InterMamba: A Visual-Prompted Interactive Framework for Dense Object Detection and Annotation

Shanji Liu, Zhigang Yang, Qiang Li, *Member, IEEE*, Qi Wang, *Senior Member, IEEE*

***Abstract*—Existing object detection methods is constrained by the high annotation costs, particularly in remote sensing due to the diversity of targets and the large scale of data. Visual-Prompted Interactive Object Detection can enhance the efficiency of data annotation by leveraging user-provided visual prompts to iteratively refine detection results. However, current interactive annotation frameworks are hindered by their reliance on simple feature fusion strategies, which limit their ability to capture fine-grained semantic relationships. Moreover, more advanced fusion methods face computational complexity challenges, making them unsuitable for high-resolution feature spaces commonly encountered in remote sensing imagery. To address these limitations, we propose InterMamba, an efficient framework for interactive object detection in remote sensing images. InterMamba integrates the VMamba backbone and a novel Cross Vision Selective Scan Module (Cross-VSSM) to achieve linear-complexity multi-scale feature fusion, reducing memory consumption while capturing fine-grained details in high-resolution feature spaces. To further enhance interaction flexibility and detection precision, a hybrid Gaussian heatmap generation method is proposed to encodes user-provided point and bounding box annotations. Meanwhile, a User Interaction Loss function further optimizes detection accuracy in dense scenarios by aligning localization and classification with user guidance. Our experiments demonstrate that InterMamba consistently outperforms existing methods in mean Average Precision (mAP). In terms of enhancing precision and reducing annotation costs, InterMamba establishes a robust solution for interactive remote sensing object detection. Code will be available at https://github.com/lsjhaha/InterMamba.**

***Index Terms*—Remote sensing, Visual-prompted interactive object detection, Cross vision selective scan, VMamba**

I. Introduction

OBJECT detection is one of the core tasks in computer vision, aiming to identify and locate objects of interest in images. The advancements in this field have been driven not only by a series of groundbreaking works [1] [2] [3], but also by the availability of large-scale annotated datasets. However, creating such datasets is both time-consuming and high-cost [4]. This challenge is particularly pronounced in the remote sensing domain, where the diversity of objects and the massive scale of data make manual annotation extremely difficult [5]. To address these issues, some methods have attempted to generate datasets using non-interactive inference models [6] [7] [8]. However, the quality of the generated data is frequently suboptimal, and these methods are constrained by their inability to flexibly and efficiently address annotation errors as they occur. To overcome the annotation bottleneck in object detection, Lee et al. proposed the C3Det model [9], which introduces a novel strategy for labeling multiple instances through multi-round user interaction, providing a new pathway for efficient data annotation. A similar motivation is reflected in the iDet3D [10] method, which focuses on interactive annotation for multiple 3D objects. These methods can be categorized as "visual-prompted interactive object detection", a field that has been scarcely explored.

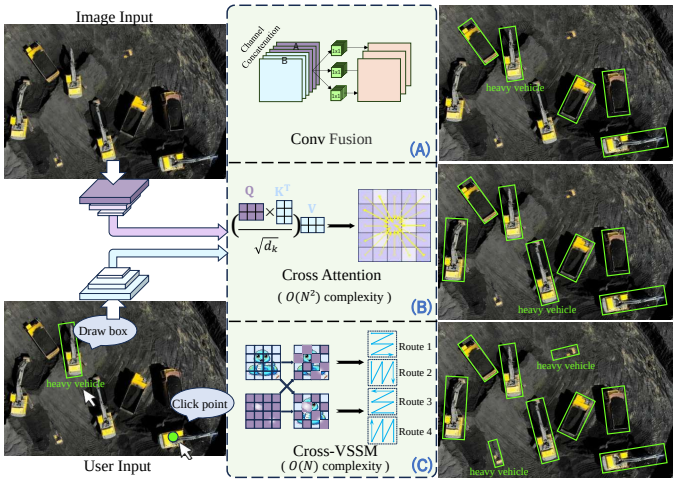


Fig. 1. The motivation of the proposed InterMamba. Excavators, classified as "heavy vehicle", are scarce in the training data, making detection challenging. The figure compares three feature fusion methods: **(A)** Simple convolution detects only user-marked excavators and does not generalize effectively. **(B)** Cross attention associates and detects targeted objects. However, it is limited to small feature maps due to high memory consumption, performing poorly on small objects. **(C)** Our proposed Cross-VSSM significantly reduces memory usage. It enables direct application to large feature maps for effective detection of both large and small excavators.

Visual-Prompted Interactive Object Detection distinguishes itself from general object detection methods by utilizing user-provided visual prompts to iteratively refine detection results through multi-round feedback, thereby enhancing data annotation efficiency and accuracy. Although C3Det provides an effective solution for interactive annotation, its reliance on simple convolutional operations for fusing image features with user-provided prompts constrains its capacity to capture

This work was supported by the National Natural Science Foundation of China under Grant 62471394 and U21B2041.

Shanji Liu is with the School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: liusj@mail.nwpu.edu.cn).

Zhigang Yang, Qiang Li, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail: zgyang@mail.nwpu.edu.cn, liqmges@gmail.com, crabwq@gmail.com) (Corresponding author: Qi Wang.)

the deep semantic information embedded in user inputs. To enhance the integration of complementary information across feature spaces, cross-attention has become a widely used approach [11] [12]. This popularity stems from the ability of cross-attention mechanisms to capture deep semantic relationships by attending to information across multiple feature spaces. Nevertheless, the $O(n^2)$ computational complexity of cross-attention presents significant memory challenges. This makes it unsuitable for directly fusing high-resolution feature maps in feature pyramid networks (FPNs). In order to alleviate this limitation, existing studies [13] often employ cross-attention at lower-resolution FPN layers for detecting large objects, while high-resolution layers are integrated only after dimensionality reduction via convolution. As illustrated in Fig. 1, excavators, despite being classified as "heavy vehicle", are underrepresented in the training data, rendering their detection particularly challenging. Simple convolution methods are limited to identifying user-annotated excavators, leaving other similar objects in the image unrecognized (Fig. 1 (B)). Cross-attention mechanisms, though effective at capturing relationships across objects, are constrained to low-resolution feature maps due to memory limitations, which reduces their effectiveness in detecting small objects (Fig. 1 (C)). Consequently, the development of more efficient cross-modal fusion modules capable of addressing small objects within high-resolution feature layers is crucial for advancing remote sensing detection models.

Recently, a new technique known as VMamba [14] has been introduced. By integrating state-space modeling into visual processing, VMamba achieves linear-complexity operations without sacrificing the ability to capture fine-grained details in high-resolution feature spaces. Inspired by these breakthroughs, we propose InterMamba, an efficient interactive annotation framework that integrates the advanced VMamba backbone into the foundational structure of interactive object detection in remote sensing images. A novel Cross-VSSM fusion module forms a key component of InterMamba, leveraging the selective state mechanism of Mamba with multi-scale feature fusion to efficiently integrate image features and user annotations. By employing a linear-complexity feature selection mechanism, Cross-VSSM significantly reduces memory requirements for high-resolution remote sensing images, effectively addressing the memory bottleneck of traditional cross-attention on high-resolution feature layers. It captures fine-grained features within annotated regions, demonstrating strong adaptability and robustness in multi-class and complex scenarios.

