FASTer: Focal Token Acquiring-and-Scaling Transformer for Long-term 3D Object Detection

Supplementary Material

A. Extra results

To ensure a fair comparison with existing two-stage methods, we deliberately employ a suboptimal Sparse-CNN as the backbone in our approach, though the backbone can be substituted with any ones. To further validate the effectiveness of our proposal-based method, we present experimental results in which a powerful transformer is used as the backbone alongside our FASTer (16-frame) as the ROI-Head. Data augmentation techniques, including ground-truth sampling, random flipping, rotation, scaling, and translation, are applied consistently. All other hyper-parameters follow the official settings without modification. We also employ a uniform center-based dense head across experiments.

As presented in Tab. 1, the results reveal that, although substituting the backbone significantly enhances proposal quality, FASTer still demonstrates substantial improvements in bounding box refinement, pushing performance closer to the upper limits achievable in Lidar-based detection tasks. Our experiments reveal that the recall rate of proposals is critical to the final optimization outcome.

B. Some Details

Details of result-level concatenation variants. In this section, we provide the details of the result-level concatenation network, in our ablation experiments. As show in Fig. 1, the Single-frame Sequence Processing is identical to that of FASTer. In the Multi-frame Sequence Processing module, after the initial adaptive scaling and Group Fusion, multiple compressed sequences are fed into the multi-layer temporal-geometric fusion module. Each sequence is then decoded using a standard Transformer decoder and their outputs from all groups are concatenated for detection. Notably, we sequentially employ BiFA in [2] and Cross Attention inspired by [1] for temporal fusion. Experiments demonstrate that the absence of Adaptive Scaling mechanism significantly increases inference latency and adversely affects detection performance. We argue that the isolation of geometric fusion and temporal fusion impedes the flow of global information, resulting in abrupt changes in feature dimensions during result-level concatenation, which restricts the model's learning capacity.

Details of staged training and Extra Point Augmentation. Focal points are acquired through learning and cannot be obtained in advance. During the first three epochs of training, points from the history frame are sampled from the full point cloud. After three epochs, we use the existing model to infer the training set and generate informative points for further training. Following the fifth epoch, this process is repeated to update informative points.

However, this significantly reduces the diversity of the training set and makes the model sensitive to the RPN. Therefore, we select hard proposals (those containing only a few points) for retention. During training, we transform the hard proposals of the two preceding and two succeeding frames to the current time. We then extract points within the transformed proposal regions from the remaining scene points and store them, along with the focal points for training. We refer to this operation as Extra Points Augmentation (EPA), which we believe mitigates FASTer's reliance on the RPN, and enhances the model's generalization on the validation set, as confirmed in. During inference, the EPA is discarded, and only informative points from the SSP are retained.

Details of various token compression methods. We experimented with various token compression methods.

For supervised token selection, a simple prediction head is added to the tokens to predict their scores, with direct supervision from the ground truth. Following [6], a transition threshold η is used to ensure smooth training. Specifically, we define point p as lying within the box obtained by scaling the length, width, and height of Ground truth Box B by a factor of a, the point-box supervision value $\hat{y}$ can be obtained as follows:

$$\hat{y} = \begin{cases} 1 & \text{if } a < 1 - \eta \\ 0 & \text{if } a > 1 + \eta \\ \frac{1+\eta-a}{2\eta} & \text{else} \end{cases} \tag{1}$$

where η, in our implementation, is set to 0.2.

For attention masking implementation, inspired by [5], we obtain the binary score for each token via Gumbel-Softmax[4] technique, while ensuring the differentiability of training. Additionally, a masking operation is applied to the attention map to enable gradient updates. During the inference stage, the attention masking operation is omitted, and tokens are selected based on their scores.

C. Visualization

As illustrated in Fig. 2, we visualize the token scores of selected instances after each dynamic scaling operation. It is observed that the scores of points within the bounding

