SimpleFusion: 3D Object Detection by Fusing RGB Images and Point Clouds

Yongchang Zhang
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
Institute of Automation
Chinese Academy of Sciences
Beijing, China
zhangyongchang2020@ia.ac.cn

Yue Guo
Institute of Automation
Chinese Academy of Sciences
Beijing, China
guoyue2013@ia.ac.cn

Hanbing Niu
University of Electronic Science and Technology of China
Chengdu, China
niuhanbing2021@std.uestc.edu.cn

Bo Zhang
Intelligent Mining Research Academy, Chinese Institute of Coal Science
China Coal Technology & Engineering Group Co., Ltd.,
Beijing, China
zbo1026@126.com

Yun Cao
Intelligent Mining Research Academy, Chinese Institute of Coal Science
China Coal Technology & Engineering Group Co., Ltd.,
Beijing, China
caoyun@mail.ccri.ccteg.cn

Wenhao He*
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
Institute of Automation
Chinese Academy of Sciences
Beijing, China
wenhao.he@ia.ac.cn

***Abstract*—Achieving robust 3D object detection by fusing images and point clouds remains challenging. In this paper, we propose a novel 3D object detector (SimpleFusion) that enables simple and efficient multi-sensor fusion. Our main motivation is to boost feature extraction from a single modality and fuse them into a unified space. Specifically, we build a new visual 3D object detector in the camera stream that leverages point cloud supervision for more accurate depth prediction; in the lidar stream, we introduce a robust 3D object detector that utilizes multi-view and multi-scale features to overcome the sparsity of point clouds. Finally, we propose a dynamic fusion module to focus on more confident features and achieve accurate 3D object detection based on dynamic weights. Our method has been evaluated on the nuScenes dataset, and the experimental results indicate that it outperforms other state-of-the-art methods by a significant margin.**

Keywords-3D object detection; multi-sensor fusion; BEV detetion; multi-scale fusion

I. INTRODUCTION

With the application of deep learning technology in computer vision and robotics, object detection methods based on image and point cloud have been developed rapidly. In recent years, 3D point clouds acquisition equipment represented by depth cameras and lidar has become increasingly low cost, making the focus of visual object detection tasks gradually shift from 2D to 3D. The aim of 3D object detection is to detect objects in 3D space. Relying on point cloud data with 3D spatial information, 3D object detection has extra depth information and focuses on identifying and positioning targets in the real world. At the same time, the development of hardware technology has also enabled more applications of 3D object detection on scenarios such as autonomous driving and industrial robot.

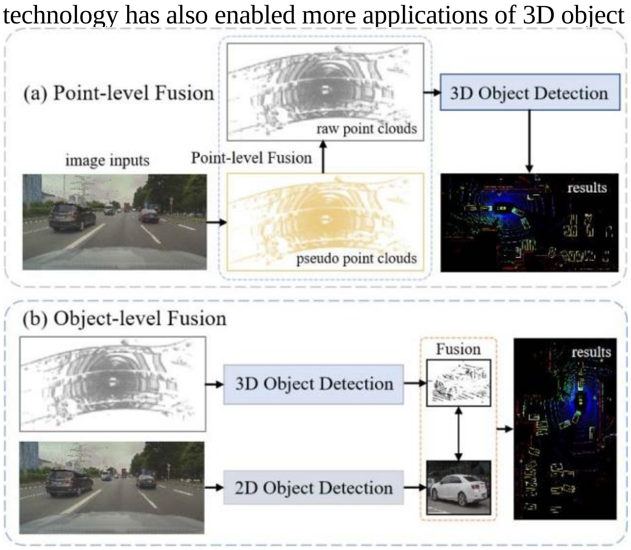

Figure 1. Existing data fusion methods on 3D object detection mainly consist of two branches. (a) point-level fusion: they fuse both features of images and point clouds at the early stage. (b) object-level fusion: they independently handle the detection results of images and point clouds.