Furthermore, we design a flexible annotation strategy that generates Gaussian heatmaps based on both points and bounding boxes. These heatmaps represent the user-provided interactive information and serve as input to the model. This enhancement allows annotator to flexibly choose points or boxes as input, which significantly improves the precision of interactions and equips the model with a loss function tailored for dense target scenarios. Through experimenting with various strategies for creating Gaussian heatmaps from points or bounding boxes, we determine the most effective approach to enhance the ability of the model to capture user intent and focus on specified regions.

Overall, InterMamba processes images through the Vmamba feature extractor to generate a feature map, while simultaneously creating Gaussian heatmaps from interactive information that guide feature refinement via a lightweight ResNet and the Dynamic Branch Weighting module. These feature maps are fused in the Cross-VSSM module to form a unified representation, which is then used to predict object bounding boxes and class labels, optimizing the model through detection and User Interaction Loss. The main contributions of this work are summarized as follows:

- The proposed Cross Vision Selective Scan Module (Cross-VSSM) facilitates effective fusion of image and guidance features through a cross selective scan strategy. This approach significantly reduces memory consumption, enabling the use of large feature maps for robust feature fusion without explicit attention mechanisms.
- A novel hybrid heatmap generation method is introduced, combining point-based and bounding box annotations to enhance flexibility and detection accuracy in high-density scenarios. This is paired with a User Interaction Loss function, which optimizes detection through a combination of localization and classification components, ensuring precise alignment with provided information.
- We apply the Vmamba backbone to vision-prompted interactive object detection, demonstrating its strong ability to model long-range dependencies and capture fine-grained feature representations. Furthermore, we propose a Dynamic Branch Weighting (DBW) module that adjusts the reliance on image and guidance features dynamically. This design reduces modality bias and enhances detection robustness across different levels of user input density.

II. Related Work

In this section, we briefly review the major works in text-prompted object detection and segmentation, visual-prompted instance segmentation, and visual-prompted object detection, highlighting their evolution, strengths, and limitations.

A. Text-Prompted Object Detection and Segmentation

Text-prompted object detection and segmentation have made significant strides. Foundational work like CLIP [15] introduced a powerful contrastive learning framework, aligning visual and textual modalities and enabling zero-shot classification. Although initially limited to image-level tasks, CLIP inspired a surge of research exploring its extension to localization and pixel-level understanding [16] [17]. In object detection, GLIP [18] reformulates detection as a grounding task, using textual prompts to condition object localization and classification, while Grounding DINO [19] incorporates text-conditioned grounding mechanisms in detection transformers to achieve zero-shot detection across diverse categories. Many other models, such as OWL-ViT [20] and DetCLIP [21], further enhance open-world detection by integrating richer textual descriptions and advanced vision-language training strategies. In the realm of segmentation, recent works have

successfully adapted text prompts to achieve pixel-level semantic understanding. CLIPSeg [22] extends CLIP with a lightweight decoder for efficient segmentation tasks, while GroupViT [23] employs hierarchical grouping within transformers to realize zero-shot segmentation using noisy image-text datasets. Furthermore, approaches like CoDe [24] employ joint vision-text decomposition strategies to refine region-word alignment, advancing the state of text-supervised semantic segmentation. Despite these advancements, text prompts struggle to directly reflect local features within images, resulting in limited precision for object localization. Textual descriptions are often too abstract and lack the capacity to capture pixel-level information, making it difficult for models to accurately target specific regions. This limitation becomes particularly pronounced in tasks that require segmentation of complex geometric shapes or fine boundaries, which can hinder the model effectiveness in practical applications.

B. Visual-Prompted Instance Segmentation

The evolution of visual-prompted instance segmentation began with early techniques like Graph-Cut [25], which laid the groundwork for energy-based segmentation methods, enabling iterative refinement of object boundaries through user-provided scribbles or bounding boxes. Building on Graph-Cut, GrabCut [26] introduced Gaussian Mixture Models and iterative optimization for improved segmentation precision. A major milestone in the field is SAM [27], which provides a general-purpose framework for segmenting any object with minimal user input. Its adaptability has made it widely popular, supporting a broad range of interactive segmentation tasks across various domains. Building upon SAM, SEEM [28] introduces multi-modal prompt handling, supporting diverse interactions such as points, boxes, scribbles, and text commands, which enhances its ability to address complex segmentation challenges. Another advancement, Semantic-SAM [29], extends the capabilities of SAM by incorporating multi-granularity segmentation. This approach enables detailed segmentation at various levels, from whole objects to smaller parts, improving both semantic understanding and precision. Other methods like RefineMask [30] and PointRend [31] enable iterative mask refinement using point-based adjustments, while FocalClick [32] further enhance accuracy, particularly in dense or cluttered environments. However, segmenting small objects remains a challenge due to their weak representation in feature maps, which often causes them to be overshadowed by more prominent regions. Even with iterative refinement, satisfying results usually require significant user input, limiting efficiency. Unlike SAM, which focuses on segmenting a single object using bounding boxes or click-based input, InterMamba supports visual prompts for multiple object categories and is designed to detect a variety of targets, offering enhanced flexibility and robustness, particularly in tasks involving small or densely packed objects.

C. Visual-Prompted Object Detection

Visual-prompted object detection leverages pre-trained models and visual prompts to enhance detection in multi-object and dense scenarios, minimizing the need for task-specific fine-tuning or extensive annotations. A seminal work by Yao [33] introduced an interactive object annotation system that incrementally trains an object detector as annotations are provided. This system minimized human effort by integrating a cost model optimized based on user studies, allowing for real-time feedback and on-the-fly object detection. Building on this interactive foundation, modern methods like C3Det [9] advances this concept by utilizing point annotations to handle dense environments filled with multiple instances, effectively mapping objects across different classes with sparse interactive information. DetPro [34] builds on this by incorporating interactive visual prompts within region proposals, refining detection precision through iterative user feedback. Expanding these techniques to 3D applications, iDet3D [10] leverages point-based inputs for object localization in LiDAR data, making it particularly well-suited for autonomous driving scenarios. Compared to text-prompted methods, visual-prompted object detection models offer a more intuitive, spatially direct interaction, allowing annotator to precisely target and adjust areas of interest, which is especially valuable in real-time or complex visual environments.

Although methods such as C3Det and SAM utilize user interactions, they struggle with efficient multi-scale feature fusion and precise adaptation to high-resolution, small-object scenarios. Unlike the aforementioned methods, our approach addresses these gaps with the Cross-VSSM module for efficient high-resolution feature integration and a hybrid Gaussian strategy to combine point and bounding box inputs, enhancing detection accuracy in complex settings.

III. Proposed Method

In this section, we describe the preliminaries and then introduces the overall structure of InterMamba model with details of each module.

