Anchor-SAM: Active Mining of Latent Anchors From SAM Encoder for Road Extraction

Wenhai Li, Xiaohui Huang, Xiaofei Yang, *Member, IEEE*, Yicong Zhou, *Senior Member, IEEE*, Jiangtao Peng, *Senior Member, IEEE*, Yifang Ban, *Senior Member, IEEE*, and Nan Jiang, *Member, IEEE*

***Abstract*—Road extraction in high-resolution remote sensing imagery remains a persistent challenge due to occlusions and complex backgrounds, which lead to fragmented road topologies. However, existing road extraction methods constrained by limited pretraining remote sensing data often lack the generalization capability to distinguish roads from complex backgrounds. To address this issue, we propose Anchor-SAM, a novel framework that actively mines latent semantic anchors embedded in the SAM encoder to guide topological reconstruction. Our approach stems from a pivotal insight: the SAM encoder is able to abstract complex scenes into sparse semantic anchors at deep layers, thereby implicitly encoding the global structural skeleton. To harness these implicit cues, we introduce the multiscale deformable context perceiver (MDCP) and the deformable Bayesian conditional interaction module (DBCIM). The MDCP explicitly utilizes spatial cues to aggregate global semantics across distributed anchors, establishing a robust initial context for the decoder. The DBCIM facilitates the diffusion of semantic cues to surrounding regions and effectively suppresses noise. Specifically, by leveraging the semantic certainty of anchors to guide deformable sampling trajectories, this mechanism proactively filters out background regions while precisely repairing fragmented road topologies. Our method achieves competitive performance on both the DeepGlobe and Massachusetts datasets. The source code will be publicly available at https://github.com/Hmbb0606/Anchor-SAM**

***Index Terms*—Deformable convolution, multiscale context perception, remote sensing imagery, road extraction, segment anything model.**

I. Introduction

ROAD extraction from high-resolution remote sensing imagery plays a critical role in applications such as urban planning, autonomous navigation, and disaster response [1]. However, despite considerable progress in deep learning techniques, accurate road segmentation in complex scenarios remains challenging. Specifically, fragmented road topology is often caused by tree occlusions, shadows, and interference from road-like background features [2]. In these occluded regions, inconsistent spectral features hinder the model's ability to utilize contextual information, resulting in discontinuous segmentation results.

Existing road extraction methods predominantly utilize ImageNet-pretrained CNNs [3] or ViTs [4] as backbones, aiming to enhance feature representation through architectural refinements [5], [6]. However, these methods are constrained by two intrinsic drawbacks. First, the local receptive fields of traditional CNNs limit the modeling of long-range dependencies. Second, standard backbones are constrained by the relatively small scale of pretraining data, typically on the order of millions. Consequently, limited by restricted receptive fields and data scale, these backbones tend to capture textural patterns rather than achieve global scene understanding. As a result, they lack the sufficient generalization capability to distinguish roads from complex backgrounds.

To address these constraints, we investigate large vision foundation models (VFMs) known for their strong generalization [7]. Recently, models such as SAM [8], empowered by billion-scale pretraining data, have shown impressive universal representation capabilities. However, current SAM-based adaptations for road extraction mostly rely on superficial feature fusion [9], [10], or simple fine-tuning [11], neglecting to deeply investigate the internal feature representation mechanisms of large foundation models. Motivated by this, in exploring the adaptation of the SAM2 encoder for road extraction, we observed a unique feature response pattern in Fig. 1(a): the backbone abstracts complex remote sensing scenes into a set of sparse semantic anchors within its deep layers. This contrasts sharply with the textural features typically generated by traditional encoders. These anchors are not confined to specific road intersections but are widely distributed across diverse semantic regions of the image. This implies that the backbone has implicitly encoded the global structural layout, demonstrating a comprehensive understanding of the scene context. Consequently, as illustrated in Fig. 1(b), we hypothesize that these distributed semantic anchors act as critical contextual cues, enabling global reasoning to maintain the connectivity of road networks.

However, the utilization of these sparse anchors is constrained by their intrinsic properties. While these anchors encode critical global cues, their implicit and sparse nature

Received 7 January 2026; revised 16 March 2026 and 1 April 2026; accepted 14 April 2026. Date of publication 17 April 2026; date of current version 23 April 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62462031 and Grant 62301174, in part by the Natural Science Foundation of Jiangxi Province under Grant 20242BAB26023, and in part by the Program for Overseas High-Level Talents in Urgent Need under Grant 20232BCJ25062. *(Corresponding authors: Xiaohui Huang; Xiaofei Yang.)*

Wenhai Li, Xiaohui Huang, and Nan Jiang are with the School of Information and Software Engineering, East China Jiaotong University, Nanchang 330013, China (e-mail: 2024068081200017@ecjtu.edu.cn; hxh016@hotmail.com; jiangnan@ecjtu.edu.cn).

Xiaofei Yang is with the School of Electronic and Communication Engineering, Guangzhou University, Guangzhou 511370, China (e-mail: xiaofeiyang@gzhu.edu.cn).

Yicong Zhou is with the Faculty of Science and Technology, University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo).

Jiangtao Peng is with Hubei Key Laboratory of Applied Mathematics, Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China (e-mail: pengjt1982@hubu.edu.cn).

Yifang Ban is with the Division of Geoinformatics, School of Architecture and the Built Environment, KTH Royal Institute of Technology, 11428 Stockholm, Sweden (e-mail: yifang@kth.se).

Digital Object Identifier 10.1109/TGRS.2026.3684896

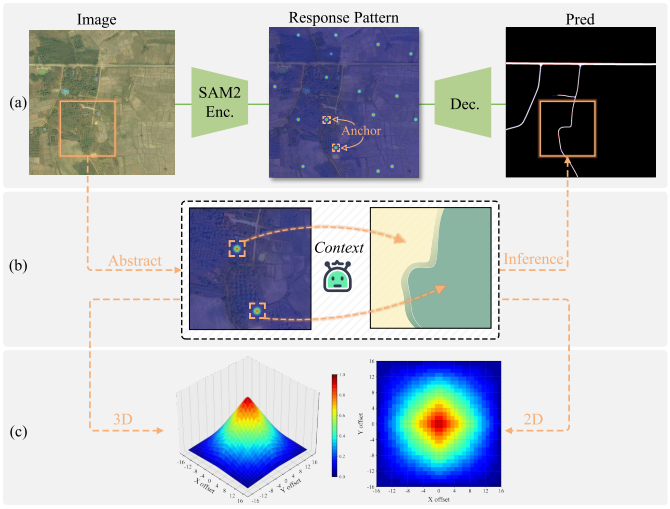

Fig. 1. Illustration of the latent semantic anchors discovered within the SAM2 encoder. (a) Input image, feature response heatmap, and final prediction. Semantic anchors are defined as the sparse high-activation regions in this heatmap. (b) Our hypothesized inference mechanism. (c) Spatial distribution of the semantic anchors visualized in 2-D and 3-D.

hinders the recovery of fine-grained pixel-level details through direct decoding. Furthermore, naive feature fusion tends to treat these prominent semantic points as considerable noise. Therefore, we need a mechanism to actively perceive and aggregate these global semantic anchors while propagating their semantic information, thereby proactively suppressing noise in regions centered around background anchors.

Building upon this distinctive feature response pattern, we propose the Anchor-SAM framework, which introduces two core components. First, to address the challenge of effectively perceiving implicit sparse anchors, we propose the multiscale deformable context perceiver (MDCP) at the bottleneck. The MDCP employs DCNv3 [12] with varying dilation rates to aggregate dispersed semantic anchors. Through the coordinated action of spatial localization via the sparse cue activator (SCA) and semantic correlation via the global relationship reasoner (GRR), it effectively integrates sparse semantic representations. Second, to address the challenge of semantic signal diffusion during feature fusion, we propose the deformable Bayesian conditional interaction module (DBCIM) within the skip connections. In the DBCIM, we introduce the concept of Bayesian maximum a posteriori (MAP) estimation, which is instantiated by the proposed Bayesian posterior estimator (BPE). Specifically, we model the semantic anchors generated by SAM2 as priors for regional semantic centroids, utilizing horizontal interaction features as observational evidence. Through probabilistic inference, the BPE generates a guidance map to optimize the directionality of vertical deformable sampling. Crucially, this mechanism utilizes anchor semantics to guide spatial propagation: background anchors actively suppress texture noise aggregation in their respective regions via the generated guidance map. The main contributions of this work are summarized as follows.

1) We reveal the intrinsic capability of the SAM2 encoder to abstract complex scenes into sparse semantic anchors. Based on this insight, we propose the Anchor-SAM framework, designed to leverage these semantic cues for road extraction.
2) We design the DBCIM. It utilizes the BPE to generate a guidance map, which optimizes vertical deformable sampling. This mechanism leverages the spatial diffusion of anchor semantics to suppress noise and repair road topology.
3) We propose the MDCP. This module employs deformable convolutions to aggregate distributed semantic anchors. Furthermore, it constructs a robust global context for the decoder by synergizing the explicit localization of the SCA with the implicit modeling of the GRR.
4) We conduct extensive experiments on the DeepGlobe and Massachusetts datasets to evaluate Anchor-SAM. The results demonstrate that our proposed method achieves competitive performance compared to state-of-the-art approaches.

II. Related Work

In this section, we review the evolution of feature encoder backbones, existing road extraction methods, and the development of deformable convolution.

A. Evolution of Feature Encoder Backbones

In remote sensing image segmentation, the encoder–decoder architecture is the dominant paradigm [13]. In this architecture, the representation capability of the feature encoder largely determines the segmentation performance. The evolution of encoders has generally progressed from convolutional neural networks (CNNs) to vision transformers (ViTs), and recently to VFMs.

