Integration Transformer for Ground-based Cloud Image Segmentation

Shuang Liu, *Senior Member, IEEE,* Jiafeng Zhang, Zhong Zhang, *Senior Member, IEEE,* Xiaozhong Cao, and Tariq S. Durrani, *Life Fellow, IEEE*

***Abstract*—Recently, convolutional neural network (CNN) dominates the ground-based cloud image segmentation task, but disregards the learning of long-range dependencies due to the limited size of filters. Although Transformer-based methods could overcome this limitation, they only learn long-range dependencies at a single scale, hence failing to capture multi-scale information of cloud image. The multi-scale information is beneficial to ground-based cloud image segmentation, because the features from small scales tend to extract detailed information while features from large scales have the ability to learn global information. In this paper, we propose a novel deep network named Integration Transformer (InTransformer), which builds long-range dependencies from different scales. To this end, we propose the Hybrid Multi-head Transformer Block (HMTB) to learn multi-scale long-range dependencies, and hybridize CNN and HMTB as the encoder at different scales. The proposed InTransformer hybridizes CNN and Transformer as the encoder to extract multi-scale representations, which learns both local information and long-range dependencies with different scales. Meanwhile, in order to fuse the patch tokens with different scales, we propose Mutual Cross-Attention Module (MCAM) for the decoder of InTransformer which could adequately interact multi-scale patch tokens in a bidirectional way. We have conducted a series of experiments on large ground-based cloud detection database TLCDD and SWIMSEG. The experimental results show that the performance of our method outperforms other methods, proving the effectiveness of the proposed InTransformer.**

***Index Terms*—ground-based cloud image segmentation, CNN, Transformer.**

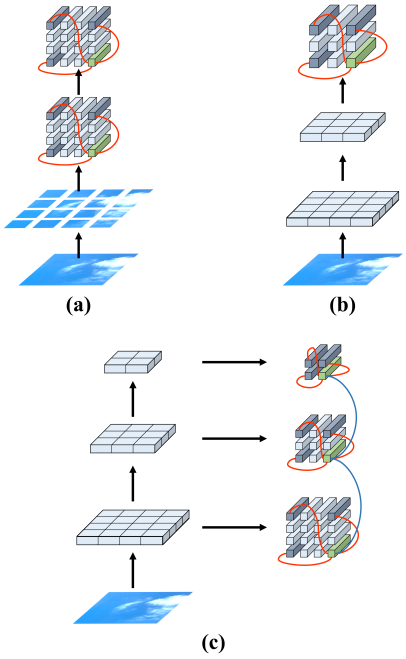

Fig. 1. Comparison of Transformer-based methods. (a) converts the input image into patch tokens, and (b) converts CNN's high-level feature maps into patch tokens. (c) learns long-range dependencies from multiple scale feature maps. The red curves denote the long-range dependencies among the patch tokens at the same scale, and the blue curves indicate cross-scale interaction. For simplicity, we only show a fraction of interaction among the patch tokens.

I. Introduction

CLOUDS are composed of water droplets, ice crystals, or a mix of them [1]–[3]. As a common natural phenomenon, clouds promote the earth's energy balance and the global hydrological cycle. Hence, accurate cloud observation is necessary for many applications such as weather forecasting, climate modeling, etc. There are two main ways for cloud observation: ground-based and satellite-based [4]–[6]. Local weather research and local environmental monitoring require ground-based cloud observation because it learns more information about the bottom of local clouds than satellite-based cloud observations [7], [8].

The WMO requires ground-based meteorological stations to record cloud data [9], and nowadays ground-based cloud observation heavily relies on human observers, whereas manual observation is subjective and may obtain inconsistent results from different observers for the same cloud image sample. Meanwhile, since shapes, structures and boundaries of clouds are variable, ground-based cloud observation is the challenging task. Hence, automatic ground-based cloud observation methods are eagerly needed. The existing ground-based cloud observation methods are roughly classified into three kinds according to their targets, i.e., cloud cover estimation [10]–[12], cloud type classification [13]–[15] and cloud base height measurement [16]–[18]. In this paper, we focus on cloud cover estimation, which is a key step of cloud observation [9]. The accurate cloud cover estimation is beneficial to weather

This work was supported by National Natural Science Foundation of China under Grant No. 62171321, Natural Science Foundation of Tianjin under Grant No. 22JCQNJC00010 and 20JCZDJC00180, and the Scientific Research Project of Tianjin Educational Committee under Grant No. 2022KJ011. (Corresponding author: Zhong Zhang.)

Shuang Liu, Jiafeng Zhang and Zhong Zhang are with Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China (e-mail: {shuangliu.tjnu, m648167095, zhong.zhang8848}@gmail.com).

Xiaozhong Cao is with the Meteorological Observation Centre, China Meteorological Administration, Beijing 100081, China (e-mail: xzhongcao@163.com).

Tariq S. Durrani is with the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow Scotland, UK (e-mail: t.durrani@strath.ac.uk).

prediction and climatic conditions understanding. Moreover, cloud cover evaluation also plays a guiding role on flight planning and aviation [19].

In order to implement cloud cover estimation, many segmentation methods are proposed so as to obtain the segmentation mask. The segmentation mask possesses the same resolution with the corresponding cloud image, and each pixel of the segmentation mask indicates that the corresponding pixel of the input image is cloud or sky as so to estimate the cloud cover [1], [20]. Since the rayleigh scattering of light causes the sky to be predominantly blue and the clouds to appear white, many traditional methods for ground-based cloud image segmentation [21]–[23] extract color information as the features for segmentation. Although these traditional methods are likely to improve performance, they still do not satisfy the actual demand.

Nowadays, various segmentation methods [24]–[26] based on convolutional neural network (CNN) [27]–[29] were proposed, because CNN has powerful representation capability. For example, Cheng et al. [30] proposed a novel two-branch deconvolutional network that can improve network performance and reduce computational complexity compared with traditional deconvolutional networks. These methods typically consist of an encoder-decoder architecture. The encoder extracts the representation features from the input image, and the decoder generates segmentation masks by enlarging the representation features. But because of the limited size of CNN filters, the learned representation features usually lack global contextual information, which is crucial for cloud image segmentation. To overcome this limitation, some methods directly modify the convolution filters including dilated convolution [31], convolution filter size increase [32], and spatial pyramid features [33]. Meanwhile, some other methods [34]–[36] combine shallow features with deep features to learn contextual information. However, these methods are difficult to model long-range dependencies because of the intrinsic local property in convolution filters.

Recently, researchers introduce Transformer [37], [38] into the computer vision field, which could learn long-range dependencies among image patches. As for image segmentation, some methods [39], [40] convert the input image or the high-level feature maps of CNN into patch tokens, and then learn long-range dependencies as shown in Fig. 1(a) and (b). However, these Transformer-based methods [41]–[43] only learn long-range dependencies at single scale, which results in insufficient dependency modeling for cloud images.

In this paper, we propose a novel method named Integration Transformer (InTransformer), which builds multi-scale long-range dependencies for ground-based cloud image segmentation as shown in Fig. 1(c). InTransformer is designed as an encoder-decoder architecture. The encoder of InTransformer consists of the CNN stage and the Transformer stage, which takes full advantage of CNN and Transformer to simultaneously learn local information and long-range dependencies at different scales. In the CNN stage, the cloud image is extracted by multiple convolutional layers to obtain feature maps with different scales. In the Transformer stage, we first transform the feature maps at each scale into the patch tokens. Many Transformer-based methods apply multi-head self-attention (MSA) for learning long-range dependencies, which is weak in mining multiple scale information. To overcome this limitation, we propose the Hybrid Multi-head Transformer Block (HMTB) to learn the multi-scale long-range dependencies, which its inputs are patch tokens with different scales.

After obtaining multi-scale patch tokens from the output of the encoder, the decoder of InTransformer aims to fuse them so as to mine cross-scale information. The cross-attention mechanism [44], [45] is usually applied to learn the interaction between different scales. However, it only learns the interaction in a unidirectional way, which results in insufficient interaction between different scales. Hence, we propose Mutual Cross-Attention Module (MCAM) to adequately interact different scale patch tokens in a bidirectional way. Specifically, the patch tokens from one scale are treated as Value (V) and Key (K) and the patch tokens from the other scale are regarded as Query (Q) and vice versa. To fuse the patch tokens from three scales, we design three MCAMs in the decoder of InTransformer. Finally, with the output of the last MCAM, we generate the segmentation mask.

