MuStD: A MultiStream Detection Network for 3D Object Detection

Muhammad Ibrahim [1], Naveed Akhtar [2], Haitian Wang[1], Saeed Anwar[3], and Ajmal Mian [1]

***Abstract*— Multimodal approaches that fuse data from LiDAR and RGB cameras have significant potential in improving 3D object detection accuracy. However, existing fusion methods often fail to fully integrate 3D geometric and RGB spatial information. To overcome these limitations, we propose MultiStream Detection (MuStD) network, designed to optimize the integration of LiDAR and RGB data. MuStD network has three parallel streams: the LiDAR-PillarNet stream for extracting sparse 2D pillar features, the LiDAR-Height Compression stream for Bird's-Eye View (BEV) features, and the 3D Multimodal (MM) stream, which combines RGB and LiDAR features using UV mapping and polar coordinate indexing. This novel architecture effectively captures both geometric and texture information, addressing the challenges of feature fusion. Extensive experiments on the challenging KITTI Object Detection Benchmark demonstrate the MuStD network's superior performance. For 3D car detection, MuStD achieved a mean precision (AP) of 80.78% in the Hard category and a mean AP of 85.3% across all categories. For 2D car detection, MuStD achieved an AP of 94.04% in the Hard category and mean AP of 96.39% . These results highlight our method's robustness in complex urban environments, advancing the state-of-the-art in multimodal 3D object detection.For detailed information, please refer to MuStD GitHub repository.**

I. INTRODUCTION

Accurate outdoor 3D object detection is crucial for reliable autonomous navigation [1], [2]. Currently, LiDAR stand out as the primary sensor to enable that [3], [4], [5]. However, sampling sparsity and partial measurements caused by occlusion compromise 3D objection detection using only the LiDAR data [6], [7], [8]. This limitation can potentially be addressed by fusing LiDAR measurements with RGB camera inputs [9], [10], [11]. The latter are known for accurate acquisition of high resolution texture information, which complements LiDAR input for improved 3D object detection.

Currently, fusion of LiDAR and RGB data for 3D object detection is gaining rapid traction [9], [10], [12], [11]. Whereas early methods have typically employed late fusion at regions of interest or Bird's Eye View (BEV) [10], [12], the more recent approaches use depth completion to enrich the LiDAR's sparse data with pseudo points [13], [11]. The virtual points help mitigating the sparsity issue, particularly for distant and occluded objects, resulting in improved geometric details for object detection.

Commonly, contemporary fusion-based 3D detection methods rely on isolated strategies, such as UV mapping or polar transformation [6], [9], [10]. The former is effective in aligning LiDAR points with 2D image; however, it fails to adequately encode the spatial relationships of point clouds. This leads to under-utilization of critical geometric details of the scene. Similarly, while polar transformations is proficient at encoding spatial orientation and distance information, it lacks in effectively integrating the rich semantic and texture information provided by RGB images [14], [15], [16]. These isolated approaches fail to effectively merge LiDAR's 3D spatial context with RGB's dense semantic information, leading to sub-optimal detection performance, especially in complex scenarios including occluded or distant objects.

In this wok, we propose a MultiStream Detection (MuStD) network that effectively integrates LiDAR point clouds with RGB images for enhanced 3D object detection. Illustrated in Fig. 1, our technique addresses multi-modal fusion by optimizing geometric and spatial information extraction and improving object orientation estimation. Our approach is structured as three parallel data processing streams where each stream eventually contributes to the fusion of rich geometric details of LiDAR input and textural information of RGB images. Our contributions are as follows.

- **Multistream Detection (MuStD) Network**: We propose the MuStD network containing three parallel streams combining LiDAR and RGB data for enhanced 3D object detection. MuStD effectively leverages the strengths of each modality - LiDAR for spatial geometry and RGB for texture detail - while mitigating their limitations through comprehensive feature fusion.
- **3D Multimodal (MM) stream**: A central innovation in our MuStD network is the 3D MM stream, which effectively integrates the UV mapping and polar coordinate indexing. UV mapping aligns 3D LiDAR points with 2D image features to capture fine-grained texture and appearance details from the RGB modality. Concurrently, polar coordinate indexing encodes the spatial orientation and depth relationships in the scene, enhancing the geometric representation of objects.
- **UV-Polar block**: We propose a novel block that projects 3D sparse convolution features onto a UV image and polar space, creating 2D grid representations. These are processed with 2D sparse convolutions and then merged with the original 3D sparse features, resulting in a unified feature set that integrates both position and orientation information of objects.

Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (project # FT210100268). and Naveed Akhtar is a recipient of the Australian Research Council Discovery Early Career Researcher Award (project number DE230101058) funded by the Australian Government, both funded by the Australian Government.

[1]Department of Computer Science, The University of Western Australia. muhammad.ibrahim@, 23815631@student, ajmal.mian@) uwa.edu.au

[2] School of Computing & Information Systems, The University of Melbourne, naveed.akhtar1@unimelb.edu.au

