Transformer-Based Sensor Fusion for Autonomous Driving: A Survey

Apoorv Singh
Motional
Carnegie Mellon University
apoorv.singh@motional.com

Abstract

Sensor fusion is an essential topic in many perception systems, such as autonomous driving and robotics. According to the dataset leaderboards, the transformers-based detection head and CNN-based feature encoder to extract features from raw sensor data has emerged as one of the top performing sensor-fusion 3D-detection-framework. In this work, we provide an in-depth literature survey of transformer-based 3D-object detection tasks in the recent past, primarily focusing on sensor fusion. We also briefly review the Vision Transformers (ViT) basics so readers can easily follow through with the paper. Moreover, we also briefly go through a few non-transformer-based, less-dominant methods for sensor fusion for autonomous driving. In conclusion, we summarize the role that transformers play in the domain of sensor fusion and also provoke future research in the field.

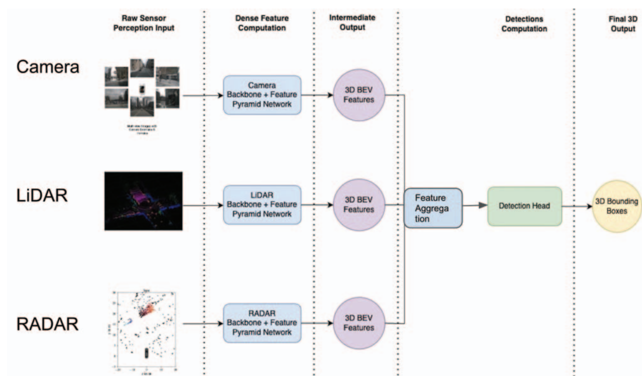

Figure 1. An overview architecture diagram of modern sensor-fusion model. Transformers-based head (Green-block); CNN-based Feature extractors (Blue-block) for predicting 3D Bird's Eye View (BEV) bounding box (Yellow-block) with intermediate BEV features per-sensor (Purple-block) as defined in Section 3. This sensor-fusion setup takes input from multi-view cameras, LiDARs and RADARs.

1. Introduction

Camera and radar fusion is a crucial technology in the field of autonomous driving. It involves combining the information from cameras and radar sensors to enhance the perception and decision-making capabilities of autonomous vehicles. Each sensor type has its own strengths and weaknesses, and by fusing their data together, autonomous vehicles can obtain a more comprehensive and accurate understanding of their surroundings.
Here's how camera and radar fusion works and why it's important:

- Sensor Complementarity: Cameras are excellent at capturing high-resolution images and identifying visual details such as lane markings, traffic signs, and objects' shapes. However, they can struggle in adverse weather conditions like heavy rain, fog, or low light. On the other hand, radar sensors are effective in detecting the speed and distance of objects, making them reliable in challenging weather and lighting conditions. By fusing the data from both sensors, the system can compensate for the limitations of each sensor type.

- Redundancy and Reliability: Having redundant sensors improves the reliability and safety of the autonomous system. If one sensor type fails or provides inaccurate information, the other sensor can validate or correct the data. This redundancy reduces the risk of misinterpretation and enhances the overall safety of the vehicle.

- Object Detection and Tracking: Combining camera and radar data enables more accurate detection and tracking of objects in the environment. For instance, a camera might detect a pedestrian, while radar can determine the pedestrian's speed and distance. By combining this information, the vehicle can better predict the pedestrian's behavior and make safer decisions.

- Filling Sensor Blind Spots: Cameras have blind spots, especially around the vehicle's perimeter. Radar sensors can help fill these blind spots by detecting objects that might be outside the camera's field of view, such as vehicles in adjacent lanes or objects approaching from behind.

To achieve effective camera and radar fusion, sophisticated sensor fusion algorithms are employed. These algorithms process and integrate the data from cameras and radar sensors to generate a unified representation of the environment. Machine learning techniques, such as deep learning and Bayesian filtering, are often used to fuse the data and make sense of the combined information.

What makes the sensor-fusion problem so hard? Sensor data of different modalities usually have large discrepancies in the data distribution and the difference in coordinate space for each sensor. For example, natively, LiDAR is in Cartesian coordinate space; RADAR is in polar coordinate space, and images are in perspective space. Spatial misalignment introduced by different coordinate frames makes it hard to merge these modalities. Another issue with multi-modal input is that there would be asynchronous timelines when the camera and LiDAR feed are available to the ML network.

While Deep-CNNs can be used to capture global context within a single modality, it is non-trivial to extend them to multiple modalities and accurately model interactions between pairs of features. To overcome this limitation, attention mechanisms of transformers are used to integrate global contextual reasoning about the 2D scene directly into the feature extraction layers of modalities. Recent advances in sequential modeling and audio-visual fusion [8] demonstrates that Transformers-based architecture is very competent in modeling the information interaction for sequential or cross-modal data.

