remote sensing

MDPI

Article

SAM2MS: An Efficient Framework for HRSI Road Extraction Powered by SAM2

Pengnian Zhang †, Junxiang Li †, Chenggang Wang and Yifeng Niu *

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; zhangpengnianizsx@nudt.edu.cn (P.Z.); lijunxiang@nudt.edu.cn (J.L.); wcg_fc@nudt.edu.cn (C.W.)
* Correspondence: niuyifeng@nudt.edu.cn
† These authors contributed equally to this work.

Highlights

What are the main findings?

- SAM2MS, a novel road extraction framework, efficiently integrates the foundational vision model SAM2 with a multi-scale subtraction module (MSSM), achieving promising performance across three benchmark datasets.
- SAM2MS demonstrates the capability to infer unannotated road segments in datasets and achieves competitive performance on unseen datasets without requiring additional training, comparable to models specifically trained on those datasets.

What is the implication of the main finding?

- SAM2MS enhances segmentation accuracy and road connectivity through the synergistic use of SAM2 and the multi-scale subtraction module.
- SAM2MS exhibits strong robustness and generalizability, delivering reliable road predictions even in challenging scenarios.

Abstract

Road extraction from high-resolution remote sensing images (HRSIs) provides critical support for downstream tasks such as autonomous driving path planning and urban planning. Although deep learning-based pixel-level segmentation methods have achieved significant progress, they still face challenges in handling occlusions caused by vegetation and shadows, and often exhibit limited model robustness and generalization capability. To address these limitations, this paper proposes the SAM2MS model, which leverages the powerful generalization capabilities of the foundational vision model, segment anything model 2 (SAM2). Firstly, an adapter-based fine-tuning strategy is employed to effectively transfer the capabilities of SAM2 to the HRSI road extraction task. Secondly, we subsequently designed a subtraction block (Sub) to process adjacent feature maps, effectively eliminating redundancy during the decoding phase. Multiple Subs are cascaded to form the multi-scale subtraction module (MSSM), which progressively refines local feature representations, thereby enhancing segmentation accuracy. During training, a training-free lossnet module is introduced, establishing a multi-level supervision strategy that encompasses background suppression, contour refinement, and region-of-interest consistency. Extensive experiments on three large-scale and challenging HRSI road datasets, including DeepGlobe, SpaceNet, and Massachusetts, demonstrate that SAM2MS significantly outperforms baseline methods across nearly all evaluation metrics. Furthermore, cross-dataset transfer experiments (from DeepGlobe to SpaceNet and Massachusetts) conducted without any additional training validate the model's exceptional generalization capability and robustness.

check for updates

Academic Editor: John Trinder

Received: 28 July 2025
Revised: 5 September 2025
Accepted: 9 September 2025
Published: 14 September 2025

Citation: Zhang, P.; Li, J.; Wang, C.; Niu, Y. SAM2MS: An Efficient Framework for HRSI Road Extraction Powered by SAM2. *Remote Sens.* **2025**, *17*, 3181. https://doi.org/10.3390/rs17183181

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Keywords: high-resolution remote sensing images (HRSI); road extraction; segment anything model 2 (SAM2); multi-scale subtraction module

1. Introduction

High-resolution remote sensing images (HRSIs) play a crucial role in various fields, such as map drawing for autonomous driving, urban planning, and vital target reconnaissance. Among these applications, with the rapid development of autonomous vehicles, the task of road extraction using HRSIs presents significant challenges [1]. This is because the complexity of unknown environments demands high-precision and real-time processing of HRSI data, which directly bolsters the capacity of autonomous vehicles to adapt to unknown environments [2,3].

Specifically, road extraction from HRSIs using semantic segmentation approaches can be formulated as a dense pixel-wise prediction task. This task requires the accurate delineation of geometrically continuous road labels across large-scale complex scenes. However, the road extraction challenge from HRSIs is far more complex than generic semantic segmentation. In HRSIs, roads frequently suffer from occlusion by vegetation or buildings. Moreover, factors such as viewing angles, post-processing artifacts, and shadows caused by natural illumination severely affect the extraction accuracy. These issues collectively pose a significant hurdle, where maintaining a balance between precisely delineating road boundaries and ensuring connectivity remains a key challenge.

To address this challenge, deep learning-based road extraction methods have emerged as an essential and revolutionary solution. Some have modified conventional convolutional architectures to better capture the elongated and curvilinear structures of roads, thereby enhancing topological feature extraction from HRSIs. Others have incorporated transformers or their variants to better model global contextual information, ultimately improving the extraction performance [4–6]. Although the aforementioned methods have achieved certain effects, they still encounter challenges when dealing with occlusions caused by vegetation and shadows. Moreover, these methods often exhibit limited model robustness and generalization capabilities. Large-scale vision foundation models (such as the segment anything model (SAM) [7]) have already shown promise in the road extraction task due to their superior generalization capabilities. Very recently, its successor, SAM2 [8] has been introduced, demonstrating even more powerful segmentation capabilities. To better address the challenges in road extraction tasks, an increasing number of researchers are exploring the adaptation of large foundation models to downstream applications. SAM_MLoRA [9] leverages LoRA theory to fine-tune SAM, establishing a robust and parameter-efficient framework for extracting urban artificial structures such as buildings and roads. Concurrently, SAM-Road [10] and SAM-Road++ [11] have emerged, which directly utilize pre-trained weights of SAM by connecting geometric–topological networks for end-to-end road graph extraction. Notably, SAM-Road++ requires the construction of an exceptionally large dataset—the Global-Scale dataset—for training. These advancements demonstrate that leveraging foundational vision models like SAM holds significant potential for enhancing the precision and robustness of segmentation methodologies. Such approaches could offer efficient solutions for specialized yet technically demanding tasks. Nevertheless, critical trade-offs between model performance, training data requirements, and generalization capabilities must be carefully addressed to ensure practical applicability across diverse urban environments.

In this paper, we present SAM2MS, a novel framework that effectively integrates SAM2 with HSRI road extraction tasks. This integration is specifically designed to enhance both model accuracy and robustness, enabling superior adaptation to diverse road extraction scenarios. The principal contributions of this study can be summarized as follows:

- Building on prior research insights, we developed a novel road extraction architecture that effectively adapts the SAM2 model to this task while achieving substantially improved performance. The key innovation of the framework lies in its seamless integration of our proposed multi-scale subtraction module (MSSM) with the SAM2 encoder, which not only outperforms baseline methods but also demonstrates remarkable capabilities in handling missing label regions and maintaining strong generalization across complex scenarios through enhanced robustness.
- Multi-level supervision strategy: We introduce a lossnet supervision architecture that establishes multi-level constraints for background suppression, contour refinement, and region-of-interest consistency. This framework forms a multi-level supervision strategy, significantly improving the transparency of both model behavior and training dynamics.
- Cross-scenario validation: The proposed SAM2MS achieves remarkable performance on three challenging large-scale datasets—the off-road-oriented DeepGlobe dataset, the urban-scene SpaceNet dataset, and the Massachusetts dataset—demonstrating its adaptability to diverse environments.

The remainder of this paper is organized as follows: Section 2 provides a concise overview of road extraction methodologies and the evolution of SAM with its applications. Section 3 elaborates on the proposed approach. Section 4 introduces the experimental datasets and performance metrics, followed by quantitative and qualitative results along with auxiliary experiments to validate the efficacy of our method. Finally, Section 5 concludes this work and discusses potential avenues for future research. Code and models are available at https://github.com/zhongwlf/SAM2MS, accessed on 8 September 2025.

2. Related Work

2.1. Traditional Machine Learning for Road Extraction

Traditional machine learning methods for road extraction tasks can be categorized into automatic and semi-automatic algorithms based on the level of human–computer interaction [12]. Semi-automatic approaches typically require predefined or interactively placed seed points during the extraction process, along with user validation of results. Representative techniques include active contour models and template matching methods [13,14]. In contrast, fully automatic algorithms necessitate parameter configurations tailored to specific image characteristics. Standard methods encompass artificial neural networks (ANNs), support vector machines (SVM), Bayesian classifiers, watershed algorithms, mean shift (MS), K-means clustering, gaussian mixture models (GMMs), superpixel segmentation, and conditional random fields (CRFs), which are applied to road extraction or post-processing stages [15–17]. While these classical methods achieve acceptable performance in structured scenarios, their scenario-dependent hyperparameters reduce automation efficiency. Additionally, limited model transferability compromises robustness across diverse environments.

2.2. CNN and Transformers for Road Extraction

2.2.1. Only CNN

Road extraction tasks can be transformed into pixel-level image segmentation problems. The early fully convolutional network (FCN) achieved semantic segmentation through end-to-end pixel classification, utilizing transposed convolution for feature map up-

sampling to restore input resolution, thereby enabling the processing of input images with arbitrary dimensions [18]. Subsequently developed encoder–decoder architectures (e.g., SegNet, UNet) optimized feature reconstruction through symmetrical structures [19–21]. MCMCNet [22] employs a semi-supervised training paradigm, introducing a guided contrastive learning module (GCLM) to enhance inter-model consistency and incorporating a road skeleton prediction head (RSPH) to improve the extraction of topologically coherent road networks. Nevertheless, this approach imposes stringent requirements on both the data distribution/quality of training datasets and the selection of pretrained models. Follow-up studies focused on optimizing internal convolutional structures to enhance model performance: D-LinkNet [23] introduced dilated convolution to construct multi-scale feature perception capabilities [24–26]. At the same time, MSMDFF-Net [27] proposed the strip convolution module (SCM) that performs long-range convolution along orthogonal directions, effectively capturing linear contextual information consistent with road morphology.

2.2.2. CNN Combined with Transformer

With the emergence of vision transformers (ViTs), transformer-based segment models have surpassed traditional convolutional networks in performance [28]. Existing hybrid architectures maintain the encoder–decoder paradigm: RADANet [29] integrates deformable transformers throughout the encoding–decoding process, combining the sparse sampling advantage of deformable convolution with spatial self-attention mechanisms to achieve multi-scale road semantic extraction. RoadFormer [30] hierarchically integrates multi-resolution deformable vision transformer features through a pyramid architecture. Seg-Road [31] employs a transformer-based encoder to capture global contextual dependencies for suppressing road fragmentation while incorporating a CNN decoder to enhance edge detail reconstruction accuracy. SwinUNet [6] integrates the hierarchical window attention mechanism with the U-shaped architecture, strengthening global feature modeling capabilities while reducing computational complexity.

