Althena

LiDAR-Camera 3D Object Detection Cross-Modal-Transformer (CMT)

MODEL CARD

Author name/s and affiliation	Till Beemelmanns, RWTH Aachen University
Contact details	till.beemelmanns@rwth-aachen.de
Link/s	Institute for Automotive Engineering
Affiliation logo	ika | RWTH AACHEN UNIVERSITY
Date	15/02/2026

Disclaimer

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for the

INTRODUCTION

MODEL DESCRIPTION

The Cross-Modal Transformer (CMT) is an end-to-end 3D multi-modal object detection architecture for autonomous driving. It avoids explicit geometric projections (like BEV) by implicitly aligning spatial data, directly taking image and point cloud tokens to output 3D bounding boxes. It is highly robust, resilient against sensor failures, achieves highest detection accuracy and is efficient.

- **Model Version:** *v0.1.0*
- **Model license:** Apache License
- **Citation:** https://arxiv.org/abs/2301.01283
- **Repository link:** Github

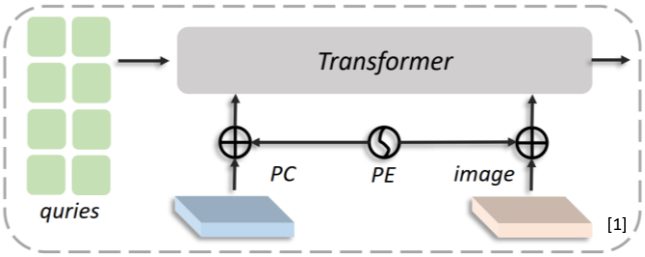

Figure 1 CMT Architecture Overview. In CMT, object queries interact in a transformer decoder directly with point cloud (PC) and image data, producing 3D bounding boxes.

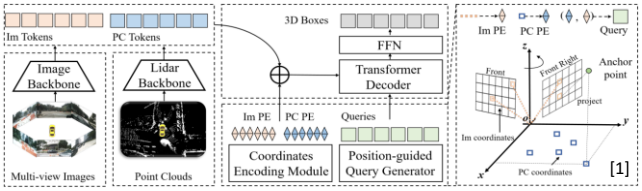

Figure 2 CMT Architecture Details. Multi-view camera images and LiDAR point clouds are processed via feature encoders to produce multi-modal tokens that are then enriched with a 3D positional encoding. 3D anchor points are projected into each modality and interact with the 3D features and produce 3D bounding boxes.

[1] "Cross Modal Transformer via Coordinates Encoding for 3D Object Dectection", Yan, Junjie and Liu, Yingfei and Sun, Jianjian and Jia, Fan and Li, Shuailin and Wang, Tiancai and Zhang, Xiangyu , ICCV 2023

MODEL DETAILS

MODEL ARCHITECTURE

- **Model type:** Multi-modal (LiDAR-Camera) transformer-based neural network.
- **Model task:** 3D Object Detection
- **Inputs:**
  - **Image Tokens:** From a 6-camera multi-view array, processed via a VoVNet backbone.
  - **Point Cloud Tokens:** From a roof-mounted LiDAR, processed via a VoxelNet backbone.

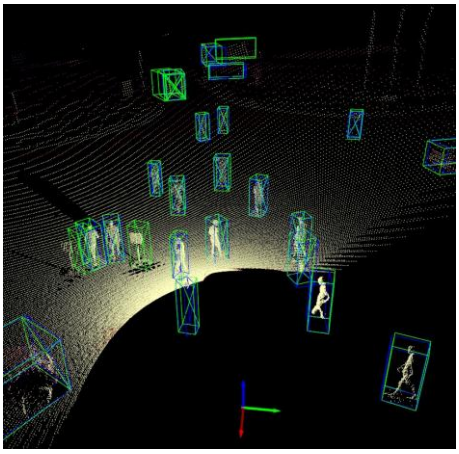

Figure 3 LiDAR Point Cloud with 3D Object Detections.

