.

transformer encoder, and MCA block. In the first stage, a pre-trained ResNet50 [48] serves as the residual block to capture local information. Note that the last layer of ResNet50 [48] is removed to reduce down-sampling operations, thus enlarging the size of the feature map to retain more details. In the second stage, transformer encoder is employed to capture global information. For the feature map $X \in \mathbb{R}^{H\times W\times C}$, the Img2Seq block first reshapes it into patches $x_p \in \mathbb{R}^{N\times D}$, where H, W, and C denote the height, width, and channel number, respectively. Here, $N = (H \times W)/P^2$ and $D = P \times P \times C$ represent the patch count and dimension. Then, Img2Seq block maps the patches x_p into latent embedding space using a learnable linear projection, obtaining patch embedding $x_{pe} \in \mathbb{R}^{N\times d}$, where d is the embedding dimension. The x_{pe} processed by the Img2Seq block is then fed into the transformer encoder for long-range dependency capture. Seq2Img block initially employs a linear projection to transform the output of the transformer encoder into patches $z \in \mathbb{R}^{N\times D}$with dimensions D, and then reshape patches z into feature maps $Z \in \mathbb{R}^{H\times W\times C}$. In the third stage, the MCA block captures and fuses multiscale contextual information from the feature maps of maximum size. After the above three encoding stages, feature maps with a size of 1/16 of the original input image are obtained. High-level features in the decoder possess robust semantic information but lack in detailed information, while low-level features in the encoder are semantically weaker but rich in detail. Consequently, two skip connections are established between the encoder's residual block and the decoder to fuse high-level semantics with low-level details, and to recover local spatial information lost due to downsampling. In the decoder, transposed convolution is employed for upsampling, restoring spatial resolution incrementally. Ultimately, a prediction mask is obtained through a sigmoid function.

C. Transformer Encoder

Transformer excels in modeling long-range dependencies within a sequence of embeddings owing to the self-attention mechanism. As shown in Fig. 5, the transformer encoder takes the patch embedding $x_{pe} \in \mathbb{R}^{N\times d}$ derived from Img2Seq block, concatenates a learnable class token, and then adds a position embedding. The transformer encoder's core consists of M transformer layers, with input tokens $t \in \mathbb{R}^{(N+1)\times d}$. Each transformer layer comprises layer norm (LN), multihead attention (MHA), and multilayer perceptron (MLP). The single-head self-attention (SHSA) projects the input t into query (Q), key (K), and value (V) via different projection matrices W^Q, W^K, and W^V, respectively. MHA extends on SHSA by using distinct projection matrices in each attention head. Specifically, for h attention heads, the Q_i, K_i, V_i of the ith attention head are obtained from the projection of matrices W_i^Q, W_i^K, $W_i^V \in \mathbb{R}^{d\times(d/h)}$, respectively. They can be formalized as follows:

$$Q_i = tW_i^Q \tag{1}$$

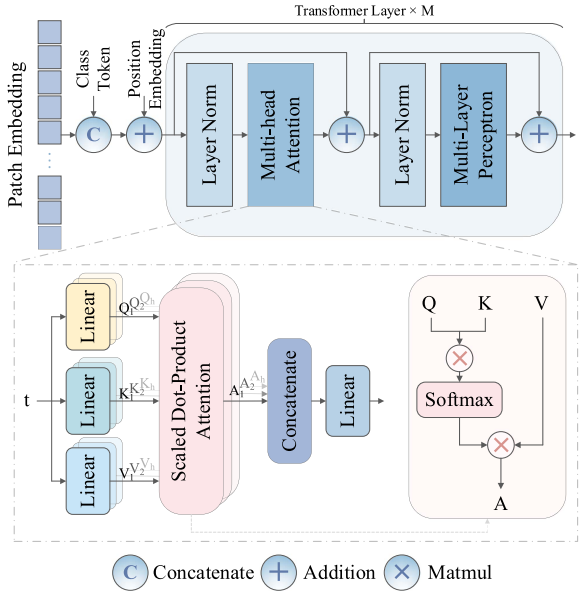

Fig. 5. Structure of the transformer Encoder.

$$K_i = tW_i^K \tag{2}$$

$$V_i = tW_i^V. \tag{3}$$

Scaled dot-product attention (SDPA) of *i*th attention head in MHA can be defined as

$$\text{Attention}\,(Q_i,\ K_i, V_i) = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{(d/h)}}\right) V_i. \tag{4}$$

The formula involves three main steps: First, the dot product between Q_i and K_i is computed and scaled by the square root of the dimension of $V_i(d/h)$. Then, the Softmax function is applied to the scaled dot product to generate the attention weights. Finally, the attention weights are multiplied by V_i to obtain the output of SDPA. The output of MHA is produced by applying a linear projection to the concatenated results of SDPA. Both M and h are set to 12 in this article. The output z_m of mth transformer layer can be expressed as

$$z_{mt} = \text{MHA}\,(\text{LN}\,(z_{m-1})) + z_{m-1} \tag{5}$$

$$z_m = \text{MLP}\,(\text{LN}\,(z_{mt})) + z_{mt}. \tag{6}$$

Here, z_{m-1} is the output from $(m-1)$th transformer layer. z_{mt} is the result of applying MHA to the output of the $(m-1)$th layer after LN, and then adding z_{m-1}. Finally, the output z_m is obtained by applying MLP to the result of LN applied to z_{m-1}, and then adding z_{mt}.

D. Multiscale Context Aggregation Block

The MCA block is proposed to fuse multiscale features. This block is composed of dilated convolution, which can enlarge the receptive field while maintaining the resolution of the feature map, preventing the loss of spatial information. Deeplabv3 [49] designed modules that use dilated convolution in cascade or parallel manner to address segmenting objects across multiple scales. Inspired by this, this article integrates both parallel and cascaded structures. As shown in Fig. 4, the MCA block incorporates four parallel branches, each cascaded by dilated convolutions with varying dilation rates. The receptive fields of the final output feature maps across these four branches stand at 37, 15, and 31, respectively. The fusion of multiscale features is achieved by summing the output feature maps of these branches with the MCA block's input feature map.

E. Category Balance Block

The CB block is designed to address class imbalance in the dataset, given that airports comprise only a minor segment of the total image. As depicted in Fig. 3, the CB block initially partitions the input large-scale SAR images and ground truth into nonoverlapping 512 × 512-pixel slices. Following this, the minimum bounding box coordinates for the airports are identified based on their labels, and the airport sections within these boxes are subsequently cut. The nonoverlapping slices bear no correlation to one another, but the pixels of objects belonging to the same class are contiguous and have inherent semantic correlation. Consequently, a 50% overlap rate is allowed between adjacent slices by setting the slicing stride to 256 pixels, ensuring semantic information continuity to some extent. Through overlapping slicing, each slice incorporates some duplicated pixel information, belonging to both the current slice and its adjacent slices. The model can leverage pixel information within the overlapping region to infer and capture continuous features in the image. Consequently, when processing each slice, the model can acquire contextual information, mitigating errors during slice stitching. Subsequently, data augmentation is employed, such as flipping and random rotations, to inject diversity within the class data. This strategy effectively alleviates the negative impact of class imbalance.