A. Network Architecture

Our task is formulated as an interactive multi-class tiny-object detection [9] problem, with the goal of identifying and classifying densely packed tiny objects in complex scenes using minimal guidance. Each training image is represented as a tuple (I, A), where $I \in \mathbb{R}^{H\times W\times 3}$ denotes an RGB image, and $A = \{a_1, a_2, \ldots, a_M\}$ represents the corresponding object annotations. Each annotation $a_i = (c_i, b_i)$ consists of a class label $c_i \in \{1, \ldots, C\}$ and a bounding box $b_i = (x_i, y_i, w_i, h_i)$. Here, (x_i, y_i) specifies the coordinates of the top-left corner of the box, and (w_i, h_i) denote the width and height, respectively.

To enhance detection performance in high-density scenarios, user-provided information is introduced during inference to provide additional guidance. A user input u may take two distinct forms, i.e. ,

$$u = \begin{cases} (u^{pos}, u^{cls}), & \text{(point input)} \\ (u^{box}, u^{cls}). & \text{(rotated bounding box input)} \end{cases}$$

In the case of point input, $u^{pos} = (u_x^{pos}, u_y^{pos})$ represents the user-specified click position in 2D space, while $u^{cls} \in$

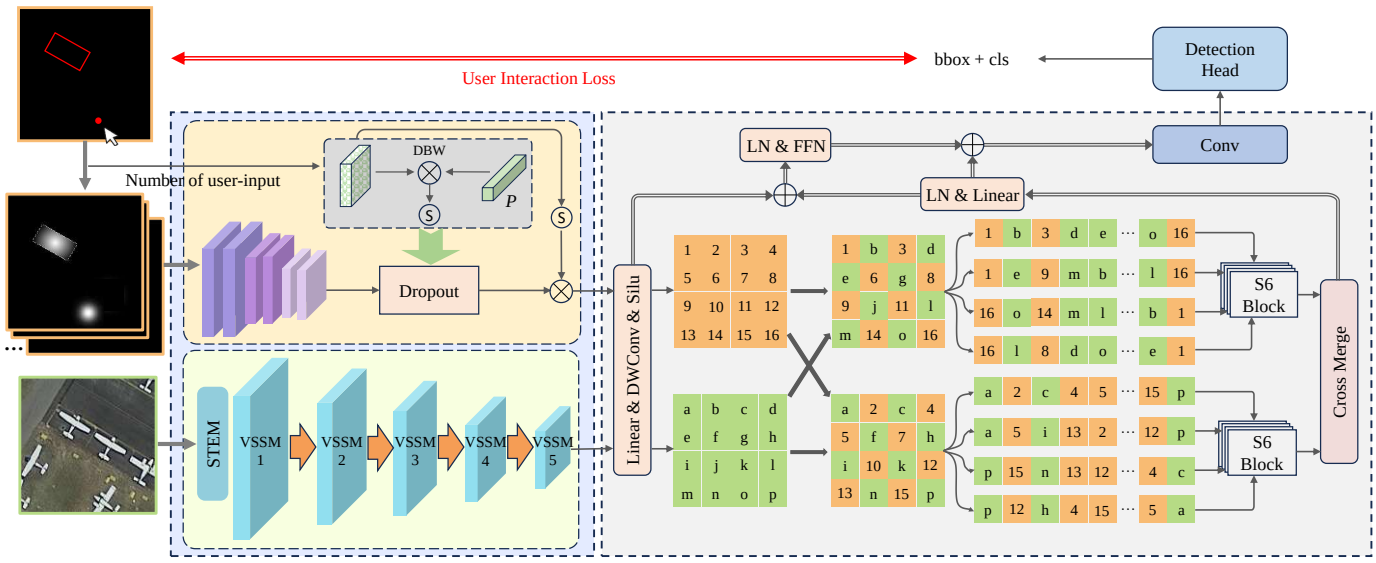


Fig. 2. Overview of the Proposed Method. Our method employs the Vmamba backbone to extract image features and generates Gaussian heatmaps based on user inputs. A lightweight ResNet extracts guidance features, while dynamic branch weighting reduces modality bias. The Cross-VSSM module divides feature maps into p×p grid blocks (e.g., 4×4), where "1, 2, 3, ..." and "a, b, c, ..." denote distinct blocks from two feature maps. These blocks are cross-fused via four-directional sequential scans. Finally, the detection head predicts bboxes and classes, optimizing the model through a combination of user interaction, detection, and classification losses.

$\{1, \ldots, C\}$ specifies the object class. For rotated bounding box input, $u^{box} = (u_x^{box}, u_y^{box}, w, h, \theta)$ includes the center coordinates (u_x^{box}, u_y^{box}), dimensions (w, h), and rotation angle θ.

The overall architecture is illustrated in Fig. 2. During training, the image is first processed by the Vmamba feature extractor (subsection III-B) to generate the feature map F_I. Meanwhile, Gaussian heatmaps based on simulated guidance (subsection III-C) are created and subsequently processed by a lightweight ResNet. The features obtained are passed through the Dynamic Branch Weighting module (subsection III-E) to produce a guidance-enhanced feature map F_U. Both F_I and F_U are input into the Cross-VSSM fusion module (subsection III-D), which constructs a unified feature representation F_{fused} that effectively capture deep semantic information and convey user intentions. Finally, a detection head processes F_{fused} to obtain object bounding boxes and class labels. The outputs generated are then used to calculate detection loss and interaction loss (subsection III-F) for model optimization.

At inference time, the model accepts a limited number of user-provided inputs and aims to maximize detection accuracy across all objects, including those that are not explicitly guided. The objective is to leverage minimal yet flexible guidance to improve detection accuracy for objects across multiple categories, even in complex scenes with densely arranged targets.

B. Preliminaries

In this study, we utilize the VMamba [14] backbone, an innovative vision model based on Vision State Space Models (VSSM), specifically designed for efficient visual feature extraction. By addressing key challenges in computational efficiency, VMamba achieves significant reductions in both computational complexity and memory usage, making it highly advantageous over traditional Transformer-based models.

VMamba is inspired by the Mamba [35] architecture, known for its strengths in long-sequence modeling using selective state spaces. Unlike Vision Transformers (ViTs) [36], where the self-attention mechanism incurs $O(N^2)$ complexity due to the need to compute pairwise interactions across all input tokens, VMamba employs VSSM to achieve linear complexity ($O(N)$). Here, N represents the number of input tokens (or pixels). This reduction is facilitated by the Cross-Scan Module (CSM) and the 2D-Selective Scan (SS2D) mechanism. Together, these components enable efficient scanning of image features along four directions, maintaining global receptive fields while avoiding the memory-intensive attention matrices characteristic of ViTs.

S6 is the core computing unit of VMamba, which efficiently models long-range dependencies based on VSSM. Mathematically, S6 is expressed as a linear time-invariant (LTI) system, i.e. ,

$$h'(t) = Ah(t) + Bu(t),$$
$$y(t) = Ch(t) + Du(t),$$

where $h(t)$ is the hidden state, and A, B, C, and D are the system matrices. The continuous-time system is then discretized to a form that is computationally manageable for deep models. The discretized solution is given by

$$h_{t+1} = e^{A\Delta t} h_t + \sum_{k=1}^{n} e^{A\Delta t_k} B_k u_k.$$

This discretized form is not only computationally efficient but also inherently compact in terms of memory requirements, as it removes the need to store large intermediate matrices commonly used in Transformer-based attention mechanisms. The SS2D mechanism enhances memory efficiency by restricting computations to specific regions of the input and avoids the redundant processes found in traditional attention systems.

