2D-CrossScan Mamba: Enhancing State Space Models with Spatially Consistent Multi-Path 2D Information Propagation

Longlong Yu[1,5*], Wenxi Li[3*], Yaoqi Sun[4†], Hang Xu[1], Chenggang Yan[1], Yuchen Guo[2†]

[1]School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China
[2]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China
[3]KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China
[4]Lishui University, Lishui, China
[5]Zhuoxi Lab, Hangzhou, China
{longlong.yu, hxu, cgyan}@hdu.edu.cn, wxli@sfs.ecnu.edu.cn, sunyq2233@163.com, yuchen.w.guo@gmail.com

Abstract

Despite recent progress in adapting State Space Models such as Mamba to vision tasks, their intrinsic 1D scanning mechanism imposes limitations when applied to inherently 2D-structured data like images. Existing adaptations, including VMamba and 2DMamba, either suffer from inconsistency between scanning order and spatial locality or restrict inter-patch communication to singular paths, hindering effective information propagation. In this paper, we propose 2D-CrossScan, a novel 2D-compatible scan framework that enables spatially consistent, multi-path hidden state propagation by integrating modified state equations over two-dimensional neighborhoods. Furthermore, we mitigate redundant information accumulation due to overlapping paths via cross-directional subtraction. To fully align with the 2D spatial structure, we introduce a multi-directional scanning strategy that starts simultaneously from all four corners of the image, enabling diverse propagation paths and better feature integration. Our approach maintains efficiency, requiring only minimal architectural changes to existing Mamba variants. Experimental results demonstrate substantial improvements in multiple visual tasks, including object detection and semantic segmentation on PANDA and COCO datasets. Compared to baseline SSM-based methods, 2D-CrossScan consistently yields better spatial representations, as confirmed by extensive effective receptive field visualizations and attention analyses. These results highlight the importance of geometry-aware state propagation and validate 2D-CrossScan as a simple yet powerful extension to SSMs for vision.

Code — https://github.com/longlong-yu/official-cross-scan

Introduction

Recent advances in sequence modeling have significantly extended the capabilities of State Space Models (SSMs), most notably through the development of Mamba (Gu and Dao 2023), which achieves strong performance on language tasks via selective state propagation mechanisms. Building upon this success, a growing body of research—such as VMamba (Liu et al. 2024b), 2DMamba (Zhang et al. 2025), Vision Mamba (Zhu et al. 2024), SaMam (Liu et al. 2025a), Spatial-Mamba (Xiao et al. 2025), and LocalMamba (Huang et al. 2024)—has explored ways to adapt Mamba to the domain of computer vision by applying it to image data structured as sequences of patches. These approaches reflect an ongoing conversation in vision modeling: how to effectively transfer sequential architectures to spatially structured domains while preserving representational power.

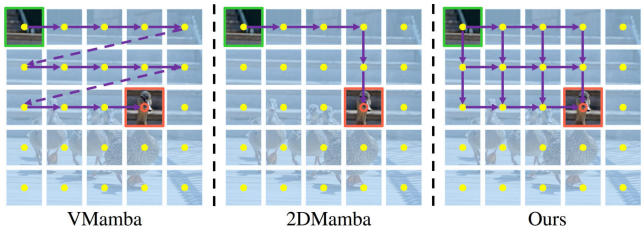

Figure 1: Information propagation paths in different methods. In **VMamba**, the information propagation path between the **starting** node and the **target** node does not align with the 2D spatial distance. Nodes closer or farther from the target node alternate in the propagation sequence, leading to non-optimal, non-shortest paths. **2DMamba** uses the shortest path in 2D space, but it relies on a single path and has temporal gaps between horizontal and vertical scans. **Our method** supports multi-path propagation in 2D space, ensures the use of the shortest paths, and synchronizes horizontal and vertical scanning directions.

However, despite this progress, existing SSM-based vision models still inherit the original 1D scanning mechanism of Mamba, which fails to respect the inherent 2D spatial geometry of image data. In 1D scan order, as shown in Fig.1 the spatial proximity between image patches is distorted—patches that are physically close may be updated non-consecutively, and long-range propagation paths may be required to connect even neighboring patches. Some methods attempt to mitigate this by introducing

*These authors contributed equally.
†These authors are co-corresponding authors.

multi-directional 1D scanning or modifying patch adjacency (e.g., DefMamba (Liu et al. 2025b), EfficientVMamba (Pei, Huang, and Xu 2025), and Hamba (Dong et al. 2024)), but such strategies still fundamentally operate on serialized views of 2D images. More notably, 2DMamba (Zhang et al. 2025) proposes a 2D scanning mechanism that leverages local neighborhood information in both row and column directions. Yet, it still confines information propagation to a single shortest path per patch and performs directional updates asynchronously, thereby introducing inconsistencies in spatial state transitions and limiting geometric expressiveness.

The purpose of this paper is to propose a new 2D-compatible scanning mechanism, 2D-CrossScan, that overcomes these theoretical and practical limitations by aligning the scan process with the true geometry of 2D space. Our method replaces Mamba's 1D scan with a true 2D scan and can also serve as a plug-in kernel for existing 1D-based variants. We address three core challenges: (1) how to associate each patch with all spatially adjacent directions simultaneously during scanning; (2) how to enable multi-path shortest-distance information propagation and aggregate such paths efficiently; and (3) how to design a 2D scan strategy that is consistent with 2D geometric structures and directionally diverse.

To solve these, we introduce 2D-CrossScan, a unified framework that modifies the state equation to aggregate hidden states from all previously scanned spatial neighbors in both horizontal and vertical directions, while suppressing redundancy caused by overlapping propagation paths. Built upon this formulation, we propose a novel multi-directional 2D scanning strategy that performs cross-scan from all four image corners—that enhances geometric diversity and aligns feature extraction with 2D spatial structure. This framework requires only minor changes to existing VMamba architectures, with the addition of four learnable weights to adaptively fuse directional features. Our method maintains computational efficiency and improves the spatial coherence and expressiveness of the learned features. Empirical results on object detection and semantic segmentation tasks across PANDA (Wang et al. 2020) and COCO (Lin et al. 2014) demonstrate clear improvements over state-of-the-art (SOTA) SSM-based baselines. Furthermore, analysis of effective receptive fields and attention maps further confirms the geometric advantages of our design.