The evolution from models exclusively employing CNNs to the integration of transformers has continuously advanced the capability of models to capture both global contextual features and local structural details. Nevertheless, training on a single dataset may lead to suboptimal generalization capabilities due to the inherent complexity and diversity of scenarios in road extraction datasets. To address this limitation, adopting large-scale heterogeneous datasets or cross-domain pre-trained models is critical to enhance the robustness and adaptability of the model across varied environments.

2.3. Segment Anything Model

Recently, SAM has emerged as a foundational model for general-purpose natural image segmentation, demonstrating exceptional performance through pre-training on the large-scale SA-1B dataset. This model enables pixel-level mask generation for target regions via interactive clicks, bounding boxes, or natural language prompts, with validated zero-shot segmentation capabilities across various vision tasks [7]. This technological breakthrough has further advanced the field of medical image segmentation. Its enhanced version, SAM2, incorporates a hierarchical encoder (Hiera backbone) for multi-scale feature extraction and extends support to video content segmentation [8]. However, these general models still face limitations in class-agnostic segmentation results when manual prompts are absent, hindering their application in specific downstream tasks. Certain researchers directly fuse predictions from the original semantic segmentation model with more accurate boundary masks generated by SAM to enhance segmentation precision. While this approach eliminates the need for additional training or architectural modifications—thereby

streamlining the deployment pipeline—it requires manually engineered fusion mechanisms that inherently compromise robustness and generalization capability [32]. Researchers have proposed two improvement strategies to address these challenges: parameter fine-tuning and model architectural optimization.

2.3.1. Parameter Fine-Tuning

The SAM-Adapter [33] incorporates domain-specific prior knowledge with the generalizable representations of SAM through mechanisms that inject task-specific knowledge. This approach allows for flexible adaptation to various tasks. In contrast, SAM-RSIS [34] employs a multi-scale adapter to adapt the pretrained vision transformer, effectively enhancing instance segmentation performance while demonstrating superior generalization capability. However, its adapter architecture entails significant complexity, leading to suboptimal performance when trained on single road extraction datasets. Meanwhile, SAM2UNet [35] merges the hierarchical encoder of SAM2 with a UNet decoder framework, resulting in notable performance enhancements in both natural and medical image segmentation tasks.

2.3.2. Model Architecture Optimization

Some methods enhance feature representation by modifying the structures of models. However, retraining requires substantial datasets and extensive computational resources. For instance, SAMUS [36] introduces a parallel CNN branch into the image encoder and uses cross-attention mechanisms to facilitate interactions between multi-modal features. SAMUNet [37] incorporates a learnable CNN auxiliary branch while preserving the generalization capabilities of the ViT batch. Additionally, it includes a multi-scale fusion module in the mask decoder to improve segmentation accuracy. Gao et al. [38] integrate an additional CNN as an adapter with the foundational FastSAM model [39], freezing parameters of FastSAM while solely training the introduced adapter and decoder components to generate refined confidence maps. Nevertheless, this approach entails architectural modifications to the original framework. Although such a design improves training efficiency by leveraging limited datasets to train the supplemental CNN module while keeping FastSAM frozen, the parallel encoder configuration suffers from inadequate feature alignment when trained on significantly divergent datasets.

In summary, adapter-based fine-tuning represents the most streamlined and parameter-efficient paradigm for road extraction tasks. However, significant challenges persist in adapting SAM-series vision foundation models to road extraction: (1) distributional discrepancy between remote sensing road datasets and the SA-1B pretraining corpus compromises feature alignment and (2) semantic granularity mismatch wherein SAM's mask decoder generates class-agnostic segmentation masks—imposing unnecessary computational overhead while failing to leverage task-specific prior knowledge inherent to binary road extraction frameworks. Crucially, SAM2 achieves breakthroughs beyond its predecessor through enhanced inference accuracy, accelerated processing speed, and superior zero-shot transfer capability. This advancement enables effective operation on unseen HRSI data without additional fine-tuning, thereby significantly enhancing adaptation efficacy for downstream remote sensing applications.

3. Method

3.1. Overview

We formulate road extraction from remote sensing imagery as a pixel-wise dense prediction task on RGB images, requiring per-pixel classification label prediction. As illustrated in Figure 1, the proposed SAM2MS model accepts $H \times W \times 3$ RGB inputs and generates an $H \times W \times 1$ binary road segmentation map. The architecture adheres

to the predominant encoder–decoder paradigm for segmentation efficiency: (1) Encoder: Incorporates a fine-tuned SAM2 image encoder retrained on remote sensing datasets for multi-scale feature extraction, followed by dimensional reduction blocks (DRBs) to standardize channel dimensions for downstream processing. (2) A multi-scale subtraction module (MSSM) is introduced between the encoder and decoder. The MSSM incorporates multiple subtraction blocks (Subs) and utilizes inter-layer skip connections to compute differential representations of features across adjacent levels, thereby effectively eliminating redundant feature replicates. This module helps reduce repetitive information in the feature maps output by the encoder, enhances discriminability across scales, and improves the delineation of fine-grained details. (3) Decoder: Recursively reconstructs road segmentation through successive upsampling operations. During training, a lossnet module provides multi-level feature supervision to optimize segmentation performance from local details to global structures.

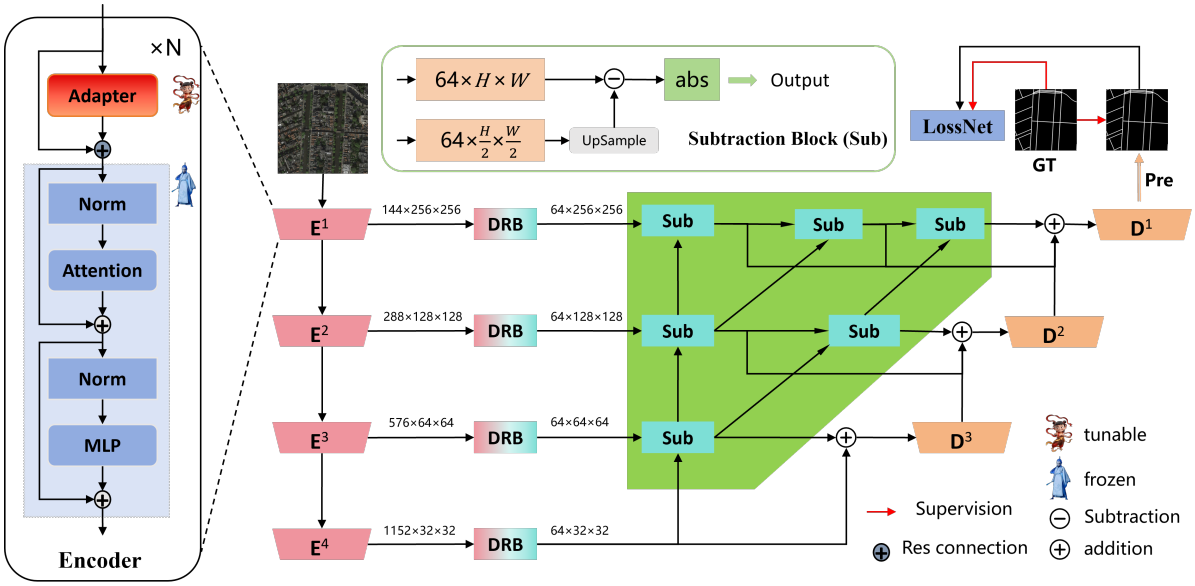


Figure 1. Architectural composition of SAM2MS: Encoder–decoder framework with adapters, dimensionality reduction blocks (DRBs), and a multi-scale subtraction module (MSSM) constructed through the cascading of multiple subtraction blocks (Subs). The MSSM module is visually highlighted in green.

3.2. Encoder–Decoder

In contrast to the original SAM2 model, SAM2MS is specifically tailored for the road extraction task and can utilize the pre-trained image encoder more efficiently. This design discards the original prompt encoder of SAM2 (i.e., the point, box, and text prompts) as they cannot be effectively positioned for prompting elongated and irregular roads in HRSIs. Furthermore, the mask decoder of SAM2 is replaced by a CNN-based mask head. This is because the mask decoder of SAM2 generates multiple category-agnostic masks, whereas road extraction only requires binary labeling (road/non-road). Our convolutional decoder not only processes the fine-tuned encoder features efficiently but also excels at capturing local details to enhance precision.

Specifically, we efficiently adapt the encoder of SAM2 via adapter modules to extract multi-level features from the H × W × 3 input, generating four hierarchical feature representations. The decoder then aggregates these lower-level features with different representations produced by a multi-scale subtraction module operating at corresponding scales. This fusion generates enhanced complementary feature representations optimized for the task.

3.3. Adapter and DRB

Fine-tuning visual foundation models for downstream tasks effectively transfers their high robustness and versatility to specific applications. While some approaches augment SAM2 by integrating a trainable encoder with its original encoder—forming a dual-branch architecture with multi-level feature fusion—this strategy substantially increases parameter count and introduces computational complexity. The intricate fusion process between the supplementary encoder and computationally intensive ViT backbone of SAM2 hinders efficient integration. To overcome this limitation, we adopt a streamlined adapter-based fine-tuning approach.

During fine-tuning, we freeze parameters of the pretrained image encoder of SAM2 and insert adapters (Figure 2) before each multi-scale block for parameter-efficient adaptation. These adapters—comprising fully connected layers with ReLU activation—efficiently embed domain knowledge from remote sensing imagery (whose distribution substantially diverges from SA-1B) while maintaining minimal computational overhead.

