# A APPENDIX

## A.1 MULTI-MODAL BRANCHES

**Camera Branch.** To extract image features from multi-view camera images, a backbone network, such as ResNet-50, is employed. These image features are then processed by a feature pyramid network (FPN), which generates multi-scale image features.

**LiDAR Branch.** FusionFormer is designed to accommodate diverse representations of multi-modal features. This study explores two different representation forms of LiDAR features, specifically BEV and voxel features. The original point cloud data is voxelized, and then processed through sparse 3D convolution operations. In one case, voxel features are obtained by encoding the volumetric representation using 3D convolution operations. In another case, the Z-axis of the features are compressed into the channel dimension, and BEV features are obtained with 2D convolution operations.

## A.2 EFFICIENCY AND COMPUTATION COST STUDY

As shown in Table 7, we compare the efficiency of FusionFormer and existing methods. The FPS and performance are tested on a single Tesla A100 GPU with the best model setting of official repositories. In comparison to BEVFusion, FusionFormer demonstrates superior performance with notable improvements of $3.1\%$ in mAP and $2.7\%$ in NDS, while maintaining a similar processing speed. These results highlight the significant advancements achieved by FusionFormer in the field of object detection.

Table 7: Efficiency comparison on the nuScenes val set. "L" is LiDAR. "C" is camera. "T" is temporal. The "-S" indicates that the model only utilizes single-frame BEV features without incorporating temporal fusion techniques.

| Methods | Modality | mAP↑ | NDS↑ | FPS |
|---|---|---|---|---|
| TransFusion | CL | 67.5 | 71.3 | 3.2 |
| BEVFusion | CL | 68.5 | 71.4 | 4.2 |
| UVTR | CL | 65.4 | 70.2 | 2.6 |
| CMT | CL | 70.3 | 72.9 | **6.0** |
| DeepInteraction | CL | 69.8 | 72.6 | 1.7 |
| FusionFormer-S | CL | 70.0 | 73.2 | 4.0 |
| FusionFormer | CLT | **71.4** | **74.1** | 3.8 |

Table 8: Computation cost comparison on the nuScenes val set. "L" is LiDAR. "C" is camera. "T" is temporal. The "-S" indicates that the model only utilizes single-frame BEV features without incorporating temporal fusion techniques.

| Methods | Modality | mAP↑ | NDS↑ | FLOPS | Params |
|---|---|---|---|---|---|
| CMT | CL | 70.3 | 72.9 | 2.17T | 77.73M |
| FusionFormer-S | CL | 70.0 | 73.2 | 2.33T | 77.55M |
| FusionFormer | CLT | **71.4** | **74.1** | 2.42T | 78.54M |

Table 9: The latency of each module in FusionFormer. The latency for the multimodal fusion encoding (MMFE) module represents the total time for 6 layers of encoding, while the time for the temporal fusion encoding (TFE) module represents the total time for 3 layers of encoding.

| | Camera Backbone | LiDAR Backbone | MMFE | TFE | Head |
|---|---|---|---|---|---|
| FusionFormer-S | 20 ms | 124 ms | 80 ms | - | 23 ms |
| FusionFormer | 20 ms | 124 ms | 79 ms | 22 ms | 23 ms |

In addition, we conducted a comparative experiment on computation cost between our method and the previous state-of-the-art method, CMT. As shown in Table 8, our method has similar FLOPS and parameters as CMT.

We also analyzed the time consumption of each module in FusionFormer on a single A100 GPU, and the results are shown in Table 9.

### A.3 Additional Ablation Study

In this section, we investigate the influence of other factors on the performance of FusionFormer. We adopt ResNet-50 as the backbone for the image branch, with an input resolution of $800\times320$ for the image and a voxel size of $0.1m$ for the point cloud branch, outputting $150\times150$ BEV features. With the exception of the temporal fusion section, all other experiments presented in this section are based on single frame.

**Temporal Fusion.** We conducted a comparative study between our proposed temporal fusion module and the concatenation method used by prior temporal fusion approaches under different temporal sequences. All models were trained for 24 epochs and the CBGS strategy was employed during the training process. The experimental results are presented in Table 10. Compared to the previous temporal fusion method using channel concatenation, our deformable attention-based temporal fusion method demonstrates better performance in 3D object detection.

Table 10: Study of the temporal fusion module on the nuScenes val set.

| | **Concate** | | **Ours** | |
|---|---|---|---|---|
| T | mAP↑ | NDS↑ | mAP↑ | NDS↑ |
| 1 | 66.48 | 70.39 | 66.48 | 70.39 |
| 2 | 67.71 | 71.01 | 67.85 | 71.20 |
| 4 | 68.18 | 71.36 | 68.24 | 71.51 |
| 8 | 68.31 | 71.49 | 68.56 | 71.66 |

**CBGS.** We evaluated the impact of utilizing the class-balanced grouping and sampling (CBGS) strategy during the training process on the model's performance. Table 11 presents the results of this comparison. The application of the CBGS strategy resulted in a balanced distribution of samples across different categories, leading to a notable enhancement in the performance of the model.

Table 11: Ablation study of the CBGS strategy on the nuScenes val set.

| CBGS | mAP↑ | NDS↑ |
|---|---|---|
| ✓ | 66.5 | 70.4 |
| | 62.7 | 67.3 |

**Modality Order.** In our proposed method, multi-modal features are sequentially fed into the fusion encoder at each layer. To evaluate the impact of the input order of multi-modal features, we compared the performance of the model with different input orders. The results are presented in Table 12. All models were trained for 24 epochs without utilizing the CBGS strategy.

