

PathMamba: A Hybrid Mamba-Transformer for Topologically Coherent Road Segmentation in Satellite Imagery

Jules Decaestecker^{*} Nicolas Vigne[†]

Thales CortAIX Labs

Abstract

Achieving both high accuracy and topological continuity in road segmentation from satellite imagery is a critical goal for applications ranging from urban planning to disaster response. State-of-the-art methods often rely on Vision Transformers, which excel at capturing global context, yet their quadratic complexity is a significant barrier to efficient deployment, particularly for on-board processing in resource-constrained platforms. In contrast, emerging State Space Models like Mamba offer linear-time efficiency and are inherently suited to modeling long, continuous structures. We posit that these architectures have complementary strengths. To this end, we introduce PathMamba, a novel hybrid architecture that integrates Mamba’s sequential modeling with the Transformer’s global reasoning. Our design strategically uses Mamba blocks to trace the continuous nature of road networks, preserving topological structure, while integrating Transformer blocks to refine features with global context. This approach yields topologically superior segmentation maps without the prohibitive scaling costs of pure attention-based models. Our experiments on the DeepGlobe Road Extraction [10] and Massachusetts Roads [27] datasets demonstrate that PathMamba sets a new state-of-the-art. Notably, it significantly improves topological continuity, as measured by the APLS metric [13], setting a new benchmark while remaining computationally competitive.

1. Introduction

Road segmentation from satellite imagery is a critical task in remote sensing with wide-ranging applications in autonomous navigation, urban planning, and disaster response [26, 27]. The primary goal is to produce a precise pixel-wise mask that delineates road networks [24]. This task is inherently difficult due to numerous challenges, including



Figure 1. An example from the Massachusetts Roads dataset illustrating the challenges of road segmentation. The input satellite image (left) contains significant occlusions from vegetation and complex shadows, while the goal is to produce a precise, topologically coherent ground-truth mask (right).

frequent occlusions from buildings and vegetation, inconsistent lighting and shadows, the thin and elongated structure of roads, and severe class imbalance between road and background pixels [9, 47]. These factors, illustrated in Figure 1, make it especially hard to maintain the topological integrity of the predicted network, often resulting in fragmented segments [28].

For years, Convolutional Neural Networks (CNNs) [30, 31] have been the foundational architecture for segmentation models. Their strong inductive biases centered on locality are effective for general feature extraction [29], but their intrinsically localized receptive fields struggle to model the long-range dependencies required to trace a road’s path across an entire complex scene [42]. To overcome this limitation, Vision Transformers (ViTs) [12] were introduced, leveraging a self-attention mechanism to capture global context. This approach led to superior performance in many vision tasks, including road segmentation [6, 40]. However, the quadratic complexity of self-attention with respect to input size makes ViTs computationally expensive and difficult to scale for high-resolution satellite imagery [36].

Recently, State Space Models (SSMs) have emerged

^{*}decaestecker.jules@gmail.com

[†]nicolas.vigne@thalesgroup.com

as a compelling alternative, challenging the dominance of Transformers [15]. Initially developed for natural language processing (NLP) to efficiently model long sequences, the Mamba architecture [14] introduced a selective scanning mechanism that combines linear-time complexity with the ability to capture long-range dependencies. This success has spurred its adaptation to computer vision, leading to a new family of models. Architectures like Vision Mamba (ViM) [53] and the hierarchical VMamba [22] have successfully adapted the sequential SSM core for 2D image processing by serializing patches into sequences, demonstrating performance on par with Transformers but with significantly better scalability.

Maintaining network connectivity is a critical challenge where standard pixel-wise losses often fall short [28]. Current approaches to improve topology typically rely on two strategies: complex post-processing steps to reconnect fragmented road segments [5, 25], or the integration of specialized topological loss functions into the training process [3, 28]. Our work pushes this limit by proposing an architectural solution designed to inherently generate more topologically coherent segmentation maps, thereby reducing the reliance on such auxiliary techniques.

We posit that an optimal architecture for road segmentation can be achieved by combining the complementary strengths of Mamba and Transformers, a direction also explored in recent hybrid models [16, 46]. Mamba’s unique advantage lies in its selective state-space mechanism [14], which allows it to dynamically compress and carry relevant contextual information along a sequence. This is not merely a matter of efficiency; it is a fundamental architectural advantage for modeling long-range dependencies. Prior work on the Long Range Arena benchmark [34] demonstrated that foundational SSMs could achieve near-perfect scores on the Path-X task, a challenge requiring a model to determine if two points are connected by a path of dashes amidst numerous distractor paths, where Transformers trained from scratch were unable to solve the task [2, 15]. The conceptual parallel between this task and tracing occluded roads is striking: the path composed of dashes creates a visually discontinuous sequence, functionally equivalent to a road that is intermittently occluded by trees and shadows. In both scenarios, the core challenge is to maintain contextual memory and infer the underlying topological connection despite interruptions in the input data. This makes Mamba’s state-space mechanism, which is designed to carry information across long sequences, exceptionally well-suited for modeling such structures. Conversely, the Transformer’s strength is its content-based global self-attention [36], which performs an all-pairs comparison of image patches. While computationally intensive, this mechanism is unparalleled for understanding complex, isotropic spatial relationships (like intersections) and

resolving ambiguities by drawing context from the entire scene at once. We propose a novel hybrid backbone where Mamba blocks are primarily used to model the long, continuous structure of roads, while strategically placed Transformer blocks provide robust global reasoning to connect disparate segments. This targeted combination aims to preserve the topological continuity that Mamba excels at, while leveraging the Transformer’s power for global scene understanding to improve overall accuracy, a core hypothesis that we validate quantitatively in our experimental analysis (see Section 4.5).