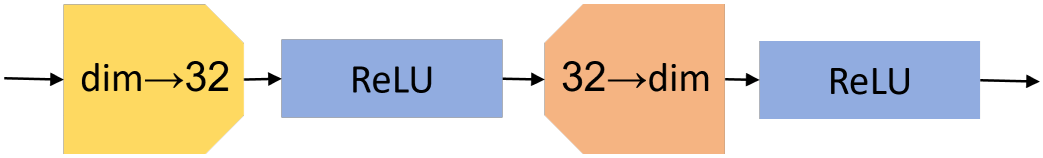


Figure 2. The schematic diagram of the adapter shows that the adapter is composed of two fully connected neural networks for upsampling and downsampling, along with ReLU activation layers. The input to the adapter is external remote sensing information, which is efficiently embedded into the SAM2 image encoder through this simplified network architecture, enabling effective processing of remote sensing data.

After extracting the encoder features, these representations are processed by four DRBs before being passed into the multi-scale subtraction module. The DRB reduces the channel dimension to 64, alleviating the computational burden of subsequent modules while enhancing the discriminative ability of these lightweight features. As shown in Figure 3, the design of the DRB combines the strengths of the multi-scale convolutional attention (MSCA) module [40], receptive field block (RFB) [41], and CoANet [5], among other techniques. We designed deep convolutions for aggregating local information to capture multi-scale contextual features through multi-branch deep strip convolutions and used 1 × 1 convolutions to model relationships between different channels. The output of the 1 × 1 convolutions is directly used as attention weights to reweight the input of the MSCA. In each branch, we employ two depth-wise strip convolutions to approximate the large kernels in standard depth-wise convolutions. The kernel sizes are set to 3, 5 and 7, respectively. This design ensures that, while expanding the receptive field, the strip convolutions remain lightweight. Additionally, since roads in scene segmentation typically appear as elongated objects, the strip convolution serves as a complement to grid convolutions, aiding in the extraction of strip-like features and enhancing the performance of the model.

3.4. Multi-Scale Subtraction Module

The efficacy of road extraction is constrained by the inherent characteristics of road regions within these images. Specifically, roads exhibit a sparse spatial distribution, occupy a small proportion of the total image area, and possess irregular, complex boundaries. These characteristics frequently lead to artifacts and discontinuities in the extracted results. To address these limitations, we employ a multi-scale subtraction module. This module enhances the characterization of fine road details, mitigates artifact generation, and improves the model's capability to address the challenges of extracting elongated and occluded road

segments. To elucidate the operational mechanism of the multi-scale subtraction module, we denote F_a and F_b as adjacent-level feature maps. A fundamental subtraction block is defined as

$$Sub = Conv(|F_a \ominus F_b|), \tag{1}$$

where $\ominus$ represents element-wise subtraction, $|\,|$ computes absolute values, and $Conv\,(\cdot)$ denotes a convolutional layer. This block provides richer feature representations for the decoder by capturing complementary information and highlighting differential characteristics between F_a and F_b. To further acquire higher-order complementary information across multiple feature levels, we horizontally and vertically concatenate multiple Subs to generate a series of differential features with varying orders and receptive fields.

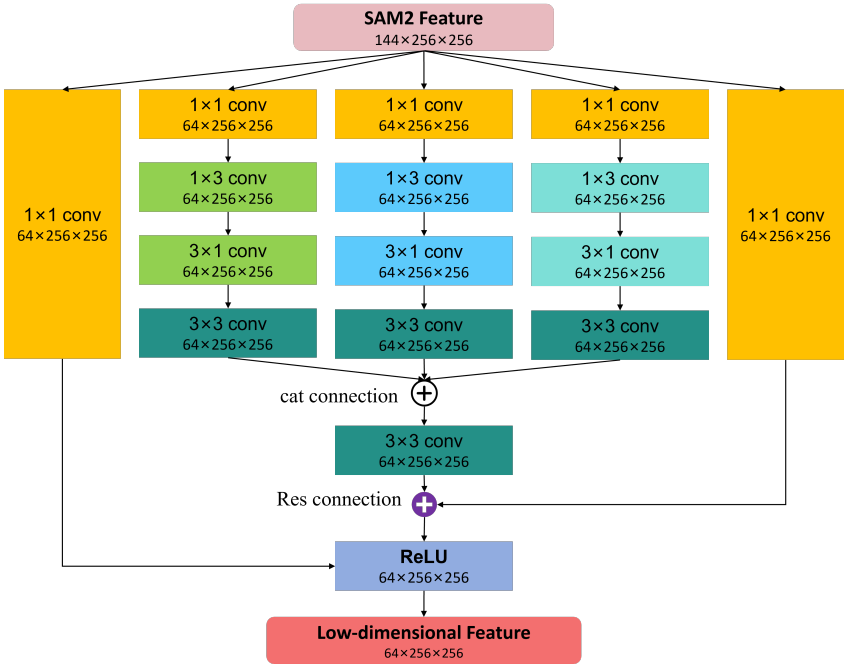


Figure 3. The schematic diagram of the dimensional reduction block indicates that the input consists of four feature layers from the SAM2 encoder. After passing through the DRB module, the feature dimension is reduced to a low-dimensional (64-dimensional) representation. The structure utilizes convolution kernels of sizes 3, 5 and 7, with the outputs from different branches being fused, expanding the receptive field while ensuring the model remains lightweight. Taking a feature map input of size 144 × 256 × 256 as an example, the channel number and feature map size annotated at each layer correspond to the dimensions of the output feature map of that layer.

3.5. Lossnet

To enable more effective deep supervision during model training, we augment standard loss functions with a lossnet. Distinct from the deep supervision paradigm in UNet++ [21], lossnet leverages a pre-trained model to extract multi-level feature representations from ground truth labels. These hierarchical feature maps provide rich supervisory signals by contrasting with corresponding model outputs. Crucially, all components within lossnet are non-trainable—their parameters remain isolated from those of the SAM2MS model. Only the final loss value computed by lossnet propagates gradients back to update SAM2MS parameters during backpropagation. The total loss function of our model is formulated as

$$\mathcal{L}_{total} = \mathcal{L}^{w}_{IoU} + \mathcal{L}^{w}_{BCE} + \mathcal{L}_{Fun}, \tag{2}$$

where $\mathcal{L}^{w}_{IoU}$ and $\mathcal{L}^{w}_{BCE}$ denote weighted IoU loss and binary cross entropy (BCE) loss, respectively, both widely adopted in segmentation tasks. We adopt parameter settings

consistent with recently proposed road extraction models, the effectiveness of which has been empirically validated in prior studies.

$$\mathcal{L}_{IoU}^{w} = 1 - \frac{\sum_{r=1}^{H}\sum_{c=1}^{W} P(r,c)G(r,c)}{\sum_{r=1}^{H}\sum_{c=1}^{W}[P(r,c) + G(r,c) - P(r,c)G(r,c)]} \tag{3}$$

where H and W represent the height and width of the predicted image. r is an integer value in the range [0, H], i.e., $r \in \mathbb{Z} \wedge 0 \leq r \leq H$. c is an integer value in the range [0, W], i.e., $r \in \mathbb{Z} \wedge 0 \leq r \leq W$. $G(r,c) \in \{0,1\}$ is the ground truth label of the pixel(r, c) and $P(r,c)$ represents the probability of predicting the pixel as road.

$$\mathcal{L}_{BCE}^{w} = -\sum_{(r,c)} [G(r,c)\log(P(r,c)) + (1 - G(r,c))\log(1 - P(r,c))] \tag{4}$$

where $G(r,c) \in \{0,1\}$ is the ground truth label of the pixel (r, c) and $P(r,c)$ represents the probability of predicting the pixel as road.

As shown in Figure 4, we utilize an ImageNet [42] pre-trained ResNet50 [43] network to extract multi-scale features from both predictions and ground truths, constructing the loss term $\mathcal{L}_{Fun}$ through feature discrepancy calculation:

$$\mathcal{L}_{Fun} = l_f^1 + l_f^2 + l_f^3 + l_f^4 + l_f^5. \tag{5}$$

Each layer loss is computed as pixel-wise L2 norm

$$l_f^i = ||f_p^i - f_g^i||_2, \quad i = 1,2,3,4,5. \tag{6}$$

where f_p^i and f_g^i represent the i-th level feature maps from predictions and ground truths, respectively.

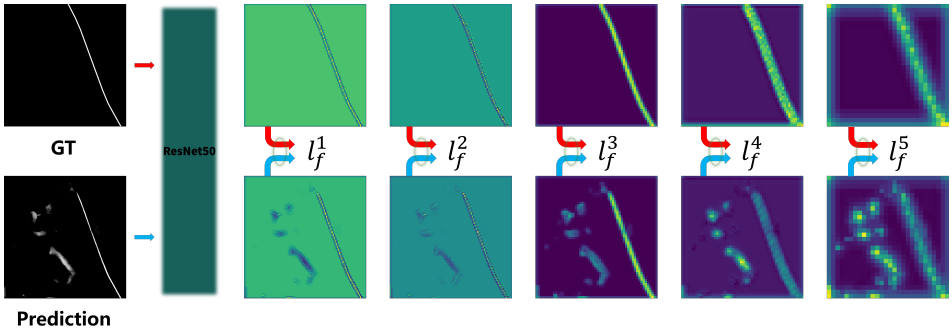


Figure 4. Parameter-free lossnet architecture: deep supervision via fixed ResNet50 backbone.

As illustrated in Figure 4, the attention to lower-level features is primarily focused on the background, and as the hierarchy increases, it gradually transitions from edge contours to road features. A deep-level supervision of background–contour–road was established to ensure contour optimization and regional consistency, thereby enhancing the interpretability of the training process.

4. Experiment

4.1. Dataset and Evaluation Metrics

DeepGlobe Dataset [44]: The dataset consists of 8570 aerial images primarily focusing on off-road environments, each with a spatial resolution of 0.5 m per pixel (1024 × 1024 pixels). Within this collection, 6226 images are annotated with pixel-level labels for road and background classes in challenging off-road scenarios. These labeled

samples are partitioned into 4326 training samples and 1900 testing samples to ensure robust model evaluation.

SpaceNet Dataset [45]: The original dataset comprises 2549 images (1300 × 1300 pixels with a per-pixel resolution of 0.5 m) collected from four cities: Khartoum, Paris, Shanghai and Las Vegas. Without altering the spatial resolution, these images were processed into 10,196 patches of 1024 × 1024 pixels, each annotated with pixel-level road/background labels. The processed dataset was then partitioned into 7096 training samples and 3100 testing samples, ensuring a strict correspondence between images and labels.