[3]The Australian National University, saeed.anwar@anu.edu.au

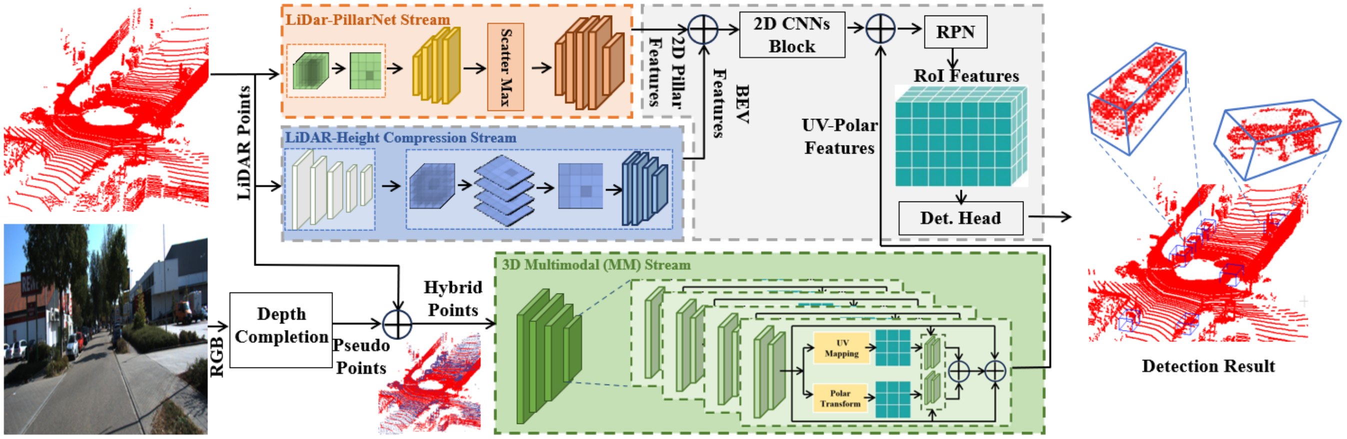

Fig. 1. Overview of the proposed MuStD network for 3D object detection comprising three key components. (1) The PillarNet stream converts LiDAR frames into 2D Pillar features, followed by an MLP and 2D sparse convolution for final feature extraction. (2) The LiDAR-Height compression stream uses 3D sparse convolutions to extract features, which are then transformed to bird's eye view (BEV) before being concatenated with UV-polar features and processed by a 2D CNN block. (3) The UV-Polar based 3D MM network processes both RGB and LiDAR data, featuring MM layers for extracting 2D and 3D sparse features. These MM features are combined with features from other streams and processed by a Detection Head for final object detection.

Extensive experiments show that the proposed MuStD network achieves state-of-the-art results on the KITTI object detection leader board.

II. Related Work

In 3D detection, LiDAR-only methods still remain popular due to the spatial precision of LiDAR input [17], [18], [19], [20], [21], Techniques such as PointPillars [22] and SECOND [23], perform real-time processing by converting point clouds into 2D pseudo-images and then apply 2D convolutions. This reduces computational cost while maintaining detection accuracy. Advanced models like PV-RCNN [24] and Voxel-RCNN [25] enhance performance by combining point-wise and voxel-wise features using sparse 3D convolutions and attention mechanism, achieving improved results.

Recent research has also focused on optimizing the integration of LiDAR and RGB data to improve 3D detection [26]. Fusing the two modalities is aimed at exploiting precise geometric information from the LiDAR input and dense textural details from the images. Early fusion methods like AVOD [27] and MV3D [1] fuse LiDAR and image features after their independent extraction. More recently, techniques such as 3D-CVF [28] align features from both modalities at the feature level. To address the sparsity of LiDAR input, methods such as SFD [5] and VirConv [29] use point cloud completion to generate dense pseudo points. However, this also introduces noise, particularly at object boundaries. VPFNet [7] refines feature fusion by selectively combining features based on spatial reliability, mitigating noise and improving detection performance.

Multimodal fusion for 3D object detection mainly faces challenges in aligning features across modalities and suppressing noise [30], [31]. NRConv [32] addresses this with noise-resistant convolutions that enhance feature extraction by reducing the influence of noisy data. The 2DPASS framework [33] improves segmentation and detection by integrating 2D semantic priors into 3D point clouds. Other advances focus on handling the complexities of real-world scenarios. For instance, [34] employs transformation-equivariant convolutions to improve robustness against rotation and reflection variations in autonomous driving. Similarly, Graph-VoI [8] uses graph neural networks to model complex object relationships, leveraging both geometric and semantic features.

Despite the advances in combining RGB and LiDAR modalities, effective outdoor 3D object detection remains a widely open problem. We integrate advanced mapping and efficient convolutions while optimizing multimodal fusion to achieve state-of-the-art performance for this task.

III. Methodology

Figure 1 illustrates the schematics of our MultiStream Detection (MuStD) network. The proposed network utilizes three parallel data processing streams. (a) 3D Multimodal (MM) stream, which processes 3D hybrid points using our proposed UV-polar block at each layer. (b) The LiDAR-Height Compression stream that extracts sparse 3D features from LiDAR frames, compresses them using a height module, and processes them with 2D CNN blocks to capture spatial relationships and object geometry. (c) The LiDAR-PillarNet stream, which converts 3D point clouds into 2D polar representations through pillar-based voxelization and neural networks, effectively extracting robust geometric features such as object orientation and localization. Details of MuStD network are given below.

A. *3D Multimodal (MM) stream*

The 3D MM stream (see Fig. 2), a key innovation of the MuStD Network, is specifically designed to enhance 3D object detection by integrating LiDAR and camera data. The novelty lies in its ability to extract 3D sparse features that include both UV and polar information, crucial for accurately determining an object's location and orientation. This stream is composed of a series of UV-Polar blocks with channel sizes of 16, 32, 64, and strides of 1, 2, 2, 2, processing hybrid 3D points through the stages discussed below.

Hybrid Points Generation: In this step, pseudo point clouds are first generated from RGB images using point cloud completion [4]. To improve efficiency, about 80% of these pseudo points are discarded and the remaining points are fused with LiDAR point clouds to create Hybrid 3D points

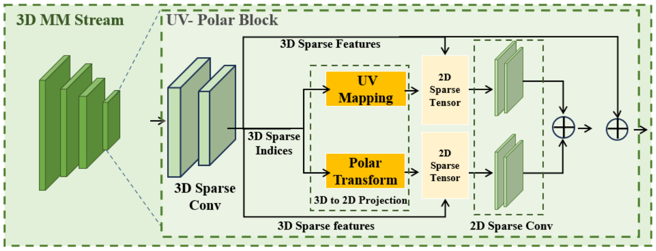

Fig. 2. 3D MM stream of MuStD network integrates UV mapping and polar coordinate indexing. UV mapping aligns 3D LiDAR points with 2D image features, capturing texture and appearance, while polar coordinate indexing encodes orientation and depth for improved geometric representation.

with enriched spatial and geometric information. The hybrid points are then processed by the 3D MM stream.

