Scene Adaptive Sparse Transformer for Event-based Object Detection

Yansong Peng[1] Hebei Li[1] Yueyi Zhang[1] Xiaoyan Sun[1,2] Feng Wu[1,2]

[1]University of Science and Technology of China

[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

{pengyansong, lihebei}@mail.ustc.edu.cn, {zhyuey, sunxiaoyan, fengwu}@ustc.edu.cn

1. Video Detection & Sparsification Results

In the accompanying multimedia file, titled **`video.mp4`**, we present visualizations corresponding to several event clips in the test set of 1Mpx. This video includes comparisons between ground truth and SAST's object detection results, showcasing SAST's high performance. It also features the visualizations of score heatmaps and selection results across different scenes, providing a clear demonstration of SAST's scene-aware adaptability. From the video, it can be observed that SAST assigns higher scores to important tokens within important windows and performs a series of operations such as self-attention, MLP, and normalization exclusively on these sparse tokens, significantly reducing computational costs.

2. Additional Experiments

2.1. Sparsity Level of SAST.

We adjust the hyper-parameters a and b to limit the sparsification of SAST and SAST-CB, resulting in 10 sparsity levels. The performance of 20 networks is illustrated in Fig. 1. SAST and SAST-CB respectively excel at sparser and denser sparsity levels, but both achieve the best results at a moderate sparsity level. We interpret these results from an information perspective, where higher information density (more effective interactions among fewer tokens) has been shown to benefit the Transformers [2]. SAST-CB achieves this by broadcasting information among selected tokens, enhancing effective information interactions within a reduced token set. However, an overly sparse network can lead to information loss due to excessive compression. Therefore, the choice between SAST and SAST-CB depends on the specific requirements of the task at hand: SAST for lower computational load and SAST-CB for enhanced detection at a slightly higher computational cost.

2.2. Weighting Method Ablation.

In Tab. 1, we conduct a comparative analysis of various functions used in the STP weighting process for transitive

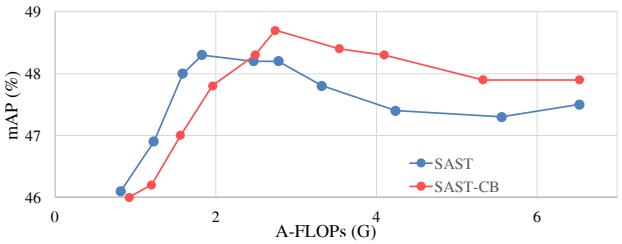

Figure 1. Setting different hyper-parameters results in different sparsity levels and performance of SAST and SAST-CB on 1Mpx. Performance does not continuously improve with increasing sparsity levels. SAST and SAST-CB each have their advantages in sparser and denser settings.

	1Mpx		Gen1	
Functions	mAP (%)	A-FLOPs (G)	mAP (%)	A-FLOPs (G)
Identity	46.5	2.3	46.1	1.2
SoftMax	40.9	1.5	42.0	0.7
Tanh	47.0	1.7	46.7	0.8
Sigmoid	**48.3**	1.8	**47.9**	0.8

Table 1. Detection performance on 1Mpx and Gen1 using different weighting functions. Sigmoid achieves the optimal results.

derivatives. The Sigmoid function, which smoothly maps scores to weights within the range of 0 to 1, assigning higher weights to more significant tokens, delivers superior performance on both datasets.

2.3. Bigger Model, Bigger Gain.

We scale up the RVT and SAST by increasing the layer count in the first, second, and fourth blocks by a factor of two and in the third block by a factor of six, producing the larger RVT-L and SAST-L variants. As depicted in Tab. 2, training these larger models on the 1Mpx and Gen1 datasets both result in performance gains. However, for RVT-L, the proportion of A-FLOPs within the total FLOPs significantly increases, which echoes the discussions of model scalability in the section ***Introduction***. At a comparable model size, SAST-L exhibits more pronounced benefits from its

		1Mpx		Gen1		
Methods	Backbone	mAP (%)	FLOPs (G)	mAP (%)	FLOPs (G)	Params (M)
RVT-L [1]	MaxViT-L [3]	47.8	23.2 (19.3)	47.6	7.8 (6.6)	33.2
Ours	SAST-L	**49.2** (+1.4)	**7.5** (**3.7**, -81%)	**48.6** (+1.0)	**2.9** (**1.7**, -74%)	33.6

Table 2. Detection performance on 1Mpx and Gen1 by training larger variants of RVT and SAST.

adaptive sparsification. It achieves an impressive mAP of 49.2% on 1Mpx and 48.6% on Gen1 datasets, with even fewer FLOPs than the pre-scaled RVT-B. This fully demonstrates the potential of our proposed sparsification method in achieving a remarkable balance between performance and computational cost for large models.