The main contributions of this work can be summarized as follows:

- An overview for Vision Transformers(ViT) background to get readers up-to-speed with the theoretical background prerequisite for going through the latest trending sensor fusion methods in Section 3
- Conduct an in-depth survey about the recent State-of-the-art(SoTA) methods for object detection tasks with sensor fusion, focusing on Transformers-based approaches in Section 4
- Go through quantitative analysis of discussed SoTA methods and provoke future research work in the space in Section 5 and 6.

2. Related Work

Fusion Levels: Recently, multi-sensor fusion arouses increased interest in the 3D-detection community. The existing approach can be classified into *detection-level*, *proposal-level*, and *point-level* fusion methods, based on how early or late in the process we fuse different modalities viz., Cameras, RADARs, LiDARs, et al.

Detection-level a.k.a. late-fusion has emerged as the simplest form of fusion since each modality can process their own BEV detections individually which can be later post-processed to aggregate and remove duplicate-detections using Hungarian cost-matching algorithm and Kalman-filtering. However, this approach cannot leverage the fact that each sensor can also contribute to different attributes within a single bounding-box prediction. CLOCS [15] leverages modalities in the form they can naturally perform detection tasks, i.e., LiDAR for 3D object detection and cameras for 2D detection tasks. It operates on both the output candidates before Non-maximum suppression. It uses geometric consistencies between two sets of predictions to eliminate False-positives (FP), as it is highly unlikely that the same FP would be detected simultaneously with different modalities.

Point-level a.k.a. early-fusion is augmenting LiDAR point-cloud with the camera features [21, 12]. This method finds hard associations between LiDAR points and images using transformation matrices. However, camera-to-LiDAR projections are semantically lossy as the point's sparsity limits fusion quality. This approach suffers when there is even a slight error in the calibration parameters of the two sensors.

Proposal-level a.k.a. deep-fusion, is the most researched-upon method in the literature. Advances in transformers [3, 22, 11] have unlocked the possibilities of how intermediate features can interact despite being cross-domain from different sensors. Representative work like MV3D [4] proposes initial bounding boxes from LiDAR features and iteratively refines them using camera features. BEVFusion [14] generates camera-based BEV features as highlighted in [17, 16, 20, 19]. Camera and LiDAR modality are concatenated in the BEV space, and a BEV decoder [26] is used to predict 3D boxes as a final output. In TransFuser [7], single-view image and LiDAR's BEV representation are fused by transformers in the encoder at various intermediate feature maps. This results in an encoder's 512-dimensional feature vector output constituting a compact local and global context representation. In addition, this paper feeds the output to a GRU (Gated Recurrent Unit) and predicts differentiable ego-vehicle way-points using L1 regression loss. 4D-Net [18], in addition to being multi-modal, adds a temporal dimension to the problem as the 4^{th} dimension. They first extract in-time features of cameras and LiDAR[9] individually. To add the different contexts of image representation, they collect image features in three levels: high-resolution image, low-resolution image, and video. Then they fuse cross-modal information using a transformation matrix to fetch 2D context given a 3D-pillar center, defined by the center-point of the BEV grid cell (x^o, y^o, z^o).

3. Transformers Based Fusion Network Background

This approach can be divided into 3-steps: 1. Apply CNN-based backbones to extract spatial features from all the modalities individually. 2. Small set of learned embeddings (object queries/ proposals) are iteratively refined in the transformers module to generate a set of predictions of 3D boxes. 3. Set-based loss is calculated over the predictions and groud-truth. This architecture is presented in Fig. 1

3.1. Backbone: Feature Extractor

Cameras: Multi-camera images are fed into the backbone network (e.g., ResNet-101) and FPN [13] and obtain features $\{\{F^{ij}_{images} \in \mathbb{R}^{C*h*w}\}_{i=1}^{N_{view}}\}_{j=1}^{M}$, where M is the number of feature levels from FPN; N_{view} is the number of cameras in surround-view; $h * w$ is the image-view feature dimensions.
LiDAR: Generally, a voxelNet[28] with 0.1m voxel size or PointPillar[9] with 0.2m pillar size is used to encode points. After 3D backbone and FPN [13], a multi-scale BEV feature maps $\{F^{j}_{lidar} \in \mathbb{R}^{C*H_j*W_j}\}_{j=1}^{M}$ is obtained.
RADARS: Let's consider N Radar points $\{r_j\}_{j=1}^{N} \in \mathbb{R}^{C_{radar}}$, where C_{radar} is the number of features of the radar points, such as location, intensity and speed. A shared MLP Φ_{radar} is used to obtain per-point features $F^i_{rad} = \Phi_{rad}(r_j) \in \mathbb{R}^C$.