Our main contributions are as follows:

- We propose a 2D-CrossScan framework that reformulates the state equation to aggregate hidden states from all previously scanned spatial neighbors, ensuring propagation along true 2D shortest paths. This enables spatially consistent updates and aligns information flow with the underlying geometry of visual data.
- We develop a multi-path aggregation mechanism that efficiently combines all possible shortest paths between patch pairs, while suppressing redundancy caused by overlapping routes via cross-directional subtraction.
- We design a novel 2D multi-directional scanning strategy, which performs scans from all four image corners simultaneously and fuses the resulting features with trainable weights. This design ensures geometric symmetry and maximizes feature diversity.
- Extensive experimental validation across PANDA and COCO benchmarks on object detection and semantic segmentation tasks, demonstrating that our method significantly outperforms SOTA SSM-based models in both accuracy and spatial coherence.

Related Work

State Space Models. State Space Models have gained attention in deep learning for sequence modeling due to their linear complexity, which offers advantages over self-attention in handling long-range dependencies. S4 (Gu, Goel, and Ré 2021) improved scalability by normalizing parameter matrices, while S5 (Smith, Warrington, and Linderman 2022) and S6 (Gu and Dao 2023) introduced MIMO systems and efficient parallel scanning algorithms. However, since SSMs were originally designed for 1D data, applying them to 2D structures like images presents challenges. Recent extensions to vision tasks, such as image classification (Liu et al. 2024b; Zhu et al. 2024), text-conditional motion synthesis (Zhang et al. 2024), and medical image segmentation (Zhu et al. 2025), still struggle with misalignment between 1D sequence models and 2D spatial relationships, hindering the capture of fine-grained spatial dependencies.

Selective Scanning Mechanism for Images. Mamba (Gu and Dao 2023) introduced a selective scanning mechanism to enhance SSM efficiency, but it still faces challenges when applied to 2D data. Several approaches have attempted to address this by incorporating multi-directional scanning (Liu et al. 2024b; Zhu et al. 2024; Xiao et al. 2025; Liu et al. 2025a; Dastani et al. 2025) or modifying patch adjacency (Huang et al. 2024; Liu et al. 2025b; Dong et al. 2024), improving the spatial relationships between patches. However, these methods continue to face challenges with misaligned scanning sequences. 2DMamba (Zhang et al. 2025) uses shortest paths between image patches but relies on single-path propagation, creating temporal gaps between horizontal and vertical scans that hinder full 2D spatial modeling (Fig.1). These challenges underscore the need for more advanced methods that fully exploit multi-path propagation and more efficiently process 2D data.

High-Resolution Object Detection. High-resolution object detection especially with datasets like PANDA (Wang et al. 2020), remains challenging. Traditional CNN-based methods struggle with dense feature extraction and long-range dependencies. Prior work has enhanced visual representations through hashing (Shen et al. 2015, 2018), cross-modal learning (Wang et al. 2017), and attention models (Gao et al. 2017). Related 3D vision advances—including Gaussian control (Lin et al. 2025), fern-based relocalization (Lu et al. 2023), and thermal Gaussian splatting (Lu et al. 2024)—offer additional insights. ACNet (Ding et al. 2019) further improves CNN kernels via asymmetric convolutions. Recent models, such as ViTs (Dosovitskiy et al. 2020), PnP-DETR (Wang et al. 2021), SparseFormer (Li et al. 2024b,a), GigaHumanDet (Liu et al. 2024a), SaccadeDet (Li et al.

2024c,d) and LSNet (Wang et al. 2025), show promise but fail to fully address spatial complexity inherent in high-resolution images. SSMs, including Mamba, hold potential for improving long-range dependency modeling, but their 1D scanning mechanisms limit their effectiveness with 2D data. Our 2D-CrossScan method addresses these limitations by enabling multi-path propagation and synchronized 2D scanning, enhancing feature extraction and improving high-resolution object detection tasks.

Preliminaries

State Equations in 1D Scanning. State Space Sequence Models (S4) (Gu, Goel, and Ré 2021) process sequential data using time-invariant continuous state equations. Mamba (Gu and Dao 2023) extends these equations to an input-dependent, time-varying form and introduces a selective scanning mechanism, achieving linear-time processing of long sequences. In practice, Mamba performs state transitions independently for each channel via 1D state equations, whose outputs are combined to produce the final representation. The discretized state equation is expressed as:

$$\begin{aligned} \mathbf{h}_t &= \overline{\mathbf{A}}_t \odot \mathbf{h}_{t-1} + \overline{\mathbf{B}}_t u_t, \\ y_t &= \mathbf{C}_t \mathbf{h}_t + D u_t, \end{aligned} \tag{1}$$

where $\mathbf{h}_t$ is the hidden state, $u_t \in \mathbb{R}$ is the input, $\overline{\mathbf{A}}_t, \overline{\mathbf{B}}_t \in \mathbb{R}^{N\times 1}$ and $\mathbf{C}_t \in \mathbb{R}^{1\times N}$ are input-dependent parameters, and $D \in \mathbb{R}$ is input-independent. N denotes the hidden dimension. Each dimension of the state updates independently, enabling parallel computation.

Attention Formulation in 1D Scanning. To analyze the selective mechanism in an attention-like manner, VMamba (Liu et al. 2024b) rewrites the 1D selective scanning state equation into an attention-style form:

$$\begin{aligned} y_t &= (\mathbf{Q}_t \odot \mathbf{w}_t)\, \mathbf{h}_0 + \sum_{i=1}^{t} (\mathbf{Q}_t \odot \mathbf{w}_t) \left(\frac{\mathbf{K}_i}{\mathbf{w}_i}\right) V_i, \\ w_i^{(n)} &= \prod_{k=1}^{i} \overline{A}_k^{(n)}, \quad n \in [0, 1, 2, \cdots, N-1], \end{aligned} \tag{2}$$

where $\mathbf{Q}_t \coloneqq \mathbf{C}_t$, $\mathbf{K}_t \coloneqq \overline{\mathbf{B}}_t$, and $V_t \coloneqq u_t$ serve as the Query, Key, and Value, and $\mathbf{w}$ denotes the selective-scan weights. This reformulation casts the 1D scan as an attention-like structure, enabling clearer analysis of node relationships and importance during scanning.