UV-Polar Block: This block serves as the foundational unit of the MM stream, designed to capture detailed geometric, orientation and spatial features of objects. This block initially processes the hybrid points using 3D sparse convolutional layers to extract high-level 3D sparse features. These features are then projected to image and polar spaces through two parallel transformations: UV Mapping and Polar Transform.

The UV Mapping process projects 3D points to a 2D plane, aligning them with their corresponding RGB image features. This is achieved by calculating the UV coordinates (u, v) as $\left(\frac{x}{z}, \frac{y}{z}\right)$, where x, y, and z represent the 3D coordinates of each point. The input 3D sparse feature, originally of size $F_{3D} \in \mathbb{R}^{H \times W \times D \times C}$, is projected onto a 2D feature space of size $F_{UV} \in \mathbb{R}^{1600 \times 600 \times C}$. The projected 2D features are then processed through a series of 2D sparse convolutional layers to capture texture-rich details, enabling the network to identify fine-grained 2D features crucial for accurate object detection.

The Polar Transform projects the 3D sparse features to polar coordinates $F_P \in \mathbb{R}^{1600 \times 600 \times C}$. These polar 2D features are then processed through 2D sparse convolutional layers to extract critical information related to the orientation and distance of objects, enhancing the network's ability to detect and localize objects within the scene. The conversion is defined as

$$(r, \theta, \phi) = (\sqrt{x^2 + y^2 + z^2}, \tan^{-1}(\frac{y}{x}), \tan^{-1}(\frac{z}{\sqrt{x^2 + y^2}})),$$

where r is the radial distance, θ is the azimuth and ϕ is the elevation angle. This representation helps manage variations in scale and rotation commonly found in real-world data.

Finally, polar, UV and input 3D sparse features are fused within the block to form a comprehensive feature set F_{MM} at each layer of the stream. The equation below represents F_{MM} for the $l+1^{\text{th}}$ layer, where $\mathbf{X}$, $\mathbf{W}$, $\mathcal{U}$, $\mathcal{P}$, $\oplus$, $\circledast$, and $\star$ denote input feature, kernel weights, UV mapping, polar transformation, concatenation, 2D sparse convolution and 3D sparse convolution operations respectively.

$$\mathbf{F}_{\text{MM}}^{(l+1)} = (\mathbf{X} \star \mathbf{W}) \oplus (\mathcal{U}(\mathbf{X}) \circledast \mathbf{W}) \oplus (\mathcal{P}(\mathbf{X}) \circledast \mathbf{W}).$$

B. LiDAR Height Compression stream

The LiDAR Height Compression stream is designed to efficiently extract geometric features from the raw LiDAR point cloud. It consists of 3D sparse convolution blocks, height compression followed by a series of 2D CNN layers.

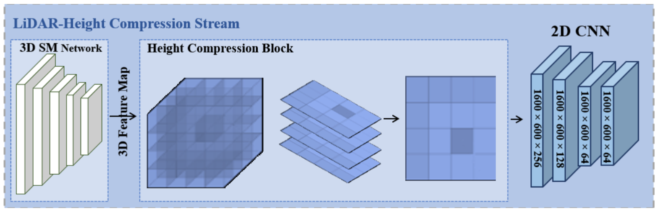

Fig. 3. LiDAR-Height Compression stream processes LiDAR points with 3D sparse CNN to extract 3D geometric feature maps which are then aggregated by a height compression block along the z-axis to get 2D feature maps for processing by a 2D CNN to reduce computational complexity.

3D Sparse Convolution Blocks: The raw LiDAR data is processed through a sequence of 3D Sparse Convolution blocks to effectively handle its high dimensionality and sparsity. The LiDAR frame is first voxelized and then passed through 3D sparse convolutional blocks with channel sizes of 16, 32, 32, 64, 64, and strides 1, 2, 2, 2, 1, respectively. Each block contains three convolution layers, with the first layer downsampling to capture essential geometric details. The final output is a 3D feature tensor $F_{3D} \in \mathbb{R}^{H \times W \times D \times C}$, where H, W, D, and C represent the height, width, depth, and the number of feature channels, respectively.

Height Compression Block: To reduce the computational complexity while retaining vital spatial information, the next stage involves compressing the height dimension of the 3D feature tensor F_{3D} (see Fig. 3). This block focuses on capturing the spatial relationships and geometry of objects by processing height information. Height compression aggregates features along the z-axis, effectively projecting the 3D feature map onto a 2D plane, forming the BEV feature map F_{BEV} [35]. This process is defined as

$$F_{BEV}(i, j, c) = \max_{k \in [1, H]} F_{3D}(i, j, k, c),$$

where $F_{BEV} \in \mathbb{R}^{W \times D \times C}$ is the compressed feature map that retains the most prominent features (maximum values) along the height dimension. The BEV feature map F_{BEV} is then refined through a series of 2D CNN layers, producing the processed features $\mathbf{F}_{\text{BEV,2D}}$ as

$$\mathbf{F}_{\text{BEV,2D}}^{(l+1)} = \sigma \left(\sum_{k=1}^{K} \mathbf{W}^{(l,k)} * \mathbf{F}_{\text{BEV}}^{(l)} + \mathbf{b}^{(l)} \right), \quad l = 1, 2, \ldots, L.$$

Here, $\mathbf{F}_{\text{BEV,2D}}^{(l+1)} \in \mathbb{R}^{H' \times W' \times C'}$ is the refined 2D BEV feature map after the l-th convolutional layer. $\mathbf{W}^{(l,k)} \in \mathbb{R}^{K \times K \times C \times C'}$ is the learnable kernel for the l-th layer, where K represents the kernel size and k indexes the filters. $\mathbf{b}^{(l)} \in \mathbb{R}^{C'}$ is the bias term. $\sigma(\cdot)$ represents the non-linear activation function ReLU. $*$ denotes the 2D convolution operation.

