SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer

CVPR2023

胡天乐

2023/07/04

目录

- **作者介绍**
- 研究动机
- 本文方法
- 实验效果
- 总结反思

作者介绍

Xuanyao Chen

Fudan University
在 fudan.edu.cn 的电子邮件经过验证 - 首页
Computer Vision Machine Learning

关注

标题	引用次数	年份
What makes multi-modal learning better than single (provably) Y Huang, C Du, Z Xue, X Chen, H Zhao, L Huang Advances in Neural Information Processing Systems 34, 10944-10956	84	2021
FUTR3D: A Unified Sensor Fusion Framework for 3D Detection X Chen, T Zhang, Y Wang, Y Wang, H Zhao Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern …	56	2022
Mutr3d: A multi-camera tracking framework via 3d-to-2d queries T Zhang, X Chen, Y Wang, Y Wang, H Zhao Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern …	18	2022
Vip3d: End-to-end visual trajectory prediction via 3d agent queries J Gu, C Hu, T Zhang, X Chen, Y Wang, Y Wang, H Zhao IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	10	2022
SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer X Chen, Z Liu, H Tang, L Yi, H Zhao, S Han IEEE Conference on Computer Vision and Pattern Recognition (CVPR)		2023

作者介绍

Hang Zhao

关注

Assistant Professor, Tsinghua University
在 csail.mit.edu 的电子邮件经过验证 - 首页

Multimodal Learning Autonomous Driving Robotics Computer Vision

标题	引用次数	年份
Scene parsing through ade20k dataset B Zhou, H Zhao, X Puig, S Fidler, A Barriuso, A Torralba Proceedings of the IEEE conference on computer vision and pattern …	2213	2017
Loss functions for image restoration with neural networks H Zhao, O Gallo, I Frosio, J Kautz IEEE Transactions on Computational Imaging 3 (1), 47-57	2116 *	2017
Scalability in perception for autonomous driving: Waymo open dataset P Sun, H Kretzschmar, X Dotiwalla, A Chouard, V Patnaik, P Tsui, J Guo, ... Proceedings of the IEEE/CVF conference on computer vision and pattern …	1570	2020
Semantic understanding of scenes through the ade20k dataset B Zhou, H Zhao, X Puig, T Xiao, S Fidler, A Barriuso, A Torralba International Journal of Computer Vision 127, 302-321	1200	2019
Through-wall human pose estimation using radio signals M Zhao, T Li, M Abu Alsheikh, Y Tian, H Zhao, A Torralba, D Katabi Proceedings of the IEEE conference on computer vision and pattern …	500	2018

- 作者介绍
- **研究动机**
- 本文方法
- 实验效果
- 总结反思

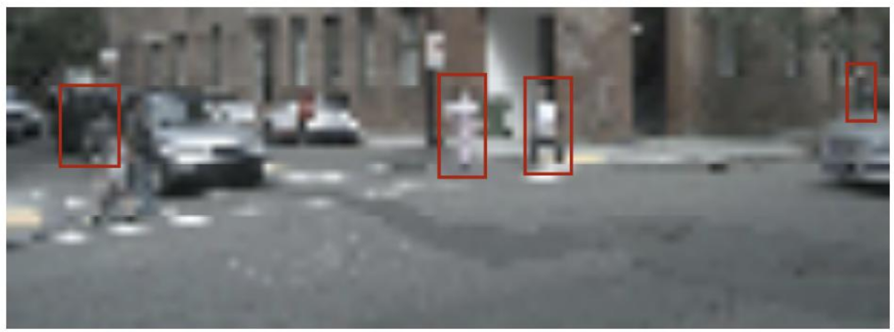
研究动机

- 近年来，Transformer 架构在计算机视觉的各项任务中都表现出令人惊艳的性能。然而，对于高分辨率图像，其计算量较大，也无法在通用硬件上进行有效的部署。

- 最简单常用的方法就是降低图像分辨率， 但这将使得模型丢失高分辨率传感器捕获的细节信息，损失模型性能。

研究动机

- 直接下采样会造成信息丢失
- activation pruning，但激活稀疏性不能轻易转化为实际加速 CNNs 的通用硬件 (如 GPU)