The contribution of the proposed method is mainly concluded in three aspects:

1) We propose InTransformer for ground-based cloud image segmentation, which hybridizes CNN and HMTB as the encoder to learn long-range dependencies from multiple scales.

2) We propose MCAM in the decoder to adequately mine cross-scale information for accurate ground-based cloud image segmentation.

3) The proposed InTransformer outperforms other methods on the large-scale cloud detection database TLCDD [46] and SWIMSEG [47], which demonstrates the effectiveness of our method.

II. Related Work

A. Ground-based cloud image segmentation

As the growing demand for ground-based cloud image segmentation, many approaches are introduced based on various hand-crafted features. Long *et al*. [21] proposed the color-based fixed threshold method, which utilizes the ratio of red to blue (R/B) of pixels. Afterwards, Heinle *et al*. [48] replaced R/B with R-B as the color-based fixed threshold for cloud image segmentation task. For single-peaked and two-peaked cloud image segmentation, Li *et al*. [22] proposed a hybrid method with fixed and adaptive thresholds. Liu *et al*. [49] first generated high-confidence labels as hard-constrained seeds. Afterwards, graph cut was employed to segment cloud image.

Recently, many approaches are proposed to learn deep features using CNN for ground-based cloud image segmentation. Dev *et al*. [50] presented CloudSegNet which employs convolution, maxpooling and upsampling as a light architecture for daytime and nighttime cloud image segmentation. Xie *et al*. [51] proposed SegCloud by introducing pooling indices in the upsampling operation, which effectively restores the loss caused by pooling. Shi *et al*. [26] proposed EFCN to apply the skip connection and the histogram equalization in order

to improve the segmentation performance. Zhang *et al*. [1] integrated the attention mechanism and the multi-scale strategy into the encoder-decoder structure so as to learn discriminative information for segmentation performance improvement. Zhou *et al*. [52] proposed TL-DeepLabV3+ to offset a limited number of training cloud images by using transfer learning for cloud segmentation.

B. Transformer

Vaswani *et al*. [37] proposed Transformer which has a powerful ability to learn long-range dependencies, so it shows excellent performance in natural language processing (NLP). Recently, researchers introduce Transformer to computer vision tasks [53], and Transformer-based methods achieve comparable performance with CNN-based methods in image segmentation [39], [54], [55], object detection [56], classification [57], [58] and so on.

The performance of Vision Transformer (ViT) [38] in image classification demonstrated that Transformer-based methods can outperform state-of-the-art CNN-based methods. Zheng *et al*. [41] utilized sequence-to-sequence prediction to handle the segmentation task and proved the feasibility of Transformer for the segmentation task. Petit *et al*. [45] inserted Transformer into U-Net to obtain long-range dependencies. Furthermore, Hatamizadeh *et al*. [43] applied Transformer as the encoder and designed the corresponding decoder for 3D medical image segmentation.

Different from the above methods which learn long-range dependencies from single scale, we apply CNN and Transformer to learn the long-range dependency information from different scales. For multi-scale tokens from the encoder output, we design MCAM as the decoder to effectively enhance their interaction for adequately mining cross-scale information.

III. Approach

In this section, we first clarify the motivation of the proposed InTransformer. Afterwards, we present the overview of the proposed InTransformer. Finally, we introduce the encoder and the decoder of InTransformer in detail.

A. Motivation

The blurred boundaries and irregular shapes of clouds cause considerable challenges for ground-based cloud image segmentation. Recently, many CNN-based methods achieve promising performance because of powerful representation capability, but disregard the long-range dependencies. As shown in Fig. 2 (b), they mainly focus on local areas. Transformer-based methods only learn long-range dependencies at a single scale by using self-attention, and the visualization of Transformer-based methods is shown in Fig. 2 (c). Such design fails to capture multi-scale information.

The multi-scale information is beneficial to cloud image segmentation, because the features from small scales tend to extract detail information and the features from large scales learn global information. Hence, we propose InTransformer to build long-range dependencies from different scales for capturing more information. The visualization of our method is shown in Fig. 2 (d), where our method focuses on more cloud regions.

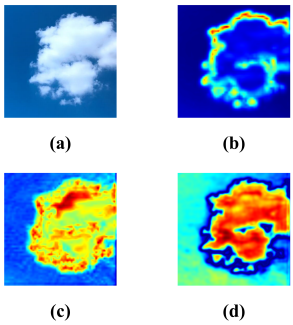

Fig. 2. Visualization of different methods. (a) a ground-based cloud image, (b) CNN-based methods, (c) Transformer-based methods with single scale, (d) our method.

B. Overall Framework

The main framework of InTransformer is shown in Fig. 3.

1) Encoder: The encoder of InTransformer consists of the CNN stage and the Transformer stage. In the CNN stage, we extract multi-scale feature maps from the ground-based cloud image. In the Transformer stage, we first transform the feature maps at each scale into the patch tokens. Afterwards, we employ HMTBs to build multi-scale long-range dependencies.

2) Decoder: We design MCAM as the decoder to fuse the patch tokens from different scales. The MCAM adequately interacts these patch tokens in a bidirectional way. We perform three MCAMs to yield cross tokens. Finally, we reshape the cross tokens, and then apply the convolution operations and the upsampling operations to generate the segmentation mask.

C. Encoder

The encoder is employed to extract the representation features from input cloud image. The representation features greatly effect the performance of the segmentation mask. The proposed InTransformer learns multi-scale representations via the CNN and the Transformer stages in the encoder, thereby mining local information and long-range dependencies, simultaneously.

1) CNN stage: In the CNN stage, we adopt stacked convolutional layers to learn multi-scale information from the ground-based cloud image. Specifically, we apply ResNet-50 (BiT) [38] as backbone to extract the feature maps with different scales. The structure of ResNet-50 (BiT) is shown in Table I. The size of the ground-based cloud map is $H \times W \times 3$, where H, W and 3 are the height, the width and the channel number, respectively. Then, we obtain the feature maps with different scales $f^i \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times 256 \cdot 2^{i-1}} (i = 1, 2, 3)$, i.e., $f^1 \in \mathbb{R}^{H/2 \times W/2 \times 256}$, $f^2 \in \mathbb{R}^{H/4 \times W/4 \times 512}$, and $f^3 \in \mathbb{R}^{H/8 \times W/8 \times 1024}$ corresponding to Stage1, Stage2, and Stage3.

