Appendix for FocalFormer3D

The supplementary materials for FocalFormer3D is organized as follows:

- Sec. A shows additional ablation studies on decoder head and latency analysis for multi-modal models.
- Sec. B gives more implementation details including network details, and extension to the multi-modal variant.
- Sec. C discusses the *prediction locality* for second-stage refinements.
- Sec. D presents some visual results for multi-stage heatmaps and 3D detection results on bird's eye view.

A. Additional Ablation Studies

Design of the decoder head. We analyze the capability of the decoder head in processing massive queries in Table A - 1. Concerning the type of cross attention, with an increasing number of queries up to 600, the computation time of cross attention module [2] (c) grows faster than deformable one [13] (e). As a result, more deformable transformer layers can be applied. In our experiments, the transformer decoder head with 6 layers obtains the best performance (66.5 mAP and 71.1 NDS) with a more affordable computation time than the cross attention modules. Furthermore, compared with point-level query embedding [1] (g), our box-level query embedding (f) achieves +0.6 points improvements with $3.7ms$ computation overhead, demonstrating the effectiveness of box-level query.

#	C.A.	#Q	#Layer	mAP	NDS	Latency
(a)	Full	200	1	65.8	70.5	7.6ms
(b)	Full	600	1	66.1	70.9	13.1ms
(c)	Full	600	2	66.3	71.1	26.2ms
(d)	Deform	200	6	65.9	70.8	14.8ms
(e)	Deform	600	2	66.2	70.7	7.6ms
(f)	Deform	600	6	66.5	71.1	17.0ms
(g)	w/o Box-pooling			65.9	70.9	–

Table A - 1. **Ablation studies for box-level deformable decoder head.** "C.A." denotes the types of cross attention layers. "# Q" represents the number of used queries. "# Layer" stands for the number of decoder layers. Latency is measured for the transformer decoder head on a V100 GPU for reference.

Latency analysis. We compare ours with other leading-performance methods in Table A - 2. It shows that FocalFormer-F outperforms the dominating methods, BEVFusion [4] and DeepInteraction [8] in terms of both performance and efficiency.

Methods	mAP	NDS	Latency
BEVFusion [4]	69.2	71.8	1610ms
DeepInteraction [8]	70.8	73.4	480ms
FocalFormer3D-F (Ours)	**71.6**	**73.9**	**363ms**

Table A - 2. **Efficiency comparison with other SOTA methods on nuScenes dataset.** Results are shown on nuScenes test set. All methods are tested on a single V100 GPU for reference.

Results on nuScenes val set. We also report the method comparisons on the nuScenes *val* set in Table A - 3.

Methods	mAP	NDS
CBGS [12]	51.4	62.6
CenterPoint [10]	59.6	66.8
LiDARMultiNet [9]	63.8	69.5
TransFusion-L^ [1]	65.2	70.2
FocalFormer3D (Ours)	66.5	71.1

Table A - 3. **Performance comparison on the nuScenes *val* set.** Results marked with ^ indicate our reproduction. The results of other compared methods on the nuScenes *val* set were obtained from their respective original papers.

B. Additional Implementation Details

Model details for nuScenes dataset. On the nuScenes dataset, the voxel size is set as $0.075m \times 0.075m \times 0.2m$ and the detection range is set to $[-54.0m, 54.0m]$ along X and Y axes, and $[-5.0m, 3.0m]$ along Z axis.We follow the common practice of accumulating the past 9 frames to the current frame for both training and validation. We train the LiDAR backbone with the deformable transformer decoder head for 20 epochs. Then, we freeze the pre-trained LiDAR backbones and train the detection head with multi-stage focal heatmaps for another 6 epochs. GT sample augmentation is adopted except for the last 5 epochs. We adopt pooling-based masking for generating Accumulated Positive Mask, where we simply select *Pedestrian* and *Traffic Cones* as the small objects.

Model details for Waymo dataset. On the Waymo dataset, we simply keep the VoxelNet backbone and FocalFormer3D detector head the same as those used for the nuScenes

dataset. The voxel size used for the Waymo dataset is set to $0.1m \times 0.1m \times 0.15m$. For the multi-stage heatmap encoder, we use pooling-based masking, selecting *Vehicle* as the large object category, and *Pedestrain* and *Cyclist* as the small object categories. The training process involves two stages, with the model trained for 36 epochs and another 11 epochs trained for the FocalFormer3D detector. We adopt GT sample augmentation during training, except for the last 6 epochs. As the Waymo dataset provides denser point clouds than nuScenes, the models adopt single-frame point cloud input [10, 1].

