A Multimodal Framework for MODT Using Enhanced Correlation-Based Affinity Metrics in Autonomous Driving

Muhammad Adeel Altaf[1], Min Young Kim[2*]
[1]School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, South Korea
Email: adeel.altaf88@knu.ac.kr
[2]School of Electronics Engineering, Research Center for Neurosurgical Robotic System,
Kyungpook National University, Daegu, South Korea
Email: minykim@knu.ac.kr

***Abstract*—3D object tracking plays a pivotal role in 3D computer vision, with significant applications in robotics, autonomous vehicles, and human-computer interaction. Despite progress, leveraging multimodal information to enhance the accuracy of multi-object detection and tracking (MODT) remains a key research challenge. To address this, we introduce a multimodal multi-object tracking (MOT) framework based on enhanced Affinity computation-based multi-object detection and tracking (ACMODT), specifically designed for autonomous driving scenarios. This framework integrates data from cameras and LiDAR sensors to deliver more reliable feature extraction and correlation for real-time tracking and detection. Our approach employs a deep neural network (DNN) that combines image (2D) and point cloud (3D) data for simultaneous object detection, tracking, and association. We design a reliable module to calculate motion and appearance relationships in 3D space while accounting for multiple occlusions and harsh weather conditions. We also create a unified data association module to optimize detection reliability, object associations, and start-end estimation. Experiments taken on the KITTI car tracking dataset and RADIATE dataset demonstrate our method achieves superior tracking accuracy and precision compared to existing approaches.**

***Index Terms*—Deep neural network, Affinity computation, Data association, Autonomous driving, Multi-object Tracking.**

I. Introduction

The importance of 3D object tracking [1], [2] has grown significantly across various fields, including human-computer interaction, robotics, and autonomous driving. Recent trends show an increasing use of sensors like LiDARs, radars, RGB cameras, and infrared sensors in vehicles. Autonomous vehicles with these multi-sensors can gather richer perceptual data, enabling safer and more dependable driving performance. For instance, Kim et al. [3] introduced EagerMOT, a multi-order data association approach that effectively integrates data from various object detection modules and modalities. Shenoi et al. developed JRMOT [4], which combines 2D camera images with 3D point cloud data for real-time tracking. Zhang et al. presented mmMOT [5], pioneering deep features from point clouds in tracking tasks. Their findings demonstrate that fusing multi-sensor data significantly enhances tracking accuracy compared to single-sensor approaches.

A major challenge in 3D object tracking lies in improving detection precision when utilizing multimodal information from multi-sensors. A conventional MOT approach comprises object detection, correlation, data association with affinity computation, and track management. To address the inherent complexities in tracking, robust affinity metrics are required, blending appearance and geometric features to manage subtle visual differences and complex motion dynamics. However, while fusion-based methods have been explored, the impact of multimodal features from multi-sensors on multi-object detection remains insufficiently studied. Prior research in 3D-MOT often prioritizes feature distance correlation while overlooking the directional correlation between features.

This study makes the following contributions:

1) This study introduces an end-to-end framework called Enhanced affinity computation multi-object detection and tracking, designed to produce 3D bounding boxes and best association scores in real-time by leveraging camera and LiDAR data with the Enhanced Boost Correlation Feature (EBcF).
2) The proposed method replaces 3D mean IoU [6] with 3D Enhanced Generalized IoU (3D-EGIoU) for geometric affinity calculations, enabling a more precise representation of the spatial relationships between objects.
3) The method is tested on the two datasets: KITTI car tracking benchmark [7] and RADIATE [8] which demonstrate significant improvements in tracking accuracy, precision, lesser ID switches (IDSW), and other key evaluation metrics compared to existing methods with great visualization in the form of 2D images.

The structure of this paper is outlined as follows: Section II provides an overview of related work on MOT methodologies. Section III details the system architecture of the proposed work. Section IV explains the affinity computation framework with its metrics. Section V presents the experimental evaluations and analyzes the results. Finally, Sections VI and VII

*Corresponding author: Min Young Kim (minykim@knu.ac.kr)

summarize the key findings in the Conclusion, limitations, and potential directions for future research respectively.