IV. Experiments

In this section, we introduce the dataset and the experimental setup. To evaluate the performance of CTMANet, we compared the proposed approach with seven state-of-the-art semantic segmentation methods, and reported quantitative results on the test set. To showcase CTMANet, we provide the visualization results and details of two airport scenes. Moreover, we conducted ablation experiments to evaluate the impact of individual blocks in the proposed architecture. Finally, we present an analysis of the impact of the ratio of positive and negative samples on the segmentation performance.

A. Dataset

The dataset used in this article contains 34 real SAR images captured by the Gaofen-3 satellite with a resolution of 0.5 m/pixel. The size of each image is about 15 000 × 10 000 pixels. And the polarization modes include horizontal–horizontal and vertical-vertical. Pixel-level labeling is performed on the SAR images using MATLAB's image labeler. For precise annotation of the airport contour, optical images from Google Earth are used as a reference. Airport regions are annotated as the airport class in white masks (255, 255, 255), while the rest are labeled as background class in black masks (0, 0, 0). An example is shown in Fig. 6.

To facilitate the experiments, the input large-scale SAR images and corresponding ground truths are split into small slices of 512 × 512 pixels via the CB block. Furthermore, to mitigate the class imbalance in the training set, data augmentation strategies are employed in CB block. The training set and test set encompass 26 783 and 7208 small images, respectively. Representative slices of the dataset are shown in Fig. 7. To better visualize the performance of the proposed model, the predicted masks for the test set are stitched back to match the original size of the large-scale image.

B. Implementation Details

The experimental parameter settings are presented in Table I. During the training phase, the SGD optimizer with a batch size at 16 is employed. The initial learning rate is set to 0.05 and gradually decreases from the initial value to the lowest point and then gradually returns to 0.05 within every 100 epochs

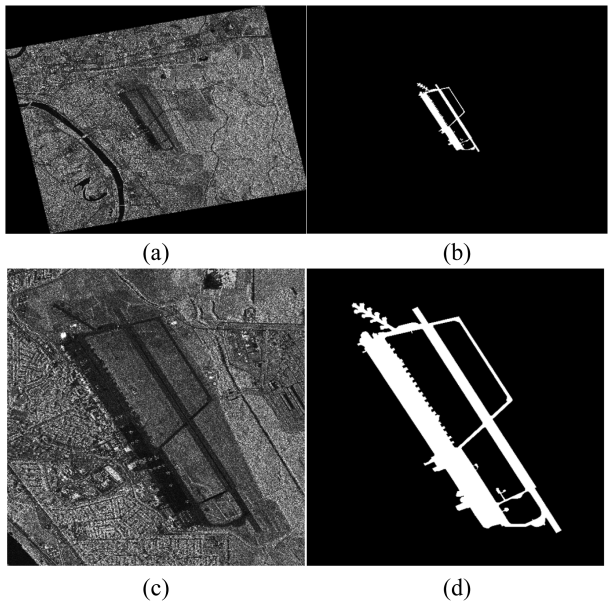

Fig. 6. Example of SAR images used in this article. (a) Original large-scale SAR image. (b) Ground truth of (a). (c) Details of the airport in (a). (d) Ground truth of (c).

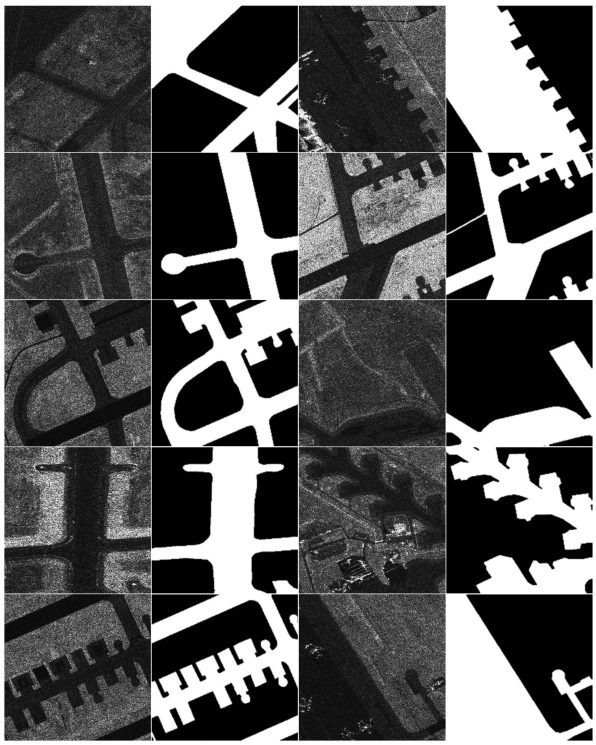

Fig. 7. Samples of the dataset.

with the CosineAnnealingLR scheduler. The network reaches convergence at approximately 500 epochs.

The sum of binary cross entropy (BCE) loss L_{BCE} and dice loss L_{Dice} are used as loss functions, which can be expressed as follows:

$$L_{\text{BCE}} = -X \log Y + (1 - X) \log (1 - Y) \tag{7}$$

TABLE I
TRAINING SETTING

Training setting	Value
Input size	512×512
Batch size	16
Max epochs	540
Optimizer	SGD
Lr_scheduler	CosineAnnealingLR
Initial learning rate	0.05
Loss function	$L_{BCE} + L_{Dice}$

$$L_{\text{Dice}} = 1 - \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{8}$$

$$L_{\text{total}} = L_{\text{BCE}} + L_{\text{Dice}} \tag{9}$$

where X denotes the ground truth and Y is the prediction of the model.

In tasks like airport extraction where the imbalance between target and background is significant, BCE loss is less effective. This is due to the imbalance between the airport ($Y = 1$) and background ($Y = 0$) pixels, where BCE loss biases the model towards the background. Dice loss measures the overlap between predictions and labels, which can address the class imbalance problem by focusing more on airport regions. However, dice loss can be prone to significant loss fluctuations and gradient instability in training, especially in binary classification with small positive samples. Therefore, we combine BCE loss and Dice loss to mitigate these issues.

C. Evaluation Metric

This article adopts Intersection over Union (IoU), mean IoU (mIoU), and F_1 score to evaluate the model's performance. IoU measures the overlap between the predicted segmentation and ground truth, computed as the ratio of the intersection to the union of the prediction and ground truth for a single category. And mIoU is the average of IoU of all categories in the dataset. F_1 score is the harmonic mean of Precision and Recall. Precision represents the proportion of pixels correctly predicted as positive (i.e., airport) among all pixels predicted as positive. And Recall indicates the proportion of pixels correctly predicted among all pixels whose ground truth is the airport. The formulas used to calculate these metrics are as follows:

$$\text{IoU} = \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} (P_{ji} - P_{ii})} \tag{10}$$

$$\text{mIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} (P_{ji} - P_{ii})} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

TABLE II
QUANTITATIVE COMPARISON OF SEGMENTATION RESULTS ON TEST SET (%)

Method	Backbone	IoU		mIoU	F_1 score
		background	airport		
DANet	ResNet50	99.88	88.33	94.10	93.80
PSPNet	ResNet50	99.87	87.34	93.60	93.24
HRNet	HRNetV2-W18	99.83	84.38	92.10	91.53
DMNet	ResNet50	99.87	87.29	93.58	93.21
Deeplabv3+	ResNet50	99.89	89.39	94.64	94.40
OCRNet	HRNetV2-W18	99.85	86.26	93.05	92.62
Segformer	MIT-B0	99.86	86.57	93.21	92.80
Ours	**ResNet50**	**99.91**	**91.57**	**95.74**	**95.59**

The best values in the column are in bold.

$$\mathrm{F}_1\ \mathrm{Score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{14}$$

where $k+1$ denotes the number of categories, and $k=1$ given that airport extraction constitutes a binary semantic segmentation task in this article. P_{ii} represents the count of pixels belonging to class I that are correctly predicted. P_{ij} is the count of pixels whose ground truth is class I but are predicted as class J. P_{ji} denotes the count of pixels that belong to class J but are predicted as class I. TP, FP, and FN correspondingly represent true positive, false positive, and false negative, respectively.