Among all the sensors in a perception system, lidars and cameras remain the two most critical sensors, precisely capturing point clouds and images about the world. Point cloud data, obtained from lidars, has rich spatial information and can provide detailed surface structure information in the real world. But its sparsity and lack of rich texture information in RGB images make small object detection difficult. Among them, the point cloud data about 3D objects cannot meet the needs of industrial applications well, and 2D RGB image data lacks spatial depth information, resulting in the inability to locate objects accurately in 3D space. Therefore, how to better use the rich texture information in the visual image and the spatial position information in the point cloud data has become the key to improving the 3D object detection algorithms.

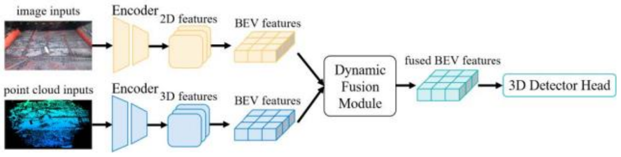

Figure 2. The main architecture of SimpleFusion. SimpleFusion contains three parts: camera stream, lidar stream, and dynamic fusion module. The camera stream predicts the depth probability distribution at the pixel level and converts the image feature space to the BEV space based on camera parameters and depth. The lidar stream uses the features in both perspective view and BEV to overcome the sparsity of point clouds and use deformable convolution to obtain object-level features for multi-scale and multi-view fusion. The multi-sensor fusion uses a simple fusion module to focus on more confident features based on dynamic weights.

Existing data fusion methods on 3D object detection mainly consist of two branches. On the one hand, they fuse both features of images and point clouds at the early stage. For example, given an input image, a depth estimation network outputs the depths of objects. And a pseudo point cloud is generated by combining the RGB image and converting it into a 3D space. Next, such a pseudo point cloud and the point cloud data collected by the radar are fused. On the other hand, post-fusion methods separately handle the detection results of images and point clouds. And the final object detection results are filtered according to their confidence. One of the biggest challenges in both branches is to find a unified representation between the lidar and the camera suitable for multi-modal fusion.

To solve the above problem, we introduce a novel multimodal data fusion framework named SimpleFusion that achieves robust and efficient 3D object detection in a shared bird's-eye view (BEV). Specifically, Our framework comprises of two separate streams, each of which encodes the raw images and point clouds into features, all within the same Bird's Eye View (BEV) space. Within the image stream, the depth estimation network is supervised by the point cloud and facilitates the conversion of 2D image features into BEV features. Meanwhile, in the lidar stream, we utilize multi-view and multi-scale features to generate more densely packed BEV features. Then we design a simple module that dynamically fuses these BEV-level features that achieve better performances. Experiments conducted on the nuScenes dataset [1] demonstrate that our proposed method, SimpleFusion, outperforms other state-of-the-art approaches significantly.

The main contributions of our work are as follows:

- We introduce a novel multi-modal data fusion framework, named Simple Fusion, which enables efficient and reliable 3D object detection.
- We build a new visual 3D object detector that leverages point cloud supervision for more accurate depth predictions.
- We present a robust 3D object detector on point clouds, which utilizes multi-scale and multi-view features to solve the problem of point cloud feature sparsity.
- We design a simple adaptive-weight fusion module that makes our framework achieve state-of-the-art performances.

II. Related Work

Our work takes insights from 3D object detection methods based on camera-only, lidar-only, and camera-lidar fusion.

3D object detection on images. In the autonomous driving field, researchers have paid many attentions on camera-only 3D object detection. DOP [2], 3DOP-Stereo [3], and Mono3D [4] are pioneer works, which mainly obtain 2D object detection results and use specific area prior information such as shape, height, and position distribution to generate final 3D object boxes. DETR3D [5] and PETR [6] use learnable object queries in the 3D space to detect the 3D objects. Instead of detecting objects in the perspective view, BEVDet [7] and M2BEV [8] are extensions of the Lift-Splat-Shoot (LSS) [9] technique that utilize multi-perspective images to extract implicit depth information, and subsequently transform the camera feature maps into Bird's Eye View (BEV) space to perform 3D object detection within the BEV feature space. In our camera stream, we build a new visual 3D object detector that leverages point cloud supervision for depth map prediction, so we obtain more accurate depth maps and generate stable and reliable BEV features.