II. Related work

A. Multi-Object Tracking

Two primary paradigms exist for addressing MOT challenges. The first is the tracking-by-detection (TBD) approach, which separates detection and tracking into independent tasks. Mostly, MOT methods adhere to the TBD approach. However, TBD-based methods face significant limitations, such as reduced performance speed and error accumulation, due to the cascading nature of data association, object detection, and their association. To address the shortcomings, the joint detection and tracking (JDT) paradigm [9] integrates these tasks into an end-to-end learning framework. Wu et al. [2] introduced an innovative online tracking model called Track-to-Detect and Segment (TraDeS), which enhances MOT by integrating tracking information back into the detection stage and incorporating a re-identification (Re-ID) loss that aligns better with detection loss. Additionally, several tracking methods based on the JDT approach have been developed, such as RetinaTrack [9], CenterTrack [10], ChainedTracker [11], JDE [12], and JMODT [13]. For instance, ChainedTracker [11] creates tracklets by linking paired boxes across consecutive frames, while Zhang et al. [5] demonstrated that leveraging correlations between detection pairs can enhance overall model performance.

Despite the performance advantages of the JDT approach, designing effective models for it remains challenging. The success of the JDT paradigm relies heavily on creating robust models that can effectively utilize multiple sensor's information. Notably, much of the research on 3D-MOT with multi-sensor fusion has underscored the critical role of accurate sensor calibration in achieving optimal tracking performance. However, there remains a gap in addressing the attribute relationships between objects, which are often overlooked in existing studies. Our proposed ACMODT method follows the JDT paradigm.

III. System Architecture

The architecture consists of multiple interconnected components designed to facilitate continuous object tracking, as illustrated in the figure 1. The system leverages a deep neural network comprising many key subnetworks, including a backbone network, a Region Proposal Network (RPN), an RCNN [14], and a PointRCNN [15]. The backbone network is responsible for extracting features from both the 2D images and the 3D point cloud data. The RPN creates initial object proposals, which are subsequently classified and refined by the RCNN, and the PointRCNN completes the process by conducting 3D object detection and segmenting individual instances. Detection results are generated using the Region of Interest (RoI) and proposal features provided by the detection network. Meanwhile, the correlation network utilizes RoI features along with the EBcF to compute re-identification (Re-ID) and start-end estimation.

In GMM-based affinity methods, an affinity matrix (a cost matrix) is constructed, where each element indicates the level of similarity between a detection and a track. This matrix incorporates affinities derived from the GMM, along with additional metrics such as classification confidence scores, which reflect the likelihood of detection being a relevant object. It also includes other evaluation measures, such as Intersection over Union (IoU) and Euclidean distance, to assess the alignment between detection and tracking predictions. Both motion and appearance information contribute to calculating object affinities, with appearance affinities represented by the softmax-ranked outputs of the Re-ID network. For motion prediction, this study utilizes the Kalman filter. GMM-based data association is used to improve the matching of detections between frames. The GMM helps model the distribution of features extracted from the detected and predicted objects. To ensure uninterrupted tracking, the track management module handles cases of occlusion or object reappearance, maintaining consistent tracking over time.

IV. Affinity Computation Framework

The shared features generated by the RPN undergo further processing to produce 3D bounding boxes and more precise association scores using data from both the camera and LiDAR sensors. This process does not alter the 2D or 3D encoding modules but instead filters the RPN features based on a predefined threshold, ensuring that object features with the same ID are standardized. This study introduces an affinity metric that combines Enhanced boost correlation features with a 3D-enhanced generalized Intersection over Union to capture appearance similarity and motion consistency better. It is expressed as follows:

$$Y_{p,q}^{aff} = \lambda \cdot B_{p,q}^{\text{app}} + \gamma \cdot B_{p,q}^{\text{3D-EGIoU}} \tag{1}$$

Where:

- $B_{p,q}^{\text{app}}$ is the affinity calculated using the Enhanced Boost Correlation Features.
- $B_{p,q}^{\text{3D-EGIoU}}$ is the geometric affinity metric using the new 3D Enhanced Generalized IoU.
- λ and γ are weighting factors where $\lambda + \gamma = 1$.

