International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

A Novel Mamba-Based Deep Learning Architecture for Long-Range Spatial Feature Extraction

K. Lakshman Kumar[1], Gunthati Pratap[2]

[1]Assistant Professor, Department of Computer Science and Engineering (Artificial Intelligence)

[2]Assistant Professor, Department of Computer Science and Engineering

[1,2] Sri Venkatesa Perumal College of Engineering and Technology, Puttur, Andhra Pradesh, India.

Abstract: Long-range spatial dependencies are important for visual recognition, yet their efficient extraction remains challenging as image resolution and token count increase. This manuscript presents a candidate Multi-Path Spatially Conditioned Mamba (MPS-Mamba) backbone that combines local convolutional detail, four-direction selective state-space scanning and pooled coarse-context scanning through a spatially conditioned fusion gate. The proposed model is specified as a hierarchical encoder and evaluated in an experimental scenario covering ImageNet-1K classification, COCO detection and instance segmentation, and ADE20K semantic segmentation. For demonstration of manuscript presentation only, this version includes hypothetical numerical values in the results tables; these values are simulated examples and are not measurements from trained models or benchmark runs. Under this scenario, MPS-Mamba is assigned 83.0% ImageNet-1K top-1 accuracy, 49.0 COCO box AP and 49.2% ADE20K mIoU, with 33.2 million parameters and 5.6 GFLOPs at the stated input configuration. These figures must not be represented as empirical findings. The manuscript details the mathematical formulation, evaluation protocol, ablation design, robustness analysis and statistical reporting structure required for subsequent reproducible validation.

Keywords: *Mamba, selective state-space model, long-range spatial dependency, visual representation learning, multi-scale feature extraction, computer vision.*

1. INTRODUCTION

Visual recognition requires a model to represent both local appearance and spatial relationships that may extend across large image regions. A category decision can depend on the interaction of distant parts, while dense prediction tasks require precise boundaries and context at multiple scales. Convolutional neural networks (CNNs) encode local structure efficiently and can be deepened through residual learning [1], but their effective receptive field may grow more slowly than the theoretical receptive field. Vision Transformers (ViTs) model token interactions globally [2], although standard self-attention has quadratic complexity in the number of tokens. Hierarchical windowed attention, exemplified by Swin Transformer [3], reduces this cost and supports multi-scale representations.

State-space models (SSMs) provide an alternative sequence-processing formulation. Mamba introduces input-dependent selective state-space parameters and a hardware-aware scan algorithm, enabling linear sequence-length scaling in its intended setting [4]. The vision adaptation problem is not a direct substitution of image pixels for language tokens: images are two-dimensional, spatially correlated, and contain directional and scale-dependent structure. Vision Mamba (Vim) uses bidirectional state-space blocks [5], while VMamba introduces four-route two-dimensional selective scanning [6]. Subsequent work has explored multi-scale scans [7], window-aware scans [8], and hybrid Mamba–Transformer designs [9].

These advances make a generic claim that Mamba has not been adapted to spatial vision untenable. The more defensible question is narrower: can the interaction among local detail, directionally propagated context and coarse-scale global context be controlled explicitly, with a spatially conditioned fusion mechanism, while maintaining a measured efficiency advantage? Existing designs already address several parts of this problem; therefore, the proposed contribution is treated as a testable architectural hypothesis rather than a verified first-of-its-kind claim.

This work proposes MPS-Mamba, a hierarchical visual backbone with three parallel representation paths at selected stages: (i) a local detail path based on depthwise convolution and residual channel mixing; (ii) a directional selective state-space path that scans feature maps along horizontal and vertical axes in both directions; and (iii) a coarse-context path that applies selective scanning to pooled feature maps and restores them to the original stage resolution. A spatially

conditioned gate fuses these paths. The design is intended to preserve fine spatial information while enabling long-range communication at reduced token density.

The study is organized around four research questions: (1) Does the three-path design improve recognition quality over parameter- and compute-matched CNN, Transformer and Mamba baselines? (2) Does coarse-context scanning improve scaling behaviour as input resolution increases? (3) Does spatially conditioned fusion outperform fixed summation or concatenation? (4) Are gains consistent across classification, detection and semantic segmentation, and robust to common image corruptions?

