# IDA-3D: Instance-Depth-Aware 3D Object Detection from Stereo Vision for Autonomous Driving

Wanli Peng* Hao Pan* He Liu Yi Sun†
Dalian University of Technology, China
{1136558142,panhao15320,lhiceu}@mail.dlut.edu.cn, lslwf@dlut.edu.cn

## Abstract

*3D object detection is an important scene understanding task in autonomous driving and virtual reality. Approaches based on LiDAR technology have high performance, but LiDAR is expensive. Considering more general scenes, where there is no LiDAR data in the 3D datasets, we propose a 3D object detection approach from **stereo vision** which does not rely on LiDAR data either as input or as supervision in training, but solely takes RGB images with corresponding annotated 3D bounding boxes as training data. As depth estimation of object is the key factor affecting the performance of 3D object detection, we introduce an Instance-Depth-Aware (IDA) module which accurately predicts the depth of the 3D bounding box's center by instance-depth awareness, disparity adaptation and matching cost reweighting. Moreover, our model is an end-to-end learning framework which does not require multiple stages or postprocessing algorithm. We provide detailed experiments on KITTI benchmark and achieve impressive improvements compared with the existing image-based methods. Our code is available at https://github.com/swords123/IDA-3D.*

## 1. Introduction

Three-dimensional object detection is one of the most important scene understanding tasks that has many applications such as autonomous driving and virtual reality. It classifies objects and estimates oriented 3D bounding boxes of physical objects from input sensor data. According to the type of sensor, 3D object detection can be divided into point cloud-based methods [6, 5, 11, 21, 30, 24, 16, 15, 28, 13], monocular image-based methods [3, 20, 19, 27, 12, 22, 1, 25, 18] and binocular image-based methods [4, 14, 23, 25]. Three-dimensional object detection based on point clouds such as those from LiDAR can achieve the best performance, but LiDAR sensors are the most expensive. Moreover, some 3D datasets do not provide LiDAR data, such as PASCAL3D+ [26], which annotates 3D bounding boxes by means of the CAD model. Monocular cameras are the cheapest and most convenient to install, but 3D detection using only a single image inherently lacks reliable depth information. Binocular cameras are not expensive and can provide denser information for small objects in the distance compared to LiDAR and inherently provide absolute depth information compared to monocular cameras. We therefore focus on stereo-based 3D object detection in this paper.

Stereo-based 3D object detection takes stereo pairs of images as input and oriented 3D bounding boxes of objects as output. Since the depth error from stereo vision increases quadratically with distance, 3D object detection without depth maps in the training process is a difficult task if solely relying on the annotated 3D bounding boxes [4, 14]. Currently, stereo-based 3D object detection without supervised depth still lags behind in localizing objects. We hope to bridge the gap between two kinds of methods, with and without depth-data supervision, so that the performance of 3D object detection with only annotated 3D bounding boxes of physical objects can approach the performance of those that use depth images as supervision.

In this work, we propose a 3D object detection approach from **stereo vision** that does not rely on LiDAR data either as input or as supervision during training but solely takes RGB images with corresponding annotated 3D bounding boxes as training data. It first extracts the objects from the background by stereo Region Proposal Network (RPN) to remove its interference on 3D object detection. Because depth estimation of an object instance is the key factor affecting the performance of 3D object detection, we design a separate instance-depth-aware (IDA) module that predicts the center depth of an object's 3D bounding box. Unlike previous stereo-based methods that compute the correspondence of each pixel between two images [10, 2], we measure the correspondence of each instance, paying more attention to the global spatial information of the object. To reduce the error of depth estimation for a distant object, we adaptively adjust the range of disparity levels in the cost volume

*The first two authors contributed equally to this work.
†Corresponding author.

according to the location of the object and transform the uniform quantization of the disparity level into nonuniform quantization. The matching cost is also reweighted to make depth estimation more discriminative by penalizing the depth levels that are not unique for an object instance and promoting the depth levels that have high probabilities. The overview of the proposed architecture is illustrated in Fig. 1.

Our main contributions are listed as follows:

• We propose a stereo-based end-to-end learning framework for 3D object detection that does not rely on depth images either as input or for training and does not require multistage or postprocessing algorithms.

• We introduce an instance-depth-aware (IDA) module that accurately predicts the depth of the 3D bounding box's center by instance-depth awareness, disparity adaptation and matching cost reweighting, thus improving the accuracy of 3D object detection.

• We provide detailed experiments on the KITTI 3D dataset [7] and achieve state-of-the-art performance compared with the stereo-based methods without depth map supervision.

