PME3D: An Adaptive and Efficient Multi-modal Feature Extraction Plug-in for 3D Object Detection

310

Tianyi Yu[1], *Lei Yu[2]

[1] School of Computing Science, University of Glasgow
[2] Sino-French Engineer School, Beihang University
***Contact:* 2883406Y@student.gla.ac.uk**

Aim & Motivation

•**Challenge:** Although 3D object detection models deliver high detection accuracy, they often suffer from low real-time inference efficiency, making them less suitable for time-critical applications.

•**Need for Faster Models**: Autonomous vehicles must detect objects rapidly without compromising performance—any delay in detection could lead to serious accidents.

•**Solution:** We address this issue by introducing a plug-in module and a pre-fusion sampler that replace and refine the original feature outputs. This enables more efficient feature fusion within the model, significantly improving inference speed.

Problem Formulation

- **Multi-modal data:** Although 3D object detection models achieve high accuracy, their limited real-time inference speed restricts their deployment in latency-sensitive scenarios.
- **Feature Extractor:** We formalize feature extraction as an efficient vision transformer process, employing an inverted encoder architecture with pixel-level attention. This design reduces computational complexity from $o(T^2 \cdot o^2)$to $o(T^2)$,significantly improving scalability.
- **Depth Estimation, Sampling and Fusion:** Feature fusion is modeled as a depth-guided process aligning BEV features from camera, LiDAR, and Radar via coordinate transformation and adaptive sampling to balance accuracy and efficiency.
- **Loss Function:** The optimization objective integrates focal loss, Smooth L1 loss, and IoU loss, combined with bipartite matching. This ensures balanced gradient flow and consistent training across modalities.

Proposed Method

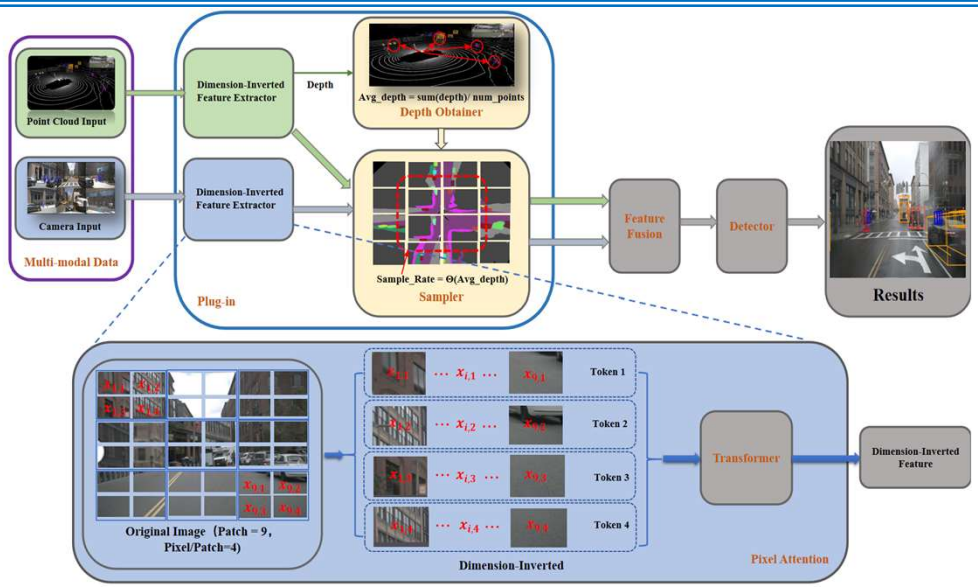

Figure A: The model processes camera, LiDAR, and radar data using an inverted transformer architecture that extracts features along guided dimensions rather than traditional patch-based methods. A depth-guided sampler adaptively selects camera features based on depth information derived from point cloud data, enabling effective multi-modal fusion. This architectural design and adaptive sampling strategy significantly enhance computational efficiency while preserving detection accuracy, setting the approach apart from conventional ViT-based methods.