C. LiDAR-PillarNet Stream

The LiDAR-PillarNet Stream (see Fig. 4) converts 3D point clouds into 2D representations using pillar-based voxelization followed by feature extraction that extracts robust geometric features for high detection accuracy, while reducing computational complexity. The pillar-based voxelized data is passed through a Multi-Layer Perceptron (MLP) followed by 2D sparse convolutions to extract features. This

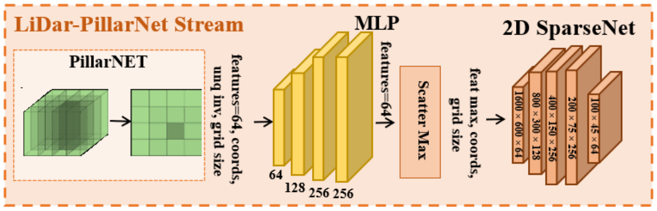

Fig. 4. The LiDAR-PillarNet architecture voxelizes raw LiDAR data into 2D pillar features, refines them with a Multi-Layer Perceptron (MLP) to extract Max Pillar Features, and processes these features through a 2D SparseNet for further refinement.

stream initially converts raw LiDAR point cloud data into vertical columns or "pillars". Each pillar aggregates LiDAR points based on their x and y coordinates, effectively discretizing the continuous 3D space into a grid representation

$$P(x, y) = \left\{ (x_i, y_i, z_i, r_i) \mid \left\lfloor \frac{x_i}{\Delta x} \right\rfloor = x, \left\lfloor \frac{y_i}{\Delta y} \right\rfloor = y \right\},$$

where, (x_i, y_i, z_i, r_i) represent the coordinates and reflectance of a LiDAR point, and Δx and Δy are the pillar sizes in the x and y dimensions. This step generates a sparse pseudo-image, where each cell in the 2D grid corresponds to a pillar containing point cloud information [36]. After discretization, the stream encodes each point within its pillar into a fixed-size feature vector relative to the pillar's center, effectively capturing both local and global spatial context.

The encoded features are refined through an MLP to capture the most relevant aspects for object detection. A Scatter Max layer then selects key features and aggregates them into a 2D grid, preserving the spatial arrangement of the pillars. This grid is processed by a 2D SparseNet, designed to handle data sparsity while extracting high-level geometric features, producing the output $\mathbf{F}_{\mathcal{S},\text{2D}}$.

$$\mathbf{F}_{\mathcal{S},\text{2D}}^{(l+1)} = \sigma \left(\sum_{k=1}^{K} \mathbf{W}^{(l,k)} \circledast \mathbf{F}_{\mathbf{max}}^{(l)} + \mathbf{b}^{(l)} \right), \quad l = 1, 2, \ldots, L.$$

where $\mathbf{F}_{\mathcal{S},\text{2D}}^{(l+1)} \in \mathbb{R}^{H \times W \times C'}$ is the output feature of PillarNet map after the l-th 2D sparse convolutional layer, $\mathbf{W}^{(l,k)} \in \mathbb{R}^{K \times K \times C \times C'}$ is the learnable kernel matrix for the k-th filter in the l-th layer, where K is the kernel size, and $\mathbf{b}^{(l)} \in \mathbb{R}^{C'}$ is the bias term for the l-th layer. The function $\sigma(\cdot)$ is the ReLU non-linear activation function and $\circledast$ denotes the sparse convolution operation. This formulation refines the feature map $\mathbf{F}_{\mathcal{S},\text{2D}}^{(l)}$, capturing the essential spatial and geometric properties of the scene.

D. Feature Fusion and Object Detection

At the final stage, features from the LiDAR-PillarNet, LiDAR-Height Compression, and 3D MM streams are integrated for robust object detection. Features from the LiDAR streams are concatenated and processed through 2D CNN layers, creating a unified feature set. These are then combined with the 3D MM stream features to generate enhanced feature maps. The final feature map, denoted as $F_{\text{H}} \in \mathbb{R}^{H' \times W' \times C'}$, is derived by fusing the 2D and 3D feature maps. Fusion is performed as follows.

$$\mathbf{F}_{\text{H}} = \sigma \left(\sum_{l=1}^{L} \mathbf{W}^{(l,k)} \circledast (\mathbf{F}_{\text{BEV,2D}} \oplus \mathbf{F}_{\mathcal{S},\text{2D}}) \right) \oplus \mathbf{F}_{\text{MM}}.$$

TABLE I

Server-generated object detection results (average precision in %) for 2D, 3D, BEV, and car orientation on the KITTI test set for Easy, Moderate, and Hard categories.

Benchmark	Easy	Moderate	Hard	mean AP
Car (Detection)	97.91	97.21	94.04	96.39
Car (Orientation)	97.88	97.03	93.74	96.22
Car (3D Detection)	91.03	84.36	80.78	85.39
Car (Bird's Eye View)	94.62	91.13	88.28	91.34

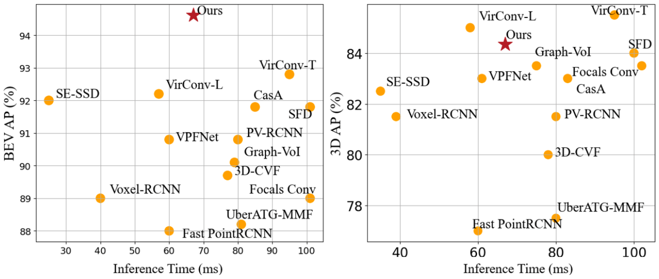

Fig. 5. Comparison of inference time (ms) versus object detection accuracy (AP) on the KITTI dataset. The left plot shows AP for BEV and the right shows for 3D detection. MuStD, marked as red star, achieves superior accuracy in both tasks while maintaining competitive inference speed.