3D object detection on point clouds. Current 3D object detection methods that are based on point clouds can be broadly categorized into two groups based on their feature modality: voxels and points. Voxel-based methods [10]mainly divide the point clouds into a regular grid and use dense convolution for voxel feature extraction. To increase the efficiency of feature extraction, SECOND adopts sparse convolution to output features from non-empty grids. Point-based detectors are born to be fully sparse. To address the disorder, sparsity, and rotation variance in point clouds, PointNet and PointNet+ utilize max pooling operations to obtain the most critical feature vector from the point clouds. VoteNet [10] groups point sets into voxels, and it uses a 3D CNN that learns voxel features to generate 3D boxes. CenterPoint uses the geometric center point to regress the parameters of the target box. Our lidar-only detector utilizes multi-scale and multi-view features to solve the problem of feature sparsity of point clouds.

Lidar-camera fusion. Current multi-sensor fusion arouses increasing interest in 3D object detection. And we classify existing works into point-level, feature-level, and proposal-level approaches. Point-level fusion methods, including PointPainting, PointAugmenting, MVP, FusionPainting, AutoAlign, and FocalSparseCNN, usually add image features onto raw point clouds and perform lidar-based object detection. Featurelevel methods fuse the multi-sensor features and output more dominant features. DeepFusion is a method that projects point cloud features onto images and subsequently employs cross-attention modules to effectively fuse the features from both modalities. BevFusion utilizes separate detection branches for cameras and radars to enhance detection stability. Unlike proposal-level fusion methods, BevFusion does not require the fetching of proposals in advance, as it is object-centric. FPointNet, F-ConvNet, and CenterFusionneed image proposals as priors. FUTR3D and TransFusiondefine object queries in 3D space and use crosstransform to fuse image features onto these proposals.

III. Method

In Figure 2, we introduce an overview of our framework, SimpleFusion, for 3D object detection. Given different in put data from sensors, modality-specific encoders extract features independently. Then these features are transformed into a unified BEV space that keeps geometric and semantic information. Finally, a dynamic feature fusion module incorporates the above representations, and a detection head outputs the results.

A. Camera Stream

In the camera stream, we design a new image encoder that can extract the multiple image features and convert them to the BEV representation, as shown in Figure 3. Specifically, the camera stream includes two parts: an image view encoder and a view projector module. An image view encoder aims to extract rich semantic features from input images. It is composed of two main components: a backbone for fundamental 2D image feature encoding, and a neck module for multi-scale feature representation. In line with LSS [9], our approach employs a ResNet as the backbone network and incorporates a Feature Pyramid Network (FPN) to utilize features at multiple scales. Additionally, we employ a view projector module to map 2D features into the 3D ego-car space. This view projector module is similar to that of BevDet [7], and utilizes the lift-splat technique [9] to project 2D features onto the BEV space. Besides, we use a depth prediction network supervised by point clouds to obtain more accurate depth values.

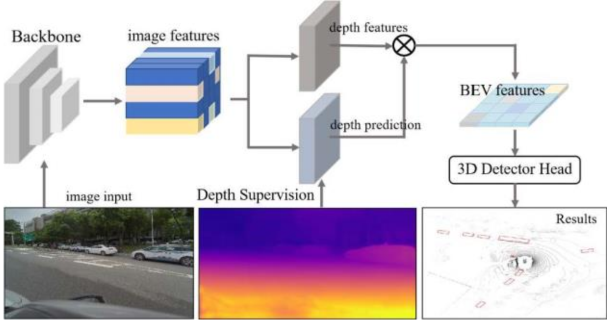

Figure 3. The main structure of the camera stream: we predict the depth probability distribution at the pixel level and convert the image feature space to the BEV space based on camera parameters and depth. Note that our depth labels come from point cloud data.

B. Lidar Stream

In the lidar stream, we present a robust 3D point cloud object detector that utilizes multi-scale and multi-view features to solve feature sparsity of point clouds, as illustrated in Figure 4. Similar to H2-3D, we extract perspective-view features based on polar coordinates and paint these features to raw point clouds. After that, our detector voxelizes the painted point clouds to reduce the Z dimension and use networks to efficiently output features in the BEV space. In practice, we utilize PointPillar as our point cloud encoder. We adopt the deformable convolutionto generate object-level features. Then we concatenate them with the BEV feature to obtain the multi-scale features.

