Multi-Stage Fusion for Multi-Class 3D Lidar Detection

Zejie Wang[1]*, Zhen Zhao[2], Zhao Jin[2], Zhengping Che[2], Jian Tang[3], Chaomin Shen[1]✉, Yaxin Peng[4]
[1]School of Computer Science *and* Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University
[2]Didi Chuxing
[3]Syracuse University
[4]Department of Mathematics, School of Science, Shanghai University
51194506037@stu.ecnu.edu.cn, alexzhaozhen@163.com
{jinzhao, chezhengping}@didiglobal.com, jtang02@syr.edu
cmshen@cs.ecnu.edu.cn, yaxin.peng@shu.edu.cn

Abstract

In autonomous driving, the robust and accurate perceptions of the environment is a fundamental and challenging task. Resorting to the advancing of different sensors such as LiDAR and Camera, the autonomous systems are able to capture and process complementary perceptual information for better detection and classifying objects. In this paper, we propose a LiDAR-Camera fusion method for multi-class 3D object detection. The proposed method makes the utmost use of data from the two sensors by multiple fusion stages, and can be learned in an end-to-end manner. First, we apply a multi-level gated adaptive fusion mechanism with the feature extraction backbone. This point-wise fusion stage assiduously exploits the image and point cloud inputs, and obtains joint semantic representations of the scene. Next, given the regions of interest (RoIs) proposed based on the LiDAR features, the corresponding Camera features are selected by RoI-based feature pooling. These features are used to enrich the LiDAR features in local regions and enhance the proposal refinement. Moreover, we introduce a multi-label classification task as an auxiliary regularization to the object detection network. Without relying on extra labels, it helps the model better mine the extracted features and discover hard object instances. The experiments conducted on the KITTI dataset have proved all our fusion strategies are effective.

*Work done primarily while Zejie Wang was an intern at Didi AI Labs, Didi Chuxing.

1. Introduction

The 3D object detection tasks [2, 29, 24, 11, 17, 20, 21, 23, 27] create both an opportunity and a challenge for the intelligent transportation industry as a whole. Demands for multi-class 3D object detection are increasing in complex traffic situations, particularly in large, metropolitan areas. As the foundational components, LiDAR and Camera are two most common sensory inputs in autonomous driving. LiDAR points provide 3D structure information, but suffer from uneven and sparse points distribution. Especially small distant objects are hardly to be recognized by a LiDAR-only model due to extremely low density points. Cameras can capture images with rich semantic features while inevitably lack depth information. The demonstration is shown in Figure 1. To fully utilize the advantage of each sensor modality, LiDAR points and image features are combined to enhance detection accuracy. However, many existing works [2, 12, 11, 20, 5] tend to focus on fusing vehicle mixture perceptual information rather than other classes, such as pedestrian and cyclist. Other fusion works [24, 21] requires additional complex networks for a priori reasoning. To fill this gap, we propose an end-to-end learnable architecture for fusing Camera and LiDAR sensors in a feature-wise manner for multi-class 3D object detection.

The structure of the fusion model can be divided into multi-view projection fusion [2, 12, 11, 20] and feature mapping fusion [24, 21, 23]. In the multi-perspective projection fusion method, most of the works adopt via perspective projection and voxelization to quantify point cloud to pseudo-image by using BEV (birds-eye-view) map. The BEV format typically comes up with a CNN detection head to predict bounding box, which can be processed by the existing mature convolutional neural network. Unfortunately,

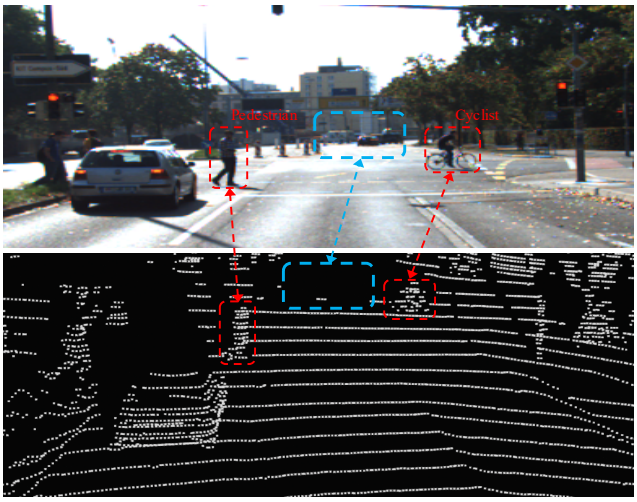

Figure 1. The problems faced by LiDAR-based multi-class 3D detection. The sparse 3D structure captured by LiDAR cannot effectively express the object. The small goals of 'Pedestrian' and 'Cyclist' are very sparse, such as the red areas. For relatively distant target points, it is almost missing, such as the blue area.

the BEV format progressively breaks down the spatial features and inevitably downscales resolution, and the conversion process is inefficient. As a result, the detection head fails to regress localization precisely due to insufficient point cloud features.