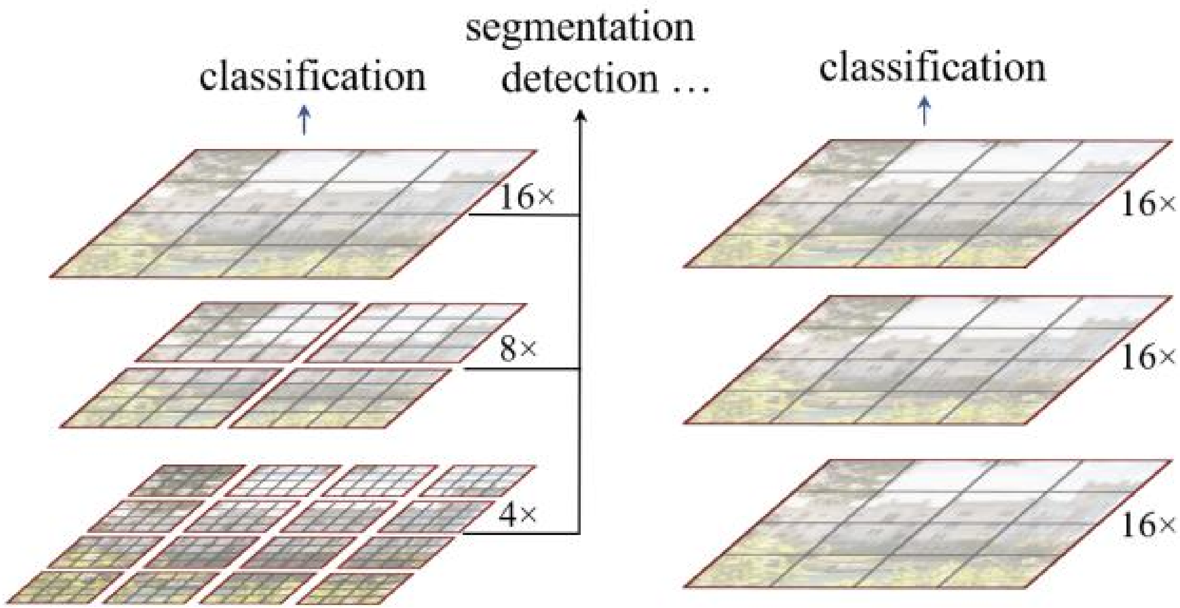


(a) **Direct Downsample**: Lower Resolution (**0.5x**), Dense (**100%**)

(b) **Window Activation Pruning**: Higher Resolution (**1.0x**), Sparse (**25%**)

