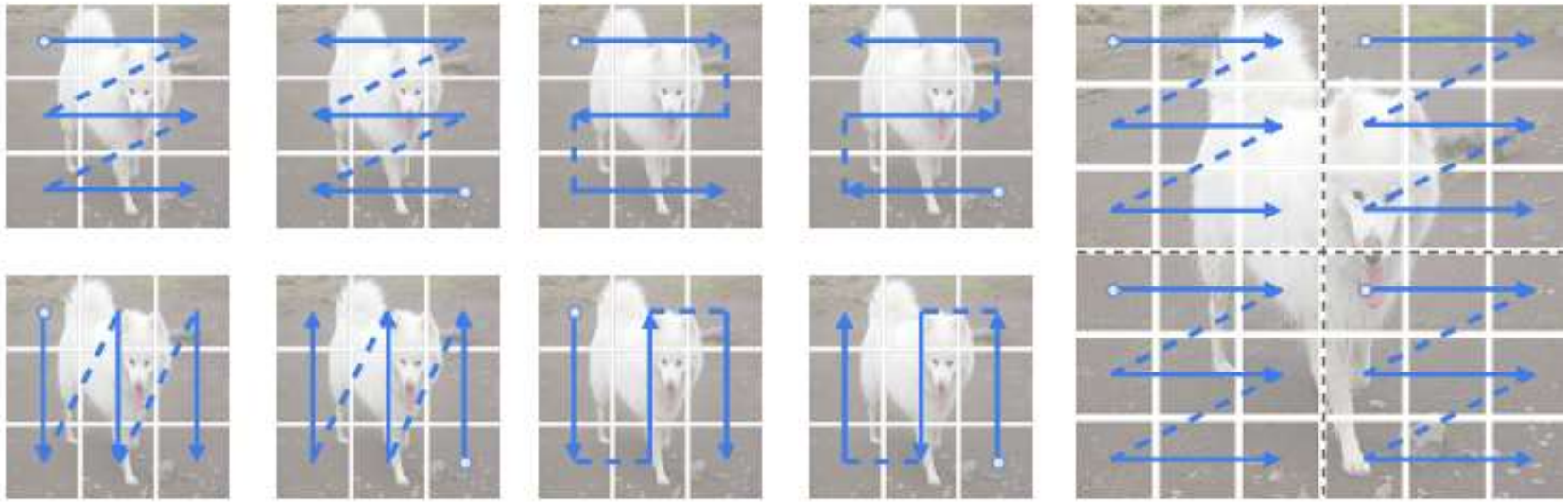
DAMamba: Vision State Space Model with Dynamic Adaptive Scan

Tanzhe Li, Caoshuo Li, Jiayi Lyu, Hongjuan Pei, Baochang Zhang,Taisong Jin, Rongrong Ji

Introduction

- Existing vision SSMs primarily leverage manually designed scans to flatten image patches into sequences locally or globally.
- This approach disrupts the original semantic spatial adjacency of the image and lacks flexibility.