D. Comparison With State-of-the-Art Methods

To validate the segmentation performance of CTMANet, it was compared with seven state-of-the-art semantic segmentation methods. The backbone of DANet [50], PSPNet [51], DMNet [52], Deeplabv3+ [53] is ResNet50 [48]; for HRNet [54], OCRNet [55], it is HRNetV2-W18; and for Segformer [56], it is MIT-B0.

Table II presents the quantitative comparison between CTMANet and these other methods on the test set. Among the other four methods utilizing ResNet50 [48] as their backbone, DANet [50] and Deeplabv3+ [53] emerged as superior performers. CTMANet rivals or even surpasses both these methods, with at least 2.18% (91.57 versus 89.39) higher IoU of the airport and 1.1% (95.74 versus 94.64) higher in mIoU. Compared with HRNet [54], the IoU of airport, mIoU, and F_1 score of the proposed method are 7.19% (91.57 versus 84.38), 3.67% (95.74 versus 92.10), and 4.06% (95.59 versus 91.53) higher, respectively. Segformer [56], which is based on the transformer, and OCRNet [55], which combines HRNet [54] and transformer, deliver comparable performance. Though CTMANet also incorporates the benefits of transformer, it surpasses these two methods by achieving at least 5% (91.57 versus 86.57), 2.53% (95.74 versus 93.21), and 2.79% (95.59 versus 92.80) higher IoU of airport, mIoU, and F_1 score, respectively. CTMANet achieves the most competitive results among all considered methods, which quantitatively demonstrates the effectiveness and superiority.

To provide a more intuitive demonstration of CTMANet's performance, we showcase visualizations and in-depth details for two airports.

TABLE III
INFORMATION OF AIRPORT I

Coordinates	Resolution	Size
99.9° W, 32.4° N	0.5	15 872×11 776

1) Results of Airport I: Table III illustrates the information about the image containing airport I. Fig. 8 shows the segmentation results on airport I. Fig. 8(a) depicts the SAR image and corresponding ground truth. Fig. 8(b)–(i) presents the fusion maps and predictions from DANet [50], PSPNet [51], HRNet [54], DMNet [52], Deeplabv3+ [53], OCRNet [55], Segformer [56], and CTMANet, respectively. The details of the airport I extraction results are shown in Fig. 9. Notably, the yellow boxes highlight regions where the background is misclassified as airport categories, whereas the red boxes represent the misclassification and omission of airport regions. This airport is situated in a farmland area with a town area to its right [see Fig. 8(a)]. The complex road network in the vicinity tends to be misclassified as the airport category due to its similar appearance. Other methods exhibit varying degrees of misclassification in these areas, as shown in the yellow boxes of Fig. 8(b)–(h). On closer examination, it can be seen that the roads surrounding the airport are often misinterpreted as airport areas by other methods [see Fig. 9(b)–(h)]. In contrast, the proposed method, exhibits almost no such errors, demonstrating its capability to better distinguish these roads from the airport areas. Meanwhile, this airport is surrounded by buildings with irregular boundaries, especially noticeable in the apron area. Other methods exhibit serious misclassification and omission in these areas, as indicated by the red boxes in Fig. 9(b)–(h). However, CTMANet aligns most accurately with the airport area, as shown in Fig. 9(i). Finally, at the splicing of slices, some methods [see Fig. 9(b), (c), (e), and (f)] show different degrees of errors, while CTMANet has almost no errors at the splicing. This demonstrates CTMANet's superiority to capture and fuse multiscale information, enhancing the overall scene understanding.

Table IV summarizes the quantitative results of these methods on airport I. The IoU of airport for the other methods is below 80%, whereas CTMANet achieves 88.91%. Moreover, CTMANet outperforms the other methods in terms of both mIoU and F_1 score.