The feature map fusion method is usually recommended to employ an image segmentation auxiliary network [1, 16, 30] to extract semantic information before fusing with LiDAR features to ensure the effective spatial structure of the point cloud. The image segmentation auxiliary network is off-the-shelf and not learnable. Specifically, it is separately pre-trained and only performs inference in the fusion network's training phase, which is efficient. [21] has shown that the precision of the whole fusion network mainly relies on the performance of the segmentation auxiliary network. Meanwhile, the recognition situation of the auxiliary segmentation network and the point cloud detection network is split, and mutual optimization cannot be realized. Moreover, obtaining segmentation mask annotations is less cost-effective than bounding box labeling in terms of human resources. It is rarely possible to apply for access to open source dataset for both image segmentation and LiDAR object annotations at same frames.

To address these challenges, we propose a novel end-to-end LiDAR-Camera fusion method, multistage fusion for multi-class 3D object detection (MSF-MC). We design a classification-aware auxiliary model which can generate image features for multi-label prediction to guide LiDAR backbone to learn more discriminative features. Unlike popular LiDAR-Camera fusion network [21, 23], instead of image segmentation annotations, the object classification labels are sufficient to provide multi-label supervision. Importantly, we apply VGG Network as the segmentation backbone to emphasize that the performance of image backbone is not critical for the LiDAR stream prediction accuracy. The fusion model leverages the attention mechanism to adaptively inject high-dimensional semantic image features into the LiDAR encoder in point-wise manner through a lightweight gated network module. As a result, the LiDAR encoder-decoder module could maintain the point cloud structural feature losslessly and efficiently and receive supplemental image semantic information. We further add image global semantic information to point cloud ROIs (regions of interest) to amplify the refinement of the proposal regression. In order to achieve the correlation between LiDAR predicted objects and corresponding image object classification, we adopt classification-based regularization mechanisms to efficiently assist point cloud detector to recognize small objects.

In summary, our key contributions are as follows:

1. This is a novel work for multi-class LiDAR-Camera fusion that can be used for end-to-end training, which does not require additional segmentation annotations.

2. We propose a multi-level gated fusion method to retain the original point cloud structure information, while provide feature supplement in the ROI.

3. We develop a prediction consistency regularization mechanism to align the feature gap between LiDAR detection and image classification.

4. We conducted extensive experiments on the authoritative dataset KITTI, which proved the effectiveness of our proposed method.

2. Related Work

LiDAR-based 3D object detection Like the development of image object detection, there are two categories in LiDAR object detection, one stage approach and two stage approach. A typical one stage model consists of a point feature extraction module and a detection module. Point feature extraction module generally produces bird eye's view or voxelization grids. The joint detection head is either 2D or 3D CNN to learn the features for 3D box prediction. Complex-yolo [19] projected point cloud to BEV format and used 2D detector. PointPillars [10] encoded point cloud with 6 statistical quantities and stacked voxel features as 'pillar'. VoxelNet [29] exploited the PointNet as the backbone to extract features for each voxel. While Yang *et al.* [26] applied a 3D CNN to a grouped voxel grid. Though one stage approach efficiently saves computing resources, as a transformed compact presentation, voxelization inevitably lose original spatial information and result in relatively low precision.