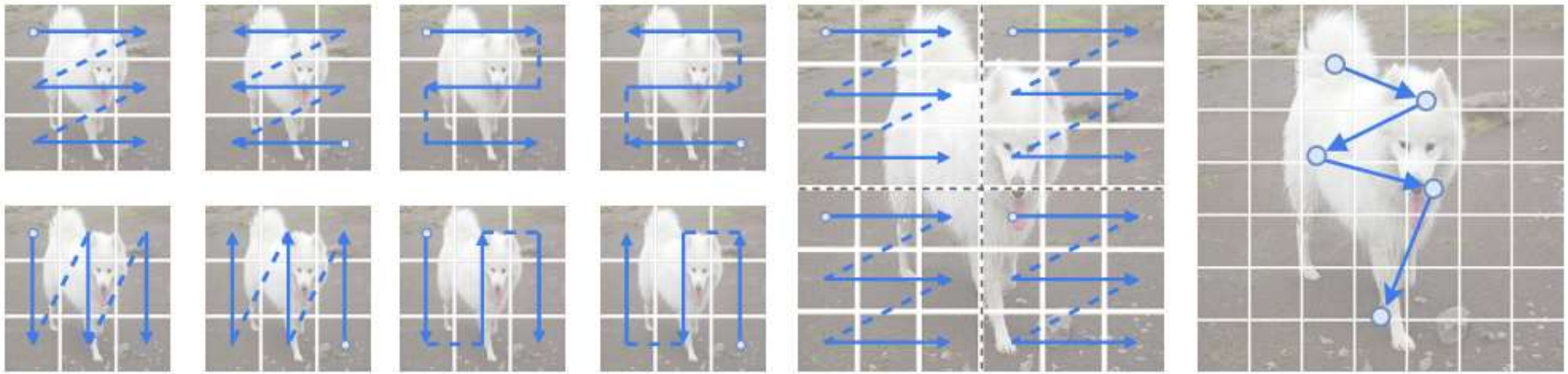


(a) Sweeping Scan (b) Continuous Scan (c) Local Scan

Introduction

- To address this limitation, we propose Dynamic Adaptive Scan (DAS), a data-driven method that adaptively allocates scanning orders and regions.

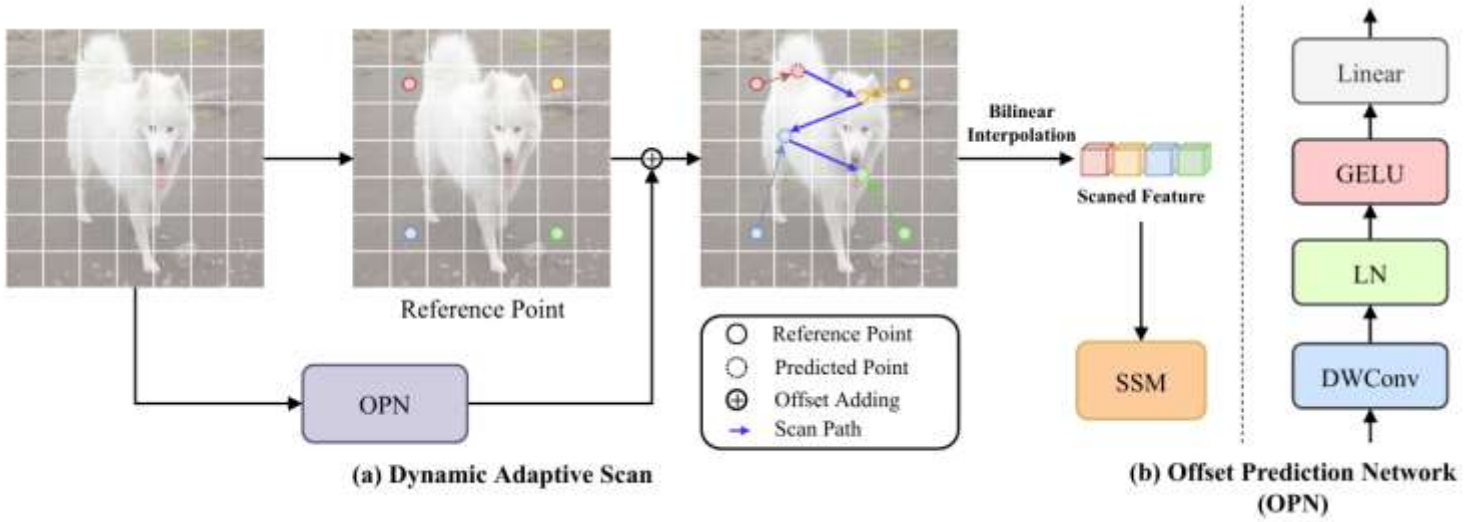
- This enables more flexible modeling capabilities while maintaining linear computational complexity and global modeling capacity.