Method

In this section, we describe 2D-CrossScan, which addresses the limitations of 1D scanning in state space models when applied to 2D visual structures. By leveraging image geometry for spatially consistent, multi-path propagation, our method enhances feature extraction for tasks such as object detection and semantic segmentation.

2D-CrossScan Framework. The primary challenge addressed by 2D-CrossScan is the misalignment between the 1D scan order and the spatial relationships in images. While traditional Mamba employs a 1D scanning mechanism that processes data sequentially, this method is not well-suited for the 2D spatial structure of images. The core issue with the 1D scan is that it results in spatially inconsistent updates, where nearby patches may rely on distant intermediaries for communication, degrading the quality of the feature representation, as illustrated in Fig.1.

To resolve this, we introduce 2D-CrossScan, a framework that modifies the state equation to better align with the 2D spatial geometry of the image. In this framework, information propagation follows the shortest spatial paths between patches, preserving the true 2D spatial relationships that are crucial for vision tasks. The updated 2D state equation is as follows:

$$\hat{\mathbf{h}}_{x,y} = \overline{\mathbf{A}}_{x,y} \odot (\mathbf{h}_{x-1,y} + \mathbf{h}_{x,y-1}) + \overline{\mathbf{B}}_{x,y} u_{x,y}, \tag{3}$$

where $\overline{\mathbf{A}}_{x,y} \in \mathbb{R}^{N\times 1}$, $\overline{\mathbf{B}}_{x,y} \in \mathbb{R}^{N\times 1}$, and (x, y) denotes the coordinates of the patch in the 2D image. This equation allows the scanning process to respect 2D spatial distances, ensuring that information propagates through the shortest possible paths between patches.

Multi-path Propagation. One of the key advantages of 2D-CrossScan is its ability to support multi-path propagation. Unlike the 1D scan, which limits information flow to a single path between patches, the 2D-CrossScan allows multiple propagation paths to be utilized for each patch pair. This is particularly useful in 2D spaces, where the number of possible paths between patches increases factorial-wise with the distance between them.

For two patches with 2D coordinates (x, y) and $(x + m, y + n)$, the number of shortest paths between them is given by the combinatorial formula::

$$C_{m+n}^{m} = \frac{(m+n)!}{m! \cdot n!}. \tag{4}$$

Directly accumulating information from all these paths would lead to factorial-level computational complexity. To address this, our framework modifies the state equation to reduce the computational overhead. Equ.3 uses a single arithmetic operation for each state transition, enabling efficient multi-path propagation and aggregation without the factorial cost.

However, redundancy arises when different propagation paths overlap. As shown in Fig.2(a), two adjacent patches that receive information at the target node may cover overlapping regions, which results in redundant information. To mitigate this, we heuristically subtract the overlapping portions during aggregation:

$$\mathbf{h}_{x,y} = \hat{\mathbf{h}}_{x,y} - \overline{\mathbf{A}}_{x,y} \odot \frac{\overline{\mathbf{A}}_{x-1,y} + \overline{\mathbf{A}}_{x,y-1}}{2} \odot \mathbf{h}_{x-1,y-1}, \tag{5}$$

which ensures that only unique information from each propagation path is considered. And the final output can be expressed as:

$$y_{x,y} = \mathbf{C}_{x,y} \mathbf{h}_{x,y} + D u_{x,y}, \tag{6}$$

where $\mathbf{C}_{x,y} \in \mathbb{R}^{1\times N}$, $D \in \mathbb{R}$, $u_{x,y} \in \mathbb{R}$. This process improves the spatial uniformity of the feature representation.

Scanning Algorithm Implementation. The scanning algorithm implementation builds upon the 2D state equation discussed above (Fig.2(b)). To compute the hidden state of each patch, we need to consider the hidden states of its