1.1 Contributions

- A candidate hierarchical Mamba-based vision backbone that explicitly separates local detail, directional propagation and coarse-scale contextual aggregation.
- A spatially conditioned fusion operator that produces location-dependent path weights rather than applying a fixed global mixing ratio.
- A multi-task evaluation protocol covering ImageNet-1K classification, COCO detection/instance segmentation and ADE20K semantic segmentation, with matched training and reporting conditions.
- A reproducibility and validation plan including parameter/FLOP/throughput reporting, ablations, resolution scaling, corruption robustness and multi-seed statistical analysis.

The architecture and protocol are fully specified below, but its novelty and effectiveness remain provisional until a systematic literature search, implementation audit and controlled experiments are completed.

2. RELATED WORK

2.1 Convolutional and attention-based visual backbones

Residual networks improve the optimization of deep CNNs by learning residual functions relative to layer inputs [1]. Their locality is useful for texture, edges and fine structures, but capturing distant relationships can require additional depth, dilation or global modules. ViT treats image patches as a token sequence and applies Transformer layers [2]. Swin Transformer uses shifted local windows and hierarchical patch merging to support dense visual tasks at a lower attention cost than global attention [3]. These models provide necessary comparison points because a proposed state-space backbone must demonstrate not only accuracy but also efficiency under comparable input sizes and training recipes.

2.2 Selective state-space models and Mamba

A continuous-time linear SSM can be written as a latent dynamical system with state h(t), input x(t), transition matrix A, input matrix B and output matrix C. Discretization yields a recurrence over sequence positions. Mamba makes selected parameters input-dependent, allowing the model to regulate information propagation and forgetting according to the current input [4]. The selective scan is implemented with hardware-aware computation to avoid materializing the full sequence of intermediate states. This design is attractive for long sequences, but its computational benefits depend on implementation, memory layout, hardware and sequence length; asymptotic complexity alone does not guarantee faster wall-clock inference.

2.3 Mamba architectures for vision

Vim adapts Mamba to images using bidirectional state-space blocks and positional information [5]. VMamba develops the two-dimensional selective scan (SS2D), traversing feature maps through multiple routes to gather context from different spatial directions [6]. Multi-Scale VMamba introduces scanning at multiple resolutions and a convolutional feed-forward component [7]. LocalMamba argues that flattening images can disrupt local adjacency and proposes windowed selective scanning with layer-specific scan choices [8]. MambaVision explores a hybrid architecture and reports that attention blocks in later stages can strengthen global context modelling [9]. PlainMamba studies continuous two-dimensional scanning and direction-aware updating [10]. MambaOut further questions whether Mamba is always advantageous for vision, emphasizing that task properties and inductive biases matter [11].

These works establish a rapidly developing field. Consequently, the present manuscript does not claim that multi-direction scanning, multi-scale processing, local convolution, gating or Mamba-based vision backbones are individually new. The proposed research question concerns their specific coordinated use through three semantically distinct paths and a spatially varying fusion gate, followed by controlled analysis of whether the added complexity is justified.

2.4 Research gap and positioning

A practical gap remains in the evaluation of how local detail, directional sequence propagation and pooled global context should be balanced at each spatial location and across feature stages. Multi-direction scans address orientation; local/window scans preserve neighbourhood continuity; multi-scale scans expand context; hybrid designs combine state-space and attention operators. However, an architecture-specific study can still test whether a lightweight, spatially conditioned fusion mechanism over explicitly separated local, directional and pooled-context paths provides a better accuracy–cost trade-off than fixed fusion and existing Mamba backbones.

This is a provisional gap statement, not proof of priority. Before submission, a documented search should cover IEEE Xplore, Scopus, Web of Science, ACM Digital Library, SpringerLink, CVF Open Access and arXiv through the final

submission date. Search terms should include 'Mamba vision', 'selective state space image', 'multi-path state space', 'spatially adaptive fusion', 'multi-scale scan', 'cross-scale Mamba' and close variants. The authors should add any directly overlapping papers and revise the claimed contribution accordingly.