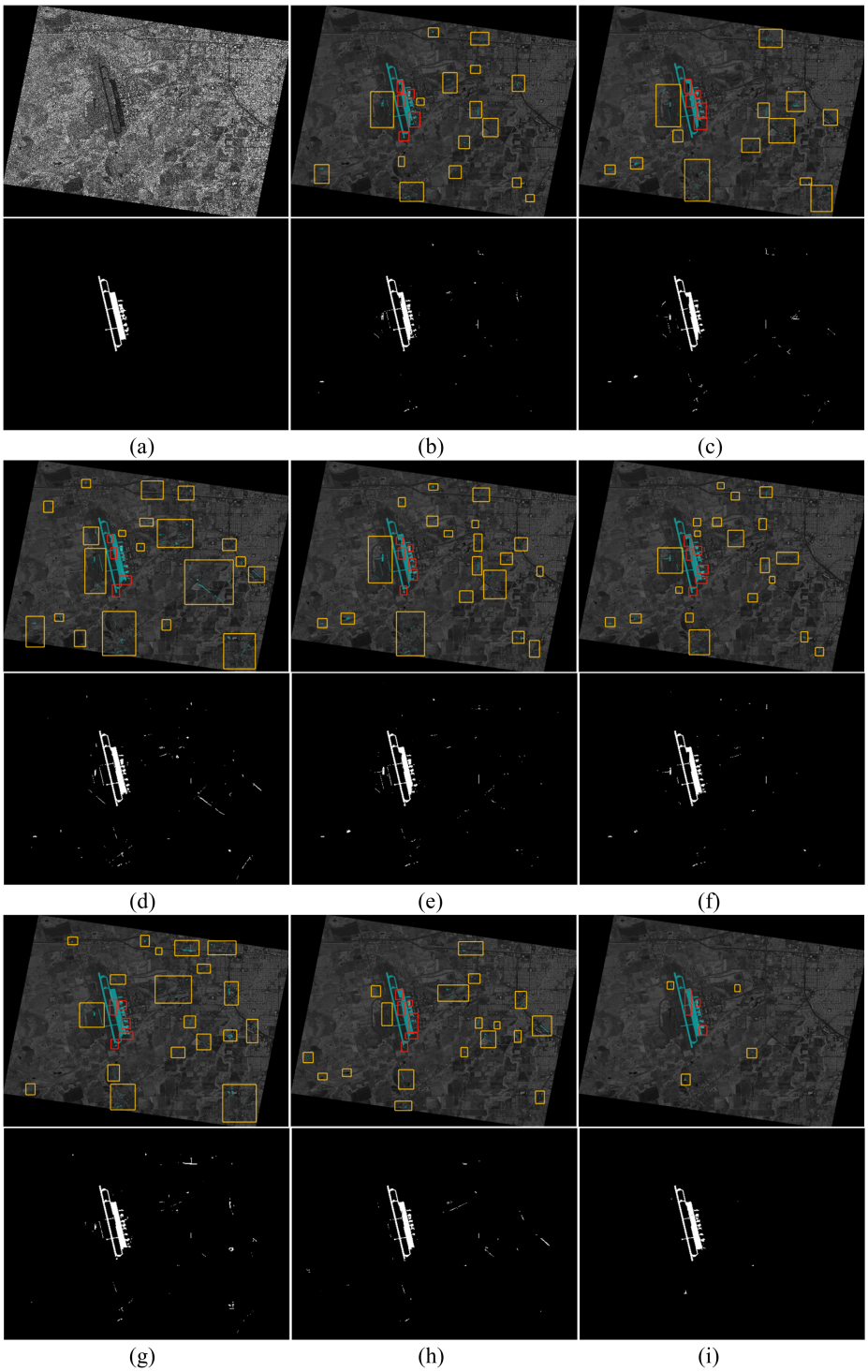

Fig. 8. Segmentation results for airport I. (a) Original SAR image and the corresponding ground truth. (b) DANet. (c) PSPNet. (d) HRNet. (e) DMNet. (f) Deeplabv3+. (g) OCRNet. (h) Segformer. (i) Ours.

2) Results of Airport II: The segmentation results of CTMANet compared to currently popular methods on airport II are visualized in Fig. 10. Table V provides information about the image. This airport is located on the outskirts of town, surrounded by buildings and roads. Most other methods inaccurately classify darker areas, like roads, as part of the airport (as seen in the yellow box). However, CTMANet does not share this issue, showing practically no false alarms.

The advantages of CTMANet are evident in the extraction results of airport II, as shown in Fig. 11. In the upper part of the airport, the airport area is prone to confusion with its surrounding nonairport areas. And this area is also where the

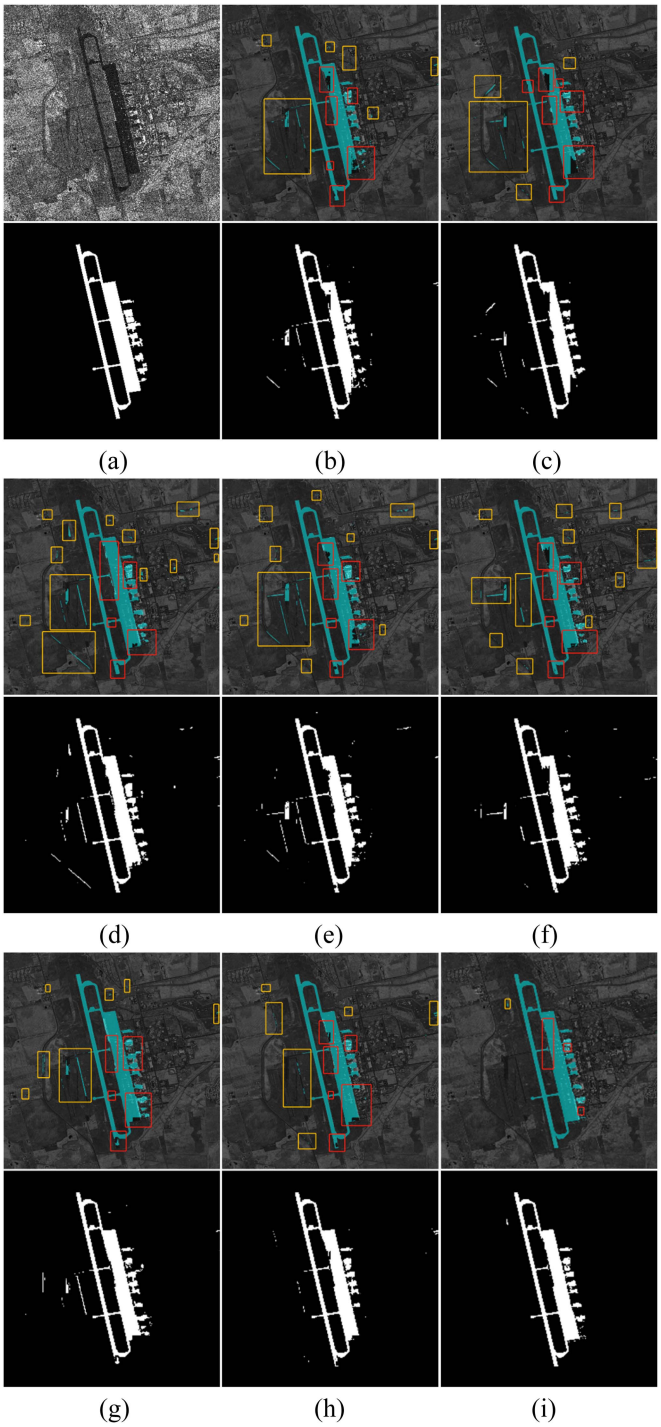

Fig. 9. Details of segmentation results for Airport I. (a) SAR image and the corresponding ground truth. (b) DANet. (c) PSPNet. (d) HRNet. (e) DMNet. (f) Deeplabv3+. (g) OCRNet. (h) Segformer. (i) Ours.

slices are spliced. However, as demonstrated in Fig. 11(b)–(h), other methods suffer from severe misclassification and omission in this region, whereas CTMANet extracts the airport area almost completely in terms of edge and detail extraction.

Table VI illustrates the quantitative comparison of segmentation results on airport II of the eight methods. Even though there are a few misclassified pixels, they are relatively small in comparison to the total number of background pixels. Therefore, the IoU of the background remains high despite the obvious misclassification. Reducing background misclassifications has a negligible effect on the IoU of the background. In this case, comparison of other evaluation metrics unambiguously shows that CTMANet outperforms the other methods, particularly in terms of the IoU of airport

The elaborate experimental results and analysis demonstrate that CTMANet can accurately and completely extract the airport region from large-scale images, while exhibiting almost no false alarms. It proves that the proposed method shows stronger robustness in complex scenes, especially in challenging regions such as those with dark visual features and messy environments.

TABLE IV
QUANTITATIVE COMPARISON OF SEGMENTATION RESULTS ON AIRPORT I (%)

Method	IoU		mIoU	F_1 score
	background	airport		
DANet	99.76	77.41	88.58	87.26
PSPNet	99.70	74.80	87.25	85.58
HRNet	99.68	73.30	86.49	84.59
DMNet	99.72	75.22	86.49	84.59
Deeplabv3+	99.77	78.86	89.31	88.18
OCRNet	99.70	74.80	87.25	85.58
Segformer	99.78	79.39	89.58	88.51
Ours	**99.89**	**88.91**	**94.40**	**94.13**

The best values in the column are in bold.

TABLE V
INFORMATION OF AIRPORT II

Coordinates	Resolution	Size
93.7° W, 32.5° N	0.5	14 848 × 11 264

TABLE VI
QUANTITATIVE COMPARISON OF SEGMENTATION RESULTS ON AIRPORT II (%)

Method	IoU		mIoU	F_1 score
	background	airport		
DANet	99.88	91.08	95.48	95.33
PSPNet	99.88	91.43	95.66	95.52
HRNet	99.87	90.36	95.11	94.94
DMNet	99.84	88.43	94.14	93.86
Deeplabv3+	99.89	91.69	95.79	95.66
OCRNet	99.88	91.43	95.66	95.52
Segformer	99.89	91.12	95.50	95.35
Ours	**99.93**	**95.23**	**97.58**	**97.55**

The best values in the column are in bold.