The final feature map is then processed by the Region Proposal Network (RPN) to generate candidate Regions of Interest (RoI) as bounding boxes. These proposals are then converted into fixed-size feature maps via RoI pooling. The detection head, consisting of fully connected layers, processes these RoI features to output class scores and refined bounding box coordinates as $(C_{\text{obj}}, B_{\text{refined}}) = \text{Det\_Head}(F_{\text{RoI}})$. Finally, non-maximum suppression (NMS) is applied to remove redundant boxes for precise object detection.

IV. Experiments

We evaluated our MuStD network on the popular KITTI 3D object detection dataset [51], which is well-suited to autonomous driving research since it contains cars. The dataset comprises 7,481 training and 7,518 test images, along with their corresponding LiDAR point clouds. Our experiments follow the standard KITTI protocol, using the specified sequences for training, validation, and testing. The model was trained for 40 epochs with a batch size of 2 and a learning rate of 1e-4, using a single RTX4090 GPU. Accuracy metrics were obtained by submitting the predictions on the test set to the official KITTI online evaluation server, ensuring standardized and transparent performance assessment. The KITTI server provides precision-recall curves, average precision (AP), and average orientation similarity (AOS) for evaluation. The server reports results across various benchmarks, including 2D and 3D detection, orientation estimation, and bird's-eye view (BEV) detection. Additionally, we conducted an ablation study on the validation set to examine the contributions of various network components. Our method was also evaluated for multi-class object detection. The proposed approach achieved outstanding results across multiple detection tasks, surpassing most existing methods in terms of accuracy and inference time. Please refer to the KITTI Benchmark Results.

Table I summarizes the overall car detection results of our method for 2D, 3D, BEV, and orientation on the KITTI test

TABLE II

Results on the KITTI test set generated by the online server. Average precision (AP) and average orientation similarity (AOS) in % are reported for 2D car detection and orientation, respectively. Best results are bolded and 2nd best underlined.

Approach	Reference	Car 2D AP				Car Orientation AOS				Time (ms)
		Easy	Moderate	Hard	mean AP	Easy	Moderate	Hard	mean AP	
MVRA-FRCNN [37]	ICCV 2019	95.87	94.98	82.52	91.12	95.66	94.46	81.74	90.62	180
CLOCs [38]	IROS 2020	96.77	96.07	91.11	94.65	96.77	95.93	90.93	94.54	100
SPA-Net [39]	PRICAI 2021	96.54	95.46	90.47	94.16	96.31	95.03	89.99	93.78	60
VoTr-TSD [40]	ICCV 2021	95.97	94.94	92.44	94.45	95.95	94.81	92.24	94.33	70
Pyramid R-CNN [41]	ICCV 2021	95.88	95.13	92.62	94.54	95.87	95.03	92.46	94.45	70
3D D-Fusion [42]	ArXiv 2022	96.54	95.82	93.11	95.16	96.53	95.76	93.01	95.10	100
CasA [43]	TGRS 2022	96.52	95.62	92.86	94.99	96.51	95.53	92.71	94.92	100
VPFNet [7]	TMM 2022	96.64	96.15	91.14	94.64	96.63	96.04	90.99	94.55	<u>60</u>
Graph-VoI [44]	ECCV 2022	96.81	<u>96.38</u>	91.20	94.80	96.81	<u>96.29</u>	91.06	94.72	70
SFD [4]	CVPR 2022	**98.97**	96.17	91.13	95.42	**98.95**	96.05	90.96	95.32	100
DVF (Voxel-RNN) [45]	WACV 2023	96.60	95.77	90.89	94.42	96.59	95.63	90.71	94.31	100
OcTr [46]	CVPR 2023	96.48	95.84	90.99	94.44	96.44	95.69	90.78	94.30	<u>60</u>
Focals Conv [47]	CVPR 2023	96.30	95.28	92.69	94.76	96.29	95.23	92.60	94.71	100
TED [48]	AAAI 2023	96.64	96.03	93.35	95.34	96.63	95.96	<u>93.24</u>	95.28	100
MLF-DET [49]	ICANN 2023	96.89	96.17	88.90	93.99	96.87	96.09	88.78	93.91	90
PVT-SSD [50]	CVPR 2023	96.75	95.90	90.69	94.45	96.74	95.83	90.58	94.38	**50**
VirConv-T [29]	CVPR 2023	<u>98.93</u>	<u>96.38</u>	<u>93.56</u>	<u>96.29</u>	<u>98.64</u>	96.01	93.12	<u>95.92</u>	90
Ours	-	97.91	**97.21**	**94.04**	**96.39**	97.88	**97.03**	**93.74**	**96.22**	**50**

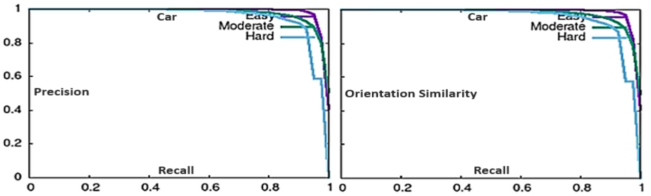

Fig. 6. Car detection results of our method on KITTI test set for easy, moderate and difficult categories generated by the online KITTI server. Left: Recall vs precision curve for 2D car detection. Right: Recall vs orientation

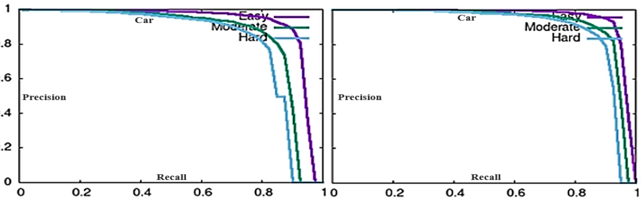

Fig. 7. 3D and BEV Car detection results of our method on the KITTI test set for Easy, Moderate and Difficult categories generated by the online KITTI server. Left: Recall and precision curve for 3D detection. Right: Recall and precision curve for BEV detection.

set, as generated by the server. Fig. 5 presents a comparison of inference time and detection accuracy for object detection methods on the KITTI dataset in 3D detection and BEV. Our method demonstrates superior performance, achieving high accuracy in both tasks while maintaining competitive inference speed. Detailed results from the experiments are provided in the following section.