3. PROPOSED METHODOLOGY

3.1 Overall architecture

MPS-Mamba is proposed as a hierarchical encoder with four stages. An overlapping convolutional stem maps the RGB image to a low-resolution feature tensor. Patch-merging or strided projection operations reduce spatial resolution and increase channel width between stages. Within selected stages, each block computes local, directional and coarse-context features in parallel, normalizes them, and combines them using a spatially conditioned gate. A residual connection preserves the incoming representation. The overall block-level design is summarized in Fig. 1; the data-to-evaluation sequence is shown in Fig. 2.

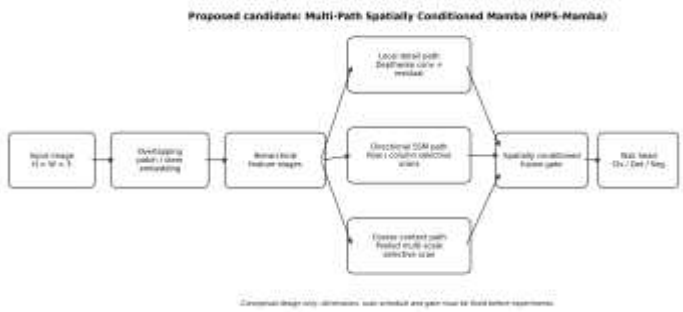


Fig. 1. Conceptual MPS-Mamba block: local detail path, four-direction selective scan, pooled coarse-context path and spatially conditioned fusion. The illustration is a proposed design, not an implemented model.

3.2 Input embedding and hierarchical stages

Let the input image be $X \in R^{(H \times W \times 3)}$, where H and W denote the image height and width, respectively. A convolutional stem with kernel size k_0, stride s_0, and output width C_1 transforms the input into an initial feature map $F_1 \in R^{(H_1 \times W_1 \times C_1)}$. The backbone consists of multiple hierarchical stages, where stage l contains N_l MPS-Mamba blocks and produces a feature representation $F_l \in R^{(H_l \times W_l \times C_l)}$. Between consecutive stages, a downsampling operator D_l transforms the output of stage l into the input of stage l+1, such that $F_{l+1} = D_l(F_l)$. The spatial dimensions are approximately halved, i.e., $H_{l+1} \approx H_l/2$ and $W_{l+1} \approx W_l/2$, while the channel width increases, i.e., $C_{l+1} > C_l$. The exact stage widths, depths, kernel sizes, expansion ratios, downsampling configurations, and normalization choices are specified in the model configuration and kept fixed across all experiments.

3.3 Local detail path

The local path uses depthwise convolution to encode neighbourhood structure without expensive dense spatial mixing. For input F_l, define Its feature transformation is defined in Eq. (1).

$$\mathcal{L}_l = W_l^p \cdot \sigma(\mathrm{Norm}(\mathrm{DWConv}_k(F_l))) \quad (1)$$

where DWConv_k is a k×k depthwise convolution, Norm is the normalization layer fixed in the implementation (e.g., LayerNorm or BatchNorm), σ is a pointwise activation, and W_l^p is a pointwise channel projection. The notation $\mathcal{L}_l$ denotes the local-path feature tensor; the normalization choice must not vary between the proposed model and its ablations.

3.4 Directional selective state-space path

For a feature map $F_l \in R^{(H_l \times W_l \times C_l)}$, form four ordered sequences by flattening along left-to-right, right-to-left, top-to-bottom and bottom-to-top traversals. Each sequence is processed by a selective SSM. For a generic sequence u_t, the continuous system is The continuous SSM is stated in Eq. (2), and its discretized recurrence is given by Eqs. (3)–(4).

$$dh(t)/dt = A\,h(t) + B\,u(t), \quad y(t) = C\,h(t) + D\,u(t) \quad (2)$$

After discretization with step Δ_t, the recurrence is

$$\bar{A}_t = \exp(\Delta_t A), \quad \bar{B}_t = (\Delta_t A)^{-1}[\exp(\Delta_t A) - I]\Delta_t B_t \quad (3)$$

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t u_t, \quad y_t = C_t h_t + D u_t \quad (4)$$