E. Ablation Study

To verify the effectiveness of each block, an ablation study was conducted. The quantitative results are displayed in Table VII and the corresponding visualizations are in Fig. 12. The baseline model is a version of CTMANet, which does not include the transformer encoder and MCA block. We assessed the impact of transformer encoder, MCA block, and CB block on segmentation performance. Additionally, the influence of backbone selection on segmentation accuracy by substituting ResNet50 [48] with ResNet101 is further investigated.

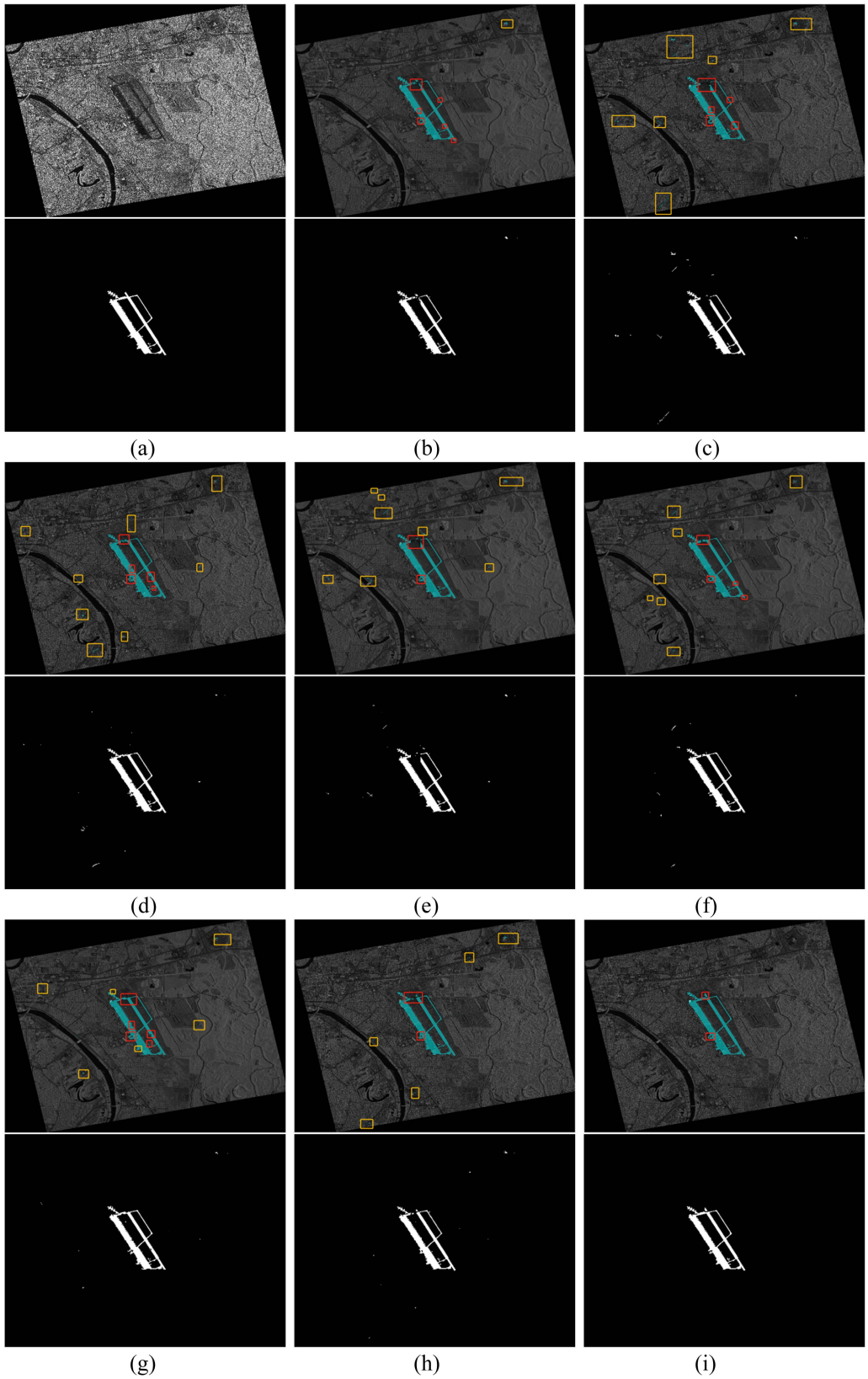

Fig. 10. Segmentation results for airport II. (a) Original SAR image and the corresponding ground truth. (b) DANet. (c) PSPNet. (d) HRNet. (e) DMNet. (f) Deeplabv3+. (g) OCRNet. (h) Segformer. (i) Ours.

TABLE VII
ABLATION STUDY ON TEST SET (%)

Baseline	Backbone	MCA Block	Transformer Encoder	CB Block	IoU		mIoU	F_1 score
					background	airport		
Baseline	ResNet50			√	99.71	74.98	87.34	85.70
	ResNet50	√		√	99.85	85.95	92.90	92.44
	ResNet50		√	√	99.90	89.89	94.89	94.67
	ResNet50	√	√		99.77	78.22	88.99	87.78
	ResNet101	√	√	√	99.90	90.73	95.32	95.14
	ResNet50	√	√	√	**99.91**	**91.57**	**95.74**	**95.59**

The best values in the column are in bold.

TABLE VIII
EFFECT OF DIFFERENT BRANCHES IN MCA BLOCK ON SEGMENTATION RESULTS (%)

	Branches of MCA Block				IoU		mIoU	F_1 score
	1	2	3	4	background	airport		
CTMANet					99.90	89.89	94.89	94.67
	√				99.90	90.23	95.07	94.86
	√	√			99.90	90.79	95.35	95.17
	√	√	√		99.91	91.10	95.50	95.34
	√	√	√	√	**99.91**	**91.57**	**95.74**	**95.59**

As illustrated in Table VII, the baseline method attains an IoU of airport, mIoU, and F_1 score of 74.98%, 87.34%, and 85.70%, respectively. Notwithstanding the multitude of misclassifications, the IoU of the background remains high due to the sheer abundance of background pixels. The visualization result of the baseline [see Fig. 12(c)] displays a misclassification of dark areas, such as water bodies, as airport regions. Moreover, the model exhibits insufficient accuracy in airport extraction, particularly at the edges.

1) Effect of MCA block: Table VII indicates that incorporating the MCA block led to a significant improvement in segmentation performance, with the IoU of airport increasing by 10.97%, mIoU by 5.56%, and F_1 score by 6.74%. Furthermore, as shown in Fig. 12(d), the introduction of the MCA block effectively reduced misclassifications in areas with dark colors, such as rivers.

In addition, we explored the effect of different branches in the MCA block on segmentation results. The four branches of the MCA block are numbered 1 to 4 from top to bottom (see Fig. 4). As shown in Table VIII, adding the first branch improves the mIoU and F_1 score by 0.18% and 0.19%, respectively. The incorporation of the second branch led to increases in mIoU and F_1 score by 0.28% and 0.31%, respectively. Adding the third branch resulted in further improvements in mIoU and F_1 score by 0.15% and 0.17%, respectively. With the addition of the fourth branch, the mIoU and F_1 score increased by 0.24% and 0.25%, respectively. Each branch captures information at different scales, so each is important for enhancing segmentation accuracy.