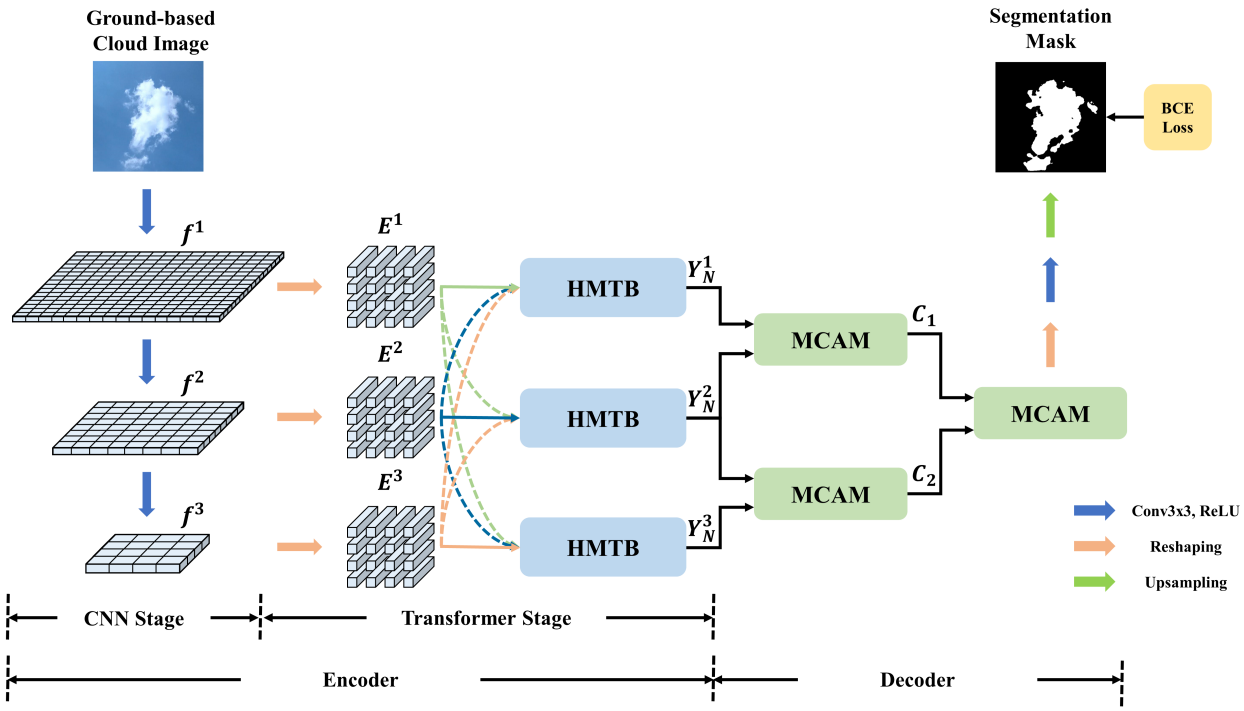

Fig. 3. The framework of the proposed InTransformer. We extract multi-scale feature maps in the CNN Stage, and then feed them into the Transformer stage to build multi-scale long-range dependencies. Afterwards, we interact the patch tokens with different scales generated by the encoder via MCAM and yield cross tokens for adequately mining cross-scale information. Finally, we generate the segmentation mask.

TABLE I
THE STRUCTURE OF RESNET-50 (BIT).

Name	Output Size	Filters	Padding
Conv1	$\frac{H}{2} \times \frac{W}{2}$	[7 × 7, 64], stride = 2	(3, 3)
Stage1	$\frac{H}{2} \times \frac{W}{2}$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 3$
Stage2	$\frac{H}{4} \times \frac{W}{4}$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 4$
Stage3	$\frac{H}{8} \times \frac{W}{8}$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 9$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 9$

2) Transformer stage: Due to the intrinsic properties of convolution, CNN tends to gradually reduce the resolution of feature maps in order to enlarge the receptive field size, which yields the feature maps with different scales. However, these multi-scale feature maps usually contain local information and lack long-range dependencies. Hence, we learn long-range dependencies in the Transformer stage from the feature maps with different scales via self-attention [38].

We first transform the feature maps at each scale extracted from the CNN stage into the patch tokens in the Transformer stage as shown in Fig. 4. Specifically, we first uniformly partition f^i into L patches and the patch size is $\frac{P}{2^{i-1}} \times \frac{P}{2^{i-1}}$. Therefore, we use different size $\frac{P}{2^{i-1}} \times \frac{P}{2^{i-1}}$ patch for feature maps with different scales. Here, $\frac{P}{2^{i-1}} \times \frac{P}{2^{i-1}}$ represents the patch size with the i-th scale, and P is a constant which controls the patch size and it is set to 16. We employ different sizes of patches for the feature maps from different scales, we obtain the multi-scale patch tokens with the same $L = (\frac{H}{2} \times \frac{W}{2})/(P \times P)$. A patch-level feature map with the size of $\frac{P}{2^{i-1}} \times \frac{P}{2^{i-1}} \times 256 \cdot 2^{i-1}$ is obtained by the patch on the feature maps.Then, we flatten it to a one-dimensional vector with the size of $1 \times \frac{P}{2^{i-1}} \cdot \frac{P}{2^{i-1}} \cdot 256 \cdot 2^{i-1}$. We employ the linear layer to project the dimensionality of the vector into D. Hence, we obtain the multi-scale patch tokens $E^i \in \mathbb{R}^{L \times D}$, where L and D are the length and hidden channel size of the patch tokens, respectively. When the feature maps are transformed into patch tokens, there is loss of pixel position information, and to compensate for the loss we add the learnable position embedding $G^i \in \mathbb{R}^{L \times D}$ into the patch tokens at each scale:

$$E^i = E^i + G^i \tag{1}$$

Nowadays, many Transformer-based methods apply multi-head self-attention (MSA) for learning long-range dependencies. MSA concatenates the outputs of multiple SAs, and each SA operation is performed on single scale patch tokens. Hence, MSA is weak in mining multiple scale information. To overcome this limitation, we propose the Hybrid Multi-head Transformer Block (HMTB), which could learn the multi-scale long-range dependencies. The HMTB consists of N hybrid multi-head transformer layers. The hybrid multi-head

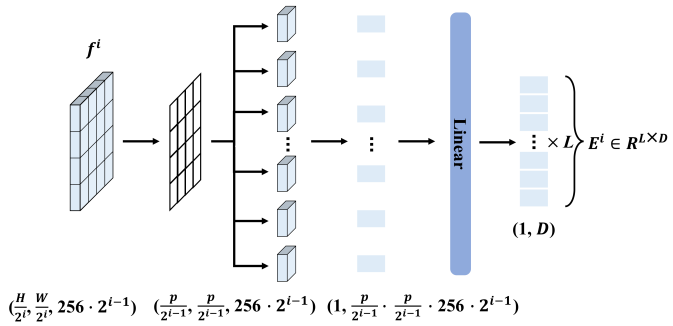

Fig. 4. The flowchart of transforming the feature maps into the patch tokens.

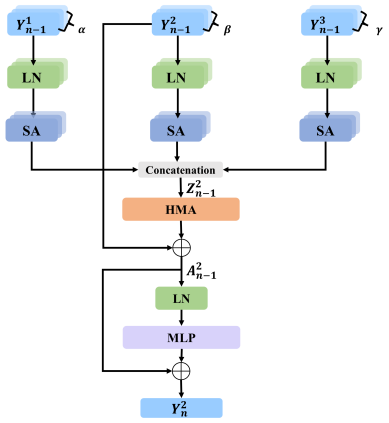

Fig. 5. The structure of hybrid multi-head transformer layer at the second scale.

transformer layers for the i-th scale take patch tokens from three scales as input, where the i-th scale patch tokens are used as the main branch. Patch tokens at each scale perform different number of SA operations.

We take the second scale as an example, and Fig. 5 shows the structure of one hybrid multi-head transformer layer at the second scale. The output of the n-th hybrid multi-head transformer layer at the second scale is formulated as:

$$Y^2_n = MLP(LN(A^2_{n-1})) + A^2_{n-1} \tag{2}$$

$$A^2_{n-1} = HMA(Z^2_{n-1}) + Z^2_{n-1} \tag{3}$$

$$\begin{aligned} Z^2_{n-1} =& Cat(SA_1(LN(Y^1_{n-1})); ...; SA_\alpha(LN(Y^1_{n-1})); \\ & ...; SA_\beta(LN(Y^2_{n-1})); ...; SA_\gamma(LN(Y^3_{n-1}))) \end{aligned} \tag{4}$$

$$M = \alpha + \beta + \gamma \tag{5}$$

where Y^1_{n-1}, Y^2_{n-1}, and Y^3_{n-1} are from three different scales of patch tokens, Cat () represents the concatenation operation, HMA represents the hybrid multi-head attention mechanism, α, β, and γ represent the head number of Y^1_{n-1}, Y^2_{n-1}, and Y^3_{n-1} respectively, and M is the total number of independent SA operations. Here, layer normalization (LN) is applied to normalize the patch tokens to accelerate the model convergence, multilayer perceptron (MLP) is used to reduce the number of parameters, the residual connections could alleviate the gradient vanishing or exploding problem.

The MLP consists of two fully connected (FC) layers with the neuron number of 3096 and 768, respectively. The first FC layer is followed by the Gaussian error linear units (GELU) and the dropout operation, and the second one is only followed by the dropout operation.

Furthermore, we propose hybrid multi-head attention mechanism to utilize spatial and multi-head information in order to selectively emphasize informative features and suppress less useful ones. Fig. 6 shows the framework of HMA. Here, g represents the decay factor, and g is a constant which controls the channel number and it is set to 4.