A. Enhanced Boost Correlation Feature

The mmMOT method [5] uses element-wise absolute, subtraction, and multiplication to compute the correlation between candidate features. Determining adjacency requires calculating the correlation for each pair of detected objects. This correlation process is not dependent on batch size, allowing it to handle data from different modalities, and it operates channel by channel to leverage the capabilities of the neural network. In JMODT [13], features that are not effective are removed using a standard IoU threshold, and absolute subtraction is applied to find the correlation between candidate features, which reflects the relationship between frames. However, these approaches do not address the direction of the features. Thus, a more comprehensive approach to feature correlation is needed

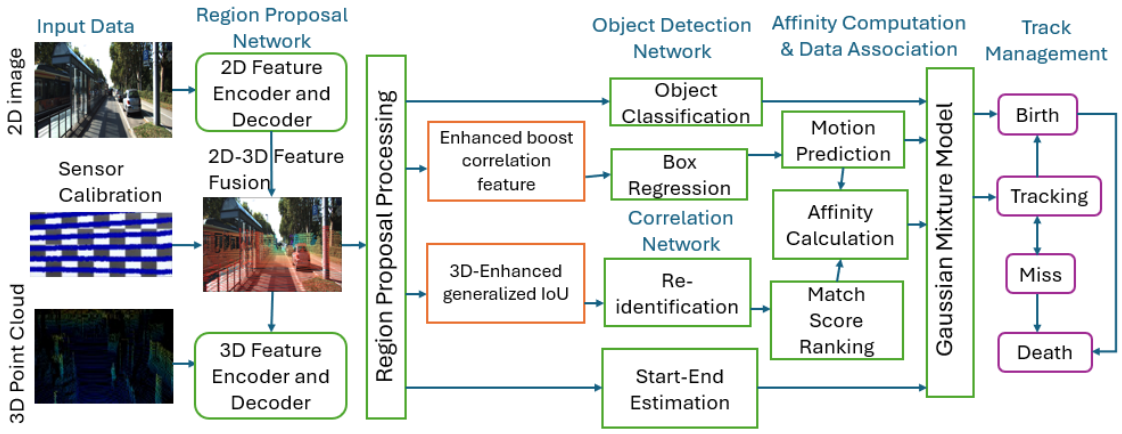

Fig. 1: Structure diagram of the affinity computation-based multi-object detection and tracking workflow using robust data association with affinity metrics of Enhanced Boost Correlation Features and three dimensional-enhanced generalized Intersection over Union.

to account for this limitation. EBcF is calculated considering features from the previous frame and the current frame for a better understanding of the object-level dependencies across frames. Unlike the traditional cosine similarity [16], this method also considers temporal consistency and feature variations. The correlation is determined using a modified similarity function, which incorporates a temporal smoothing factor. The equation is defined as follows:

$$EBcF_{p,q} = (1-\delta)\,\|G_p - G_q\| + \delta \cdot \frac{G_p \cdot G_q}{\|G_p\|\|G_q\|} \quad (2)$$

Where:

- G_p and G_q are the features of bounding boxes.
- $G_p \cdot G_q$ is the dot product between the two feature vectors.
- $\|G_p\|$ and $\|G_q\|$ are the magnitudes of the feature vectors.
- δ is a temporal consistency weighting factor that ensures smoother feature correlations across frames.

This enhanced version provides a more robust correlation score that accounts for both spatial differences and temporal variations, thus improving association robustness in multi-object tracking.