2) Effect of Transformer Encoder: Table VIII demonstrates that after employing the transformer encoder independently, the model achieves 14.91%, 7.55%, and 8.97% improvements on IoU of airport, mIoU, and F_1 score, respectively. Fig. 12(e) shows a marked reduction in the number of false alarms in the background part, maintaining only a marginal fraction of misclassifications, thereby enhancing the accuracy of segmentation.

3) Effect of CB Block: the effect of the class balance strategy in CB block is also verified. Fig. 12(f) illustrates that the model exhibits poorer performance without the class balance strategy. Owing to the severe disproportion between positive and negative samples in the original dataset, the model fails to adequately learn features of the airports. Moreover, it demonstrates a significant reduction in errors at the splice of the image when using the CB block [see Fig. 12(h)] compared to not using it [see Fig. 12(f)], highlighting the importance of the CB block in ensuring the continuity of semantic information between neighboring slices.

Furthermore, the impact of the proportion of positive and negative samples on segmentation accuracy is explored. Fig. 13 provides details for three datasets exhibiting different proportions. The original dataset contains only 666 airport slices with a positive-to-negative sample ratio of approximately 1:20. To overcome the issue of class imbalance, the class balance strategy is used in the CB block to expand the dataset and acquire two additional datasets with positive-to-negative sample ratios of roughly 1:4 and 1:1.

The segmentation results for the different datasets are displayed in Fig. 14. Notably, with the ratio at 1:4, there is a marked improvement in mIoU and F_1 score by 4.65% and 5.5%, respectively. Furthermore, the ratio of 1:1 proved to be the most advantageous for all evaluation metrics, enhancing the mIoU by 6.75% and the F_1 score by 7.81% in comparison to the original dataset.

CTMANet achieves the best performance [see Fig. 12(h)], which is closest to the ground truth with almost no false alarms. The effectiveness of the joint blocks is fully demonstrated in ablation experiments. And ResNet50 [48] is still adopted as the backbone because our replacement of the backbone with

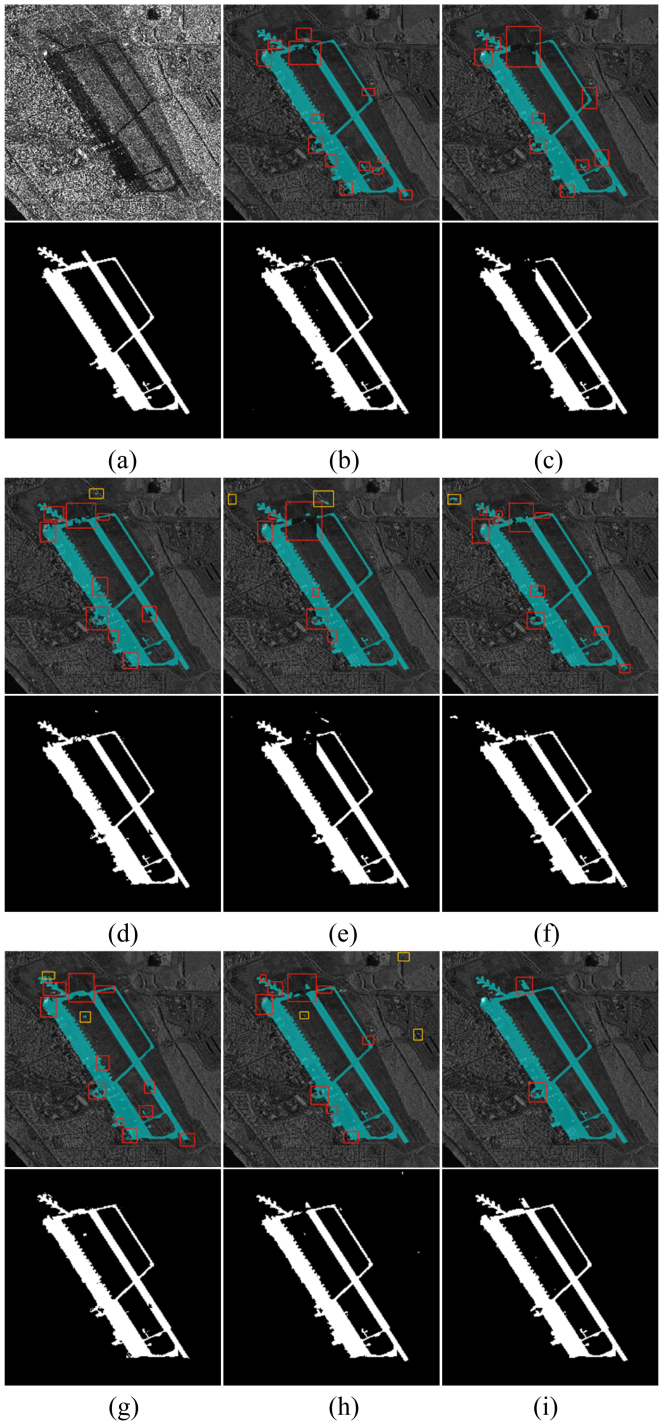

Fig. 11. Details of segmentation results for Airport II. (a) SAR image and the corresponding ground truth. (b) DANet. (c) PSPNet. (d) HRNet. (e) DMNet. (f) Deeplabv3+. (g) OCRNet. (h) Segformer. (i) Ours.

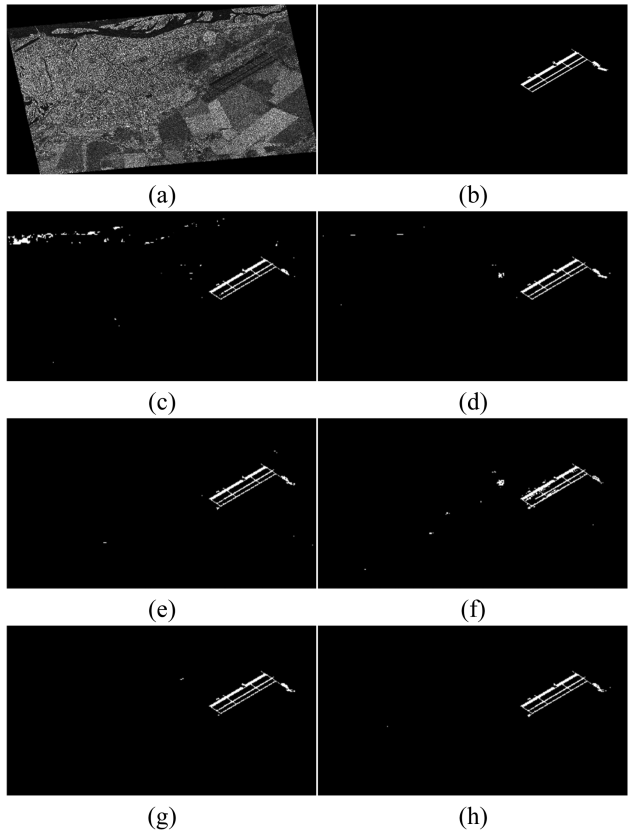

Fig. 12. Segmentation results of ablation study. (a) SAR image. (b) Ground truth. (c) Baseline. (d) Baseline + MCA block. (e) Baseline + transformer encoder. (f) CTMANet without CB block. (g) CTMANet (ResNet101). (h) CTMANet.