Method	Stage	ALL (mAPH)		VEH (AP/APH)		PED (AP/APH)		CYC (AP/APH)	
		L1	L2	L1	L2	L1	L2	L1	L2
CenterPoint-4f[8]	1	74.30	68.71	76.68/76.13	69.09/68.58	78.58/75.51	71.25/68.37	72.20/71.28	70.10/69.20
+FASTer	2	81.49	75.92	83.21/82.77	75.81/75.37	85.94/83.35	79.14/76.61	79.09/78.37	76.80/76.10
DSVT-4f [7]	1	81.3	75.6	81.8 / 81.4	74.1 / 73.6	85.6 / 82.8	78.6 / 75.9	80.4 / 79.6	78.1 / 77.3
+FASTer	2	82.99	77.26	84.08/83.71	77.03/77.60	86.88/84.42	80.44/78.16	81.69/80.85	79.93/76.02
Scatterformer-4f[3]	1	81.5	76.5	82.7 / 81.9	75.0 / 74.5	86.5 / 83.7	80.2 / 77.5	79.8 / 79.0	78.5 / 77.4
+FASTer	2	83.55	78.39	84.81/84.38	77.76/77.34	87.65/85.32	81.19/78.88	81.85/80.97	80.03/78.97

Table 1. Extra comparative experiments on the validation set of Waymo Open Dataset.

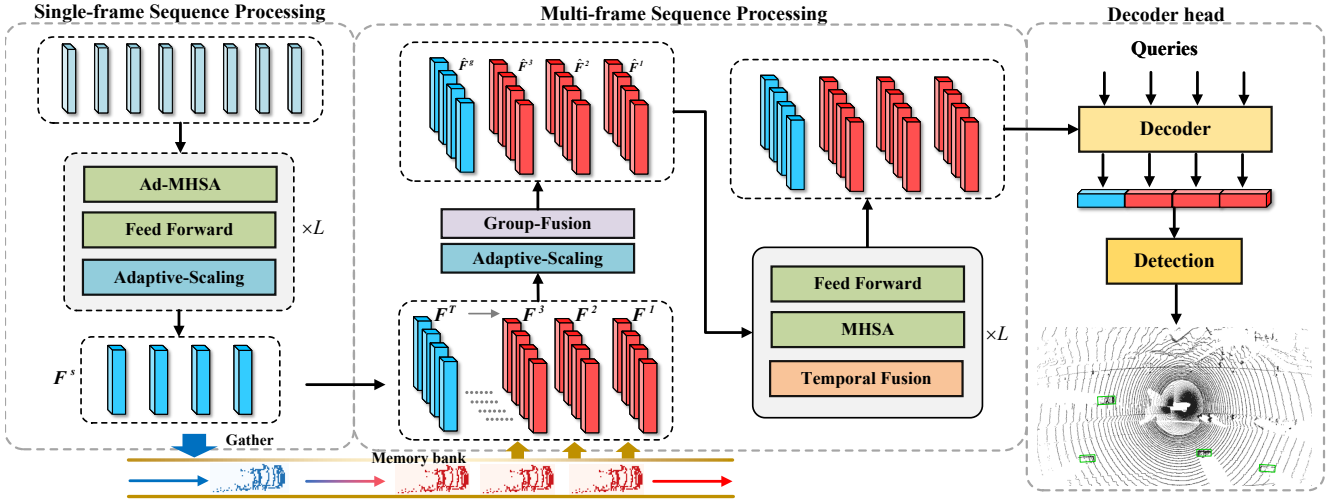

Figure 1. The details of the result-level concatenation network in our ablation experiments. The Single-frame Sequence Processing is identical to that of FASTer. In the Multi-frame Sequence Processing, multiple compressed sequences are fed into the multi-layer temporal-geometric fusion module. Each sequence is then decoded using a Transformer decoder and thehe outputs from all groups are concatenated for detection. Notably, we sequentially employ BiFA in [2] and Cross Attention inspired by [1] for temporal fusion.

box are significantly higher than those outside, which aligns with our general understanding. Interestingly, the score distribution in the initial layer is relatively concentrated and continuous, indicating that the model primarily focuses on local features. As the dynamic scaling progresses through multiple layers, the overall token score distribution becomes increasingly scattered and chaotic. This suggests that the model is gradually shifting its perspective from local details to a more global understanding of the instances.

D. Extra experiments

In this section, we give some other ablation experiments. The mAPH on L1 and L2 are reported in default.

D.1. Extra ablation experiments.

In Tab. 2, we present additional ablation study results. It can be observed that the staged training strategy enables the model to progressively focus on focal tokens from historical frames. Additionally, EPA provides an improvement of approximately 0.25, which is significant for LiDAR-based 3D object detection, further validating our hypothesis.

ST	EPA	mAPH(L1)	mAPH(L2)
✓	✓	81.49	75.92
✓	✗	81.24(-0.25)	75.68(-0.24)
✗	✗	81.13(-0.37)	75.59(-0.33)

Table 2. Ablation experiments. ST and EPA denote the proposed staged training and Extra Points Augmentation strategies, respectively.