## 2. Related Work

There are two scenarios for 3D object detection: indoor and outdoor. This work briefly reviews the recent literature on outdoor autonomous driving based on LiDAR point clouds, monocular images and stereo images.

**Point Clouds for 3D Object Detection.** Because LiDAR can provide the three-dimensional information of objects, it is naturally used as input to detect 3D objects and is currently the most accurate method for 3D object detection. LiDAR data on 3D object detection can be represented in various ways, including direct 3D point clouds [21, 24], 3D volumetric forms [30, 6], and 2D-front-view or bird's-eye-view images [13, 11, 16, 15, 28, 5]. Point clouds and volumetric forms can fully utilize the 3D information of the object and have a one-to-one relation with the 3D pose. However, they are both high-dimensional and computational inefficient methods. To reduce the high dimensionality of 3D representations, some works project 3D point clouds to front-view or birds-eye-view images, leveraging mature convolutional networks for 3D object localization.

**Monocular Images for 3D Object Detection.** Because monocular cameras are cheaper and more flexible to install than LiDAR or stereo cameras, 3D object detection using monocular cameras naturally becomes a requirement for industry and academia. Some recent techniques [3, 19, 12, 20] extend the state-of-the-art 2D object detector to regress the orientation of the object's 3D bounding box and its dimensions by transferring 2D detection, orientation, and scale estimation into 3D space. Others [27, 22, 1] localize 3D objects from a monocular RGB image via geometric reasoning in both the observed 2D projection and the unobserved depth dimension. Because monocular image inherently has scale ambiguity in depth estimation and the error in depth estimation increases greatly with distance, its performance is not as good as the other two methods.

**Stereo Pairs of Images for 3D Object Detection.** 3D object detection relying on LiDAR has achieved high accuracy. However, LiDAR sensors are expensive and therefore cannot enable intelligent driving systems to be used by millions of users. With the development of deep learning methods, there is hope that the accuracy of 3D object detection based on stereo vision will be improved. Stereo-based 3D object detection methods usually create 2D or 3D object proposals with extra geometric constraints [4, 14], which are then used to regress the object pose. 3DOP [4] makes use of the 3D point cloud features estimated from a stereo camera pair to estimate each 3D candidate proposal by a greedy algorithm; then, a 3D object detection network taking 3D object proposals as input is presented to predict accurate 3D bounding boxes. Stereo R-CNN [14] uses a coarse-to-fine 3D bounding box estimation method, where a stereo RPN is exploited to predict 2D left-right boxes, sparse keypoints, viewpoints, and object dimensions for calculating a coarse 3D object bounding box; then, the Stereo R-CNN recovers the accurate 3D bounding box by a region-based photometric alignment using left and right RoIs.

The accuracy of depth-based 3D object detection relies heavily on the quality of depth estimation. Some works have focused on stereo-based depth estimation to obtain the true 3D bounding box of objects. For example, LiDAR data have been incorporated into the training step as supervision for depth estimation [25, 29], which is called pseudo-LiDAR [25]. Pseudo-LiDAR introduces a two-step approach for stereo-based 3D object detection. It first estimates the depth map from stereo vision using the supervised depth information generated by LiDAR and then converts the depth map into a 3D point cloud and takes advantage of existing LiDAR-based models for 3D object detection.

Different from the above methods, considering that some 3D datasets, such as PASCAL3D+ [26], do not have LiDAR data, we only take RGB images with corresponding annotated 3D bounding boxes as training data, and propose an entire instance depth aware 3D object detection approach. Our approach is closely related to recent work [14] that uses solely training data of RGB images and corresponding annotated 3D bounding boxes. This prior work defines four semantic keypoints indicating the four corners at the bottom of the 3D bounding box. The 3D box can be solved by minimizing the reprojection error of the 2D boxes by perspective keypoints in the first stage; then, the method