The m-th SA operation in the n-th hybrid multi-head transformer layer at the i-th scale is defined as:

$$SA_m(Z^i_{n-1}) = softmax(\frac{Q^i_m {K^i_m}^T}{\sqrt{d}})V^i_m \tag{6}$$

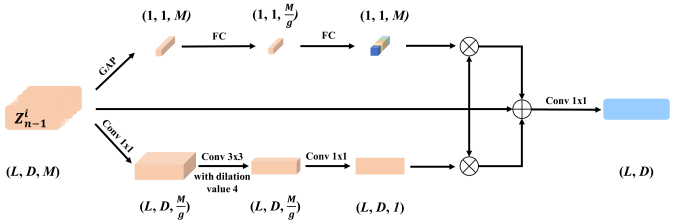

Fig. 6. The framework of HMA.

$$Q^i_m = Z^i_{n-1}W^i_{qm}, K^i_m = Z^i_{n-1}W^i_{km}, V^i_m = Z^i_{n-1}W^i_{vm} \tag{7}$$

where $W^i_{qm}, W^i_{km}, W^i_{vm} \in \mathbb{R}^{D\times d}$ are three independent trainable linear projection for the m-th SA operation.

D. Decoder

The outputs of the encoder are the patch tokens with different scales, and therefore the decoder is designed to fuse these patch tokens to obtain the segmentation mask. Although the existing cross-attention [44], [45] could fuse the patch tokens from multiple scales, this operation performs the unidirectional interaction, that is the patch tokens from one scale only treat as V, K or Q. This results in insufficient multi-scale fusion.

To solve the above limitation, we propose Mutual Cross-Attention Module (MCAM) in the decoder of InTransformer which could model the bidirectional interaction of the patch tokens from two scales. We design three MCAMs in the decoder of InTransformer as shown in Fig. 4. The first two MCAMs are used for the fusion of multi-scale patch tokens, and the last MCAM is used to interact their outputs.

Fig. 7 shows the framework of MCAM. The output of MCAM at the i-th MCAM ($i = 1, 2$) is defined as:

$$\begin{aligned} MCA(Y^i_N, Y^{i+1}_N) = & Cat(CA_1(Y^i_N, Y^{i+1}_N); ...; \\ & CA_M(Y^i_N, Y^{i+1}_N))W^i_d \end{aligned} \tag{8}$$

$$CA_m(Y^i_N, Y^{i+1}_N) = softmax(\frac{Q^{i+1}_m {K^i_m}^T}{\sqrt{d}})V^i_m \tag{9}$$

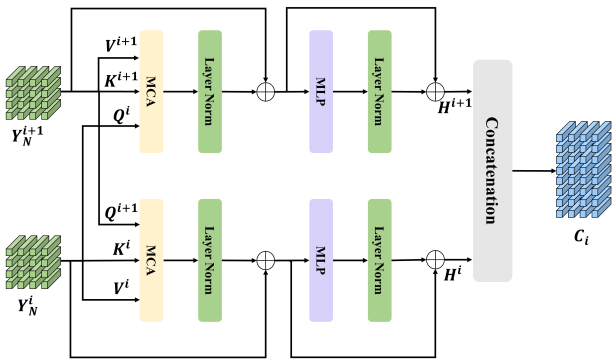

Fig. 7. The framework of MCAM.

$$Q_m^{i+1} = Y_N^{i+1} W_{qm}^{i+1}, K_m^i = Y_N^i W_{km}^i, V_m^i = Y_N^i W_{vm}^i \quad (10)$$

where $W_d^i \in \mathbb{R}^{M \cdot d \times D}$ is the trainable linear projection, and $W_{qm}^i, W_{km}^i, W_{vm}^i \in \mathbb{R}^{D \times d}$ are three independent trainable linear projection for the m-th CA operation. Q, K and V are from the patch tokens with different scales and are bidirectionally interacted via the two-stream structure of MCAM, so that the multi-scale information can be adequately mined.

We feed Y_N^1 and Y_N^2 into first MCAM to generate cross tokens $C_1 \in \mathbb{R}^{2L \times D}$. Meanwhile, Y_N^2 and Y_N^3 are fed into another MCAM to yield cross tokens $C_2 \in \mathbb{R}^{2L \times D}$. Finally, C_1 and C_2 are fed into MCAM to obtain the final cross tokens $C \in \mathbb{R}^{4L \times D}$. We change the cross tokens C to obtain the segmentation mask by using reshaping, convolution and upsampling operations. Finally, the binary cross-entropy (BCE) loss is used as the optimization objective, which forces the predicted distribution to gradually approximate the ground truth distribution. It is formulated as:

$$Loss = -\frac{1}{H}\sum_{i=1}^{H}[u_i log(g_i) + (1-u_i) log(1-g_i)] \quad (11)$$

where H is the total number of pixels in the cloud image. u_i and g_i indicate the ground-truth and predicted labels on the i-th pixel, respectively.

IV. Experiments

In order to comprehensively evaluate the performance of InTransforme, we conduct a series of experiments on the TJNU Large Scale Cloud Detection Database (TLCDD) and Singapore Whole sky IMaging SEGmentation Database (SWIMSEG). Firstly, we present the database and details of the implemented experiments. Afterwards, we demonstrate the effectiveness of InTransformer with different experiments, i.e., ablation studies and comparison experiments with other methods. Finally, we analyze several essential parameters of InTransformer.

A. Databases

TLCDD [46] consists of 5000 ground-based cloud images and corresponding ground-truth segmentation masks, which are carefully captured and annotated by professional technicians, i.e., meteorologists and cloud-related researchers. It is often used in the cloud image segmentation task. All images in the database are collected over two years in Tianjin, Hainan, Liaoning, Jiangsu, Sichuan, Gansu, Shandong, Hebei, and Anhui. This shows the diversity of cloud image samples, which provides convincing experimental results. The vision sensor captures the cloud images, and then we reshape the resolution to 512×512. There are 4208 cloud samples and 792 cloud samples on TLCDD for training and testing, respectively. Some ground-based cloud images and the corresponding ground-truth segmentation masks in TLCDD are shown in Fig. 8.

The SWIMSEG [47] dataset contains 1013 ground-based cloud images and the corresponding ground-truth segmentation masks, which were released by National University of Singapore.

B. Implementation Details and Evaluation Criteria

Before feeding the ground-based cloud images into InTransformer, we first preprocess these samples. The preprocessing operations include horizontal flipping, normalization, and random gaussian blur. Specifically, horizontal flipping is implemented with a probability of 0.5, normalization is performed by means and standard deviation values, and random gaussian blur is conducted with a probability of 0.5.

As for the optimizer, we employ stochastic gradient descent (SGD) [59]. The initialized learning rate and the momentum are set to 0.01 and 0.9 respectively, and the weight decay is set to 0.001. In the experiments, the epoch number is 150. Furthermore, we use the "poly" learning rate decay strategy [33], [60] to update the learning rate. Notably all experiments adopt the same parameter settings and the same data augmentation methods for fair comparison.

We treat ResNet-50 (BiT) [38] as backbone of the CNN stage. We set the total number of independent SA operations M in hybrid multi-head transformer layer to 16 and the number of independent CA operations of MCA to 12.

To quantitatively evaluate the performance of different methods, we adopt F-score (F), Recall (R), Precision (P), Accuracy (A) and intersection over union (IoU) evaluation metrics, which are commonly used in the image segmentation task.

C. Experimental Results

1) Ablation Studies: The advantages of the proposed InTransformer are that the encoder learns multi-scale long-range dependencies and MCAM is designed as the decoder to mine cross-scale information. To demonstrate their roles in InTransformer, we perform ablation experiments.

The C-C version. C-C only utilizes the CNN in the encoder and decoder. Its architecture is similar to U-Net [61]. Specifically, we use ResNet-50 (BiT) [38] to extract multi-scale features in the encoder, and then utilize high-scale features to obtain the segmentation mask in the decoder by skip connections.