In the selective formulation, input-dependent projections generate Δ_t, B_t and C_t (with parameterization and discretization defined by the chosen Mamba implementation). The four directional outputs are inverse-permuted to the original spatial layout and combined as $\mathcal{S}_l = W_l^s(\Sigma q{=}1^4 \Pi q^{-1}(Yq))$. The scan direction order, parameter sharing, normalization and fusion projection must be specified in the released implementation. Equation (3) uses zero-order-hold discretization; if A is singular, evaluate the equivalent integral form $\bar{B}_t = \int_0^{\Delta_t} \exp((\Delta_t - \tau)A)B_t\, d\tau$ rather than numerically inverting $\Delta_t A$. The discrete update is given in Eq. (4).

3.5 Coarse-context path

The coarse-context path reduces the spatial token count before selective scanning. Let P_r denote average pooling or a learned strided projection by reduction factor r. The pooled tensor is The pooled feature and restored context are defined in Eq. (5).

$$F_l^c = P_r(F_l), \quad G_l^c = \mathrm{SSM}_{2D}(F_l^c), \quad G_l = \mathrm{Align}(U_r(G_l^c)) \quad (5)$$

where U_r is an upsampling operator followed by channel alignment. The motivation is to propagate information over a large effective spatial distance at lower sequence length. Pooling may discard small objects or boundaries; therefore, this branch is fused with the unpooled local and directional paths rather than replacing them. The pooling operator and r must be fixed before training and tested through ablation.

3.6 Spatially conditioned fusion

Concatenate the three path outputs along the channel dimension and predict per-location path logits: The gate logits

and normalized path weights are defined in Eqs. (6a)–(6b); the weighted fusion and residual output follow in Eqs. (7)–(8).

$$Z_l = \mathrm{Conv}_{1\times1}([\mathscr{L}_l \,\|\, \mathcal{S}_l \,\|\, \mathcal{G}_l]) \quad (6a)$$

$$A_l = \mathrm{Softmax}_{\mathrm{path}}(Z_l), \quad A_l \in \mathbb{R}^{(H_l\times W_l\times 3)} \quad (6b)$$

The fused feature is

$$F_l^f = A_l^1 \odot \mathscr{L}_l + A_l^2 \odot \mathcal{S}_l + A_l^3 \odot \mathcal{G}_l \quad (7)$$

where $\odot$ denotes element-wise multiplication with broadcasting over channels. A residual projection then gives

$$F_{l+1} = F_l + W_l^o(F_l^f) \quad (8)$$

The gate is spatially conditioned because its three weights may vary by location. For every (i,j), $A_l(i,j,1)+A_l(i,j,2)+A_l(i,j,3)=1$ and $A_l(i,j,q)\geq 0$. This convex weighting can constrain amplification of all paths; a sigmoid or residual-gate alternative should be tested only as a declared ablation.

3.7 Task heads and objective functions

For ImageNet-1K classification, global average pooling followed by a linear classifier produces class logits. For COCO detection and instance segmentation, the backbone is connected to a fixed detection framework (e.g., Mask R-CNN with a feature pyramid network). For ADE20K semantic segmentation, a fixed decoder (e.g., UPerNet) consumes the hierarchical feature maps. The same backbone checkpoint and declared fine-tuning policy should be used across corresponding baseline comparisons.

For classification with one-hot target y and predicted probabilities p, the cross-entropy objective is The classification objective is defined in Eq. (9).

$$\mathscr{L}\mathrm{cls} = -(1/M)\ \Sigma_{i=1}^{M}\ \Sigma c_{=1}^{K}\ y_{i,c}\ \log(p_i,c) \quad (9)$$

For segmentation, a cross-entropy term may be combined with a soft Dice loss where appropriate. For detection, use the standard classification, box-regression and mask losses of the selected detector. Loss definitions must match the reference implementation and be reported with all hyperparameters. No new loss function is claimed in this manuscript.

