Binary quantization vision transformer for effective segmentation of red tide in multi-spectral remote sensing imagery.

XIE, Y., HOU, X., REN, J., ZHANG, X., MA, C. and ZHENG, J.

2025

Binary Quantization Vision Transformer for Effective Segmentation of Red Tide in Multi-spectral Remote Sensing Imagery

Yefan Xie, Xuan Hou, Jinchang Ren*, *Senior Member, IEEE,* Xinchao Zhang, Chengcheng Ma,and Jiangbin Zheng*

***Abstract*—As a global marine disaster, red tides pose serious threats to marine ecology and the blue economy, making their monitoring crucial for preventing harmful algal blooms and protecting the marine environment. In this study, satellite remote sensing was utilized to provide timely, large-scale, and continuous observation capabilities, overcoming the high cost and spatial and temporal limitations of in-situ monitoring. However, existing remote sensing-based methods often exhibit coarse segmentation granularity and suffer from high computational complexity. To overcome these challenges, we propose a novel bi-modal multi-spectral dynamic offset binary quantization visual transformer (DoBi-SWiP-ViT) that utilizes the ViT for global feature aggregation and parameter quantization for efficient segmentation. With the Bi-modal Swin-ViT with Unified Perceptual Parsing architecture, our model integrates data from multiple spectral bands to achieve fine-grained segmentation of large-scale remote sensing images. Additionally, we introduce a dynamic magnitude offset binary quantization ViT block to reduce the parameter redundancy and improve the computational efficiency. In addition, we validated the performance of our model through extensive comparative experiments on high-resolution imagery datasets of sea surface red tides collected from different satellite platforms. The results show that our proposed DoBi-SWiP-ViT has significantly improved the mean accuracy (mAcc) of the segmentation results. For the two test areas acquired from different satellite platforms, the improvements are 8.78% and 10.18%, respectively. This has demonstrated the superior performance of our model in detecting the red tides from high-resolution visible images, highlighting its effectiveness in capturing complex patterns and subtle features in multi-spectral imagery.**

***Index Terms*—Red Tide, Segmentation, Binary Quantization, Vision Transformer, Remote Sensing, Multi-spectral Imagery**

I. Introduction

Y. Xie and J. Zheng are with the School of Computer Science, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, Northwestern Polytechnical University, Xi'an 710072, China.

X. Hou is with the School of Computer Science, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Department of Computer Science, Faculty of Business and Physical Sciences, Aberystwyth University, Aberystwyth SY23 3DB, U.K.

J. Ren is with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K.

X. Zhang is with the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, 450001, China.

C. Ma is with the School of Software, Northwestern Polytechnical University, Shaanxi, 710129, China.

*Corresponding authors are J. Ren (jinchang.ren@ieee.org) and J. Zheng (zhengjb@nwpu.edu.cn))

As a global marine disaster, red tides pose significant risks to the marine ecology, aquaculture and blue economy. Therefore, monitoring red tides is crucial for preventing and reducing the hazards of harmful algal blooms, which is essential for protecting the marine environment. Traditional on-site monitoring collects data of marine environmental elements through fixed-point observation [1] and mobile observation [2]. Although these methods tend to have local spatial and temporal continuity and high accuracy, they are limited by the reliability of sensors and the trustworthiness of data [3], [4]. Considering the diverse spatial distribution of red tides and their rapid rate of change, they often fail to meet the requirements of large-scale timely monitoring. Satellite remote sensing has the advantages of timely, large-scale and continuous observation [5], [6], which is conducive to the rapid location of hazardous areas and impact levels of red tide, where the accurate location of such areas can also guide the ground staff to advance the response speed of protection and specific actions to mitigate the hazard [7].

Conventional remote sensing methods indicate the presence of red tide by identifying changes in water colour caused by algal blooms, using ocean colour data captured from platforms such as Landsat, MODIS and Sentinel satellites. Indices such as the Red Tide Index (RI) [7], P. Donghaiense Index (PDI), and Diatom Index (DI) [3], as well as a series of improved RI algorithms [9], [10], have been developed for this purpose. Alternatively, spectral analysis methods use specific spectral bands to detect chlorophyll-A, bio-optical properties of seawater, or fluorescence line height (FLH) as alternative indicators to determine the presence of red tides [11]–[13]. In recent years, deep learning methods have significantly advanced the intelligent interpretation of remote sensing images by fully leveraging their spectra, textures, and fine features[14]–[18]. The encoder-decoder architecture proposed in U-Net [19] has been beneficial, effectively capturing local features at different scales, enhancing spatial detail and structural recovery in images. However, the local-focused features limit its ability to represent the global context, which is crucial when processing large-scale images. Furthermore, the high parameter redundancy inherent in deep neural networks has led to to significant computational costs [14], [20].

To address these challenges, we propose a binary quantization Vision Transformer (ViT) for red tide segmentation in multi-spectral satellite imagery, by using a feature fusion scheme with a unified perceptual parsing architecture to further

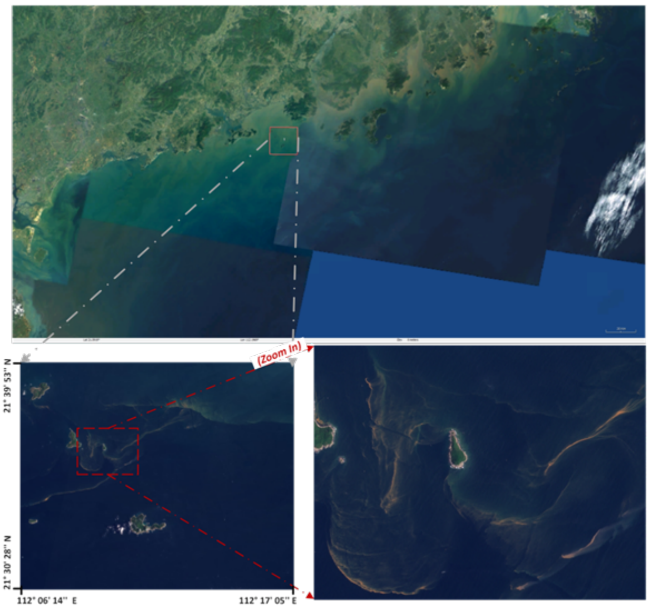


Fig. 1. Example of colour imagery of the red tide in the monitoring area (Area 2) from the PlanetScope satellite [8]. The top portion of the image is colour imagery at a large scale, processed by stitching sequentially collected imagery by coordinate correction. The bottom left image is one of the imagery from the experimental dataset, with the red dashed box showing an example image after the cropped process.

enhance the fine feature extraction capability of the model. To address the issues associated with the large model complexity of the ViT structure and the unified perceptual structure, we propose a dynamic magnitude offsets binary quantised ViT (DoBi-ViT) block structure to reduce the parameters. Moreover, we conducted extensive comparative experiments on high-resolution imagery datasets of sea surface red tides collected from various satellite platforms to validate our method. The results demonstrate the superior performance of our model in red tide segmentation, highlighting its ability to capture complex patterns and subtle features in multi-spectral imagery.

The major contributions of the model can be highlighted as follows.

1) We propose a bi-modal Vision Transformer with unified perceptual parsing architecture, significantly enhancing the ViT to extract fine-grained semantic details and improve its performance in high-resolution and large-scale scenes.
2) To address the model complexity associated with ViT structures and unified perceptual architectures, we have designed a dynamic offset binary-ViT block structure. This design reduces the overall parameter footprint of the model and enhances its efficiency.
3) A high-resolution dataset of red tide imagery was collected from three public satellite platforms. Unlike existing methods that crop and split data from single imagery, it includes scenes from different sea areas, times, and outbreak scales for experimental training and validation. Our model demonstrated superior performance and was validated through extensive comparative experiments.

II. Related Work

A. Deep Learning Based Red Tide Segmentation

Red tide segmentation and monitoring methods can be broadly categorized into in-situ surveys and remote sensing-based techniques [21]. Given its extensive spatial coverage and short revisit intervals, remote sensing technology has become a pivotal tool for red tide monitoring and segmentation [22]–[24]. Due to the sensitivity of spectral-based monitoring methods to the associated segmentation thresholds and the advancements in ground resolution of remote sensing, various deep learning techniques have been applied for red tide segmentation [25]. Jiang et al. [20] employed a deep confidence network model to detect red tide using airborne hyperspectral remote sensing data. Li et al. [15] proposed a red tide extraction method based on deep learning with Unmanned Aerial Vehicle (UAV) remote sensing images. Lee et al. [26] combined the high loss sample mining method with the ResNet and Geostationary Ocean Color Imager (GOCI) image data for red tide segmentation. Zhao et al. [25] proposed a red tide segmentation method based on the U-Net using HY-1D satellite Coastal Zone Imager (CZI) data. Shen et al. [27] proposed a progressive CNN-transformer alternating reconstruction network (PCTARN), which introduces the global-local dynamic priors and stacks lightweight convolutional modules at different levels to efficiently enhance the reconstruction quality of red tide hyperspectral data, thereby facilitating red tide species identification. However, existing methods often encounter challenges such as insufficient emphasis on the global features, excessive granularity in semantic segmentation, and parameter redundancy, which hinder both the efficiency and accuracy of red tide monitoring. To address these issues, we propose an alternative approach that reduces the number of parameters in ViT-based models through parameter quantization. This approach significantly reduces model parameters while maintaining the strong global feature extraction capability, enabling high-precision red tide segmentation in large-scale remote sensing images.

B. Vision Transformer (ViT)

The Transformer was initially proposed for machine translation tasks [28]. In the field of Natural Language Processing (NLP), Transformer-based approaches have achieved state-of-the-art performance across various tasks. Around the same period, before the introduction of Transformer architecture into the field of Computer Vision (CV), researchers had already recognised the potential of attention mechanisms for enhancing the capabilities of neural networks. These efforts included employing self-attention to improve the performance of conventional Convolution Neural Networks (CNNs) by enabling them to adaptively focus on features of interest within an image. In [29], skip connections with additive attention gates were integrated into a U-shaped architecture for medical image segmentation. However, the core component of the model was still based on convolutional constructions, indicating that it remained fundamentally CNN-based.

