PEFT-DML: Parameter-Efficient Fine-Tuning Deep Metric Learning for Robust Multi-Modal 3D Object Detection in Autonomous Driving (Student Abstract)

Abdolazim Rezaei, Mehdi Sookhak
Department of Computer Science
6300 Ocean Dr
Texas A&M University
Corpus Christi, TX 78412 USA
arezaei@islander.tamucc.edu, m.sookhak@ieee.org

Abstract

This study introduces PEFT-DML, a parameter-efficient deep metric learning framework for robust multi-modal 3D object detection in autonomous driving. Unlike conventional models that assume fixed sensor availability, PEFT-DML maps diverse modalities (LiDAR, radar, camera, IMU, GNSS) into a shared latent space, enabling reliable detection even under sensor dropout or unseen modality–class combinations. By integrating Low-Rank Adaptation (LoRA) and adapter layers, PEFT-DML achieves significant training efficiency while enhancing robustness to fast motion, weather variability, and domain shifts. Experiments on benchmarks nuScenes demonstrate superior accuracy.

Introduction

Reliable detection of moving 3D objects is fundamental for autonomous driving but there are still challenges by the fast motion, unstable environmental conditions, and sensor limitations (Qian, Lai, and Li 2022). To overcome these issues, we propose PEFT-DML, which unifies LiDAR, radar, camera, IMU, and GNSS into a shared latent space. Using LoRA and adapters, our proposed model, PEFT-DML, achieves robust detection which is modality-agnostic with reduced training cost. Following prior Deep Metric Learning (DML) formulations (Dullerud et al. 2022; Peng and Wang 2022), the objective is to learn a projection function $\phi(\cdot)$ that maps inputs from heterogeneous modalities into a shared embedding space, where intra-class samples cluster closely and inter-class samples remain well separated. This embedding alignment enables consistent cross-modal representation learning which forms the basis for our parameter-efficient PEFT–DML framework that extends DML principles to robust multi-modal 3D object detection.

The proposed PEFT-DML framework surpasses recent studies in multi-modal 3D object detection and cooperative perception. Unlike 3ML-DML framework (Dullerud et al. 2022), which requires fixed modality availability, PEFT-DML generalizes across unseen modality–class combinations through a unified latent space that enables zero-shot cross-modal detection. In addition, CRKD (Zhao, Song, and Skinner 2024) focuses on distillation between camera and radar but it is fragile when one sensor fails. PEFT-DML instead supports any subset of sensors which ensures robustness under partial sensor dropout.

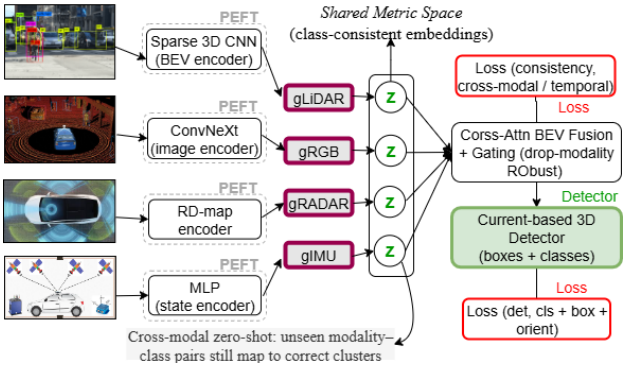

Figure 1: The PEFT-DML pipeline unifies various modalities into a shared latent space with LoRA and adapter layers.

RoboFusion (Song et al. 2024) leverages computationally heavy Visual Foundation Models which limits efficiency. In contrast, PEFT-DML achieves comparable robustness in lightweight LoRA-based fine-tuning. The authors in (Chae, Kim, and Yoon 2024) introduce weather-aware gating but still requires both modalities at inference whereas PEFT-DML performs robust detection even when one or more modalities are missing.

PEFT-DML Framework

We propose **PEFT-DML**, a framework for robust 3D object detection in autonomous driving (Figure 1). The model unifies heterogeneous modalities where backbone encoders remain frozen to preserve pretrained features, while lightweight *LoRA* and adapter layers enable efficient fine-tuning. Each modality is mapped by a projection head into a normalized d-dimensional embedding, where intra-class features cluster closely and inter-class features remain distinct. Cross-attention and gating modules fuse embeddings, ensuring flexibility under sensor dropout. A detection head then predicts 3D bounding boxes and class labels.

Training is guided by a joint multi-objective loss $\mathcal{L} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{met}}\mathcal{L}_{\text{metric}} + \lambda_{\text{cons}}\mathcal{L}_{\text{consistency}}$ where **Detection Loss**

Method	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
UVTR	37.2	52.2	0.612	0.256	0.385	0.664	0.125
X3KD	45.6	56.1	0.506	0.253	0.414	0.366	0.131
UniDistill	29.6	39.3	0.637	0.257	0.492	1.084	0.167
CRKD	48.7	58.7	0.404	0.253	0.425	0.376	0.111
RoboFusion	54.3	67.1	0.338	0.229	0.382	0.367	0.102
3D-LRF	45.2	57.8	0.337	0.226	0.398	0.375	0.118
PEFT-DML	**62.2**	**71.7**	**0.316**	**0.206**	**0.346**	**0.339**	**0.093**

Table 1: Table 1: PEFT-DML achieves the best performance across all metrics,

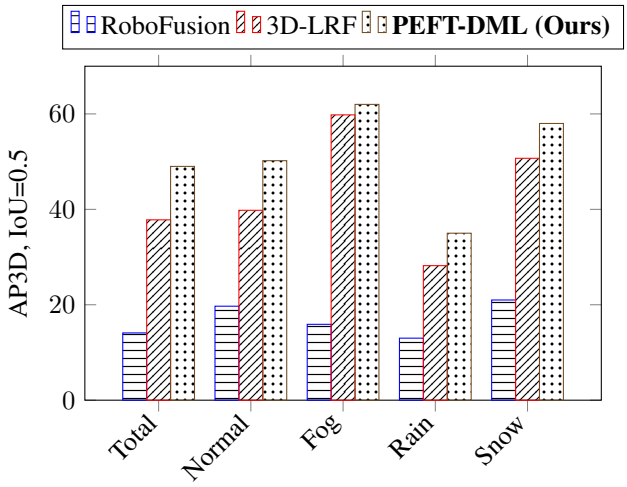

Figure 2: Comparison over different climate conditions.