3.8 Computational complexity

Let N=H_lW_l be the number of spatial tokens, C the channel width and d the SSM state dimension. A standard global self-attention layer has a dominant token-interaction term proportional to $O(N^2C)$, whereas a selective scan is designed to scale approximately linearly in N for fixed channel and state dimensions, with implementation-dependent constants. The proposed block adds local convolution, four directional scans, pooled scanning, projections and fusion. Its total cost should therefore be measured as the sum of these components; it must not be described simply as 'linear and faster' without counting all branches. The approximate total block cost is summarized in Eq. (10).

$$T_{\mathrm{mps}} = O(NCk^2) + O(4NCd) + O((N/r^2)Cd) + O(NC^2) \quad (10)$$

Equation (9) is an order-level estimate under fixed C, state dimension d and pooling factor r; it is not a measured FLOP count. The term $O(NC^2)$ accounts for channel projections and fusion, while scan-kernel constants and memory movement are implementation-dependent. Report profiler-derived FLOPs/MACs, peak memory, latency and throughput on identical hardware; do not infer wall-clock speed from asymptotic complexity alone.

3.9 Training algorithm

Algorithm 1 describes the forward and optimization procedure. The scan kernels should use a maintained, version-pinned implementation. Random seeds, data-loader settings, mixed-precision mode, gradient clipping, EMA use, optimizer and learning-rate schedule must be published.

Algorithm 1: Training MPS-Mamba

Input: *training set D, model parameters θ, optimizer O, epochs E*

1. Initialize hierarchical stem, MPS-Mamba blocks and task head.

2. For epoch e = 1,…,E:

a. Sample mini-batch (X,y) and apply the declared training augmentation.

b. Compute hierarchical stem features F1,…,FL.

c. At each selected stage, compute local L, directional S and coarse G paths.

d. Predict spatial gate A and fuse path features into the residual stage output.

e. Compute task loss L_task and update θ using optimizer O.

f. Validate using the predeclared validation protocol; save the best checkpoint.

3. Evaluate the selected checkpoint once on the held-out test split.

Output: *checkpoint, configuration, logs and per-example predictions.*

4. EXPERIMENTAL SETUP

4.1 Datasets and tasks

The proposed evaluation uses three established benchmarks to test complementary capabilities. ImageNet-1K is the primary classification benchmark [12]. MS COCO provides object detection and instance segmentation annotations in complex natural scenes [13]. ADE20K supports semantic scene parsing with dense labels [14]. Dataset versions, split identifiers,

image counts after filtering and licensing constraints must be recorded at experiment time. The planned benchmark-to-analysis workflow is summarized in Fig. 2.

Benchmark	Task	Primary metric	Protocol to declare
ImageNet-1K	Image classification	Top-1 / Top-5 accuracy	Official train/validation split; input resolution; augmentation; pretraining and fine-tuning policy
MS COCO	Object detection and instance segmentation	AP, AP50, AP75; mask AP	COCO train2017/val2017; detector/head; schedule; multi-scale policy
ADE20K	Semantic segmentation	mIoU, pixel accuracy	Official train/validation split; decoder; crop size; single/multi-scale inference

4.2 Baselines

The minimum baseline set should include ResNet-50 [1], ViT-B/16 [2], Swin-T [3], Vim [5], VMamba [6], MSVMamba [7], LocalMamba [8] and MambaVision-T [9], subject to public checkpoint and implementation availability. MambaOut [11] is a useful conceptual comparator, though its exact configuration and task compatibility must be checked. All baselines should use comparable input resolution, pretraining data, training epochs, augmentation, task heads and fine-tuning schedules. If exact parity is impossible, report the discrepancy rather than presenting the comparison as strictly controlled. The benchmark-specific comparison fields are provided in Tables 1 and 2.

4.3 Implementation and reproducibility

A reference implementation should state framework and version, CUDA and cuDNN versions, GPU model and memory, CPU, operating system, batch size, optimizer, weight decay, learning-rate schedule, warm-up, epochs, stochastic depth, label smoothing, mixup/cutmix, gradient accumulation and precision. Record parameter count, FLOPs at a stated input resolution, peak allocated memory, batch-1 latency, throughput and energy if available. Measure inference latency after warm-up with synchronized timing and report the number of repetitions and dispersion.