The two-stage approach lifts a 3D region of interests in

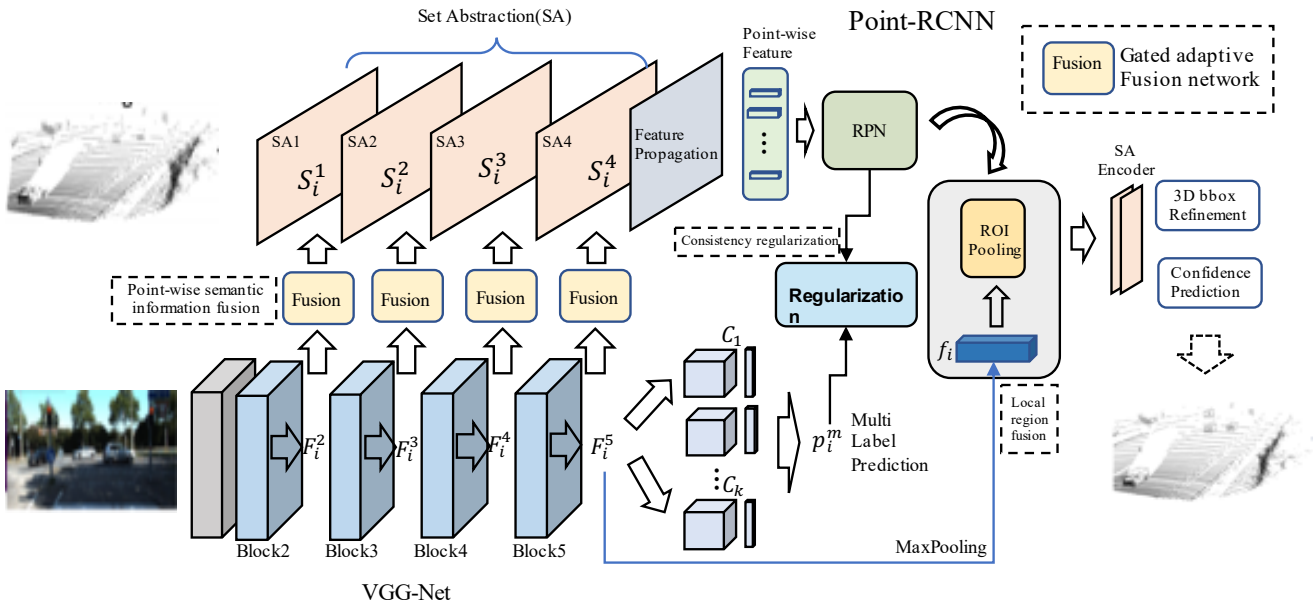

Figure 2. The structure of the proposed multistage fusion for multi-class 3D object detection model(MSF-MC). The proposed model uses PointRCNN as the 3D object detection network, and the VGG-16 structure as the backbone network for image feature extraction to achieve 3-stage fusion. The point-wise semantic information fusion structure is used to realize the point-wise fusion of multi-level image depth semantic information and point cloud features. The object region fusion structure is used to achieve the feature fusion of the object region. At the same time, the consistency regularization structure realizes the multi-label prediction results to implement the point cloud prediction consistency regularization mechanism.

stage one and refine the regression in stage two. Part-a^2 net [18] generated ROIs from voxel-based feature extraction layer and adopted intra-object part-aware analysis to enrich ROI features. Votenet [13] utilized PointNet++ [15] as the backbone and introduced the Hough voting principle to group deep features. Yang *et al.*[28] used PointNet++ [15] as well to keep spatial information and divided ROIs into voxel grid to fit regular CNN. Generally, the two-stage approach utilizes PointNet++[15] as the backbone, especially SA layers and FP layers, to generate proposal for the refinement module. It achieves better accuracy by learning more fine-grained features from the proposal.

Multi-sensor 3D object detection In the past years, there has been a rapid rise in the use of multi-sensors. Some works [14, 22] proposed to produce ROIs from image and applied the PointNet backbone to extract corresponding LiDAR features. One obvious limitation was 3D frustum view was extruded from 2D region, the image backbone dominated the performance of 3D detection. Chen *et al.* [2] gathered image, front view point cloud and bird view point cloud as the threefold input branches to generate 3D proposal. The variant research [9] proposed image features in the proposal generation stage. These works had a cumbersome structure that involved different backbones for each view though the optimization were end-to-end.

There is a trend that researchers are paying more attention on image semantic information in fusion work rather than relying on image detection. Huang *et al.* [5] merged point cloud features with image semantic features in a point-wise way. Xie *et al.* [23] proposed an attention fusion module to solely merge 3D proposals and image segmentation masks. Vora *et al.* [21] applied a semantic segmentation network obtaining pixel-wise image segmentation scores, to decorate point cloud for fine-grained semantic understanding. The above works inspire us to take advantages of global semantic features from images.

3. Method

In this section, we propose a multi-stage fusion model for multi-class 3D object detection (MSF-MC) that can be used for end-to-end training. We obtain two different sensory inputs from LiDAR and Camera. Given the $\{x_{[i]}, b_{[i]}, c_{[i]}\}_{i=1}^{n}$ donate the LiDAR data set containing n frames, where $x_{[i]}$ presents the point cloud at i-th frame, $b_{[i]}$ and $c_{[i]}$ are the bounding box location and object classification relatively. Meanwhile, the calibrated Camera dataset $\{z_{[i]}, c_{[i]}\}_{i=1}^{n}$ can be obtained, where $z_{[i]}$ represents image data corresponding to $x_{[i]}$. Because the $z_{[i]}$ are synchronized with $x_{[i]}$, it is possible to use $c_{[i]}$ for supervised learn-

ing of the image content without introducing additional annotations. For the convenience of the following presentation, we use x and z to represent the current i-th point cloud and image respectively. We aim to thoroughly use two different sensory inputs to gain the understanding of complementary features, and make the 3D object detection more precise.