Our main contributions are:

- A novel hybrid Mamba-Transformer backbone that sequentially leverages Mamba’s continuity modeling and the Transformer’s global context aggregation within a single stage.
- State-of-the-art performance on the DeepGlobe Road Extraction benchmark across IoU, F1-score, and the topology-focused APLS metric [13].
- The highest reported APLS score on the challenging Massachusetts Roads dataset, demonstrating superior topological integrity in the predicted road networks.
- Comprehensive ablation studies validating our architectural design choices and demonstrating the complementary contributions of the Mamba and Transformer components.

2. Related Work

Pre-Deep Learning Methods. Before the widespread adoption of deep learning, road extraction from aerial and satellite imagery was tackled using a variety of classical computer vision and machine learning techniques [38]. These methods often relied on handcrafted features and the explicit modeling of road properties. Prominent approaches included probabilistic and graph-based models, which were combined to identify and structure road networks [35]. This family of methods included the use of geometric-stochastic models to find main roads [4] and Gibbs point processes to handle complex road geometries [32]. Complementary techniques focused on recovering line networks by detecting junctions [7] or using active contour models, also known as snakes, to delineate road boundaries based on scale space analysis [20]. While foundational, these methods often struggled with the complex variations and occlusions present in real-world imagery, paving the way for data-driven deep learning solutions.

CNN-based Road Segmentation. Early and influential approaches to road segmentation relied on fully convolutional networks. Architectures like U-Net [30] and DeepLab [8] set strong baselines. D-LinkNet [52] specifically tailored its design for road extraction by using dilated convolutions in its center path to enlarge the receptive field

and preserve spatial information, achieving state-of-the-art results at the time. More recent works continue to build on these foundations, introducing refinements such as attention mechanisms and novel fusion strategies to further improve performance [17, 37]. However, the inherently local nature of convolutions remains a limitation for capturing global road network topology, a challenge that persists even in modern CNN designs [1].

Transformer-based Models. Vision Transformers (ViTs) [12] and their hierarchical variants like the Swin Transformer [23] introduced global self-attention to vision, overcoming the limited receptive fields of CNNs [55]. Models like SegFormer [40] and Swin-Unet [6] extended this paradigm to semantic segmentation, demonstrating strong performance on road extraction benchmarks [19, 49]. Despite their success, the quadratic complexity of self-attention poses a significant computational barrier for high-resolution remote sensing applications [36].

State Space Models in Vision. State Space Models (SSMs) have recently been adapted for vision tasks as an efficient alternative to Transformers. Their theoretical strength in handling long sequences was empirically validated by precursors like S4 [15] on the demanding Long Range Arena benchmark [34]. Notably, on the Path-X benchmark within this suite, which requires identifying a continuous logical path from a series of visual dashes, S4 achieved a high score while Transformer-based models trained from scratch failed. The task’s use of a dashed path serves as a strong proxy for real-world continuity challenges, such as tracing road networks fragmented by occlusions. This result highlights that the core SSM mechanism is inherently superior for tasks demanding long-range spatial continuity. Mamba [14] builds upon this success by introducing a selective scan mechanism that further enhances the ability to model these dependencies with linear-time complexity. This has led to vision-specific adaptations like Vision Mamba (ViM) [53] and the hierarchical VMamba [22]. VMamba, in particular, uses a Cross-Scan Module to traverse image patches in multiple directions, effectively capturing 2D spatial context by applying the sequential SSM mechanism along these generated paths. Concurrently, other works have explored Mamba’s potential specifically for remote sensing. For instance, RS-Mamba [50] proposed a specialized architecture that also leverages Mamba for segmentation and has shown competitive results. Hybrid architectures like MambaVision [16] have also emerged, combining Mamba and Transformer blocks to leverage their complementary strengths. Our work builds on this direction but proposes a specific sequential arrangement tailored for the unique challenges of road segmentation.

Approaches for Topological Coherence. Pixel-wise metrics like IoU often fail to penalize topological errors, such as disconnected road segments, which are critical for navigation applications. This has motivated research into methods that explicitly optimize for topology [13]. These approaches typically fall into two categories: complex post-processing steps to reconnect fragmented segments [5], or the integration of specialized topological loss functions into the training process [28]. Architectural innovations, such as the graph-based reasoning in SPIN [3], often function as refinement modules that operate on features from a standard backbone to enforce connectivity. While effective, these methods are primarily reactive, designed to repair topological errors after they occur. Our work diverges from this paradigm by proposing a preventative, architectural solution. We hypothesize that by integrating a mechanism inherently suited to modeling continuity, the Mamba SSM, directly into the backbone, the network can learn to generate feature representations that are already topologically sound, thereby reducing the dependency on complex loss functions or post-processing heuristics.

3. Methodology

3.1. State Space Model Preliminaries

At its core, PathMamba is built upon the State Space Model (SSM) framework. This section briefly reviews its foundational principles, as detailed in the original Mamba paper [14] and adapted for vision in architectures like VMamba [22]. A continuous-time SSM maps a 1D input signal $x(t) \in \mathbb{R}$ to an output $y(t) \in \mathbb{R}$ via a latent state $h(t) \in \mathbb{R}^N$. The system is governed by a set of linear ordinary differential equations:

$$h'(t) = Ah(t) + Bx(t) \quad (1)$$

$$y(t) = Ch(t) + Dx(t) \quad (2)$$

where $A \in \mathbb{R}^{N \times N}$ is the state matrix and $B \in \mathbb{R}^{N \times 1}$, $C \in \mathbb{R}^{1 \times N}$, $D \in \mathbb{R}$ are projection matrices.

A crucial characteristic of modern SSMs, including Mamba, is that they operate as single-input single-output (SISO) systems. To handle the multi-channel feature maps common in computer vision, where an input can be considered $x(t) \in \mathbb{R}^D$, the SSM processes each of the D channels independently. This results in a much larger effective state size of $N \times D$, where N is the state expansion factor. This expansion of the latent state is a key element that allows the model to capture complex relationships within information-dense data like images [14].