研究动机

- Swin Transformer：各窗口重要性相同，计算量大
- 与卷积不同，窗口注意力是在窗口上自然分批的，这使得通过窗口级激活修剪实现真正的加速成为可能



(a) Swin Transformer (ours)　　(b) ViT

研究动机

- 本文提出了一种名为 SparseViT 的方法，它可以在保持分辨率不变的基础上，减少基于窗口的ViT的计算复杂性。

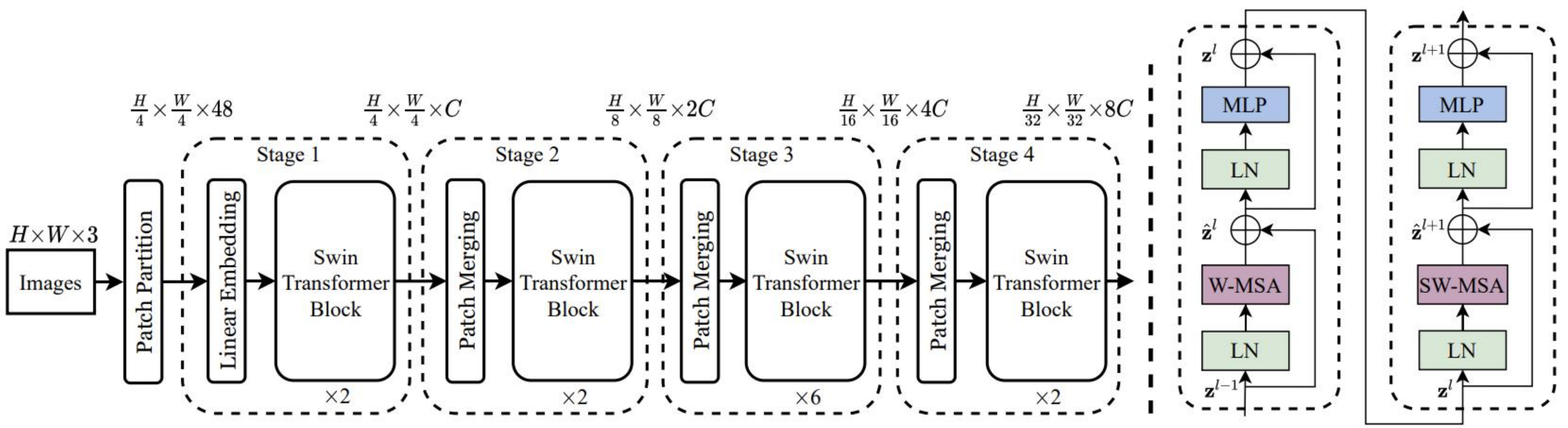
- 通过对不同层分配不同的修剪比例，可以在 60% 的稀疏率下实现 50% 的延迟降低。

- SparseViT 还应用了进化搜索来有效地找到最优的层次稀疏配置。

- 作者介绍
- 研究动机
- **本文方法**
- 实验效果
- 总结反思

Swin Transformer

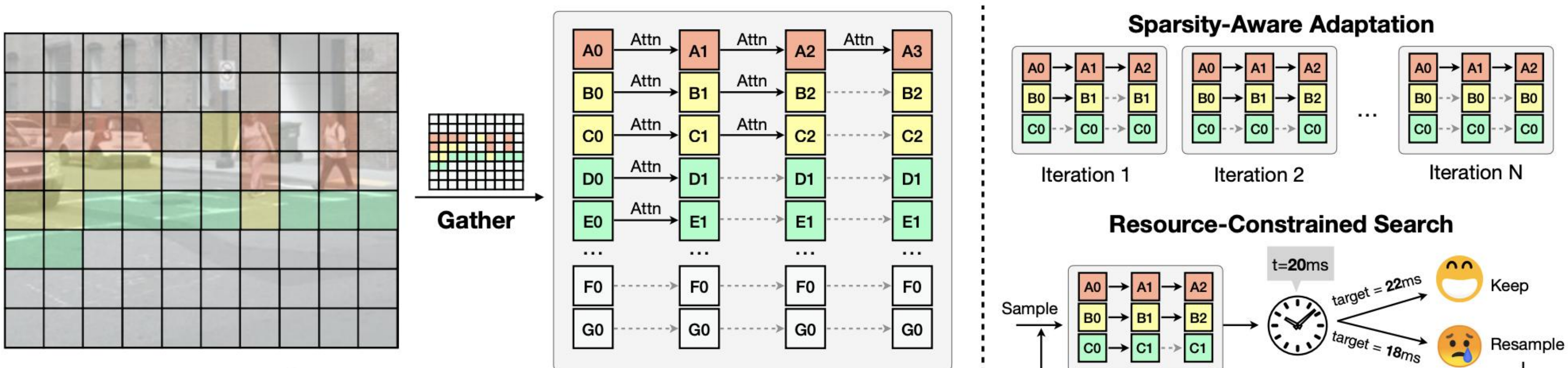
- MHSA(multi-head self-attention)以窗口为单位
- FFN(feed-forward layer)和LN(layer normalization)修改为逐窗口(window-wise)执行