Early road extraction methods primarily relied on classic CNN architectures, such as ResNet [14] and DeepLabv3+ [15]. Owing to the effective translation invariance and local inductive bias, this backbone paradigm remains prevalent in current research [16]. For instance, recent methods like TransRoadNet [17] and CGCNet [18] continue to employ ResNet encoders to capture high-frequency details such as road edges. However, their limited receptive fields restrict the modeling of long-range dependencies. This creates challenges when processing long and narrow roads, often leading to fragmented results in complex scenarios.

To address the limitation of local receptive fields, ViTs [19] and hybrid architectures have become mainstream. For example, RoadCT [20] and CMTFNet [21] combine CNNs and transformers in a parallel manner. The former uses hybrid encoders to extract complementary features, while the latter integrates local details and global semantics via a multiscale attention fusion module. However, transformers suffer from quadratic computational complexity. To solve this, state-space models (SSMs) have recently gained attention [22]. Zhao et al. [23] proposed an omnidirectional selective scan mechanism, which achieves efficient global modeling with linear complexity. Subsequently, FDMamba [24] integrates frequency-aware modeling with a rotation-aware Mamba mechanism to improve the connectivity of elongated road structures.

While these architectures improve feature extraction, they rely on supervised pretraining with limited datasets, restricting their generalization. Recently, VFMs like SAM [8] and DINOv2 [25] have learned robust visual representations from billion-scale data, offering new solutions for domain discrepancies in remote sensing. Furthermore, the rapid advancement of VFMs has spurred the development of foundation models tailored specifically for the remote sensing domain. For instance, SpectralGPT [26] leverages large-scale spectral data to construct versatile pretrained representations. Similarly, SeaMo [27] integrates multiseasonal remote sensing information to enhance the perception of surface dynamics, while Fleximo [28] focuses on flexible multimodal data fusion to adapt to diverse downstream tasks.

However, existing applications of these large models in specific tasks remain relatively preliminary. Specifically, UrbanSAM [29] employs an invariance adapter to handle scale variations in complex urban scenes, whereas other approaches directly apply simple fine-tuning [30]. These generic feature extraction strategies often overlook the sparse semantic cues embedded in the deep features of foundation models. In fact, effectively mining and employing these latent anchors is critical for accurately representing and reconstructing fragmented road topological structures.

B. Road Extraction From RSIs

Road extraction remains a challenging task due to inherent characteristics such as long spans, narrow widths, and frequent occlusions. To address these challenges, existing research primarily follows two paradigms: enhancing contextual feature representation and imposing geometric topology constraints.

To enhance contextual representation, Zhou et al. [31] utilized dilated convolutions with varying dilation rates to capture multiscale context. Xu et al. [32] developed a gated self-attention mechanism to model long-range dependencies and suppress noise, thereby improving performance in complex backgrounds. To address insufficient context and feature interference, Hua et al. [33] proposed a flexible multiscale feature extractor that enhances context perception with low computational cost. Furthermore, Yang et al. [34] developed a semantic-spatial feature refinement network that employs global–local context decoupling and refines shallow features with deep semantic guidance to suppress background noise and enhance topological consistency.

However, feature enhancement alone is often insufficient to fully resolve topological breaks caused by occlusions. Consequently, the second paradigm focuses on restoring topological relationships via prior constraints. Hu et al. [35] employed a multitask learning framework with an auxiliary road localization task to recover missing predictions. Deng et al. [36] utilized a direction-aware module and strip convolutions to explicitly reinforce connectivity during decoding. Similarly, Mei et al. [37] designed a connectivity loss function to encourage the model to capture pixel-wise neighborhood relationships. Furthermore, Chen et al. [38] introduced a focused masked image modeling strategy to learn latent dependencies between occluded and visible segments, thereby mitigating occlusion effects. To further improve geometric accuracy, Qi et al. [39] proposed dynamic snake convolution, which integrates continuity constraints into the kernel design to adaptively capture slender road features.

However, both paradigms primarily rely on extracting texture features from the image, rather than deeply understanding the scene. Consequently, they struggle to maintain topological continuity when visual evidence is ambiguous or occluded. They typically overlook the semantic anchors embedded in VFMs. To address this, we propose MDCP and DBCIM to effectively activate these anchors for road extraction.

C. Deformable Convolution

In road extraction tasks, roads typically exhibit elongated and curvilinear geometric features. However, standard CNNs are constrained by fixed geometric sampling grids, making it difficult to precisely capture these diverse and irregular shapes. To address this limitation, the deformable convolutional network (DCN) series has been widely adopted. Dai et al. [40] proposed DCNv1, which learns adaptive 2-D offsets from input features to dynamically adjust sampling positions according to the orientation of targets. Although DCNv1 improves sampling flexibility, its sampling points occasionally stray into background regions, introducing irrelevant noise. Based on DCNv1, Zhu et al. [41] introduced a modulation mechanism in DCNv2. By learning additional modulation scalars to weight each sampling point, DCNv2 enables the network to regulate the intensity of feature responses.

Subsequently, to address the demand for operator efficiency in large-scale vision tasks, Wang et al. [12] proposed DCNv3. This method proposes the design principles of depthwise separable convolution and multigroup attention mechanisms, significantly reducing the computational burden. Nevertheless, the general-purpose DCNv3 operator still employs a 2-D planar sampling strategy, which is not entirely compatible with the high aspect ratio and linear extension characteristics inherent to roads. Addressing this geometric mismatch, Yu et al. [42] proposed DSAN, a more adaptive strip-wise deformation scheme. DSAN decomposes the 2-D sampling process of DCNv3 into two orthogonal 1-D strip convolutions and utilizes a spatial attention mechanism to replace the original modulation mask.

Building upon these advancements, we incorporate the deformable mechanism into the design of our MDCP and DBCIM modules. However, unlike previous methods that rely solely on local texture features to drive deformation, we leverage semantic anchors from the foundation model as high-level guidance. By adaptively focusing on these sparse yet critical cues, our method achieves precise road network reconstruction even in complex backgrounds.

III. Proposed Method

In this section, we introduce the overall architecture of the proposed Anchor-SAM. Subsequently, we describe the DBCIM and MDCP in detail, along with the specific compositions of the encoder and decoder.

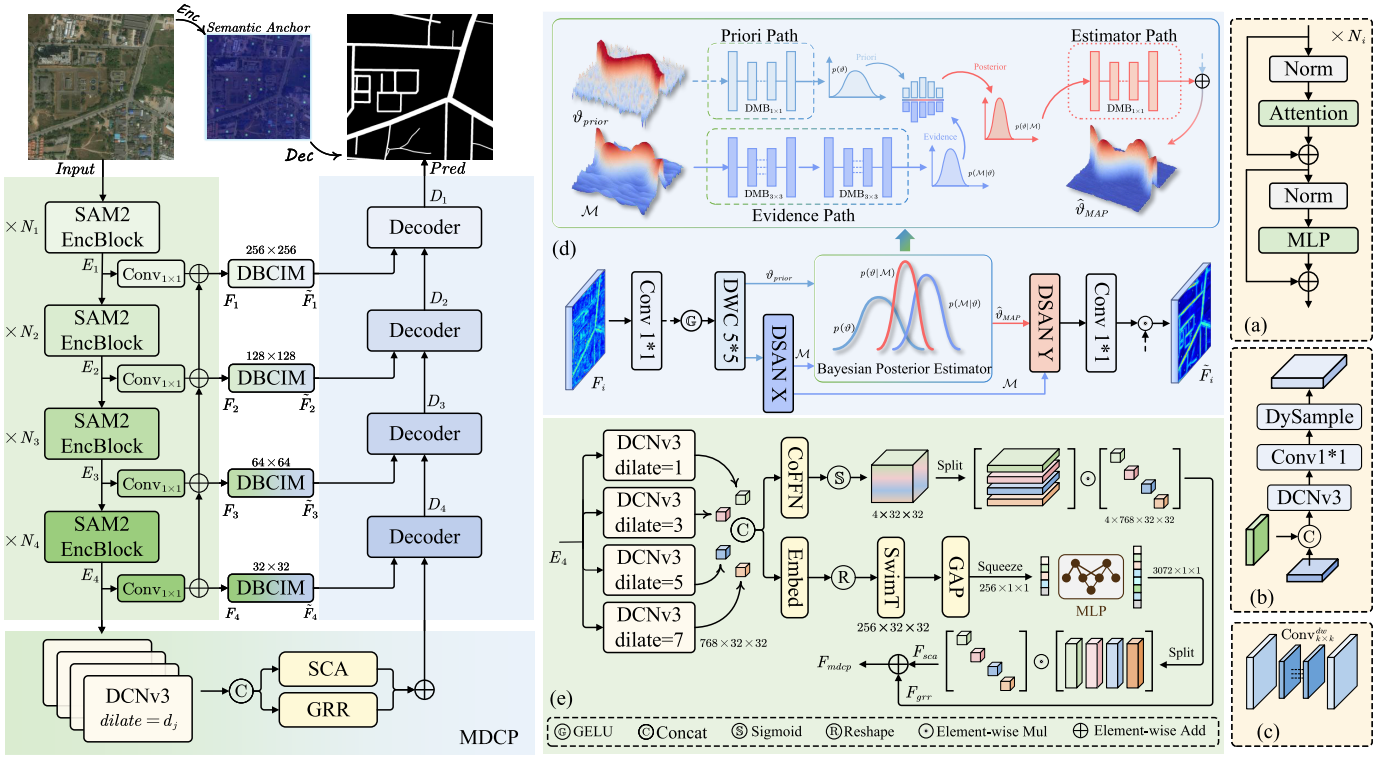

Fig. 2. Overall framework of the proposed Anchor-SAM. It is composed of five components. (a) EncBlock for hierarchical feature extraction. (b) DBCIM for leveraging anchor priors via Bayesian estimation to suppress noise. (c) MDCP for aggregating sparse semantic anchors to establish a robust global context. (d) Decoder for progressively reconstructing the continuous road network. (e) DMB for serving as the fundamental probabilistic unit within DBCIM.