TRAINING DATA

- **Dataset description:** Trained on the nuScenes dataset, featuring 1,000 driving scenes (20s each) sampled at 2 Hz, containing 360-degree multi-modal telemetry and 3D bounding box annotations for 23 semantic classes.
- **Data splitting:** Standard nuScenes partitioning: mutually exclusive training, validation, and hidden test sets.
- **Data augmentation techniques:** Random 3D coordinate space augmentations (random global scaling, rotation, translation) synchronized with multi-view camera inputs. Masked-Modal Training, where modalities are masked during training.

EVALUATION DATA

- **Dataset description:** nuScenes hold-out validation and hidden test sets.

Outputs: 3D Bounding Boxes and Classes

- **Classification:** Predicted category for each detected object
- **Regression:** 3D bounding box parameters including position (x, y, z), dimensions (length, width, height), and orientation

INTENDED USE

USE CASES AND USERS

Intended use cases:

- **Urban traffic scenarios with mixed participants:** The model can detect cars, pedestrians, trucks, trailers, and two-wheelers in dense environments such as intersections, highways, and city streets, supporting collision avoidance and path planning.
- **Adverse conditions:** Since CMT is designed for robustness, it can help perception under challenging situations like occlusions, partial visibility, or sensor noise.
- **Fleet monitoring / cooperative driving research:** It can be used in R&D settings for evaluating collaborative perception or improving autonomous driving safety.

Target audience or users of the model:

- **Self-driving car developers and researchers:** AI researchers, perception engineers, and regulatory auditing bodies
- **Automotive companies / suppliers:** To evaluate robustness in production LiDAR-based systems.

Users to be analyzed by the model:

- **Traffic participants:** All road users within the public domain, e.g. pedestrians, cyclists, trucks and vehicles.

LIMITATIONS

- **Limitations or restrictions of the intended use:**
  - The model performance depends on the characteristics of the **nuScenes training dataset**. Generalization to very different environments (e.g., rural scenes, tunnels, or drastically different traffic distributions) may be limited.
  - The LiDAR-camera sensor setup specifications (e.g., FoV, resolution, camera sensor, number of LiDAR beams) should be close to the setup used during data collection.
  - If the mounting position of the Camera-LiDAR sensor setup is changed the accuracy of the model will suffer.
- **Guidance on how to address limitations or restrictions:**
  - Use LiDAR-camera sensor setup as in the training setup.
  - Do not change the sensor mounting position.
  - Deploy alongside complementary sensors like radar.
- **Identification of potential biases:**
  - The model may be biased because of limitations of diversity in data collection where some types of scenarios and objects are not well represented.

EVALUATION AND PERFORMANCE

METRICS

- **Performance Metrics:** nuScenes Detection Score (NDS), mean Average Precision (mAP) utilizing 3D center distance thresholds.
- **True Positive Metrics:**
  - **Average Translation Error (ATE)**: Euclidean center distance in 2D in meters.
  - **Average Scale Error (ASE)**: Calculated as *1 - IOU* after aligning centers and orientation.
  - **Average Orientation Error (AOE)**: Smallest yaw angle difference between prediction and ground-truth in radians. Orientation error is evaluated at 360 degree for all classes except barriers where it is only evaluated at 180 degrees.
  - **Average Velocity Error (AVE)**: Absolute velocity error in m/s. Velocity error for barriers and cones are ignored.
  - **Average Attribute Error (AAE)**: Calculated as *1 - acc*, where acc is the attribute classification accuracy. Attribute error for barriers and cones are ignored.
- **Relevant thresholds or criteria:** Achieving high NDS/mAP while maintaining an inference speed that exceeds the hardware sampling rate.
- **Evaluation method:** Blind, distributed evaluations submitted securely to the official nuScenes test server.

RESULTS

- **Results of the performance evaluation:**
  - **CMT-L (LiDAR-only):** 70.1% NDS, 65.3% mA
  - **CMT (Multi-modal):** 74.1% NDS, 72.0% mAP
- **CMT (Multi-modal) Class-Wise Results:**