(a) Sweeping Scan (b) Continuous Scan (c) Local Scan (d) Dynamic Adaptive Scan

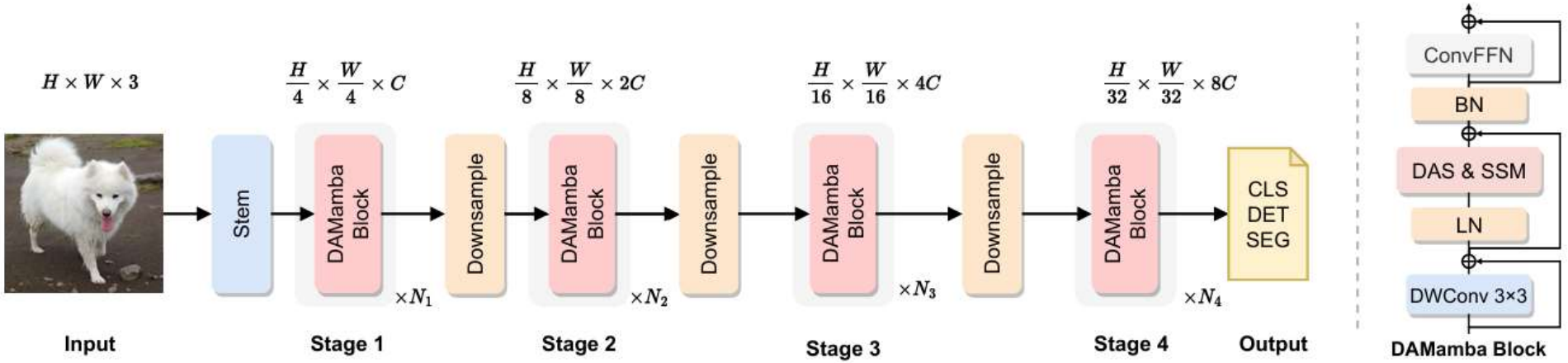
Methodology

- DAS starts by defining a set of learnable positions, with initial values corresponding to the original locations of each patch.
- Then, through a learnable offset prediction network (OPN), a set of offset values is generated for each patch.
- The predicted patches are arranged from top to bottom and left to right based on their original positions, dynamically forming a new sequence



Methodology

- Based on the proposed DAS, we develop a powerful vision Mamba model, termed DAMamba.DAMamba can serve as a versatile vision backbone for various vision tasks.



Models	Channels	Blocks
DAMamba-F	[48, 96, 192, 256]	[2, 2, 10, 2]
DAMamba-T	[80, 160, 320, 512]	[3, 4, 12, 5]
DAMamba-S	[96, 192, 384, 512]	[4, 8, 20, 6]
DAMamba-B	[112, 224, 448, 640]	[4, 8, 25, 8]

Image Classification on ImageNet-1K

XIAMEN UNIVERSITY

Table 2: Results of DAMamba and the current state-of-the-art backbones on ImageNet-1K. All the models are trained and tested at 224 × 224 resolution.