D.2. Effects of grouping strategies

We design experiments to investigate the impact of different grouping strategies. As shown in Tab. 3, we observed that, modifying the equidistant grouping to a single-stride grouping results in a significant decrease in detection scores. This finding indicates that our hierarchical grouping and fusion model is far from a mere aggregation of multi-frame point clouds. Instead, its rational and structured fusion strategy enhances the model's ability to extract long-term temporal features effectively.

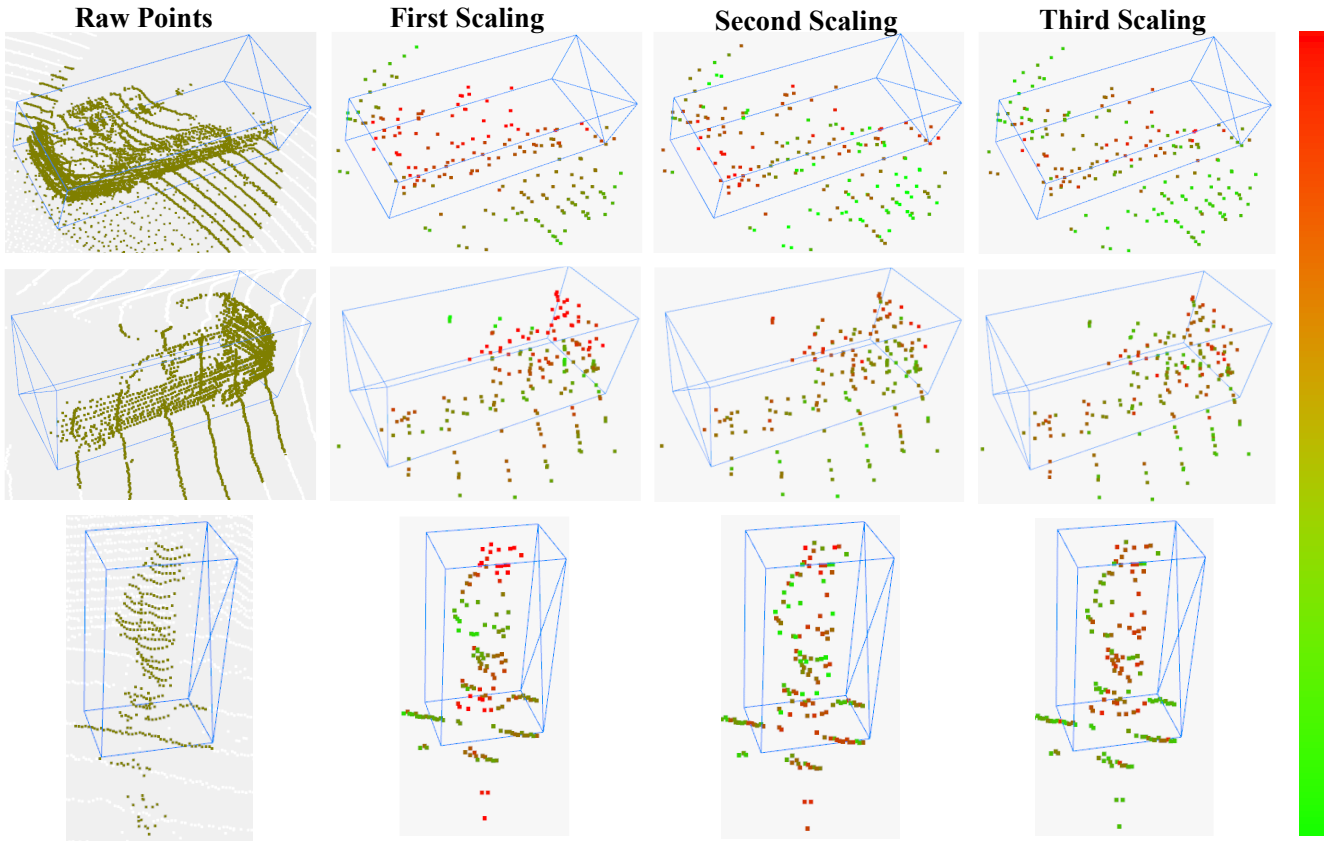

Figure 2. Token scores derived from several instances following different layers of adaptive scaling.