3.2. Query Initialization

In seminal work [3], sparse queries are learned as a network parameter and represent the entire training data. This query type takes longer, i.e., more sequential decoder layers (typically qty. 6) to iteratively converge to the actual 3d-objects in the scene. However, recently input-dependent queries[25] are considered as a better initialization strategy. This strategy can bring a 6-layered transformer decoder down to even a single-layered decoder layer. Transfusion [1] uses center-heatmap as queries, and BEVFormer [11] introduced dense queries as an equally-spaced BEV grid.

3.3. Transformers Decoder

Repeated blocks of Transformer decoders are used sequentially to refine object proposals in a ViT model, where each block consists of self-attention and cross-attention layers. *Self-attention* between object queries does pairwise reasoning between different object candidates. *Cross-attention* between the object queries and the feature-map aggregates relevant context into the object queries based on the learned attention mechanism. Cross-attention is the slowest step in the chain because of the huge feature size, but techniques [29] had been proposed to reduce the attention window. After these sequential decoders, d-dimensional refined queries are independently decoded with an FFN layer following [26]. FFN predicts the center-offset from the query position $\delta x, \delta y$, bounding box height as z, dimensions l, w, h as $log(l), log(w), log(h)$, yaw angle α as $\sin(\alpha)$ and $\cos(\alpha)$ and velocity as v_x, v_y; lastly per-class probability $\hat{p} \in [0,1]^K$ is predicted for K semantic classes.

3.4. Loss Computations

Bipartite matching between set-based predictions and ground truths through the Hungarian algorithm is used, where matching cost is defined by:

$$C_{match} = \lambda_1 L_{cls} + \lambda_2 L_{reg} + \lambda_2 L_{IoU} \quad (1)$$

where, L_{cls} is a binary cross-entropy loss; L_{reg} is an L1 loss; L_{IoU} is a box IoU loss. $\lambda_1, \lambda_2, \lambda_3$ are network hyper-parameters.

4. Transformers-based Sensor Fusion

TransFusion [1] tackles modality misalignment issue with the soft association of features. The first decoder layer constitutes sparse-queries generation from LiDAR BEV features. The second decoder layer enriches LiDAR queries with the image features with soft associations by leveraging locality inductive bias with cross-attention only around the bounding box decoded from the query. They also have an image-guided query initialization layer.
FUTR3D [5] is closely related to [22]. It is robust to any number of sensor modalities. MAFS (Modality Agnostic Feature sampler) takes in the 3D queries and aggregates features from multi-view cameras, high-res lidars, low-res lidars, and radars. Specifically, it first decodes the query to get a 3D coordinate, which is then used as an anchor to gather features from all the modalities iteratively. BEV features are used for LiDAR and cameras; however, for RADARS, the top-k nearest radar points are picked in MAFS. For each query i, all these features F are concatenated as below where Φ is an MLP layer:

$$F^i_{fused} = \Phi_{fused}(F^i_{lidar} \oplus F^i_{camera} \oplus F^i_{radar}) \quad (2)$$

CMT: Cross-Modal Transformers [23] encodes 3D coordinates into the multi-modal tokens by the *coordinates encoding*. The queries from the *position-guided query generator* interact with the multi-modal tokens in the transformer decoder and then predict the object parameters. *Point-based query denoising* is further introduced to accelerate the training convergence by introducing local prior.

Table 1. Performance comparison on the nuScenes **test set**. The metrics key is defined in Section 5.

Methods	NDS(%) ↑	mAP(%) ↑	mATE(cm) ↓	mASE(%) ↓	mAOE(rad) ↓	mAVE(cm/s) ↓	mAAE(%) ↓
CMT [23]	73.0	70.4	29.9	24.1	32.3	24.0	11.2
BEVFusion [14]	72.9	70.2	26.1	23.9	32.9	26.0	13.4
TransFusion [1]	71.7	68.9	25.9	24.3	35.9	28.8	12.7
UVTR [10]	71.1	67.1	30.6	24.5	35.1	22.5	12.4

UVTR: Unifying Voxel-based Representation with Transformer [10] unifies multi-modality representations in the voxel space for accurate and robust single or cross-modality 3D detection. Modality-specific space is first designed to represent different inputs in the voxel space without height compression to alleviate semantic ambiguity and enable spatial connections. This is a more complex and information-packed representation than the other BEV approaches. For *image-voxel space*, perspective view features are transformed into the predefined space with view-transform, following [17]. CNN-based voxel encoder is introduced for multi-view feature interactions. For *point-voxel space*, 3D points can be naturally transformed into voxels. Sparse convolutions are used over these voxel features to aggregate spatial information. With accurate positions in the point cloud, the semantic ambiguity in z direction is much reduced compared to the images.