The C+T-C version. We design C+T-C as the encoder-decoder architecture, which utilizes single scale patch tokens to build long-range dependencies. Specifically, we extract

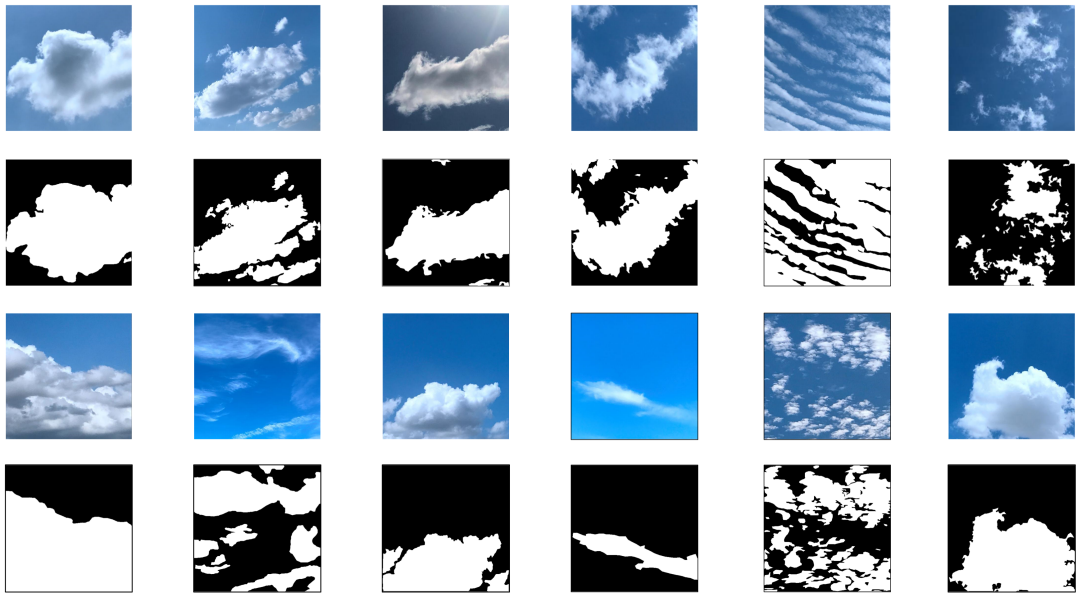

Fig. 8. Some cloud images and the corresponding ground-truth masks.

TABLE II
COMPARISON WITH DIFFERENT ABLATION METHODS

Methods	F	R	P	A	IoU
C-C	68.26	79.99	63.35	74.95	59.33
C+T-C	70.02	80.98	66.38	84.68	60.93
C+H-CA	75.25	82.13	74.33	88.07	68.17
C+T-MCAM	78.67	83.12	77.60	91.38	71.63
InTransformer	**79.34**	**83.67**	**79.27**	**92.15**	**73.84**

multi-scale feature maps in the CNN stage, and then apply the transformer layer to build long-range dependencies on the last scale CNN feature maps. Afterwards, we utilize reshaping, convolution and bilinear interpolation operations to obtain the segmentation mask. The architecture is similar to TransUNet [45], except that there is no skip connection.

The C+H-CA version. We implement C+H-CA to learn multi-scale long-range dependencies by using the encoder of InTransformer including the CNN stage and the Transformer stage. Meanwhile, we apply the CA module in the decoder to fuse the output of the encoder. Finally, the segmentation mask is generated.

The C+T-MCAM version. We design the C+T-MCAM encoder-decoder structure. The structure is similar to InTransformer, except that we replace the hybrid multi-headed transformer layer with the transformer layer to learn multi-scale long-range dependencies in the encoder.

Table II presents the experimental results of different ablation methods on TLCDD. We can draw some conclusions in the table. Firstly, the proposed method achieves the best performance in all five evaluation criteria, which demonstrates the effectiveness of different components in InTransformer. Secondly, all Transformer-based methods, i.e., C+T-C, C+H-CA, C+T-MCAM and InTransformer, outperform C-C. It indicates that learning long-range dependencies by using Transformer is beneficial for ground-based cloud image segmentation. Thirdly, C+H-CA, C+T-MCAM and InTransformer surpass C+T-C, which verifies multi-scale long-range dependencies could improve the segmentation performance. Fourthly, the performance of InTransformer is better than that of C+H-CA because MCAM adequately mines cross-scale information in the decoding process. Finally, InTransformer outperforms C+T-MCAM, demonstrating that the proposed HMTB can effectively mine scale information and perform multi-head feature recalibration.

2) Comparisons with Other Methods: We compare InTransformer with other methods. In order to comprehensively evaluate the performance, comparison methods are composed of traditional methods and deep learning methods. The traditional methods mainly apply color features as thresholds, i.e., R/B (0.6) [21], (B-R)/(B+R) (Otsu) [23], B-R (Otsu) [23], and B/R (Otsu) [23].

Deep learning methods consist of CNN-based methods and Transformer-based methods. CNN-based methods include FCN [62], CloudSegNet [50], U-Net [61], SegCloud [51], PSPNet [33], and FLA [63] which are widely used in the segmentation task. For Transformer-based methods, we apply TransUNet [45], DC-Swin [64], and UNetFormer [65] as comparison methods.

As shown in Table III, the performance of the proposed InTransformer outperforms other methods on all evaluation criteria on TLCDD. Specifically, it exceeds the second best experimental result by 6.56%, 1.95%, 6.88%, 5.68%, and 9.83% in Precision, Recall, F-score, Accuracy, and IoU, respectively. Furthermore, deep learning methods usually outperform traditional methods because deep learning employs multi-layer networks to mine discriminative features of ground-

TABLE III
COMPARISON WITH OTHER METHODS ON TLCDD

Methods	F	R	P	A	IoU
R/B (0.6) [21]	46.12	51.59	69.47	71.76	36.48
B/R (Otsu) [23]	57.26	77.48	55.98	67.72	45.39
B-R (Otsu) [23]	50.80	61.47	57.91	66.92	38.34
(B-R)/(B+R) (Otsu) [23]	59.11	69.60	63.00	73.61	47.23
FCN [62]	57.00	73.77	63.20	66.49	46.75
CloudSegNet [50]	57.79	77.61	64.46	64.59	47.78
U-Net [61]	67.32	80.43	68.80	74.13	58.16
SegCloud [51]	66.95	81.50	68.35	73.06	57.76
PSPNet [33]	67.00	77.75	68.74	78.64	57.43
FLA [63]	71.62	81.04	71.49	86.47	63.39
TransUNet [45]	70.37	81.03	72.39	84.93	64.01
DC-Swin [64]	71.06	81.49	70.53	85.52	62.55
UNetFormer [65]	72.78	81.72	69.67	85.68	63.87
InTransformer	**79.34**	**83.67**	**79.27**	**92.15**	**73.84**

TABLE IV
COMPARISON WITH OTHER METHODS ON SWIMSEG

Methods	F	R	P	A	IoU
R/B (0.6) [21]	73.28	74.01	81.07	74.37	65.36
B/R (Otsu) [23]	74.59	76.58	77.41	79.16	67.44
B-R (Otsu) [23]	78.89	78.38	84.60	80.95	67.42
(B-R)/(B+R) (Otsu) [23]	79.35	74.26	82.18	81.91	69.82
FCN [62]	78.38	80.17	76.42	82.31	73.88
CloudSegNet [50]	78.36	83.41	83.24	86.99	75.06
U-Net [61]	84.05	81.43	85.34	86.11	74.24
SegCloud [51]	84.86	86.13	84.89	86.88	76.29
PSPNet [33]	85.26	84.58	86.47	88.09	78.52
FLA [63]	81.56	83.71	87.94	85.97	72.44
TransUNet [45]	86.31	81.03	87.59	88.19	77.93
DC-Swin [64]	86.78	88.58	88.06	88.21	78.45
Unetformer [65]	86.53	88.09	89.95	89.01	79.60
InTransformer	**88.39**	**90.28**	**91.64**	**91.31**	**82.97**

based cloud images. Transformer-based methods are generally better than CNN-based methods due to building the long-range dependencies.

To demonstrate the generalization ability of the proposed InTransformer, we conduct a series of comparison experiments on SWIMSEG. The results of different comparison experiments are listed in Table IV. From the table, we can see that our method achieves the best results in all five metrics compared with state-of-the-art methods, which demonstrates the superiority of our method. Furthermore, many methods have lower performance on TLCDD than that of SWIMSEG, which indicates that TLCDD is more challenging for ground-based cloud image segmentation.

Furthermore, we analyze the computational complexity, training time, the number of parameters and the running time of different methods on TLCDD as shown in Table V. The Floating Point Operations (FLOPs) is usually applied to measure the computational complexity. Note that the traditional methods do not acquire the training time and do not compute the number of parameters due to the lack of training phases.

From table V, we can see that the training time generally increases with the number of parameters. The running time of CNN-based methods are more than those of the traditional methods. It is because CNN-based methods require large number of parameters to learn deep features. The Transformer-based methods have more running time than the CNN-based methods because the self-attention interacts all patch tokens to obtain the attention matrix.