B. 3D Enhanced Generalized IoU

This study proposes a new geometric affinity measure called 3D-EGIoU. This measure extends the traditional IoU with additional penalties for bounding boxes that are misaligned or have large aspect ratio discrepancies, leading to better accuracy when calculating overlap between 3D bounding boxes. The 3D-EGIoU is defined as follows:

$$3D - EGIoU_{p,q} = \text{Average}\,(IoU_{3D} + GIoU_{3D} + EIoU_{3D}) \quad (3)$$

Where:

- IoU_{3D} represents the basic 3D Intersection over Union between the two bounding boxes.
- $GIoU_{3D}$ represents the Generalized Intersection over Union in 3D space, which adds a penalty for bounding boxes that are not overlapping but are spatially close.
- $EIoU_{3D}$ represents the Enhanced Intersection over Union in 3D space, which takes into account not only the aspect ratios of the bounding boxes but also their relative orientation in 3D space, thus providing a better measure for similarity in complex scenes.

This measure is designed to handle various overlapping situations while considering object orientation and aspect ratio, making it more comprehensive than traditional IoU metrics.

C. Affinity Metric Computation

Unlike traditional methods that compute appearance affinity solely based on the distance between camera-LiDAR fusion features, this approach incorporates the Enhanced boosted correlation features into the calculation. This study potentially improves the accuracy and robustness of the multi-object tracking framework by leveraging richer appearance information through EBcF and a more comprehensive geometric similarity metric using 3D-EGIoU. The pseudocode for the affinity computation is outlined in Algorithm 1.

V. Experimental Study

The experiments were conducted on a system equipped with an Intel(R) Core (TM) i7-8700K CPU and a TITAN RTX GPU with 24 GB of memory. The implementation was carried out using Python and PyTorch. A pre-trained EPNet detection model [17] was utilized, and the correlation network was trained for 40 epochs with a batch size of 4. The training process employed the AdamW optimizer [18] with a cosine annealing learning rate of 2×10^{-4}. In this study, GMM was adopted for data association, with its parameters aligned and compared with the MIP-based data association used in JMODT [13]. While computing affinities, the parameters were optimized based on empirical evaluation to achieve their best

Algorithm 1: Affinity with EBcF and 3D-EGIoU

Input: Detection measurements E, tracks L, and proposal features $H = \{H_i, i \in E \cap L\}$.

Output: Refined affinities $Y_{p,q}^{aff}, p \in E, q \in L$

for *each track* $q \in L$ **do**
  Predict the 3D bounding box C_q for track q using Kalman Filter;
  for *each detection* $p \in E$ **do**
    Feature correlation: $F_{p,q} \leftarrow |G_p - G_q|$;
    Appearance Re-ID affinity for feature $F_{p,q}$: $B_{p,q}^{\text{app}} \leftarrow$ Appearance Re-ID$(F_{p,q})$;
    3D bounding box p: C_p;
    $B_{p,q}^{\text{3D-EGIoU}} \leftarrow$ 3D-EGIoU(C_p, C_q);
$B^{\text{app}} \leftarrow \{B_{p,q}^{\text{app}}, p \in E, q \in L\}$;
$B^{\text{3D-EGIoU}} \leftarrow \{B_{p,q}^{\text{3D-EGIoU}}, p \in E, q \in L\}$;
$P \leftarrow \text{Softmax}(B^{\text{app}})$ along columns;
$Q \leftarrow \text{Softmax}(B^{\text{app}})$ along rows;
$B^{\text{app}} \leftarrow \frac{1}{2}(P + Q)$;
$Y_{p,q}^{aff} \leftarrow \lambda B^{\text{app}} + \gamma B^{\text{3D-EGIoU}}$;

performance. For instance, the appearance score weight was set to 2, the classification score weight was assigned a value of 100, the IoU score weight was set to 10, the distance measure weight was adjusted to 10, and the start-end probability weight was defined as 1.

In terms of data association, the weights for classification, affinity, and spatial embedding were set to 100, 22, and 1, respectively, and these values were fine-tuned using cross-validation. The classification threshold for filtering detections with low confidence was set to a high value of 0.80 to eliminate unreliable detection results.