To be used in deep learning, this system must be discretized. Using a timestep Δ , the continuous parameters are transformed into discrete counterparts \bar{A}, \bar{B} via a dis-

cretization rule like the Zero-Order Hold (ZOH):

$$\bar{A} = \exp(\Delta A) \quad (3)$$

$$\bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I)B \quad (4)$$

The discrete-time SSM is then given by:

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t \quad (5)$$

$$y_t = \bar{C}h_t i \quad (6)$$

Following the convention in recent works [14, 22], we omit the direct feedthrough matrix D as its function is captured by residual connections in the overall block architecture.

Mamba enhances this formulation with a selective scan mechanism, where the B, C matrices and the timestep Δ are dynamically generated from the input data itself. This makes the SSM input-dependent and time-variant, allowing the model to selectively focus on or ignore parts of the input sequence. This selectivity breaks the property of time-invariance, which precludes the use of standard convolution. Mamba therefore employs an efficient parallel scan algorithm for computation instead [14].

3.2. Architectural Foundations

Our architecture is built upon the hierarchical design popularized by Swin Transformer [23] and VMamba [22]. The model processes an input image through four stages, progressively downsampling the spatial resolution while increasing the channel dimension. This multi-scale feature extraction is crucial for semantic segmentation, as it allows the model to capture both fine-grained details and high-level contextual information simultaneously [8, 21, 48].

The core component in the Mamba-based stages is the Visual State Space (VSS) block, illustrated in Figure 2. It follows a MetaFormer structure [44], which separates token-mixing from channel-mixing. The token-mixing is performed by a 2D-aware SSM module, while a standard Feed-Forward Network (FFN) handles channel-mixing. To apply the inherently sequential SSM to 2D image data, we employ the cross-scan strategy from VMamba. This involves serializing the image patches by scanning them along four cardinal directions, allowing the SSM to capture comprehensive spatial context.

3.3. Hybrid Mamba-Transformer Backbone

The primary contribution of this work is a novel hybrid backbone that strategically combines Mamba and Transformer blocks. Our design is motivated by the hypothesis that Mamba and Transformers have complementary strengths for road segmentation. This hypothesis is directly informed by prior work on long-range benchmarks. The task of road segmentation, especially with occlusions, is functionally analogous to the Path-X benchmark [34],

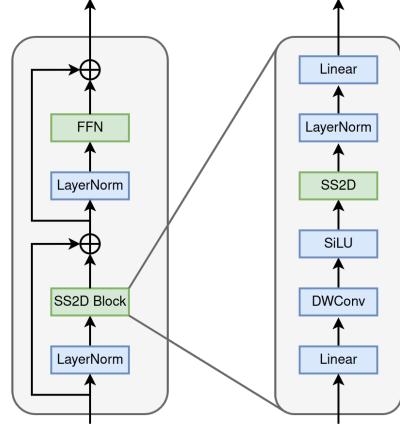


Figure 2. Architecture of a Visual State Space (VSS) block, which forms the basis of our Mamba stages. It uses an SSM for token-mixing and an FFN for channel-mixing.

where the challenge is to identify a logically continuous path composed of visually discontinuous dashes. These interruptions are a direct proxy for the challenge of tracing a road network that is intermittently hidden by occlusions like trees and shadows. Given that SSMs have proven uniquely capable of solving this task while Transformers trained from scratch have failed [15], we posit that Mamba’s architecture is fundamentally better suited for establishing topological continuity. As confirmed in our analysis in Section 4.5, Mamba excels at modeling these long, continuous structures. In contrast, Transformers excel at global contextual reasoning, making them powerful for improving overall classification accuracy. Our proposed hybrid stage is therefore designed to harness both capabilities.

We propose a four-stage hierarchical backbone, as shown in Figure 3. The first, second, and fourth stages are composed entirely of VSS blocks. The key innovation lies in the **third stage**, which is a hybrid design. It begins with a series of VSS blocks to efficiently process features and model continuity, followed by a series of standard Transformer blocks (Multi-Head Self-Attention).

The sequential arrangement within our hybrid stage, VSS blocks followed by Transformer blocks, is a deliberate design choice. The initial Mamba blocks act as efficient continuity modelers, tracing the elongated structure of roads and propagating contextual information along their paths. This process generates feature maps that are already imbued with a strong sense of topological structure. The subsequent Transformer blocks then operate on these semantically enriched features. Their global self-attention mechanism is perfectly suited to resolve complex spatial ambiguities, such as multi-road intersections, and to integrate information from disconnected regions of the image, effectively refining the continuous paths identified by Mamba.

This arrangement allows the model to first capture continuous features efficiently with Mamba at a medium resolution. Then, the Transformer blocks operate on these feature maps to perform global context aggregation. Placing the computationally intensive attention mechanism in a deeper stage, where the spatial resolution is reduced, makes it computationally tractable while maximizing its impact.

3.4. Segmentation Head

To generate the final segmentation mask, we use the UperNet decoder head [39] with all our backbone experiments. UperNet effectively fuses multi-scale features from the backbone’s four stages using a combination of a Pyramid Pooling Module (PPM) [48] and a Feature Pyramid Network (FPN) [21]. This creates a feature representation that is rich in both high-level semantic context and low-level spatial detail. Using a consistent, powerful decoder allows us to focus our analysis on the performance of the encoder backbones.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our model on two public benchmarks:

- **DeepGlobe Road Extraction Dataset** [10]: Contains 6,226 satellite images (1024×1024) with a Ground Sampling Distance (GSD) of 0.5 m, covering diverse geographic regions.
- **Massachusetts Roads Dataset** [27]: Comprises 1,171 aerial images (1500×1500) with a GSD of 1 m, covering urban and suburban areas. A key challenge is its annotation style, where all roads are rasterized to a fixed 7-pixel width.