A. Overview

The overall architecture of Anchor-SAM is illustrated in Fig. 2, following an encoder–decoder paradigm. We utilize the pretrained SAM2 hierarchical architecture as the backbone. The input image is processed by the encoder to generate multiscale features at four hierarchical levels, denoted as $E = \{E_1, E_2, E_3, E_4\}$. Through lateral connections within the FPN structure, the encoder produces a feature set with unified channels, $F = \{F_1, F_2, F_3, F_4\}$. Subsequently, the deepest feature E_4 is fed into the MDCP to capture sparse internal semantic cues and establish an initial global context. Meanwhile, the features F_i serve as inputs for the DBCIM via skip connections. The DBCIM utilizes a Bayesian posterior mechanism to suppress background noise and produce refined features $\tilde{F}_i$. Finally, the decoder fuses the features from the preceding decoding stage with $\tilde{F}_i$ to progressively restore spatial resolution and output the road segmentation mask.

B. Encoder

We employ the pretrained SAM2 encoder for feature extraction. It consists of the hierarchical backbone and the FPN. The backbone processes the input image to generate hierarchical features at four levels, denoted as $E = \{E_1, E_2, E_3, E_4\}$. Subsequently, the FPN projects these multiscale features into a unified channel dimension of 256 via top–down pathways. Specifically, this process starts with the deepest feature E_4 as the initial basis. For shallower features, we fuse their corresponding lateral projections with the upsampled results from the preceding deeper level. Let F_i denote the aligned feature at the ith level. This process is defined as follows:

$$F_4 = \mathrm{Conv}_{1\times1}(E_4) \tag{1}$$

$$F_i = \mathrm{Conv}_{1\times1}(E_i) + \mathcal{U}_{2\times}\left(F_{i+1}\right), \quad i \in \{3, 2, 1\} \tag{2}$$

where $\mathcal{U}_{2\times}(\cdot)$ represents the $2\times$ nearest-neighbor upsampling. Finally, the encoder yields the feature set $F = \{F_1, F_2, F_3, F_4\}$ and provides aligned inputs for the subsequent modules.

C. Deformable Bayesian Conditional Interaction Module

To globally propagate the sparse semantic anchors of SAM2 while circumventing the quadratic complexity of 2-D attention, existing approaches typically adopt an axial decoupled strategy. However, this sequential decomposition leads to information loss, as the vertical sampling step lacks access to the structural details captured during the horizontal step. To address this limitation, we design a Bayesian-inspired interaction mechanism that dynamically integrates horizontal structural evidence to construct an optimal guide map for the subsequent vertical deformation.

We define the semantic anchors as the latent geometric state ϑ and the horizontally sampled features as the observed structural evidence $\mathcal{M}$. To probabilistically model the sparse nature of these anchors, we decompose the global state ϑ into spatially independent local states ϑ_i, and formulate the prior belief $P(\vartheta)$ as a pixel-wise Bernoulli distribution

$$P(\vartheta_i|p_i) = p_i^{\vartheta_i}(1-p_i)^{1-\vartheta_i} \tag{3}$$

where $\vartheta_i \in \{0,1\}$ denotes the discrete existence state of an anchor at location i. Correspondingly, $p_i \in (0,1)$ denotes the probability of a semantic anchor existing at that location, which is continuously approximated via a Sigmoid activation in our network. Instead of explicitly defining the likelihood function $P(\mathcal{M}|\vartheta)$, we adopt an implicit probabilistic modeling perspective [43], where complex conditional distributions are parameterized directly through deep neural representations. Specifically, we draw inspiration from the multiplicative nature of Bayes' theorem to integrate the prior and the evidence

$$P(\vartheta|\mathcal{M}) \propto P(\mathcal{M}|\vartheta)\, P(\vartheta) \tag{4}$$

where $P(\vartheta|\mathcal{M})$ is the posterior state and $P(\mathcal{M}|\vartheta)$ is the likelihood of the evidence $\mathcal{M}$. To achieve MAP inference guided by this proportional relationship, we design a BPE to model this conditional interaction. By employing an element-wise gating mechanism, the prior probability acts as a soft gate to dynamically filter the horizontal structural evidence $\mathcal{M}$. This translates the theoretical Bayesian update into a differentiable feature recalibration, yielding the optimized geometric state $\hat{\vartheta}_{\text{MAP}}$ for the subsequent vertical deformation. To physically instantiate the BPE, we introduce the distribution modeling block (DMB), a transformation unit employing a bottleneck structure

$$\text{DMB}_{k\times k}(\mathbf{x}) = \text{Conv}^{c}_{1\times 1}\left(\text{Conv}^{dw}_{k\times k}\left(\text{Conv}^{\frac{c}{r}}_{1\times 1}(\mathbf{x})\right)\right) \tag{5}$$

where $\mathbf{x}$ is the input tensor, k is the kernel size, and r denotes the channel compression ratio.

The DBCIM performs the feature interaction via a structured sequential procedure. First, the input feature F_i encodes the prior ϑ_{prior} and collects horizontal structural evidence $\mathcal{M}$. In this step, ϑ_{prior} serves as both content and offset guidance

$$\vartheta_{prior} = \text{Conv}^{dw}_{5\times 5}\left(\text{GELU}\left(\text{Conv}_{1\times 1}(F_i)\right)\right) \tag{6}$$

$$\mathcal{M} = \text{DSCN}_x\left(\text{feat} = \vartheta_{prior}, \text{guide} = \vartheta_{prior}\right). \tag{7}$$

Subsequently, we design an asymmetric BPE to perform the MAP inference. The BPE models the prior $P(\vartheta)$ via a shallow DMB and the likelihood $P(\mathcal{M}|\vartheta)$ via a deep stacked DMB. The optimal state is computed by fusing these components through a residual gating mechanism, which approximates the conditional posterior expectation of the features

$$\hat{\vartheta}_{\text{MAP}} = \vartheta_{prior} + \rho\left(\sigma\left(\text{DMB}_{1\times 1}\left(\vartheta_{\text{prior}}\right)\right) \odot \text{DMB}^{\times 2}_{3\times 3}(\mathcal{M})\right) \tag{8}$$

where σ denotes the Sigmoid activation function. This allows the prior to act as a differentiable soft gate, and $\odot$ represents the element-wise Bayesian interaction. Crucially, we employ a residual connection by adding the initial ϑ_{prior} to the update term. Finally, ρ is instantiated as a $\text{DMB}_{1\times 1}$ to serve as the belief projection layer. It maps the updated probabilistic posterior back into the geometric offset domain, thereby guiding the vertical deformable convolution. Thereafter, the vertical reconstruction and feature recalibration are computed as follows:

$$Y = \text{DSCN}_y\left(\text{feat} = \mathcal{M}, \text{guide} = \hat{\vartheta}_{MAP}\right) \tag{9}$$

$$\tilde{F}_i = \text{Conv}_{1\times 1}(Y) \odot \text{Conv}_{1\times 1}(F_i) \tag{10}$$

where Y denotes the content reconstructed via vertical deformable convolution, and $\tilde{F}_i$ represents the final refined feature. This process enables the DBCIM to efficiently propagate sparse semantic anchors along posterior-optimized trajectories, thereby suppressing background noise and repairing fragmented road topology.

D. Multiscale Deformable Context Perceptor

The heatmap visualization in Section IV-E demonstrates that encoder features manifest as sparse semantic cues starting from deeper levels. Consequently, the MDCP utilizes the deepest feature E_4 as input to consolidate these discrete anchors into a robust global context. Specifically, the module employs four parallel branches configured with increasing dilation rates $d_j \in \{1,3,5,7\}$ to capture multiscale information, providing progressive initial receptive fields to adaptively aggregate spatial context

$$branch_j = \text{DCN}_j(E_4), \quad j \in \{1,2,3,4\} \tag{11}$$

where $branch_j$ denotes the feature extracted by the jth branch, and DCN_j represents the deformable convolution [12] with the corresponding dilation rate. To effectively utilize these multiscale features, we design two parallel perception branches.

1) Spatial Cue Activator: We employ the SCA to precisely localize semantic cues. First, the module aggregates multiscale features to generate a spatial saliency map $\mathbf{W}_{sca}$, which perceives key cues in the spatial dimension

$$\mathbf{W}_{sca} = \sigma\left(\text{ConvFFN}\left(\text{Cat}\left(branch_j\right)\right)\right) \tag{12}$$

where Cat$(\cdot)$ denotes channel concatenation, ConvFFN$(\cdot)$ is a bottleneck convolution, and σ is the sigmoid activation function. Subsequently, the generated weights are split and applied to each branch to enhance spatial responsiveness

$$F_{sca} = \sum_{j=1}^{4} \text{Split}_j(\mathbf{W}_{sca}) \odot branch_j \tag{13}$$

where $\text{Split}_j(\cdot)$ slices the weight map along the channel dimension to match the jth branch, and $\odot$ denotes the element-wise product for spatial reweighting.