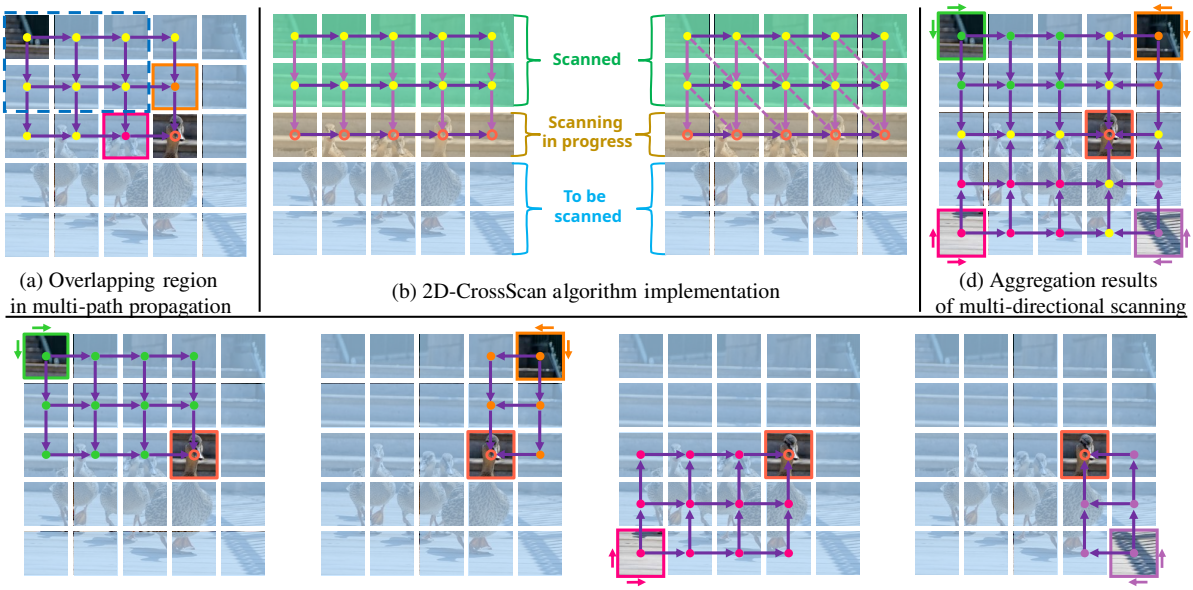

Figure 2: Illustration of the 2D-CrossScan architecture. **(a)** In 2D, different propagation paths share many intersecting nodes, creating overlapping regions (**blue dashed** box). **(b) Left**: scanning without overlap suppression—each row is processed sequentially, and hidden states are passed to the next row (Equ.3). **Right**: scanning with overlap suppression—overlapping values are removed using diagonal nodes during state transitions (Equ.5). **(c)** Multi-directional scanning from different image corners. **(d)** Final aggregation of all scan directions; except for the target row and column, all other regions remain independent and non-overlapping.

neighboring patches in both the horizontal and vertical directions. Therefore, we scan the image row by row, processing the unscanned patches in the topmost row at each step. Once all nodes in the current row have been processed, the hidden states of these nodes are passed to the next row directly beneath them.

In addition to handling vertical propagation, we also address the redundancy caused by multi-path propagation. As previously described, we subtract the overlapping information from different directions during the hidden state computation. After scanning the current row, the hidden states of the scanned nodes are passed to the next diagonal node in the lower-right direction, ensuring smooth information propagation across the 2D grid.

Multi-directional Scanning Strategy. To further enhance feature representation, we propose a multi-directional scanning strategy that aligns with the geometric properties of the 2D space. Unlike traditional methods, which typically use 1D scanning in multiple directions (e.g., scanning first horizontally and then vertically), our method performs a 2D scan in all directions simultaneously. Each scan considers both horizontal and vertical dimensions, ensuring that the spatial relationships between patches are captured in their full 2D context.

The results from each of these 2D scans are then aggregated to form a unified feature representation. The regions scanned from different directions do not overlap, except for the horizontal and vertical directions of the target node. This ensures that the features are spatially coherent and adequately cover the entire 2D image. Fig.2(c) and (d) illustrate the process of multi-directional scanning. In Fig.2(c), we show the individual scanning processes in each of the four directions. Each direction performs a 2D scan, processing both horizontal and vertical neighbors simultaneously, and ensuring that the entire spatial context is captured in each scan. By scanning from the four corners and processing both dimensions simultaneously, we maximize the diversity and richness of the features extracted from the image. In Fig.2(d), we show the aggregation of the results from the four scanning directions. The feature maps from the top-left, top-right, bottom-left, and bottom-right scans are combined to produce a unified feature map that captures spatial relationships across all directions. This aggregation ensures that the final representation is more complete and spatially consistent.

To combine the features from the different scanning directions, we introduce four trainable weight coefficients. These weights ($\omega_{[\cdot]}$) adaptively fuse the directional features, improving the ability of the model to capture complex spatial relationships. The aggregation formula is as follows:

$$y = \omega_{tl} \cdot y_{tl} + \omega_{tr} \cdot y_{tr} + \omega_{bl} \cdot y_{bl} + \omega_{br} \cdot y_{br}, \tag{7}$$

where t, b, l, and r represent the top, bottom, left, and right scanning directions, respectively.

Method	Backbone	AP_{total}	AP_s	AP_m	AP_l
Pedestron (Hasan et al. 2021)	HRNet	0.699	0.224	0.671	0.788
YOLOX (Ge et al. 2021)	CSPDarknet	0.695	0.211	0.635	0.815
GigaDet (Chen et al. 2022)	CSP-DarkNet-53	0.684	0.210	0.599	0.762
SaccadeDet (Li et al. 2024c)	CSPDarkNet	0.760	0.340	0.751	0.808
ClusDet (Yang et al. 2019)	ResNet-50	0.718	0.219	0.696	0.782
FasterRCNN (Fan et al. 2022)	ResNet-50	0.705	0.203	0.712	0.760
PAN (Fan et al. 2022)	ResNet-50	0.715	0.256	0.719	0.768
CentripetalNet (Dong et al. 2020)	Hourglass-104	0.626	0.273	0.601	0.696
GigaHumanDet (Liu et al. 2024a)	Hourglass-52	0.796	0.485	0.788	0.848
Dynamic-Head (Dai et al. 2021)	Swin-T	0.592	0.165	0.537	0.694
Dynamic-Head+DEG (Song et al. 2021)	PVT-DEG	0.575	0.154	0.508	0.695
Dynamic-Head+SparseFormer (Li et al. 2024b)	SparseFormer	0.771	0.364	0.740	0.863
DINO (Zhang et al. 2022)	Swin-T	0.606	0.367	0.612	0.649
DINO+DEG (Song et al. 2021)	PVT-DEG	0.582	0.339	0.578	0.624
DINO+SparseFormer (Li et al. 2024b)	SparseFormer	0.780	0.508	0.781	0.823
DINO+Ours	2D-CrossScan-T	0.829	0.521	0.810	0.875

Table 1: Performance comparison on PANDA.

Method	Params	FLOPs	TP	RT	GPU	AP_{total}
VMamba-T	50M	261G	0.230	30.2h	6.8G	$80.86_{\pm 0.12}$
2DMamba-T	50M	262G	0.225	31.0h	6.9G	$80.93_{\pm 0.12}$
2D-CrossScan-T	50M	262G	0.224	31.5h	7.3G	$82.80_{\pm 0.10}$

Table 2: Performance comparison of Mamba-based methods on PANDA. FLOPs are computed at 1280×800. AP_{total} is reported as the mean ± std over 3 runs. TP denotes throughput (img/s), and RT denotes training time.