Evaluation Metrics. We use standard pixel-wise metrics for the road class: Intersection over Union (IoU) and F1-Score. To specifically assess the topological integrity of the road network, we use the Average Path Length Similarity (APLS) metric [13]. To compute this metric, the binary segmentation masks (both predicted and ground-truth) are first converted into graph structures via morphological skeletonization. APLS compares shortest path lengths in these graphs, yielding a score from 0 (dissimilar) to 1 (identical). It is defined as:

$$\text{APLS} = 1 - \frac{1}{N} \sum \min \left\{ 1, \frac{|L(a, b) - L(a', b')|}{L(a, b)} \right\}$$

where N is the number of paths in the ground truth graph, $L(a, b)$ is a path length in the ground truth, and $L(a', b')$ is the corresponding path length in the prediction.

Implementation Details. Our model and all baselines are implemented using the MMSegmentation v1.2.2 framework. For fair comparison, all models benchmarked against the state-of-the-art are initialized with weights pretrained on ImageNet-1K and subsequently on ADE20K. The ADE20K benchmark was run following its standard training protocol. All ablation studies were trained from scratch on the DeepGlobe dataset. We use the AdamW optimizer with a learning rate of 6×10^{-5} and a weight decay of 0.01. The loss function is a combination of binary cross-entropy and Dice loss, a hybrid approach effective for balancing pixel-wise supervision with robustness to class imbalance [41]. The learning rate follows a linear warm-up for 1,500 iterations, followed by a polynomial decay schedule. Models were trained for 160k iterations on DeepGlobe and 80k iterations on the smaller Massachusetts Roads dataset to prevent overfitting. Experiments were conducted on NVIDIA A100 GPUs.

Baselines. We compare our model against similarly-sized, state-of-the-art architectures. Our selected baselines include: a strong CNN in DeepLabv3+ [8] with a ResNet-101 backbone; leading Transformer models like SegFormer [40] with an MIT-B3 backbone and Swin-UperNet [23] (Swin-T backbone); and recent Mamba-based architectures such as VMamba-T [22] and MambaVision-T [16], both using the UperNet decoder. This selection provides a comprehensive comparison across different architectural paradigms.

4.2. Generalization Benchmarks

To first establish the general feature extraction capability of our backbone, we benchmark it on the ImageNet-1K classification task [11]. As shown in Table 1, PathMamba achieves a Top-1 accuracy of **83.1%**, outperforming other leading architectures of a similar scale. This demonstrates the fundamental strength of our hybrid design.

Table 1. Performance on the **ImageNet-1K** validation set. Best in **bold**, second in underline.

Method	Backbone	Params (M)	Top-1 Acc. (%)
Swin [23]	Swin-T	29	81.3
VMamba [22]	VMamba-T	30	82.6
MambaVision [16]	MambaVision-T	35	<u>82.7</u>
Ours	PathMamba	31	83.1

We further validate its versatility for dense prediction tasks on the challenging ADE20K dataset [51]. As shown in Table 2, our model achieves a new state-of-the-art mIoU of **48.6%**. This result surpasses leading Mamba and Transformer-based architectures, demonstrating that our hybrid design is a powerful and versatile feature extractor effective for general-purpose segmentation beyond remote sensing.

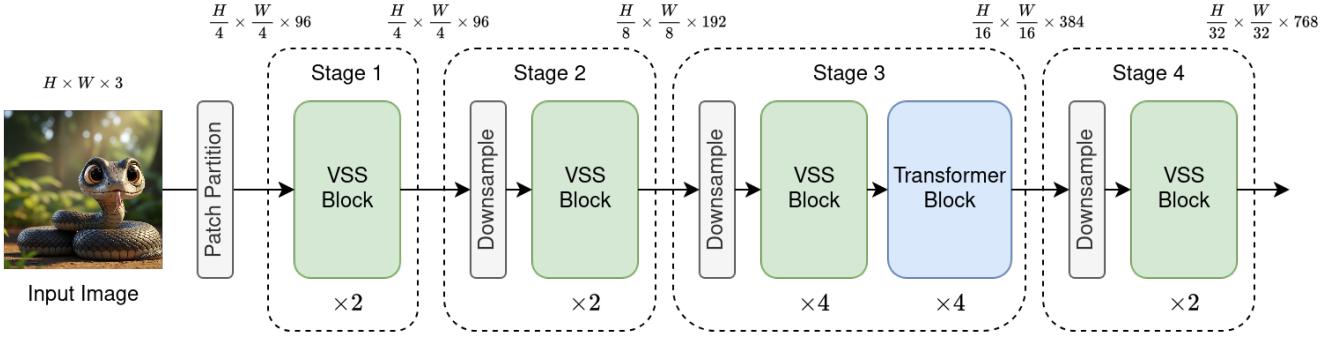


Figure 3. The architecture of our proposed hybrid Mamba-Transformer backbone. Stages 1, 2, and 4 consist of VSS (Mamba) blocks. Stage 3 is a hybrid stage containing a sequence of VSS blocks followed by standard Transformer (Attention) blocks to integrate continuity modeling with global context aggregation. All stages are connected to a UperNet decoder head.

Table 2. Performance on the **ADE20K** validation set (single-scale mIoU). Our model is compared against other architectures of a similar scale. Best in **bold**, second in underline.

Method	Backbone	Params (M)	FLOPs (G)	mIoU (%)
Swin-UperNet [23]	Swin-T	59.0	940	44.4
DeepLabV3+ [8]	R-101-D8	60.1	1017	45.5
MambaVision [16]	MambaVision-T	61.7	933	46.0
SegFormer [40]	MIT-B3	44.6	239	47.8
VMamba [22]	VMamba-T	60.9	942	<u>47.9</u>
Ours	PathMamba	61.7	946	48.6

Table 3. Results on the **Massachusetts Roads** test set. Our model achieves the best topological score (APLS) while remaining competitive on pixel-wise metrics and inference speed. Best in **bold**, second in underline.