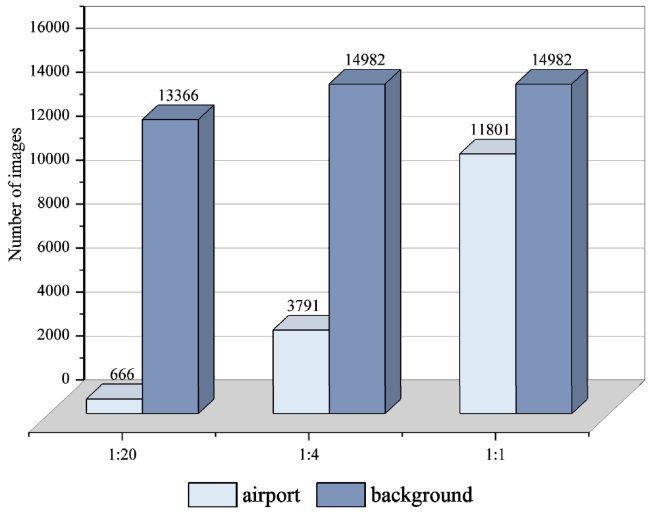

Fig. 13. Datasets with different proportions of positive and negative samples.

ResNet101 [48] did not show any particular impact on the segmentation accuracy.

V. Discussion

This article proposes CTMANet, a CNN-transformer hybrid network with a MCA block, designed to address the challenges posed by complex contexts and the fine structure of airport in airport extraction. CNN is adept at capturing local information because of its inductive biases: locality and translation invariance. But fall short in global context comprehension. In contrast, transformer excels at capturing long-range dependencies and can obtain global representations from shallow layers [57], but is struggle to capture local information. CTMANet combines both

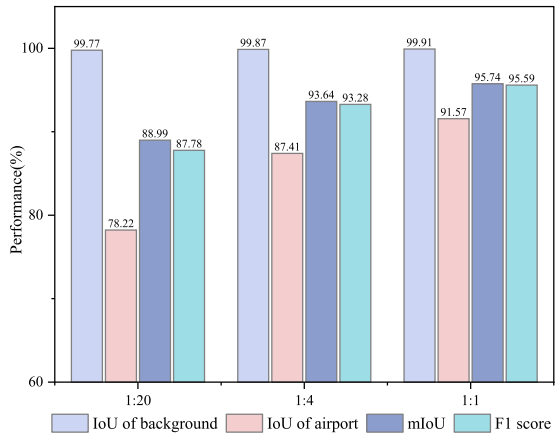

Fig. 14. Segmentation performance of datasets with proportions of positive and negative samples.

of these network architectures, aiming to effectively capture and utilize local detail and global context for airport extraction.

The approach is validated through comparative experiments and ablation studies. Experiments reveal that for long-span targets within airports, CNN-based methods such as PSPNet, DANet, and DeepLabv3+ (see Fig. 9) suffer from significant misclassifications and omissions. Conversely, transformer-based network like Segformer shows no severe errors for such targets. For smaller-scale targets like aprons, the segmentation accuracy of CNN-based methods surpasses that of Segformer. The proposed method achieves accurate segmentation for both long-span and small-scale targets.

Ablation studies (see Fig. 12) showcase a long-span river in the original image. The baseline model [see Fig. 12(c)], a CNN-based architecture without the transformer encoder and MCA block, shows effective segmentation for smaller airport targets, indicating CNN's proficiency in capturing local information. However, it significantly misclassifies the long-span river and the airport runway. Adding the transformer encoder to the baseline [see Fig. 12(e)] eliminates misclassifications in the river and improves airport runway segmentation accuracy. These experiments illustrate the transformer's superior capability to capture long-range dependencies for identifying long-span targets, while also highlighting the CNN's aptitude in capturing fine-grained local information for smaller-scale targets. Considering the challenges posed by varying target scales, this article further introduced the MCA block. Ablation studies reveal that adding the MCA block to the baseline [see Fig. 12(d)] significantly improves the misclassification of the long-span river and restores finer details of the airport, demonstrating the module's effectiveness in capturing and fusing multiscale information.

Another challenge is the small proportion of airport areas in large-scale SAR images, leading to an abundance of background images in slices. Ablation experiments [see Fig. 12(f)] show that without category balancing, the model heavily biases towards background areas, resulting in accurate background segmentation but severe errors in airport areas. To address this, this article introduced the CB block, which improves the segmentation accuracy by implementing class balancing.

In summary, this article proposes a novel airport extraction framework, stemming from perspectives of capturing global and local information, multiscale information fusion, and positive-negative sample class balance. This method enhances the model's comprehension of complex scenes, and improves segmentation precision for objects of various scales.

VI. Conclusion

In this article, a method for fine-grained airports extraction from large-scale SAR images is proposed, aiming to address the challenges posed by the complex backgrounds and detailed airport structures. A novel semantic segmentation network, CTMANet, is introduced, which combines CNN and transformer to effectively capture both local and global contextual information. The three-stage encoder includes CNNs for local feature capture, transformers for long-range dependency processing, and an MSA block for multiscale contextual feature fusion. Skip connections between the encoder and decoder facilitate the fusion of low-level details and high-level semantics, while transposed convolution is employed for upsampling in the decoder. Additionally, a CB block is incorporated to alleviate the class imbalance caused by the small proportion of airports in large-scale images. Experiments with various state-of-the-art methods on a real large-scale SAR dataset demonstrate the effectiveness and superiority of CTMANet. Future work will focus on optimizing and adapting the proposed method for broader applications.

References

[1] J. Tu, F. Gao, J. Sun, A. Hussain, and H. Zhou, "Airport detection in SAR images via salient line segment detector and edge-oriented region growing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 314–326, 2021, doi: 10.1109/JSTARS.2020.3036052.

[2] Y. Xu, M. Zhu, S. Li, H. Feng, S. Ma, and J. Che, "End-to-end airport detection in remote sensing images combining cascade region proposal networks and multi-threshold detection networks," *Remote Sens.*, vol. 10, no. 10, Sep. 2018, Art. no. 1516.

[3] S. Li, Y. Xu, M. Zhu, S. Ma, and H. Tang, "Remote sensing airport detection based on end-to-end deep transferable convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1640–1644, Apr. 2019.

[4] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Jul. 2019.

[5] F. Gao, T. Huang, J. Sun, J. Wang, A. Hussain, and E. Yang, "A new algorithm for SAR image target recognition based on an improved deep convolutional neural network," *Cogn. Comput.*, vol. 11, pp. 809–824, Jun. 2019.

[6] X. Wei, W. Zheng, C. Xi, and S. Shang, "Shoreline extraction in SAR image based on advanced geometric active contour model," *Remote Sens.*, vol. 13, no. 4, Feb. 2021, Art. no. 642.

[7] L. Chen et al., "A new framework for automatic airports extraction from SAR images using multi-level dual attention mechanism," *Remote Sens.*, vol. 12, no. 3, Feb. 2020, Art. no. 560.

[8] W. Xiong, J. Zhong, and Y. Zhou, "Automatic recognition of airfield runways based on Radon transform and hypothesis testing in SAR images," in *Proc. 5th Glob. Symp. Millimeter-Waves*, 2012, pp. 462–465.