	Strategy 1	Strategy 2	Strategy 3
Group1	$\{1, 5, 9, 13\}$	$\{1, 2, 3, 4\}$	$\{1, 2, 3, 4\}$
Group2	$\{2, 6, 10, 14\}$	$\{5, 6, 7, 8\}$	$\{1, 3, 5, 7\}$
Group3	$\{3, 7, 11, 15\}$	$\{9, 10, 11, 12\}$	$\{1, 4, 7, 10\}$
Group4	$\{4, 8, 12, 16\}$	$\{13, 14, 15, 16\}$	$\{1, 5, 9, 13\}$
mAPH	81.49/75.92	80.92/75.37	81.01/75.44

Table 3. Comparison of different grouping strategies. Each cell contains the index of the corresponding sequence within the complete temporal sequences of its respective group.

β_1 \ β_2	0.6	0.5	0.4	0.3
0.6	81.7 / 76.0	81.3 / 75.6	80.2 / 74.2	78.7 / 74.5
0.5	81.7 / 75.9	81.2 / 75.6	80.2 / 74.2	78.8 / 74.6
0.4	81.2 / 75.4	80.6 / 75.1	79.8 / 74.3	78.2 / 74.0
0.3	80.6 / 74.7	79.9 / 74.5	79.1 / 73.7	77.5 / 73.2

Table 4. Comparison of scaling ratios, β_1 for SSP, and β_2 for MSP.

D.3. Effects of Scaling Ratio

In the main body of our paper, for clarity, we assume that each adaptive scaling operation reduces a sequence to half of its original length, though this reduction ratio can be adjusted, enhancing the flexibility of FASTer. Specifically, let β_1 and β_2 denote the reduction ratios for Ad-MHSA in SSP and MSP respectively. By varying these ratios while maintaining a constant K, we present the comparative experimental results in Tab. 4. We observe that larger scaling ratios generally enhance performance by preserving more information, but excessive values can lead to redundant tokens. Our analysis indicates that the model is more sensitive to β_2 than β_1, as multi-frame sequences inherently encapsulate richer information. This suggests that frame count or point cloud density is a critical factor in determining the model's performance limit.

D.4. Robustness Analysis

To assess the robustness of FASTer under partial data loss, we randomly discard a subset of points or boxes from the historical frames during the inference phase, without additional training. The results are presented in Tab. 5.

Compared to boxes, FASTer exhibits a relatively higher sensitivity to points, while the impact of dropping boxes is minimal. This demonstrates that, within the context of

Drop Rate	Point Drop		Box Drop	
	FASTer	MSF	FASTer	PTT
0	81.49/75.92	80.20/74.62	81.49/75.92	80.20/74.60
0.1	81.15/75.56	79.51/73.40	81.43/75.84	80.03/74.42
0.2	80.80/75.19	78.67/72.73	81.30/75.72	79.82/74.14
0.3	80.30/74.67	77.58/71.54	81.15/75.58	79.58/73.89

Table 5. Comparison of the model's robustness when points or boxes are randomly dropped with a specified probability.

region-guided temporal fusion, points play a more significant role and have a greater influence on the model performance than boxes. We hypothesize that points provide richer semantic information, whereas boxes mainly contribute limited geometric information.

In comparison with other methods, MSF[2], which relies solely on historical boxes, shows greater sensitivity to points, while PTT, which relies solely on historical boxes, exhibits stronger sensitivity to boxes than FASTer. These validate that FASTer maintains robust performance even in the presence of sensor malfunctions and data loss.

References

[1] Xuesong Chen, Shaoshuai Shi, Benjin Zhu, Ka Chun Cheung, Hang Xu, and Hongsheng Li. Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection. In *European Conference on Computer Vision*, pages 680–697. Springer, 2022. 1, 2

[2] Chenhang He, Ruihuang Li, Yabin Zhang, Shuai Li, and Lei Zhang. Msf: Motion-guided sequential fusion for efficient 3d object detection from point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5196–5205, 2023. 1, 2, 4

[3] Chenhang He, Ruihuang Li, Guowen Zhang, and Lei Zhang. Scatterformer: Efficient voxel transformer with scattered linear attention. *arXiv preprint arXiv:2401.00912*, 2024. 2

[4] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 1

[5] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 1

[6] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 1

[7] Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang. Dsvt: Dynamic sparse voxel transformer with rotated sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13520–13529, 2023. 2

[8] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 2