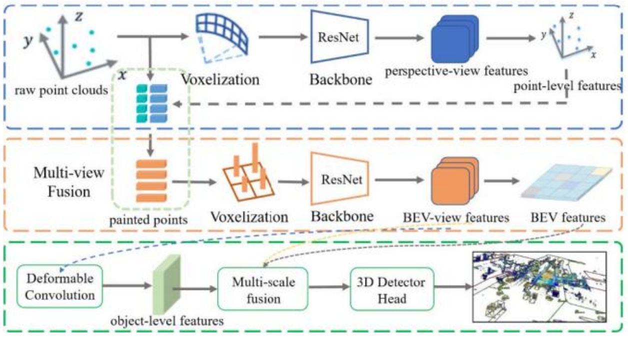

Figure 4. The main structure of our lidar stream: we use the perspectiveview and the BEV features to overcome the sparsity of point clouds and use deformable convolution to obtain object-level features for multi-scale and multi-view fusion.

C. Dynamic Fusion Module

The camera and lidar streams respectively provide the image feature $\left(F_{img} \in R^{X \times Y \times C_{img}}\right)$ and point cloud feature $\left(F_{pts} \in R^{X \times Y \times C_{pts}}\right)$ in BEV space. To efficiently integrate features from multiple sensors, we introduce a dynamic fusion module. As depicted in Figure 5, when given two feature maps of the same dimensions, the conventional approach would be to concatenate the features. However, our proposed fusion method employs a straightforward channel attention module to emphasize critical features. This process can be mathematically expressed as follows:

$$F_i^{fusion} = MLP\left(CAT\left(w_i^{img} F_i^{img}, w_i^{pts} F_i^{pts}\right)\right) \tag{1}$$

Adaptive weights $\left(w_i^{img}, w_i^{pts}\right)$ are generated according to the respective input features:

$$\left(w_i^{img}, w_i^{pts}\right) = \sigma\left(MLP\left(F_i^{img}, F_i^{pts}\right)\right) \tag{2}$$

where MLP is a multi-layer perception, CAT means the concatenation operation, and σ denotes the Sigmoid function. In practice, we process all pairs of grid features in a batch in parallel, using 1 × 1 convolution instead of MLP.

D. Detection Head

Our framework outputs fused features in BEV space, and we can use the common 3D detection heads from previous works. At the same time, this also demonstrates the generalization ability of SimpleFusion. To reduce the time consumption of our model, we adopt CenterHead as our final detection head.

IV. Experiments

In this section, we describe our experimental setup and evaluate the performance of our detector against other state-of-the-art methods using the nuScenes dataset [1].

The comparison results demonstrate the effectiveness and robustness of our proposed method.

A. Experimental Settings

Dataset. The nuScenes dataset comprises 1000 scenes, each of which is captured by a 32-beam lidar that generates point clouds. Additionally, there are six cameras with surrounding views that capture images of the scenes. The dataset is divided into three subsets: 700 scenes for training, 150 scenes for validation, and 150 scenes for testing.

TABLE I. EXTENSIVE COMPARISONS ON **NUSCENES** DATASET WITH **PILLAR SIZE (0.2, 0.2).** NOTE THAT **C** DENOTES CAMERA-BASED METHODS.