Backbone	Params	FLOPs	Top-1
Swin-T (Liu et al. 2021)	28M	4.5G	81.3
ConvNeXt-T (Liu et al. 2022)	29M	4.5G	82.1
VMamba-T (Liu et al. 2024b)	30M	4.9G	82.6
LocalVMamba-T (Huang et al. 2024)	26M	5.7G	82.7
QuadMamba-S (Xie et al. 2024)	31M	5.5G	82.4
MSVMamba-T (Shi, Dong, and Xu 2024)	32M	5.1G	83.0
DefMamba-S (Liu et al. 2025b)	32M	4.8G	83.5
Spatial-Mamba-T (Xiao et al. 2025)	27M	4.5G	83.5
2DMamba-T (Zhang et al. 2025)	30M	4.9G	82.8
2D-CrossScan-T	30M	4.9G	83.5

Table 3: Performance comparison on ImageNet-1K with an image size of 224.

Experiments

Datasets. We train and evaluate our models on PANDA (Wang et al. 2020), ImageNet-1K (Deng et al. 2009), COCO (Lin et al. 2014), and ADE20K (Zhou et al. 2017). PANDA contains high-resolution real-world scenes (13 training and 5 testing) for object detection. ImageNet-1K is used for image classification, COCO for object detection and instance segmentation, and ADE20K for semantic segmentation.

Implementation Details. For PANDA, we replace the backbone of SparseFormer (Li et al. 2024b) with 2D-CrossScan and follow its original 36-epoch training schedule. For the other datasets, we modify VMamba (Liu et al. 2024b) by substituting its scanning module with 2D-CrossScan, keeping the training strategy unchanged except for adding four trainable coefficients to fuse results from different scan directions. PANDA experiments are conducted on 8×4090 GPUs, and all others on $8 \times A100$ GPUs.

Backbone	Params	FLOPs	AP^b	AP^m
Swin-T	48M	267G	0.427	0.393
ConvNeXt-T	48M	262G	0.442	0.401
VMamba-T	50M	271G	0.473	0.427
LocalVMamba-T	45M	291G	0.467	0.422
QuadMamba-S	55M	301G	0.467	0.424
MSVMamba-T	52M	275G	0.469	0.425
DefMamba-S	-	268G	0.475	0.428
Spatial-Mamba-T	46M	261G	0.476	0.429
2DMamba-T †	49M	247G	0.472	0.426
2D-CrossScan-T	49M	247G	0.477	0.429

Table 4: Performance comparison on COCO using Mask R-CNN (He et al. 2017). FLOPs are computed at 1280×800. † denotes the results obtained from our experiments, where we followed the 2DMamba method exactly.

Results on PANDA. For object detection tasks on the PANDA dataset, we present the results in Tab.1 and Tab.2. Our experiments show that the 2D-CrossScan method outperforms SOTA methods on high-resolution images. Although our scan kernel remains a linear attention mechanism, extending attention from 1D to 2D inevitably increases FLOPs and memory. Nonetheless, our method shows clear advantages on high-resolution tasks. This performance improvement is attributed to the ability of our method to propagate and update information based on the 2D spatial geometry of the data, allowing it to more effectively capture the spatial relationships in the image.

Results on ImageNet-1K, COCO and ADE20K. We evaluate our method on the ImageNet-1K (Tab.3), COCO (Tab.4), and ADE20K (Tab.5) datasets across different tasks. The results show that 2D-CrossScan achieves competitive performance, surpassing 2DMamba, a method based on another 2D scanning approach, as well as other 1D-based vari-

Backbone	Params	FLOPs	mIoU(SS)	mIoU(MS)
Swin-T	60M	945G	0.445	0.458
ConvNeXt-T	60M	939G	0.460	0.467
VMamba-T	62M	949G	0.479	0.488
LocalVMamba-T	57M	970G	0.479	0.491
QuadMamba-S	62M	961G	0.472	0.481
MSVMamba-T	63M	953G	0.479	0.485
DefMamba-S	65M	946G	0.488	0.496
Spatial-Mamba-T	57M	936G	0.486	0.494
2DMamba-T	62M	950G	0.486	0.493
2D-CrossScan-T	62M	950G	0.488	0.497

Table 5: Performance comparison on ADE20K using UperNet (Xiao et al. 2018). FLOPs are computed at 512 × 2048. "SS" and "MS" denote single-scale and multi-scale testing, respectively.

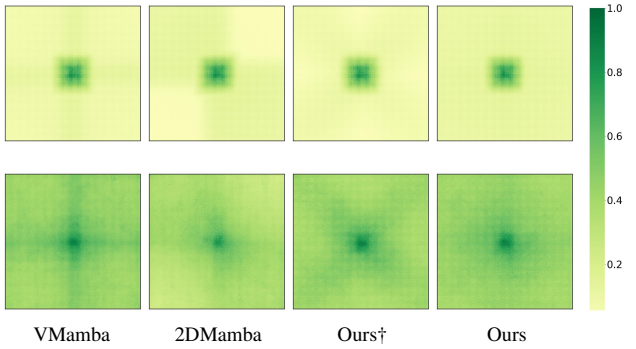

Figure 3: Comparison of the Effective Receptive Fields of our method and other scanning methods. Pixels with higher intensity indicate larger responses relative to the central pixel. † denotes the case where overlapping information from multi-path propagation has not been suppressed.

ants of Mamba. This demonstrates that our 2D-CrossScan method, which better aligns with the 2D spatial geometry of the image, is able to extract features more effectively and improve model performance.

ERF Analysis. The Effective Receptive Field (ERF) is computed as the gradient of the output with respect to each input pixel, indicating which regions influence the feature representation (Luo et al. 2016; Ding et al. 2022). We compare the ERFs of VMamba, 2DMamba, and 2D-CrossScan (with and without redundancy suppression) for both pretrained and finetuned models. As shown in Fig.3, VMamba's ERF concentrates along horizontal and vertical bands around the target pixel due to its multi-directional 1D scanning, which misaligns with the 2D spatial structure. 2DMamba shows similar horizontal–vertical biases, inheriting VMamba's 1D scanning behavior and producing a noticeable upper-left to lower-right preference.