Massachusetts Dataset [46]: The original dataset comprises 1171 aerial images (1500 × 1500 pixels) primarily covering the Massachusetts region. Road centerlines acquired from the OpenStreetMap project were rasterized to generate target maps, producing 7-pixel-wide linear features without smoothing. All images were resampled to a uniform resolution of 1 pixel per square meter. Following identical preprocessing protocols, the imagery was tiled into 4615 patches of 1024 × 1024 pixels, with 3115 patches designated for model training and the remaining 1500 reserved for testing.

Dataset Annotation and Partitioning: As illustrated in Figure 5, fundamental differences exist in the annotation paradigms employed by DeepGlobe, SpaceNet, and the Massachusetts datasets. DeepGlobe utilizes manual tracing along actual road boundaries—an approach that enhances the fidelity of deep learning-based road extraction methods to real-world data. In contrast, SpaceNet and Massachusetts adopt a road centerline annotation paradigm, generating training masks by applying fixed-width buffers to centerlines. Consequently, the resulting road masks may not fully encompass the complete road surface but prioritize the representation of topological connectivity and macro-morphological characteristics.



Figure 5. Annotation exemplars for the Massachusetts, SpaceNet, and DeepGlobe datasets are shown sequentially from left to right, with the top row presenting original imagery and the bottom row depicting corresponding annotations. Red markings highlight critical annotation details.

To ensure fairness and unbiased partitioning between training and testing sets, we implemented a MapReduce-inspired strategy. The cropped datasets were first shuffled and divided into subgroups, each containing 100 images. Within each subgroup, images were partitioned in a 7:3 ratio, allocating 70% for training and 30% for testing. The training

and testing subsets from all subgroups were then aggregated to form the final training and testing sets.

As shown in Figure 6, we performed dimensionality reduction analysis using t-SNE and principal component analysis (PCA) on the finally divided datasets. From both the training and test sets, features such as road density, road edge characteristics, image color attributes, and road texture features (quantified using the gray-level co-occurrence matrix (GLCM), including metrics such as contrast, dissimilarity, homogeneity, and energy) were extracted for joint analysis. The resulting two-dimensional visualizations from both methods consistently revealed that the training and test data points from the three datasets are closely interwoven in the reduced dimensional space. This overlapping phenomenon fully reflects their intrinsic distribution similarity, thereby validating the feasibility of the division method we adopted and ensuring the reliability of subsequent experimental results.

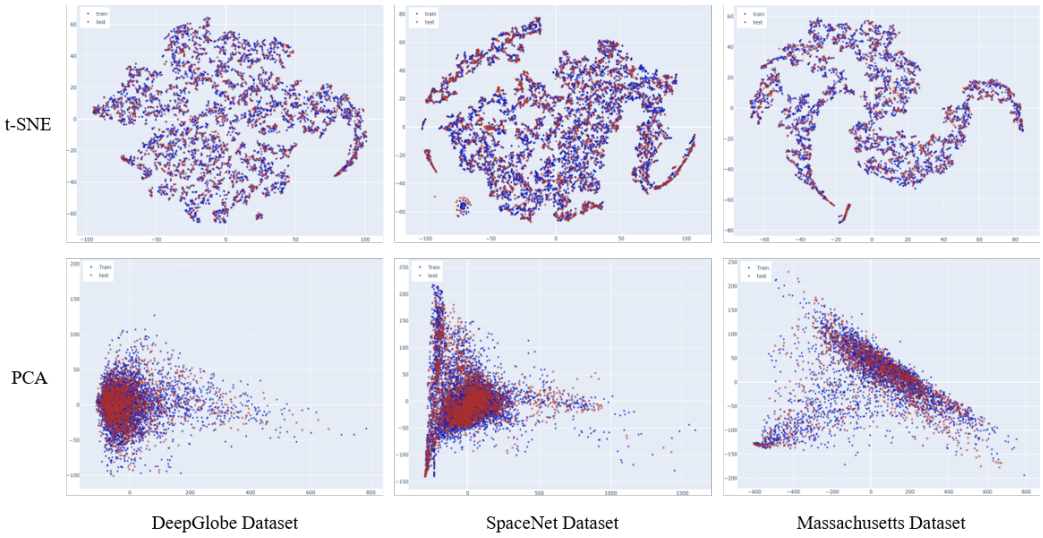


Figure 6. Dimensionality reduction analysis was performed on the training and test sets using both t-SNE and PCA. By extracting multiple features from each dataset, we conducted a joint analysis of these datasets. As shown in the figure, the resulting two-dimensional visualization illustrates the distribution of the data in the reduced space: blue points represent the training set, while red points denote the test set. For all three datasets, the training and test data exhibit a closely interwoven distribution pattern in this space.

Evaluation Metrics: To comprehensively evaluate the performance of road segmentation methods, we follow the convention of most segmentation-based approaches and employ a set of metrics covering pixel-level accuracy, regional consistency, and error analysis.

We define the pixel classification metrics with the following notations: TP (True Positives): number of correctly predicted road pixels; TN (True Negatives): number of correctly predicted background pixels; FP (False Positives): number of background pixels incorrectly predicted as roads; FN (False Negatives): number of road pixels incorrectly predicted as background.

- Pixel-Level Accuracy Metrics

We utilize precision and recall to measure the ability of the model to control false positives and false negatives. The F1 score, which is the harmonic mean of precision and recall, is used to assess the balance of classification performance comprehensively. The larger the value of these metrics, the better the model performance. The formulas are defined as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

$$Recall = \frac{TP}{TP + FN}, \tag{8}$$

$$F1 = \frac{2TP^2}{2TP^2 + TP(FP + FN)}. \tag{9}$$

- Regional Consistency Metrics

We employ the mean intersection over union (mIoU) to measure the geometric overlap of segmentation boundaries. Furthermore, the mean dice coefficient (mDice) is employed to evaluate pixel-level matching accuracy, providing greater sensitivity to the pixel-wise matching of small road objects. The larger the value of these metrics, the better the model performance. The formulas are defined as follows:

$$mIoU = \frac{1}{N+1} \sum_{n=0}^{N} \frac{TP}{TP + FP + FN}, \tag{10}$$

Here, N denotes the total number of pixels.

$$mDice = \frac{2TP}{2TP + FP + FN}. \tag{11}$$

- Error Analysis Metrics

We use the mean absolute error (MAE) to assess the overall error level of the model quickly. MAE provides an intuitive measure of global error by calculating the mean of pixel-level absolute differences between the predicted (P_i) and ground-truth (G_i) segmentation maps. This metric is particularly suitable for rapid performance evaluation on large-scale datasets (lower values indicate better performance). The formula is defined as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |P_i - G_i|. \tag{12}$$

- Road Label Overlap Degree

We introduce a new road overlap metric method, used for quantitative analysis in Section 4.5. The formulation of this metric stems from the fundamental discrepancy in annotation protocols between SpaceNet and DeepGlobe datasets: while DeepGlobe employs precise road boundary delineation, SpaceNet adopts homogeneous linear annotations resembling road centerlines. This annotation discrepancy results in severely degraded performance of DeepGlobe-trained models on the SpaceNet dataset. The proposed metric calculates the ratio of overlapping pixels (i.e., True Positives, TP) between predicted and annotated roads relative to the total annotated road pixels (denoted as P_{road}) per image, followed by arithmetic mean computation across the test set. The formal definition is expressed as

$$\text{RLOD} = \frac{1}{N} \sum_{n=1}^{N} \frac{TP_n}{P_{road}}. \tag{13}$$

4.2. Implementation Details

Our implementation is based on PyTorch 2.1.0 with the AdamW optimizer. The initial learning rate is set to 1×10^{-4} and decays to 1×10^{-7} via a cosine annealing scheduler (CosineAnnealingLR), whose cycle length matches the training epochs. All experiments utilize distributed data parallelism across 8 NVIDIA RTX 4090 GPUs (24 GB VRAM each),

maintaining a total batch size of 16 (2 per GPU). The computation of FLOPs and parameter counts for the model was conducted on an NVIDIA RTX 4060 GPU. Models are trained for the whole 50 epochs without early stopping. The data augmentation pipeline includes random rotation $\pm 30^{\circ}$, horizontal flipping ($p = 0.5$), and bilinear interpolation-based resizing to standardize input resolution to 1024×1024 pixels.

4.3. Comparison with Other Methods

We conducted systematic comparative experiments for the proposed SAM2MS against four representative road extraction categories: UNet architectures [20,21,23–25], topology-aware models [26,27], transformer-based methods [6,31], and fine-tuned SAM variants [35].

Figure 7 comparatively demonstrates the inference results of baseline models versus our proposed SAM2MS across three datasets. Columns (1) to (3) present results on the DeepGlobe dataset, featuring urban viaducts, rural roads, and extreme off-road environments. Although UNet/UNet++-based models maintain road connectivity and accuracy through specialized modules, SAM2MS exhibits superior robustness against tree occlusions and elevated highway interference. Our method not only accurately predicts annotated road segments but also infers unlabeled road portions, as shown in column (3). Columns (4) to (6) display results on the SpaceNet dataset, highlighting challenges such as tree occlusion and remote sensing image distortions. Compared to columns (4) and (5), both SAM2MS and SAM2UNet produce predictions consistent with ground truth while preserving connectivity, whereas traditional methods like UNet suffer significant fragmentation. Column (6) illustrates severe radiometric distortion caused by data processing artifacts: solid red boxes mark occluded and shadowed regions, while dashed boxes indicate missing areas. Notably, SAM2MS consistently demonstrates stronger noise resistance and cross-scene generalization. Columns (7) and (8) show results on the Massachusetts dataset, where dense road networks and extensive building interference blur boundaries between road and non-road areas. Models such as MSNet, M2SNet, and Seg-Road exhibit poor grayscale representation in non-road regions but successfully infer reverse roads (e.g., dense road networks in the lower-right corner of column (8)). This behavior correlates with their high accuracy but subregional performance metrics (e.g., mIoU and mDice). In column (8), SAM2MS also underperforms, failing to fully reconstruct the road network. This indicates that predicting dense road structures remains a challenge for our model and requires further improvement.