Driven by the success of the Transformer across various fields and the application of attention mechanisms in CV, a pioneering Vision Transformer (ViT) was introduced in [30], marking the first image recognition model built upon the Transformer mechanism. Compared with CNN-based methods, a notable drawback of ViT is that it requires pre-training on its own large dataset. To mitigate the challenges associated with training ViT, Deit [31] proposed several training strategies that enable ViT to perform effectively on ImageNet. Recently, several additional works have been made based on ViT [32]–[34]. Among these, it is worth highlighting that an efficient and effective hierarchical ViT, called Swin Transformer, is proposed as a vision backbone in [32]. Leveraging the shifted windows mechanism, the Swin Transformer achieved state-of-the-art performance in various vision tasks, including image classification, object detection, and semantic segmentation. Swin-UNet [35] is the first pure Transformer-based U-shaped architecture, comprising an encoder, bottleneck, decoder, and skip connections. The Swin Transformer block constitutes the core of the encoder, bottleneck, and decoder. The process begins by dividing input images into non-overlapping patches, each treated as a token. These tokens are processed by the Transformer-based encoder to extract deep feature representations. The decoder, equipped with a patch-expanding layer, upsamples the extracted features and integrates them with multi-scale features from the encoder via skip connections. This integration restores the spatial resolution of the feature maps, enabling precise segmentation. However, a common limitation of ViTs is their high computational demand, which can pose a bottleneck in resource-constrained environments.

C. Parameter Quantization

With the advancement of deep neural networks (DNNs), the number of parameters and computational costs have grown substantially. To tackle the challenges of deploying large models on resource-constrained platforms, parameter quantization has emerged as a widely adopted solution. This technique compresses DNNs by replacing weights and activations with low-bit representations, enabling significant reductions in model size while preserving the original network structure.[36]–[43].

The binary quantization paradigm represents an extreme form of parameter quantization, where both the weights and activations in DNNs are constrained to 1-bit representations. Compared to their full-precision counterparts, binary quantization replaces multiplication operations with bitwise operations, offering the potential to reduce network size by a factor of 32. Xnor-Net [42] proposed that a real-valued scaling factor could be implemented to each output channel of the binary convolution for accuracy improvement, which has become a common practice for binary networks. Bi-real-Net [44] argued that the real-valued skip connection presents the basis of binary networks, and they suggested converting the down-sampling layer to full precision values, trading negligible computational complexity for improved accuracy. Xnor-Net++ [45] proposed using PReLU to smooth the gradient approximation. Wang et al. [46] proposed leveraging reinforcement learning to model channel correlations, enabling better preservation of the sign output of the convolution. Ding et al. [47] introduced a set of regularisers into the loss function to constrain activation values and ensure proper gradient flow. Alizadeh et al. [48] perform validation tests on the impact aspects of gradient clipping and batch-norm momentum. Xu et al. [49] proposed using a rectified clamp unit (ReCU) to leverage the information entropy and quantization error relationships in the Binary Neural Network (BNN). RB-Net [50] was proposed to achieve a balance between the accuracy and efficiency in object classification tasks by introducing reshaped point-wised convolution (RPC) and integrated balanced activation (BA). Despite these advancements, binary quantization continues to face significant accuracy degradation, particularly in pixel-level segmentation tasks. This will be addressed in our model as detailed in the next Section.

III. The Proposed Method

The overall structure of the proposed bi-modal multi-spectral dynamic offset binary quantization ViT (DoBi-SWiP-ViT) segmentation model is illustrated in Fig. 2. At the input stage, similar to the conventional visual DNNs, a cropping operation is first applied to the collected imagery. The cropped image serves as the model input, and following feature aggregation through the binary-quantised ViT backbone with dynamic offsets, the features are fused and processed at multiple levels via the Unified Perceptual Parsing (UPP) module. This process culminates in the generation of the final semantic segmentation results. The detailed implementation of the key modules is described below.

A. Bi-modal Swin-ViT with Unified Perceptual Parsing

The decision to use a Transformer as the primary feature extractor is motivated by its robust capability to capture global features, making it particularly well-suited to tasks such as red tide segmentation, where accurately identifying key features across the entire image is crucial. Conventional Vision Transformers (ViTs) utilize multi-head self-attention mechanisms (MSA) for global feature aggregation. However, the computational complexity of MSA increases quadratically with the length of the input sequence ($O(n^2 \cdot d)$, where n is the sequence length and d is the feature dimension). This imposes a substantial computational burden, particularly when processing large-scale datasets or high-resolution imagery. Therefore, the adoption of the Swin Transformer is driven by its ability to achieve effective feature aggregation while markedly enhancing computational efficiency. The Swin Transformer introduces a window-based multi-head self-attention (W-MSA) mechanism, which confines attention computations to non-overlapping local windows. This approach significantly reduces computational complexity, particularly when the window size M is considerably smaller than the input dimension n, yielding a reduced complexity of $O(n^2 \cdot d/M^2)$. This represents a substantial improvement over conventional MSA. The reduction in complexity not only improves the computational efficiency of the model, but also enables it to handle large-scale visual tasks more effectively by accommodating larger input sizes.

The remote sensing image acquisition sensor on satellite platforms are equipped to capture multi-spectral data, including visible and near-infrared (NIR) bands. In particular,

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3540784

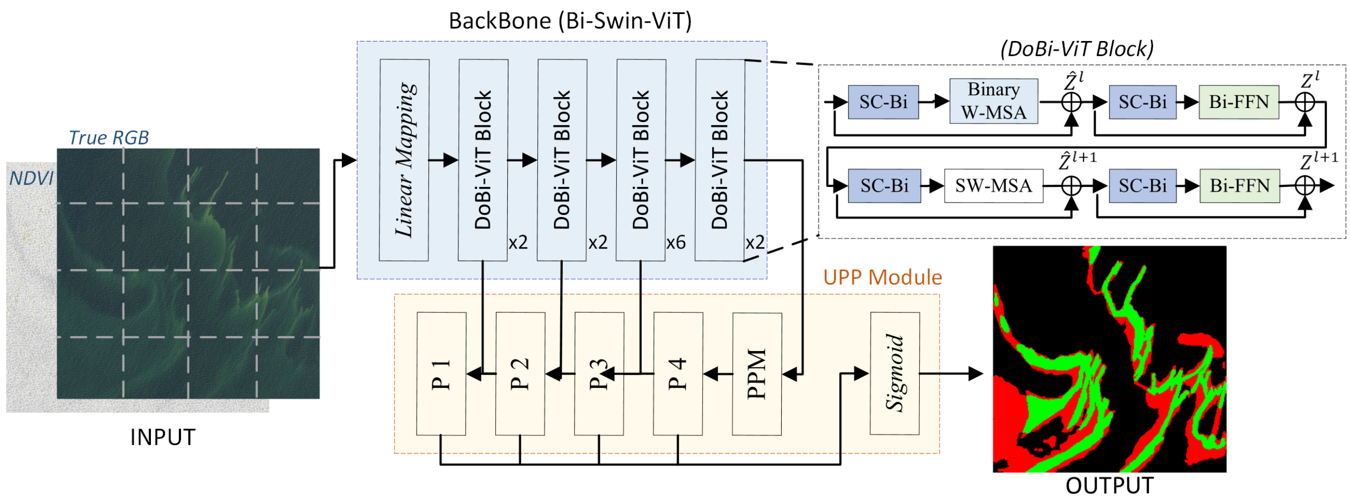


Fig. 2. The overall structure of the bi-modal multi-spectral dynamic offset binary quantization ViT (DoBi-SWiP-ViT) segmentation model. In the UPP module, the pyramid pooling module (PPM) has four different pyramid scales with layer sizes of 1×1, 2×2, 3×3, and 6×6, preserving global context information at multiple scales. P1, P2, P3, and P4 denote multi-level feature maps of $(1/4, 1/8, 1/16, 1/32)$ scales, which are derived through upsampling, downsampling, and convolution operations on the feature maps extracted from corresponding layers of the ViT. The processed feature maps are then combined element-wise to produce the fused feature map for the subsequent layer. Additionally, the Bi-FFN in the DoBi-ViT block comprises two binary linear projection layers and a single activation layer.

the NIR bands offer critical insights into the physical and biological properties of water bodies that are imperceptible through standard visible spectroscopy. Specifically, NIR wavelengths have the capability to penetrate water to a certain depth, enabling the detection of subtle variations in water composition and temperature that serve as indicators of red tide events [51]. Furthermore, the NIR band exhibits high sensitivity to chlorophyll and other pigments associated with algal blooms, which are the primary components of red tides [52]. Building on this, the model design proposed in this paper incorporates a bi-modal multi-spectral input structure within the traditional Swin-ViT architecture. This design leverages bi-modal data sources derived from different spectral bands of multi-spectral images, utilising satellite-acquired multi-spectral information to enrich the feature representation of visible images for red tide monitoring. By adopting this dual-input approach, the model effectively captures additional physical characteristics of the scene, thereby enhancing its overall feature representation capability.

Building upon this foundation, the design of the multi-spectral data input facilitates enhanced feature extraction by leveraging dual input data, enabling the model to integrate features from multiple spectral bands for a more comprehensive understanding of the scene. The proposed dual-stream Bi-Swin-ViT architecture processes these multi-spectral inputs in a coordinated manner, improving the model's robustness to variations in input data. In addition to combining the red, green, and blue (RGB) bands to synthesise true RGB images, the Normalised Difference Vegetation Index (NDVI), specifically designed for detecting Harmful Algal Blooms (HABs) in water bodies, is incorporated to provide supplementary feature information. The calculation of NVDI involves the irradiance of the red and NIR bands, as shown in Eq. 1. Under the multi-input scenario, data augmentation methods applied to images from different spectral bands are maintained consistently to ensure uniformity across modalities.

$$\mathrm{NDVI} = (B_{\mathrm{NIR}} - B_{\mathrm{Red}})/(B_{\mathrm{NIR}} + B_{\mathrm{Red}}) \tag{1}$$

In the backbone component utilising the ViT, the input image is treated as a sequence of tokens by dividing it into small blocks and embedding positional information during preprocessing. At this stage, the input image is partitioned into multiple fixed-size patches, which are linearly projected to form a sequence before being fed into the Transformer. In this paper, the patch size is set to 4, and the window size is set to 7. This approach enables each patch to capture information from the entire image, rather than being constrained to its local region. Additionally, the positional encoding process enhances the model's capacity to perceive global information. The above processing steps for the input image are formulated as in Eq. 2.

$$\mathbf{z}_0 = [\mathbf{x}_{\mathrm{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos} \tag{2}$$

where $\mathbf{z}_0$ denotes the initial input sequence. $\mathbf{x}_{\mathrm{class}}$ represents the class token, which encodes the class-specific information of the input sequence. $\mathbf{x}_p^i$ corresponds to the input feature at patch i. The feature embedding matrix $\mathbf{E}$ maps the input features to a high-dimensional space, with $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C)\times D}$. The positional embedding matrix $\mathbf{E}_{\mathrm{pos}}$, which represents positional information for each input position, is defined as $\mathbf{E}_{\mathrm{pos}} \in \mathbb{R}^{(N+1)\times D}$.