Method	Modality	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
MonoDIS	C	30.4%	73.8%	26.3%	54.6%	155.3%	13.4%	38.4%
CenterNet	C	30.6%	71.6%	26.4%	60.9%	142.6%	65.8%	32.8%
FCOS3D	C	29.5%	80.6%	26.8%	51.1%	131.5%	17.0%	37.2%
DETR3D	C	30.3%	86.0%	27.8%	43.7%	96.7%	23.5%	37.4%
BevDet-R50	C	28.6%	72.4%	27.8%	59.0%	87.3%	24.7%	37.2%
BevDet-Tiny	C	31.2%	69.1%	27.2%	52.3%	90.9%	24.7%	39.2%
PETR-R50	C	31.3%	76.8%	27.8%	56.4%	92.3%	22.5%	38.1%
Ours	C (0.2, 0.2)	**32.7%**	**65.6%**	**27.3%**	59.2%	**75.0%**	27.0%	**38.7%**

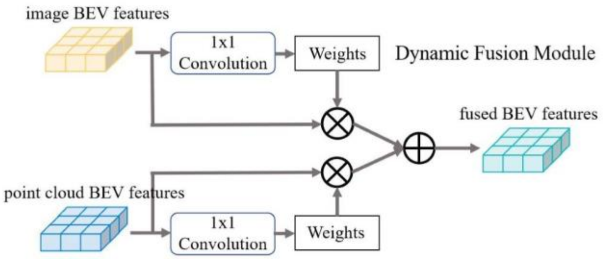

Figure 5. The dynamic fusion module. We design a simple fusion module to to focus on more confident features based on dynamic weights.

Implementation details. To fairly compare different methods, we define the detection region within 51.2 meters on the ground plane and select AdamW as the optimizer, where the learning rate is 2e-4. Our network runs on the open-source MMDetection3D. The chosen pillar size is (0.2, 0.2, 10), and our training contains two steps.

We first train the lidar stream and camera stream respectively. Then we train SimpleFusion in both streams for another five epochs, in which we freeze parameters in backbones from the two trained streams. As for testing, we follow settings of the lidar-only detector.

Evaluation metrics. The nuScenes dataset provides official predefined evaluation metrics, including mean Average Precision (mAP), Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity (AVE), Average Attribute Error (AAE), and NuScenes Detection Score (NDS). We employ these metrics to evaluate the performance of our detector across ten detection classes. mAP measures precision and recall, while NDS is a composite metric that evaluates the detection capacity comprehensively by incorporating other indicators. Other metrics such as scale, orientation, translation, velocity, and attribute evaluate the accuracy of positive results.

TABLE II. EXTENSIVE COMPARISONS ON NUSCENES DATASET WITH **PILLAR SIZE (0.2, 0.2)**. NOTE THAT **L** MEANS LIDAR-BASED METHODS.

Method	Modality	mAP↑	NDS↑
WYSIWYG	L	0.350	0.419
PointPillar	L	0.401	0.550
PMPNet	L	0.454	0.531
SSN	L	0.467	0.582
CenterPoint	L	0.491	0.597
Ours	L (0.2, 0.2)	**0.513**	**0.613**
PointPainting	C + L	0.464	0.581
3D-CVF	C + L	0.527	0.623
SimpleFusion	C + L (0.2, 0.2)	**0.537**	**0.653**

B. Comprehensive Comparisons

Comparisons with camera-based methods. We compare our camera stream with existing visual 3D object detection approaches. For fair comparisons, we use the detection head of our camera stream from BevDet with the same settings and compare on validation dataset in nuScenes. As shown in Table I, the detector using our camera stream outperforms state-of-the-art methods on nearly all the evaluation metrics (above 1.14% increase on mAP). Our detector utilizes point clouds to supervise depth prediction during training and only input images for testing. Existing visual 3D object detectors mainly rely on the final loss to train the depth prediction, which is prone to more errors. Experimental results show the superiority of our method.

Comparisons with lidar-based methods. We compare our SimpleFusion with existing 3D object detectors based on multi-sensor fusion. As shown in Table II, SimpleFusion has a better performance compared to popular fusion methods(above 1% increase on mAP and NDS). At the same time, it has a significantly improvement compared to the singlesensor methods. Specifically, we adopt an adaptive fusion module, which can dynamically fuse the features of multiple sensors to obtain a better performance. Experimental results demonstrate the superiority of our method.

C. Ablation Study

Effectiveness of our fusion method. To verify the effectiveness of our fusion method, we conduct experiments with a single sensor, including camera-only, lidar-only, and multisensor fusion. As listed in Table III, our fusion method achieves better performances than those using single sensors. Experimental results prove the superiority of our dynamic fusion: the lidar stream plays a more crucial role than the camera stream in 3D object detection, but the added visual features can further improve the effect of 3D object detection.