2) Global Relation Reasoner: Building upon spatial activation, the GRR aims to further establish semantic consistency among discrete anchors. First, we utilize a Swin transformer to capture long-range dependencies and aggregate them into a global semantic vector v_{global}

$$v_{global} = \text{AvgPool}\left(\text{Swin}\left(\text{Cat}\left(branch_j\right)\right)\right) \tag{14}$$

where Swin$(\cdot)$ represents the Swin transformer block used for global reasoning, and AvgPool$(\cdot)$ compresses the spatial information into a global descriptor. Then, this vector is projected to recalibrate the channel responses of each branch

$$F_{grr} = \sum_{j=1}^{4} \text{Split}_j\left(\sigma\left(\text{MLP}\left(v_{global}\right)\right)\right) \odot branch_j \tag{15}$$

where MLP maps the global vector to channel-wise weights.

Finally, the MDCP fuses the spatially activated and semantically reasoned features to provide a robust initial bottleneck input for the decoder

$$F_{mdcp} = F_{sca} + F_{grr}. \tag{16}$$

The resulting feature F_{mdcp} encapsulates a comprehensive representation wherein internal semantic interest points are both explicitly activated in space and logically verified in semantics, thereby effectively guiding the subsequent road network reconstruction.

E. Decoder

The decoder is responsible for fusing the global context cues from the MDCP with the road detail features from the DBCIM to reconstruct the high-resolution road network. We employ deformable convolutions in the decoder to dynamically adjust the receptive field according to the road shape. Finally, DySample [44] is adopted as the upsampling operator to maintain clear boundaries while restoring spatial resolution.

The decoding process proceeds through four stages. In each stage, we concatenate the features decoded from the upper level with the refined skip connection $\tilde{F}_i$. The formula for the decoding phase at stage i is

$$D_i = \text{DySample}\left(\text{DCN}\left(\text{Cat}\left(D_{i+1}, \tilde{F}_i\right)\right)\right) \tag{17}$$

where D_{i+1} represents the feature from the previous deeper stage (with the initial input $D_5 = F_{mdcp}$). After restoring the feature map to the original resolution, a final convolution layer projects the features to generate the binary road segmentation mask.

IV. Experiments

In this section, we evaluate Anchor-SAM on the DeepGlobe and Massachusetts datasets. First, we introduce the experimental setup and implementation details. Then, we elucidate the discovery of semantic anchors via heatmap visualizations. Finally, we compare with state-of-the-art methods and validate the core components through ablation studies.

A. Datasets

To comprehensively evaluate our proposed Anchor-SAM method, we conducted experiments on the DeepGlobe [45] and Massachusetts [46] datasets. These two datasets offer strong complementarity in spatial resolutions and geographical topographical features, serving to validate our model's capability in extracting road topology under complex occluded environments.

The DeepGlobe dataset provides 6226 high-resolution image pairs with a spatial resolution of 0.5 m/pixel and dimensions of 1024×1024 pixels. Collected primarily from Thailand, Indonesia, and India, this imagery presents an inherent topographical challenge because dense vegetation frequently occludes the road networks. For our experiments, we established an 8:1:1 split ratio by randomly allocating 4981 image pairs for training, 623 for validation, and 622 for testing.

The Massachusetts road dataset encompasses diverse urban, suburban, and rural landscapes characterized by severe interference from complex building environments. It comprises 1171 aerial image pairs with a resolution of 1 m/pixel and a larger size of 1500×1500 pixels. While the official configuration designates 1108 image pairs for training, 14 for validation, and 49 for testing, we identified and manually rectified inherent errors in the original ground truth across 364 training and two validation image pairs to prevent performance degradation, as illustrated in Fig 3. Subsequently, these corrected image pairs were cropped into overlapping 1024×1024 patches to align with the model input requirements.

Fig. 3. Error samples in the Massachusetts dataset. (a) Original imagery from the Massachusetts dataset. (b) Erroneous mask result (blank regions incorrectly labeled as road). (c) Corrected mask result (blank regions corrected to background).

B. Implementation Details

Our model was implemented using the PyTorch 2.2.0 framework, and all experiments were conducted on a single NVIDIA RTX 4090 GPU platform.

We employed the AdamW optimizer combined with a cosine annealing scheduler, where the minimum learning rate was set to 1×10^{-6}. The weight decay was consistently set to 1×10^{-4}, and the batch size was fixed at 2. We adopted dataset-specific initial learning rates: 1×10^{-4} for the DeepGlobe dataset and 2×10^{-4} for the Massachusetts dataset. Additionally, we utilized automatic mixed precision (AMP) to conserve GPU memory and accelerate training. For model initialization, the encoder was loaded with pretrained weights from SAM2. We employed a full fine-tuning strategy to ensure the model effectively adapted to the feature distribution of the road extraction task. The evaluation metrics follow those used in [14].

To supervise the network training, we utilize a hybrid loss function composed of binary cross-entropy loss [47] and Dice loss [48]

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{Dice} + \beta \cdot \mathcal{L}_{BCE} \tag{18}$$

where α and β are balancing coefficients, both of which are set to 1 in our experiments.

Furthermore, we developed four variants based on the SAM2 encoder scales: Anchor-SAM-T, -S, -B, and -L. These variants are distinguished by the block counts in their four hierarchical stages, denoted as $\{N_1, N_2, N_3, N_4\}$. Specifically, the configurations range from the lightweight tiny $\{1, 2, 7, 2\}$ and small $\{1, 2, 11, 2\}$ versions to the deeper base $\{2, 3, 16, 3\}$ and large $\{2, 6, 36, 4\}$ models. For comprehensive performance evaluation, the selected baseline models encompass classic CNN-based architectures such as UNet [13], DeepLabv3+ [15], and D-LinkNet [31]. Furthermore, we incorporate representative approaches from the last three years to cover various structural paradigms. These include recent CNN-based

TABLE I
ROAD EXTRACTION RESULTS ON THE DEEPGLOBE ROAD DATASET AND THE MASSACHUSETTS ROAD DATASET

Method	DeepGlobe Dataset (1024×1024)					Massachusetts Dataset (1024×1024)				
	Precision ↑	Recall ↑	F1-score ↑	IoU ↑	APLS ↑	Precision ↑	Recall ↑	F1-score ↑	IoU ↑	APLS ↑
$UNet_{2015'}$ [13]	78.15	78.56	78.35	64.41	65.63	75.92	75.89	75.90	61.17	60.14
$Deeplabv3+_{2018'}$ [15]	81.81	77.32	79.50	65.98	69.19	78.10	73.95	75.97	61.25	61.43
$D\text{-}LinkNet_{2018'}$ [31]	81.72	80.79	81.25	68.42	72.57	80.12	76.55	78.30	64.34	69.58
$TransRoadNet_{2022'}$ [17]	81.03	81.47	81.25	68.42	71.21	80.91	74.95	77.81	63.68	68.33
$CMTFNet_{2023'}$ [21]	80.36	78.65	79.50	65.98	69.74	80.81	76.21	78.44	64.53	69.99
$C^2Net_{2024'}$ [14]	78.63	78.49	78.56	64.69	68.26	80.99	74.90	77.82	63.69	71.08
$StripUnet_{2024'}$ [16]	81.61	80.05	80.82	67.81	71.95	81.92	74.27	77.91	63.81	69.07
$RS\text{-}Mamba_{2024'}$ [23]	82.21	82.76	82.49	70.20	76.16	80.41	77.35	78.85	65.08	70.05
$CGCNet_{2025'}$ [18]	81.47	79.61	80.53	67.41	69.01	79.86	77.18	78.50	64.61	65.34
$AFDANet_{2025'}$ [36]	81.42	79.07	80.23	66.99	71.15	81.58	75.07	78.19	64.19	70.48
$RSAM\text{-}Seg_{2025'}$ [30]	76.25	78.42	77.32	63.02	61.50	77.45	76.18	76.81	62.35	66.66
Anchor-SAM-T	83.97	83.25	83.61	71.84	76.16	80.81	76.44	78.56	64.69	69.40
Anchor-SAM-S	84.57	83.21	83.88	72.24	77.01	80.48	77.47	78.95	65.22	71.94
Anchor-SAM-B	83.98	84.18	84.08	72.53	77.07	79.77	**78.36**	79.06	65.37	72.04
Anchor-SAM-L	**85.01**	**84.53**	**84.77**	**73.57**	**80.26**	**82.48**	77.90	**80.11**	**66.82**	**72.91**

methods like C^2Net [14], StripUnet [16], and CGCNet [18], as well as CNN-transformer hybrid architectures represented by TransRoadNet [17], CMTFNet [21], and AFDANet [36]. Finally, we extend our comparisons to SSMs like RS-Mamba [23], and SAM-based adaptations including RSAM-Seg [30].

C. Experimental Results and Analysis

In this section, we conduct comparative experiments on the DeepGlobe and Massachusetts datasets to evaluate our proposed framework against current state-of-the-art methods.

The comparative results of Anchor-SAM and its variants on both datasets are presented in Table I, where Anchor-SAM demonstrates consistently superior performance. In the DeepGlobe dataset, Anchor-SAM-L achieves the best results with an IoU of 73.57% and an $F1$-score of 84.77%. Compared to RS-Mamba, which performs competitively among the baselines, Anchor-SAM-L achieves an IoU improvement of approximately 3.37%. Notably, the recently proposed RSAM-Seg achieves relatively lower performance, obtaining an IoU of 63.02 on the DeepGlobe dataset. This unexpected underperformance likely stems from its generalized feature adaptation approach designed for diverse remote sensing tasks. Conversely, our framework effectively adapts the SAM2 encoder to complex remote sensing scenes, contributing to the steady performance gains. Additionally, the lightweight Anchor-SAM-T outperforms state-of-the-art methods such as TransRoadNet and C^2Net with an APLS of 76.16, thereby validating the structural effectiveness of our design.