However, this mechanism results in the model lacking a sufficiently clear perceptual field for different patch blocks during feature extraction, making it less effective than CNNs at understanding local feature information. To address this limitation, we introduce the Unified Perceptual Parsing (UPP) module in the feature extraction phase to mitigate the issue

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3540784

of local feature blurring as shown in Fig. 3. This is achieved through cross-level multi-stage feature fusion operations, ensuring effective information flow between deep and shallow features. Functionally similar to the traditional feature pyramid structure, its purpose is to extract high-level semantic features from input images via a multi-scale feature fusion mechanism, enabling the capture of information representations across different scales.

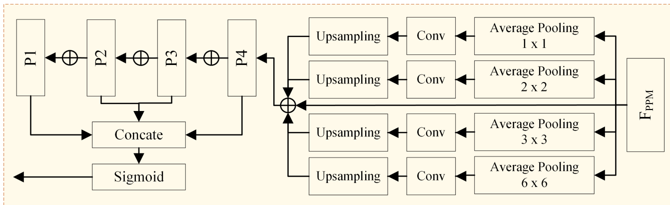


Fig. 3. The detailed structure of the UPP module employs a multi-scale feature fusion mechanism, facilitating the flow of information between deep and shallow features through cross-level fusion operations.

Specifically, this architecture enhances the detail recovery capability and ensures the semantic representation consistency of multi-scale features by incorporating a Feature Pyramid Network (FPN) and a Pyramid Pooling Module (PPM) following the feature extractor. The FPN enriches the semantic information of features through a top-down pathway and lateral connections, enabling the network to effectively handle objects of varying scales. Meanwhile, the PPM captures global context by pooling features across various regions, improving the model's comprehension of the background and large objects. Additionally, the integration of a global pooling operation within this module provides a global feature representation of the entire image. This fusion of global contextual information with local features provides richer semantic information and more accurate segmentation predictions. The concatenation and channel up-sampling steps involve up-sampling all pooled results to the same spatial dimensions, concatenating them along the channel-wise, and adjusting the channel dimensions via convolutional layers to align with the input requirements of subsequent layers.

In summary, we propose the Bi-modal Swin-ViT framework with a Unified Perceptual Parsing module, designed to incorporate more diverse and effective information by integrating bi-modal multi-spectral inputs and advanced feature parsing. This architecture maintains computational efficiency and maximises the extraction and utilisation of contextual information. By leveraging the Transformer mechanism and the unified perceptual parsing approach, the model effectively captures both global and local features, significantly enhancing semantic understanding through their seamless integration. This synergy enables the Bi-modal Swin-ViT and UPP module framework to deliver superior performance in image segmentation tasks, particularly in scenarios demanding high precision and adaptability.

B. Dynamic Offset Binary-ViT block

ViT models are characterised by hierarchical and partitioned self-attention mechanisms, which effectively capture global dependencies and contextual information across images. However, these models are computationally intensive, particularly in large-scale applications requiring timely processing. To address these challenges, we propose applying binary quantization to specific components within the ViT architecture. This approach aims to reduce computational complexity and memory usage while maintaining the model's ability to accurately represent complex image features. Implementing binary quantization accelerates inference speed and facilitates deployment on resource-constrained platforms.

Binary neural networks primarily implement binary processing of the network structure for classification tasks [39], [44], [53], [54], which proves that the feature extraction backbone part of the network already has sufficient representational capacity. However, in semantic segmentation tasks, due to the strict requirement of the sensitivity of the parameter in the segmentation task, the vanilla binary segmentation network can lead to severe performance deterioration situation [55]. Furthermore, addressing the binarisation of ViT models involves unique challenges due to their reliance on the attention mechanism to capture global information and their substantial parameter requirements during training. While the attention mechanism provides ViT models with strong representational capabilities for image tasks, simple binarisation often compromises the accurate representation of complex attention weights. This results in a loss or blurring of global information, ultimately hindering the model's ability to effectively comprehend the overall image structure.

To address this challenge, we propose a learnable Dynamic Offset Binary-ViT (DoBi-ViT) block structure, as illustrated in Fig. 4. This design introduces an additional SC-Bi module both before and after the W-MSA, enabling the binarised transformer blocks to achieve lower quantization error through the introduction of dynamic offset parameters. Compared to conventional binarisation methods, this approach enhances suitability for high-precision segmentation tasks. Subsequently, the feature stream, processed via residual skip connections, serves as input to the binary feed-forward network (Bi-FFN), where further refined feature representations are extracted through the multi-head self-attention mechanism.

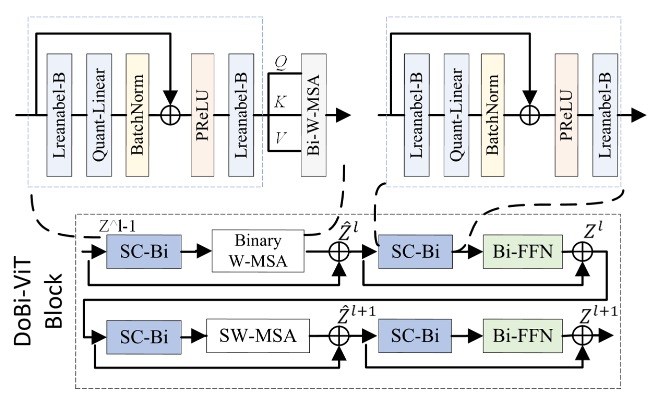


Fig. 4. The SC-Bi design in a learnable Dynamic Offset Binary-ViT (DoBi-ViT).

For each matrix multiplication in the forward phase, a sign function is applied to each input activation X and weight W. A threshold vector σ_X is applied to the real-valued inputs prior to applying the sign function, allowing these inputs to account for distributional shifts. For the weights, the threshold $\mu(W)$ is determined by computing the mean value of all elements within the matrix, as suggested in [56]–[58]. For the activations, the threshold parameter is optimised through back-propagation to minimise the task loss as in [59], [60]. The matrix multiplication output $Y(X)$ in a binary ViT block is calculated as shown in Eq. 3.

$$Y(X) = \frac{1}{n}||W||_1 Rsign(X) \otimes sign(W - \mu(W)) \quad (3)$$

where $Rsign = sign(X+\sigma_X)$ as described in [59]. $\otimes$ denotes the binary convolution, which can be implemented using bit-wise operations such as XNOR and Pop-Count.

In each binary fully connected layer (BiFC) of a binary transformer, we introduce a residual connection that directly links the input to the output of the linear layer, as shown in Eq. 4. This residual connection is designed to preserve information from the previous layer, consistent with the methodology of [44], [59]. Furthermore, all layer normalisation in the ViT model is replaced with Batch Normalization (BN) [61], since all linear layers have a normalisation layer after it, as in [62]. This substitution facilitates faster inference and training compared to layer normalisation.

$$\text{BiFC(X)} = \text{RPReLU}(\text{BN}(Y(X)) + R(X)) \quad (4)$$

where X denotes the input of the layer, $R(\cdot)$ represents the residual connection, and $\text{BN}(Y(X))$ refers to the output of the linear layer. The $\text{RPReLU}(\cdot)$ activation function, as proposed by [59], is applied following each residual connection.

During the back-propagation process, we follow the principle of binary quantization and use the Straight-Through Estimator (STE) [44], to approximate the derivative of the sign function with respect to the input, as presented in Eq. 5.

$$\frac{\partial sign(x)}{\partial x} = \begin{cases} 1 & \text{if } |x| \leq 1 \\ 0 & otherwise \end{cases} \quad (5)$$

Based on the aforementioned settings, we propose a learnable Dynamic Offset Binary-ViT (DoBi-ViT) block structure within the SC-LB-Bi architecture. To address the issue of feature information collapse caused by the linear layer in the quantization model, we introduce a trainable bias (Learnable-B), which performs a sensitivity shift operation on the features during training. This adjustment enhances the diversity of quantization thresholds across different channels by transforming the single sign function into a soft-threshold sign function. Additionally, we adopt the Rectified Parameter exponential Linear Unit (RPeLU) activation function, which introduces further diversity to the hard-threshold quantization originally achieved by the sign function. These modifications improve the model's ability to handle threshold uniformity during training, enhancing its performance after quantization.

In the multi-head attention mechanism, for each head h, the received input features are transformed into three branches: Query (Q), Key (K), and Value (V), which are used for subsequent processing. In the Bi-W-MSA module of the binary transformer with N_H attention heads, the output from the batch normalisation is subsequently used to compute the Query, Key, and Value matrices, denoted as Q_h, K_h, and V_h, where h represents as each attention head. The specific formulation is provided in Eq. 6. In this case, $Q_h, K_h, V_h \in \mathbb{R}^{(N+1)\times D_h}$, where D_h represents the dimensionality of the vectors in each head. Additionally, $D_h = D/N_H$, where D is the total dimensionality of the input feature representation and N_H is the number of attention heads.

$$\begin{aligned} Q_h &= \text{BiFC}_{Q_h}(\hat{H}_h) \\ K_h &= \text{BiFC}_{K_h}(\hat{H}_h) \\ V_h &= \text{BiFC}_{V_h}(\hat{H}_h) \end{aligned} \quad (6)$$

The outputs from all heads are concatenated and processed by the fully connected layer, BiFC_{out}, to compute the multi-head attention output [53]. These are additionally incorporated into the head output to retain the information from the query, key, and value matrices. A primary residual connection is applied to the output of the MHA, as described in Eq. 7.

$$\begin{aligned} F \ &= \text{BiFC}_{out}(Cat(B_1, ..., B_n)) + H \\ B_n &= \text{RPReLU}(\text{BN}(P_h \cdot Rsign(V_h)) + Q_h + K_h + V_h \\ P_h &= \alpha \cdot \lfloor \Theta(\text{Softmax}(\frac{Rsign(Q_h) \cdot Rsign(K_h^T)}{\sqrt{D_h}}), 0, 1)/\alpha \rceil \end{aligned} \quad (7)$$

where P_h denotes the attention matrix derived through the scaled dot-product operation. B_n represents the output of individual heads within the MHA. The learnable scaling factor α, optimised using the method outlined in [63], dynamically adjusts the output range and sensitivity. Specifically, α ensures that the attention weights remain balanced across varying data distributions, enabling the model to better adapt to the dynamic range of quantised data characteristics. The threshold function $\Theta(x, \rho_1, \rho_2)$ constrains the output to the interval defined by ρ_1 and ρ_2. To enhance the robustness and suppress the noise, the output undergoes discretization via the round-to-nearest-integer function $\lfloor \cdot \rceil$. This rounding mechanism minimizes the effect of minor numerical fluctuations on the attention weights, thereby improving the stability of the model during training and inference. The rounding mechanism further facilitates the model's focus on specific attention patterns, minimizing the impact of minor numerical fluctuations on the attention weights and enhancing robustness. This allows the model to balance different attention weights more effectively, thereby boosting overall performance.