Qualitative Results

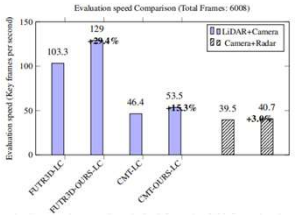

Fig B. Based on the 6008-frame benchmark from the nuScenes test set, our method achieves an inference time of 129 seconds in the LiDAR+Camera setting, representing a **29.4% improvement in efficiency** when integrated with FUTR3D, and a **15.3% gain** when integrated with CMT.

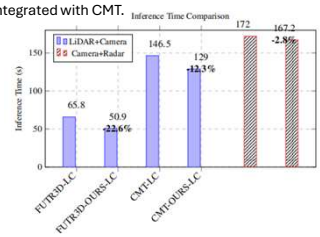

Fig C. This figure compares inference times between baseline models and our enhanced method. Negative percentages indicate performance improvements. In the LiDAR+Camera configuration, **FUTR3D-OURS runs 22.6% faster** than the original FUTR3D, where "OURS" refers to the baseline augmented with our plug-in module.

Results (Speed & Performance)

Method	FPS-500	FPS-3500	FPS-5000	FPS-full
UVTR [23]	-	-	-	1.77
SparseFusion [24]	-	-	-	4.38
BEVFusion [30]	-	-	-	9.58
TransFusion [1]	-	-	-	6.51
DAL-Base [18]	-	-	-	10.69
CMT-R50 [45]	-	-	-	10.72
CMT-OURS	15.82	15.5	16.44	**24.90**

Figure D: The table compares the runtime efficiency (FPS) of various fusion models under different input conditions. Only **CMT-OURS** was evaluated across all four FPS scenarios and consistently outperformed other methods, achieving the highest overall speed at **24.90 FPS** in the full input setting. This highlights its superior scalability and real-time capability.

Results (Performance only)

Method	Modality	NDS ↑
3D-CVF[49]	L+C	62.3
PointPainting[38]	L+C	61.0
FUTR3D[6]	L+C	66.1
FUTR3D-OURS	L+C	66.3
MVP[42]	L+C	70.5
FusionPainting[44]	L+C	71.6
TransFusion[1]	L+C	71.7
CMT[45]	L+C	70.1
CMT-OURS	L+C	62.0
TransCAR[34]	C+R	52.2
CRAFT[19]	C+R	48.3
FUTR3D[6]	C+R	50.3
CenterFusion[32]	C+R	44.9
FUTR3D-OURS	C+R	48.5

Figure E: Comprehensive evaluation of sensor-fusion algorithms for 3D object detection on the nuScenes test set; **CMT/FUTR3D-OURS** denotes the baseline augmented with our plug-in module.

Ablation Research

Modality	Baseline Model	S	i	NDS	Time (s)	Evaluation Speed (1/s)
L+C	FUTR3D	×	×	66.2	65.8	91.31
L+C	FUTR3D	×	✓	67.0	49.6	121.35
L+C	FUTR3D	✓	✓	66.3	50.9	118.25
R+C	FUTR3D	×	×	50.3	80.8	74.49
R+C	FUTR3D	×	✓	45.8	58.0	103.77
R+C	FUTR3D	✓	✓	48.5	77.2	78.0
L+C	CMT	×	×	70.3	146.5	41.1
L+C	CMT	×	✓	63.4	132	45.6
L+C	CMT	✓	✓	62.1	129.6	45.2
L+C	CMT	✓	×	61.5	129.1	46.4

Figure F. Overall detection performance (NDS and Time) on the validation set for models with ResNet101 and ResNet50 backbones under different Self-adaptive Sampler (S) and iTransformer (i) configurations.

Conclusion & Future Work

We propose a plug-in architecture for efficient multimodal 3D object detection, combining a self-adaptive sampler and dimension-inverted transformers. Experiments on nuScenes show competitive accuracy with notable speed gains across sensor types and hardware. The method is especially effective for LiDAR and robust with radar, advancing real-time perception for autonomous driving. Future work will extend to more datasets and baseline models.