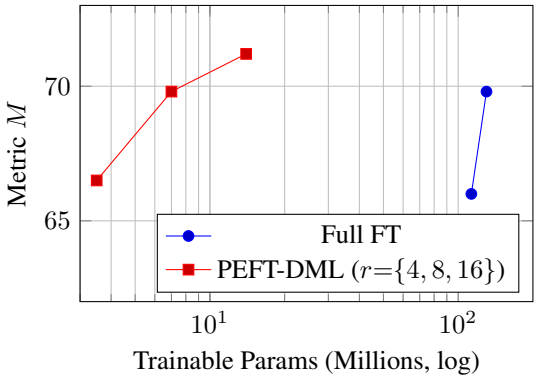

Figure 3: PEFT-DML achieves nearly the same or slightly higher accuracy than Full Fine-Tuning (Full-FT) while updating less than 10% of the parameters.

is $\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} = \text{FocalCE}(y, \hat{y}) + \text{IoU}(b, \hat{b}) + \|o - \hat{o}\|_1$. This combines focal classification loss, IoU-based bounding box regression, and orientation regression.

Metric Alignment Loss, is $\mathcal{L}_{\text{metric}} = \max\big(0, d(z_i, z_j) - d(z_i, z_k) + \alpha\big)$. A triplet loss encourages embeddings z_i and z_j from the same class to be closer than z_i and z_k from different classes.

In addition, **Consistency Loss** is $\mathcal{L}_{\text{consistency}} = \|z_t - z_{t+1}\|_2^2$ which enforces temporal stability across adjacent frames and consistency across modalities.

Together, these objectives enable **cross-modal zero-shot generalization** by mapping them into the shared latent space and comparing them with embeddings from known modalities.

Experiments

Experiments on nuScenes dataset demonstrate that PEFT-DML achieves superior accuracy, robustness, and parameter efficiency compared to state-of-the-art baselines.

Figure 2: PEFT-DML consistently achieves the highest AP3D scores across all weather conditions, demonstrating superior robustness compared to RoboFusion and 3D-LRF. Furthermore, PEFT-DML in Figure 3 achieves higher accuracy than full fine-tuning while requiring far fewer trainable parameters, demonstrating superior parameter efficiency. Table 1 compares PEFT-DML with six recent methods across the nuScenes, using detection and error metrics. The results demonstrate that PEFT-DML outperforms all baselines. It achieves the highest mAP (62.2) and NDS (71.7), reflecting superior detection accuracy and overall performance. In terms of localization and geometry, PEFT-DML attains the lowest mATE (0.316) and mASE (0.206), indicating more precise and stable bounding boxes. It also outperforms others in orientation and velocity estimation, with the lowest mAOE (0.346), mAVE (0.339), and mAAE (0.093).

For clarity, the weighting coefficients in the joint loss were empirically set to balance detection, metric, and temporal objectives ($\lambda_{\text{triplet}} = 0.5$, $\lambda_{\text{xmod}} = 0.2$, $\lambda_{\text{temp}} = 0.1$). LoRA ranks were chosen as $r = 8$ for high-dimensional backbones (camera/LiDAR) and $r = 4$ for compact modalities (radar/IMU/GNSS) to ensure parameter efficiency. Under true sensor-dropout tests on the nuScenes dataset, the model obtained over 85% of its full-sensor accuracy. Internal ablations further showed that LoRA, adapter layers, and each loss term contribute incrementally, confirming that both the parameter-efficient design and multi-objective training together enhance robustness.

Conclusion

In conclusion, the proposed PEFT-DML framework delivers a modality-agnostic 3D object detection for autonomous driving. Utilizing different sensors in a shared latent space and employing PEFT, our model outperforms other models.

Acknowledgments

This research is supported by the US Department of Transportation (USDOT) Tier-1 University Transportation Center (UTC) Transportation Cybersecurity Center for Advanced Research and Education (CYBER-CARE-Grant No. 69A3552348332).

References

Chae, Y.; Kim, H.; and Yoon, K.-J. 2024. Towards robust 3d object detection with lidar and 4d radar fusion in various weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15162–15172.

Dullerud, N.; Roth, K.; Zhang, H.; Jin, Q.; Hartvigsen, T.; and Ghassemi, M. 2022. An Integrated Multi-Label Multi-Modal Framework in Deep Metric Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Withdrawn submission.

Peng, Y.; and Wang, Y. 2022. Leaf disease image retrieval with object detection and deep metric learning. *Frontiers in Plant Science*, 13: 963302.

Qian, R.; Lai, X.; and Li, X. 2022. 3D object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796.

Song, Z.; Zhang, G.; Liu, L.; Yang, L.; Xu, S.; Jia, C.; Jia, F.; and Wang, L. 2024. RoboFusion: Towards robust multi-modal 3D object detection via SAM. *arXiv preprint arXiv:2401.03907*.

Zhao, L.; Song, J.; and Skinner, K. A. 2024. Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15470–15480.