D. Visual Results

Consistent with the previous findings, Anchor-SAM-L demonstrates competitive performance on the Massachusetts dataset, yielding an $F1$-score of 80.11%, an IoU of 66.82%, and an APLS of 72.91. Notably, compared to other methods, Anchor-SAM-B demonstrates a high recall of 78.36% on this dataset. Such a prominent recall indicates an enhanced ability to minimize false negatives, directly aligning with the design motivation of our MDCP module. By actively aggregating distributed semantic anchors, the model captures long-range dependencies more effectively, thereby facilitating the repair of topological breaks in occluded road regions. Furthermore, we observe that Anchor-SAM maintains high precision alongside high recall. For instance, Anchor-SAM-L achieves a precision of 85.01% on the DeepGlobe dataset. This balance can be attributed to the DBCIM introduced in the skip connections, which assists the model in distinguishing road-like background interference, consequently reducing false positives.

TABLE II
QUANTITATIVE EVALUATION OF COMPUTATIONAL COMPLEXITY AND INFERENCE SPEED FOR DIFFERENT METHODS

Method	Params (M)	FLOPs (G)	FPS
$UNet_{2015'}$ [13]	31.04	1751.61	23.19
$Deeplabv3+_{2018'}$ [15]	61.39	2088.67	19.65
$D\text{-}LinkNet_{2018'}$ [31]	31.10	268.69	63.48
$TransRoadNet_{2022'}$ [17]	36.01	647.47	42.18
$CMTFNet_{2023'}$ [21]	30.07	261.97	47.70
$C^2Net_{2024'}$ [14]	61.51	980.43	27.15
$StripUnet_{2024'}$ [16]	58.69	504.46	13.66
$RS\text{-}Mamba_{2024'}$ [23]	50.38	298.22	13.65
$CGCNet_{2025'}$ [18]	**8.51**	**222.09**	**65.19**
$AFDANet_{2025'}$ [36]	29.59	351.08	28.08
$RSAM\text{-}Seg_{2025'}$ [30]	350.29	1429.95	3.10
Anchor-SAM-T	49.94	432.09	18.61
Anchor-SAM-S	57.03	493.96	17.65
Anchor-SAM-B	96.66	768.44	15.24
Anchor-SAM-L	251.91	1879.55	8.97

As shown in Table II, to comprehensively assess the practical value of Anchor-SAM, we compare its model complexity and inference speed against existing methods. Anchor-SAM-S consists of 57.03 M parameters, representing a smaller scale compared to the 350.29 M parameters of the RSAM-Seg. Additionally, Anchor-SAM-S achieves an inference speed of 17.65 FPS, vastly outperforming RSAM-Seg

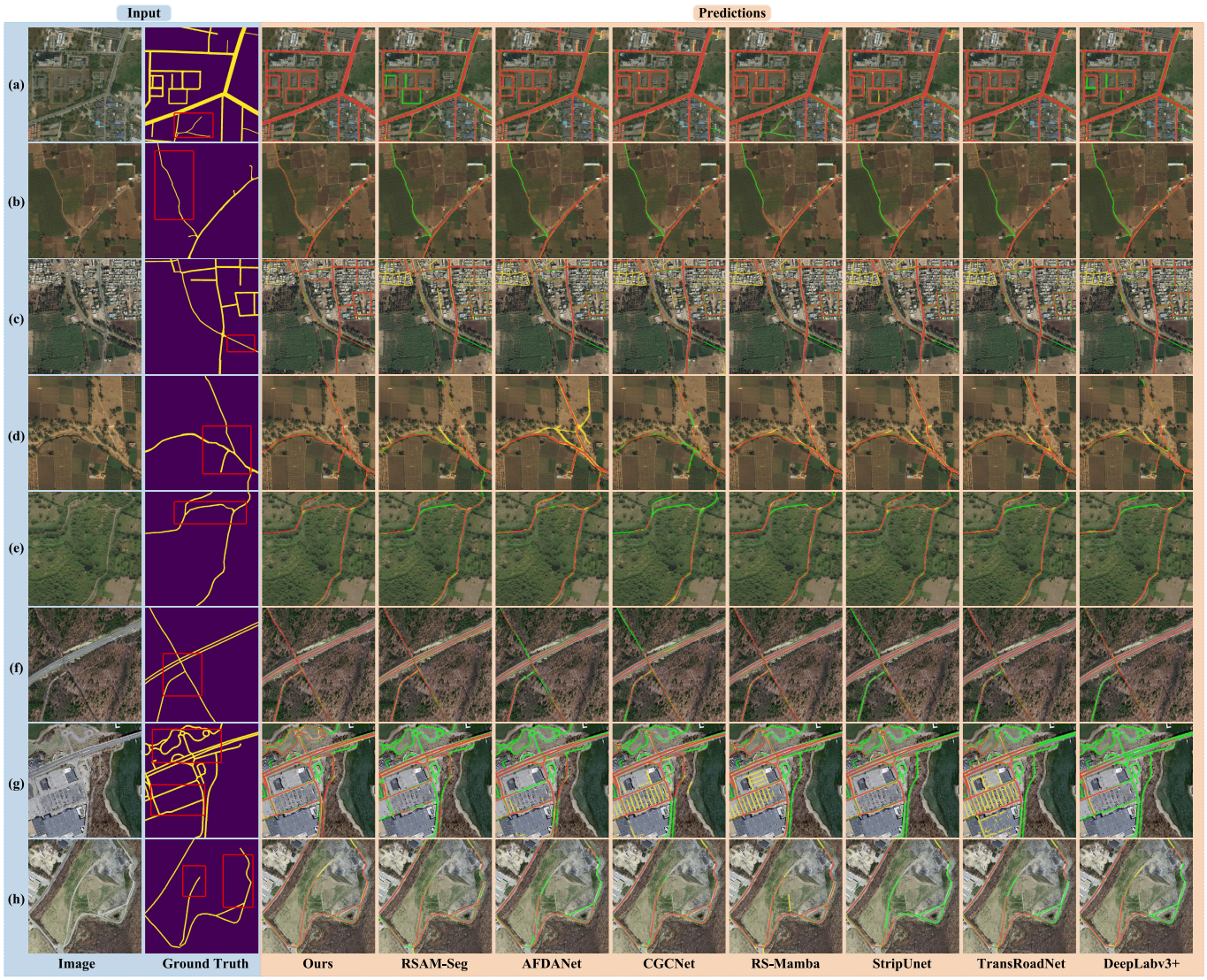

Fig. 4. Comparative results of different network models on (a)–(e) DeepGlobe dataset and (f)–(h) Massachusetts dataset. True positives, false positives, and false negatives are marked in red, yellow, and green, respectively. Predictions are overlaid on the original images for better visualization.

at 3.10 FPS and RS-Mamba at 13.65 FPS. This represents a speed improvement of nearly 29.3% over RS-Mamba while maintaining a comparable model scale. Furthermore, the lightweight Anchor-SAM-T achieves higher detection accuracy with a computational cost of 432.09 GFLOPs, which is considerably lower than the massive 1429.95 GFLOPs required by RSAM-Seg. Overall, these results highlight the superiority of our proposed model in terms of efficiency and practicality compared to existing road extraction methods.

Representative scenes from the DeepGlobe and Massachusetts datasets were selected to benchmark Anchor-SAM-T against other competing methods. To provide a qualitative visualization of our model's performance, we superimposed the predicted results onto the original images, as illustrated in Fig. 4.

1) Performance in Urban Areas: Urban environments frequently feature abundant regular geometric structures. For example, the narrow building gaps in Fig. 4(a) and (c) and parking markings in Fig. 4(g) pose notable background interference. Anchor-SAM demonstrates robust discriminative power in such complex scenarios, as evidenced by the visual results. In particular, as shown in Fig. 4(g), false detections on the parking lines are effectively suppressed, while the irregular road regions in the upper portion are captured with high accuracy.

2) Scenes With Background Similarity: As observed in Fig. 4(b), the linear features of field ridges often induce visual confusion. Consequently, methods such as RSAM-Seg and CGCNet exhibit noticeable discontinuities in these regions, suggesting limitations in capturing subtle features. Similarly, in Fig. 4(d), the background color bears a high resemblance to the road surface, leading AFDANet to generate a substantial number of false positives. In contrast, our proposed method maintains improved connectivity and accuracy in these areas. This enhanced context capture capability is largely attributed to the effective aggregation of semantic anchors by the MDCP. This design allows the model to effectively leverage contextual information and perform joint inference to obtain more accurate road regions.