Following the aforementioned operations, the residual output F is normalised through a batch normalisation layer. It is subsequently passed through a binary feed-forward network (Bi-FFN) layer, comprising two binary fully connected (BiFC) layers. Finally, a residual connection is applied to the Bi-FFN output, yielding $R = \text{Bi-FFN}(\text{BN}(F)) + F$, which serves as the input for the subsequent DoBi-ViT block.

TABLE I
OVERVIEW OF SATELLITE SYSTEMS: SENTINEL-2, LANDSAT-8, AND PLANETSCOPE

Satellite System	Sensor Name	Spectral Bands	Resolution	Orbital Altitude	Number of Satellites	Revisit Period
Landsat-8 [64]	Operational Land Imager (OLI) Thermal Infrared Sensor (TIRS)	OLI: 9 Bands TIRS: 2 Bands	OLI: 30m TIRS: 100m	705 km	1	16 days
Sentinel-2 [65]	Multi-spectral Imagery (MSI)	13 Bands	10m, 20m, 60m	786 km	2	5 days
PlanetScope [8]	Dove Satellites	8 Bands	3-5 m	400 km	Over 120	1 day

IV. EXPERIMENTS

A. Datasets

The datasets for this study were collected from the open-access Landsat-8 [64], Sentinel-2 [65], and PlanetScope [8] satellite platforms. The sensor specifications and revisit cycles of these satellites are summarised in Table I. Among these satellites, Landsat-8 continues the decades-long tradition of the Landsat programme, capturing high-quality and detailed surface features of the Earth. Equipped with advanced sensors, it provides critical data that support long-term environmental change studies. Sentinel-2, managed by the European Space Agency (ESA), consists of two satellites with multi-spectral imaging capabilities. These satellites deliver high-resolution imagery valuable for applications such as vegetation monitoring, soil and water analysis, urban planning, and disaster management. Notably, Sentinel-2 is the first optical satellite series to incorporate three "red-edge" bands, offering crucial insights into vegetation health and conditions. Meanwhile, PlanetScope, operated by Planet Labs, consists of a constellation of over 120 "Dove" satellites. This system can image the entire land surface of the Earth daily, with a total acquisition capacity of 200 million square kilometres. It is particularly well-suited for rapid responses to natural disasters, agricultural monitoring, and urban development initiatives.

Considering the distribution characteristics of different objects in the study area, multiple locations at different times were selected to construct the training sample dataset, as summarised in Table II. The data were collected from diverse regions across different temporal periods, with the ground truth determined through visual interpretation. To prepare the dataset for model training, the images and their corresponding labels were divided into cropped images of size 512×512 pixels using a sliding window approach, matching the input size of the network. This process yielded a total of 143 samples, which were partitioned into training and validation datasets based on different imagery areas. To ensure sufficient training data and mitigate the risk of overfitting, data augmentation techniques, including horizontal, vertical, and diagonal flipping, were applied to the input images during the training phase. For testing, two separate imagery regions not included in the training dataset were selected. To maintain consistency with the training process, the same sliding window strategy was applied to the test images, dividing each image into cropped images of 512×512 pixels.

Additionally, the two test areas were selected from different satellite platforms to represent distinct red tide conditions, allowing for a comprehensive evaluation of the model's robustness across various scenarios and data sources. Specifically, Test Area A, originated from the Sentinel-2 satellite [65], serves as a critical baseline, utilising independent remote sensing imagery that is temporally and spatially distinct from the training dataset, to assess the model's capability for generalised red tide detection. This approach contrasts with prior studies [66] that often relied on cropping the training and testing data from the same scene. Test Area B, sourced from the PlanetScope satellite [8] and characterised by distinct temporal and spatial conditions, was introduced to further examine the model's robustness and adaptability across diverse contexts. Experimental results indicate that our model consistently outperforms several state-of-the-art methods across both test areas, demonstrating its superior robustness and generalisation capability.

B. Metrics

The performance of the proposed method was evaluated using three criteria, including the mean Intersection over Union (mIoU), the mean Dice Coefficient (mDice), and pixel-based mean Pixel Accuracy (mAcc). The mIoU evaluates the overlap between the model's predictions and ground truth labels by computing the ratio of intersection to the union of the predicted and ground truth regions. Meanwhile, mDice provides another measure of segmentation accuracy, reflecting the degree of overlap between predicted and ground truth regions. It computes the ratio of intersection to the average size of both regions. On the other hand, mAcc assesses the model's pixel-level classification accuracy by determining the ratio of correctly classified pixels to the total number of pixels. The calculation methods for these metrics are presented in Eq. 8.

$$
\begin{aligned}
\text{mIoU} &= \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i+FN_i} \\
\text{mDice} &= \frac{1}{N}\sum_{i=1}^{N}\frac{2\times TP_i}{2\times TP_i+FP_i+FN_i} \\
\text{mAcc} &= \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i}
\end{aligned} \tag{8}
$$

where N represents the number of total classes, which in this case is 2, representing the background and red tide classes. Specifically, TP_i and FP_i denote the True Positive and False Positive for class i, respectively, representing the number of pixels correctly or incorrectly identified as belonging to class i. FN_i refers to False Negative, which is the number of pixels of class i incorrectly identified as another class. TN_i denotes True Negative, which is the number of pixels correctly identified as not belonging to class i.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3540784

TABLE II
OVERVIEW OF SPECIFIC ACQUISITION AREA COORDINATE INFORMATION

Satellite	Resolution	Date	Imagery	Location	Usage
Landsat-8 [64]	30m/px	2020.08.18	LC08_L1TP_116037_20200818_20200920_02_T1	Yellow Sea	Train & Valid
Sentinel-2 [65]	10m/px	2020.08.15	S2A_MSIL2A_20200815T021611_N9999_R003	Yellow Sea	Train & Valid
		2021.02.14	2021-02-14-00_00_2021-02-14-23_59_Sentinel-2_L2A	Vietnam	Train & Valid
		2021.02.23	2021-02-23-00_00_2021-02-23-23_59_Sentinel-2_L2A	Vietnam	Train & Valid
		2020.08.15	Sentinel-2 L2A 2020-08-15-02	Yellow Sea	Test (Area A)
PlanetScope [8]	3m/px	2021.02.22	20210222_030742_62_227a	Guangdong	Train & Valid
		2022.04.10	20220410_021109_59_241f	Guangdong	Test (Area B)

TABLE III
COMPARISON OF THE STATE-OF-THE-ART METHODS FOR THE TEST AREA A.

Method	Base	Param(M)	mIoU	mDice	mAcc
GF1_RI [24]	Index	-	47.69	49.68	50.19
U-Net [19]	CNN	34.5	65.05	75.29	88.79
DeeplabV3 [67]		28.9	64.92	75.03	85.42
DeeplabV3+ [68]		41.2	63.95	73.47	73.07
RDU-Net [25]		34.7	63.38	72.56	68.23
VM-UNet [69]	SSM	27.4	66.97	76.94	85.22
Swin-UNet [35]	ViT	27.1	66.99	46.47	72.43
Swin-ViT [32]		58.9	56.43	67.16	89.39
Vanilla Bi-ViT [55]	Bi-ViT	13.4	52.13	62.37	81.12
BiViT [54]		15.4	58.61	69.25	**89.74**
BinaryViT [70]		22.6	65.13	74.93	78.56
Ours-Binary		22.6	**68.41**	**78.32**	87.34

C. Comparison with the state-of-the-art Methods

In the comparative experiments with state-of-the-art (SOTA) methods, we selected several commonly used methods in semantic segmentation tasks, including U-Net [19], Deeplabv3 [67], Deeplabv3+ [68], and Swin-UNet [35]. Additionally, we included methods specifically designed for red tide segmentation tasks, such as GF1_RI [24] and RDU-Net [25]. Furthermore, we incorporated a recent method based on Selective State-Spaces Models (SSM) [69] for comparative testing. For the binary quantization comparison, we also evaluated several typical methods [54], [55], [70] to assess performance. To validate the robustness of our method across multiple scenarios and datasets, we conducted experiments using remote sensing images collected from different satellite platforms. The differences in acquisition time and location for Areas A and B are detailed in Table II, with harmful algal bloom conditions in Area B significantly differing from those in Area A. A detailed comparison of our method with state-of-the-art (SOTA) methods in Areas A and B are presented in Table III and Table IV, respectively. Additionally, a visual comparison of experimental results is shown in Figure 5 and 6. These results demonstrate that our method achieves excellent segmentation performance across diverse image scenarios, highlighting its effectiveness on various satellite platforms and conditions. This robust performance underscores the adaptability and accuracy of our approach in different remote sensing environments.