These innovations enable VMamba to scale exceptionally well, process high-resolution inputs with lower memory usage, and deliver superior performance in experimental evaluations. In object detection tasks involving densely packed remote sensing images, VMamba proves to be an ideal backbone for this application domain.

C. Hybrid Heatmap Generation

Traditional point annotation methods often fail to provide enough information to optimize detection performance in dense object scenarios. This underscores the urgent need for a new annotation approach. In this section, we propose a novel hybrid heatmap generation method that integrates point-based and bounding box annotations. Alongside this, we introduce a User Interaction Loss aimed at enhancing detection performance in dense object scenarios. Unlike traditional point-only annotation methods, this approach offers greater flexibility and achieves a significant improvement in detection accuracy.

For each object category, a Gaussian heatmap is generated based on user input. These category-specific heatmaps are subsequently stacked to form a composite representation, which is then processed by the feature extractor. The following methods are used for heatmap generation.

Point Annotations: For point-based annotations, we generates a Gaussian heatmap centered on the user-specified location, defined as follows, i.e. ,

$$H(x,y) = \exp\left(-\frac{(x-u_x^{pos})^2 + (y-u_y^{pos})^2}{2\sigma^2}\right),$$

where σ controls the spread of the Gaussian function. This approach enables the model to focus on small object regions with minimal user intervention.

Bounding Box Annotations: For bounding box annotations, we adopt a hybrid method that combines a Gaussian distribution with a binary mask. First, a binary mask $M(x,y)$ is created to define the object boundaries, where pixels inside the bounding box are assigned a value of 1, and those outside are assigned 0. The hybrid heatmap is then formulated as

$$H_{\text{hybrid}}(x,y) = M(x,y) + \exp\left(-\frac{(x-u_x^{box})^2 + (y-u_y^{box})^2}{2\sigma^2}\right).$$

This hybrid approach leverages the binary mask to provide sharp boundary information, while the Gaussian component smooths transitions between the object and background. This method is particularly advantageous in crowded scenes where precise boundary delineation is crucial yet difficult to achieve solely with user input. After generating the Gaussian heatmaps for all object categories, these heatmaps are stacked together to form a multi-channel input representation. This composite input is then passed through the feature extractor, enabling it to learn category-specific spatial patterns and contextual relationships effectively.

D. Cross Vision Selective Scan Module

Visual-Prompted Interactive Object Detection introduces an additional visual prompt branch on top of traditional object detection. Consequently, the image features must be fused with user input features before being passed to the detection head. Conventional feature fusion modules, however, often fail to fully utilize user annotation information, which results in suboptimal performance in dense object detection scenarios. To address this issue, it is crucial to develop a new module that optimizes the integration of image features and user inputs. In this section, we present a novel Cross-VSSM (Cross Vision Selective Scan Module) architecture, specifically designed to fuse image features with user inputs for interactive object detection in remote sensing. Based on the previously introduced Vision State Space Models (VSSM) within the VMamba backbone, Cross-VSSM improves feature fusion by linking the deep semantic information of image features with that of user annotations. Existing research indicates that organizing patch features without explicit down-sampling can preserve global information effectively [37]. Building on this insight, Cross-VSSM incorporates a more robust mechanism for feature fusion.

The core idea behind Cross-VSSM is an innovative cross-scan strategy, which alternates between scanning two feature maps in complementary patterns to ensure global integration of information from both maps. Each feature map is partitioned into smaller patches using a grid structure, with a step size p that determines the spacing between patches. In this approach, each feature map is scanned at intervals defined by p, ensuring that only selected patches are used for fusion. Specifically, during the scanning of feature map F_A, patches are retrieved by skipping every p-th element, while feature map F_B contributes the remaining patches. Mathematically, the feature map F_A is sliced at intervals defined by p, producing patches O_A such that:

$$O_A = F_A[:, m :: p, n :: p],$$

where m and n are determined by the offset and the desired angular pattern. Similarly, feature map F_B is sliced at corresponding intervals to extract patches O_B. The selective scanning process ensures that the global context of both feature maps is maintained, while redundant information is minimized.

At the next stage, the roles of the feature maps are reversed, and the same scanning operation is performed. In this phase, feature map B and feature map A contribute to the feature fusion. This bidirectional interaction ensures that each feature map contributes both unique and shared global information, each step follows the same SS2D mechanism described in the VMamba backbone section, which generate four feature sequences. These sequences are processed using separate S6 blocks in parallel and then reshaped and merged to form the output map. The final output map passes through a series of modules, including Layer Normalization (LN), Linear, and Feed-Forward Networks (FFN), before being processed by a convolutional layer for dimensionality reduction, ensuring the correct feature dimensionality for the subsequent detection head.

Mathematically, the fusion process in Cross-VSSM can be represented as follows. Let F_A and F_B denote the feature maps. The Cross-VSSM operation can be expressed as

$$F_{\text{cross}} = \text{Conv}\left(\text{Scan}(F_A, F_B) \oplus \text{Scan}(F_B, F_A)\right),$$

where $\text{Scan}(F_A, F_B)$ and $\text{Scan}(F_B, F_A)$ represent the cross-scan operations described above, and $\oplus$ denotes the channel-wise concatenation of the two scanned feature maps. The final convolutional layer Conv reduces the dimensionality of the fused feature map.

Cross-VSSM achieves efficient feature fusion while retaining the linear complexity of VSSM, as discussed in Section subsection III-B. This design significantly reduces memory consumption, enabling the processing of high-resolution feature maps and enhancing performance in resource-intensive remote sensing tasks.

E. Dynamic Branch Weighting Module

In our initial experiments, we observe a significant modality bias issue during the multi-modal fusion process. Specifically, when user input is sparse (e.g., only one or two points are provided), the model becomes overly reliant on these sparse input areas, leading to detection performance that is even worse than when the model solely relies on the image modality.

To alleviate this issue, a Dynamic Branch Weighting Module is proposed. It can adaptively adjusts the dropout rate based on the total number of user-specified interest points and bounding boxes. This mechanism enables the model to control its reliance on different modal features. When the interactive information is sparse, the module increases the dropout rate to guide the model to focus more on the global context provided by the image modality. Conversely, as the number of user inputs increases, the dropout rate decreases and reinforces the attention of the model to user-defined regions of interest by maintaining a more focused and consistent processing approach.

Let n represent the total number of interest points and bounding boxes, the dynamic dropout rate, p_{drop}, is calculated using the following softmax function, i.e. ,

$$p_{\text{drop}} = 0.5 \cdot \text{softmax}(-\alpha \cdot n),$$

where α is a scaling factor that controls the rate of change of dropout with respect to n. The softmax function ensures that the dropout rate p_{drop} is constrained within the range $[0, 0.5]$ and decreases as n increases. When the user input is sufficiently abundant, p_{drop} approaches zero, allowing the model to focus entirely on the user-defined regions of interest.