A. 2D car detection and orientation results on KITTI test set

We evaluate the performance of MuStD on the KITTI test set for 2D car detection and orientation estimation. The KITTI server reports results for three difficulty levels: Easy, Moderate, and Hard, using AP for 2D detection and AOS for orientation similarity. Table II compares our method to existing state of the art. We can see that our method outperforms all others in the Moderate and Hard categories for 2D detection as well car orientation. For the Easy category, our method achieves results very close to the top performer. Hence, the overall mean AP of our method is still the best for 2D car detection and car orientation. These results underscore the effectiveness of our method, particularly in challenging (Hard and Moderate) scenarios with occluded or distant objects. Fig 6 visually illustrates the precision-recall and orientation similarity curves for the three difficulty levels. Our method is one of the fastest, with an inference time of only 50 milliseconds. Notice that our close competitors in accuracy, e.g. VirConv-T and SDF, are much slower.

B. 3D and BEV detection results on KITTI Dataset

Table III compares our method for 3D and bird's eye view (BEV) detection on the KITTI test set (server generated results) with existing state of the art. The proposed MuStD network achieves the best performance in the Hard category for 3D and BEV car detection. In the Easy and Moderate cases, our method achieves comparable results. For 3D detection, MuStD obtained 80.78% AP in the Hard category and a mean AP of 85.3% across all categories. For BEV, MuStD achieved 88.28% AP in the Hard category and a mean AP of 91.34%. These results highlight the effectiveness of our method in integrating LiDAR and RGB data and refining features through advanced mapping and indexing techniques. Precision-recall curves in Fig. 7 further demonstrate high precision across varying recall levels, reinforcing the efficacy of our multimodal fusion strategy.

C. Multi-class results on KITTI dataset

We also evaluate MuStD network on the KITTI validation set for multi-class 3D object detection. We test three classes that are most important for autonomous driving namely, 'Car,' 'Pedestrian,' and 'Cyclist' across across the three difficulty levels. Table V compares our results to Voxel-RCNN and VirConvT on the AP metric. Our method consistently outperforms the competitors on all three categories achieving a mean AP of 91.29% for Car, 68.32% for Pedestrian, and 80.12% for Cyclist. These results demonstrate our method's effectiveness in capturing geometric structures and spatial-texture features equally well for small and large objects in complex urban environments. Consistent performance across classes highlights the robustness and generalizability of our method in real-world autonomous driving scenarios.

TABLE III

RESULTS ON THE KITTI OBJECT DETECTION TEST SET, GENERATED BY THE ONLINE KITTI SERVER. THE AVERAGE PRECISION (AP) IN % IS REPORTED FOR CAR 3D AND BIRD'S-EYE VIEW (BEV) DETECTION IN ALL CATEGORIES. BEST RESULTS ARE BOLDED AND 2ND BEST UNDERLINED.

Approach	Reference	Modality	Car 3D AP				Car BEV AP				Time
			Easy	Moderate	Hard	mean AP	Easy	Moderate	Hard	mean AP	(ms)
PV-RCNN [24]	CVPR 2020	LiDAR	90.25	81.43	76.82	82.83	94.98	90.65	86.14	90.59	80*
Voxel-RCNN [25]	AAAI 2021	LiDAR	90.90	81.62	77.06	83.19	94.85	88.83	86.13	89.94	40
CT3D [52]	ICCV 2021	LiDAR	87.83	81.77	77.16	82.25	92.36	88.83	84.07	88.42	70*
SE-SSD [53]	CVPR 2021	LiDAR	91.49	82.54	77.15	83.73	95.68	91.84	86.72	91.41	30
BtcDet [54]	AAAI 2022	LiDAR	90.64	82.86	78.09	83.86	92.81	89.34	84.55	88.90	90
CasA [43]	TGRS 2022	LiDAR	91.58	83.06	80.08	84.91	95.19	91.54	86.82	91.18	86
Graph-Po [8]	ECCV 2022	LiDAR	91.79	83.18	77.98	84.32	95.79	92.12	87.11	91.01	60
berATG-MMF [55]	CVPR 2019	LiDAR+RGB	88.40	77.43	70.22	78.02	93.67	88.21	81.99	87.29	80
3D-CVF [28]	ECCV 2020	LiDAR+RGB	89.20	80.05	73.11	80.79	93.52	89.56	82.45	88.51	75
Focals Conv [47]	CVPR 2022	LiDAR+RGB	90.55	82.28	77.59	83.47	92.67	89.00	86.33	89.33	100*
VPFNet [7]	TMM 2022	LiDAR+RGB	91.02	83.21	78.20	84.14	93.94	90.52	86.25	90.24	62
Graph-VoI [44]	ECCV 2022	LiDAR+RGB	**91.89**	83.27	77.78	84.31	95.69	90.10	86.85	90.88	76
SFD [4]	CVPR 2022	LiDAR+RGB	91.73	84.76	77.92	84.80	**95.64**	91.85	86.83	91.44	98
VirConv-L [29]	CVPR 2023	LiDAR+RGB	91.41	**85.05**	80.22	**85.56**	95.53	**91.95**	87.07	**91.52**	56
Ours	-	LiDAR+RGB	91.03	84.36	**80.78**	85.39	94.62	91.13	**88.28**	91.34	67

TABLE IV

ABLATION STUDY ON THE KITTI OBJECT DETECTION VALIDATION SET USING DIFFERENT FUSION/COMPONENT COMBINATIONS OF OUR PROPOSED METHOD. THE AVERAGE PRECISION (AP) IN % IS REPORTED FOR 3D CAR DETECTION AND 2D CAR DETECTION. LIDAR-HC = LIDAR HEIGHT COMPRESSION STREAM.