Model	Type	Params (M)	FLOPs (G)	Top-1 (%)
ConvNeXt V2-F [57]	CNNs	5	0.8	78.0
Vim-Ti [69]	SSMs	7	1.5	76.1
LocalVim-Ti [69]	SSMs	8	1.5	76.2
EffcientVMamba-T [41]	SSMs	6	0.8	76.5
DAMamba-F (ours)	SSMs	6	1.3	**79.1**
SLaK-T [34]	CNNs	30	5.0	82.5
ConvNeXt V2-T [57]	CNNs	29	4.5	82.5
InceptionNeXt-T [62]	CNNs	28	4.2	82.3
MambaOut-Tiny [61]	CNNs	27	4.5	82.7
UniRepLKNet-T [8]	CNNs	31	4.9	83.2
Swin-T [36]	ViTs	29	4.5	81.3
CSwin-T [9]	ViTs	23	4.3	82.7
Agent-Swin-T [15]	ViTs	29	4.5	82.6
DAT-T [58]	ViTs	29	4.6	82.0
PVTv2-B2 [55]	ViTs	26	4.0	82.0
ClusterFormer-Tiny [32]	ViTs	28	-	81.5
Slide-PVT-S [39]	ViTs	23	4.0	81.7
NAT-T [16]	ViTs	28	4.3	83.2
QFormer$_A$-T [66]	ViTs	29	4.6	82.5
PartialFormer-B3 [52]	ViTs	36	3.4	83.0
StructViT-S-8-1 [28]	ViTs	24	5.4	83.3
Vim-S [69]	SSMs	26	5.1	80.5
VMamba-T [35]	SSMs	30	4.9	82.6
PlainMamba-L2 [60]	SSMs	25	8.1	81.6
LocalVMamba-T [25]	SSMs	26	5.7	82.7
EffcientVMamba-T [41]	SSMs	33	4.0	81.8
DAMamba-T (ours)	SSMs	26	4.8	**83.8**
SLaK-S [34]	CNNs	55	9.8	83.8
InceptionNeXt-S [62]	CNNs	49	8.4	83.5
MambaOut-Small [61]	CNNs	48	9.0	84.1
UniRepLKNet-S [8]	CNNs	56	9.1	83.9
Swin-S [36]	ViTs	50	8.7	83.0
Agent-Swin-S [15]	ViTs	50	8.7	83.7
NAT-S [16]	ViTs	51	7.8	83.7
PVTv2-B4 [55]	ViTs	63	10.1	83.6
DAT-S [58]	ViTs	50	9.0	83.7
ClusterFormer-Small [32]	ViTs	49	-	83.4
QFormer$_A$-S [66]	ViTs	51	8.9	84.0
BiFormer-B [68]	ViTs	57	9.8	84.3
PartialFormer-B4 [52]	ViTs	64	6.8	83.9
StructViT-B-8-1 [28]	ViTs	52	12.0	84.3
TransNeXt-Small [44]	ViTs	50	10.3	84.7
VMamba-S [35]	SSMs	50	8.7	83.6
PlainMamba-L3 [60]	SSMs	50	14.4	82.3
LocalVMamba-S [25]	SSMs	50	11.4	83.7
DAMamba-S (ours)	SSMs	45	10.3	**84.8**
ConvNeXt V2-B [57]	CNNs	89	15.4	84.3
SLaK-B [34]	CNNs	95	17.1	84.0
InceptionNeXt-B [62]	CNNs	87	14.9	84.0
MambaOut-Base [61]	CNNs	85	15.8	84.2
Swin-B [36]	ViTs	88	15.4	83.5
CSwin-B [9]	ViTs	78	15.0	84.2
Agent-Swin-B [15]	ViTs	88	15.4	84.0
NAT-B [16]	ViTs	90	13.7	84.3
PVTv2-B5 [55]	ViTs	82	11.8	83.8
FLatten-Swin-B [14]	ViTs	89	15.4	83.8
DAT-B [58]	ViTs	88	15.8	84.0
QFormer$_A$-B [66]	ViTs	90	15.7	84.1
TransNeXt-Base [44]	ViTs	90	18.4	84.8
VMamba-B [35]	SSMs	89	15.4	83.9
DAMamba-B (ours)	SSMs	86	16.3	**85.2**