LIFT: LiDAR Image Fustin Transformer [27] is capable to align the 4D spatiotemporal cross-sensor information. In contrast to [18], it exploits integrated utilization of sequential multi-modal data. For sequential data processing, they use the prior vehicle pose to remove the effects of ego-motion between temporal data. They encode the lidar frames and camera images as sparsely located BEV grid features and propose a sensor-time 4D attention module to capture mutual correlation.

DeepInteraction: [24], follows a slightly different approach than its other counterparts. It claims that the previous approaches are structurally restricted due to their intrinsic limitations of potentially dropping off a large fraction of modality-specific representational strengths due to largely imperfect information fusion into a unified representation as in [21, 14]. Instead of deriving a fused single-BEV representation, they learn and maintain two modality-specific representations throughout to enable inter-modality interaction so that both information exchange and modality-specific strengths can be achieved spontaneously. They refer to it as a multi-input-multi-output (MIMO) structure, taking as input two modality-specific scene representations independently extracted by LiDAR and image backbones and producing two refined representations as output. This paper includes DETR3D-like [22] queries which are sequentially updated from LiDAR and vision features with sequential cross-attention layers in the transformer-based decoder layer.

Auto-align [6] they model a mapping relationship between image and point-cloud with a learnable alignment map instead of establishing a deterministic correspondence for sensor projection matrix as done in the other approaches. This map enables the model to automate the alignment of non-homogeneous features in a dynamic data-driven manner. They leverage cross-attention modules to adaptively aggregate pixel-level image features for each voxel.

5. Quantitative Analysis

Here we compare previously discussed methods on nuScenes [2], a large-scale multi-modal dataset, which is composed of data from 6 cameras, 1 LiDAR, and 5 RADARs in Table 1. This dataset has 1000 scenes total and is divided into $700/150/150$ scenes as train/validation/test sets, respectively. **Cameras:** Each scene has $20s$ video frame with 12 FPS. 3D bounding boxes are annotated at $0.5s$. Each sample includes six cameras. **LiDAR:** A 32-beam LiDAR with 20FPS is also annotated at every $0.5s$. **Metrics**: We follow the nuScenes official metrics. Keys are as follows: nuScenes Detection Score (NDS), mean Average Precision (mAP), mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error(mAOE), mean Average Velocity Error (mAVE) and mean Average Attribute Error (mAAE).

6. Conclusion

For the autonomous vehicle's perception reliability, accurate 3D-object detection is one of the key challenges we must solve. Sensor fusion helps make these predictions more accurate by leveraging the pros of all the sensors on the platform. Transformers have emerged as one of the top methods to model these cross-modal interactions, especially when sensors operate in different coordinate spaces, making it impossible to align perfectly.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022.

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Gi-

ancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[5] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*, 2022.

[6] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinghong Jiang, Feng Zhao, Bolei Zhou, and Hang Zhao. Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection. *arXiv preprint arXiv:2201.06493*, 2022.

[7] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[8] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[9] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

[10] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *arXiv preprint arXiv:2206.00630*, 2022.

[11] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 1–18. Springer, 2022.

[12] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 641–656, 2018.

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[14] 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.

[15] Su Pang, Daniel Morris, and Hayder Radha. Clocs: Camera-lidar object candidates fusion for 3d object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10386–10393. IEEE, 2020.

[16] Jongwoo Park, Apoorv Singh, and Varun Bankiti. 3m3d: Multi-view, multi-path, multi-representation for 3d object detection. *arXiv preprint arXiv:2302.08231*, 2023.

[17] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020.

[18] AJ Piergiovanni, Vincent Casser, Michael S Ryoo, and Anelia Angelova. 4d-net for learned multi-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15435–15445, 2021.

[19] Apoorv Singh. Vision-radar fusion for robotics bev detections: A survey. *arXiv preprint arXiv:2302.06643*, 2023.

[20] Apoorv Singh and Varun Bankiti. Surround-view vision-based 3d detection for autonomous driving: A survey. *arXiv preprint arXiv:2302.06650*, 2023.

[21] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020.

[22] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022.

[23] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer via coordinates encoding for 3d object dectection. *arXiv preprint arXiv:2301.01283*, 2023.

[24] 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.

[25] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.

[26] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[27] Yihan Zeng, Da Zhang, Chunwei Wang, Zhenwei Miao, Ting Liu, Xin Zhan, Dayang Hao, and Chao Ma. Lift: Learning 4d lidar image fusion transformer for 3d object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17151–17160, 2022.

[28] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 4490–4499, 2018.

[29] 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.