Effectiveness of adaptive fusion module. To evaluate the effectiveness of the adaptive fusion module, we experimented with several fusion methods, including direct addition, maximum, concatenation, and dynamic weighting, in our SimpleFusion framework. As shown in Table. IV, the results show the effectiveness of our dynamic weight fusion. General fusion methods such as addition and maximum cannot obtain effective information in fusion, but dynamic fusion can effectively obtain key parts of features in the fusion process and give it a larger weight to achieve more robust 3D object detection.

TABLE III. EFFECTIVENESS OF OUR FUSION METHOD.

Method	Modality	mAP↑	NDS↑
Camera-Only	C	0.327	0.387
Lidar-Only	L	0.513	0.613
SimpleFusion	C + L	**0.537**	**0.653**

TABLE IV. EFFECTIVENESS OF ADAPTIVE FUSION MODULE.

Fusion Method	**Modality**	**mAP↑**	**NDS↑**
Addition	C + L	0.509	0.613
MaxPooling	C + L	0.513	0.607
Concatnation	C + L	0.521	0.638
Our Default Setting	C + L	**0.537**	**0.653**

V. CONCLUSION

This paper proposes a novel 3D object detector to dynamically fuse multi-sensor features. In the camera stream, we predict the depth probability distribution at the pixel level and convert the image feature space to the BEV space based on camera parameters and depth map. In the lidar stream, we use the perspective view and the BEV features to overcome the sparsity of point clouds and use deformable convolution to obtain object-level features for multi-scale and multi-view fusion. For multi-sensor fusion, we integrate a simple fusion module to focus on more confident features based on dynamic weights. Experiments on nuScenes dataset show that our method outperforms other state-of-the-art models by a large margin.

REFERENCES

[1] 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,” Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 20), IEEE Press, Jun. 2020, pp. 11618–11628, doi:10.1109/CVPR42600.2020.01164.

[2] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, “3D object proposals using stereo imagery for accurate object class detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, May. 2018, pp. 1259–1272, doi: 10.1109/TPAMI.2017.2706685.

[3] Y. Liu, T. Wang, X. Zhang, and J. Sun, “PETR: Position embedding transformation for multi-view 3d object detection,” in Computer Vision–ECCV 2022, vol 13687, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner Eds. Cham: Springer, 2022, pp. 531–548.

[4] J. Philion and S. Fidler, “Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D,” in Computer Vision – ECCV 2020, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 194–210.

[5] Y. Yan, Y. Mao, and B. Li, “Second: Sparsely embedded convolutional detection,” Sensors, vol. 18, no. 10, Oct. 2018, p. 3337, doi:10.3390/s18103337.

[6] T. Yin, X. Zhou, and P. Kr¨ahenb¨uhl, “Center-based 3D object detection and tracking,” Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 21), IEEE Press, Jun. 2021, pp. 11779–11788, doi:10.48550/arXiv.2006.11275.

[7] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, “PointPainting: Sequential fusion for 3D object detection,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 20), IEEE Press, Jun. 2020, pp. 4604–4612, doi:10.1109/CVPR42600.2020.00466.

[8] C. Wang, C. Ma, M. Zhu, and X. Yang, “PointAugmenting: Cross-modal augmentation for 3D object detection,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 21), IEEE Press, Jun. 2021, pp. 11794–11803, doi:10.1109/CVPR46437.2021.01162.

[9] T. Yin, X. Zhou, and P. Kr¨ahenb¨uhl, “Multimodal virtual point 3D detection,” Proc. Advances in Neural Information Processing Systems 2021 (NIPS 21), vol. 34, MIT Press, Dec. 2021, pp. 16494–16507, doi:10.1109/CVPR42600.2020.00466.

[10] S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, and L. Zhang, “FusionPainting: Multimodal fusion with adaptive attention for 3D object detection,” Proc. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC 21), IEEE Press, Sept. 2021, pp. 3047–3054, doi:10.1109/ITSC48978.2021.9564951.