This dynamic adjustment mechanism effectively alleviates the modality bias issue, preventing the model from excessively relying on sparse input regions when the user provides fewer points. Instead, it ensures that the model combines global image modality information with the intent of the user, thus enhancing the robustness and overall detection accuracy.

F. User Interaction Loss

To accommodate the new input format in subsection III-C and enhance responsiveness to user inputs, we propose a User Interaction Loss (UIL) composed of a localization loss and a classification loss.

The localization loss L_{loc} applies exclusively to bounding box inputs, ensuring alignment between predicted bounding boxes and user-provided annotations. For a rotated bounding box u^{box} and predicted boxes $\hat{b}_i$, the localization loss is defined as

$$L_{loc} = \frac{1}{N_U} \sum_{k=1}^{N_U} \mathbb{I}\left(\text{IoU}(u^{box}, \hat{b}^*) > \tau\right) \cdot \left(1 - \text{IoU}(u^{box}, \hat{b}^*)\right),$$

where N_U is the total number of bounding box inputs, τ is the IoU threshold, and $\hat{b}^*$ represents the matched predicted box.

The classification loss L_{cls} handles both bounding box and point inputs. For bounding box inputs, the classification loss $L_{\text{bbox-cls}}$ is implemented as a standard cross-entropy loss to ensure accurate class predictions for matched bounding boxes. For point inputs, the classification loss enforces category consistency within a region of radius r centered at the user-provided point $u^{pos} = (x_c, y_c)$. The loss applies to all predicted bounding boxes $\hat{b}_i$ satisfying $\|\text{center}(\hat{b}_i) - u^{pos}\|_2 \leq r$, i.e. ,

$$L_{\text{point-cls}} = -\frac{1}{N_U} \sum_{k=1}^{N_U} \sum_{\hat{b}_i \in \mathcal{R}(u_k^{pos}, r)} y_{k,c} \log p_{k,c},$$

where $\mathcal{R}(u_k^{pos}, r)$ denotes the set of bounding boxes within radius r of the point u_k^{pos}, and $y_{k,c}$ is the one-hot encoded class label for the k-th point input.

The total User Interaction Loss is

$$L_{\text{UIL}} = \alpha L_{loc} + \beta L_{\text{bbox-cls}} + \gamma L_{\text{point-cls}},$$

where α, β, and γ are hyperparameters controlling the balance between the localization and classification losses. Finally, the total loss integrates UIL with standard object detection losses, i.e. ,

$$L_{\text{total}} = L_{\text{UIL}} + L_{\text{bbox}} + L_{\text{class}},$$

where L_{bbox} and L_{class} are the bounding box regression and classification losses used in standard object detection models.

IV. Experiments

In this section, we propose Tiny-DIOR, a new dataset that overcomes the limitations of existing datasets in terms of target distribution and annotation density. On the basis, we conduct extensive experiments to validate the effectiveness of our proposed methods, including performance comparisons, memory consumption analysis, and ablation studies.

A. Datasets

To address the need for datasets suitable for interactive annotation tasks in remote sensing imagery, we introduce the new dataset Tiny-DIOR. This dataset complements Tiny-DOTA [9], which is currently the only dataset available for such tasks. Both Tiny-DOTA and Tiny-DIOR focus on remote sensing imagery, offering diverse scenarios while meeting the requirements of interactive annotation.

Fig. 3. Demonstration of the response by our method to incremental interactive information. (a) Initial result after the user inputs a single label, showing preliminary detections. (b) Improved output after adds a second label, refining the category of a misclassified object by annotating a bounding box for a heavy vehicle. (c) Final output after three additional annotations, guiding the model to detect previously missed hard-to-recognize objects. (d) and (e) illustrate the effectiveness of our method in a different scenario. The visual prompt input quantities and positions for C3Det are kept consistent with ours, but only point inputs are used. The results show that, C3Det experiences a higher missed detection rate compared to our method.

1) Tiny-DOTA Dataset: Tiny-DOTA is a subset of the DOTA v2.0 dataset, tailored for object detection in an interactive setting. It emphasizes eight small object classes, known for their detection difficulty due to size and density. The dataset reorganizes the training and validation sets of DOTA v2.0 into new subsets, with all images uniformly cropped to 1,024×1,024 pixels to ensure consistency.

2) Tiny-DIOR Dataset: To enhance the generality of our experiments and evaluate the performance of the method across different scenarios, we construct a new interactive object dataset, Tiny-DIOR, based on the DIOR-R [38] dataset. DIOR-R is a remote sensing dataset that includes 23,463 remote sensing images and 190,288 target instances. However, its original annotations exhibit imbalanced target distributions, with certain classes having sparse instances, making it less suitable for our task. We analyze the target distributions in DIOR-R and select five small-object categories—ship, vehicle, storagetank, airplane, and tenniscourt. However, after filtering, we find that many images still contain only one or two instances of a given class. Such sparsity conflicts with the logic of interactive object detection, which relies on annotating a small number of targets to generalize to similar objects. As shown in Table I, this issue is also observed in Tiny-DOTA. To improve its suitability, we further filter the dataset by retaining only images containing at least three instances of the same class. This additional filtering step removes 3,707 images, resulting in a more balanced and practical dataset for interactive tasks. Finally, the dataset is split into training, validation, and testing subsets, with a ratio of 70%, 10%, and 20%, respectively.

TABLE I
COMPARISON OF STATISTICS BETWEEN TINY-DOTA AND TINY-DIOR. FOR TINY-DIOR, "NUM. PATCHES" INCLUDES IMAGES BEFORE REMOVING "LOW-INSTANCE FILES," WHEREAS FOR TINY-DOTA, IT REFLECTS THE COUNT AFTER REMOVAL.

	Num. classes	Num. patches			Low-instance files (1–2/category)
		train	val	test	
Tiny-DOTA	8	11,198	1,692	2,823	4,490
Tiny-DIOR	5	6,810	851	1,703	3,707

B. Implementation Details

During training, we simulate user-provided information based on ground-truth annotations to closely approximate real-world interaction scenarios. First, a random number N_u is sampled from a uniform distribution, with its maximum value adjustable based on dataset characteristics. Next, the actual number of interactions, K, is defined as the minimum of N_u and N_a, where N_a represents the number of available objects in the current image. Finally, for each annotated object, either the object center or bounding box, combined with its class information, is selected as simulated input. We employ the AdamW optimization method to refine the weights. VMamba-S [14] is selected as our image backbone, while ResNet-18 [1] serves as the guiding backbone. Our experiment follows the hyperparameter settings detailed in Swin [39], with a learning rate of 0.001, a batch size of 2, and 12 epochs. The hyperparameters α, β, σ and γ are set to 0.5, 1.0, 0.5, and 0.75. All models were implemented in PyTorch [40] on a computer with an Intel Xeon E5-2680 v4 @ 2.40GHz CPU and two Nvidia RTX 3090 GPU.

TABLE II
PERFORMANCE ON TINY-DOTA COMPARISON ACROSS DIFFERENT NUMBERS OF USER INPUTS. ALL VALUES REPRESENT MAP WITH AN IOU THRESHOLD OF 0.5. THE BEST RESULTS ARE BOLD.