3D MM	LiDAR-HC	LiDAR-Pillar	LiDAR	RGB	Car 3D AP				Car 2D AP			
					Easy	Moderate	Hard	mean AP	Easy	Moderate	Hard	mean AP
✓	✓	✓	✓	✓	95.21	93.56	90.09	92.95	98.42	97.75	94.07	96.08
✓	✓		✓	✓	92.77	90.19	87.30	90.75	95.11	95.81	92.27	94.39
✓		✓	✓	✓	91.33	89.14	86.15	88.87	94.67	93.53	90.70	92.97
	✓		✓		84.19	80.01	77.30	80.50	87.40	85.95	80.72	84.69
✓			✓	✓	91.38	89.60	87.37	89.45	93.89	90.41	89.65	91.32
	✓	✓	✓		88.48	86.07	83.93	86.83	91.06	89.34	86.73	89.04

TABLE V

COMPARISON WITH STATE-OF-THE-ART ON THE KITTI VALIDATION SET FOR 3D DETECTION OF MULTIPLE CLASSES.

Class	Method	3D Detection (AP in %)			
		Easy	Moderate	Hard	mean
Car	Voxel-RCNN [25]	89.39	83.83	87.73	86.32
	VirConv-T [29]	**94.98**	89.96	88.13	91.02
	Ours	94.21	**91.56**	**90.09**	**91.29**
Pedestrian	Voxel-RCNN [25]	70.55	62.92	57.35	63.61
	VirConv-T [29]	73.32	66.93	60.38	66.88
	Ours	**75.45**	**68.11**	**63.40**	**68.32**
Cyclist	Voxel-RCNN [25]	89.86	71.13	66.67	75.89
	VirConv-T [29]	90.04	73.90	69.06	77.00
	Ours	**91.89**	**76.56**	**71.91**	**80.12**

D. Ablation study on KITTI validation set

To assess the impact of various components in the MuStD Network, we conducted an ablation study on the KITTI dataset, evaluating different configurations for 3D and 2D detection tasks. The study focused on key components: LiDAR-PillarNet, LiDAR Height Compression, and the 3D MM stream, along with LiDAR and RGB modalities. As shown in Table IV, MuStD achieved the best results with mean AP scores of 92.95% for 3D detection and 96.08% for 2D detection. Removing the LiDAR-PillarNet stream led to approximately a 2% drop in mean AP for both detection tasks, highlighting its role in capturing geometric details. Eliminating LiDAR Height Compression reduced AP to 88.87% and 92.97% for 3D and 2D detection, demonstrating its importance in preserving spatial characteristics. Excluding the 3D MM Network resulted in a significant performance decline, with AP dropping to 80.50% and 84.69%, underscoring the necessity of integrating RGB and LiDAR data. LiDAR-only configurations yielded notably lower scores, confirming the critical role of multimodal fusion. These results emphasize the importance of combining all components for optimal detection performance.

V. CONCLUSION

We introduced the Multistream Detection Network (MuStD), a novel approach that enhances 3D object detection by integrating LiDAR point cloud data with RGB image information. Utilizing UV mapping and polar coordinate indexing, our method improves the extraction of geometric and spatial-texture features through a unified 2D representation. The network, consisting of three parallel streams—the LiDAR-PillarNet Stream, LiDAR Height Compression Stream , and 3D MM Stream—demonstrated superior performance on the KITTI benchmark. Experimental results highlighted the high accuracy and robustness of MuStD across various detection tasks, validating its effectiveness in addressing challenges related to detection loss, computational efficiency, and multimodal feature fusion. This work marks a significant advancement in 3D object detection for autonomous driving, providing a reliable solution for real-time navigation in complex urban environments.

REFERENCES

[1] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1907–1915.

[2] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[3] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *CVPR*, 2021, pp. 7077–7087.

[4] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, and D. Cai, "Sparse fuse dense: Towards high quality 3d detection with depth completion," in *CVPR*, 2022.

[5] B. Wu, S. He, Z. Yan, W. Zeng, and L. Zhang, "Sparse fuse dense: Towards high quality 3d detection with depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5412–5421.

[6] H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, "Virtual sparse convolution for multimodal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21 653–21 662.

[7] W. Wang, J. Shen, Z. Wu, T. He, J. Zhang, Z. Jiang, and G. H. Lee, "Vpfnet: Virtual point based feature fusion network for 3d object detection," in *IEEE Transactions on Multimedia*, vol. 24, 2022, pp. 3487–3497.

[8] Y. Yang, X. Sun, Z. Zhang, K. Jia, and W. Zeng, "Graph-voi: Graph neural network based voxel information aggregation for 3d object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 678–695.

[9] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, "Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion," *IEEE Transactions on Multimedia*, vol. 25, 2023.

[10] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. 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 (CVPR)*, 2022, pp. 1090–1099.

[11] T. Guan, J. Wang, S. Lan, R. Chandra, Z. Wu, L. Davis, and D. Manocha, "M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers," *arXiv preprint arXiv:2104.11896*, 2021.

[12] H. Meng, C. Li, G. Chen, L. Chen, and A. Knoll, "Efficient 3d object detection based on pseudo-lidar representation," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1953–1964, 2024.

[13] Y. Tian, X. Zhang, X. Wang, J. Xu, J. Wang, R. Ai, W. Gu, and W. Ding, "Acf-net: Asymmetric cascade fusion for 3d detection with lidar point clouds and images," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 2, 2023.

[14] D. Wang, X. Cui, X. Chen, Z. Zou, T. Shi, S. Salcudean, Z. J. Wang, and R. Ward, "Multi-view 3d reconstruction with transformer," *arXiv preprint arXiv:2103.12957*, 2021.

[15] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020, pp. 11 621–11 631.

[16] L. Ma, Y. Li, J. Li, W. Tan, Y. Yu, and M. A. Chapman, "Multi-scale point-wise convolutional neural networks for 3d object segmentation from lidar point clouds in large-scale environments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 821–836, 2021.

