Supplementary to UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

In our supplementary material, we offer in-depth insights into our network architecture, training methodologies, and ablation baselines, which can be found in Section A. Moreover, we present an extensive examination of our robustness experiments in Section B. Lastly, we address the limitations of UniTR in Section C.

A. Implementation Details

A.1. Network Architecture

Tokenizers. In outdoor perception scenarios, we consider two input modalities: multi-view camera images and LiDAR point clouds. For image inputs, we first take the raw images $X^I \in \mathbb{R}^{6\times256\times704\times3}$ captured by six cameras and split them into non-overlapping patches using a patch-splitting module, similar to ViT [5]. Each patch serves as a "token" with its feature being a concatenation of raw pixel RGB values. In our implementation, we employ an 8×8 patch size, resulting in a feature dimension of $8 \times 8 \times 3 = 192$. A linear embedding layer is then applied to these features, projecting them to an arbitrary dimension denoted as C. The output of the image tokenizer is $\mathcal{T}^I \in \mathbb{R}^{M\times C}$, where M represents the token number.

Regarding LiDAR point clouds X^P, we utilize the standard dynamic voxel feature encoding tokenizer [12] as implemented by OpenPCDet [10]. We use a grid size of $(0.3m, 0.3m, 8.0m)$ for detection and $(0.4m, 0.4m, 8.0m)$ for segmentation to generate LiDAR voxels, $\mathcal{T}^P \in \mathbb{R}^{N\times C}$. By employing these two tokenizers, the multi-modal inputs can be converted to $\mathcal{T} \in \mathbb{R}^{(M+N)\times C}$, which includes N point cloud tokens and M image tokens for subsequent intra-modal transformer blocks.

Multi-modal Backbone. Our UniTR features a single-stride, pillar-based multi-modal backbone that starts with one weight-sharing intra-modal transformer block for parallel processing of modal-wise representation learning. Subsequently, three inter-modal transformer blocks bridge different modalities and establish connections among them by alternating between 2D and 3D partitioning configurations. The block configuration adopted in this paper is $\{\text{intra}, \text{inter}_{2D}, \text{inter}_{2D}, \text{inter}_{3D}\}$. The window sizes for both $L^P \times W^P \times H^P$ and $L^I \times W^I \times 1$ are (30, 30, 1), and the maximum number of tokens assigned to each set (τ) is set to 90 for all modalities. All attention modules are equipped with 8 heads, 128 input channels, and 256 hidden channels. For the inter-modal block (3D), the pseudo grid points size, $L^S \times W^S \times H^S$, is set to $360 \times 360 \times 20$.

A.2. Ablation Baselines

Effect of 2D & 3D fusion. The base competitor is the lidar-only variant of our model with four intra-modal blocks [11] and transfusion head [1]. For a fair comparison, we only switch the fusion algorithm while keeping all other settings remain unchanged. The number of the intra- and inter-modal blocks is summarized in Table 1.

Modality	Intra-B	Inter-B (2D)	Inter-B (3D)	BEVLSS	NDS	mAP
L	4	0	0		70.5	65.9
C+L	3	0	1		72.0	68.5
C+L	3	1	0		72.5	69.0
C+L	2	1	1		72.9	69.8
C+L	1	2	1		73.1	70.0
C+L	1	2	1	✓	73.3	70.5

Table 1. The number of the intra- and inter-modal blocks on the ablation of 2D & 3D fusion. Camera (C), LiDAR (L).

Effect of parallel intra-modal transformer block. The 1^{st} and 2^{nd} rows are the image-only and lidar-only baselines with four intra-modal blocks. To evaluate the effectiveness of our weight-sharing approach, we conducted experiments with both serial and parallel multi-modal variants, where only a BEV unifier was used without our proposed 2D&3D fusion strategies. This allowed us to better isolate the impact of the weight-sharing approach on its own, separate from the strong fusion strategies. All latency measurements are taken on the same workstation with an A100 GPU. Note that the latency reported in Table 4 of the main paper only refers to the transformer backbone, without including the modality-specific tokenizers and partitioning.

A.3. Training Schemes

As stated in the main paper, previous approaches for multi-modal fusion usually involve two-step training strategies with separate single-modal pre-training and joint multi-modal post-training for fusion. In contrast, our UniTR can be directly trained with a one-step end-to-end training

scheme, where the data augmentations of the image and lidar are aligned. We follow the matching strategies of BEV and image space data augmentation used in BEVFusion [8], *e.g.*, random rotation, translation, and flip. To synchronize 2D-3D joint GT-AUG, we use the same implementation of cross-modal copy-paste proposed in [3] with a Mix-up Ratio $\alpha = 0.7$ and add fade strategy at the last two epochs. Our UniTR backbone is pre-trained on both ImageNet [4] and nuImage [2] datasets. We train all experiments using the AdamW optimizer [9] on 8 A100 GPUs with weight decay 0.03, a one-cycle learning rate policy [6], and a maximum learning rate of 3e-3. For 3D object detection, we used a batch size of 24 and trained for 10 epochs, while for BEV map segmentation, we used a batch size of 24 and trained for 20 epochs. All inference times were measured on the same workstation (single A100 GPU and AMD EPYC 7513 CPU).

B. Robustness Against Sensor Failure

B.1. LiDAR Malfunctions.

To assess the robustness of our framework, we conducted experiments on the nuScenes validation set under conditions where objects cannot reflect LiDAR points. Such situations may arise during rainy weather when the reflection rate of certain common objects falls below the LiDAR system's threshold. To simulate this scenario, we employed the same dropping strategy as [7]: each frame has a 50% chance of dropping objects, and each object has a 50% chance of dropping the LiDAR points it contains.

As shown in the main paper, our UniTR outperforms both the LiDAR-only stream and previous fusion approaches, such as BEVFusion [7], in terms of accuracy when detectors are evaluated without robustness augmentation. Furthermore, our method exhibits significant improvements when the detectors are fine-tuned on the robust augmented training set, outpacing BEVFusion by a substantial margin.

B.2. Camera Malfunctions.

We performed additional experiments to evaluate the robustness of our UniTR backbone against camera malfunctions in three scenarios outlined in [7]: i) missing front camera, ii) missing all cameras except the front, and iii) 50% of camera frames stuck. As demonstrated in the main paper, UniTR surpasses other LiDAR-camera fusion methods and even camera-only methods in these scenarios, showcasing its resilience against camera malfunctions.

C. Limitation

Despite its notable performance and processing speed in multi-modal 3D perception, our UniTR has certain limitations that warrant attention. First, as it inherits features from DSVT, UniTR is primarily a single-stride backbone designed for outdoor BEV perception, which constrains its adaptability to various other 3D perception tasks, such as indoor 3D perception. Second, UniTR is mainly focused on jointly processing different sensor types without considering the compatibility of transformation modalities for diverse scenarios. For instance, it does not accommodate switching to LiDAR-only, image-only, or image-LiDAR encoders during the inference stage. The design of a more modality-flexible backbone utilizing a Mixture-of-Experts approach remains an open challenge for the 3D perception community.

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*, 2022.

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[3] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. *arXiv preprint arXiv:2207.10316*, 2022.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[6] Sylvain Gugger. The 1cycle policy. https://sgugger.github.io/the-1cycle-policy.html, 2018.

[7] 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. 2022.

[8] 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. In *ICRA*, 2023.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[10] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020.

[11] 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. *CVPR*, 2023.

[12] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *CoRL*, 2020.