A. Evaluation Metrics

This proposed ACMODT method assessed the performance using the KITTI car tracking dataset [7]. This dataset includes 29 testing and 21 training sequences, comprising forward-facing camera 2D images and LiDAR 3D point cloud data. Each ground truth annotation in the dataset contains a unique ID associated with a 3D bounding box. An object is considered a true positive (TP) only if its 2D-IoU [19] exceeds 0.5. Following the KITTI evaluation standards, we used metrics such as Multiple Object Tracking Precision (MOTP), Multiple Object Tracking Accuracy (MOTA), Multiple Object Detection Accuracy (MODP), Multiple Object Detection Accuracy (MODA), False Negative (FN), False Positive (FP), Mostly Tracked (MT), Mostly Lost (ML), fragmentation (Frag), and ID-switches (IDSW) to assess MOT performance [20].

B. Quantitative and Qualitative Results

In comparison to other published methods, including AB3DMOT, mmMOT, JRMOT, JMODT, and BcMOT, our approach ACMODT demonstrated improved accuracy and precision in terms of multiple-object tracking and detection by surpassing all of these methods across all evaluated indicators in the vehicle-tracking benchmark tests on the KITTI dataset as shown in the table I. The evaluation results, presented in table I, are based on the MOTA metric [20]. The quantitative

TABLE I: The car tracking performance comparison on the KITTI dataset is based on results reported in respective research papers and data obtained from the KITTI public leaderboards.

Method	AB3DMOT	mmMOT	JRMOT	JMODT	BcMOT	**Ours**
TBD	✓	✓	✓	✗	✗	✗
JDT	✗	✗	✗	✓	✓	✓
MOTA ↑	83.92%	84.77%	85.70%	86.27%	86.53%	**88.56%**
MOTP ↑	85.30%	85.21%	85.48%	85.41%	85.37%	**89.61%**
MODA ↑	83.95%	85.60%	85.98%	86.40%	86.66%	**88.73%**
MODP ↑	88.21%	88.28%	88.42%	88.32%	88.29%	**90.18%**
FP ↓	978	711	772	772	1,248	**589**
FN ↓	4,542	4,243	4,049	3,433	3,341	**962**
MT ↑	66.77%	73.23%	71.85%	77.38%	78.31%	**86.57%**
ML ↓	9.08%	2.77%	4.00%	2.92%	2.62%	**1.38%**
IDSW ↓	10	284	98	45	45	**2**
Frag ↓	199	753	372	585	626	**130**
Runtime ↓	0.05 s	0.02 s	0.07 s	0.01 s	0.01 s	0.01 s

results compare the performance of various MOT methods. MOTA represents the overall tracking accuracy, considering false positives, false negatives, and identity switches. Our method achieves the highest value of 88.56%, outperforming the second-best method (BcMOT) at 86.53%, indicating superior tracking accuracy. Besides, our approach achieves the highest score (89.61%), showcasing improved object localization precision compared to other methods, such as 85.48% by JRMOT. Similarly to MODA, our method achieves the best result (88.73%), indicating the method's high detection accuracy, closely followed by BcMOT (86.66%). In MODP, it leads with a score of 90.18%, reflecting excellent precision in detecting and bounding objects. Our method also minimizes false positives (589), significantly lower than the baseline BcMOT (1,248), showing better detection reliability. Our ACMODT method also minimizes errors, with the lowest FN (962), IDSW (2), and Frag (130), highlighting its reliability and tracking consistency. It achieves the highest percentage of MT by 86.57% and the lowest percentage of ML by 1.38%, demonstrating its robust ability to maintain accurate trajectories. Furthermore, the runtime per frame is 0.01 seconds, comparable to efficient methods like JMODT and BcMOT, ensuring real-time applicability. These results emphasize the proposed method's state-of-the-art performance, enabled by effective multimodal fusion and robust optimization, making it highly suitable for 3D object tracking applications in autonomous driving.