Method	1 Input	5 Inputs	10 Inputs	15 Inputs	20 Inputs
Faster R-CNN [41]	60.19	60.19	60.19	60.19	60.19
Swin Transformer [39]	60.71	60.71	60.71	60.71	60.71
Vmamba [14]	61.82	61.82	61.82	61.82	61.82
RTMDet [42]	69.45	69.45	69.45	69.45	69.45
DecoupleNet [43]	70.06	70.06	70.06	70.06	70.06
LWGANet [44]	70.79	70.79	70.79	70.79	70.79
DCFL [45]	**73.43**	73.43	73.43	73.43	73.43
C3Det [9]	63.55	69.02	72.31	73.39	73.82
C3Det-Vmamba	64.31	70.22	73.24	73.56	74.60
Ours (points)	66.20	72.23	74.41	75.93	77.52
Ours (bboxes)	67.68	**74.23**	**76.61**	**78.35**	**79.26**

TABLE III
PERFORMANCE ON TINY-DIOR COMPARISON ACROSS DIFFERENT NUMBERS OF USER INPUTS. ALL VALUES REPRESENT MAP WITH AN IOU THRESHOLD OF 0.5. THE BEST RESULTS ARE BOLD.

Method	1 Input	5 Inputs	10 Inputs	15 Inputs	20 Inputs
Faster R-CNN [41]	79.43	79.43	79.43	79.43	79.43
Swin Transformer [39]	82.38	82.38	82.38	82.38	82.38
Vmamba [14]	82.84	82.84	82.84	82.84	82.84
RTMDet [42]	83.26	83.26	83.26	83.26	83.26
DecoupleNet [43]	83.01	83.01	83.01	83.01	83.01
LWGANet [44]	85.15	85.15	85.15	85.15	85.15
DCFL [45]	**86.92**	86.92	86.92	86.92	86.92
C3Det [9]	83.05	86.01	87.99	88.10	88.15
C3Det-Vmamba	83.38	86.42	88.13	88.24	88.49
Ours (points)	84.65	87.44	89.21	89.42	89.47
Ours (bboxes)	85.11	**88.44**	**89.81**	**90.02**	**90.07**

C. Evaluation Procedure

To assess the effectiveness of our method in utilizing incremental annotations, we design an evaluation procedure that simulates real-world interactions. As shown in Fig. 3, our model progressively refines detection results through incremental annotations, effectively addressing challenges in dense object scenarios. As indicated by the red boxes in Fig. 3 (b), the model leverages visual cues to resolve misclassification issues. Similarly, as highlighted in Fig. 3 (c) and (e), it utilizes visual prompts to mitigate missed detections. Compared to C3Det, our method achieves a lower missed detection rate after adding visual prompts. The evaluation procedure detailed below closely mimics this iterative process.

In our model evaluation, we simulate up to 20 annotation interactions for each image sample. During each simulation, a ground-truth object is randomly selected (without replacement) as the annotation target. Its center location or bounding box information is randomly used as the simulated input. If all ground-truth objects in the image have already been used as inputs, no additional input is provided. This process generates a sequence of predicted bounding boxes corresponding to each interaction, constructing a test set comprising 20 stages of user inputs. To ensure consistent evaluation conditions across different models, we preprocess the simulated inputs before testing, ensuring that all models receive identical inputs during evaluation.

We evaluate model performance using mean Average Precision (mAP) with an IoU threshold of 0.5. As shown in Table II and Table III, the results are reported for different numbers of inputs (e.g., 1, 5, 10, 15, and 20). These results demonstrate how incremental annotations progressively improve detection accuracy, particularly in challenging scenarios with dense objects or sparsely annotated data.

D. Comparison with Baseline Methods

To comprehensively validate our method, we conduct comparative experiments against several baseline methods. Considering that our approach supports both point and bounding box inputs, while baseline method C3Det only support point inputs, we also evaluate a point-input-only version of our model for a fair comparison. The methods compared are as follows:

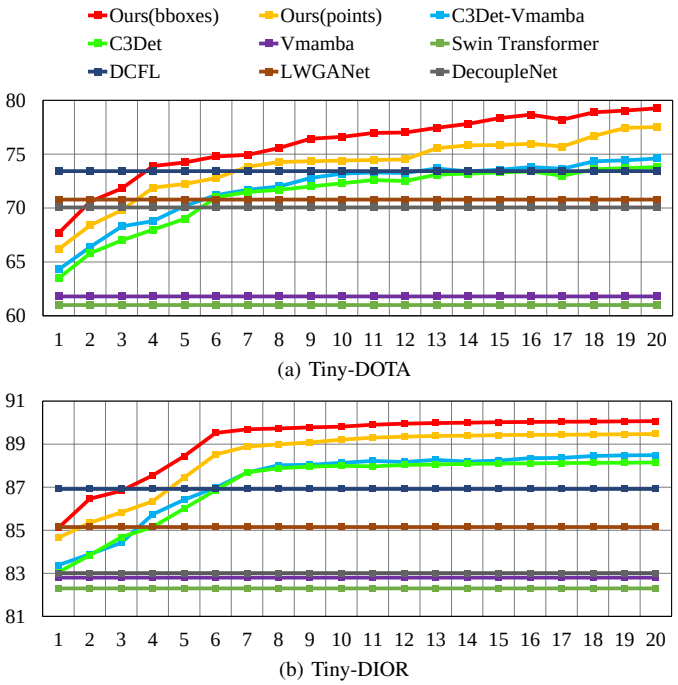


Fig. 4. InterMamba performance on Tiny-DOTA and Tiny-DIOR datasets compared to baseline methods. General object detection methods are unresponsive to user clicks, causing their performance curves to remain flat.

Our model (full): The complete model supporting both point and bounding box inputs.

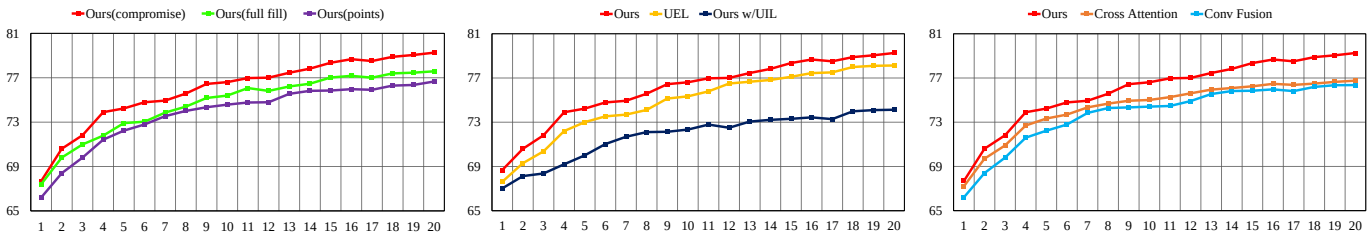
Our model (points only): The same model evaluated using only point inputs for direct comparison with methods like C3Det.

C3Det: The full interactive detection model based on a Feature Pyramid Network (FPN) [46], supporting only point inputs.

C3Det (VMamba backbone): A variant of C3Det where the ResNet-50 [1] backbone is replaced with our proposed VMamba backbone.

Other models: Standard models that remain consistent with their original implementations.