The proposed InTransformer processes a ground-based cloud image with a running time of 72.06ms, that is our method could process 13 ground-based cloud images per second. It takes about 2 minutes for the acquisition device to collect a ground-based cloud image in the weather station. Hence, the proposed method could satisfy the actual application demand. It is reasonable to apply the proposed InTransformer for ground-based cloud image segmentation when trading off the performance and the running time. Note that the running time analysis of all comparison experiments is performed on a workstation equipped with NVIDIA RTX 3090Ti GPUs.

3) Visualization: We visualize some segmentation results of different methods as shown in Fig. 9. From the figure, we can see that our method generates more accurate segmentation masks than other methods, especially for illumination regions, thin clouds, thick clouds, etc.

Illumination is challenging for the ground-based cloud image segmentation task. From the red rectangle in the first row of the figure, we can see that InTransformer is more robust to illumination than other methods. The purple and green rectangles in Fig. 9 represent thick and thin cloud areas, respectively. The thick cloud areas appear darker due to the stacked cloud distribution. From the purple rectangle in the second row of the figure, the proposed InTransformer identifies complex thick clouds more easily than other methods. Thin cloud refers to a form of cloud that is light and somewhat transparent with low optical depth. Thin cloud has relatively low contrast between cloud and sky, and therefore it is challenging for segmentation. In terms of the green rectangles in the last four rows from Fig. 9, we can see that the proposed InTransformer could segment the thin cloud areas correctly compared to other methods. In a word, InTransformer could handle the hard segmented cloud pixels effectively.

4) Parameters Analysis: We study the influence of two important parameters on the performance of InTransformer in this section.

a) Number of HMTBs: We employ the HMTBs in the encoder to build long-range dependencies at each scale. We conduct the experiments with different number of HMTBs, and the results are shown in Fig. 10. From this figure, we can see that the performance increases with the number of HMTBs. Hence, the number of HMTBs is 3.

b) Number of Hybrid Multi-head Transformer Layers: We

TABLE V
THE TRAINING TIME, COMPUTATIONAL COMPLEXITY, NUMBER OF PARAMETERS, AND RUNNING TIME COMPARISON OF DIFFERENT METHODS

Methods	Training time (h)	Parameters (M)	FLOPs (G)	Running time (ms)
R/B (0.6)	-	-	-	10.57
B/R (Otsu)	-	-	-	12.12
B-R (Otsu)	-	-	-	11.84
(B-R)/(B+R) (Otsu)	-	-	-	13.25
FCN [62]	3.1	13.15	60.93	35.48
CloudSegNet [50]	4	15.46	57.31	39.73
U-Net [61]	4.6	17.27	160.51	49.54
SegCloud [51]	5.3	33.75	168.73	50.89
PSPNet [33]	6.7	46.70	184.73	52.26
FLA [63]	12.5	66.99	451.37	68.94
TransUNet [45]	16.9	107.48	60.87	73.53
DC-Swin [64]	18	111.72	51.72	76.15
Unetformer [65]	16.3	96.85	69.96	68.59
InTransformer	17.4	109.87	167.58	72.06

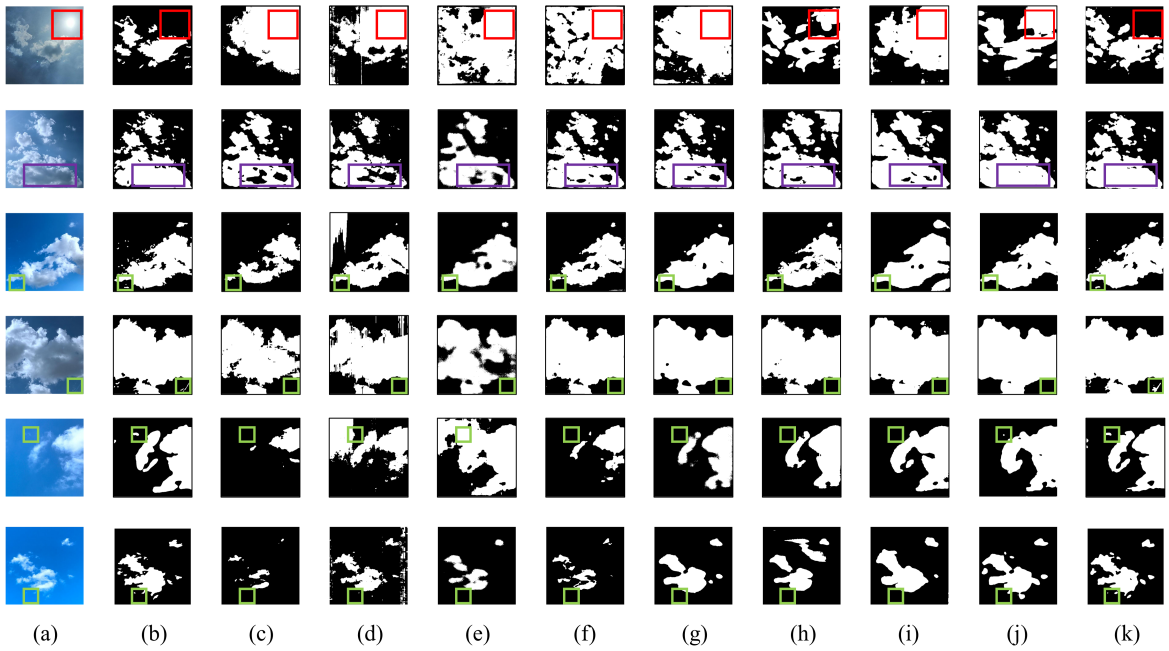

Fig. 9. The segmentation results of different methods. (a) Ground-based cloud images. (b) Ground-truth segmentation masks. (c) R/B (0.6). (d) (B-R)/(B+R)(Otsu). (e) FCN. (f) U-Net. (g) SegCloud. (h) TransUNet. (i) FLA. (j) DC-Swin. (k) InTransformer. The figure is best viewed in color with PDF magnification.

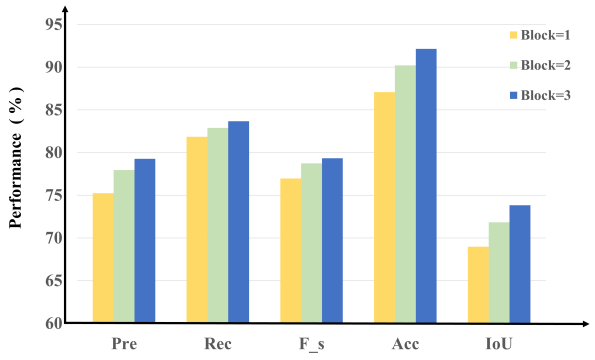

Fig. 10. The performance of InTransformer with different number of Hybrid Multi-head Transformer Blocks.

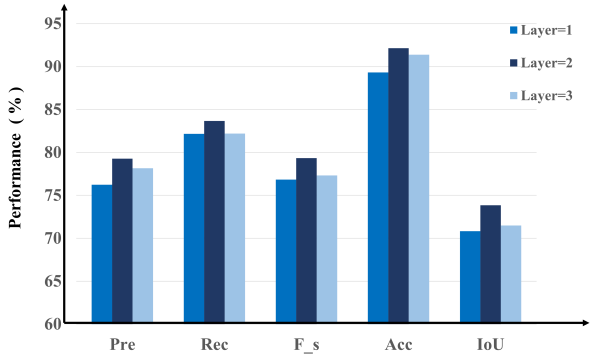

Fig. 11. The performance of InTransformer with different number of hybrid multi-head transformer layers.