Method	IoU (%)	F1 (%)	APLS	Params (M)	FLOPs (G)	FPS
SegFormer [40]	66.94	80.20	78.40	44.6	796	6.17
MambaVision [16]	67.31	80.46	78.00	61.7	2188	9.67
DeepLabV3+ [8]	67.51	80.61	78.48	60.2	2288	10.2
Vanilla-VMamba [22]	67.76	80.78	78.32	53.5	2058	5.22
Swin-UperNet [23]	68.20	81.10	79.53	58.9	2098	11.08
VMamba [22]	68.26	<u>81.14</u>	<u>79.66</u>	61.8	2111	13.8
Ours	68.20	81.09	79.88	61.8	2119	<u>9.22</u>

Table 4. Results on the **DeepGlobe Road Extraction** test set. Our hybrid model sets a new state-of-the-art across all metrics.

Method	IoU (%)	F1 (%)	APLS (%)	Params (M)	FLOPs (G)	FPS
Swin-UperNet [23]	70.29	82.55	75.74	59.0	940	22.8
DeepLabV3+ [8]	70.35	82.59	75.38	60.1	1017	23.2
MambaVision [16]	70.55	82.73	77.11	61.7	933	25.2
Vanilla-VMamba [22]	71.67	83.50	79.72	53.5	919	15.0
SegFormer [40]	71.94	83.68	78.82	44.6	239	21.2
VMamba [22]	<u>72.08</u>	<u>83.78</u>	<u>79.42</u>	60.9	942	26.8
Ours	72.19	83.85	80.03	61.7	946	<u>22.9</u>

4.3. Comparison with State-of-the-Art

As shown in Table 3 and Table 4, our hybrid architecture achieves outstanding performance.

On **Massachusetts Roads**, our model obtains the highest APLS score of **79.88%**, demonstrating its superior ability to generate topologically correct road networks. This is particularly significant given the dataset's fixed-width annotation artifact, which makes APLS a more reliable metric than IoU or F1-score for evaluating functional quality.

On **DeepGlobe Road Extraction**, our model establishes a new state-of-the-art, outperforming all baselines across all metrics. It achieves an IoU of **72.19%**, an F1-score of **83.85%**, and a top APLS score of **80.03%**. This consistent superiority confirms our hypothesis that combining Mamba's continuity modeling with the Transformer's global reasoning leads to more accurate and complete segmentations, all while maintaining a comparable computational footprint to other leading models.

4.4. Ablation Studies

We conducted extensive ablation studies on the DeepGlobe dataset to validate our design choices. For computational efficiency and a clear assessment of architectural changes, all models in this section were trained from scratch. As expected, this protocol results in performance metrics that are lower than their fully pretrained counterparts (e.g., our model achieves 69.10% IoU from scratch vs. 72.19% with pretraining). However, the relative performance differences between configurations provide a valid basis for our conclusions.

Impact of Hybrid Stage Arrangement. Our investigation into the arrangement of Mamba ('m') and attention ('a') blocks, detailed in Table 5, reveals a nuanced trade-off between pixel-wise accuracy (IoU/F1) and topological continuity (APLS). No single configuration achieved superiority across all metrics. Instead, our proposed 'mmmm-aaaa' design and the alternating 'ma-ma-ma-ma' configuration emerge as the two top-performing models, forming a Pareto front.

Specifically, our 'mmmm-aaaa' architecture achieves the best performance on the primary segmentation metrics, with

an IoU of **69.10%** and an F1-score of **81.72%**. The ‘mma-ma-ma’ configuration, while performing slightly worse on these metrics, secures a marginal lead in its APLS score. Given that the gains our model achieves in IoU and F1-score are more significant than its minor deficit in APLS, our proposed architecture offers the most compelling and well-balanced performance profile. We therefore select the ‘mmmm-aaaa’ configuration as our final design.

Table 5. Ablation on the arrangement of Mamba (‘m’) and attention (‘a’) blocks in the hybrid stage. Our proposed ‘mmmm-aaaa’ configuration achieves the best performance on pixel-wise metrics (IoU/F1).

Hybrid Stage Config.	IoU (%)	F1 (%)	APLS (%)
mmmm-aaaa (Ours)	69.10	81.72	75.33
ma-ma-ma-ma	68.84	81.54	75.44
am-am-am-am	68.88	81.57	75.02
aaaa-mm-mm	68.85	81.55	74.64
mmmmmmmm-a	68.64	81.40	74.79

Necessity of the SSM Component. To verify the importance of Mamba’s core SSM mechanism, especially in light of recent work questioning its necessity for all vision tasks [43], we replaced the SSM component in the VSS blocks with a simple identity mapping. As shown in Table 6, removing the SSM from any part of the network leads to a significant drop in performance. Removing it entirely results in a catastrophic performance collapse, with the APLS score dropping by over 7 points. This definitively demonstrates that the selective state space mechanism is critical for achieving high-quality topological results in this domain.

Table 6. Ablation on the removal of the SSM component from different stages. The SSM is critical for high performance.

SSM Removed from Stages	IoU (%)	F1 (%)	APLS (%)
- (Baseline)	69.10	81.72	75.33
Stage 1, 2	68.72	81.46	74.63
All Stages	66.98	80.22	68.34

Impact of Scan Strategy. To apply the sequential SSM to 2D images, a scanning strategy is required to serialize the image patches. Our architecture adopts the cross-scan method from VMamba [22] as its default. To validate this choice, we compared it against a range of alternative scanning patterns from recent literature, including uni/bi-directional scans [22], vertical-2D scan [45], omni-directional scan [50], local scan [18], and fractal scan [33]. As shown in Table 7, while cross-scan achieves the best overall performance, the differences between the top methods are minor. This suggests that our model is largely robust

to the specific choice of scanning path, a finding that aligns with other recent studies questioning the criticality of complex scan paths for vision tasks [54]. This indicates that the core hierarchical and hybrid nature of our architecture is the primary driver of its strong performance.

Table 7. Ablation on different 2D scan strategies. The Cross-Scan method used in our model provides the best overall performance.