Table 12: Ablation study of the input order of modality features on the nuScenes val set.

| Order | mAP↑ | NDS↑ |
|---|---|---|
| LC | 62.7 | 67.3 |
| CL | 62.5 | 67.1 |

**Voxel Size.** We conducted an experiment to evaluate the impact of voxel size on the performance of our proposed method. The results are presented in Table 13. The models were trained for 24 epochs and did not use the CBGS strategy.

Table 13: Ablation study of the voxel size on the nuScenes val set.

| Voxel size | mAP↑ | NDS↑ |
|---|---|---|
| 0.075m | 63.2 | 67.8 |
| 0.100m | 62.5 | 67.1 |

**Image Size.** We conducted an experiment to evaluate the impact of image size on the performance of our proposed method. The results are presented in Table 14. The models were trained for 24 epochs and did not use the CBGS strategy.

Table 14: Ablation study of the image size on the nuScenes val set.

| Image size | mAP↑ | NDS↑ |
|---|---|---|
| $1600\times600$ | 64.4 | 68.1 |
| $800\times320$ | 62.5 | 67.1 |

### A.4 QUALITATIVE DETECTION RESULTS

In this section, we showcase further detection results of FusionFormer on the nuScenes test set. As depicted in Figure 7, FusionFormer exhibits exceptional performance in detecting objects at long distances. Notably, the incorporation of the temporal fusion module enables FusionFormer to effectively recall occluded objects by leveraging the fused historical frame BEV features. This capability proves valuable in scenarios where objects may be partially or fully obstructed. The presented results highlight the robustness and effectiveness of FusionFormer in addressing the challenges of object detection in complex environments.

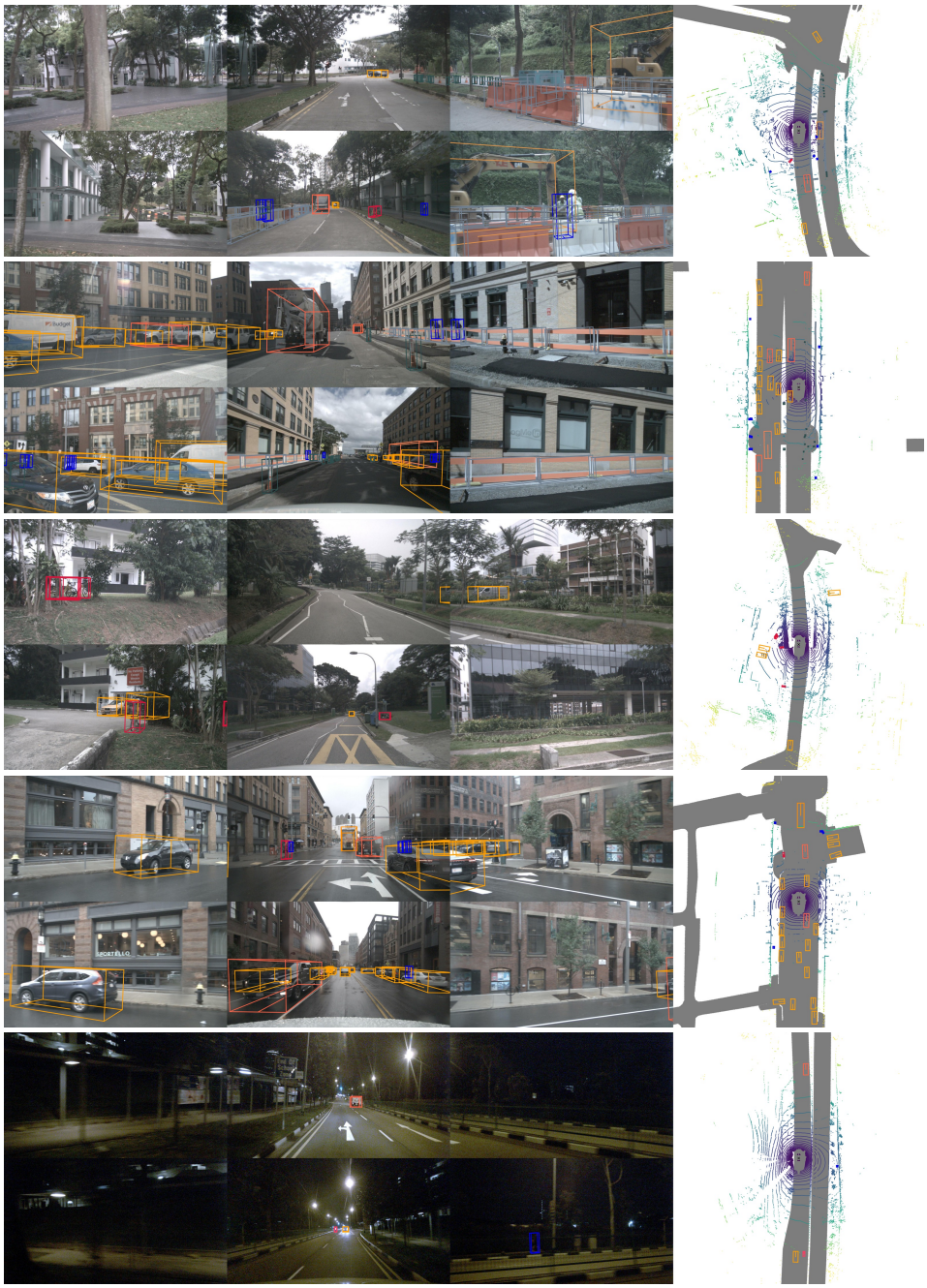

Figure 7: **More qualitative detection results in the nuScenes test set.** Bounding boxes with different colors represent Cars(•), Pedestrians(•), Bus(•) and Truck(•).