In the field of MODT, challenges such as occlusion make both detection and tracking highly complex. Objects can be partially or fully occluded for a period, whether in 2D image data or 3D point clouds. First, we select random frames within a sequence of 0001 from the KITTI car tracking dataset. Our multi-object tracking algorithm is designed to handle

complex urban traffic conditions, especially in fully occluded objects, which can be visualized in figure 2. The algorithm effectively identifies and tracks vehicles, assigning distinct IDs (e.g., IDs 15, 2, 6, and 17) and maintaining continuity despite challenges like multiple occlusions and overlapping object trajectories. The bounding boxes highlight the spatial positioning, motion, and orientation of the detected objects within the scene. For instance, ID 6, representing a vehicle on the right, is tracked continuously, showcasing the method's capability to handle steady movement over time. Color-coded bounding boxes (e.g., red, green, yellow, and cyan) distinguish between objects, improving visual clarity. Additionally, the algorithm prioritizes motion prediction and re-identification mechanisms, ensuring that objects temporarily occluded or entering the field of view are seamlessly incorporated into the tracking process.

Next, we choose the frames from the RADIATE dataset to showcase multi-object tracking results under various weather conditions such as foggy, rainy, and dark scenarios as shown in the figure 3. It underscores the robustness and adaptability of the tracking algorithm across diverse environmental scenarios. In (a) foggy condition 1 and (b) foggy condition 2, the tracking system demonstrates resilience to significant visibility constraints by accurately detecting and tracking vehicles (blue bounding boxes), leveraging LiDAR for depth information, and maintaining stable object IDs such as ID 4 and ID 3. Under (c) rainy condition 1 and (d) rainy condition 2, the tracker effectively handles moderate visibility reduction and reflections, consistently generating precise 3D bounding boxes and maintaining smooth object trajectories across frames. In (e) dark condition 1 and (f) dark condition 2, the algorithm compensates for poor illumination using LiDAR data, ensuring accurate object detection (red and blue bounding boxes) and maintaining stable IDs like ID 9 and ID 13, highlighting its ability to handle motion prediction and affinity calculations.

VI. Conclusion

In conclusion, the proposed tracking method demonstrates robust and accurate multi-object tracking capabilities, focusing specifically on vehicles in diverse and challenging real-world scenarios. The multimodal data utilization (combining 2D and 3D data) ensures reliable performance in adverse conditions, while its ability to manage occlusions and process real-time updates enhances its adaptability to dynamic environments. Experimental results highlight the method's superior detection accuracy, consistent ID assignment, and trajectory continuity, proving its suitability for autonomous driving applications in unpredictable weather conditions.

VII. Limitations and Future works

This study has some limitations. First, it focuses only on tracking vehicles, which means it doesn't include other important road users such as pedestrians, cyclists, bikes, etc. This could affect how accurate the tracking is in real-world situations where these other road users are also present. Second, although the system works well in real time, it requires a lot of computing power. This might make it difficult to use on devices with limited resources, like low-cost sensors or older hardware.

Future work can address these limitations by expanding object detection to include diverse road users, improving robustness under extreme conditions, optimizing the framework for low-power hardware, and validating the method on broader datasets to ensure adaptability and scalability. The RPN can also be enhanced to better represent pedestrians and cyclists by adapting the anchor box sizes and aspect ratios to align with the typical shapes and dimensions of these road users. Additionally, incorporating multi-scale feature fusion can improve the detection of smaller or more dynamic objects, such as pedestrians, ensuring greater accuracy and robustness for real-world autonomous driving applications.

Acknowledgment

The Basic Science Research Program supported this work through the National Research Institute Foundation of Korea (NRF) funded by the Ministry of Education(2021R1A6A1A03043144) and the Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0020536, The Competency Development Program for Industry Specialist). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A2C2008133).

References

[1] X. Weng, Y. Wang, Y. Man, and K. M. Kitani, "Gnn3dmot: Graph neural network for 3d multi-object tracking with 2d-3d multi-feature learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6499–6508.

[2] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan, "Track to detect and segment: An online multi-object tracker," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 352–12 361.