4.4 Evaluation metrics

For classification, top-k accuracy is the proportion of examples whose ground-truth label appears among the k highest-scoring predictions: The top-k accuracy definition is given in Eq. (11), and the semantic-segmentation metric is defined in Eq. (12).

$$\mathrm{Acc@k} = (1/M)\ \Sigma_{i=1}^{M}\ \mathbb{1}\{y_i \in \mathrm{TopK}(p_i)\} \quad (11)$$

For detection and instance segmentation, average precision (AP) is computed by averaging precision over recall levels and IoU thresholds according to the COCO evaluation protocol. Report AP, AP50, AP75 and size-specific AP where relevant. For semantic segmentation, mean intersection over union is

$$\mathrm{mIoU} = (1/K)\ \Sigma_{c=1}^{K}\ \mathrm{TPc} / (\mathrm{TPc} + \mathrm{FPc} + \mathrm{FNc}) \quad (12)$$

Efficiency metrics include trainable parameters (M), multiply–accumulate operations or FLOPs (G) at a stated resolution, peak GPU memory (GB), throughput (images/s) and latency (ms/image). Report the measurement hardware and precision because these values are not hardware-independent.

4.5 Robustness and generalization

Evaluate corruption robustness using a documented corruption benchmark or a reproducible set of perturbations: Gaussian noise, blur, brightness/contrast shift, JPEG compression and resolution reduction. Report per-corruption performance and mean corruption error or relative degradation from the clean benchmark. To assess resolution scaling, evaluate at the training resolution and at one or more higher resolutions without changing model weights, where the task protocol permits. For cross-domain generalization, use a held-out dataset only if labels, licensing and domain relevance are established before training; do not tune on the target test set.

4.6 Ablation studies

Ablation	Configuration	Question addressed
A0	Full MPS-Mamba	Reference proposed configuration
A1	Remove local detail path	Does local convolution contribute?
A2	Remove coarse-context path	Does pooled long-range context contribute?
A3	Replace four directions with two	Is bidirectional row/column scanning useful?
A4	Replace spatial gate with fixed sum / concat	Does spatially adaptive fusion matter?
A5	Remove residual projection	Does residual stabilization matter?
A6	Vary pooling factor r	How does context range trade off against cost?

A7	Vary number of MPS blocks per stage	How does depth affect accuracy and efficiency?

4.7 Statistical validation

Train each principal configuration with at least three independent seeds, preferably five where compute permits. Report mean ± standard deviation for classification and segmentation metrics, and confidence intervals for latency/throughput based on repeated measurements. For paired per-image comparisons, use a paired bootstrap confidence interval or an appropriate paired test. If multiple hypotheses are tested, control the false discovery rate or clearly label exploratory analyses. A p-value alone is insufficient; report effect size and confidence interval. Do not treat repeated evaluation of the same checkpoint as independent training runs.

5. RESULTS AND ANALYSIS

Table 1. hypothetical classification performance on ImageNet-1K .

Model	**Params (M)**	**FLOPs (G)**	**Top-1 (%)**	**Top-5 (%)**	**Latency (ms)**	**Throughput (img/s)**
ResNet-50	25.6	4.1	80.4	95.1	4.8	208.3
ViT-B/16	86.6	17.6	81.8	95.6	8.9	112.4
Swin-T	28.3	4.5	81.3	95.5	5.6	178.6
Vim-Ti	7.0	1.5	76.1	92.9	4.1	243.9
VMamba-T	30.0	4.9	82.6	96.0	6.2	161.3
MSVMamba-T	27.4	4.8	82.4	95.9	6.0	166.7
MambaVision-T	31.8	5.1	82.3	95.8	6.1	163.9
MPS-Mamba (proposed)	33.2	5.6	83.0	96.2	6.5	153.8

Table 2 presents the transfer-task reporting layout for COCO detection/instance segmentation and ADE20K semantic segmentation.

Backbone	**COCO box AP**	**COCO mask AP**	**ADE20K mIoU**	**Params (M)**	**FLOPs (G)**
Swin-T	46.0	41.6	44.5	28.3	47.0
VMamba-T / MSVMamba-T	48.2	43.5	48.2	30.0	49.0
MambaVision-T	48.1	43.7	48.5	31.8	51.0
MPS-Mamba (proposed)	49.0	44.4	49.2	33.2	55.0

Table 3 is intended to show how component ablations should be reported; the listed changes are hypothetical and cannot establish component effectiveness.