In the comparison experiment at Test Area A, conventional CNN methods [19], [67], [68] serve as baselines. While these methods are effective in general segmentation tasks, they struggle to capture global context, which is crucial for accurate red tide segmentation in complex remote sensing images. The GF1_RI method, which utilises radiometric indices, performs poorly, achieving an mIoU of 47.69% and an mDice of 49.68%. This highlights that relying solely on the traditional fixed index-based method set according to the spectrum is insufficient for achieving the fine-grained segmentation required for red tide monitoring in large-scale and multi-scenario environments. By employing a dynamically shifting binary quantization ViT block, our method achieves the highest performance, with an mIoU of 68.41%, an mDice of 78.32%, and an mAcc of 87.34%. The proposed binary ViT framework outperforms methods utilising the same framework. Furthermore, compared to models based on the standard ViT framework, the quantised version significantly reduces parameter usage, thereby enhancing both inference and deployment efficiency. The corresponding quantization loss of the binary quantised model with its counterpart has been further examined through subsequent ablation experiments. The experimental results in Test Area B demonstrate that our method has achieved superior performance compared with similar methods, with mIoU, mDice, and mAcc of 56.39%, 62.24%, and 60.18%, respectively. This indicates that our method exhibits strong adaptability across different red tide monitoring scenarios. Additionally, it can be observed that the performance of the GF1_RI method, based on a specific index design, is superior in Area B compared to Area A. This discrepancy highlights the intrinsic limitations of index-based approaches in diverse regional contexts. Relying heavily on predefined thresholds and fixed band combinations, index-based methods need to be frequently calibrated under specific environmental conditions. These static parameters will inevitably limit the adaptability of such methods, hindering their robust performance in complex environmental regions or under changing environmental conditions. To further elucidate this phenomenon, additional ablation experiments are detailed in Sec. IV-D.

The experimental results demonstrate that our proposed method for red tide segmentation achieves significant improvements over the existing approaches. By leveraging the Vision Transformer (ViT) mechanism, the method effectively addresses the limitations of conventional CNN-based methods in capturing global context, enabling superior global feature aggregation. The integration of multiple spectral bands through a bi-modal multi-spectral combination further enhances feature extraction, resulting in finer segmentation granularity and improved accuracy. Incorporating the UPP module ensures robust feature extraction, refining the segmentation process and boosting overall performance. Additionally, our dynamic offset

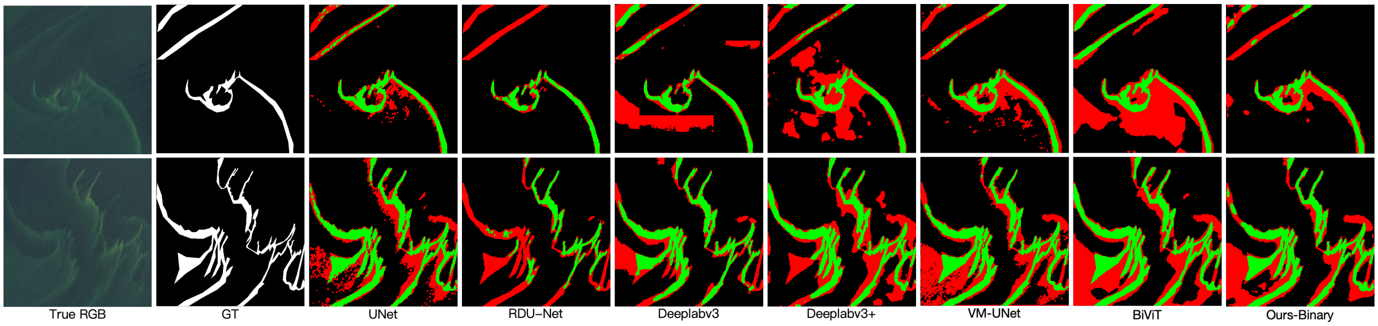


Fig. 5. Visualisation of the comparison results from various methods on the test imagery (denoted as test area A) obtained from the Sentinel-2A satellite. The green, red, and black colours represent correctly detected positive samples, incorrectly detected positive samples, and background areas, respectively. (Best to view in colour)

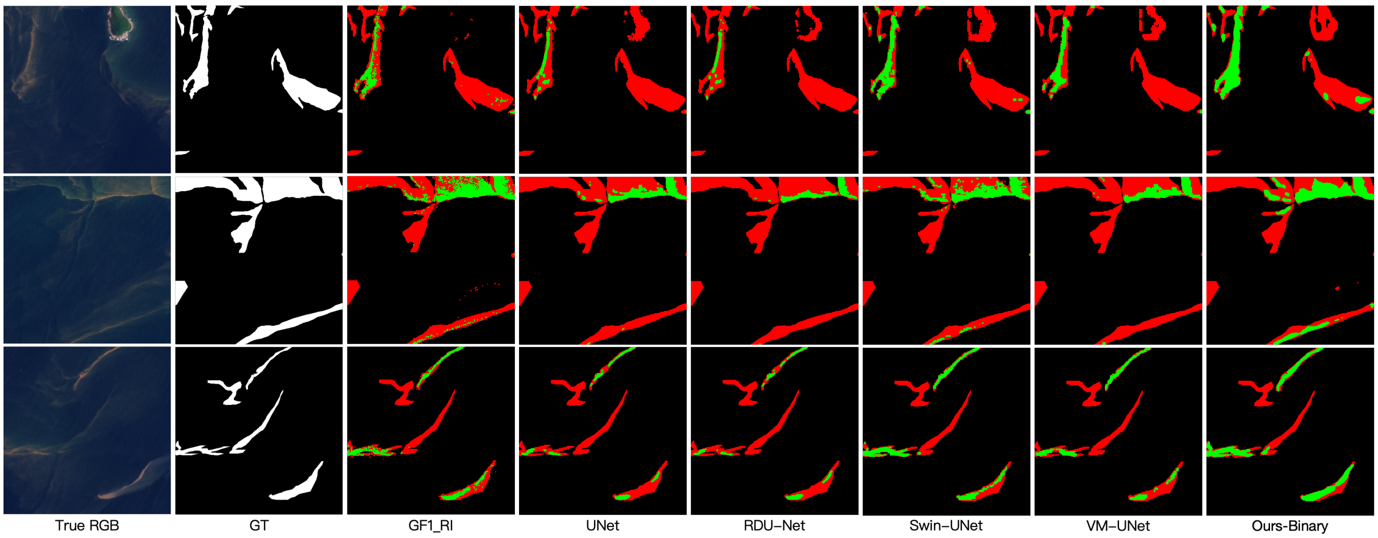


Fig. 6. Visualisation of the results comparing different methods applied to the test imagery (referred to as test area B) from PlanetScope satellite. The green, red, and black regions indicate correctly detected positive samples, incorrectly detected positive samples, and background areas, respectively. (Best to view in colour)

TABLE IV
COMPARISON OF THE STATE-OF-THE-ART METHODS ON THE STUDY AREA B.

Method	Base	Param(M)	mIoU	mDice	mAcc
GF1_RI [24]	Index	-	54.48	59.01	55.36
U-Net [19]	CNN	34.5	51.74	54.63	53.51
DeeplabV3 [67]		28.9	50.93	52.99	52.01
DeeplabV3+ [68]		41.2	50.74	52.69	51.87
RDU-Net [25]		34.7	51.38	53.83	52.72
VM-UNet [69]	SSM	27.4	52.66	56.14	54.30
Swin-UNet [35]	ViT	27.1	54.49	59.31	57.64
Swin-ViT [32]		58.9	50.82	52.71	51.72
Vanilla Bi-ViT [55]	Bi-ViT	13.4	50.03	51.26	50.87
BiViT [54]		15.4	49.61	50.44	50.38
BinaryViT [70]		22.6	49.29	49.65	50.00
Ours-Binary		22.6	**56.39**	**62.24**	**60.18**

binary quantization approach reduces parameter redundancy and enhances computational efficiency without compromising segmentation quality. This combination of advanced techniques results in a robust and accurate solution for remote sensing image segmentation, particularly in the context of red tide segmentation.

D. Ablation Study

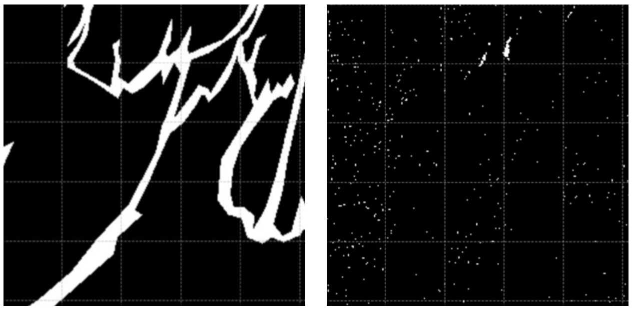
First, we designed a set of comparative experiments using combinations of visible and other multi-spectral bands as inputs to validate the effectiveness of the proposed dual-modal multi-spectral fusion approach in enhancing semantic segmentation performance. Specifically, we considered some widely used indices in remote sensing, including the Normalised Difference Vegetation Index (NDVI) and the Normalised Difference Water Index (NDWI). The calculation method for the NVDI index is presented in Eq. 1 and the NDWI index is presented in Eq. 9.

$$\mathrm{NDWI} = (B_{\mathrm{Green}} - B_{\mathrm{NIR}})/(B_{\mathrm{Green}} + B_{\mathrm{NIR}}) \tag{9}$$

In the comparative experiments, we compared the results of using only True RGB inputs against that of combining True RGB with different spectral band indices. Specifically, this involved comparing the results obtained from RGB images alone with those achieved by concatenating spectral band indices with RGB inputs. The detailed experimental results, shown in Table V, reveal that the multi-spectral combination inputs outperform the RGB-only inputs. This outcome

TABLE V
COMPARISON OF USING TRUE RGB AND AN EXPONENTIAL INDEX COMPOSED OF DIFFERENT SPECTRAL BANDS AS PAIRWISE COMBINATION INPUTS TO THE MODEL ON THE RESULTS OF RED TIDE SEGMENTATION.

Input Modal	mIoU	mDice	mAcc
True RGB	56.43	67.16	89.39
+ NDVI	**60.71**	**71.12**	**90.90**
+ NDWI	60.45	70.06	75.37



(a) Ground Truth (b) Prediction

Fig. 7. Visual comparison between the prediction masks generated using the GF1_RI index and the corresponding ground truth (GT) masks.