Extension to multi-modal fusion model. We provide more details on the extension of FocalFormer3D to its multi-modal variant. Specifically, the image backbone network utilized is ResNet-50 following TransFusion [1]. Rather than using more heavy camera projection techniques such as Lift-split-shot [6] or BEVFormer [3], we project multi-view camera features onto a predefined voxel grid in the 3D space [7]. The BEV size of the voxel grid is set to 180×180, in line with $8\times$ downsampled BEV features produced by VoxelNet [11]. The height of the voxel grid is fixed at 10.

To obtain camera features for BEV LiDAR feature, we adopt a cross-attention module [8] within each pillar. This module views each BEV pixel feature as the query and the projected camera grid features as both the key and value. The generated camera BEV features are then fused with LiDAR BEV features by an extra convolutional layer. This multi-modal fusion is conducted at each stage for the multi-stage heatmap encoder. We leave the exploration of stronger fusion techniques [4, 8, 5] as future work.

C. Prediction Locality of Second-Stage Refinement

Recent 3D detectors have implemented global attention modules [1] or fusion with multi-view camera [4, 8] to capture larger context information and improve the detection accuracy. However, we observe a limited regression range (named as *prediction locality*) compared to the initial heatmap prediction. To analyze their second-stage ability to compensate for the missing detection (false negatives), we visualize the distribution of their predicted center shifts $\delta = (\delta_x, \delta_y)$ in Fig. A - 1 for several recent leading 3D detectors, including the LiDAR detectors (CenterPoint [10], TransFusion-L [1]) and multi-modal detectors (BEVFusion [4], DeepInteraction [8]). Statistics of center shift ($\sigma_\delta < 0.283m$ illustrate almost all predictions are strongly correlated with their initial positions (generally less than 2 meters away), especially for LiDAR-only detectors, such as CenterPoint and TransFusion-L.

The disparity between small object sizes (usually $< 5m \times 5m$) and extensive detection range (over $100m \times 100m$ meters) limits the efficacy of long-range second-stage refinement, despite the introduction of global operations and perspective camera information. Achieving a balance between long-range modeling and computation efficiency for BEV detection is crucial. FocalFormer3D, as the pioneer in identifying false negatives on the BEV heatmap followed by local-scope rescoring, may provide insights for future network design.

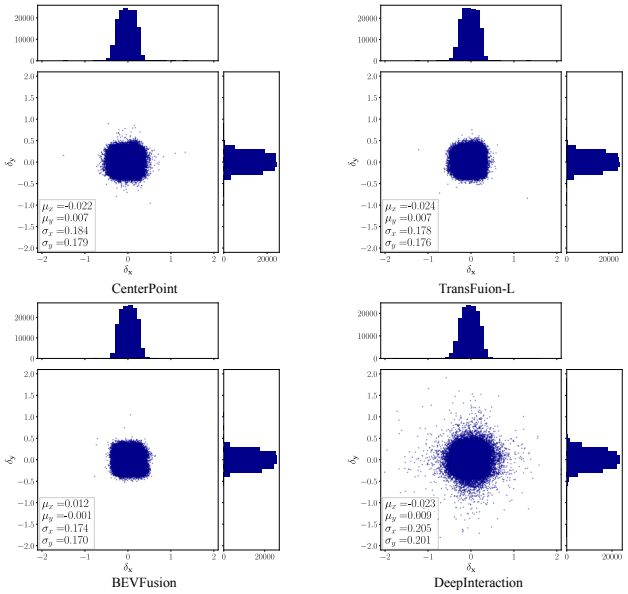

Figure A - 1. Object center shifts (δ_x, δ_y) distribution without normalization between initial heatmap response and final object predictions. The unit is a meter.

D. Example Visualization

Example visualization of multi-stage heatmaps and masking. We present a visual illustration of the multi-stage heatmap encoder process in Fig. A - 2.

Qualitative results. Fig. A - 3 shows some visual results and failure cases of FocalFormer3D on the bird's eye view. Although the average recall $\text{AR}_{<1.0m}$ reaches over 80%, some false negatives are still present due to either large occlusion or insufficient points. Also, despite accurate center prediction, false negatives can arise due to incorrect box orientation. Further exploration of a strong box refinement network is left for future work.