TABLE VI
THE PERFORMANCE OF DIFFERENT METHODS WITH DIFFERENT NUMBER OF TRAINING EPOCHS ON TLCDD

Epochs	Methods	F	R	P	A	IoU
130	SegCloud	65.87	81.38	68.24	72.75	56.95
	FLA	70.11	80.13	70.54	84.89	62.47
	Unetformer	72.19	81.23	68.11	83.28	62.83
	InTransformer	78.15	82.81	77.93	90.76	71.91
140	SegCloud	66.95	81.50	68.35	73.06	57.76
	FLA	71.12	80.87	71.12	85.92	63.18
	Unetformer	72.78	81.72	69.67	85.68	63.87
	InTransformer	78.83	82.98	78.65	91.49	72.82
150	SegCloud	66.35	81.38	68.23	72.83	57.24
	FLA	71.62	81.04	71.49	86.47	63.39
	Unetformer	72.34	81.56	68.50	84.08	63.11
	InTransformer	79.34	83.67	79.27	92.15	73.84
160	SegCloud	66.10	81.26	68.12	72.15	57.09
	FLA	71.23	80.93	71.34	86.23	63.21
	Unetformer	71.79	81.05	67.68	83.03	62.58
	InTransformer	78.95	83.43	78.87	91.83	73.12
170	SegCloud	65.41	80.12	67.35	71.92	56.15
	FLA	70.85	80.38	70.79	85.53	62.98
	Unetformer	70.74	80.66	66.21	83.08	61.65
	InTransformer	78.38	83.24	78.32	91.21	72.31

study the number of hybrid multi-head transformer layers for each HMTB. The experimental results are shown in Fig. 11, where we can see that the performance is best when the number of hybrid multi-head transformer layers is set to 2.

c) Influence of Different Loss Functions: We compare the influence of different loss functions on the performance of the proposed InTransformer, as shown in Fig. 12. From the figure, we can see that the proposed InTransformer achieves the best performance when the BCE loss is treated as the loss function.

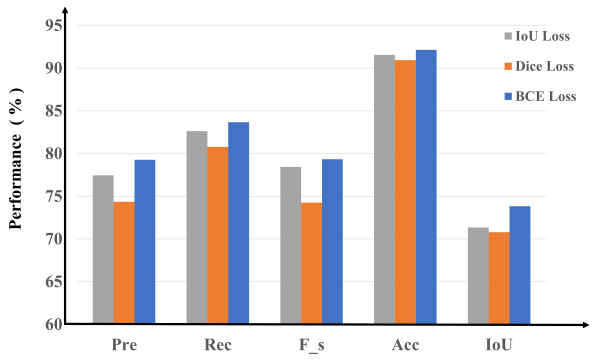

Fig. 12. The performance of InTransformer with different loss functions.

d) Number of Training Epochs: We evaluate the effect of different training epochs on the performance of different methods. As shown in Table VI, we can draw some conclusions. Firstly, the proposed InTransformer achieves the best performance when the training epoch is set to 150. Secondly, the performance of all methods decreases with increasing number of training epochs after reaching the peak.

e) Number of Training Samples: Table IV-C4 shows the performance of different methods with different proportions of training samples. From the table we can draw some conclusions. Firstly, the proposed InTransformer achieves the best performance on 100% training samples. Secondly, the performance of all methods improve with the increase of the number of training samples. Finally, the proposed InTransformer achieves the best performance in the same training sample proportion.

TABLE VII
THE PERFORMANCE OF DIFFERENT METHODS WITH DIFFERENT NUMBER OF TRAINING SAMPLE PROPORTIONS ON TLCDD

Training Samples (%)	Methods	F	R	P	A	IoU
60	SegCloud	59.68	74.68	53.65	61.63	49.66
	FLA	63.76	78.65	59.12	70.95	55.18
	Unetformer	64.13	78.75	59.82	72.36	55.89
	InTransformer	69.72	81.53	63.95	81.82	60.54
70	SegCloud	62.01	78.26	58.85	69.36	52.98
	FLA	66.98	79.58	63.15	75.58	57.25
	Unetformer	67.22	80.42	62.23	77.14	57.36
	InTransformer	73.78	81.78	68.75	85.76	65.18
80	SegCloud	65.35	79.05	63.84	71.92	55.35
	FLA	70.26	80.07	69.95	83.89	61.12
	Unetformer	68.96	80.76	66.26	78.98	59.58
	InTransformer	75.35	82.55	72.13	88.08	68.18
90	SegCloud	66.04	80.19	64.86	72.65	56.03
	FLA	70.89	80.56	70.92	85.13	62.86
	Unetformer	70.35	81.56	68.98	81.34	60.56
	InTransformer	77.25	82.92	77.26	90.86	71.12
100	SegCloud	66.95	81.50	68.35	73.06	57.76
	FLA	71.62	81.04	71.49	86.47	63.39
	Unetformer	72.78	81.72	69.67	85.68	63.87
	InTransformer	79.34	83.67	79.27	92.15	73.84

TABLE VIII
THE PERFORMANCE OF INTRANSFORMER WITH DIFFERENT RATIOS OF α, β, AND γ IN HYBRID MULTI-HEAD TRANSFORMER LAYER.

α:β:γ	F	R	P	A	IoU
1:1:1	78.28	82.20	77.76	91.06	71.20
1:2:1	**79.34**	**83.67**	**79.27**	**92.15**	**73.84**
1:4:1	78.15	82.67	77.22	90.88	71.11

f) Different Ratios of α, β and γ in the Hybrid Multi-head Transformer Layer: α, β and γ represent the head number of Y_{n-1}^{i-1}, Y_{n-1}^{i}, and Y_{n-1}^{i+1}, respectively. In order to investigate the influence of the performance of InTransformer with different ratios of ,,and in the hybrid multi-head transformer layer, we perform experiments, and the results are listed in Table VIII. From this table, we can see that the performance of InTransformer is best when the ratio of α, β, and γ is set to $1:2:1$.

V. CONCLUSION

In this paper, we have proposed InTransformer for ground-based cloud image segmentation. Specifically, the encoder of InTransformer consists of the CNN stage and the Transformer stage. We propose HMTB to replace the tansformer layer in Transformer stage, which allows the proposed InTransformer to take full advantage of CNN and Transformer to simultaneously learn local information and long-range dependency information at different scales. Meanwhile, we design MCAMs as the decoder of InTransformer to fuse the multi-scale features from the encoder and mine cross-scale information. Finally, we apply the convolution and upsampling operations to obtain

the segmentation mask. We have conducted the experiments on TLCDD and SWIMSEG, and the experimental results have demonstrated the effectiveness of the proposed InTransformer. In the future, we will apply the proposed InTransformer to the satellite-based observation to demonstrate the generalization ability of the proposed method. Furthermore, we will study the lightweight Transformer-based model so as to reduce the complexity for ground-based cloud image segmentation.

REFERENCES

[1] Z. Zhang, S. Yang, S. Liu, B. Xiao, and X. Cao, "Ground-based cloud detection using multiscale attention convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[2] Q. Li, W. Lu, J. Yang, and J. Z. Wang, "Thin cloud detection of all-sky images using markov random fields," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 3, pp. 417–421, 2012.

[3] S. Liu, L. Duan, Z. Zhang, X. Cao, and T. S. Durrani, "Multimodal ground-based remote sensing cloud classification via learning heterogeneous deep features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7790–7800, 2020.

[4] C. Shi, Y. Zhou, B. Qiu, D. Guo, and M. Li, "Cloudu-net: A deep convolutional neural network architecture for daytime and nighttime cloud images segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 10, pp. 1688–1692, 2021.

[5] L. Ye, Z. Cao, and Y. Xiao, "Deepcloud: Ground-based cloud image categorization using deep convolutional features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5729–5740, 2017.

[6] Y. Wang, C. Wang, C. Shi, and B. Xiao, "A selection criterion for the optimal resolution of ground-based remote sensing cloud images for cloud classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1358–1367, 2019.

[7] S. Dev, F. M. Savoy, Y. H. Lee, and S. Winkler, "Nighttime sky/cloud image segmentation," in *Proceedings of the IEEE International Conference on Image Processing*, 2017, pp. 345–349.

[8] L. Ye, Z. Cao, Y. Xiao, and Z. Yang, "Supervised fine-grained cloud detection and recognition in whole-sky images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7972–7985, 2019.

[9] "Manual on the observation of clouds and other meteors," *Available online: https://cloudatlas.wmo.int/en/home.html*, 2017.

[10] Z. Zhang, S. Yang, S. Liu, X. Cao, and T. S. Durrani, "Ground-based remote sensing cloud detection using dual pyramid network and encoder–decoder constraint," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022.

[11] S. Liu, J. Zhang, Z. Zhang, X. Cao, and T. S. Durrani, "Transcloudseg: Ground-based cloud image segmentation with transformer," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6121–6132, 2022.

[12] S. Dev, Y. H. Lee, and S. Winkler, "Color-based segmentation of sky/cloud images from ground-based cameras," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 1, pp. 231–242, 2017.