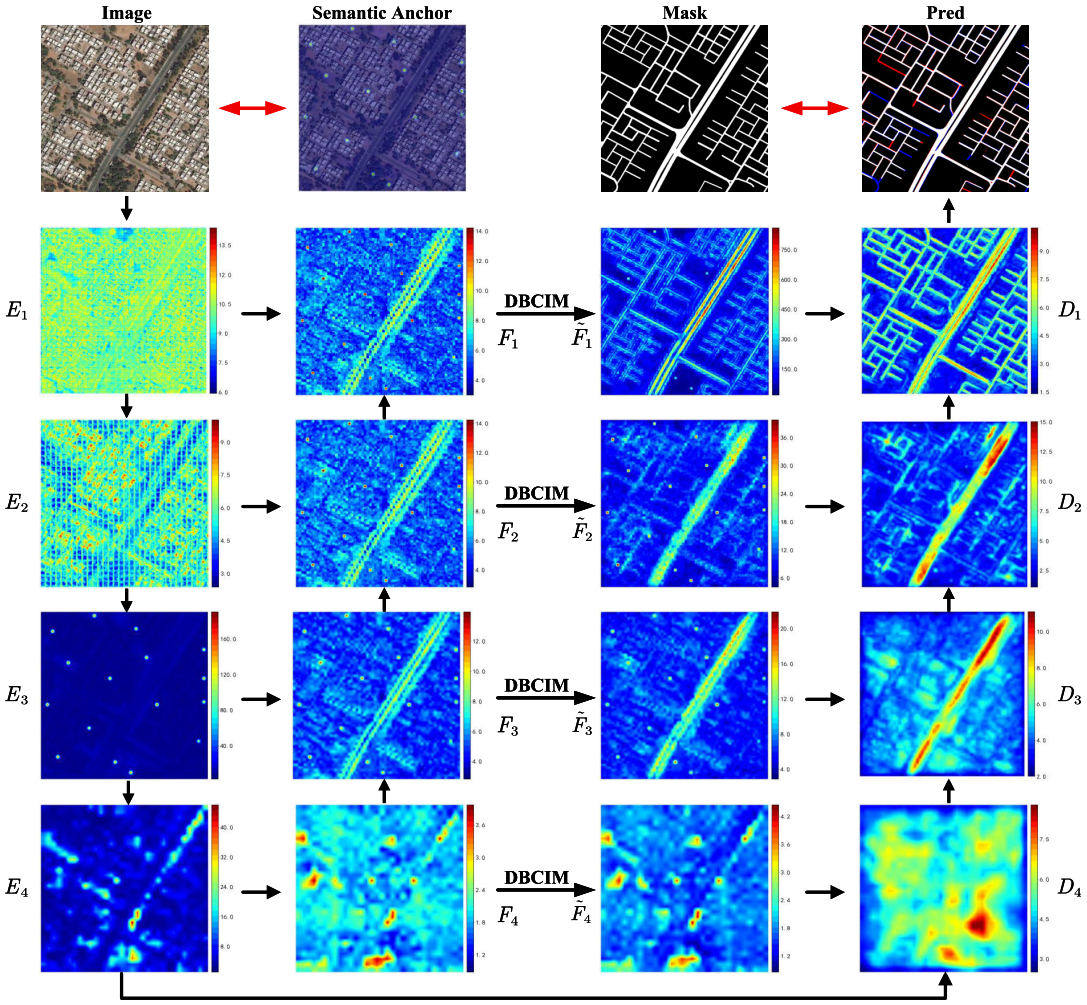

Fig. 5. Visualization of feature activations within Anchor-SAM-T. The figure displays the hierarchical evolution of features, covering the SAM2 encoder layers (EncBlock and FPN), skip connections, and decoder outputs. In the heatmaps, red indicates high activation values, while blue represents low activation.

3) Scenes With Tree Occlusion: In Fig. 4(c) and (e), roads are severely occluded by dense vegetation or building shadows, presenting challenging scenarios. When the road surface is obscured, comparative methods struggle to maintain long-range semantic consistency, leading to fragmented predictions. This observation suggests that models relying primarily on local visual texture features are susceptible to errors when explicit visual information is missing. In contrast, Anchor-SAM preserves robust topological integrity in these occluded regions. This implies that the model moves beyond the reliance on local texture features and effectively leverages the implicitly encoded global structural context to infer occluded paths, thereby facilitating the restoration of topological connectivity.

E. Heatmap Analysis

To facilitate an understanding of the internal representations within Anchor-SAM, we visualize the feature maps from key stages using heatmaps, as illustrated in Fig. 5. The model takes an *Image* as input and generates a final prediction *Pred*. From left to right, the columns correspond to: the encoder output features E_i, the FPN features F_i, the features $\tilde{F}_i$ output by the DBCIM, and the decoder reconstruction features D_i.

1) Evolution of Encoder Features: As illustrated in Fig. 5, at the shallow stages (E_1, E_2), the feature maps progressively exhibit distinct grid-like textures. This phenomenon is fundamentally attributed to the window attention mechanism inherent in the hierarchical architecture, where the image is partitioned into nonoverlapping windows to calculate local attention. The elevated activation values at the grid boundaries suggest that the model is sensitive to feature discontinuities at window margins, while simultaneously capturing high-frequency texture details within the windows, such as building edges and road boundaries. As the network deepens from E_1 to E_2, these grid patterns become more pronounced, indicating that the model effectively aggregates local textures into preliminary structural units.

2) Emergence of Semantic Anchors: As the network deepens, we observe distinct transitions in the heatmaps. At the E_3 stage, rigid grid patterns dissipate to give way to sparse high-activation regions as shown in Fig. 5. This phenomenon is primarily attributed to the introduction of global attention blocks in EncBlock and the substantial expansion of the receptive field inherent to the deepening hierarchy. Consequently, the model surpasses the limitations of local windows, shifting

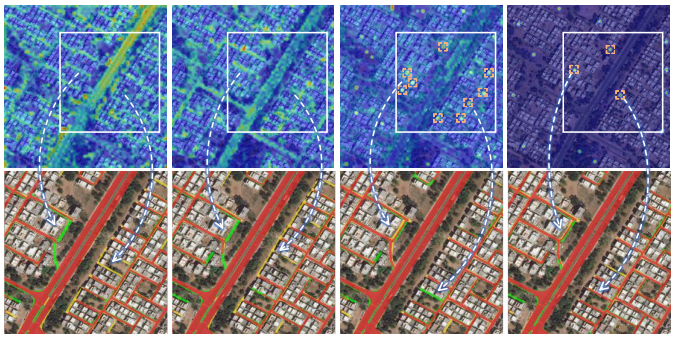

Fig. 6. Comparison of E_3 stage heatmaps and local predictions for different Anchor-SAM scales. (Left to right) Correspond to Anchor-SAM-T, S, B, and L. Predictions are overlaid on raw images for better visualization. Colors indicate TP (red), FP (yellow), and FN (green).

its focus toward broader semantic information. We refer to these distinct high-activity regions as semantic anchors. In particular, their distribution is not limited to road intersections, but extends across dense clusters of buildings and road margins. This pattern indicates that the encoder effectively captures global semantic context, which in turn facilitates the accurate segmentation of road regions.

3) Denoising and Topology Repair via DBCIM: A comparison between the aligned features F_i and the refined features $\tilde{F}_i$ highlights the noise suppression efficacy of the DBCIM. While F_i facilitates channel alignment, it retains residual grid artifacts and background noise inherited from the encoder. For instance, in F_2 and F_3, vertical and horizontal stripe patterns remain visible. However, following DBCIM processing, this noise is effectively suppressed, thereby enhancing the distinctness of the road regions.

4) Dissipation of Semantic Anchors: We observe in Fig. 5 that semantic anchors are suppressed in shallow DBCIM stages yet remain prominent in deeper layers, eventually fading during the decoding phase. This phenomenon is likely attributed to two primary factors. First, the deformable convolutions within the decoder diffuse the information from these anchor points across various spatial directions, leading to their gradual disappearance. Second, the GRR within the MDCP module exerts an implicit influence. Since the GRR operates primarily along the channel dimension, its effects are challenging to visualize directly. However, theoretically, the GRR leverages Swin transformer blocks to establish contextual connections between regions of interest and their surroundings, thereby facilitating joint inference for global semantics.

F. Properties of Semantic Anchors

To further explore the physical characteristics of semantic anchors in Anchor-SAM, we visualize the heatmaps of models across different parameter scales. As indicated by the white boxes in Fig. 6, the heatmaps of Anchor-SAM-T and Anchor-SAM-S primarily focus on the road surface itself, exhibiting a distinct tendency toward texture dependence. In contrast, the heatmaps of Anchor-SAM-B and Anchor-SAM-L reveal a unique semantic anchor mechanism, gradually breaking away from the local texture paradigm. This indicates that an increase in parameters drives a paradigm shift in the model, transitioning from low-level texture feature extraction to high-level global scene understanding.

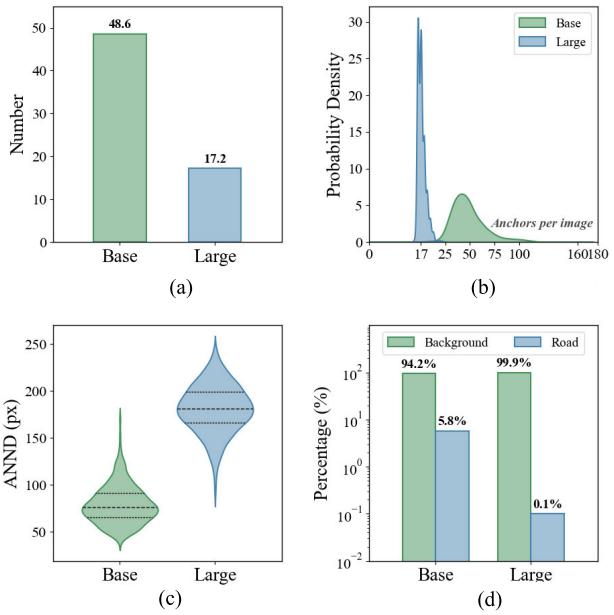

Fig. 7. Statistical analysis of the properties of semantic anchors extracted by Anchor-SAM-B and Anchor-SAM-L on the DeepGlobe test set. (a) The average number of semantic anchors per image reflects the sparsity of these anchors. (b) Distribution of anchor counts is evaluated across the entire dataset. (c) Spatial distance distribution between anchors demonstrates their spatial discreteness. (d) Semantic directionality of the anchors illustrates their localization distribution on road versus background regions.