The principle of MSF-MC utilizes the gated adaptive network to accomplish the effective fusion of multi-level image semantic information and point cloud features. To additionally refine the ROIs local feature, image global semantic information is concatenated with ROI module in the LiDAR stream. we also adopt multi-task training strategy to perform the consistency regularization of cross-modality object class prediction. As mentioned in the related work, we select a widely used point-based 3D object detector, PointRCNN, as the detection model in the LiDAR stream. In terms of image stream, to simplify the training and to emphasize the effectiveness of our fusion method, the lightweight VGG-16 model is exploited as the backbone for multi-label learning and semantic feature acquisition. Unlike previous work [23, 21], the efficiency of image backbone won't dominate our fusion performance. Its structure is shown in Figure 2.

3.1. Point-wise Semantic Information Fusion

In multi-class object detection scenarios, it is crucial that the learned image and point cloud features include effective local regions. So we propose to employ point-wise regional self-attention fusion principle in the first stage to realize sensitivity to effective regions. We exploit the mapping strategy to leverage the correspondence relationship between LiDAR points and image pixels in a feature wise manner. LiDAR detection branch consists of four set abstraction (SA) layers as the feature encoding module. Given the point cloud x, we gain the point cloud feature set $\left\{S^l\right\}$ ($l \in [1, 2, 3, 4]$) from these set abstraction layers. Concurrently we obtain the deep semantic features $\left\{F^l\right\}$ ($l \in [2, 3, 4, 5]$) for the the last four layers of the VGG backbone. In order to establish the mapping correspondence of $S^l \rightarrow F^{l+1}$, where $l \in [1, 2, 3, 4]$.

In the same way as in [6, 5], we leverage bilinear interpolation and calibration matrix M to project 3D point to corresponding image feature where we can sample 2D feature map and accordingly obtain the feature set $\left\{V^l\right\}$ ($l = 2, 3, 4, 5$), an example of which is shown in Figure 3. Specifically, for a specific point e in a given point cloud space, according to the mapping matrix and bilinear interpolation method, we can obtain its corresponding position $\hat{e}$ in the feature map output by the Camera image branch. Therefore, we can further obtain point-wise feature V^l by sampling the size area of the image feature F^l. The number of points in V^{l+1} is exactly the same as the number

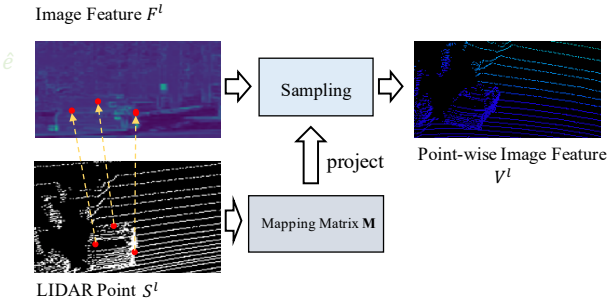

Figure 3. Project 3D point to corresponding image feature. Using the mapping matrix M, the point cloud is projected into the image, and the image features are sampled at the corresponding pixel positions.

of points in S^l. With this strategy, the original space of the point cloud can be preserved and depth semantic information can be introduced at the same time. As a consequence, we could apply multi-level fusion on $\left\{S^l\right\}$ and $\left\{V^l\right\}$, which is proved more effective in the following experiment section.

On account of illumination, occlusion and truncation, discrepant representation might occur during cross modal perception fusion [6, 5]. Inspired by [4, 7], we propose a gated adaptive fusion mechanism which is able to evaluate the relevance of point cloud feature and image feature accordingly. Detailed demonstration is shown in Figure 4. Specifically, by concatenating point-wise image feature V^{l+1} and point cloud feature S^l, we get a compressed feature vector SV^l, and obtain weighted feature map from $W = \sigma(MLP(SV^l)$, where MLP denotes shared weighted feature extraction network and σ is sigmoid activation function. To balance the complementary feature, we design two boosting attention equations $S_w^l = S^l \odot W$ and $V_w^{l+1} = V^{l+1} \odot (1 - W)$, where $\odot$ presents element wise production. We further merge above two weighted feature to update LiDAR feature S^l:

$$S^l = S_w^l \oplus V_w^{l+1} \tag{1}$$

where $\oplus$ stands for merge operation. Self-attention feature fusion learning is expected to enhance the relevance of S^l and F^{l+1}, in order to achieve effective fusion in original spatial structure.