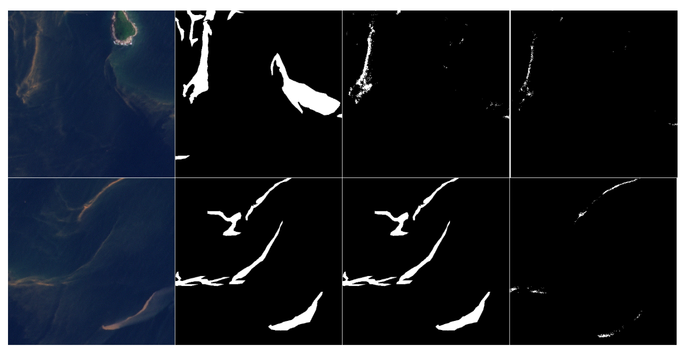
highlights the enhanced feature representation capability provided by bi-modal multi-spectral data. While visible light images offer rich colour and texture information, spectral band data contribute physical characteristics beyond the visible spectrum, enabling the model to perform effectively in more complex environments. Moreover, the bi-modal data input significantly enhances the model's generalisation ability. With inputs constrained to a single data source, the model may exhibit heightened sensitivity to specific types of interference or noise. However, combining diverse data sources allows the model to learn across a broader range of scenarios, thereby enhancing robustness in various imagery applications collected from diverse satellite platforms.

Furthermore, despite being specifically designed for red tide segmentation, the GF1_RI method [24] exhibited significant performance degradation in our experiments, as clearly seen in the visual comparison between the predictions from the GF1_RI index and the Ground Truth (GT) mask shown in Fig. 7. While index-based approaches are computationally efficient and demonstrate high sensitivity in certain conditions, their fixed-threshold mechanism, akin to hard-thresholding, lacks the necessary adaptability to varying environmental conditions and imagery from different satellite platforms. This limitation is particularly pronounced in Test Area A, where such a rigid approach struggles to maintain accuracy across diverse scenarios.

Additionally, given the relatively strong performance observed in Test Area B, we complemented the results of the ablation test by visualising the impact of threshold setting variations on the segmentation results. This was tested by mapping the index-based radiance values used in the GF1_RI method at different threshold settings to the binarised segmentation mask, thereby illustrating their effect on performance, as illustrated in Fig. 8. It is evident that the settings of different hard threshold values influence the results. However, manually adjusting this value for each detection region is impractical. This limitation aligns with the inherent constraints of the index-based method, which assesses the performance based on single-scene imagery. Consequently, the index-based method is only suitable for specific regions and categories of monitoring tasks, with limited generalisability across datasets from different satellite platforms. While index-based methods [24] undoubtedly have their limitations, the multi-source data they utilize are valuable for feature representation. Therefore, we incorporate the spectral information used by index-based methods as an input source into multispectral feature sets, providing complementary information for improved red tide segmentation.

TABLE VI
THE PERFORMANCE COMPARISON OF DIFFERENT PARAMETER CONFIGURATIONS ON SWIN-UNET [35] WAS CONDUCTED, WITH THE PATCH WINDOW SIZE CONSISTENTLY SET TO 8 ACROSS ALL EXPERIMENTS. THE "W/O" (WITHOUT) AND "W" (WITH) DENOTE THE ABSENCE OR PRESENCE OF THE UPP MODULE DESIGN, RESPECTIVELY.

Module	Crop Size	Max Epoch	mDice	mAcc
w/o	256	150	73.49	88.68
w/o	256	200	67.79	84.48
w/o	512	150	46.67	72.43
w/o	512	200	38.84	72.09
w	256	150	**79.06**	89.34
w	512	150	75.13	**92.34**



(a) Input (b) GT (c) thresh=0.3 (d) thresh=0.5

Fig. 8. Visualisation results of the impact of mapping the index-based radiance values at different threshold settings to the segmentation mask.

Building on the results of our previous experiments, we aimed to enhance the feature extraction capabilities of the backbone component in the Swin-ViT structure. Initially, it was observed that employing the Swin-UNet [35] architecture alone led to severe serration in the segmentation results due to the inherent mechanism of the Vision Transformer (ViT), as illustrated in Fig. 9. This serration posed significant challenges in accurately delineating the intricate edge variations characteristic of red tide phenomena in aquatic environments. To mitigate this issue, we designed a series of comparative experiments aimed at evaluating the influence of various adjustable parameters, such as crop size and epoch size.

The hyperparameters were kept constant in these experiments, with the patch window size fixed to 8. The experimental results, presented in Table VI, reveal that crop size has a negligible effect on the segmentation results, while increasing

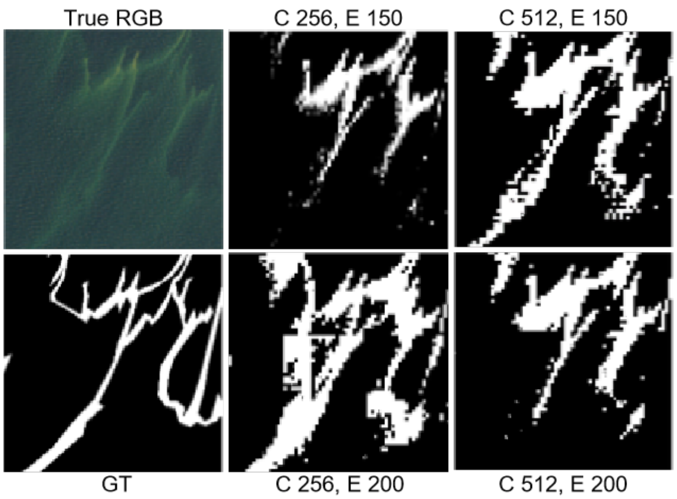


Fig. 9. Visualisation results of the serration phenomenon observed in the results of Swin-UNet with different parameter designs, where the UPP module is not included. "C" denotes the crop size, and "E" refers to the training epoch.

the epoch size exacerbates the overfitting issue. Moreover, the performance in the table reflects the overfitting phenomenon, where the model's performance on the test set significantly decreases as the number of training epochs increases. Based on these observations, we have limited the maximum number of training epochs to effectively mitigate the overfitting and improve the model performance. We have also conducted a series of ablation experiments on the Max Epochs setting, with the specific results shown in Fig. 10. It can be observed that model performance gradually improves and reaches an optimal state between 100 and 150 epochs. Beyond 150 epochs, the model performance begins to decline due possibly to the increasing issue of overfitting. Therefore, we set the maximum number of training epochs to 150 in our experiments and used this as a benchmark for further comparisons of the crop size.

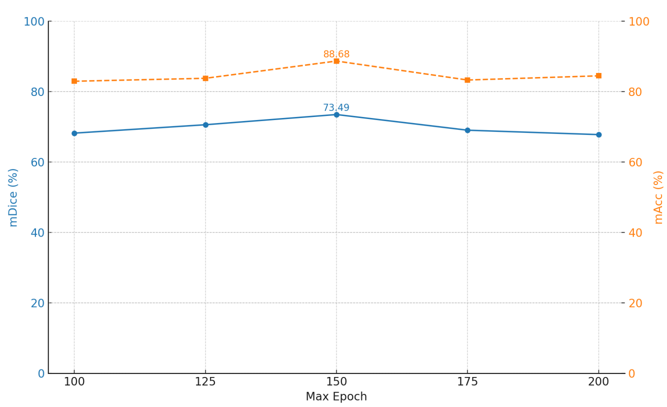


Fig. 10. Impact of different Max Epoch settings during the training phase on model performance

For the quantization aspect, we considered the magnitude of the influence of binary network modules with different design structures on the Swin Transformer-based segmentation framework. This investigation aimed to understand how these different structures influence the performance gap when compared to the FP32 structure and the design of the binary structure with the minimum gap loss is excavated, enabling the possibility of binary inference while maintaining the segmentation accuracy of the model.

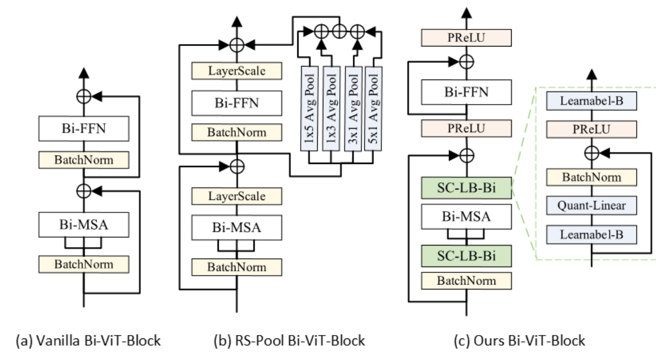


Fig. 11. The comparison of ours and conventional binary ViT architecture.

The Vanilla Binary ViT structure, illustrated in Fig. 11 (a), represents the simplest approach to binarising the model. This process involves a direct replacement of the Multi-Head Attention (MHA) and Feed-Forward Network (FFN) components in the Transformer architecture with their binary equivalents. As evident from the experimental results, this structure leads to significant performance degradation. The main issue with this design is its inability to preserve the precision required for effective feature extraction, which is essential for high-sensitivity, fine-grained visual tasks. Consequently, the network struggles to learn and model subtle differences in the data, resulting in suboptimal performance, particularly in applications that require capturing complex patterns and intricate details.

As illustrated in Fig. 11 (b), the multi-scale aware multi-pooling structure within the binary ViT block enhances the model's ability to perceive images at different scales and details by performing pooling operations at various scales. This helps capture diverse features, improving the model's recognition accuracy in complex scenes. However, the multi-pooling structure can lead to excessive smoothing of features. As noted in previous studies [71], [72], downsampling-based pooling operations are inherently lossy. The primary purpose of the pooling layer is to reduce the spatial dimensions of the feature map, thereby improving computational efficiency and facilitating the extraction of higher-level semantic features. However, the pooling process may reduce the spatial resolution of the feature map by aggregating pixel values within local receptive fields (e.g., the maxima or averages). Consequently, this operation can inevitably result in the gradual loss of local detail information, which will in turn affect the segmentation accuracy.

When examining the segmentation results in detail, it becomes evident that the incorporation of a multi-layer pooling structure tends to blur the edges, as illustrated in Fig. 12. This smoothing effect, caused by the multiple pooling layers, results in inaccurate segmentation in certain detail-rich regions. During the downsampling process, critical local details in these areas are smoothed out, diminishing the model's ability

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3540784

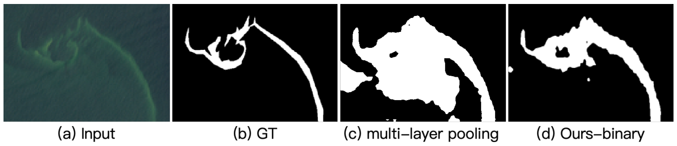


Fig. 12. Impact of the multi-pooling structure on local boundary details in high-precision segmentation.