Object Class	AP	ATE (m)	ASE (1-IOU)	AOE (rad)	AVE (m/s)	AAE (1-acc)
Car	0.880	0.172	0.134	0.046	0.216	0.118
Truck	0.633	0.333	0.174	0.045	0.269	0.123
Bus	0.754	0.279	0.156	0.035	0.391	0.313
Trailer	0.654	0.451	0.150	0.828	0.182	0.095
Construction vehicle	0.373	0.617	0.372	0.946	0.096	0.061
Pedestrian	0.879	0.148	0.289	0.277	0.186	0.112
Motorcycle	0.791	0.203	0.219	0.183	0.496	0.043
Bicycle	0.606	0.222	0.261	0.380	0.233	0.032
Traffic cone	0.847	0.131	0.328	n/a	n/a	n/a
Barrier	0.782	0.237	0.269	0.031	n/a	n/a

- **Performance on specific subsets of dataset:** n/a
- **Performance metrics interpretation:** n/a

ETHICAL CONSIDERATIONS

FAIRNESS

- **Fairness metrics:**
  Standard statistical fairness metrics (e.g., demographic parity) cannot be calculated due to nuScenes' "fairness through blindness" policy (omission of demographic labels). Audits rely on proxy metrics like variance across lighting and geography
- **Remaining biases or limitations:**
  Potential unmeasured bias against minority populations or individuals with darker skin tones under low-light conditions remains mathematically untestable without retroactively annotating the dataset.

PRIVACY

- **Data collecting and storing:** Raw public telemetry undergoes automated cryptographic blurring of human faces and vehicle license plates before storage using high-recall 2D object detectors. Furthermore, LiDAR data inherently preserves privacy; due to the extreme sparsity and limited spatial resolution of the point clouds, it only captures geometric outlines and does not capture personal information, gender, or visual identification.
- **Users' access to data:**
  Captured individuals cannot opt-out beforehand but can utilize a public privacy take-down form to request manual redaction if automated blurring fails.

ACCOUNTABILITY

- **Model's decisions transparency through explainability techniques:**
  Relies on advanced saliency maps and visual explanations generated by aggregating multi-head attention weights across all transformer layers. Tracing 3D queries back to sensor inputs helps visualize network prioritization.
- **Reviewing and auditing processes:** Deployment in the EU is classified as "high-risk" under the AI Act, requiring rigorous third-party conformity assessments and data governance audits.
- **Limitations or uncertainties:**
  Ambiguous legal consequences regarding liability.

USAGE AND LIMITATIONS

USAGE AND RECOMMENDATIONS

- **Model usage instructions:**
  The model can run **as a ROS2 Jazzy node** in an automated-driving stack **or as a standalone inference service/script**. Preprocessing steps for LiDAR and camera data are required; Multi-modal inputs demand precise intrinsic/extrinsic calibration matrices; see the inference repository/configs for camera pipeline, voxel sizes, parameters and operations.

- **Input data format and preprocessing:**
  - Point Cloud Array: Fields: (X, Y, Z, Intensity), the coordinates must be in the LiDAR Frame
  - Multi-View Camera images, normalized and pre-processed
  - Extrinsic / Intrinsic calibration parameters

- **Output interpretation:**
  - The model predicts **oriented 3D bounding boxes** with **classification** scores.
  - **Regression: box center** (x, y, z), **dimensions** (l, w, h), **velocities** (v_x, v_y) and **orientation** (yaw).
  - **Classification:** Per-box class probabilities (car, pedestrian, truck, trailer, two-wheeler, …). Predicts classes according to the nuScenes dataset class labels.
  - **Post-processing** uses **confidence thresholds** to filter low confidence boxes.

LIMITATIONS AND RISKS

- **Technical limitations:**
  - **Domain Shift**: A domain shift w.r.t. to the shape of vehicles might hurt the performance (North American Cars vs. European Cars); Special Types of vehicles (On Rails, Trains) might not be detected.
  - **Different Sensor**: The model might not be applicable to other sensor types, sensor configurations or sensor orientations; A similar input distribution as the train dataset is necessary.
  - **Bad weather conditions** (snow, rain, fog): might disturb the LiDAR and camera sensor and thus the model might not be able to make reliable detections.

- **Ethical or legal limitations and restrictions:**
  - The EU AI Act classifies these systems as high-risk components. Commercial deployment demands rigorous type approval, compliance with General Safety Requirements (Regulation (EU) 2019/2144), mandatory continuous human oversight, and absolute resilience against cyber-attacks, exposing deployers to extreme legal liability.