3.2. Local Region Fusion

In the previous section we were mainly concerned with point-wise fusion of corresponding features in the same space. Since the global semantic information of the image can directly and effectively express the scene, we suggest that in the second stage, the target area feature of ROI-Pooling extracted by the $RCNN$ module in PointRCNN should be fused with the global depth semantic in-

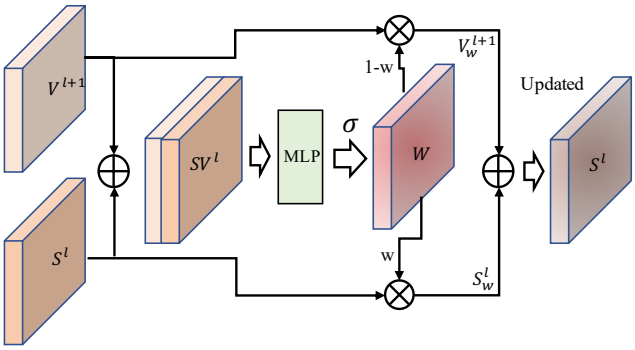

Figure 4. Gated adaptive fusion network. The point-wise features obtained from two different sensors are fused using relevant attention features to update the point cloud features.

formation of the image to enhance target information and achieve more precise regression prediction. In the image feature extraction stage, z can obtain the feature $F^5 \in \mathbb{R}^{H\times W\times C}$ through the last layer of the feature extractor to generate the global depth semantic feature vector $f = MaxPooling(F^5) \in \mathbb{R}^{1\times C}$. Similarly, in the current point cloud x, the ROI-Pooling region features $R = \{r_1, r_2, \ldots, r_t\}$ can be obtained, where t represents the number of features obtained. At this time, there is no conflict between the global semantic information and the characteristics of the target area, so we suggest that the global semantic information of the image can be directly supplemented with the object region information:

$$R^* = \{MLP(r_i \oplus f)\}_{i=0}^{t} \tag{2}$$

where $\oplus$ represents a cascading operation on two features. We use the fused object region features R^* as the input of the final refined regression task for training update.

3.3. Multi-Label Prediction Auxiliary Regularization

In this section, we introduce the multi-label training method and the accomplishment of the consistency regularization on the third stage.

Multi-label learning To make the image feature extractor learn fine-coarse semantic information, we send z to it and obtain F^5 as the input of multi-label classification. The multi-label classifier is composed of K separate binary classification sub-module, $C_1, C_2, \ldots, C_k$, to predict K different classes. Each sub-module is made up of 3×3 kernel, 512 channel convolution layer and a $sigmoid$ activation function. The labels c of image z could be converted to $y^c \in \mathbb{R}^{k\times 1}$ via a fixed $[0, 1]$ encoding transformation. Then the multi-label classifier can be learned by minimizing the cross-entropy loss:

$$L^{ml}{}_i = y^{cT} log(p^m) + (1 - y^c)^T log(1 - p^m) \tag{3}$$

The output prediction vector $p^m \in \mathbb{R}^{1\times k}$ consists of k object classification possibilities. Through multi-label classification loss training, we expect to be able to achieve effective extraction of image features and obtain predictions containing objects.

Cross-modal multi-objective category regularization In the RPN module of PointRCNN, we can obtain the N point-wise features of the current point cloud x. These features use the classifier to obtain the corresponding categorical variables to construct the segmentation part of the background before and after for precise adjustment. The prediction probability matrix $Q \in \mathbb{R}^{N\times K}$ of K categories can be obtained on the N point cloud features. Further, we can get the multi-category prediction probability vector $q^m = Max(Q) \in \mathbb{R}^{1\times K}$ of the current point cloud x, where Max is max operation, which takes $N \times K$ matrix as the input and return the vector of each column's max value. We adopt the distribution-wise asymmetric measure - KL divergence (D_{kl}) to accomplish the prediction consistency between two different sensor modalities. We prospect the usage of KL divergence to make prediction robust and avoid incorrect prediction. In order to maintain the consistency between the prediction generated by the 3D detector and the prediction generated by the image multi-label target recognition, we suggest to use $Softmax$ function to normalize the vector p^m and q^m. The loss function of cross modal regularization is defined as following:

$$L^{kl}{}_i = D_{kl}(p^m||q^m) + D_{kl}(q^m||p^m) \tag{4}$$

3.4. Overall Learning

We introduce three-stage fusion method to promote the performance of 3D object detection. We group 3D object detection loss $L^{3d}{}_i$, image multi-label classification loss $L^{ml}{}_i$ and cross-modal consistency regularization loss $L^{kl}{}_i$ to achieve multi-task training. The total loss function on all data is interpreted as:

$$L^{all} = \sum_{i=0}^{n} (L^{3d}{}_i + \lambda L^{ml}{}_i + \mu L^{kl}{}_i) \tag{5}$$

Among them, λ and μ are trade-off parameters for weighing different loss conditions. We use Adaptive Moment Estimation (Adam) [8] optimization algorithm for training.