TABLE VII
COMPARISON OF DIFFERENCE BINARY QUANTIZATION BLOCKS FOR THE TEST AREA A.

Method	#bit (W/A)	mIoU	mDice	mAcc
Ours-FP32	32/32	69.35	79.22	89.76
Vanilla	1/1	52.13	62.37	81.12
Multi-Pooling		65.13	74.93	78.56
Ours		**68.41**	**78.32**	**87.34**

to accurately delineate the boundaries and causing a decline in the segmentation accuracy. Consequently, while the multi-pooling structure can enhance feature recognition across different scales, it may also lead to the loss of critical local boundary information. This limitation is particularly significant in tasks that require precise boundary delineation and the preservation of intricate structural details, such as the segmentation of fine-grained objects or applications demanding high spatial resolution.

Our DoBi-ViT Block design is an enhanced version based on the W-MSA and SW-MSA mechanisms of the Swin-Transformer, as shown in Fig. 11 (c). By incorporating the SC-LB-Bi module before and after the W-MSA, the features entering the Bi-W-MSA can achieve better binarisation processing results, mitigating performance loss during the binarisation phase. Additionally, our model is tailored to specific tasks and does not use a complete binarisation approach that would degrade model performance. Instead, it reduces the redundant parameter bandwidth of common modules at key points, enabling potential embedded deployment of the model.

Based on the analysis of the aforementioned structural variations, we conducted a series of ablation experiments to assess the performance of the proposed various binarised ViT blocks. The detailed experimental results are presented in Table VII. All quantisation schemes were implemented using our FP32 model (Ours-FP32), with the vanilla binary file serving as the baseline for performance comparison. Additionally, the multi-pooling block method, which enhances feature extraction and segmentation accuracy by aggregating information from various image regions, is included as a comparative case. Our proposed method outperforms both the vanilla and multi-pooling methods, achieving mIoU, mDice, and mAcc scores of 68.41%, 78.32%, and 87.34%, respectively. Notably, compared to the multi-pooling method, our approach demonstrates a 3.38% increase in mIoU, a 3.39% increase in mDice, and an 8.18% increase in mAcc. This superior performance is attributed to integrating dynamic magnitude offset binary quantization in our method, which effectively reduces parameter redundancy and enhances computational efficiency while maintaining high segmentation quality.

To comprehensively evaluate the performance of our model during the inference phase, we compared it with Swin-UNet [35], Swin-ViT [32], Vanilla Bi-ViT [55], and BinaryViT [70], using the ViT-based Swin-UNet as the baseline. This comparison effectively highlights the differences in inference efficiency among various Transformer-based models, particularly when applied to large-scale, high-resolution imagery. As shown in Table VIII, the inference time for Swin-UNet, serving as the baseline, was 0.0863s. With an execution time of 0.0545s per iteration, Swin-ViT demonstrated a notable 36.9% improvement in the inference efficiency over the Swin-UNet. In addition, Binary ViT-based models, including the Vanilla Bi-ViT and BinaryViT, showed different levels of inference efficiency. Vanilla Bi-ViT achieved an execution time of 0.0290s, representing a 66.4% improvement over Swin-UNet, while BinaryViT exhibited an inference time of 0.0636s, resulting in only a 26.3% improvement. The observed variation in BinaryViT's performance may be attributed to the multi-layer average pooling operations it used, resulting in extra computational bottlenecks and impact the inference speed.

TABLE VIII
QUANTITATIVE EXPERIMENTAL RESULTS ON THE EXECUTION TIME OF OUR MODEL WITH THE COMPARISON METHODS FOR THE TEST AREA A.

Method	Backbone	Execution Time (per iteration)	Δ (Improvements)
Swin-UNet [35]	ViT-based	0.0863 s	-
Swin-ViT [32]		0.0545 s	-36.9%
Vanilla Bi-ViT [55]	Binary ViT-based	0.0290 s	-66.4%
BinaryViT [70]		0.0636 s	-26.3%
Ours-Binary		0.0393 s	-54.5%

By incorporating further optimisations to the Binary ViT architecture, our DoBi-SWiP-ViT achieved the competitive inference performance with an execution time of 0.0393s per iteration, a 54.5% improvement over the Swin-UNet. This significant performance improvement underscores that our model can not only offer significant advantages in the inference speed but also effectively reduce the inference latency while maintaining the high segmentation accuracy. The comparative analysis has clearly validated the superiority of our model in terms of the inference efficiency, strong applicability and great potential for real-world applications.

V. CONCLUSION

This paper proposes a binary quantization Vision Transformer for the effective segmentation of red tide in multi-spectral remote sensing imagery, addressing the challenges of monitoring red tide hazards. By integrating bi-modal and cross-level feature fusion UPP modules along with an efficient binarisation mechanism for ViT, our DoBi-SWiP-ViT facilitates the effective segmentation of red tides in remote sensing imagery. This approach not only seamlessly integrates the global and local semantic information to ensure the extraction of fine-grained semantic features, but also offers a computationally efficient inference solution for transformer frameworks, which are typically characterised by high parameter consumption. Furthermore, we have curated a dataset for segmenting harmful algal blooms in seawater bodies, comprising high-resolution imagery obtained from sensors on open-access satellite platforms. Based on this dataset, we conducted comparative analyses with state-of-the-art methods and various recently proposed techniques for red tide segmentation. The results demonstrate the superior segmentation performance of our proposed DoBi-SWiP-ViT, achieving finer segmentation granularity compared to state-of-the-art methods. The integration of bi-modal and cross-level feature fusion modules within the ViT framework effectively balances global and local semantic information, which is essential for accurately segmenting complex and varied patterns in remote sensing imagery. Moreover, the introduction of a dynamic magnitude offset binary quantization mechanism effectively reduces the computational burden of the ViT, offering a lightweight solution without sacrificing accuracy. This is particularly important in large-scale remote sensing applications, where computational resources are often limited. With the reduced revisiting cycles of remote sensing satellites, this study aims to enable early monitoring and prompt response to red tide outbreaks, thereby enhancing the speed of protection and mitigation of harmful algal blooms.

ACKNOWLEDGMENT

This work is supported in part by the scholarship from the China Scholarship Council (CSC) under the Grant CSC No.202206290108.

REFERENCES

[1] S. K. Moore, N. J. Mantua, and E. P. Salathe Jr, "Past trends and future scenarios for environmental conditions favoring the accumulation of paralytic shellfish toxins in puget sound shellfish," *Harmful Algae*, vol. 10, no. 5, pp. 521–529, 2011.

[2] K. Cheng, S. N. Chan, and J. H. Lee, "Remote sensing of coastal algal blooms using unmanned aerial vehicles (uavs)," *Marine Pollution Bulletin*, vol. 152, p. 110 889, 2020.

[3] X. Lou and C. Hu, "Diurnal changes of a harmful algal bloom in the east china sea: Observations from goci," *Remote Sensing of Environment*, vol. 140, pp. 562–572, 2014.

[4] H. F. Tolie, J. Ren, and E. Elyan, "Dicam: Deep inception and channel-wise attention modules for underwater image enhancement," *Neurocomputing*, vol. 584, p. 127 585, 2024.

[5] E. Zhang, H. Zong, X. Li, *et al.*, "Icsf: Integrating inter-modal and cross-modal learning framework for self-supervised heterogeneous change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[6] Z. Fang, J. Ren, J. Zheng, *et al.*, "Dual teacher: Improving the reliability of pseudo labels for semi-supervised oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[7] B. Tao, Z. Mao, H. Lei, *et al.*, "A novel method for discriminating prorocentrum donghaiense from diatom blooms in the east china sea using modis measurements," *Remote Sensing of Environment*, vol. 158, pp. 267–280, 2015.

[8] P. L. PBC, *Planet application program interface: In space for life on earth*, Planet, 2021–2022. [Online]. Available: https://api.planet.com.

[9] J. Zhao, M. Temimi, S. A. Kitbi, *et al.*, "Monitoring habs in the shallow arabian gulf using a qualitative satellite-based index," *International Journal of Remote Sensing*, vol. 37, no. 8, pp. 1937–1954, 2016.

[10] M. Moradi and K. Kabiri, "Red tide detection in the strait of hormuz (east of the persian gulf) using modis fluorescence data," *International Journal of Remote Sensing*, vol. 33, no. 4, pp. 1015–1028, 2012.

[11] G. A. Carvalho, P. J. Minnett, L. E. Fleming, *et al.*, "Satellite remote sensing of harmful algal blooms: A new multi-algorithm method for detecting the florida red tide (karenia brevis)," *Harmful algae*, vol. 9, no. 5, pp. 440–448, 2010.

[12] C. Hu, F. E. Muller-Karger, C. J. Taylor, *et al.*, "Red tide detection and tracing using modis fluorescence data: A regional example in sw florida coastal waters," *Remote Sensing of Environment*, vol. 97, no. 3, pp. 311–321, 2005.

[13] J. Shin, K. Kim, Y. B. Son, *et al.*, "Synergistic effect of multi-sensor data on the detection of margalefidinium polykrikoides in the south sea of korea," *Remote Sensing*, vol. 11, no. 1, p. 36, 2018.

[14] Z. Xu, Y. Zhou, X. Wang, *et al.*, "U-net for urban green space classification in gaofen-2 remote sensing images," *Journal of Image and Graphics*, vol. 26, no. 3, 700713, 2021.

[15] J. Li, Q. Xing, X. Zheng, *et al.*, "Noctiluca scintillans red tide extraction method from uav images based on deep learning," *Journal of Computer Applications*, vol. 42, no. 9, pp. 2969–2974, 2022.

[16] Y. Li, J. Ren, Y. Yan, *et al.*, "Cbanet: An end-to-end cross-band 2-d attention network for hyperspectral change detection in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[17] P. Ma, J. Ren, G. Sun, *et al.*, "Multiscale superpixelwise prophet model for noise-robust feature extraction in hyperspectral images," *IEEE transactions on geoscience and remote sensing*, vol. 61, pp. 1–12, 2023.

[18] H. F. Tolie, J. Ren, R. Chen, *et al.*, "Blind sonar image quality assessment via machine learning: Leveraging micro-and macro-scale texture and contour features in the wavelet domain," *Engineering applications of artificial intelligence*, vol. 141, p. 109 730, 2025.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-assisted Intervention (MICCAI) 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241.