Results and Analysis: The experimental results demonstrate that our method outperforms all baseline methods across various interactive click scenarios, particularly in terms of mean Average Precision (mAP). As shown in Fig. 4 (a) and Table II, when using both point and bounding box inputs, our approach achieve an mAP of 79.26 at 20 clicks, showcasing



(a) Heatmap Generation Method (b) User Interaction Loss (c) Cross-VSSM

Fig. 5. Ablation study of InterMamba on Tiny-DOTA. The graphs show the impacts of (a) Heatmap Generation Method, (b) User Interaction Loss, and (c) Cross-VSSM module.

its ability to effectively integrate diverse user input types. Under point-only input conditions, our method exhibits robust performance with an mAP of 77.52, significantly surpassing the 73.8 mAP achieved by C3Det. Similar experimental results can also be observed in Fig. 4 (b) and Table III. Furthermore, we have also compared our method with recent rotation-based general object detection approaches. These results demonstrate that while our method may initially lag behind these methods with fewer visual prompts, it surpasses them with five or more prompts. Comparisons of backbone networks reveal that replacing backbone in Swin Transformer [39] with VMamba results in approximately a 0.8 mAP improvement, highlighting the ability of VMamba to model global and local contextual information.

During testing, minor fluctuations in mAP values are observed across all models at specific user input stages (e.g., 12 or 17 clicks in Fig. 4 (a)), due to variations in Tiny-DOTA dataset distributions. These fluctuations, caused by the exclusion of image samples with insufficient annotations as the number of user clicks increases, do not materially affect overall trends or performance evaluations.

E. Ablation Study

To validate the impact of our proposed modular design and loss functions on model performance, we conduct three ablation studies on the Tiny-DOTA dataset. These studies specifically analyze the effect of UIL and compare the efficacy of different feature fusion methods.

Comparison of Gaussian Map Generation Methods: In Fig. 5 (a), we compare three approaches for generating Gaussian maps of bounding boxes. (1) Hybrid Approach (Our Method): This method combines Gaussian gradients and region filling within the bounding box to create a mixed heatmap, as described in detail earlier. (2) Full-Fill Approach: The entire bounding box region is filled with uniform values, ignoring gradient transitions. (3) Point-Input Approach: Gaussian gradient maps are generated solely from the center point of the bounding box.

The results show that the hybrid approach significantly improves detection performance. Compared to the full-fill method, the hybrid approach balances global region information and boundary details, enhancing precision in object detection. Relative to the point-input method, the hybrid approach retains the precise guidance from point annotations while providing additional spatial constraints via bounding box information, which improves the understanding of object shape and scope by the feature extractor. This demonstrates the superiority of a combined approach that integrates precise localization with global range information for dense object detection tasks in remote sensing scenarios.

Validation of User Interaction Loss Effectiveness: Fig. 5 (b) compares models trained with UIL, User-input Enforcement Loss (UEL) [9], and without any user-guided loss. The results demonstrate that UIL consistently improves performance across all click counts and shows significant advantages in both sparse and dense annotation scenarios. These findings highlight the superior effectiveness of UIL in leveraging user input compared to UEL.

The UIL incorporates user-annotated intentions into model predictions, ensuring high consistency between outputs and the intended annotations. This enhancement improves robustness to sparse inputs (e.g., fewer clicks) and its ability to capture multiple objects in complex scenarios. Furthermore, UIL effectively guides the model to focus on user-specified critical regions when click counts are low, and prevents over-reliance on redundant inputs when click counts are high by introducing constraints. These findings confirm the adaptability and generalization capabilities of our method across different interaction scenarios.

Comparison of Feature Fusion Methods: In Fig. 5 (c), we compare the performance of three feature fusion methods. (1) Our Cross-VSSM Approach: This novel Cross-VSSM module integrates global and local contextual information, achieving efficient multi-scale feature fusion with linear complexity. (2) Cross Attention Approach: Cross-attention mechanisms are introduced at the low-resolution layers of the Feature Pyramid Network (FPN), while convolution is used for high-resolution layers due to memory limitations. (3) All-Convolution Fusion Approach: Traditional convolutional operations are applied to all feature layers for fusion.

The results demonstrate that InterMamba achieves the highest average detection accuracy in multi-category scenarios, with a significant performance advantage over the other two methods. This shows that InterMamba fuses fine-grained information from high-resolution feature maps effectively while maintaining the ability to model global context through low-resolution layers. Compared to the Cross Attention approach, InterMamba reduces computational complexity, alleviating the memory bottleneck associated with high-resolution layers and retaining high-precision target information. In contrast to the all-convolution approach, InterMamba uses state-space models to capture complex relationships across multi-scale features,

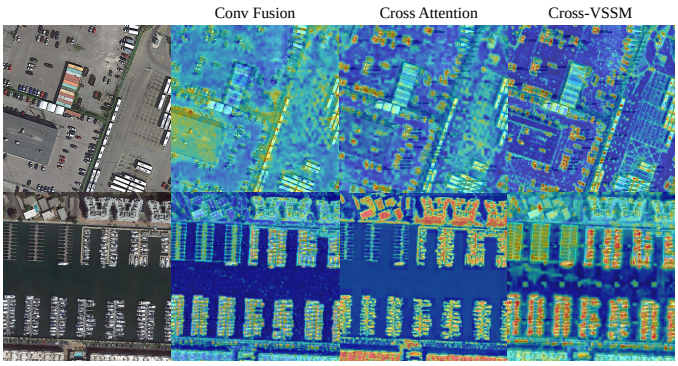


Fig. 6. The CAM visual results of different fusion methods. The figure is selected from the DOTA-tiny test set, illustrating the enhanced attention of Cross-VSSM to dense object regions compared to cross-attention mechanisms and Conv Fusion.

TABLE IV
MEMORY CONSUMPTION OF FEATURE FUSION METHODS ON FPN OUTPUTS. EXPERIMENTS WERE CONDUCTED WITH A BATCH SIZE OF 2 AND THE NUMBER OF HEADS IN THE MULTI-HEAD ATTENTION SET TO 2.

FPN Layer	Feature Shape	Memory Consumption (MB)	
		Cross-VSSM	Cross Attention
P0 (top)	$2, 256, 256, 256$	5,141.78	72,192 (estimate)
P1	$2, 256, 128, 128$	1,285.83	8,824.82
P2	$2, 256, 64, 64$	321.36	595.59
P3	$2, 256, 32, 32$	80.40	48.23
P4 (bottom)	$2, 256, 16, 16$	20.14	5.77

which enhances its robustness and adaptability in interactive object detection tasks. Additionally, the Class Activation Map (CAM) visualizations presented in Fig. 6 further highlight the superior capability of the Cross-VSSM approach in concentrating attention on dense object regions. In the first image, compared to the other two fusion methods, our approach exhibits a more precise focus on key objects, such as vehicles and ships. This advantage becomes even more evident in the second image.

F. Memory Consumption Comparison for Feature Fusion Modules

The memory consumption of Cross Attention and Cross-VSSM modules is analyzed. For Cross Attention, the memory grows quadratically with the spatial resolution due to the calculation of the attention weight matrix, whereas Cross-VSSM scales linearly with the feature map size. Table IV summarizes the memory usage for both modules, calculated using the feature shapes of FPN outputs.