For 2D-CrossScan, before suppressing redundant information, we observe a strong bias near the two diagonal regions. This bias arises because most multi-path overlaps occur near the diagonals. With redundancy suppression, this diagonal bias disappears, yielding a more uniform ERF across the 2D plane. The receptive field is now more consistent, with bias distributed according to the distance from the center pixel, decreasing as the distance increases. This demonstrates that 2D-CrossScan effectively captures spatial relationships across the entire image.

Attention Analysis. Following the approach in VMamba (Liu et al. 2024b), we rewrite the state equation in an attention-like form. The attention forms for 2DMamba (Zhang et al. 2025) are given as:

$$\begin{aligned} y_{x,y} = &\left[\mathbf{Q}_{x,y} \odot \tilde{\mathbf{w}}\left(0,0;x,y\right)\right]\mathbf{h}_{0,0} \\ &+\sum_{j=1}^{y}\sum_{i=1}^{x}\left[\mathbf{Q}_{x,y} \odot \tilde{\mathbf{w}}\left(i,j;x,y\right)\right]\mathbf{K}_{i,j}V_{i,j}, \end{aligned} \tag{8}$$

$$\tilde{w}^{(n)}\left(i,j;x,y\right) = \prod_{k=i+1}^{x}\overline{A}_{k,j}^{(n)}\prod_{s=j+1}^{y}\overline{A}_{x,s}^{(n)}, \tag{9}$$

where $\mathbf{Q}_{x,y} \coloneqq \mathbf{C}_{x,y}, \mathbf{K}_{x,y} \coloneqq \overline{\mathbf{B}}_{x,y}, V_{x,y} \coloneqq u_{x,y}, n \in [0,1,2,\cdots,N-1]$. And the attention formulations for 2D-CrossScan can be expressed as:

$$\begin{aligned} y_{x,y} = &\left[\mathbf{Q}_{x,y} \odot \mathbf{w}\left(0,0;x,y\right)\right]\mathbf{h}_{0,0} \\ &+\sum_{j=1}^{y}\sum_{i=1}^{x}\left[\mathbf{Q}_{x,y} \odot \mathbf{w}\left(i,j;x,y\right)\right]\mathbf{K}_{i,j}V_{i,j}, \end{aligned} \tag{10}$$

$$w_{x,y}^{(n)} = \begin{cases} \overline{A}_{x,y}^{(n)} \cdot w_{x,y-1}^{(n)}, & i = x, j < y \\ \overline{A}_{x,y}^{(n)} \cdot w_{x-1,y}^{(n)}, & i < x, j = y \\ \overline{A}_{x,y}^{(n)} \cdot \left(w_{x,y-1}^{(n)} + w_{x-1,y}^{(n)}\right) & \\ \quad - \frac{\overline{A}_{x-1,y}^{(n)} + \overline{A}_{x,y-1}^{(n)}}{2} & \\ \quad \cdot \overline{A}_{x,y}^{(n)} \cdot w_{x-1,y-1}^{(n)}, & i < x, j < y \\ 0, & \text{others} \end{cases}. \tag{11}$$

Unlike 1D scanning, where the query (Q) and key (K) are independently related to the start and end positions, the 2D scanning method jointly models the relationship between start and end positions within the attention formulations.

With multi-path redundancy suppression, the attention formulation incorporates more hidden states with stronger coupling, yielding a more comprehensive integration of spatial information. Fig.4 shows the attention activation maps for different methods. Compared to VMamba, 2D-CrossScan focuses more on the areas directly related to the target, suppressing regions that are weakly correlated with the target but are closer to the target node in the 1D scan sequence, as seen in VMamba. 2DMamba still exhibits attention biases along the diagonal due to its use of 1D scanning, where horizontal and vertical scans are inconsistent in time. As a result, in the second scan, nodes closer to the target node receive stronger attention. In contrast, 2D-CrossScan distributes attention weights in a way that aligns with the 2D spatial distances, capturing the relationships between nodes in a manner that better reflects the 2D geometry.

Ablation Study

Multi-path Redundancy Suppression. In Tab.6, we present the results of an ablation experiment that evaluates the impact of suppressing multi-path redundancy. Without this suppression, the feature maps become overly large,

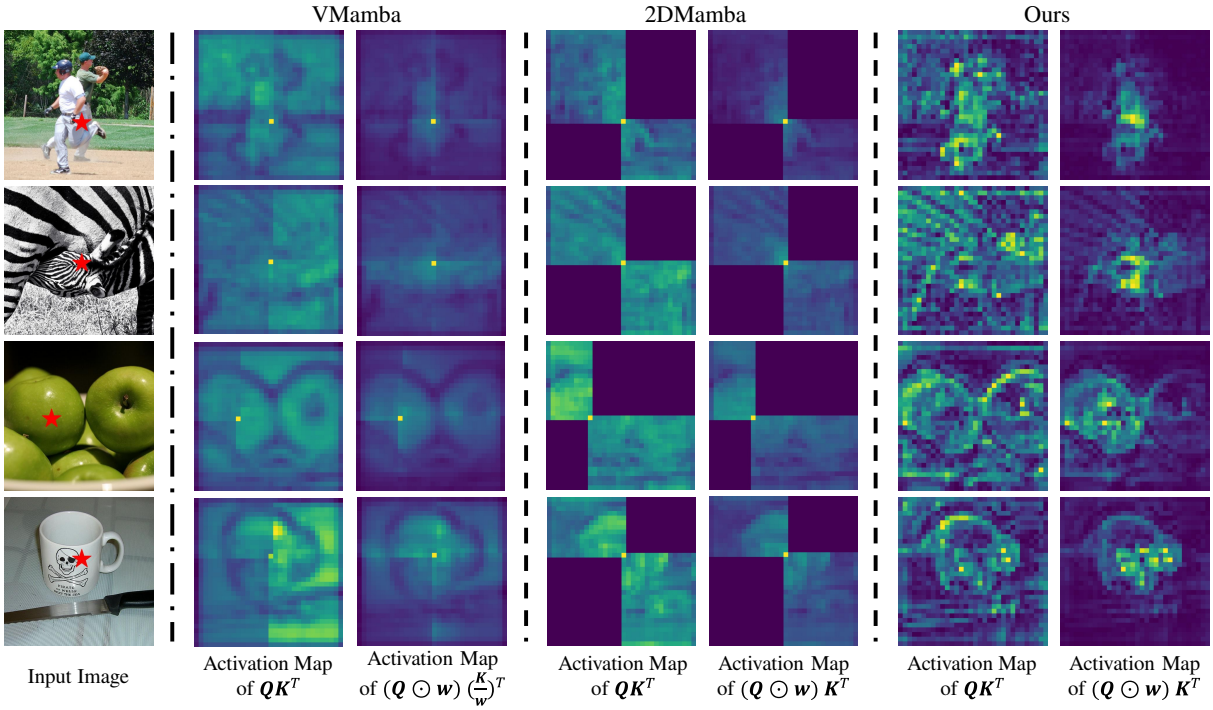

Figure 4: Comparison of attention activation maps across different models (with query nodes marked by red stars).