(a) Architecture (b) Two Successive Swin Transformer Blocks

Window Activation Pruning

- 计算每个窗口激活的L2范数作为重要性值
- 从值最高的窗口收集特征进行self-attention
- 复制未选中窗口的特征以保留信息

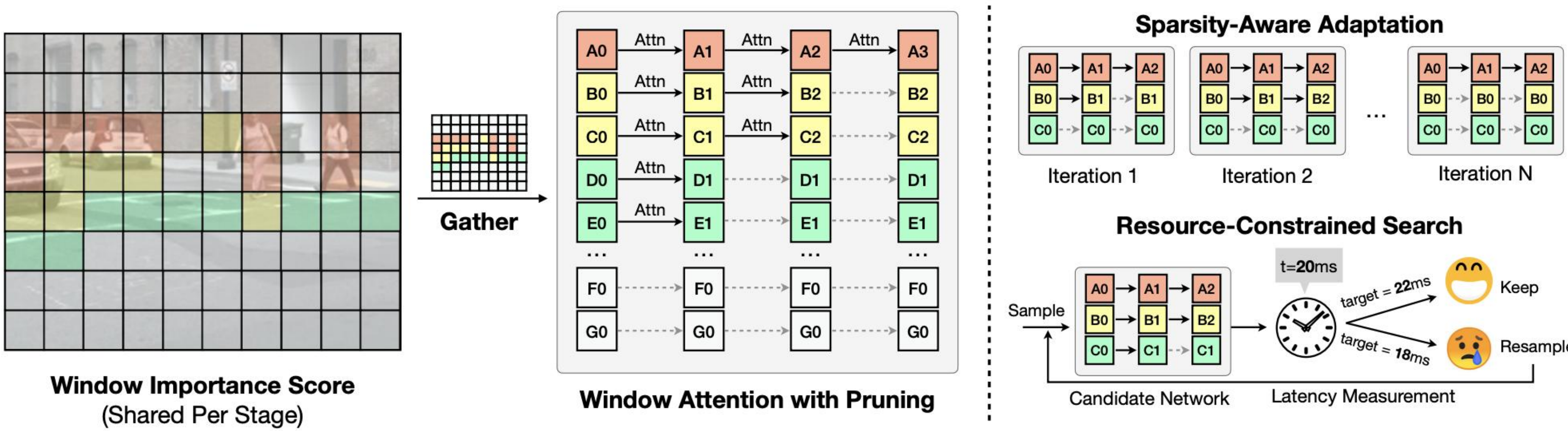


Window Importance Score (Shared Per Stage)