As shown in Tables 1–3, MAE was normalized (preserving four decimal places and scaled by 100).

Table 1. Cross-model quantitative evaluation on DeepGlobe datasets benchmark.

Method	DeepGlobe Dataset [44]							
	Param. (M)	**FLOPs (G)**	**Prec.**	**Recall**	**F1**	**mIoU**	**mDice**	**MAE ↓**
UNet (*2015*)	31.02	875.81	74.83	76.97	75.89	59.86	73.44	2.10
UNet++ (*2018*)	47.19	3202.89	80.22	52.06	63.14	44.58	58.25	2.64
D-LinkNet (*2018*)	217.64	481.25	72.95	72.48	72.71	62.88	75.75	1.90
MSNet (*2021*)	27.69	143.91	80.77	76.81	78.74	63.04	76.65	1.86
M2SNet (*2023*)	27.69	144.41	81.95	74.99	78.32	63.74	76.26	1.84
Seg-Road (*2023*)	28.68	314.41	71.78	80.78	76.02	60.42	73.98	2.29
SwinUNet (*2023*)	27.14	123.59	**83.18**	<u>68.72</u>	75.26	60.30	73.84	1.84
SGCNNet (*2022*)	42.73	1234.41	69.49	78.52	73.73	57.51	71.47	2.23
MSMDFFNet (*2024*)	603.21	39.26	74.49	80.26	77.27	62.22	75.30	1.96
SAM2UNet (*2024*)	863.72	216.41	<u>74.26</u>	**83.16**	78.46	63.64	76.47	2.08
SAM2MS (*Ours*)	867.28	217.11	79.50	81.32	**80.31**	**64.24**	**77.93**	**1.80**

Downward arrows (↓) indicate that lower values correspond to better performance.

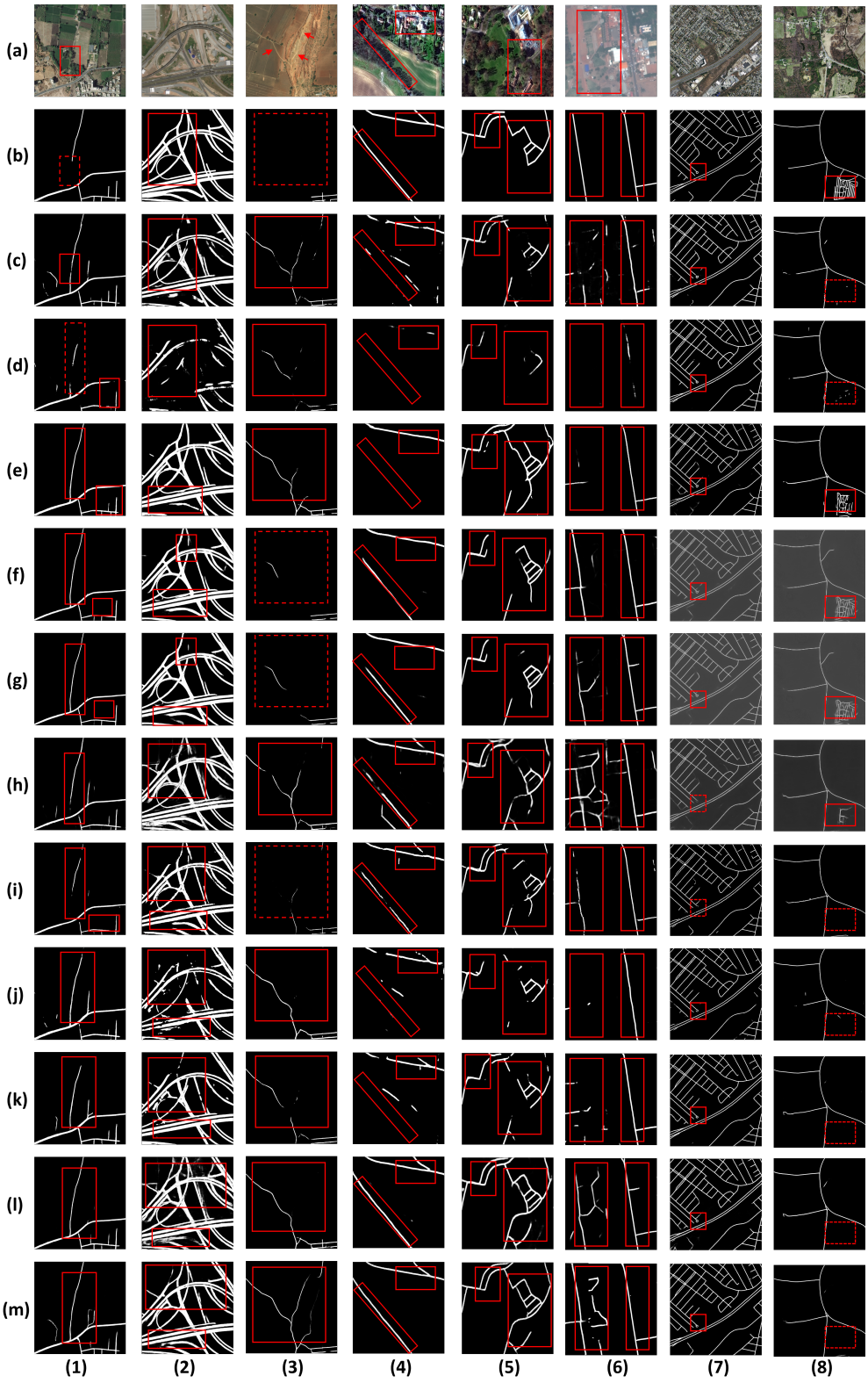


Figure 7. Comparative results: Representative outcomes from the three datasets demonstrate performance variations among the models. Row (a) displays the actual remote sensing images, while row (b) presents the corresponding ground truth labels. Rows (c–l) sequentially show the test results of baseline models: UNet, UNet++, D-LinkNet, MSNet, M2SNet, Seg-Road, SwinUNet, SGCNNet, MSMDFFNet and SAM2UNet. Row (m) showcases our proposed SAM2MS method. Samples (1–3) originate from the DeepGlobe dataset, samples (4–6) from SpaceNet, and samples (7,8) from Massachusetts. Solid red borders highlight areas with significant occlusion and shadows, whereas dashed borders emphasize missing regions in either ground truth annotations or inference results.

Table 2. Cross-model Quantitative Evaluation on SpaceNet Datasets Benchmark.

Method	SpaceNet Dataset [45]							
	Param. (M)	FLOPs (G)	Prec.	Recall	F1	mIoU	mDice	MAE ↓
UNet (*2015*)	31.02	875.81	60.61	56.61	58.54	42.01	55.56	5.28
UNet++ (*2018*)	47.19	3202.89	58.91	26.40	36.46	21.62	30.45	5.10
D-LinkNet (*2018*)	217.64	481.25	62.58	63.55	63.06	47.94	61.82	3.77
MSNet (*2021*)	27.69	143.91	65.98	58.76	62.16	46.24	60.19	3.83
M2SNet (*2023*)	27.69	144.41	66.34	57.99	61.89	46.03	59.96	3.73
Seg-Road (*2023*)	28.68	314.41	60.07	66.14	62.96	46.27	60.39	4.36
SwinUNet (*2023*)	27.14	123.59	**69.84**	53.18	60.38	44.74	58.94	**3.36**
SGCNNet (*2022*)	42.73	1234.41	57.63	50.63	53.90	38.64	51.26	4.52
MSMDFFNet (*2024*)	603.21	39.26	60.41	60.84	60.62	45.50	59.09	3.89
SAM2UNet (*2024*)	863.72	216.41	67.11	57.43	61.90	45.60	59.24	4.06
SAM2MS (*Ours*)	867.28	217.11	62.34	**67.16**	**64.66**	**48.52**	**62.57**	3.73

Downward arrows (↓) indicate that lower values correspond to better performance.

On the DeepGlobe dataset, SAM2MS outperforms all baselines in F1, mIoU, mDice and MAE. It achieves improvements of 1.57 (F1) and 1.28 (mDice) over the second-best method (MSNet, F1 = 78.74), demonstrating its balanced capabilities. SAM2MS achieves the highest recall (81.32) on DeepGlobe but slightly lower precision (79.50 vs. SwinUNet's 83.18), indicating a design focus on reducing false negatives at a potential cost of increased false positives.

Notably, SAM2UNet and SwinUNet achieve peak performance in precision and recall, respectively, yet exhibit suboptimal results in the complementary metric. This dichotomy stems from the exclusive reliance of SwinUNet on transformer connections within its U-shaped architecture, harnessing global context to enhance precision, whereas SAM2UNet incorporates convolutional decoders to strengthen fine-grained detail representation. Crucially, SAM2MS strikes an optimal balance by synergizing global modeling with local refinement, demonstrating superior metric equilibrium despite marginally lower absolute values in individual precision and recall measures. Compared to methods (MSMDFFNet, SAM2UNet) proposed in the year 2024, SAM2MS improves F1 by 3.04 and reduces MAE by 0.16 on DeepGlobe, validating its architectural efficacy.

As evidenced in Tables 2 and 3, our model maintains superior performance across multiple metrics on both SpaceNet and Massachusetts datasets. This demonstrates enhanced robustness and generalizability beyond single-dataset validation contingencies. Nevertheless, all models exhibit performance disparities when compared to results achieved on DeepGlobe, attributable to dual factors: intrinsic dataset characteristics and inter-dataset heterogeneity. Comprehensive analysis of these cross-dataset variations is provided in Section 4.5.

Table 3. Cross-model quantitative evaluation on Massachusetts road datasets benchmark.