Scan Strategy	IoU (%)	F1 (%)	APLS (%)
Cross-Scan (Ours) [22]	69.10	81.72	75.33
Vertical-2D [45]	68.94	81.61	75.18
Local-Scan [18]	68.71	81.45	74.69
Omni-Scan [50]	68.77	81.50	74.66
Bi-directional [22]	69.00	81.66	74.62
Uni-directional [22]	68.79	81.51	73.94
Fractal-Scan [33]	68.41	81.24	73.78

On the Importance of Pretraining for Topology. The performance gap between our models trained from scratch and their pretrained counterparts is most pronounced on the APLS metric. This aligns with recent findings that long-sequence models benefit immensely from data-driven priors acquired during large-scale pretraining [2]. In the context of road segmentation, these priors provide a rich, general understanding of common visual concepts like trees and buildings. A model trained from scratch may learn to associate roads with simple textural features, causing it to terminate a segment when faced with an occlusion. In contrast, a pretrained model can leverage its world knowledge to recognize the occluding object and infer the road’s continuous path through the interruption. This ability is paramount for maintaining the topological integrity measured by APLS.

4.5. Validating the Architectural Hypothesis

To validate our foundational hypothesis regarding the complementary strengths of Mamba and Transformers, we compared different backbone configurations at the stage level. We benchmarked a pure Mamba backbone against hybrids where entire Mamba stages were replaced by Transformer stages. The results, summarized in Table 8, reveal a distinct trade-off between pixel-level and topology-level performance.

The results provide strong quantitative support for our hypothesis. The pure Mamba backbone ($m \rightarrow m \rightarrow m \rightarrow m$) achieves the highest APLS score, confirming its proficiency in modeling the continuous nature of road networks. Conversely, replacing later Mamba stages with Transformer stages ($m \rightarrow m \rightarrow a \rightarrow a$) boosts the IoU score, demonstrating the Transformer’s superior capability for global context aggregation, which aids in overall pixel classification. However, this gain in pixel-wise accuracy comes at the cost of

Table 8. Analysis of architectural contributions at the stage level on the DeepGlobe dataset. Each letter represents a full stage in the four-stage backbone, composed entirely of either Mamba ('m') or Transformer ('a') blocks. Best values are in **bold**.

Backbone Stage Configuration	IoU (%)	F1 (%)	APLS (%)
m → m → m → m (All-Mamba)	68.25	81.13	74.17
m → m → m → a	67.88	80.87	73.94
m → m → a → m	68.90	81.58	73.68
m → m → a → a (Late-Attention)	69.16	81.77	71.90

a significant drop in topological continuity. This trade-off motivates our final architecture, which integrates these components within a single stage to achieve a more optimal balance.

4.6. Qualitative Analysis

To visually substantiate our quantitative results, we present a qualitative analysis in Figure 4 and Figure 5. These figures illustrate our model’s superior ability to preserve road network integrity in challenging real-world scenarios.

Figure 4 showcases several examples where our hybrid model excels compared to strong baselines like SegFormer and VMamba. Across different scenes, our model demonstrates a clear advantage in generating complete and continuous road masks, successfully avoiding the fragmentation that plagues the other methods. This is particularly evident in cases with complex intersections or where fine-grained road details are present.

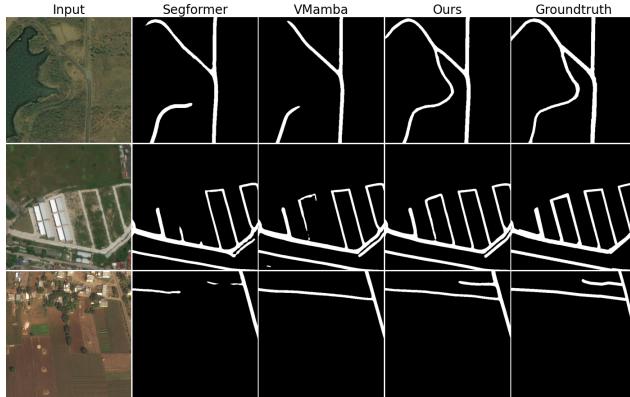


Figure 4. Qualitative comparison of our model against Segformer and VMamba. In the second row, both baselines produce disconnected masks, whereas our model correctly identifies the continuous road network. In the third row, our model accurately segments fine-grained details under tree occlusion, which are missed by the other methods.

Figure 5 focuses on a difficult case with a rural road of inconsistent texture, which often causes standard models to lose track of the path. While both SegFormer and VMamba produce fractured predictions, our model suc-

cessfully maintains a continuous path. This visual success is directly quantified by the accompanying APLS scores, where our model achieves a dramatic improvement over the baselines, reinforcing the conclusion that our hybrid design leads to more reliable road segmentation.

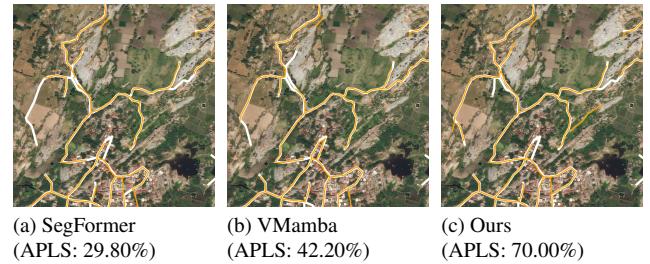


Figure 5. A challenging sample featuring a rural road with inconsistent texture. Our model (c) maintains connectivity where baselines (a, b) fail due to surface variability, resulting in a significantly higher APLS score. (White = Ground Truth, Orange = Prediction).

5. Conclusion

In this paper, we introduced PathMamba, a novel hybrid Mamba-Transformer architecture for road segmentation from satellite imagery. By strategically combining the linear-time efficiency of Mamba for modeling continuous features with the powerful global reasoning of Transformers, our model leverages their combined strengths to yield superior performance.

Our experiments demonstrate that this hybrid design sets a new state-of-the-art on the DeepGlobe benchmark and, most importantly, achieves the highest topological accuracy (APLS score) on both the DeepGlobe Road Extraction and Massachusetts Roads datasets. This confirms that our architecture produces more coherent and functionally useful road networks. Comprehensive ablation studies validated our specific design choice of using Mamba blocks to first model continuity before refining features with self-attention.