[9] N. Di, M. Zhu, and Y. Wang, "Real time method for airport runway detection in aerial images," in *Proc. Int. Conf. Audio Lang. Image Process.*, 2008, pp. 563–567.

[10] Ü. Budak, U. Halıcı, A. Şengür, M. Karabatak, and Y. Xiao, "Efficient airport detection using line segment detector and fisher vector representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1079–1083, May 2016.

[11] Ö. Aytekin, U. Zöngür, and U. Halici, "Texture-based airport runway detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 471–475, Aug. 2012.

[12] C. Tao, Y. Tan, H. Cai, and J. Tian, "Airport detection from large IKONOS images using clustered SIFT keypoints and region information," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 128–132, Jul. 2010.

[13] X. Wang, Q. Lv, B. Wang, and L. Zhang, "Airport detection in remote sensing images: A method based on saliency map," *Cogn. Neurodynamics*, vol. 7, pp. 143–154, Sep. 2013.

[14] L. Zhang and Y. Zhang, "Airport detection and aircraft recognition based on two-layer saliency model in high spatial resolution remote-sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1511–1524, Nov. 2016.

[15] D. Zhao, Y. Ma, Z. Jiang, and Z. Shi, "Multiresolution airport detection via hierarchical reinforcement learning saliency model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2855–2866, Mar. 2017.

[16] X. Dong, J. Tian, and Q. Tian, "A feature fusion airport detection method based on the whole scene multispectral remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1174–1187, Jan. 2022.

[17] F. Chen, R. Ren, T. Van de Voorde, W. Xu, G. Zhou, and Y. Zhou, "Fast automatic airport detection in remote sensing images using convolutional neural networks," *Remote Sens.*, vol. 10, no. 3, Mar. 2018, Art. no. 443.

[18] S. Yin, H. Li, and L. Teng, "Airport detection based on improved faster RCNN in large scale remote sensing images," *Sens. Imag.*, vol. 21, pp. 1–13, Oct. 2020.

[19] Z. Men, J. Jiang, X. Guo, L. Chen, and D. Liu, "Airport runway semantic segmentation based on DCNN in high spatial resolution remote sensing images," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 361–366, Nov. 2020.

[20] S. Tan, L. Chen, Z. Pan, J. Xing, Z. Li, and Z. Yuan, "Geospatial contextual attention mechanism for automatic and fast airport detection in SAR imagery," *IEEE Access*, vol. 8, pp. 173627–173640, 2020.

[21] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[22] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.

[23] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, Art. no. 25430.

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

[25] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2021, Art. no. 5518615.

[26] 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.*, 2020, pp. 213–229.

[27] 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. Representations*, Jan. 2020.

[28] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "Mpvit: Multi-path vision transformer for dense prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7287–7296.

[29] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 17864–17875, Dec. 2021.

[30] 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.*, 2022, pp. 1290–1299.

[31] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 4408715.

[32] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5522214.

[33] X. Zhang, S. Cheng, L. Wang, and H. Li, "Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 2000415.

[34] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5622519.

[35] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jun. 2023, Art. no. 5514415.

[36] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 4408820.

[37] X. Bai, Y. Han, D. Wu, and F. Zhang, "The automatic detection method for airport runway," in *Proc. Int. Congr. Image Signal Process.*, 2015, pp. 1015–1019.

[38] Z. Kou, Z. Shi, and L. Liu, "Airport detection based on line segment detector," in *Proc. IEEE Conf. Comput. Vis. Remote Sens.*, 2012, pp. 72–77.

[39] D. Wang, Q. Liu, Q. Yin, and F. Ma, "Fast Line Segment Detection and Large Scene Airport Detection for PolSAR," *Remote Sens.*, vol. 14, no. 22, Nov. 2022, Art. no. 5842.

[40] D. Liu, L. He, and L. Carin, "Airport detection in large aerial optical imagery," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 761–764.

[41] S. Zhang, Y. Lin, X. Zhang, and Y. Chen, "Airport automatic detection in large space-borne SAR imagery," *J. Syst. Eng. Electron.*, vol. 21, no. 3, pp. 390–396, Jun. 2010.

[42] N. Liu, Z. Cui, Z. Cao, Y. Pi, and S. Dang, "Airport detection in large-scale SAR images via line segment grouping and saliency analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 434–438, Jan. 2018.

[43] P. Zhang, X. Niu, Y. Dou, and F. Xia, "Airport detection on optical satellite images using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1183–1187, Jun. 2017.

[44] T. Zhu, Y. Li, Q. Ye, H. Huo, and T. Fang, "Integrating saliency and ResNet for airport detection in large-size remote sensing images," in *Proc. 2nd Int. Conf. Image Vis. Comput.*, 2017, pp. 20–25.

[45] Z. Xiao, Y. Gong, Y. Long, D. Li, X. Wang, and H. Liu, "Airport detection based on a multiscale fusion feature for optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1469–1473, Jul. 2017.

[46] B. Cai, Z. Jiang, H. Zhang, D. Zhao, and Y. Yao, "Airport detection using end-to-end convolutional neural network with hard example mining," *Remote Sens.*, vol. 9, no. 11, Nov. 2017, Art. no. 1198.

[47] R. Datla, V. Chalavadi, and C. K. Mohan, "A multimodal semantic segmentation for airport runway delineation in panchromatic remote sensing images," in *Proc. 14th Int. Conf. Mach. Vis.*, 2022, pp. 46–52.

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

[49] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

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

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

[52] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3562–3572.

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

[54] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

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

[56] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, vol. 34, pp. 12077–12090.

[57] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12116–12128.

Keyu Wu (Student Member, IEEE) received the B.S. degree in computer science and technology from Fuzhou University, Fuzhou, China, in 2021. She is currently working toward the M.S. degree in artificial intelligence with the Key Laboratory of Information Science of Electromagnetic Waves, Fudan University, Shanghai, China.

Her research interests include computer vision, remote sensing image processing, and deep learning application.

Feng Cai (Student Member, IEEE) received the B.S. degree in computer science and technology from Fuzhou University, Fuzhou, China, in 2021. He is currently working toward the M.S. degree in artificial intelligence with the Key Laboratory of Information Science of Electromagnetic Waves, Fudan University, Shanghai, China.

His research interests include computer vision, remote sensing image processing, and deep learning applications.

Haipeng Wang (Senior Member, IEEE) received the B.S. and M.S. degrees in mechanical and electronic engineering from the Harbin Institute of Technology, Harbin, China, in 2001 and 2003, respectively, and the Ph.D. degree in environmental systems engineering from the Kochi University of Technology, Kochi, Japan, in 2006.

He was a Visiting Researcher with the Graduate School of Information, Production and Systems, Waseda University, Fukuoka, Japan, in 2008. He is currently a Professor with the Key Laboratory of Electromagnetic Wave Information Science (MoE), Department of Communication Science and Engineering, School of Information Science and Engineering, Fudan University, Shanghai, China. His research interests include synthetic aperture radar image processing, deep learning, and its applications to SAR images.

Dr. Wang has been a member of the Technical Program Committee of the IEEE Geoscience and Remote Sensing Symposium (IGARSS) since 2011. He was a recipient of the Dean Prize of the School of Information Science and Engineering, Fudan University, in 2009, 2017, and 2021. He is an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.