Model	SR	AP_{total}	AP_s	AP_m	AP_l
2D-CrossScan-T		0.802	0.488	0.790	0.852
2D-CrossScan-T	✓	0.829	0.521	0.810	0.875

Table 6: Ablation study on redundant information. SR denotes the suppression of redundant overlapping information in multi-path propagation.

Model	AP_{total}	AP_s	AP_m	AP_l
2D-UnidiScan-T	0.782	0.469	0.776	0.837
2D-BidiScan-T	0.806	0.502	0.795	0.856
2D-CrossScan-T	0.829	0.521	0.810	0.875

Table 7: Ablation study on scanning strategies.

making it difficult to complete training. To address this, we multiplied the right side of Equ.3 by a factor of 0.5 in our ablation experiment. The results show that suppressing redundant information significantly improves model performance.

Multi-directional Scanning. Tab.7 shows the results from experiments using different scan directions. Our four-directional scanning strategy outperforms both single-direction and two-direction scanning strategies, demonstrating that scanning from four different corners allows the model to perceive the entire 2D space more effectively, leading to better accuracy.

Weight Coefficients. Tab.8 reports the effect of adding weight coefficients when aggregating features from different scan directions. These coefficients accelerate early-stage convergence but lead to only marginal final accuracy gains. This indicates that while they help adjust directional contributions during training, they have limited impact on the model's overall performance after convergence.

Model	TW	AP_{total}	AP_s	AP_m	AP_l
2D-CrossScan-T		0.826	0.518	0.808	0.874
2D-CrossScan-T	✓	0.829	0.521	0.810	0.875

Table 8: Ablation study on trainable weight coefficients. TW denotes the use of trainable weight coefficients in the aggregation of multi-directional scanning results.

Conclusions

We propose 2D-CrossScan, a framework that adapts Mamba to 2D spatial features. Our method aligns information propagation with true spatial distances and aggregates signals across all valid paths while suppressing redundant overlaps, leading to more spatially uniform features. A multi-directional scanning strategy—performed from all four image corners in parallel—further captures both horizontal and vertical dependencies. Experiments demonstrate significant gains across different tasks, surpassing existing methods. Our approach improves feature extraction by more effectively capturing the 2D spatial geometry, and we believe it has the potential to become a valuable tool for computer vision applications. Several directions remain for future work, including a formal unbiasedness analysis of Eq.5, exploring alternative suppression schemes, and extending the method to broader tasks.

Acknowledgments

This work was supported by the National Science and Technology Major Project (No. 2022ZD0119402), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province (Nos. 2024C01142, 2024C01107, 2023C01030, 2023C01046), the National Natural Science Foundation of China (No. U21B2013), the National Key Research and Development Program of China (No. 2023YFB4502800), the Zhejiang Provincial Natural Science Foundation of China (No. LQN25F020015), and the Key R&D Program of Xinjiang, China (No. 2022B01006).

References

Chen, K.; Wang, Z.; Wang, X.; Gong, D.; Yu, L.; Guo, Y.; and Ding, G. 2022. Towards real-time object detection in gigapixel-level video. *Neurocomput.*, 477: 14–24.

Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; and Zhang, L. 2021. Dynamic head: Unifying object detection heads with attentions. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 7373–7382.

Dastani, S.; Bahri, A.; Yazdanpanah, M.; Noori, M.; Osowiechi, D.; Hakim, G. A. V.; Beizaee, F.; Cheraghalikhani, M.; Mondal, A. K.; Lombaert, H.; et al. 2025. Spectral State Space Model for Rotation-Invariant Visual Representation Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23881–23890.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 248–255. Ieee.

Ding, X.; Guo, Y.; Ding, G.; and Han, J. 2019. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 1911–1920.

Ding, X.; Zhang, X.; Han, J.; and Ding, G. 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 11963–11975.

Dong, H.; Chharia, A.; Gou, W.; Vicente Carrasco, F.; and De la Torre, F. D. 2024. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. In *Adv. Neural Inform. Process. Syst.*, volume 37, 2127–2160.

Dong, Z.; Li, G.; Liao, Y.; Wang, F.; Ren, P.; and Qian, C. 2020. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 10519–10528.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, J.; Liu, H.; Yang, W.; See, J.; Zhang, A.; and Lin, W. 2022. Speed up object detection on gigapixel-level images with patch arrangement. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 4653–4661.

Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; and Shen, H. T. 2017. Video captioning with attention-based LSTM and semantic consistency. *IEEE Trans. Multimedia*, 19(9): 2045–2055.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Hasan, I.; Liao, S.; Li, J.; Akram, S. U.; and Shao, L. 2021. Generalizable pedestrian detection: The elephant in the room. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 11328–11337.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2961–2969.

Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024. Localmamba: Visual state space model with windowed selective scan. In *Proc. Eur. Conf. Comput. Vis.*, 12–22. Springer.

Li, W.; Guo, Y.; Zheng, J.; Lin, H.; Ma, C.; Fang, L.; and Yang, X. 2024a. Bridging the gap between object detection in close-up and high-resolution wide shots. *Comput. Vis. Image Underst.*, 249: 104181.