Future work could explore scaling the architecture, investigating more advanced Mamba modules, and focusing on deployment-oriented optimizations like quantization to further enhance its suitability for real-world, resource-constrained applications in remote sensing.

References

- [1] Amin Abdollahi, Biswajeet Pradhan, Nagesh Shukla, Subrata Chakraborty, and Ahmed Alamri. A review of deep learning-based road extraction from remote sensing images. *Remote Sensing*, 13(23):4985, 2021. 3
- [2] Ido Amos, Jonathan Berant, and Ankit Gupta. Never train from scratch: Fair comparison of long-sequence models requires data-driven priors. *arXiv preprint arXiv:2310.02980*, 2024. 2, 7

- [3] Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, and Vishal M. Patel. SPIN Road Mapper: Extracting Roads from Aerial Images via Spatial and Interaction Space Graph Reasoning for Autonomous Driving. *arXiv preprint arXiv:2109.07701*, 2021. 2, 3
- [4] M. Barzohar and D.B. Cooper. Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation. *IEEE TPAMI*, 18(7):707–721, 1996. 2
- [5] Fayyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. RoadTracer: Automatic Extraction of Road Networks from Aerial Images. *arXiv preprint arXiv:1802.03680*, 2018. 2, 3
- [6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 1, 3
- [7] Dengfeng Chai, Wolfgang Forstner, and Florent Lafarge. Recovering line-networks in images by junction-point processes. In *CVPR*, 2013. 2
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv preprint arXiv:1606.00915*, 2017. 2, 4, 5, 6
- [9] Guangliang Cheng, Ying Wang, Shibiao Xu, Hongzhen Wang, Shiming Xiang, and Chunhong Pan. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3322–3337, 2017. 1
- [10] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *CVPRW*, 2018. 1, 5
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2021. 1, 3
- [13] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv preprint arXiv:1807.01232*, 2019. 1, 2, 3, 5
- [14] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*, 2024. 2, 3, 4
- [15] Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces. *arXiv preprint arXiv:2111.00396*, 2022. 2, 3, 4
- [16] Ali Hatamizadeh and Jan Kautz. MambaVision: A Hybrid Mamba-Transformer Vision Backbone. *arXiv preprint arXiv:2407.08083*, 2025. 2, 3, 5, 6
- [17] Hao He, Danning Yang, Shaohua Wang, Yan Wang, and Chong Wang. Road extraction convolutional neural network with embedded attention mechanism for remote sensing imagery. *Remote Sensing*, 14(9):2036, 2022. 3
- [18] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. LocalMamba: Visual State Space Model with Windowed Selective Scan. *arXiv preprint arXiv:2403.09338*, 2024. 7
- [19] Xiaoling Jiang, Yinyin Li, Tao Jiang, Junhao Xie, Yilong Wu, Qianfeng Cai, Jinhui Jiang, Jiaming Xu, and Hui Zhang. RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 113:102987, 2022. 3
- [20] I. Laptev, H. Mayer, T. Lindeberg, W. Eckstein, C. Steger, and A. Baumgartner. Automatic extraction of roads from aerial images based on scale space and snakes. *Machine Vision and Applications*, 12(1):23–31, 2000. 2
- [21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. *arXiv preprint arXiv:1612.03144*, 2017. 4, 5
- [22] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166*, 2024. 2, 3, 4, 5, 6, 7
- [23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv preprint arXiv:2103.14030*, 2021. 3, 4, 5, 6
- [24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1
- [25] Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *ICCV*, pages 3438–3446, 2017. 2
- [26] Volodymyr Mnih and Geoffrey E. Hinton. Learning to Detect Roads in High-Resolution Aerial Images. In *ECCV*, pages 210–223, 2010. 1
- [27] Volodymyr Mnih and Geoffrey E. Hinton. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013. 1, 5
- [28] Agata Mosinska, Pablo Márquez-Neila, Mateusz Koziński, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *CVPR*, 2018. 1, 2, 3
- [29] Vinod Nandakumar, Arush Tagade, and Tongliang Liu. Why do cnns excel at feature extraction? a mathematical explanation. *arXiv preprint arXiv:2307.00919*, 2023. 1
- [30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv preprint arXiv:1505.04597*, 2015. 1, 2
- [31] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *arXiv preprint arXiv:1605.06211*, 2016. 1
- [32] Radu Stoica, Xavier Descombes, and Josiane Zerubia. A Gibbs Point Process for Road Extraction from Remotely Sensed Images. *International Journal of Computer Vision*, 57(2):121–136, 2004. 2