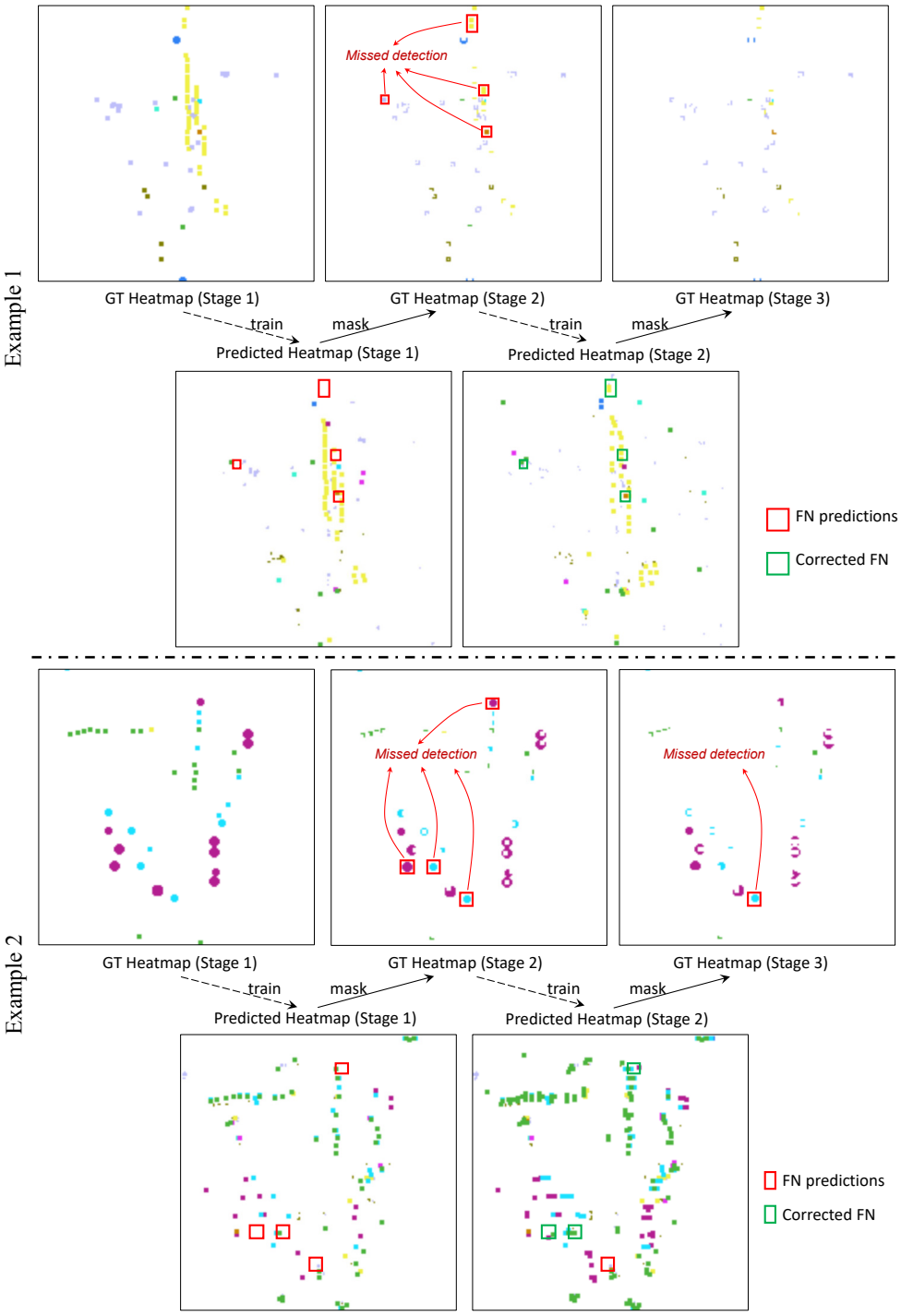

Figure A - 2. **Example visualization of multi-stage heatmap encoder process on the bird's eye view**. The process of identifying false negatives operates stage by stage. We show different categories with different colors for visualization. The top three subfigures display the ground-truth center heatmaps at each stage, highlighting the missed object detections. The two subfigures below display the positive mask that shows positive object predictions. The scene ids are "4de831d46edf46d084ac2cecf682b11a" and "825a9083e9fc466ca6fdb4bb75a95449" from the nuScenes *val* set. We recommend zooming in on the figure for best viewing.