Window Attention with Pruning

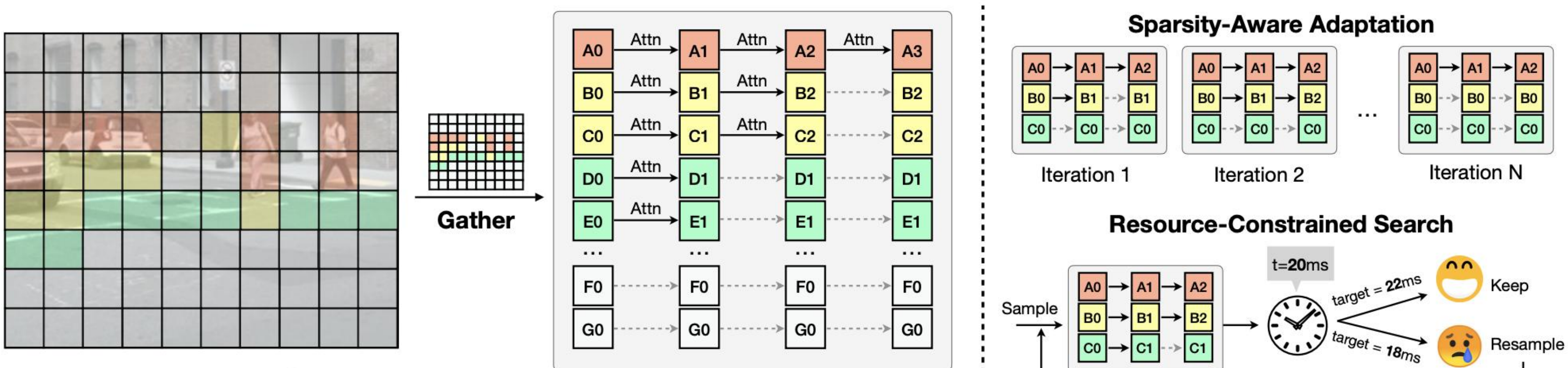
Mixed-Sparsity Configuration Search

- 不同层对稀疏性要求不同
- 每个Swin块从{0%，10%，…，80%}中选择稀疏率，且每层不下降



Mixed-Sparsity Configuration Search

- 每次迭代时对不同层的激活稀疏性进行采样
- 利用evolutionary search探索最佳分层稀疏配置
- finetuning得到的配置直至收敛



Window Importance Score
(Shared Per Stage)
Window Attention with Pruning

- 作者介绍
- 研究动机
- 本文方法
- **实验效果**
- 总结反思

实验效果

- 3D目标检测消融实验
- 分别降低分辨率和宽度，性能均不如SparseViT

Backbone	Resolution	Width	#MACs (G)	Latency (ms)	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
Swin-T	256×704	1×	140.8	36.4	31.2	69.1	27.2	52.3	90.9	24.7	39.2
SparseViT (Ours)	288×792	1×	113.9	34.5	**32.0**	72.8	27.2	53.8	79.4	25.7	**40.1**
Swin-T (R224)	224×616	1×	78.5	23.0	29.9	71.8	27.4	60.9	79.0	26.0	38.4
Swin-T (W0.6×)	256×704	0.6×	56.0	22.6	29.9	69.9	27.5	59.9	81.4	25.8	38.5
SparseViT (Ours)	256×704	1×	78.4	23.8	**31.2**	70.9	27.5	58.7	83.1	27.2	**38.9**
Swin-T (R192)	192×528	1×	67.1	18.7	28.7	74.3	27.9	59.5	76.7	27.8	37.7
Swin-T (W0.4×)	256×704	0.4×	20.4	17.6	27.6	74.2	27.9	63.4	91.0	26.2	35.5
SparseViT (Ours)	256×704	1×	58.6	18.7	**30.0**	72.0	27.5	59.7	81.7	26.6	**38.3**

Table 1. Results of monocular 3D object detection on nuScenes.

实验效果

- 2D实例分割消融实验
- 在各种输入分辨率下，SparseViT 的计算量始终优于 baseline

Backbone	Resolution	Width	#MACs (G)	Latency (ms)	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Swin-T	640×640	1×	161.8	46.6	42.0	63.3	45.7	38.3	60.3	40.9
Swin-T (R576)	576×576	1×	149.5	41.3	41.0	62.1	44.9	37.2	59.0	39.6
Swin-T (W0.9×)	640×640	0.9×	122.3	41.8	40.4	61.9	43.8	37.1	58.9	39.8
SparseViT (Ours)	672×672	1×	139.5	41.3	**42.4**	63.3	46.4	**38.5**	60.3	41.3
Swin-T (R544)	544×544	1×	119.8	34.8	40.5	61.2	43.8	36.8	58.2	39.1
Swin-T (W0.8×)	640×640	0.8×	90.5	35.9	39.4	60.7	42.8	36.4	57.9	38.8
SparseViT (Ours)	672×672	1×	116.5	34.1	**41.6**	62.5	45.5	**37.7**	59.4	40.2
Swin-T (R512)	512×512	1×	117.5	32.9	39.6	60.1	43.4	36.0	57.0	38.2
Swin-T (W0.6×)	640×640	0.6×	63.4	31.7	38.7	60.2	41.6	35.7	57.0	38.0
SparseViT (Ours)	672×672	1×	105.9	32.9	**41.3**	62.2	44.9	**37.4**	59.1	39.7

Table 2. Results of 2D instance segmentation on COCO.