[20] Z. Jiang, Y. Ma, T. Jiang, *et al.*, "Research on the extraction of red tide hyperspectral remote sensing based on the deep belief network (dbn)," *Journal of Ocean Technology*, vol. 38, no. 2, pp. 1–7, 2019.

[21] D. Blondeau-Patissier, J. F. Gower, A. G. Dekker, *et al.*, "A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans," *Progress in Oceanography*, vol. 123, pp. 123–144, 2014.

[22] X. Xu, D. Pan, Z. Mao, *et al.*, "A new algorithm based on the background field for red tide monitoring in the east china sea," *Acta Oceanologica Sinica*, vol. 33, pp. 62–71, 2014.

[23] J. Zhao and H. Ghedira, "Monitoring red tide with satellite imagery and numerical models: A case study in the arabian gulf," *Marine Pollution Bulletin*, vol. 79, no. 1-2, pp. 305–313, 2014.

[24] R.-J. Liu, J. Zhang, B.-G. Cui, *et al.*, "Red tide detection based on high spatial resolution broad band satellite data: A case study of gf-1," *Journal of Coastal Research*, vol. 90, no. SI, pp. 120–128, 2019.

[25] X. Zhao, R. Liu, Y. Ma, *et al.*, "Red tide detection method for hy- 1d coastal zone imager based on u-net convolutional neural network," *Remote Sensing*, vol. 14, no. 1, p. 88, 2021.

[26] H. Lee, H. Kwon, and W. Kim, "Generating hard examples for pixel-wise classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 9504–9517, 2021.

[27] Y. Shen, P. Zhong, X. Zhan, *et al.*, "Progressive cnn-transformer alternating reconstruction network for hyperspectral image reconstruction—a case study in red tide detection," *International Journal of Applied Earth Observation and Geoinformation*, vol. 134, p. 104 129, 2024.

[28] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[29] J. Schlemper, O. Oktay, M. Schaap, *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image analysis*, vol. 53, pp. 197–207, 2019.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[31] H. Touvron, M. Cord, M. Douze, *et al.*, "Training data-efficient image transformers & distillation through attention," in *Int. Conf. on Machine Learning*, 2021, pp. 10 347–10 357.

[32] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. the IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 10 012–10 022.

[33] W. Wang, E. Xie, X. Li, *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. the IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 568–578.

[34] K. Han, A. Xiao, E. Wu, *et al.*, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 908–15 919, 2021.

[35] H. Cao, Y. Wang, J. Chen, *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. the European Conference Computer Vision*, Springer, 2022, pp. 205–218.

[36] Y. Bhalgat, J. Lee, M. Nagel, *et al.*, "Lsq+: Improving low-bit quantization through learnable offsets and better initialization," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, 2020, pp. 696–697.

[37] T.-W. Chin, R. Ding, C. Zhang, *et al.*, "Towards efficient model compression via learned global ranking," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 1518–1528.

[38] K. Wang, Z. Liu, Y. Lin, *et al.*, "Haq: Hardware-aware automated quantization with mixed precision," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 8612–8620.

[39] I. Hubara, M. Courbariaux, D. Soudry, *et al.*, "Binarized neural networks," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[40] B. Jacob, S. Kligys, B. Chen, *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2018, pp. 2704–2713.

[41] J. Kim, Y. Bhalgat, J. Lee, *et al.*, "Qkd: Quantization-aware knowledge distillation," *arXiv preprint arXiv:1911.12491*, 2019.

[42] M. Rastegari, V. Ordonez, J. Redmon, *et al.*, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *Proc. the European Conference on Computer Vision*, Springer, 2016, pp. 525–542.

[43] D. Zhang, J. Yang, D. Ye, *et al.*, "Lq-nets: Learned quantization for highly accurate and compact deep neural networks," in *Proc. the European Conference on Computer Vision*, 2018, pp. 365–382.

[44] Z. Liu, B. Wu, W. Luo, *et al.*, "Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm," in *Proc. the European Conference on Computer Vision*, 2018, pp. 722–737.

[45] A. Bulat and G. Tzimiropoulos, "Xnor-net++: Improved binary neural networks," *arXiv preprint arXiv:1909.13863*, 2019.

[46] Z. Wang, J. Lu, C. Tao, *et al.*, "Learning channel-wise interactions for binary convolutional neural networks," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 568–577.

[47] R. Ding, T.-W. Chin, Z. Liu, *et al.*, "Regularizing activation distribution for training binarized deep networks," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 11 408–11 417.

[48] M. Alizadeh, J. Fernández-Marqués, N. D. Lane, *et al.*, "An empirical study of binary neural networks' optimisation," in *Int. Conf. on Learning Representations*, 2018.

[49] Z. Xu, M. Lin, J. Liu, *et al.*, "Recu: Reviving the dead weights in binary neural networks," in *Proc. the*

IEEE/CVF Int. Conf. Computer Vision, 2021, pp. 5198–5208.

[50] C. Liu, W. Ding, P. Chen, *et al.*, "Rb-net: Training highly accurate and efficient binary neural networks with reshaped point-wise convolution and balanced activation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6414–6424, 2022.

[51] K. Dörnhöfer and N. Oppelt, "Remote sensing for lake research and monitoring–recent advances," *Ecological Indicators*, vol. 64, pp. 105–122, 2016.

[52] W. Guan, M. Bao, X. Lou, *et al.*, "Monitoring, modeling and projection of harmful algal blooms in china," *Harmful Algae*, vol. 111, p. 102 164, 2022.

[53] P.-H. C. Le and X. Li, "Binaryvit: Pushing binary vision transformers towards convolutional models," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023, pp. 4664–4673.

[54] Y. He, Z. Lou, L. Zhang, *et al.*, "Bivit: Extremely compressed binary vision transformers," in *Proc. the IEEE/CVF Int. Conf. Computer Vision*, 2023, pp. 5651–5663.

[55] Z. Liu, B. Oguz, A. Pappu, *et al.*, "Bit: Robustly binarized multi-distilled transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 303–14 316, 2022.

[56] H. Qin, Y. Ding, M. Zhang, *et al.*, "Bibert: Accurate fully binarized bert," *arXiv preprint arXiv:2203.06390*, 2022.

[57] H. Qin, R. Gong, X. Liu, *et al.*, "Forward and backward information retention for accurate binary neural networks," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 2250–2259.

[58] H. Qin, X. Zhang, R. Gong, *et al.*, "Distribution-sensitive information retention for accurate binary neural network," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 26–47, 2023.

[59] Z. Liu, Z. Shen, M. Savvides, *et al.*, "Reactnet: Towards precise binary neural network with generalized activation functions," in *Proc. the European Conference on Computer Vision*, Springer, 2020, pp. 143–159.

[60] Z. Xu and R. C. Cheung, "Accurate and compact convolutional neural networks with trained binarization," *arXiv preprint arXiv:1909.11366*, 2019.

[61] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. on Machine Learning*, pmlr, 2015, pp. 448–456.

[62] Z. Yao, Y. Cao, Y. Lin, *et al.*, "Leveraging batch normalization for vision transformers," in *Proc. the IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 413–422.

[63] S. K. Esser, J. L. McKinstry, D. Bablani, *et al.*, "Learned step size quantization," *arXiv preprint arXiv:1902.08153*, 2019.

[64] U.S. Geological Survey, *Landsat-8 data*, Courtesy of the U.S. Geological Survey, 2020. [Online]. Available: https://landsat.usgs.gov/landsat.

[65] C. S.-2. (by ESA), *Msi level-1c toa reflectance product. collection 1*, https://doi.org/10.5270/S2_-742ikth, European Space Agency, 2021.

[66] R. Liu, Y. Xiao, Y. Ma, *et al.*, "Red tide detection based on high spatial resolution broad band optical satellite data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 131–147, 2022.

[67] L.-C. Chen, G. Papandreou, F. Schroff, *et al.*, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[68] L.-C. Chen, Y. Zhu, G. Papandreou, *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. the European Conference Computer Vision*, 2018, pp. 801–818.

[69] J. Ruan and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *arXiv preprint arXiv:2402.02491*, 2024.

[70] P.-H. C. Le and X. Li, "Binaryvit: Pushing binary vision transformers towards convolutional models," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, Jun. 2023, pp. 4665–4674.

[71] H. Gholamalinezhad and H. Khosravi, "Pooling methods in deep neural networks, a review," *arXiv preprint arXiv:2009.07485*, 2020.

[72] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects," in *Joint European conference on machine learning and knowledge discovery in databases*, Springer, 2022, pp. 443–459.

Yefan Xie received the M.Sc degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2020. He is currently working toward the Ph.D. degree in computer science at Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, lightweight networks, remote sensing, and model quantization.

Xuan Hou received the M.Sc. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2020. She is currently pursuing the joint Ph.D. degree in computer science with the School of Computer Science, Northwestern Polytechnical University and the Department of Computer Science, and Faculty of Business and Physical Sciences, Aberystwyth University, Aberystwyth, U.K. Her research interests include change detection, deep learning, and remote sensing.

Jinchang Ren (Senior Member, IEEE) received the B.Eng., M.Eng., and D.Eng. degrees from Northwestern Polytechnical University, Xi'an, China, in 1992, 1997, and 2000, respectively, and the Ph.D. degree from the University of Bradford, Bradford, U.K., in 2019. He is currently a Professor with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K. His research interests include image processing, computer vision, machine learning, and big data analytics. Dr. Ren acts as an Associate Editor for several international journals, including IEEE Transactions on Geoscience and Remote Sensing (TGRS) and the Journal of the Franklin Institute.

Xinchao Zhang is currently working toward the Ph.D. degree in computer science at School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China. His research interests include computer vision and Information Perception and Fusion.

Chengcheng Ma is currently pursuing the joint Ph.D. degree at the School of Software, Northwestern Polytechnical University, Xi'an, China. His research interests include integrated circuits and microsystem design for aviation technology.

Jiangbin Zheng received a Ph.D. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2002. Since 2009, he has been a Professor and a Ph.D. Supervisor with the School of Computer Science, Northwestern Polytechnical University. He has authored/co-authored more than 100 peer-reviewed journal/conference papers covering a wide range of topics in pattern recognition, machine learning, and big data analytics.
He broadly researches in areas such as intelligent information processing, visual computing, 3D reconstruction, multimedia signal processing, big data, and soft engineering.