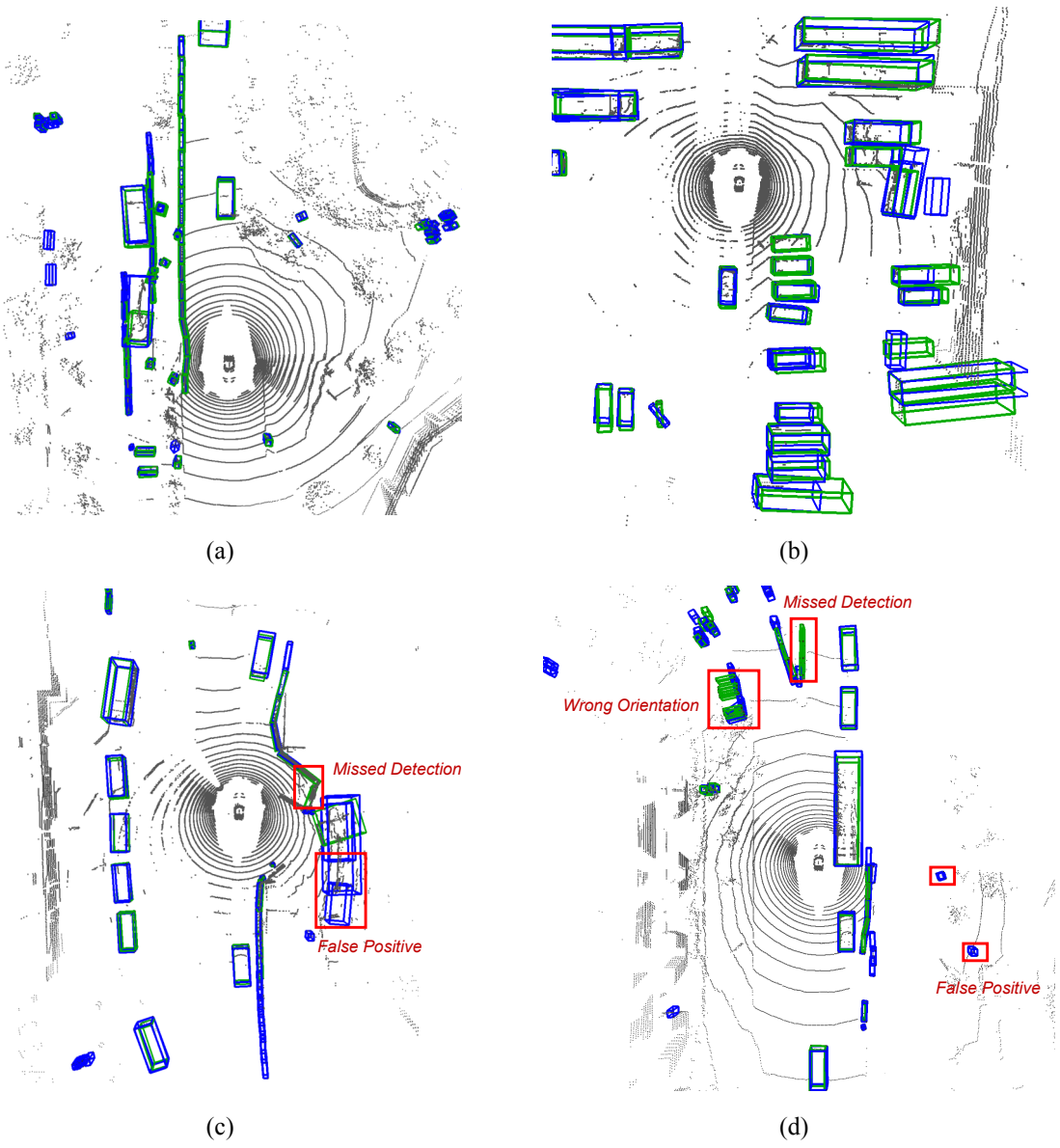

Figure A - 3. **Visual results and failure cases.** The green boxes represent the ground truth objects and the blue ones stand for our predictions. We recommend zooming in on the figure for best viewing.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, pages 1080–1089, 2022. 1, 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 1

[3] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chong-

hao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 2

[4] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. In *Advances in Neural Information Processing Systems*. 1, 2

[5] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 2

[6] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210. Springer, 2020. 2

[7] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. In *British Machine Vision Conference*, 2019. 2

[8] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *arXiv preprint arXiv:2208.11112*, 2022. 1, 2

[9] Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarmultinet: Towards a unified multi-task network for lidar perception. *arXiv preprint arXiv:2209.09385*, 2022. 1

[10] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 1, 2

[11] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 2018. 2

[12] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 1

[13] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1