4. Experiments

We evaluated our proposed fusion method (MSF-MC) on the KITTI 3D object detection dataset [3]. In the following,

we introduce the dataset and model implementation details in Sec.4.1. In Sec.4.2, we provide the evaluation results of our model on KITTI. Extensive ablation experiments are performed in Sec.4.3. Finally, we provide qualitative visualization results in Sec.4.4.

4.1. Dataset and Implementation Details

The KITTI dataset We evaluate our work on KITTI, the widely used 3D object detection dataset, which has 7,481 training images and 7518 testing images. The training set was annotated on both point cloud data and image data, and testing result must be uploaded to official test server. Following the same splitting protocol as [17], we further divide the training set to 3712 training samples and 3769 validation samples. We train our model on three commonest classes, 'Car', 'Pedestrian' and 'Cyclist'. We use average precision (AP) metric as the comparison method, which is calculating recall at 40. In our experiment, we provide the result for each official difficulty standard, easy, moderate and hard, of both validation samples and testing set. The difficulty standard is defined by size, occlusion and truncation. We also follow the official IoU threshold protocol, 0.7 for 'Car', 0.5 for 'Pedestrian' and 'Cyclist'.

Implementation details We adopt the common settings in [17]. We use the Adam optimizer to train the network for 80 epochs. The initial learning rate, momentum parameter and weight decay are set to 0.005, 0.01, and 0.9, respectively. Set the λ=0.01, and μ=0.05. In order to align the input of the network, 16384 points are obtained in the viewable area of the Camera as the input of PointRCNN. For scenes with less than 16384 points, we randomly repeat these points to obtain 16384 points. We select the point cloud located between the ranges of (0m, 70.4m),(-40m, 40m), and (-3m, 1m) along the x,y and z axes, and delete the remaining areas of invisible points. The four set abstraction layers subsample and encode the point cloud with sizes of 4096, 1024, 256, and 64, respectively. We resize the image size to [1280, 384] as the input of the VGG Network.

Data augment Many LiDAR-based object detection methods adopt the data augment [25, 10, 17] to increase the performance. This data augment strategy sampled all the annotated object from point cloud, and pasted some of them randomly to the scene which has relatively few amount of objects. However, we abandon this effective augment strategy during the whole training phase. The correct matching between the original point cloud and image pixel must be guaranteed. Any point-cloud-based data augment method will dilapidate the matching relationship because it won't generate corresponding images.

4.2. Results on the KITTI Dataset

We adopt the 3D/BEV indicator task to show our performance. We first compare with our own replicated object detection detector that can give multi-category confidence at the same time, and the results on the validation sets are shown in Table 1. Our proposed MSF-MC method is better than PointRCNN in all three categories of AP indicators, and the improvement is obvious for 'Pedestrian' and 'Cyclist'. The average performance of 'Pedestrian' and 'Cyclist' on the three difficulty levels increased by 2.29% and 6.85%, respectively. We can also find that our model still has excellent performance even in the face of the difficulty level of 'hard'. This shows that our proposed method is indeed effective and can significantly improve the categories, which are inherently small and difficult to detect.

We also provide a performance demonstration of BEV metrics, as shown in Table 2. Our proposed MSF-MC also showed excellent performance. Significant improvement for 'Car', 'Pedestrian' and 'Cyclist'. It further illustrates the effectiveness of our method.

4.3. Ablation Study

In this section, we conduct the comprehensive analysis to evaluate different proposed methodologies on the 'Car', 'Pedestrian' and 'Cyclist' classes.

Investigations on different fusion stages To figure out the effectiveness of different methodologies, we demonstrate the ablation study in Table 3. When we only use the point-wise semantic information fusion module (SF), the average performance of 'Pedestrian' and 'Cyclist' at the three difficulty levels is increased by 2.94% and 2.42%, respectively, compared to the baseline. This shows that the use of the SF can effectively supplement the semantic features of the point cloud, making the model more conducive to detecting sparse object. The module of local region fusion (RF) has a more obvious contribution to 'Pedestrian', with an average performance increase of 6.19%. When multi-objective category regularization (MR) module is adopted, the average performance of 'Pedestrian' and 'Cyclist' is increased by 3.88% and 2.23% respectively, further proving that this mutually influencing regularization mechanism can effectively promote Camera recognition results to help the 3D detection network to be sensitive to sparse small objects. When we integrate all of these modules, we get the best overall performance at the moment. However, it is worth mentioning that the category accuracy of 'Pedestrian' has a corresponding decline, but it is still much higher than the baseline. At the same time, there is a huge improvement in 'Cyclist'. It shows that the model is more inclined to detect 'Cyclist' under the condition of ensuring that the other two categories are at a higher level.