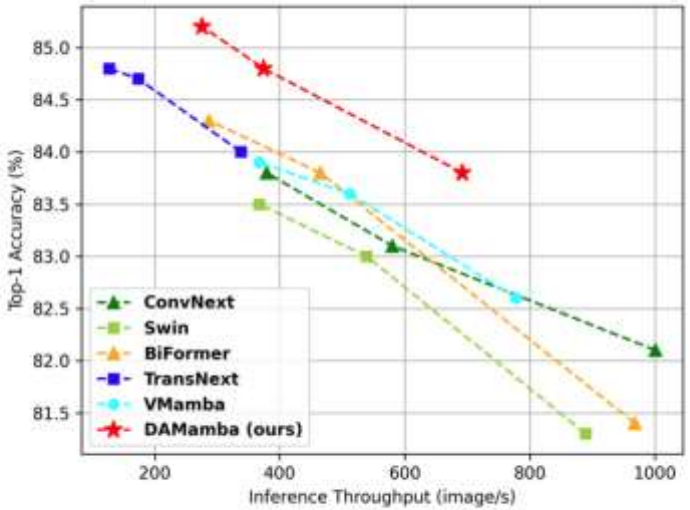


Figure 4: **The trade-off between ImageNet-1K top-1 accuracy and inference throughput.** All the models are trained under the DeiT training hyperparameters. The inference throughput is measured on an NVIDIA RTX 3090 GPU with a batch size 128.

Object Detection and Instance Segmentation on COCO2017

XIAMEN UNIVERSITY

Table 3: Comparison of object detection and instance segmentation performance on COCO with Mask R-CNN detector. FLOPs are calculated with input resolution of 1280 × 800.

Mask R-CNN 1× schedule								
Backbone	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	#Param.	FLOPs
Swin-T	42.7	65.2	46.8	39.3	62.2	42.2	48M	267G
DAT-T	44.4	67.6	48.5	42.4	66.1	45.5	48M	272G
CSWin-T	46.7	68.6	51.3	42.2	65.6	45.4	42M	279G
ConvNeXt-T	44.2	66.6	48.3	40.1	63.3	42.8	48M	262G
PVTv2-B2	45.3	66.1	49.6	41.2	64.2	44.4	45M	309G
QFormer$_h$-T	45.9	68.5	50.3	41.5	65.2	44.6	49M	-
PartialFormer-B3	45.0	-	-	40.9	-	-	54M	248G
BiFormer-S	47.8	69.8	52.3	43.2	66.8	46.5	-	-
MambaOut-T	45.1	67.3	49.6	41.0	64.1	44.1	43M	262G
VMamba-T	47.3	69.3	52.0	42.7	66.4	45.9	50M	271G
LocalVMamba-T	46.7	68.7	50.8	42.2	65.7	45.5	45M	291G
DAMamba-T	**48.5**	**70.3**	**53.3**	**43.4**	**67.2**	**46.7**	45M	284G
Swin-S	44.8	68.6	49.4	40.9	65.3	44.2	69M	354G
Agent-Swin-S	47.2	69.6	52.3	42.7	66.6	45.8	-	364G
DAT-S	47.1	69.9	51.5	42.5	66.7	45.4	69M	378G
CSWin-S	47.9	70.1	52.6	43.2	67.1	46.2	54M	342G
ConvNeXt-S	45.4	67.9	50.0	41.8	65.2	45.1	70M	348G
PVTv2-B3	47.0	68.1	51.7	42.5	65.2	45.7	63M	397G
BiFormer-B	48.6	70.5	53.8	43.7	67.6	47.1	-	-
MambaOut-S	47.4	69.1	52.4	42.7	66.1	46.2	65M	354G
VMamba-S	48.7	70.0	53.4	43.7	67.3	47.0	70M	349G
LocalVMamba-S	48.4	69.9	52.7	43.2	66.7	46.5	69M	414G
DAMamba-S	**49.8**	**71.2**	**54.7**	**44.5**	**68.4**	**48.2**	65M	395G
Swin-B	46.9	69.2	51.6	42.3	66.0	45.5	88M	496G
CSwin-B	48.7	70.4	53.9	43.9	67.8	47.3	88M	496G
ConvNeXt-B	47.0	69.4	51.7	42.7	66.3	46.0	107M	486G
PVTv2-B5	47.4	68.6	51.9	42.5	65.7	46.0	102M	557G
ViT-Adapter-B	47.0	68.2	51.4	41.8	65.1	44.9	102M	557G
MambaOut-B	47.4	69.3	52.2	43.0	66.4	46.3	100M	495G
VMamba-B	49.2	71.4	54.0	44.1	68.3	47.7	108M	485G
DAMamba-B	**50.6**	**71.9**	**55.5**	**44.9**	**68.9**	**48.7**	105M	520G

Table 6: Comparison of object detection and instance segmentation performance on COCO with Mask R-CNN detector. FLOPs are calculated with input resolution of 1280 × 800.