3. Additional Visualizations

We extend our visualization analysis to the 1Mpx and Gen1 datasets, as shown in Fig. 2 and Fig. 3. These visualizations encompass the original event data, score heatmaps, and the selection results of the windows and tokens. The supplementary visualizations reinforce our findings from the main text, providing further evidence of the network's scene-aware adaptability in assigning higher scores to important tokens and adjusting sparsity levels in response to the scene complexity.

4. Additional Implementation Details.

In Tab. 3, we list the default choices of key hyper-parameters, facilitating the replication of our study. These parameters are shared for both the 1Mpx and Gen1 datasets.

a	b	p	Batch Size	Steps	Learning Rate
0.0002	0.099	1.0	32	600000	0.00056

Table 3. Default hyper-parameters used for training SAST and SAST-CB on 1Mpx and Gen1.

References

[1] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *CVPR*, 2023. 2

[2] Nam Hyeon-Woo, Kim Yu-Ji, Byeongho Heo, Doonyoon Han, Seong Joon Oh, and Tae-Hyun Oh. Scratching visual transformer's back with uniform attention. In *ICCV*, 2023. 1

[3] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, 2022. 2

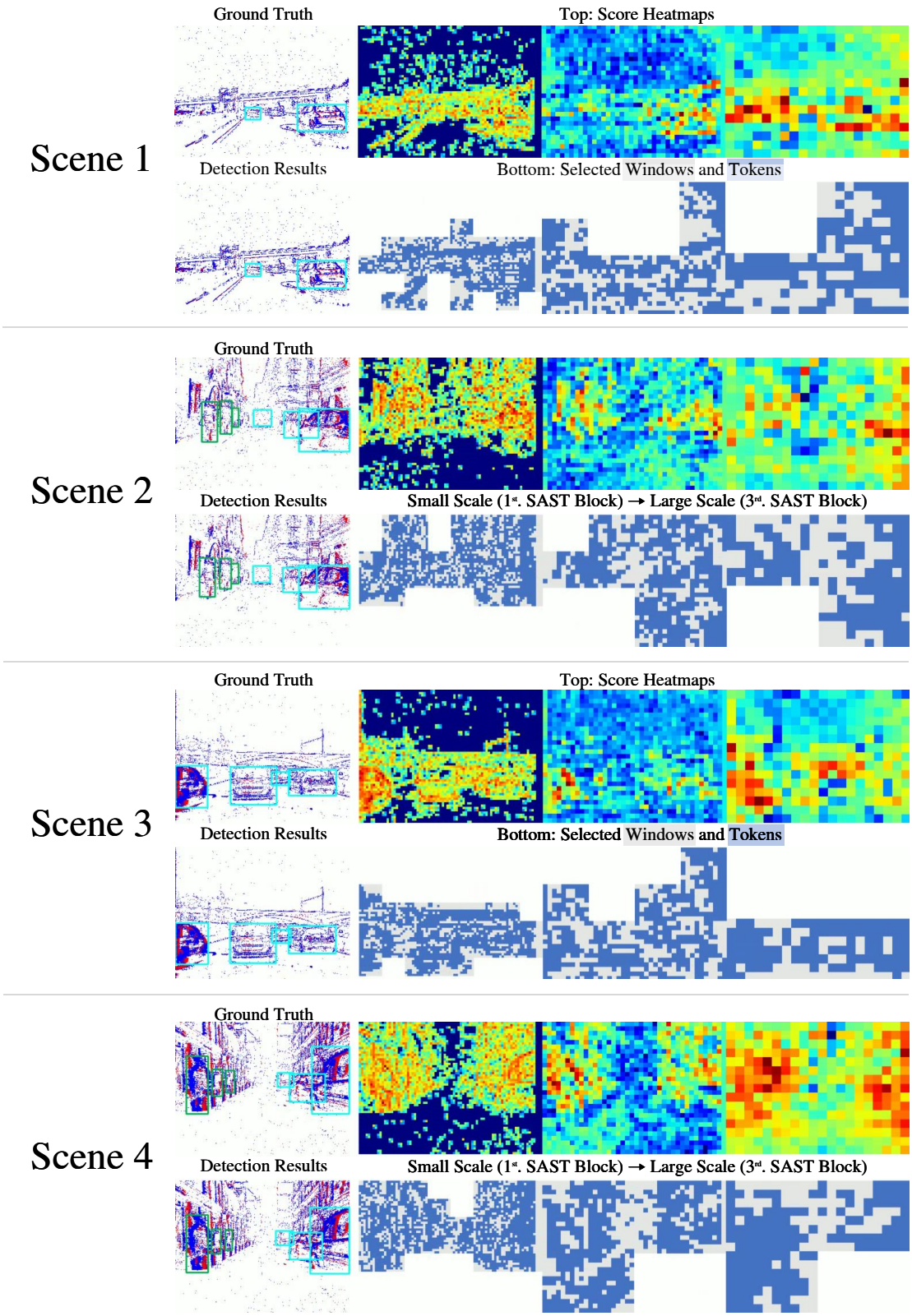

Figure 2. Additional Visualizations of original events, score heatmaps, and selection results under four scenes in Gen1. As the network progresses through subsequent SAST blocks, featuring multiple downsampling stages, the scale (receptive field) of tokens expands.