[3] A. Kim, A. Ošep, and L. Leal-Taixé, "Eagermot: 3d multi-object tracking via sensor fusion," in *2021 IEEE International conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 315–11 321.

[4] A. Shenoi, M. Patel, J. Gwak, P. Goebel, A. Sadeghian, H. Rezatofighi, R. Martin-Martin, and S. Savarese, "Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 335–10 342.

[5] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi, and C. C. Loy, "Robust multi-modality multi-object tracking," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2365–2374.

[6] N. F. Gonzalez, A. Ospina, and P. Calvez, "Smat: Smart multiple affinity metrics for multiple object tracking," in *Image Analysis and Recognition: 17th International Conference, ICIAR 2020, Póvoa de Varzim, Portugal, June 24–26, 2020, Proceedings, Part II 17*. Springer, 2020, pp. 48–62.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[8] M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "Radiate: A radar dataset for automotive perception in bad weather," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1–7.

[9] Z. Lu, V. Rathod, R. Votel, and J. Huang, "Retinatrack: Online single stage joint detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 668–14 678.

[10] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European conference on computer vision*. Springer, 2020, pp. 474–490.

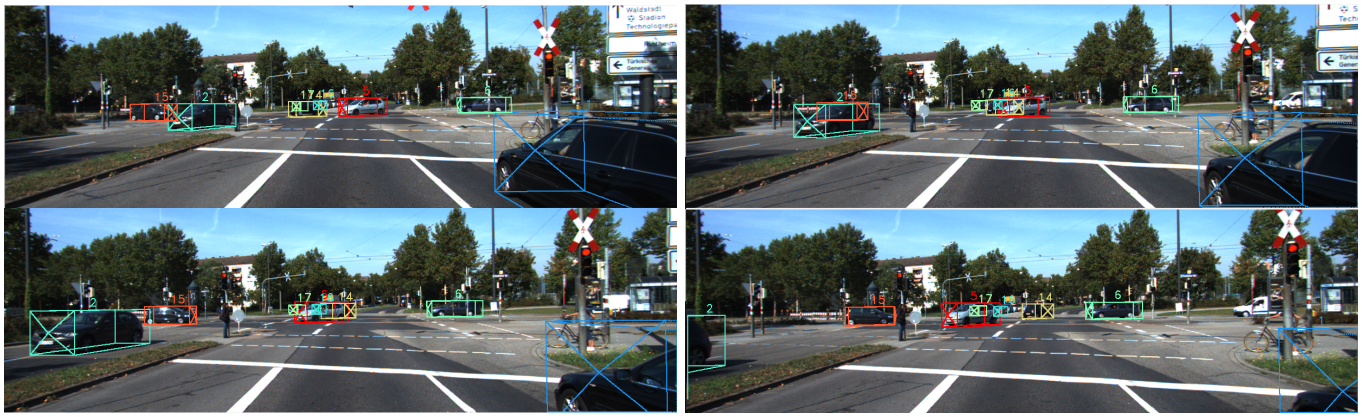

Fig. 2: Consistent vehicle tracking across consecutive frames of KITTI dataset in urban traffic scenarios where multiple occlusions occur across the road.

(a) Foggy Condition 1

(b) Foggy Condition 2

(c) Rainy Condition 1

(d) Rainy Condition 2

(e) Dark Condition 1

(f) Dark Condition 2

Fig. 3: Visualization of multi-object tracking results under various weather conditions like foggy, rainy, and dark. Each subfigure shows bounding boxes detected in challenging environmental scenarios: Insights from the RADIATE Dataset.

[11] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 145–161.

[12] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *European conference on computer vision*. Springer, 2020, pp. 107–122.

[13] K. Huang and Q. Hao, "Joint multi-object detection and tracking with camera-lidar fusion for autonomous driving," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 6983–6989.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[15] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 770–779.

[16] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Asian conference on computer vision*. Springer, 2010, pp. 709–720.

[17] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 35–52.

[18] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[19] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 784–799.

[20] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.