For Cross Attention, the quadratic complexity of the attention weight matrix, $O(n^2)$, leads to exponential growth in memory consumption for high-resolution feature maps. This is particularly evident in the top FPN layer (P0), where memory usage reaches 72,192 MB due to the large spatial dimensions. In contrast, Cross-VSSM employs a linear complexity approach, resulting in significantly reduced memory requirements. For instance, at the P0 layer, Cross-VSSM requires only 5,141.78 MB, which is approximately 93% lower than Cross Attention. These results demonstrate that Cross-VSSM is highly efficient in handling the fusion of high-resolution feature maps, making it more suitable for memory-constrained scenarios.

V. CONCLUSION

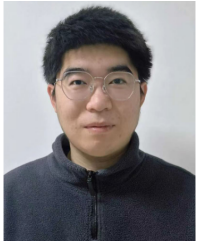
In this paper, we present a novel approach to interactive remote sensing object detection by introducing the VMamba model into this domain. Our innovative Cross-VSSM architecture effectively integrates image features with user input heatmap features, enhancing the overall performance of the annotation process. Additionally, we propose a heatmap generation method capable of combining point and bounding box inputs, accompanied by a tailored UIL function that optimizes the interaction between the system and the annotator. Our experimental results demonstrate that our method significantly reduces annotation costs while maintaining high-quality outputs compared to existing approaches. This contribution not only advances the state-of-the-art in interactive annotation but also provides a scalable solution that can be applied in various real-world settings. Future research can explore further refinements to our model, such as improving interaction robustness and expanding the applicability of our framework to additional tasks within the realm of computer vision.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
[2] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo *et al.*, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 44, no. 11, pp. 7778–7796, 2021.
[3] Q. Li, M. Zhang, Z. Yang, Y. Yuan, and Q. Wang, "Edge-guided perceptual network for infrared small target detection," *IEEE Trans. Geosci. Remote Sensing*, 2024.
[4] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *Int. J. Comput. Vis. (IJCV)*, vol. 77, pp. 157–173, 2008.
[5] Q. Li, Q. Wang, and X. Li, "Exploring the relationship between 2d/3d convolution for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 10, pp. 8693–8703, 2021.
[6] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis. (IJCV)*, vol. 128, no. 7, pp. 1956–1981, 2020.
[7] L. Cai, Z. Zhang, Y. Zhu, L. Zhang, M. Li, and X. Xue, "Bigdetection: A large-scale benchmark for improved object detector pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 4777–4787.
[8] S. Liu, Y. Ma, X. Zhang, H. Wang, J. Ji, X. Sun, and R. Ji, "Rotated multi-scale interaction network for referring remote sensing image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 26 658–26 668.
[9] C. Lee, S. Park, H. Song, J. Ryu, S. Kim, H. Kim, S. Pereira, and D. Yoo, "Interactive multi-class tiny-object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 14 136–14 145.
[10] D. Choi, W. Cho, K. Kim, and J. Choo, "idet3d: Towards efficient interactive object detection for lidar point clouds," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, no. 2, 2024, pp. 1335–1343.
[11] Q. Wang, Z. Yang, W. Ni, J. Wu, and Q. Li, "Semantic-spatial collaborative perception network for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sensing*, pp. 1–1, 2024.
[12] Z. Yang, W. Zhang, Q. Li, W. Ni, J. Wu, and Q. Wang, "C2Net: Road extraction via context perception and cross spatial-scale feature interaction," *IEEE Trans. Geosci. Remote Sensing*, 2024.
[13] Y. Ji, H. Zhang, Z. Jie, L. Ma, and Q. J. Wu, "CASNet: A cross-attention siamese network for video salient object detection," *IEEE Trans. Neural Netw. Learn. Syst. (TNNLS)*, vol. 32, no. 6, pp. 2676–2690, 2020.

[14] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "VMamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.
[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PmLR, 2021, pp. 8748–8763.
[16] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2022, pp. 12 888–12 900.
[17] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao, "Glipv2: Unifying localization and vision-language understanding," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 36 067–36 080, 2022.
[18] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10 965–10 975.
[19] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.
[20] G. Xiuye, L. Tsung-Yi, K. Weicheng, and C. Yin, "Open-vocabulary object detection via vision and language knowledge distillation," in *Int. Conf. Learn. Represent. (ICLR)*, 2022.
[21] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. Xu, and H. Xu, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 9125–9138, 2022.
[22] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 7086–7096.
[23] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18 134–18 144.
[24] J.-J. Wu, A. C.-H. Chang, C.-Y. Chuang, C.-P. Chen, Y.-L. Liu, M.-H. Chen, H.-N. Hu, Y.-Y. Chuang, and Y.-Y. Lin, "Image-text co-decomposition for text-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 26 794–26 803.
[25] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 23, no. 11, pp. 1222–1239, 2001.
[26] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut" interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.
[27] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4015–4026.
[28] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2024.
[29] F. Li, H. Zhang, P. Sun, X. Zou, S. Liu, J. Yang, C. Li, L. Zhang, and J. Gao, "Semantic-sam: Segment and recognize anything at any granularity," *arXiv preprint arXiv:2307.04767*, 2023.
[30] G. Zhang, X. Lu, J. Tan, J. Li, Z. Zhang, Q. Li, and X. Hu, "Refinemask: Towards high-quality instance segmentation with fine-grained features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 6861–6869.
[31] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9799–9808.
[32] X. Chen, Z. Zhao, Y. Zhang, M. Duan, D. Qi, and H. Zhao, "Focalclick: Towards practical interactive image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1300–1309.
[33] A. Yao, J. Gall, C. Leistner, and L. Van Gool, "Interactive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 3242–3249.
[34] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 14 084–14 093.
[35] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
[36] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[37] X. Pei, T. Huang, and C. Xu, "Efficientvmamba: Atrous selective scan for light weight visual mamba," *arXiv preprint arXiv:2403.09977*, 2024.
[38] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–11, 2022.
[39] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10 012–10 022.
[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.
[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 39, no. 6, pp. 1137–1149, 2016.
[42] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "RTMDet: an empirical study of designing real-time object detectors," *arXivorg*, vol. abs/2212.07784, 2022.
[43] X. Lai, Z. Tian, X. Xu, Y. Chen, S. Liu, H. Zhao, L. Wang, and J. Jia, "DecoupleNet: Decoupled network for domain adaptive semantic segmentation," *Lect. Notes Comput. Sci. (LNCS)*, pp. 369–387, 2022.
[44] W. Lu, S.-B. Chen, C. H. Ding, J. Tang, and B. Luo, "LWGANet: A lightweight group attention backbone for remote sensing visual tasks," *arXiv preprint arXiv:2501.10040*, 2025.
[45] C. Xu, J. Ding, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Dynamic coarse-to-fine learning for oriented tiny object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7318–7328.
[46] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

Shanji Liu is currently pursuing the M.S. degree with the School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and remote sensing image processing.



Zhigang Yang is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing and computer vision.

Qiang Li (Member, IEEE) is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, machine learning, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/)