We conduct a quantitative statistical analysis to investigate the physical properties of the semantic anchors in Anchor-SAM-B and Anchor-SAM-L. As illustrated in Fig. 7, this analysis covers three dimensions: anchor quantity, spatial discreteness, and semantic directionality.

1) Anchor Quantity: Regarding the number of semantic anchors, as shown in Fig. 7(a), Anchor-SAM-L extracts an average of 17.2 anchors per image, compared to the 48.6 anchors of Anchor-SAM-B. Furthermore, the probability density distribution in Fig. 7(b) indicates that the anchor count distribution of Anchor-SAM-L is more concentrated. These results confirm the sparsity of semantic anchors.

2) Spatial Distribution: To evaluate the spatial arrangement of anchors, we employ the average nearest neighbor distance (ANND) to measure the pixel-level Euclidean distance among them. Fig. 7(c) demonstrates that the semantic anchors of Anchor-SAM-L generally avoid aggregation in local regions. Instead, they maintain larger distances from each other. This demonstrates their strong spatial discreteness compared to the Base model.

3) Semantic Allocation: Regarding the semantic placement of anchors, Fig. 7(d) reveals an interesting pattern. Most semantic anchors in both Anchor-SAM-B and Anchor-SAM-L are located in the background regions rather than on the road bodies. This indicates that the spatial distribution of semantic anchors is physically separated from their target regions. They act differently from traditional center pixels that are densely distributed inside the roads.

TABLE III
ABLATION RESULTS OF THE PROPOSED DBCIM AND MDCP COMPONENTS IN ANCHOR-SAM ON THE DEEPGLOBE AND MASSACHUSETTS DATASETS

Method	Components			DeepGlobe					Massachusetts				
	SAM Enc.	DBCIM	MDCP	Precision	Recall	F1	IoU	APLS	Precision	Recall	F1	IoU	APLS
Baseline	✓			82.05	79.16	80.58	67.48	66.98	81.34	74.52	77.78	63.64	66.85
+ DBCIM	✓	✓		82.57	80.47	81.51	68.79	66.34	79.26	77.21	78.22	64.23	65.60
+ MDCP	✓		✓	82.58	79.37	80.94	67.98	66.80	81.00	75.59	78.20	64.20	66.01
Ours	✓	✓	✓	**85.01**	**84.53**	**84.77**	**73.57**	**80.26**	**82.48**	**77.90**	**80.11**	**66.82**	**72.91**

TABLE IV
INTERNAL ABLATION RESULTS OF THE MDCP ON THE DEEPGLOBE AND MASSACHUSETTS DATASETS

Method	Components			DeepGlobe					Massachusetts				
	DCN	SCA	GRR	Precision	Recall	F1	IoU	APLS	Precision	Recall	F1	IoU	APLS
w/o DCN		✓	✓	82.37	80.26	81.30	68.49	65.40	80.29	76.29	78.24	64.26	66.18
w/o SCA	✓		✓	81.40	80.92	81.16	68.29	64.88	80.68	75.47	77.98	63.91	64.39
w/o GRR	✓	✓		82.41	80.47	81.43	68.68	65.37	79.94	76.47	78.17	64.16	65.46
Ours	✓	✓	✓	**85.01**	**84.53**	**84.77**	**73.57**	**80.26**	**82.48**	**77.90**	**80.11**	**66.82**	**72.91**

Based on the quantitative analysis above, we summarize the physical meaning of semantic anchors. The model utilizes sparse and spatially discrete anchors distributed in background regions to establish scene understanding. This approach helps the network mitigate the impact of occlusions and infer the road topology from a global perspective.

We further evaluate the impact of anchor distribution on segmentation performance via heatmaps and prediction maps. As shown in Fig. 6, the anchors generated by Anchor-SAM-B are noticeably denser than those of Anchor-SAM-L. However, in the segmentation results indicated by the arrows, Anchor-SAM-B shows more missed detection regions. We attribute this to the global representation capability of Anchor-SAM-B remaining in a transitional stage, making it difficult to effectively merge redundant local anchors. This limitation leads to mutual interference in the feature space.

G. Ablation Study

To comprehensively evaluate the efficacy of the core components and internal mechanisms proposed in the Anchor-SAM framework, we performed detailed ablation studies on the DeepGlobe and Massachusetts datasets. All experiments were conducted using the Anchor-SAM-T model.

1) Effectiveness Analysis of Core Components: We designated the network utilizing only the pretrained SAM2 encoder as the baseline model. As shown in Table III, the baseline exhibited relatively high APLS across both datasets. However, recall and IoU were comparatively low, with a recall of 79.16% and an IoU of 67.48% on DeepGlobe. This indicates that while the SAM2 encoder yields high-quality semantic features, it requires additional mechanisms for effective context aggregation and detail recovery.

Upon incorporating the DBCIM into the baseline, the recall on the DeepGlobe dataset improved by 1.31% to reach 80.47%, while the IoU increased to 68.79%. This suggests that the DBCIM effectively suppresses noise during the encoding process, thereby improving the capability to recall road details.

TABLE V
INTERNAL ABLATION RESULTS OF THE DBCIM ON THE DEEPGLOBE AND MASSACHUSETTS DATASETS

Dataset	Method	Precision	Recall	F1	IoU	APLS
DeepGlobe	w/o BPE	82.57	79.77	81.15	68.28	63.62
	Ours	**85.01**	**84.53**	**84.77**	**73.57**	**80.26**
Massachusetts	w/o BPE	78.72	76.93	77.81	63.68	65.11
	Ours	**82.48**	**77.90**	**80.11**	**66.82**	**72.91**

Integrating the MDCP alone raised the IoU on DeepGlobe to 67.98% while maintaining precision at a high level (82.58%). This validates that the MDCP effectively constructs a robust global context by aggregating sparse semantic anchors, thus enhancing the model's perception of the overall road topology. When both DBCIM and MDCP are integrated, the model achieves superior performance across all metrics on both datasets. This demonstrates the strong complementarity between the two modules in feature enhancement and contextual reasoning.

2) Analysis of Mdcp Internal Mechanisms: We conducted further ablation studies on the MDCP, with results presented in Table IV. Replacing the parallel DCNv3 branches with standard convolutions led to a decrease in IoU on the Massachusetts dataset to 64.26%. This illustrates that utilizing DCNv3 with varying dilation rates to actively perceive multiscale context is essential for capturing distributed semantic anchors. Meanwhile, removing either the SCA or the GRR resulted in performance degradation. Specifically, excluding the SCA reduced the IoU on Massachusetts to 63.91%, confirming the role of explicit spatial activation in guiding feature focus.

3) Effectiveness of the BPE: We assessed the advantage of the Bayesian-inspired interaction mechanism within the DBCIM through comparative experiments. As shown in Table V, removing the BPE and directly utilizing the initial prior ϑ_{prior} as the guidance feature for vertical sampling yields an IoU of 68.28% on the DeepGlobe dataset. In contrast,

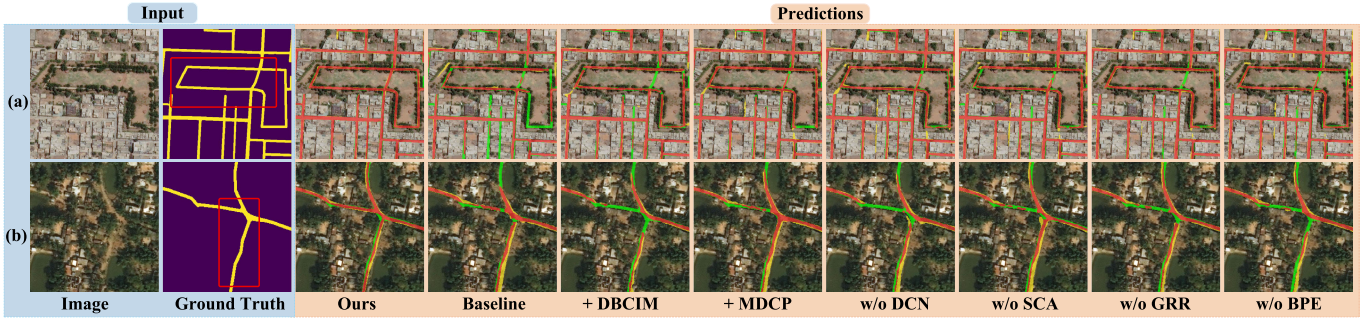

Fig. 8. Qualitative comparison of ablation experiments on the DeepGlobe dataset. True positives, false positives, and false negatives are marked in red, yellow, and green, respectively. Predictions are overlaid on the original images for better visualization. (a) and (b) Representative samples from the DeepGlobe dataset.

employing the BPE to dynamically integrate structural evidence achieves an IoU of 73.57%. This 5.29% improvement in IoU confirms that dynamically integrating horizontal structural evidence effectively mitigates the information loss in sequential decomposition, yielding an optimal guide map for vertical sampling.

4) Qualitative Analysis of the Proposed Components: We provide a qualitative comparison of the ablation configurations on the DeepGlobe dataset. In Fig. 8(a), the scenes contain complex backgrounds and roads severely occluded by dense vegetation or buildings. The baseline model is easily affected by these factors, resulting in obvious background noise and fragmented road segments. Adding the DBCIM effectively removes false positives, but topological breaks still remain. In contrast, using the MDCP alone reconnects the fragmented roads but leaves some false positives. However, when both modules work together, false positives are greatly reduced, and road connectivity is highly improved. Furthermore, the internal ablation results show that removing specific components leads to clear performance drops. For example, without the SCA and GRR modules in Fig. 8(b), false negatives and topological breaks still exist. Relying only on the prior without the BPE module cannot effectively reduce false positives. When all proposed modules are combined, most prediction errors are removed, and road connectivity is further enhanced. Therefore, our model achieves the best visual results in road extraction.