实验效果

- 2D语义分割消融实验
- 精度与分辨率几乎不变，能达到1.3倍的速度

Backbone	Resolution	Latency (ms)	mIoU
Swin-L	1024×2048	329.5	83.3
Swin-L (R896)	896×1792	256.5	82.8
SparseViT (Ours)	1024×2048	250.6	**83.2**

Table 3. Results of 2D semantic segmentation on Cityscapes.

实验效果

- SparseViT 提供了比 baseline 更好的精度–效率 trade–off

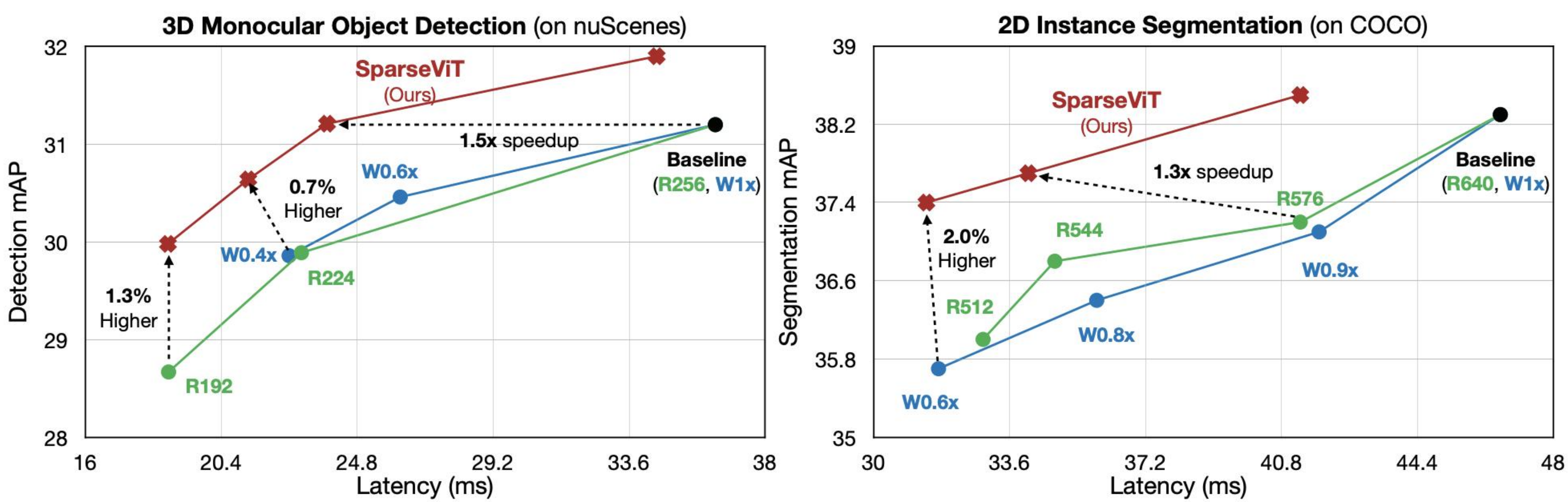


Figure 3. SparseViT delivers a significantly better accuracy-efficiency trade-off than the baselines with reduced resolutions and widths on monocular 3D object detection (**left**) and 2D instance segmentation (**right**).

实验效果

- 修剪效果可视化

Figure 5. SparseViT effectively prunes irrelevant background windows while retaining informative foreground windows. Each window's color corresponds to the number of layers it is executed. Brighter colors indicate that the model has executed the window in more layers.

- 作者介绍
- 研究动机
- 本文方法
- 实验效果
- **总结反思**

总结反思

- 本文在Swin-Transformer的基础上提出了一种新的方法：SparseViT

- 采用窗口激活修剪，引入稀疏性感知自适应，使用进化搜索来找到最佳的分层稀疏配置

- SparseViT在单目3D对象检测、2D实例分割和2D语义分割中分别实现了1.5倍、1.4倍和1.3倍的测量加速，同时几乎不损精度