Method	Massachusetts Dataset [46]							
	Param. (M)	FLOPs (G)	Prec.	Recall	F1	mIoU	mDice	MAE ↓
UNet (*2015*)	31.02	875.81	69.46	65.84	67.60	**50.64**	63.95	4.65
UNet++ (*2018*)	47.19	3202.89	71.40	62.51	66.66	49.51	62.66	4.13
D-LinkNet (*2018*)	217.64	481.25	69.03	66.34	67.66	52.22	65.36	2.73
MSNet (*2021*)	27.69	143.91	60.11	52.22	55.89	23.15	31.85	4.88
M2SNet (*2023*)	27.69	144.41	59.95	52.76	56.12	22.75	31.37	6.09
Seg-Road (*2023*)	28.68	314.41	60.45	51.03	55.34	24.10	33.33	4.18
SwinUNet (*2023*)	27.14	123.59	**78.80**	56.15	65.57	49.70	63.10	**2.64**
SGCNNet (*2022*)	42.73	1234.41	71.85	52.64	60.77	44.17	57.87	3.09
MSMDFFNet (*2024*)	603.21	39.26	71.94	60.88	65.95	49.79	63.21	2.80
SAM2UNet (*2024*)	863.72	216.41	65.37	61.21	63.22	49.41	62.76	4.33
SAM2MS (*Ours*)	867.28	217.11	72.71	**69.39**	**71.01**	50.66	**64.10**	2.93

Downward arrows (↓) indicate that lower values correspond to better performance.

4.4. Ablation Studies

We systematically validate component effectiveness in SAM2MS through three progressive experimental groups: Initial tests with SAM2-Hiera tiny (38.9 M)/small (46 M)/base+ (80.8 M)/large (224.4 M) encoders demonstrate scale effects, where SAM2-Hiera-large's visual prior knowledge transfer elevates F1-score to 80.31%. As shown in Table 4, this ablation study compares the impact of different backbones and the use of adapter on performance. Key findings include the following:

Effectiveness of Adapter: When using the adapter, all backbones achieve significant improvements in F1, mIoU, and mDice, along with reduced MAE. For example, SAM2-Large with the adapter increases F1 from 74.99 to 80.31 and reduces MAE from 2.29 to 1.80, demonstrating the capability of the adapter to enhance feature adaptation.

Impact of Backbone Scale: Model scale positively correlates with performance. Without the adapter, SAM2-Large slightly outperforms others (e.g., F1 = 74.99). With the adapter, it shows the largest gains (F1 improvement of 5.32), indicating that larger models benefit more from the adapter. As shown in Table 5, this ablation study investigates the impact of different lossnet architectures. Key findings include the following:

Table 4. Ablation studies on backbones and adapters.

Backbones	Adapter	Evaluation Metrics			
		F1	mIoU	mDice	MAE ↓
SAM2-Tiny	–	74.36	57.77	71.76	2.29
SAM2-Tiny	✓	76.81	61.11	74.47	2.16
SAM2-Small	–	74.19	57.52	71.60	2.39
SAM2-Small	✓	77.26	61.83	75.01	2.09
SAM2-Base+	–	74.39	57.91	71.89	2.29
SAM2-Base+	✓	78.11	63.03	76.00	2.04
SAM2-Large	–	74.99	58.56	72.44	2.29
SAM2-Large	✓	80.31	64.24	77.93	1.80

Downward arrows (↓) indicate that lower values correspond to better performance.

Table 5. Ablation study on lossnet.

Backbones	Adapter	Lossnet	Evaluation Metrics		
			F1	mIoU	MAE ↓
SAM2-Large	✓	–	78.77	64.08	2.00
SAM2-Large	✓	VGG16	78.89	64.21	1.84
SAM2-Large	✓	ResNet50	80.31	64.24	1.80

Downward arrows (↓) indicate that lower values correspond to better performance.

Necessity of lossnet: As evidenced in Figure 8, the inference results progressively converge toward ground truth labels through iterative optimization within the lossnet framework. The outputs of the model evolve from exhibiting significant fragmentation and ghosting artifacts in early stages to progressively sharpened predictions that achieve close alignment with the reference annotations. Introducing lossnet (e.g., VGG16 [47] or ResNet50 [43], pretrained on the ImageNet [42] dataset) improves all metrics compared to the baseline without lossnet (F1 = 78.77 → 78.89/80.31), indicating that lossnet enhances supervisory learning.

Impact of lossnet architecture: ResNet50 as lossnet outperforms VGG16 significantly. For instance, F1 increases from 78.89 (VGG16) to 80.31, and MAE decreases from 1.84 to 1.80, demonstrating the superiority of deep residual networks in feature extraction and loss supervision.

Impact of MSSM and DRB: This set of ablation studies was conducted on the DeepGlobe dataset using the optimal SAM2-Large backbone as the encoder to ensure consistent

data comparison. (1) Quantitative metrics in Table 6 to reflect the importance of the MSSM. When this key component is omitted, both the mIoU and MAE metrics show noticeable degradation compared to the complete model, indicating increased errors in distinguishing between road and non-road areas in the final predictions. These results demonstrate the essential role of the MSSM in refining segmentation outcomes—such as suppressing artifacts and clarifying boundaries—by eliminating redundant information across multi-scale features. (2) Although the removal of the DRB component only resulted in a 1.59% decrease in F1-score, occasional gradient explosion was observed during training. Eliminating the DRB forced the model to rely on a simple linear layer to reduce channel dimensions when transmitting features from the encoder to the MSSM, leading to suboptimal transitions from high- to low-channel feature maps. To achieve better metric performance without the DRB, the number of training epochs was increased from 50 to 80.

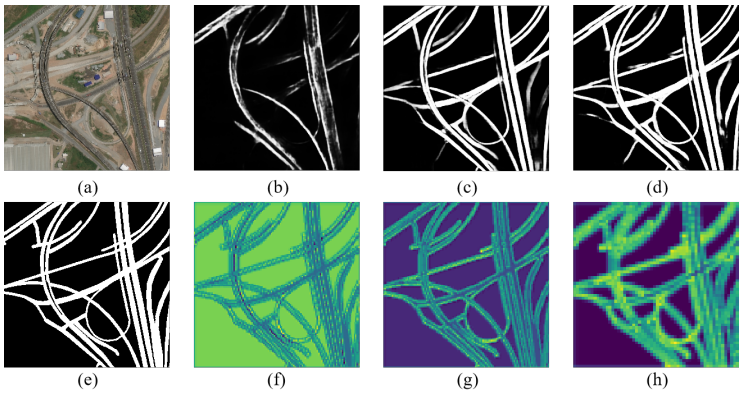


Figure 8. Progressive training visualization. To demonstrate model evolution during training, we present a randomly selected test image alongside its ground truth (**a**,**e**). Successive inference results from early-stage (**b**), intermediate (**c**), and final (**d**) training phases illustrate performance progression. Complementary multi-level supervision maps generated by lossnet (f-h) highlight critical refinement processes: background suppression (**f**), edge refinement (**g**) and region-of-interest enhancement (**h**).

Table 6. Ablation study on MSSM and DRB.

Backbones	MSSM	DRB	Evaluation Metrics		
			F1	mIoU	MAE ↓
SAM2-Large	–	✓	77.44	61.64	3.18
SAM2-Large	✓	–	78.72	63.16	1.94
SAM2-Large	✓	✓	80.31	64.24	1.80

Downward arrows (↓) indicate that lower values correspond to better performance.

4.5. Cross-Dataset Validation

To demonstrate the robustness of visual foundation models and their potential for remote sensing tasks, we establish a cross-dataset validation framework for comparative analysis. Specifically, models trained on DeepGlobe undergo direct inference testing on the SpaceNet test set (denoted D2S) and Massachusetts test set (denoted D2M). These are rigorously compared against models trained and tested within the same target domain (denoted S2S for SpaceNet and M2M for Massachusetts).

Key considerations include the following: (1) Underlying distribution shift: DeepGlobe primarily features unpaved roads in non-urban environments—characterized by narrow, vegetation-occluded pathways—whereas SpaceNet and Massachusetts focus on structured urban contexts dominated by asphalt roads with limited suburban coverage. (2) Data volume disparity: SpaceNet contains approximately 1.6× to 2× more training samples than

DeepGlobe and Massachusetts. (3) Resolution and annotation characteristics: Significant variations exist in ground sampling distance (GSD) and labeling protocols (Section 3.1).

As evidenced in Figure 9, both SAM2UNet and our SAM2MS achieve near-perfect alignment with SpaceNet road labels. Crucially, SAM2MS surpasses SAM2UNet in topological fidelity—precisely reconstructing true road geometry while eliminating spurious artifacts—while maintaining superior connectivity.