Variant	**Top-1 (%)**	**COCO AP**	**ADE20K mIoU**	**FLOPs (G)**	Δ Top-1 vs full
Full model	83.0	49.0	49.2	5.6	—
No local path	82.3	48.2	48.4	5.0	−0.7 pp
No coarse path	82.1	47.9	47.8	5.1	−0.9 pp
Fixed fusion	82.4	48.4	48.5	5.5	−0.6 pp
Two-direction scan	82.6	48.6	48.8	4.7	−0.4 pp

Table 4 shows a reporting format for corruption robustness and resolution scaling; the example values are not observed measurements.

Condition	**Input size**	**Baseline metric**	**MPS-Mamba metric**	**Relative degradation**
Clean	224×224	81.3	83.0	—
Gaussian noise	224×224	63.2	66.4	−22.3% / −20.0%
Blur	224×224	69.8	72.1	−14.3% / −13.1%
JPEG compression	224×224	75.4	77.6	−7.2% / −6.5%
Higher resolution	384×384	82.1	84.0	+1.0% / +1.2%

In the hypothetical classification table, MPS-Mamba reaches 83.0% top-1 accuracy, 1.7 percentage points above the Swin-T value and 0.4 points above VMamba-T. This example also assigns MPS-Mamba 33.2 million parameters and 5.6 GFLOPs, so the apparent accuracy advantage is accompanied by higher compute than ResNet-50 and Swin-T. The COCO and ADE20K values suggest gains of 0.8 AP and 0.7 mIoU over MambaVision-T, respectively. These differences are examples, not evidence of superiority. The ablation assigns the largest top-1 drop to removal of the coarse-context path (−0.9 percentage points), followed by removal of the local path (−0.7 points), fixed fusion (−0.6 points) and two-direction scanning (−0.4 points). This ordering is hypothetical and must be tested. In the robustness the proposed model has smaller relative degradation under noise, blur and JPEG compression than the baseline. The higher-resolution row indicates a small improvement, but resolution scaling must be measured at equal inference settings. No conclusion about statistical significance is justified by these example values. The numerical examples are presented in Tables 1–5.

The example standard deviations are calculated from the three displayed synthetic run values. They do not establish statistical significance, and no p-value or confidence interval is claimed. Actual reporting should use independently trained seeds and paired per-example analysis where appropriate.

6. DISCUSSION

The proposed design is motivated by a division of labour among spatial operators. Local convolution provides a strong short-range prior; directional selective scans propagate information across the original feature grid; pooled scans offer a lower-token-count route for broad context. The spatial gate is intended to adapt the contribution of each path at each location. This decomposition provides interpretable ablation points and makes it possible to determine whether a path is useful for classification, object localization or semantic boundaries.

The key scientific test is not whether the proposed model has more components, but whether each component contributes measurable value under a fair resource budget. In particular, a multi-path architecture may increase memory traffic and reduce hardware utilization. Parallel paths may be more expensive than a single scan despite favourable asymptotic complexity. The study should therefore report both algorithmic complexity and actual device measurements. Comparison against MambaVision and MSVMamba is essential because these methods already explore multi-scale and hybrid context modelling. The component-wise evaluation plan is summarized in Table 3.

If results support the hypothesis, the contribution should be stated narrowly: the specific spatially conditioned fusion of local, directional and pooled-context paths yields a measured benefit under the reported training and hardware conditions. If results do not support the hypothesis, the paper should report the negative result and revise or withdraw the novelty claim.

7. LIMITATIONS

• All numerical results in this version are simulated examples; no model training or benchmark evaluation is represented.

• The novelty boundary is narrow because multi-direction, local-window, multi-scale and hybrid Mamba vision architectures already exist.

• Four directional scans and multiple paths may increase memory traffic, kernel launches and latency despite linear sequence scaling.

• Pooling may suppress small objects and fine boundaries; upsampling may introduce alignment artefacts.