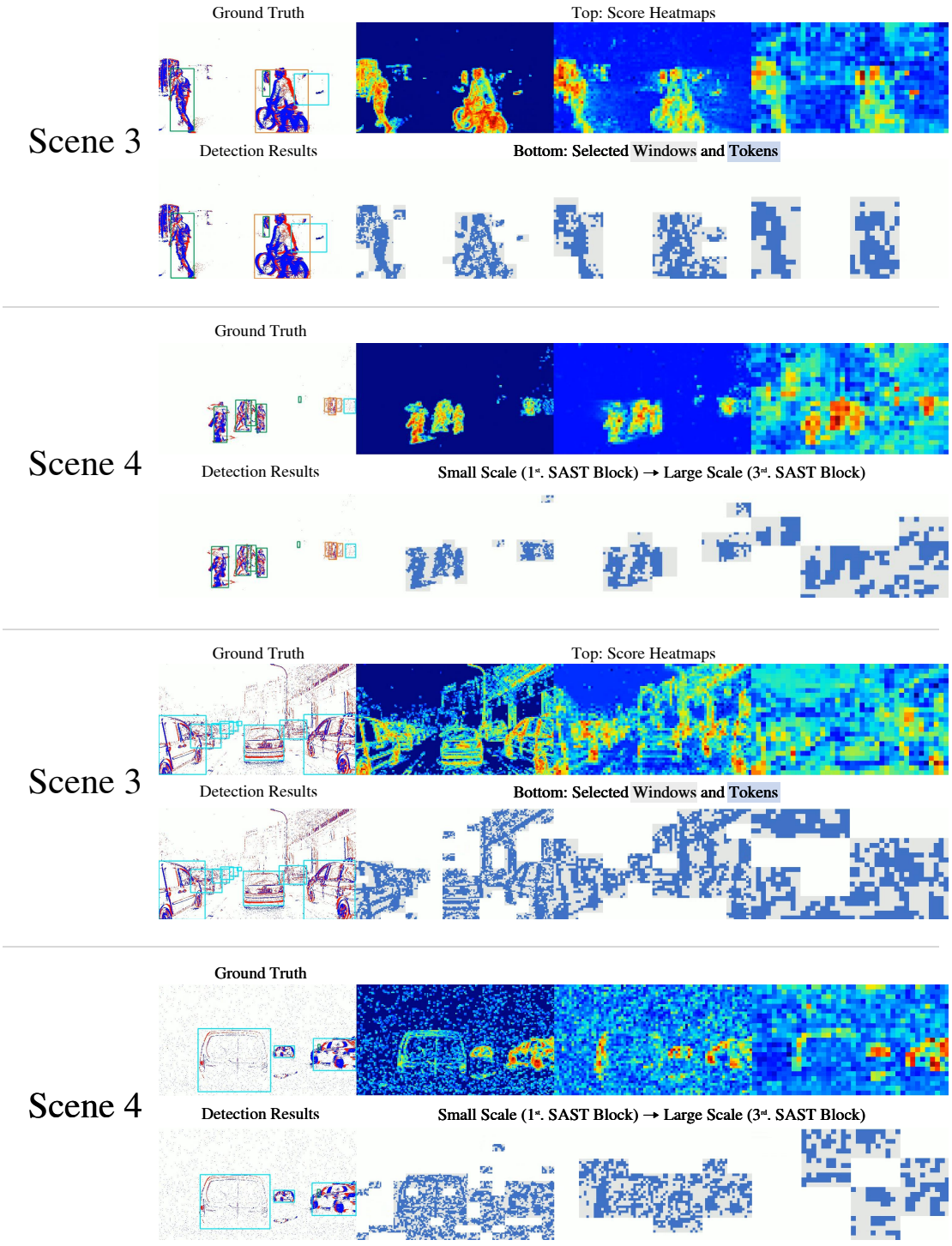

Figure 3. Additional Visualizations of original events, score heatmaps, and selection results under four scenes in 1Mpx. As the network progresses through subsequent SAST blocks, featuring multiple downsampling stages, the scale (receptive field) of tokens expands.