[17] Y. Zhang, L. Wang, C. Fu, Y. Dai, and J. M. Dolan, "Encode: a deep point cloud odometry network," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 14 375–14 381.

[18] A. D. Pon, J. Ku, C. Li, and S. L. Waslander, "Object-centric stereo matching for 3d object detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 8383–8389.

[19] J. Fang, D. Zhou, J. Zhao, C. Wu, C. Tang, C.-Z. Xu, and L. Zhang, "Lidar-cs dataset: Lidar point cloud dataset with cross-sensors for 3d object detection," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14 822–14 829.

[20] X. Li, F. Wang, N. Wang, and C. Ma, "Frame fusion with vehicle motion prediction for 3d object detection," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 4252–4258.

[21] C. R. Qi, Y. Zhou, M. Najibi, P. Sun, K. Vo, B. Deng, and D. Anguelov, "Offboard 3d object detection from point cloud sequences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6134–6144.

[22] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 19, pp. 12 697–12 705.

[23] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," in *Sensors*, vol. 18, no. 10, 2018, p. 3337.

[24] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 529–10 538.

[25] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1201–1209.

[26] T. Ji and L. Xie, "Vision-aided localization and navigation for autonomous vehicles," in *2022 IEEE 17th International Conference on Control & Automation (ICCA)*. IEEE, 2022, pp. 599–604.

[27] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–8.

[28] J. Yoo, J. Kim, S. Lee, M. Roh, K. M. Choi, and T.-K. Choi, "3d-cvf: Generating joint camera and lidar features for robust 3d object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 282–300.

[29] H. Wang, J. Li, K. Zhang, and Y.-X. Wang, "Virtual convolution for lidar-based 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13 511–13 520.

[30] L. Gao, H. Xiang, X. Xia, and J. Ma, "Multisensor fusion for vehicle-to-vehicle cooperative localization with object detection and point cloud matching," *IEEE Sensors Journal*, vol. 24, no. 7, pp. 10 865–10 877, 2024.

[31] S. Y. Alaba and J. E. Ball, "A survey on deep-learning-based lidar 3d object detection for autonomous driving," *Sensors*, vol. 22, no. 24, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/24/9577

[32] H. Wang, B. Li, X. Song, H. Li, and M. Liu, "Nrconv: Noise-resistant convolution for point cloud processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 12 011–12 020.

[33] Y. Ye, Y. Wang, X. Yang, S. Wang, Z. Huang, B. Feng, and K. Jia, "2dpass: 2d priors assisted semantic segmentation of 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1609–1618.

[34] H. Zhou, X. Wang, L. Chen, X. Luo, H. Zhang, and K. Wu, "Ted: Transformation-equivariant 3d detector," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023, pp. 2434–2442.

[35] A. Jhaldiyal and N. Chaudhary, "Semantic segmentation of 3d lidar data using deep learning: a review of projection-based methods," *Applied Intelligence*, vol. 53, pp. 6844–6855, 2023.

[36] Y. Li, Y. Zhang, and R. Lai, "Tinypillarnet: Tiny pillar-based network for 3d point cloud object detection at edge," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, 2023.

[37] H. M. Choi, H. Kang, and Y. Hyun, "Multi-view reprojection architecture for orientation estimation," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[38] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020.

[39] Y. Ye, "Spanet: Spatial and part-aware aggregation network for 3d object detection," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2021, pp. 308–320.

[40] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," in *ICCV*, 2021.

[41] J. Mao, M. Niu, H. Bai, X. Liang, H. Xu, and C. Xu, "Pyramid r-cnn: Towards better performance and adaptability for 3d object detection," in *ICCV*, 2021.

[42] Y. Kim, K. Park, M. Kim, D. Kum, and J. W. Choi, "3d dual-fusion: Dual-domain dual-query camera-lidar fusion for 3d object detection," *arXiv preprint arXiv:2211.13529*, 2022.

[43] H. Wu, J. Deng, C. Wen, X. Li, and C. Wang, "Casa: A cascade attention network for 3d object detection from lidar point clouds," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[44] H. Yang, Z. Liu, X. Wu, W. Wang, W. Qian, X. He, and D. Cai, “Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph,” in *ECCV*, 2022.

[45] A. Mahmoud, J. S. Hu, and S. L. Waslander, “Dense voxel fusion for 3d object detection,” *WACV*, 2023.

[46] C. Zhou, Y. Zhang, J. Chen, and D. Huang, “Octr: Octree-based transformer for 3d object detection,” in *CVPR*, 2023.

[47] Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, “Focal sparse convolutional networks for 3d object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[48] H. Wu, C. Wen, W. Li, R. Yang, and C. Wang, “Transformation-equivariant 3d object detection for autonomous driving,” in *AAAI*, 2023.

[49] Z. Lin, Y. Shen, S. Zhou, S. Chen, and N. Zheng, “Mlf-det: Multi-level fusion for cross- modal 3d object detection,” in *International Conference on Artificial Neural Networks*. Springer, 2023, pp. 136–149.

[50] H. Yang, W. Wang, M. Chen, B. Lin, T. He, H. Chen, X. He, and W. Ouyang, “Pvt-ssd: Single-stage 3d object detector with point-voxel transformer,” in *CVPR*, 2023.

[51] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[52] H. Sheng, S. Cai, Y. Liu, B. Deng, J. Huang, X.-S. Hua, and M.-J. Zhao, “Improving 3d object detection with channel-wise transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2743–2752.

[53] W. Zheng, W. Tang, L. Jiang, and C.-W. Fu, “Se-ssd: Self-ensembling single-stage object detector from point cloud,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 494–14 503.

[54] Q. Xu, Y. Zhong, and U. Neumann, “Behind the curtain: Learning occluded shapes for 3d object detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2893–2901.

[55] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, “Multi-task multi-sensor fusion for 3d object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7345–7353.