Mask R-CNN 3× MS schedule								
Backbone	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	#Param.	FLOPs
Swin-T	46.0	68.1	50.3	41.6	65.1	44.9	48M	267G
PVTv2-B2	47.8	69.7	52.6	43.1	66.8	46.7	45M	309G
ConvNeXt-T	46.2	67.9	50.8	41.7	65.0	44.9	48M	262G
NAT-T	47.7	69.0	52.6	42.6	66.1	45.9	48M	258G
QFormer$_h$-T	47.5	69.6	52.1	42.7	66.4	46.1	49M	-
VMamba-T	48.8	70.4	53.5	43.7	67.4	47.0	50M	271G
LocalVMamba-T	48.7	70.1	53.0	43.4	67.0	46.4	45M	291G
DAMamba-T	**50.4**	**71.4**	**55.5**	**44.8**	**68.6**	**48.6**	45M	284G
Swin-S	48.2	69.8	52.8	43.2	67.0	46.1	69M	354G
PVTv2-B3	48.4	69.8	53.3	43.2	66.9	46.7	65M	397G
ConvNeXt-S	47.9	70.0	52.7	42.9	66.9	46.2	70M	348G
NAT-S	48.4	69.8	53.2	43.2	66.9	46.5	70M	330G
QFormer$_h$-S	49.5	71.2	54.2	44.2	68.3	47.6	70M	-G
VMamba-S	49.9	70.9	54.7	44.2	68.2	47.7	70M	349G
LocalVMamba-S	49.9	70.5	54.4	44.1	67.8	47.4	69M	414G
DAMamba-S	**51.2**	**72.1**	**56.1**	**45.1**	**69.2**	**49.1**	65M	395G
ConvNeXt-B	48.5	70.1	53.3	43.5	67.1	46.7	108M	486G
Swin-B	48.6	70.0	53.4	43.3	67.1	46.7	107M	496G
PVTv2-B5	48.4	69.2	52.9	42.9	66.6	46.2	102M	557G
DAMamba-B	**51.4**	**72.3**	**56.4**	**45.3**	**69.5**	**48.9**	105M	520G

Semantic Segmentation on ADE20K

XIAMEN UNIVERSITY

Table 4: Comparison of semantic segmentation on ADE20K with UPerNet segmentor. FLOPs are calculated with input resolution of 512 × 2048. 'SS' and 'MS' represent single-scale and multi-scale testing, respectively.