[13] X. Li, B. Qiu, G. Cao, C. Wu, and L. Zhang, "A novel method for ground-based cloud image classification using transformer," *Remote Sensing*, vol. 14, no. 16, pp. 3978–3979, 2022.

[14] A. Taravat, F. Del Frate, C. Cornaro, and S. Vergari, "Neural networks and support vector machine algorithms for automatic cloud classification of whole-sky ground-based images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 666–670, 2015.

[15] W. Zhu, T. Chen, B. Hou, C. Bian, A. Yu, L. Chen, M. Tang, and Y. Zhu, "Classification of ground-based cloud images by improved combined convolutional network," *Applied Sciences*, vol. 12, no. 3, pp. 1570–1571, 2022.

[16] M. C. Allmen and W. P. Kegelmeyer Jr, "The computation of cloud-base height from paired whole-sky imaging cameras," *Journal of Atmospheric and Oceanic Technology*, vol. 13, no. 1, pp. 97–113, 1996.

[17] E. Kassianov, C. N. Long, and J. Christy, "Cloud-base-height estimation from paired ground-based hemispherical observations," *Journal of Applied Meteorology*, vol. 44, no. 8, pp. 1221–1233, 2005.

[18] N. B. Blum, B. Nouri, S. Wilbert, T. Schmidt, O. Lünsdorf, J. Stührenberg, D. Heinemann, A. Kazantzidis, and R. Pitz-Paal, "Cloud height measurement by a network of all-sky imagers," *Atmospheric Measurement Techniques*, vol. 14, no. 7, pp. 5199–5224, 2021.

[19] S. A. Rajini and G. Tamilpavai, "Classification of cloud/sky images based on knn and modified genetic algorithm," in *Proceedings of the International Conference on Intelligent Computing and Communication for Smart World*, 2018, pp. 1–8.

[20] W. Li, Z. Zou, and Z. Shi, "Deep matting for cloud detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8490–8502, 2020.

[21] C. N. Long, J. M. Sabburg, J. Calbó, and D. Pagès, "Retrieving cloud characteristics from ground-based daytime color all-sky images," *Journal of Atmospheric and Oceanic Technology*, vol. 23, no. 5, pp. 633–652, 2006.

[22] Q. Li, W. Lu, and J. Yang, "A hybrid thresholding algorithm for cloud detection on ground-based color images," *Journal of Atmospheric and Oceanic Technology*, vol. 28, no. 10, pp. 1286–1296, 2011.

[23] J. Yang, W. Lu, Y. Ma, and W. Yao, "An automated cirrus cloud detection method for a ground-based cloud image," *Journal of Atmospheric and Oceanic Technology*, vol. 29, no. 4, pp. 527–537, 2012.

[24] C. Shi, Y. Zhou, and B. Qiu, "Cloudraednet: residual attention-based encoder–decoder network for ground-based cloud images segmentation in nychthemeron," *International Journal of Remote Sensing*, vol. 43, no. 6, pp. 2059–2075, 2022.

[25] D. Makwana, S. Nag, O. Susladkar, G. Deshmukh, S. C. Teja R, S. Mittal, and C. K. Mohan, "Aclnet: an attention and clustering-based cloud segmentation network," *Remote Sensing Letters*, vol. 13, no. 9, pp. 865–875, 2022.

[26] C. Shi, Y. Zhou, B. Qiu, J. He, M. Ding, and S. Wei, "Diurnal and nocturnal cloud segmentation of all-sky imager (asi) images using enhancement fully convolutional networks," *Atmospheric Measurement Techniques*, vol. 12, no. 9, pp. 4713–4724, 2019.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[29] C. Cheng, H. Li, J. Peng, W. Cui, and L. Zhang, "Deep high-order tensor convolutional sparse coding for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[30] C. Cheng, H. Li, and L. Zhang, "Two-branch deconvolutional network with application in stereo matching," *IEEE Transactions on Image Processing*, vol. 31, pp. 327–340, 2022.

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[32] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4353–4361.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[34] Q. Zhou, Z. Qu, and F.-r. Ju, "A lightweight network for crack detection with split exchange convolution and multi-scale features fusion," *IEEE Transactions on Intelligent Vehicles*, pp. 1–11, 2022.

[35] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, pp. 3–11.

[36] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the International Conference on 3D Vision*, 2016, pp. 565–571.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, 2017.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[39] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[40] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[41] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[42] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[43] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[44] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, and P. Luo, "Multi-compound transformer for accurate biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 326–336.

[45] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," in *Proceedings of the International Workshop on Machine Learning in Medical Imaging*, 2021, pp. 267–276.

[46] *https://github.com/zhongzhang8848/TJNU-Large-Scale-Cloud-Detection-Database*.

[47] S. Dev, Y. H. Lee, and S. Winkler, "Color-based segmentation of sky/cloud images from ground-based cameras," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 1, pp. 231–242, 2017.

[48] A. Heinle, A. Macke, and A. Srivastav, "Automatic cloud classification of whole sky images," *Atmospheric Measurement Techniques*, vol. 3, no. 3, pp. 557–567, 2010.

[49] S. Liu, Z. Zhang, B. Xiao, and X. Cao, "Ground-based cloud detection using automatic graph cut," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 6, pp. 1342–1346, 2015.

[50] S. Dev, A. Nautiyal, Y. H. Lee, and S. Winkler, "Cloudsegnet: A deep network for nychthemeron cloud image segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 12, pp. 1814–1818, 2019.

[51] W. Xie, D. Liu, M. Yang, S. Chen, B. Wang, Z. Wang, Y. Xia, Y. Liu, Y. Wang, and C. Zhang, "Segcloud: a novel cloud image segmentation model using a deep convolutional neural network for ground-based all-sky-view camera observation," *Atmospheric Measurement Techniques*, vol. 13, no. 4, pp. 1953–1961, 2020.

[52] Z. Zhou, F. Zhang, H. Xiao, F. Wang, X. Hong, K. Wu, and J. Zhang, "A novel ground-based cloud image segmentation method by using deep transfer learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[53] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *arXiv preprint arXiv:2111.06091*, 2021.

[54] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[55] F. Zhu, Y. Zhu, L. Zhang, C. Wu, Y. Fu, and M. Li, "A unified efficient pyramid transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2667–2677.

[56] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[57] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[58] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 558–567.

[59] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1139–1147.

[60] F. Zhang, Y. Chen, Z. Li, Z. Hong, J. Liu, F. Ma, J. Han, and E. Ding, "Acfnet: Attentional class feature network for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6798–6807.

[61] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, pp. 234–241.

[62] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[63] Q. Song, J. Li, C. Li, H. Guo, and R. Huang, "Fully attentional network for semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2280–2288.

[64] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[65] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

Shuang Liu (Senior Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014.

She is a Professor with Tianjin Normal University, Tianjin, China. She has published over 60 articles in major international journals and conferences. Her research interests include remote sensing, computer vision, and deep learning.

Jiafeng Zhang is a master student at Tianjin Normal University, Tianjin, China. His research interests include ground-based cloud analysis and deep learning.

Zhong Zhang (Senior Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

He is a Professor with Tianjin Normal University, Tianjin, China. He has published about 110 articles in international journals and conferences such as the IEEE Transactions on Geoscience and Remote Sensing, IEEE Transactions on Fuzzy Systems, Pattern Recognition, IEEE Transactions on Circuits Systems Video Technology, IEEE Transactions on Information Forensics and Security, Signal Processing (Elsevier), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), AAAI Conference on Artificial Intelligence (AAAI), and International Conference on Image Processing (ICIP). His research interests include remote sensing, computer vision, and deep learning.

Xiaozhong Cao is a Professor at Meteorological Observation Centre in China Meteorological Administration. He received the Ph.D. degree in automatic control theory and application from Institute of Automation, Chinese Academy of Sciences in 1996. His current research interests include the theory of meteorological observation and climate change, and the automatic meteorological observation.

Tariq S. Durrani is Research Professor at University of Strathclyde, Glasgow Scotland. His research covers AI, Signal Processing and Technology Management. He has authored 350 publications; supervised 45 PhDs. He is a Fellow of the: IEEE, UK Royal Academy of Engineering, Royal Society of Edinburgh, IET, and the Third World Academy of Sciences. He was elected Foreign Member of the Chinese Academy of Sciences and the US National Academy of Engineering in 2021 and 2018, respectively.