Table 1. The comparison of 3D mAP results of our proposed method with Point-RCNN.

Method	Car			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointRCNN[17]	88.78	77.67	75.15	66.26	55.59	48.45	75.38	52.25	48.24
MSF-MC (Ours)	89.63	80.06	75.83	66.69	58.18	56.21	82.36	59.17	58.71

Table 2. The comparison of BEV results of our proposed method with Point-RCNN.

Method	Car			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointRCNN[17]	92.97	86.28	82.05	72.16	83.82	55.79	75.44	54.69	52.04
MSF-MC (Ours)	93.42	86.97	84.54	74.51	82.75	55.74	86.36	66.66	59.60

Ablation study of SF The SF module uses the idea of multi-scale to integrate the features of LiDAR point cloud and image. Semantic content information is more prevalent at higher levels of the extraction network. Therefore, we begin to merge the SF module from the higher-level features and gradually transition to the lower-level feature fusion. We conduct ablation study on the SF module, and the experimental results are reported in Table 4. We can see that when the third and fourth layers are fused simultaneously, the overall performance of the model is optimal. But when added to the lower level ($l = 2, 3$), it actually degrades its overall performance, which has a significant negative impact on the 'Pedestrian' category. This indicates that the higher semantic features of the image are more suitable for fusion operation. It is worth noting that, in order to ensure the integrity of the model, the SF module in the above experiment adopts the method of 4-layers simultaneous fusion.

Parameter sensitivity on λ and μ The multi-objective category regularization mechanism is based on the multi-label prediction output through the regularization loss L_{kl}. The label information output by the image classification network and the category information from the LiDAR point cloud are mutually influencing. In order to study the influence of these loss components, we conduct a more comprehensive ablation study.

In Figures 5 and 6, we conducted sensitivity analysis on the two trade-off parameters λ and μ. μ controls the weight of multi-label feature fusion, while λ controls the degree of focusing on hard-to-classify examples. Other parameters are set to their default values.

Figure 6 shows the sensitivity of our proposed model to the changes of λ parameters. We adjust λ by fixing $\mu = 0.05$. We can see that when the parameter λ changes from 0.001 to 0.01, the basic trend of detection performance of the model for the three categories is constantly improving. When the value of λ is further increased, it can be found that the model performance is in a downhill stage. It reveals that in the fusion mechanism of LiDAR-Camera, the semantic features learned by multi-label learning have an obvious correlation to the detection performance of 3D point cloud targets, which can also further explain why many workers now adopt a very powerful prior model to fuse point cloud. Figure 5 shows the change range of the parameter μ when $\lambda = 0.01$ is fixed. We can find that

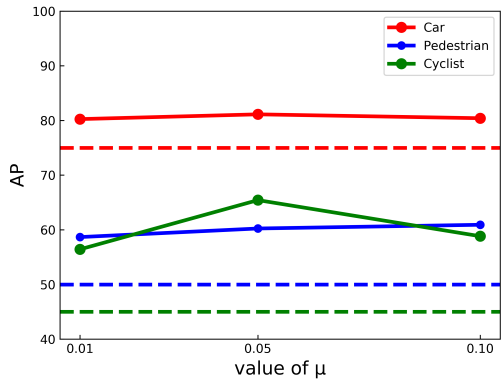

Figure 5. Parameter sensitivity analysis on μ.

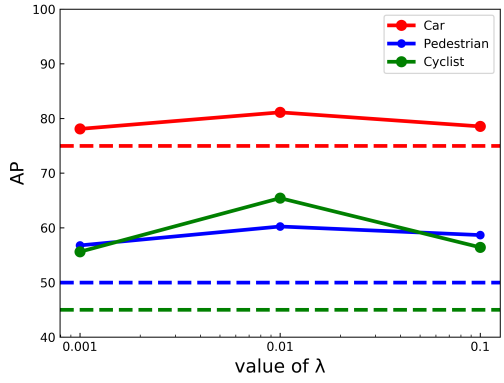

Figure 6. Parameter sensitivity analysis on λ.