Li, W.; Guo, Y.; Zheng, J.; Lin, H.; Ma, C.; Fang, L.; and Yang, X. 2024b. Sparseformer: Detecting objects in hrw shots via sparse vision transformer. In *Proc. ACM Int. Conf. Multimedia.*, 4851–4860.

Li, W.; Zhang, R.; Lin, H.; Guo, Y.; Ma, C.; and Yang, X. 2024c. SaccadeDet: A Novel Dual-Stage Architecture for Rapid and Accurate Detection in Gigapixel Images. In *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 392–408. Springer.

Li, W.; Zhang, R.; Lin, H.; Guo, Y.; Ma, C.; and Yang, X. 2024d. SaccadeMOT: Enhancing Object Detection and Tracking in Gigapixel Images via Scale-Aware Density Estimation. In *Proc. Eur. Conf. Artif. Intell.*, 145–152.

Lin, L.; Lu, R.; Chen, Q.; Ren, H.; Lu, M.; Sun, Y.; Yan, C.; and Xue, A. 2025. VGNC: Reducing the Overfitting of Sparse-view 3DGS via Validation-guided Gaussian Number Control. In *Proc. ACM Int. Conf. Multimedia.*, 4474–4483.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comput. Vis.*, 740–755. Springer.

Liu, C.; Wei, H.; Yang, J.; Liu, J.; Li, W.; Guo, Y.; and Fang, L. 2024a. Gigahumandet: Exploring full-body detection on gigapixel-level images. In *Proc. AAAI Conf. Artif. Intell.*, volume 38, 10092–10100.

Liu, H.; Wang, L.; Zhang, Y.; Yu, Z.; and Guo, Y. 2025a. SaMam: Style-aware State Space Model for Arbitrary Image Style Transfer. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 28468–28478.

Liu, L.; Zhang, M.; Yin, J.; Liu, T.; Ji, W.; Piao, Y.; and Lu, H. 2025b. Defmamba: Deformable visual state space model. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 8838–8847.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024b. Vmamba: Visual state space model. In *Adv. Neural Inform. Process. Syst.*, volume 37, 103031–103063.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 11976–11986.

Lu, R.; Chen, H.; Zhu, Z.; Qin, Y.; Lu, M.; Yan, C.; et al. 2024. ThermalGaussian: Thermal 3D Gaussian Splatting. In *Proc. Int. Conf. Learn. Represent.*

Lu, R.; Zhu, Z.; Fu, S.; Chen, S.; Wang, T.; Yan, C.; and Xu, F. 2023. Self-supervised camera relocalization with hierarchical fern encoding. *IEEE Trans. Instrum. Meas.*, 73: 1–12.

Luo, W.; Li, Y.; Urtasun, R.; and Zemel, R. 2016. Understanding the effective receptive field in deep convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, volume 29.

Pei, X.; Huang, T.; and Xu, C. 2025. Efficientvmamba: Atrous selective scan for light weight visual mamba. In *Proc. AAAI Conf. Artif. Intell.*, volume 39, 6443–6451.

Shen, F.; Shen, C.; Liu, W.; and Tao Shen, H. 2015. Supervised discrete hashing. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 37–45.

Shen, F.; Xu, Y.; Liu, L.; Yang, Y.; Huang, Z.; and Shen, H. T. 2018. Unsupervised deep hashing with similarity-adaptive and discrete optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12): 3034–3044.

Shi, Y.; Dong, M.; and Xu, C. 2024. Multi-scale vmamba: Hierarchy in hierarchy visual state space model. In *Adv. Neural Inform. Process. Syst.*, volume 37, 25687–25708.

Smith, J. T.; Warrington, A.; and Linderman, S. W. 2022. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*.

Song, L.; Zhang, S.; Liu, S.; Li, Z.; He, X.; Sun, H.; Sun, J.; and Zheng, N. 2021. Dynamic grained encoder for vision transformers. In *Adv. Neural Inform. Process. Syst.*, volume 34, 5770–5783.

Wang, A.; Chen, H.; Lin, Z.; Han, J.; and Ding, G. 2025. LSNet: See Large, Focus Small. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 9718–9729.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-modal retrieval. In *Proc. ACM Int. Conf. Multimedia.*, 154–162.

Wang, T.; Yuan, L.; Chen, Y.; Feng, J.; and Yan, S. 2021. Pnp-detr: Towards efficient visual analysis with transformers. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 4661–4670.

Wang, X.; Zhang, X.; Zhu, Y.; Guo, Y.; Yuan, X.; Xiang, L.; Wang, Z.; Ding, G.; Brady, D.; Dai, Q.; et al. 2020. Panda: A gigapixel-level human-centric video dataset. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 3268–3278.

Xiao, C.; Li, M.; ZHANG, Z.; Meng, D.; and Zhang, L. 2025. Spatial-Mamba: Effective Visual State Space Models via Structure-Aware State Fusion. In *Proc. Int. Conf. Learn. Represent.*

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proc. Eur. Conf. Comput. Vis.*, 418–434.

Xie, F.; Zhang, W.; Wang, Z.; and Ma, C. 2024. Quadmamba: Learning quadtree-based selective scan for visual state space model. In *Adv. Neural Inform. Process. Syst.*, volume 37, 117682–117707.

Yang, F.; Fan, H.; Chu, P.; Blasch, E.; and Ling, H. 2019. Clustered object detection in aerial images. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 8311–8320.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, J.; Nguyen, A. T.; Han, X.; Trinh, V. Q.-H.; Qin, H.; Samaras, D.; and Hosseini, M. S. 2025. 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 3583–3592.

Zhang, Z.; Liu, A.; Reid, I.; Hartley, R.; Zhuang, B.; and Tang, H. 2024. Motion mamba: Efficient and long sequence motion generation. In *Proc. Eur. Conf. Comput. Vis.*, 265–282. Springer.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 633–641.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: efficient visual representation learning with bidirectional state space model. In *Proc. Int. Conf. Mach. Learn.*, 62429–62442.

Zhu, X.; Wang, W.; Zhang, C.; and Wang, H. 2025. Polyp-mamba: A hybrid multi-frequency perception gated selection network for polyp segmentation. *Inf. Fusion*, 115: 102759.