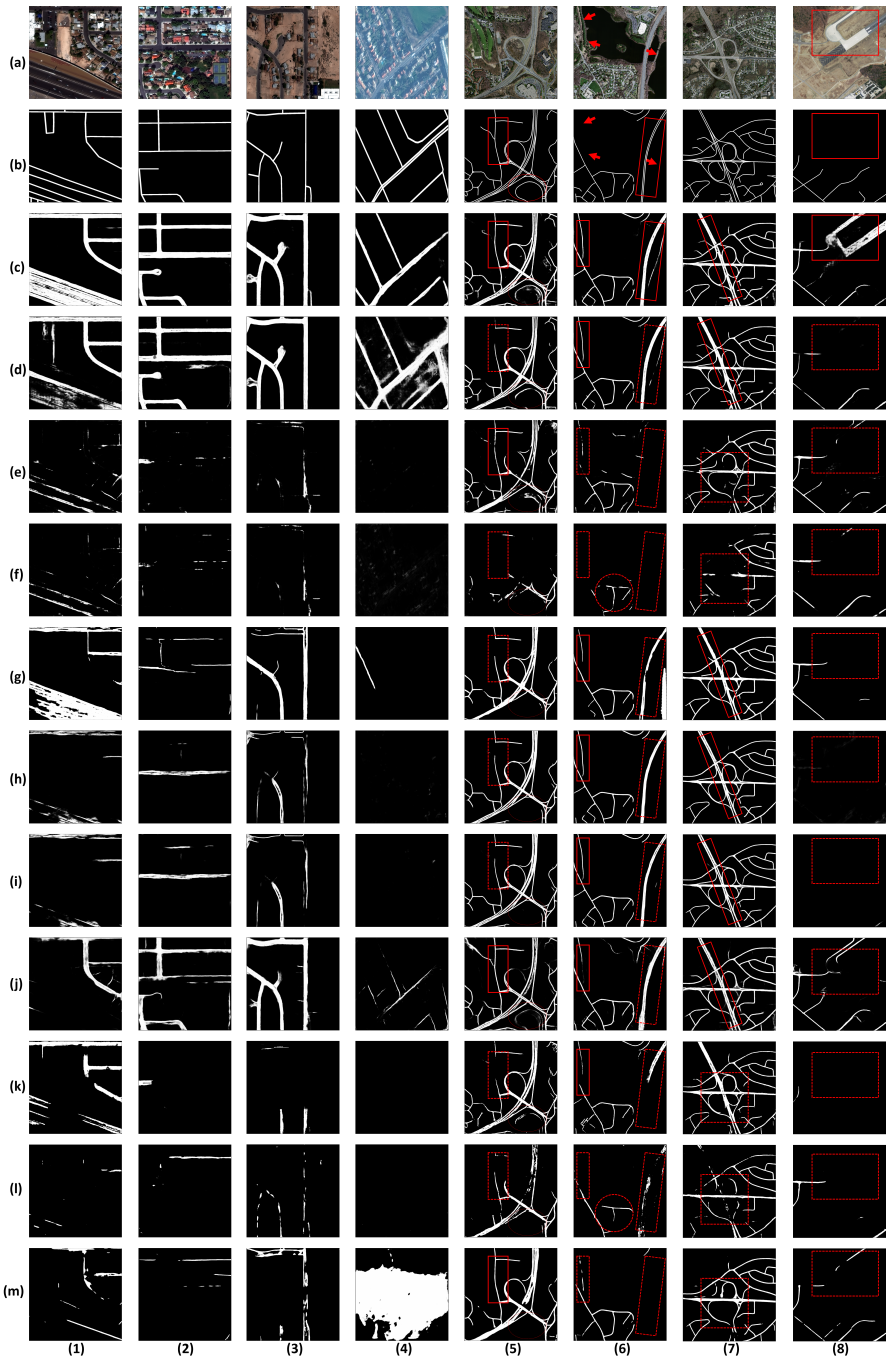


Figure 9. Cross-dataset comparative results: Models trained on the DeepGlobe dataset were directly evaluated on the test sets of SpaceNet and Massachusetts (denoted as D2S and D2M, respectively). Columns (1–4) present partial inference results for D2S, while columns (5–8) correspond to D2M. Row (a) displays the actual remote sensing images, and row (b) presents the corresponding ground truth labels. Rows (c–l) sequentially demonstrate the test results of baseline models: UNet, UNet++, D-LinkNet, MSNet, M2SNet, Seg-Road, SwinUNet, SGCNNet, MSMDFFNet and SAM2UNet. Finally, row (m) showcases the performance of our proposed SAM2MS method.

Furthermore, the red arrows and boxes highlight the important parts of the reasoning results. Due to comparable data volumes and low intra-class variance between DeepGlobe and Massachusetts, domain transfer proves more effective. This is reflected in the marginal RLOD gap between D2M and M2M configurations.

Integrated analysis of the data distributions in DeepGlobe, SpaceNet, and Massachusetts (Figure 10) reveals distinct clustering patterns: DeepGlobe and Massachusetts exhibit compact intra-class distributions with substantial inter-dataset divergence. Conversely, SpaceNet demonstrates a fragmented distribution characterized by multiple dispersed clusters and high intra-class variance, presenting significant challenges for model training and evaluation. These observations underscore the critical importance of cross-dataset experiments for comprehensively assessing model robustness and generalization capability.

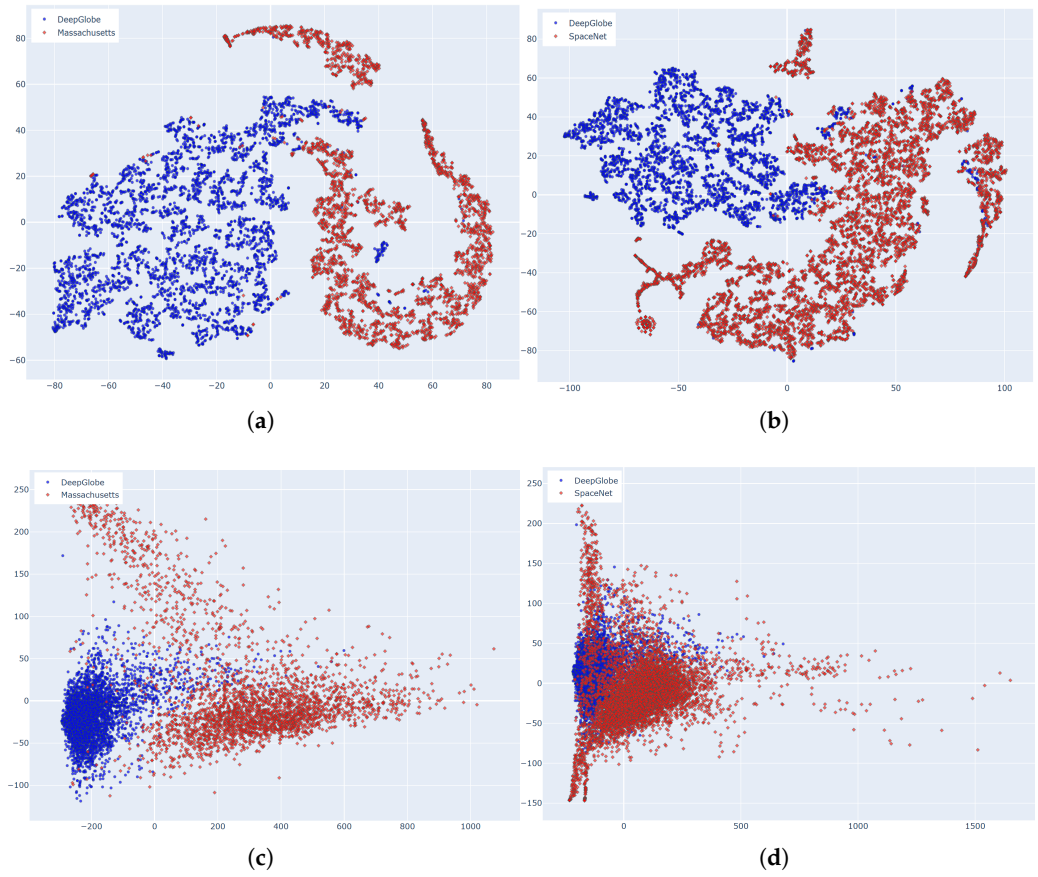


Figure 10. The spatial distribution relationships among the DeepGlobe, SpaceNet, and Massachusetts datasets were analyzed using both t-SNE (**a**,**b**) and PCA (**c**,**d**) methods, with feature scatter points visualized accordingly: samples from the DeepGlobe dataset are represented by blue scatter points, whereas samples from the Massachusetts and SpaceNet datasets are depicted using red scatter points.

As evidenced in Table 7, the cross-dataset experiments reveal profound domain shift dynamics: conventional models (e.g., UNet++) suffer catastrophic domain adaptation collapse in D2S and D2M scenarios, whereas the SAM2 series—leveraging the generalized representation power of vision foundation models—achieves breakthrough cross-domain performance with SAM2UNet (55.78%) and SAM2MS (60.84%), surpassing the best non-SAM model Seg-Road (37.29%) by over 23 percentage points. For in-domain performance (S2S), SAM2MS dominates all baselines at 68.45% RLOD, validating its architectural superiority. Crucially, SAM2MS exhibits exceptional cross-domain robustness—manifesting a mere 7.61% D2S-S2S performance gap versus the 72.3% average degradation in traditional models (e.g., the plummet of SwinUNet from 53.01% to 18.94%). This superiority stems from the multi-scale subtraction module's precise modeling of road topology, effec-

tively mitigating feature distribution shifts induced by annotation discrepancies (boundary delineation in DeepGlobe vs centerline labeling in SpaceNet).

Table 7. Quantitative analysis of cross-dataset experiments.

Method	D2S	S2S	D2M	M2M
	RLOD	RLOD	RLOD	RLOD
UNet (*2015*)	10.21	57.46	50.51	66.60
UNet++ (*2018*)	5.32	26.00	16.53	61.75
D-LinkNet (*2018*)	18.96	63.41	51.50	66.26
MSNet (*2021*)	15.94	59.31	51.53	58.73
M2SNet (*2023*)	18.69	58.55	52.44	59.00
Seg-Road (*2023*)	37.29	67.78	58.50	58.19
SwinUNet (*2023*)	18.94	53.01	52.10	55.98
SGCNNet (*2022*)	8.14	50.18	26.41	52.73
MSMDFFNet (*2024*)	8.02	60.37	45.93	61.14
SAM2UNet (*2024*)	55.78	63.64	63.62	62.94
SAM2MS (*Ours*)	60.84	68.45	67.16	69.61

4.6. Limitations

While the proposed model demonstrates solid performance in the aforementioned experiments, certain limitations remain that warrant further investigation in subsequent work.

Dataset Quality Sensitivity: While SAM2MS leverages the SAM2 visual foundation model to achieve enhanced generalization capability and robustness, it remains highly sensitive to dataset quality. Models trained on high-quality annotated datasets often achieve competitive performance on other datasets without the need for additional training. Conversely, due to the nature of supervised learning, poor annotation quality can significantly impair the training process, leading to inference results that fall short compared to models trained on well-annotated data.

Resolution Sensitivity: Both SAM2MS and the adopted baseline models were trained exclusively on 1024 × 1024 pixel images. This fixed input scale leads to suboptimal adaptation to the inherent resolution variations present in high-resolution remote sensing datasets. Empirically, while DeepGlobe and SpaceNet imagery share a resolution of 0.5 m per pixel, Massachusetts imagery features a resolution of 1.0 m per pixel. This discrepancy correlates with observed performance degradation, where nearly all models demonstrate inferior results on the Massachusetts dataset compared to DeepGlobe.