Table 3. The ablation study on SF, RF, MR on val set.

Ablation			Car			Pedestrian			Cyclist		
SF	RF	MR	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
			88.78	77.67	75.15	66.26	55.59	48.45	75.38	52.25	48.24
✓			88.95	77.70	75.28	69.79	60.63	51.72	76.78	55.26	51.10
	✓		88.50	77.60	75.16	72.17	62.85	53.88	76.03	54.22	51.51
		✓	89.23	77.71	75.18	69.09	60.15	52.73	77.42	53.91	51.22
✓	✓	✓	89.63	80.06	75.83	66.69	58.18	56.21	82.36	59.17	58.71

Table 4. Detection results of SF with deployment on different layers. Use l to represent the point-wise semantic information fusion of the point cloud feature S^l of the l-th layer and the image semantic feature F^{l+1} of the l+1-th layer.

SF				mAP		
l=1	l=2	l=3	l=4	Car	Pedestrian	Cyclist
			✓	79.65	58.95	65.63
		✓	✓	80.72	58.41	66.23
	✓	✓	✓	81.18	54.91	65.02
✓	✓	✓	✓	81.84	60.36	66.75

the multi-objective category regularization mechanism controlled by μ is more sensitive to the category of 'Cyclist'.

4.4. Compare with other fusion method

In Table 5, it can be found that our proposed method has no obvious difference in the category of 'Car' compared with the latest methods,but it is worth noting that our method is a multi-class end-to-end training method, mainly focusing on the other two classes of small target sparse point clouds. Our focus is on fusion methods, rather than optimizing the basic model of 3D detection.

Table 5. Performance comparison of 3D AP(Car) with previous methods on KITTI validation sets

Method	3D AP(Car)		
	easy	moderate	hard
MV3D [2]	71.29	62.68	56.56
ContFuse [12]	82.54	66.22	64.04
AVOD-FPN [9]	84.41	74.44	68.65
F-PointNet [14]	83.76	70.92	63.85
PI-RCNN [23]	88.27	78.53	77.75
PointPainting [21]	88.38	77.74	76.76
MSF-MC (Ours)	88.14	77.48	75.92

5. Conclusion

In this paper, we study the limitations of current LiDAR-Camera fusion works and propose a multistage fusion for multi-label 3D object detection task. We have implemented end-to-end training by adopting a three-stage fusion method without the need to introduce additional annotations. In the first stage, the multi-level gated adaptive fusion mechanism is adopted to achieve point-wise fusion of the features of point cloud and image, so as to ensure the spatial structure of point cloud and introduce effective image semantic information. In the second stage, the camera features are fused with point cloud ROI-pooling to enhance the integrity of proposals. We also introduce the third stage, using the multi-label classification task as the auxiliary regularization of the object detection network to achieve the consistency of category recognition. The experimental analysis shows that our proposed method is very effective.

Acknowledgements

This work was supported in part by National Key Research and Development Program of China (grant number 2018YFF01013402), Shanghai Science and Technology Innovation Action Plan (20511100200), and Science and Technology Commission of Shanghai Municipality (14DZ2260800).

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[2] 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.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3D object detection. In *European Conference on Computer Vision*, pages 35–52. Springer, 2020.

[6] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12605–12614, 2020.

[7] Jaekyum Kim, Junho Koh, Yecheol Kim, Jaehyung Choi, Youngbae Hwang, and Jun Won Choi. Robust deep multi-modal learning based on gated information fusion network. In *Asian Conference on Computer Vision*, pages 90–106. Springer, 2018.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[10] 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.

[11] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019.

[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] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019.

[14] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from RGB-D data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018.

[15] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

[16] Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–733, 2018.

[17] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019.

[18] Shaoshuai Shi, Zhe Wang, Xiaogang Wang, and Hongsheng Li. Part-aˆ 2 net: 3D part-aware and aggregation neural network for object detection from point cloud. *arXiv preprint arXiv:1907.03670*, 2(3), 2019.

[19] Martin Simony, Stefan Milzy, Karl Amendey, and Horst-Michael Gross. Complex-yolo: An Euler-region-proposal for real-time 3d object detection on point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[20] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019.

[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] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. *arXiv preprint arXiv:1903.01864*, 2019.

[23] Liang Xie, Chao Xiang, Zhengxu Yu, Guodong Xu, Zheng Yang, Deng Cai, and Xiaofei He. Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12460–12467, 2020.

[24] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2018.

[25] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018.

[26] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3D object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018.

[27] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11040–11048, 2020.

[28] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3D object detector for point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1951–1960, 2019.

[29] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, pages 4490–4499, 2018.

[30] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019.