- [33] Lv Tang, HaoKe Xiao, Peng-Tao Jiang, Hao Zhang, Jinwei Chen, and Bo Li. Scalable Visual State Space Model with Fractal Scanning. *arXiv preprint arXiv:2405.14480*, 2024. 7
- [34] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long Range Arena: A Benchmark for Efficient Transformers. *arXiv preprint arXiv:2011.04006*, 2020. 2, 3, 4
- [35] Cem Unsalan and Beril Sirmacek. Road Network Detection Using Probabilistic and Graph Theoretical Methods. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4441–4453, 2012. 2
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 1, 2, 3
- [37] Simon Wachter, Armin Rieger, and Andreas Zell. Road segmentation of aerial images using residual-attention-duck-net. In *2023 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6, 2023. 3
- [38] Wei Wang, Na Yang, Yifei Zhang, Fuan Wang, Tian Cao, and Peter Eklund. A review of road extraction from remote sensing images. *Journal of Traffic and Transportation Engineering (English Edition)*, 3(3):271–282, 2016. 2
- [39] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. *arXiv preprint arXiv:1807.10221*, 2018. 5
- [40] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv preprint arXiv:2105.15203*, 2021. 1, 3, 5, 6
- [41] Hongzhang Xu, Hongjie He, Ying Zhang, Lingfei Ma, and Jonathan Li. A comparative study of loss functions for road segmentation in remotely sensed road datasets. *International Journal of Applied Earth Observation and Geoinformation*, 116:103159, 2023. 5
- [42] Abolfazl Younesi, Mohsen Ansari, MohammadAmin Fazli, Alireza Ejlali, Muhammad Shafique, and Jörg Henkel. A comprehensive survey of convolutions in deep learning: Applications, challenges, and future trends. *IEEE Access*, 2024. 1
- [43] Weihao Yu and Xinchao Wang. MambaOut: Do We Really Need Mamba for Vision? *arXiv preprint arXiv:2405.07992*, 2024. 7
- [44] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. MetaFormer Is Actually What You Need for Vision. *arXiv preprint arXiv:2111.11418*, 2022. 4
- [45] Jingwei Zhang, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S. Hosseini. 2DMamba: Efficient State Space Model for Image Representation with Applications on Giga-Pixel Whole Slide Image Classification. *arXiv preprint arXiv:2412.00678*, 2025. 7
- [46] Xun Zhang, Weipeng Ou, Xiaokun Wu, and Cheng Zhang. Mhs-vit: Mamba hybrid self-attention vision transformers for traffic image detection. *Plos one*, 20(6):e0325962, 2025. 2
- [47] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 1
- [48] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. *arXiv preprint arXiv:1612.01105*, 2017. 4, 5
- [49] Ling Zhao, Jianing Zhang, Xiujun Meng, Wenming Zhou, Zhenshi Zhang, and Chengli Peng. Road Extraction Method of Remote Sensing Image Based on Deformable Attention Transformer. *Symmetry*, 16(4):468, 2024. 3
- [50] Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. RS-Mamba for Large Remote Sensing Image Dense Prediction. *arXiv preprint arXiv:2404.02668*, 2024. 3, 7
- [51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing Through ADE20K Dataset. In *CVPR*, pages 633–641, 2017. 5
- [52] Lichen Zhou, Chuang Zhang, and Ming Wu. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In *CVPRW*, pages 192–1924, 2018. 2
- [53] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv preprint arXiv:2401.09417*, 2024. 2, 3
- [54] Qinfeng Zhu, Yuan Fang, Yuanzhi Cai, Cheng Chen, and Lei Fan. Rethinking Scanning Strategies With Vision Mamba in Semantic Segmentation of Remote Sensing Imagery: An Experimental Study. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:18223–18234, 2024. 7
- [55] Siyuan Zuo, Yuxiang Xiao, Xiaojie Chang, and Xing Wang. Vision transformers for dense prediction: A survey. *Knowledge-Based Systems*, 253:109552, 2022. 3

PathMamba: A Hybrid Mamba-Transformer for Topologically Coherent Road Segmentation in Satellite Imagery

Supplementary Material

6. Reproducibility Details

This supplementary section provides additional details regarding our experimental setup to ensure the reproducibility of our results. We cover dataset splits, data preprocessing and augmentation, and the specific hyperparameter configuration for our main model. Our implementation is built upon the MMSegmentation v1.2.2 framework. We intend to release our code and pretrained models to facilitate further research.

6.1. Datasets and Preprocessing

We used standard, publicly available splits for all datasets to ensure fair and consistent comparison.

DeepGlobe Road Extraction. We followed the official split from the CVPR 2018 challenge. For data augmentation, we applied random horizontal and vertical flips during training. Input images are normalized using a mean of ‘[123.675, 116.28, 103.53]’ and a standard deviation of ‘[58.395, 57.12, 57.375]’.

Massachusetts Roads. We used the standard split defined in the original paper [27]. Input images were resized to 1532×1532 pixels. The only data augmentation used was random horizontal and vertical flips.

ADE20K. We used the standard dataset split and followed the classical preprocessing pipeline for this benchmark, which includes resizing, random cropping to 512×512 , random horizontal flipping, and normalization.

6.2. Hyperparameter Configuration

For full transparency, we provide a detailed summary of the key hyperparameters used for our main PathMamba model on the DeepGlobe dataset.

Backbone (PathMamba). The encoder architecture is configured as follows:

- **Architecture:** A four-stage hierarchical backbone with depths of ‘(2, 2, 8, 2)’ and an initial embedding dimension of ‘96’.
- **Stage Configuration:** Stages 1, 2, and 4 consist of Mamba (VSS) blocks. Stage 3 is our proposed hybrid stage with a sequential ‘mmmm-aaaa’ configuration (4 Mamba blocks followed by 4 Transformer blocks).
- **Regularization:** A drop path rate of ‘0.2’ is applied.

- **Patch Embedding:** A patch size of ‘4x4’ is used.

Decoder and Loss Function.

- **Decoder:** We use the UperNet (‘UPerHead’) decoder, which fuses features from all four backbone stages. It employs pyramid pooling with scales of ‘(1, 2, 3, 6)’.
- **Loss Function:** The total loss is a sum of two components, each with a weight of 1.0: a pixel-wise Focal Loss and a region-based Dice Loss. This hybrid loss is applied to both the main and auxiliary heads.

Training and Optimization.

- **Optimizer:** We use the AdamW optimizer with a learning rate of 6×10^{-5} , betas of ‘(0.9, 0.999)’, and a weight decay of ‘0.01’.
- **Learning Rate Schedule:** A linear warm-up schedule is used for the first 1,500 iterations, increasing the learning rate from 10^{-6} to the base learning rate. This is followed by a polynomial decay schedule for the remainder of the training.
- **Training Duration:** All models on the DeepGlobe dataset are trained for a total of 160,000 iterations.
- **Batch Size:** A batch size of 4 per GPU is used.

6.3. Experimental Protocol

Training Runs and Variation. Due to the significant computational cost associated with training large-scale vision models, all experimental results reported in the main paper are from a single, complete training run for each model configuration. This is a common practice in the field for experiments of this scale. As a result, measures of variation such as standard deviation or error bars are not applicable, as they would require multiple training runs which were computationally prohibitive.