Model Size: Taking an HRSI input of size 1024 × 1024 as an example, although an efficient fine-tuning strategy for visual foundation models has been adopted, this approach inevitably increases the overall parameter count. As a result, our SAM2MS model contains over 800 million parameters and exhibits computational requirements exceeding 200 GFLOPs on an RTX 4060 GPU. While achieving competitive performance on evaluation metrics, the computational efficiency remains an aspect with significant room for improvement. Consequently, deployment on resource-constrained edge computing platforms, such as Raspberry Pi or drones, may present substantial challenges.

5. Conclusions

This paper proposes a novel SAM2MS architecture for road extraction from high-resolution remote sensing images, with innovative designs to significantly enhance generalization capability and robustness against occlusions, shadows, and noise interference. For the first time, parameter-efficient fine-tuning via an adapter-based strategy successfully transfers the foundational vision model (SAM2) to road extraction tasks, substantially advancing the performance boundaries of foundational models in remote sensing domains. A cross-layer differential feature enhancement mechanism is innovatively developed, where

multi-scale subtraction operations eliminate feature redundancy while refining detail representation. Furthermore, a non-parametric lossnet with triple constraints (background suppression, contour refinement, and region consistency) establishes a multi-level supervision strategy. Evaluations on three datasets demonstrate superior performance over baseline methods across key metrics, with ablation studies confirming the efficacy of the SAM2 backbone, adapter module, and lossnet. Cross-dataset transfer experiments validate exceptional generalization capabilities.

Future work will focus on (1) exploring knowledge distillation mechanisms between general vision models and road morphology priors to optimize model-data equilibrium via semi-/unsupervised learning; (2) developing lightweight edge-computing variants for real-time deployment in vehicle navigation and autonomous driving path planning.

Author Contributions: Conceptualization, P.Z. and J.L.; methodology, P.Z.; software, P.Z. and C.W.; validation, P.Z. and J.L.; formal analysis, P.Z.; investigation, P.Z. and C.W.; resources, P.Z., Y.N. and J.L.; data curation, P.Z.; writing—original draft preparation, P.Z.; writing—review and editing, J.L. and C.W.; visualization, P.Z., C.W. and J.L.; supervision, Y.N. and J.L.; Project administration, Y.N. and J.L.; funding acquisition, Y.N. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (NSFC) grant numbers 61790565 and 62576349.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data presented in this study are available on request from the author Pengnian Zhang (zhangpengnianizsx@nudt.edu.cn). Part of the data are not publicly available due to our laboratory's confidentiality agreement and policies.

Acknowledgments: The authors gratefully acknowledge Jinjing Zhao and Kunzhong Miu for their insightful discussions, and the reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5612822. [CrossRef]
2. Wang, X.; Jin, X.; Dai, Z.; Wu, Y.; Chehri, A. Deep learning-based methods for road extraction from remote sensing images: A vision, survey, and future directions. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 55–78. [CrossRef]
3. Xu, Z.; Liu, Y.; Sun, Y.; Liu, M.; Wang, L. Rngdet++: Road network graph detection by transformer with instance segmentation and multi-scale features enhancement. *IEEE Robot. Autom. Lett.* **2023**, *8*, 2991–2998. [CrossRef]
4. Li, T.; Ye, S.; Li, R.; Fu, Y.; Yang, G.; Pan, Z. Topology-aware road extraction via multi-task learning for autonomous driving. In Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24–28 September 2023; pp. 2275–2281.
5. Mei, J.; Li, R.-J.; Gao, W.; Cheng, M.-M. CoANet: Connectivity attention network for road extraction from satellite imagery. *IEEE Trans. Image Process.* **2021**, *30*, 8540–8552. [CrossRef] [PubMed]
6. Ge, C.; Nie, Y.; Kong, F.; Xu, X. Improving road extraction for autonomous driving using swin transformer unet. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 1216–1221.
7. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Roll, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo W.Y. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026.
8. Ravi, N.; Gabeur, V.; Hu, Y.T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Roll, C.; Gustafson, L.; et al. Sam 2: Segment anything in images and videos. *arXiv* **2024**, arXiv:2408.00714.
9. Lu, X.; Weng, Q. Multi-LoRA fine-tuned segment anything model for urban man-made object extraction. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5935411. [CrossRef]

10. Hetang, C.; Xue, H.; Le, C.; Yue, T.; Wang, W.; He, Y. Segment anything model for road network graph extraction. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 16–22 June 2024; pp. 2556–2566. [CrossRef]
11. Yin, P.; Li, K.; Cao, X.; Yao, J.; Liu, L.; Bai, X.; Zhou, F.; Meng, D. Towards satellite image road graph extraction: A global-scale dataset and a novel method. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 1527–1537. [CrossRef]
12. Lian, R.; Wang, W.; Mustafa, N.; Huang, L. Road extraction methods in high-resolution remote sensing images: A comprehensive review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5489–5507. [CrossRef]
13. Maarir, A.; Bouikhalene, B. Roads detection from satellite images based on active contour model and distance transform. In Proceedings of the 2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV), Beni Mellal, Morocco, 29 March–1 April 2016; pp. 94–98.
14. Lin, X.; Xie, W.; Zhang, L.; Sang, H.; Shen, J.; Cui, S. Semi-automatic road extraction from high resolution satellite images by template matching using Kullback–Leibler divergence as a similarity measure. *Int. J. Image Data Fusion* **2022**, *13*, 316–336. [CrossRef]
15. Abdollahi, A.; Bakhtiari, H.R.R.; Nejad, M.P. Investigation of SVM and level set interactive methods for road extraction from google earth images. *J. Indian Soc. Remote Sens.* **2018** *46*, 423–430. [CrossRef]
16. Maurya, R.; Gupta, P.R.; Shukla, A.S. Road extraction using k-means clustering and morphological operations. In Proceedings of the 2011 International Conference on Image Information Processing, Shimla, India, 3–5 November 2011; pp. 1–6.
17. Palanivel, E.; Selvan, S. Unsupervised multispectral Gaussian mixture model-based framework for road extraction. *J. Indian Soc. Remote Sens.* **2025**, *53*, 373–388. [CrossRef]
18. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
19. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Munich, Germany, 5–9 October 2015; pp. 234–241.
21. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–11.
22. Gao, L.; Zhou, Y.; Tian, J.; Cai, W.; Lv, Z. MCMCNet: A Semi-supervised Road Extraction Network for High-resolution Remote Sensing Images via Multiple Consistency and Multi-task Constraints. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4410416. [CrossRef]
23. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
24. Zhao, X.; Zhang, L.; Lu, H. Automatic polyp segmentation via multi-scale subtraction network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*; Springer International Publishing: Cham, Switzerland, 2021; pp. 120–130.
25. Zhao, X.; Jia, H.; Pang, Y.; Lv, L.; Tian, F.; Zhang, L.; Sun, W.; Lu, H. M^2SNet: Multi-scale in multi-scale subtraction network for medical image segmentation. *arXiv* **2023**, arXiv:2303.10894.
26. Zhou, G.; Chen, W.; Gui, Q.; Li, X.; Wang, L. Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5614115. [CrossRef]
27. Wang, Y.; Tong, L.; Luo, S.; Xiao, F.; Yang, J. A multi-scale and multi-direction feature fusion network for road detection from satellite imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5615718. [CrossRef]
28. Zhang, P.; Li, J.; Dai, C.; Niu, Y. BNW: Multi-level road extraction tasks methods—A review. In Proceedings of the International Conference on Autonomous Unmanned Systems, Shenyang, China, 19–21 September 2024; pp. 370–392.
29. Dai, L.; Zhang, G.; Zhang, R. RADANet: Road augmented deformable attention network for road extraction from complex high-resolution remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602213. [CrossRef]
30. Jiang, X.; Li, Y.; Jiang, T.; Xie, J.; Wu, Y.; Cai, Q.; Jiang, J.; Xu, J.; Zhang, H. RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102987. [CrossRef]
31. Tao, J.; Chen, Z.; Sun, Z.; Guo, H.; Leng, B.; Yu, Z.; Wang, Y.; He, Z.; Lei, X.; Yang, J. Seg-Road: A segmentation network for road extraction based on transformer and CNN with connectivity structures. *Remote Sens.* **2023**, *15*, 1602. [CrossRef]
32. Qiao, Y.; Zhong, B.; Du, B.; Cai, H.; Jiang, J.; Liu, Q.; Yang, A.; Wu, J.; Wang, X. Sam enhanced semantic segmentation for remote sensing imagery without additional training. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5610816. [CrossRef]

33. Chen, T.; Zhu, L.; Deng, C.; Cao, R.; Wang, Y.; Zhang, S.; Li, Z.; Sun, L.; Zang, Y.; Mao, P. Sam-adapter: Adapting segment anything in underperformed scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 3367–3375.
34. Luo, M.; Zhang, T.; Wei, S.; Ji, S. SAM-RSIS: Progressively adapting SAM with box prompting to remote sensing image instance segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4413814. [CrossRef]
35. Xiong, X.; Wu, Z.; Tan, S.; Li, W.; Tang, F.; Chen, Y.; Li, S.; Ma, J.; Li, G. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv* **2024**, arXiv:2408.08870. [CrossRef]
36. Lin, X.; Xiang, Y.; Zhang, L.; Yang, X.; Yan, Z.; Yu, L. Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. *arXiv* **2023**, arXiv:2309.06824. [CrossRef]
37. Yang, S.; Bi, H.; Zhang, H.; Sun, J. SAM-UNet: Enhancing Zero-Shot Segmentation of SAM for Universal Medical Images. *arXiv* **2024**, arXiv:2408.09886.
38. Gao, J.; Zhang, D.; Wang, F.; Ning, L.; Zhao, Z.; Li, X. Combining SAM With Limited Data for Change Detection in Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5614311. [CrossRef]
39. Zhao, X.; Ding, W.; An Y.; Du, Y.; Yu, T.; Li, M.; Tang, M. Fast segment anything. *arXiv* **2023**, arXiv:2306.12156. [CrossRef]
40. Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.; Cheng, M.-M.; Hu, S.-M. Segnext: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156.
41. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
42. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
45. Etten, A.V.; Lindenbaum, D.; Bacastow, T.M. Spacenet: A remote sensing dataset and challenge series. *arXiv* **2018**, arXiv:1807.01232.
46. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto (Canada): Toronto, ON, Canada, 2013.
47. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.