V. Conclusion

We propose Anchor-SAM, a novel framework actively mining latent semantic anchors within the SAM encoder to address road fragmentation caused by occlusions. Leveraging the insight that the encoder acts as a global structural skeleton, we design MDCP to aggregate global semantics and DBCIM to guide topological reconstruction while suppressing noise. Extensive experiments on the DeepGlobe and Massachusetts datasets demonstrate that Anchor-SAM achieves superior performance over state-of-the-art methods.

Despite the performance improvements achieved by Anchor-SAM, the current framework exhibits certain limitations. To mitigate the substantial computational overhead of the Anchor-SAM series, we intend to explore knowledge distillation techniques, transferring this anchor mechanism into lightweight networks to reduce deployment costs. Furthermore, statistical analysis reveals that these sparse semantic anchors are predominantly distributed across background regions. To fully unleash their potential as global contextual cues, our subsequent research will explicitly model the spatial interactions between these background anchors and the foreground road network, thereby further enhancing topological connectivity in complex scenarios.

References

[1] X. Wang, X. Jin, Z. Dai, Y. Wu, and A. Chehri, "Deep learning-based methods for road extraction from remote sensing images: A vision, survey, and future directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 13, no. 1, pp. 55–78, Mar. 2025.

[2] C. Wang, J. Lu, and Z. Chen, "BMDCNet: A satellite imagery road extraction algorithm based on multilevel road feature," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

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

[4] C. Li et al., "Efficient self-supervised vision transformers for representation learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[5] D. Yu and S. Ji, "Long-range correlation supervision for land-cover classification from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4409814.

[6] J. Geng, S. Song, and W. Jiang, "Dual-path feature aware network for remote sensing image semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3674–3686, May 2024.

[7] S. Lu et al., "Vision foundation models in remote sensing: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 13, no. 3, pp. 190–215, Sep. 2025.

[8] A. Kirillov et al., "Segment anything," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3992–4003.

[9] C. Hetang, H. Xue, C. Le, T. Yue, W. Wang, and Y. He, "Segment anything model for road network graph extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 2556–2566.

[10] P. Yin et al., "Towards satellite image road graph extraction: A global-scale dataset and a novel method," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 1527–1537.

[11] W. Feng, F. Guan, C. Sun, and W. Xu, "Road-SAM: Adapting the segment anything model to road extraction from large very-high-resolution optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[12] W. Wang et al., "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14408–14419.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[14] Z. Yang et al., "C^{2}Net: Road extraction via context perception and cross spatial-scale feature interaction," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5647011.

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

[16] X. Ma, X. Zhang, D. Zhou, and Z. Chen, "StripUnet: A method for dense road extraction from remote sensing images," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 11, pp. 7097–7109, Nov. 2024.
[17] Z. Yang, D. Zhou, Y. Yang, J. Zhang, and Z. Chen, "TransRoadNet: A novel road extraction method for remote sensing images via combining high-level semantic feature and context," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[18] P. Liu et al., "CGCNet: Road extraction from remote sensing image with compact global context-aware," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5638312.
[19] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[20] W. Liu, S. Gao, C. Zhang, and B. Yang, "RoadCT: A hybrid CNN-transformer network for road extraction from satellite imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
[21] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2004612.
[22] J. Jiao et al., "VMamba: Visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 103031–103063.
[23] S. Zhao, H. Chen, X. Zhang, P. Xiao, L. Bai, and W. Ouyang, "RS-mamba for large remote sensing image dense prediction," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5633314.
[24] Z. Wang, S. Yuan, R. Li, N. Xu, Z. You, and D.-S. Huang, "FDMamba: Frequency-driven dual-branch mamba network for road extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5643419.
[25] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, Jan. 2024.
[26] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.
[27] X. Li, C. Li, G. Vivone, and D. Hong, "SeaMo: A season-aware multimodal foundation model for remote sensing," *Inf. Fusion*, vol. 125, Jan. 2026, Art. no. 103334.
[28] X. Li, C. Li, P. Ghamisi, D. Hong, J. A. Benediktsson, and J. Chanussot, "FlexiMo: A flexible remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, 2026, Art. no. 5606516.
[29] C. Li et al., "UrbanSAM: Learning invariance-inspired adapters for segment anything models in urban construction," 2025, *arXiv:2502.15199*.
[30] J. Zhang, Y. Li, X. Yang, R. Jiang, and L. Zhang, "RSAM-seg: A SAM-based model with prior knowledge integration for remote sensing image semantic segmentation," *Remote Sens.*, vol. 17, no. 4, p. 590, Feb. 2025.
[31] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.
[32] Q. Xu, C. Long, L. Yu, and C. Zhang, "Road extraction with satellite images and partial road maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4501214.
[33] Z.-T. Hua, S.-B. Chen, W. Lu, J. Tang, and B. Luo, "Multiscale adaptive decoder and diversity selection network for road extraction in remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4411713.
[34] Z. Yang et al., "Semantic–spatial feature refinement network for road extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, 2026, Art. no. 5609710.
[35] J. Hu, J. Gao, Y. Yuan, J. Chanussot, and Q. Wang, "LGNet: Location-guided network for road extraction from satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5619112.
[36] X. Deng, M. Li, and Z. Chen, "AFDANet: An adaptive full-stage feature fusion and directional-aware network for road extraction in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–5, 2025.
[37] J. Mei, R.-J. Li, W. Gao, and M.-M. Cheng, "CoANet: Connectivity attention network for road extraction from satellite imagery," *IEEE Trans. Image Process.*, vol. 30, pp. 8540–8552, 2021.
[38] H. Chen, L. Yang, Q. Jia, and W. Xiong, "RoadFocusNet: Road extraction from remote sensing imagery using focused transformer and focused masked image modeling," *Int. J. Digit. Earth*, vol. 18, no. 2, Dec. 2025, Art. no. 2549435.
[39] Y. Qi, Y. He, X. Qi, Y. Zhang, and G. Yang, "Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6070–6079.
[40] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[41] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9300–9308.
[42] Z. Yu et al., "Deformable spatial attention networks: Enhancing lightweight convolutional models for vision tasks," *TechRxiv*, Mar. 2025.
[43] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.
[44] W. Liu, H. Lu, H. Fu, and Z. Cao, "Learning to upsample by learning to sample," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6004–6014.
[45] I. Demir et al., "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.
[46] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 210–223.
[47] Q. Li, X. Jia, J. Zhou, L. Shen, and J. Duan, "Rediscovering BCE loss for uniform classification," 2024, *arXiv:2403.07289*.
[48] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. Int. Workshop Deep Learn. Med. Image Anal., 7th Int. Workshop Multimodal Learn. Clin. Decis. Support*, Sep. 2017, pp. 240–248.

Wenhai Li received the bachelor's degree in engineering from Maanshan University, Maanshan, China, in 2024. He is currently pursuing the master's degree in computer science and technology with East China Jiaotong University, Nanchang, China.

He is currently a Researcher with the Institute of Data Science and Deep Learning, East China Jiaotong University. His research interests include deep learning and road extraction.

Xiaohui Huang received the B.Eng. and master's degrees from Jiangxi Normal University, Nanchang, China, in 2005 and 2008, respectively, and the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2014.

From 2017 to 2018, he was a Visiting Scholar with the College of Computing and Data Science, Nanyang Technological University, Singapore. He is currently a Professor of Computer Science at East China Jiaotong University, Nanchang. His research interests include deep learning, remote sensing image analysis, and intelligent transportation.

Xiaofei Yang (Member, IEEE) received the B.S. degree from Suihua University, Suihua, China, in 2011, and the M.S. and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 2014 and 2019, respectively.

From 2020 to 2023, he was a Post-Doctoral Researcher with the Department of Computer and Information Sciences, University of Macau, Macau, China. He is currently with the School of Electronic and Communication Engineering, Guangzhou University, Guangzhou, China. His research interests are in the areas of semisupervised learning, deep learning, remote sensing, transfer learning, and graph mining.

Yicong Zhou (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tufts University, Medford, MA, USA, in 2010.

He is a Professor at the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE) and was recognized as one of "Highly Cited Researchers" in 2020, 2021, 2023, and 2024. He serves as an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Jiangtao Peng (Senior Member, IEEE) received the B.S. and M.S. degrees from Hubei University, Wuhan, China, in 2005 and 2008, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor at the Faculty of Mathematics and Statistics, Hubei University. His research interests include deep learning, computer vision, and remote sensing image analysis.

Yifang Ban (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Nanjing University, Nanjing, China, in 1984 and 1987, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996.

Before joining the KTH Royal Institute of Technology (KTH), Stockholm, Sweden, in 2004, she was a tenured Associate Professor at York University, Toronto, ON, Canada. She is currently the Chair Professor and the Director of the Division of Geoinformatics, KTH, and the Associate Director of the Digital Futures, Stockholm. Her research interests include Earth observation big data analytics, deep learning, and their applications in road network extraction, urban land cover classification, and monitoring urbanization.

Nan Jiang (Member, IEEE) received the Ph.D. degree in computer science from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2008.

He is currently a Professor with the Department of Internet of Things and the Director of the Intelligent Sensor Networks Laboratory, East China Jiaotong University, Nanchang, China. His research interests include the Internet of Things, intelligent transportation systems, and remote sensing for smart cities.