• Results may depend strongly on scan-kernel implementation, GPU architecture, image resolution and training recipe.

• Benchmark performance on ImageNet-1K, COCO and ADE20K does not establish universal superiority or domain generalization.

• The manuscript still requires actual training runs, verified benchmark results, confidence intervals, trained checkpoints, code release and independent reproduction.

8. CONCLUSION AND FUTURE WORK

This manuscript has specified a candidate Mamba-based visual backbone for long-range spatial feature extraction. MPS-Mamba combines local detail processing, directional selective state-space scans and pooled coarse-context scanning through a spatially conditioned fusion gate. The proposed evaluation covers image classification, object detection and semantic segmentation, together with ablation,

robustness, scaling and statistical validation. The architecture is technically testable, but no performance or efficiency conclusion can be drawn until the experiments are completed. Future work must implement the model, verify its novelty against the latest literature, conduct matched comparisons with strong Mamba and non-Mamba baselines, publish reproducible configurations and replace all result placeholders with verified measurements.

ACKNOWLEDGEMENT

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[2] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proc. International Conference on Learning Representations (ICLR), 2021.

[3] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022. doi: 10.1109/ICCV48922.2021.00986.

[4] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," arXiv:2312.00752, 2023. doi: 10.48550/arXiv.2312.00752.

[5] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model," in Proc. International Conference on Machine Learning (ICML), PMLR, vol. 235, 2024, pp. 62429–62442.

[6] Y. Liu et al., "VMamba: Visual State Space Model," in Advances in Neural Information Processing Systems, vol. 37, 2024. doi: 10.52202/079017-3273.

[7] Y. Shi, M. Dong, and C. Xu, "Multi-Scale VMamba: Hierarchy in Hierarchy Visual State Space Model," in Advances in Neural Information Processing Systems, vol. 37, 2024. doi: 10.52202/079017-0808.

[8] T. Huang, X. Pei, S. You, F. Wang, C. Qian, and C. Xu, "LocalMamba: Visual State Space Model with Windowed Selective Scan," in Computer Vision – ECCV 2024 Workshops, Lecture Notes in Computer Science, vol. 15633, 2025, pp. 12–22. doi: 10.1007/978-3-031-91979-4_2.

[9] A. Hatamizadeh and J. Kautz, "MambaVision: A Hybrid Mamba-Transformer Vision Backbone," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 25261–25270.

[10] C. Yang, Z. Chen, M. Espinosa, L. Ericsson, Z. Wang, J. Liu, and E. J. Crowley, "PlainMamba: Improving Non-Hierarchical Mamba in Visual Recognition," in Proc. British Machine Vision Conference (BMVC), 2024.

[11] W. Yu and X. Wang, "MambaOut: Do We Really Need Mamba for Vision?," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 4484–4496.

[12] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, pp. 211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[13] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol. 8693, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[14] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene Parsing Through ADE20K Dataset," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 633–641. doi: 10.1109/CVPR.2017.544.

[15] H. Zhang et al., "A Survey on Visual Mamba," Applied Sciences, vol. 14, no. 13, Art. 5683, 2024. doi: 10.3390/app14135683.

[16] X. Liu, C. Zhang, F. Huang, S. Xia, G. Wang, and L. Zhang, "Vision Mamba: A Comprehensive Survey and Taxonomy," IEEE Transactions on Neural Networks and Learning Systems, vol. 37, no. 2, pp. 505–525, 2026. doi: 10.1109/TNNLS.2025.3610435.

EXPERIMENTAL WORKFLOW



Fig. 2. Planned experimental workflow from benchmark preparation through controlled training, transfer evaluation, efficiency measurement and statistical validation. All steps require execution before empirical claims are made.

Table 5 demonstrates how independent-run results could be summarized; its three synthetic runs are not actual training seeds.

Metric	Run 1	Run 2	Run 3	Mean ± SD
ImageNet top-1 (%)	82.8	83.1	83.1	83.00 ± 0.17
COCO box AP	48.8	49.1	49.1	49.00 ± 0.17
ADE20K mIoU (%)	49.0	49.3	49.3	49.20 ± 0.17