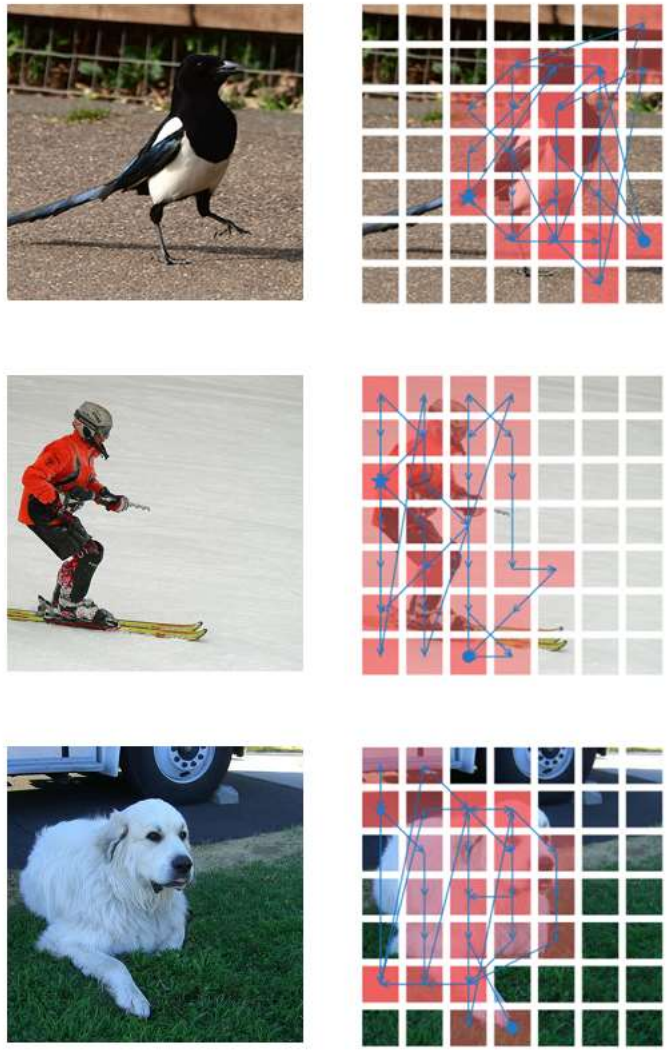
Method	mIoU (SS)	mIoU (MS)	#Param.	FLOPs
UniRepLKNet-T	48.6	49.1	61M	946G
ConvNeXt-T	46.0	46.7	60M	939G
Swin-T	44.4	45.8	60M	945G
Agent-Swin-T	46.7	-	61M	954G
NAT-T	47.1	48.4	58M	934G
QFormer$_h$-T	46.9	48.1	61M	-
PartialFormer-B3	47.0	-	65M	923G
BiFormer-S	49.8	50.8	-	-
MambaOut-T	47.4	48.6	54M	938G
VMamba-T	48.0	48.8	62M	949G
LocalVMamba-T	47.9	49.1	57M	970G
DAMamba-T	**50.3**	**51.2**	55M	937G
UniRepLKNet-S	50.5	51.0	86M	1036G
Swin-S	47.6	49.5	81M	1039G
Agent-Swin-S	48.1	-	81M	1043G
ConvNeXt-S	48.7	49.6	82M	1027G
NAT-S	48.0	49.5	82M	1010G
QFormer$_h$-S	48.9	50.3	82M	-
PartialFormer-B3	48.3	-	95M	1005G
BiFormer-B	51.0	51.7	-	-
MambaOut-S	49.5	50.6	76M	1032G
VMamba-S	50.6	51.2	82M	1028G
LocalVMamba-S	50.0	51.0	81M	1095G
DAMamba-S	**51.2**	**52.0**	75M	1050G
Swin-B	48.1	49.7	121M	1188G
Agent-Swin-B	48.7	-	121M	1196G
ConvNeXt-B	49.1	49.9	122M	1170G
NAT-B	48.5	49.7	123M	1137G
QFormer$_h$-B	49.5	50.6	123M	-
MambaOut-B	49.6	51.0	112M	1178G
VMamba-B	51.0	51.6	122M	1170G
DAMamba-B	**51.9**	**52.3**	117M	1178G

Table 5: Ablation studies on DAMamba-F for module designs.

Module design	#Param. (M)	FLOPs (G)	Top-1 acc (%).
Baseline	5.31M	1.20G	77.7
+ DAScan	5.41M	1.23G	78.3
+ Convpos	5.43M	1.24G	78.6
+ ConvFFN	5.52M	1.26G	79.1

Visualization

XIAMEN UNIVERSITY



(a) Input image. (b) Scan Path.

Figure 5: Visualization of the Dynamic Adaptive Scan, where the blue pentagram represents the start of the scan and the blue circle represents the end of the scan.

Thanks for listening!