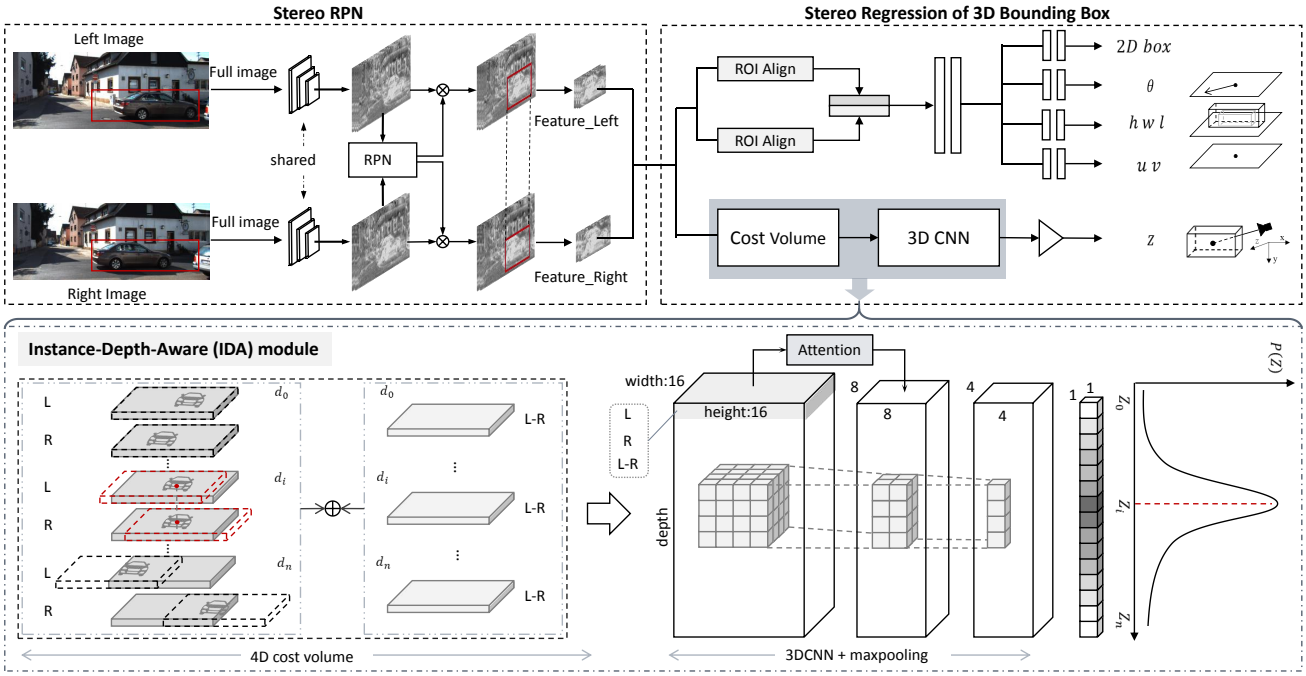



Figure 1. Overview of the proposed IDA-3D. Top: Stereo RPN takes a pair of left and right images as input and outputs corresponding left-right proposal pairs. After stereo RPN, we predict position, dimensions and orientation of 3D bounding box. Bottom: Instance-depth-aware module builds a 4D cost volume and performs 3DCNN to estimate the depth of a 3D bounding box center.

solves the disparity of the 3D bounding box center for further dense 3D-box alignment in the second stage. In contrast to this prior work, our approach achieves better performance in 3D object detection by designing a novel end-to-end instance-depth-aware module that directly predicts the single-variable depth $z$ of the 3D bounding box's center, rather than making predictions by additional keypoints and dense 3D box alignment. Instead of pixel-level post-processing, our instance depth estimation module makes the features extracted by the 3D convolution focus on the object by end-to-end training, reducing the interference of cluttered background on the depth estimation of object.

## 3. Method

We propose a stereo 3D object detection approach where the training data are solely RGB images with corresponding annotated 3D bounding boxes, without taking depth obtained from LiDAR as input or as intermediate supervision. Rather than design any step of the stereo algorithm by hand, we learn an end-to-end mapping from an image pair to object 3D bounding boxes using deep learning. Instead of constructing a machine learning architecture as a black box, we recognize that the 3D object detection error stems entirely from the error in depth estimation $z$ of a 3D bounding box center, thus we separately design a regression model to obtain the instance depth. In this paper, the coordinate $z$ of a 3D bounding box center is also called instance depth. Furthermore, we guide architecture design of object depth estimation by instance-depth awareness, disparity adaptation and matching cost reweighting. Therefore, we learn an effective Instance-Depth-Aware 3D Object Detection model (IDA-3D). Our architecture is illustrated in Fig. 1. In the remainder of this section, we discuss each component in detail.

### 3.1. Overview of Object Instance Detection

We first extract a pair of Regions of Interest (RoI) for each object in the left and right images with the stereo RPN module inspired by [14], the purpose of which is to avoid the complex matching of all pixels between the left and right images and eliminate the adverse effect of background on object detection. The stereo RPN creates an union RoI for each object whose size and location are the same on the left and right images so that the union RoI ensures the starting points of each pair of RoIs. After applying RoIAlign [8] on left and right feature maps respectively, the left and right RoI features are then concatenated and fed into the stereo regression network to predict position, orientation and dimensions of 3D bounding box respectively, where the position of the 3D bounding box can be represented by its center position $(x, y, z)$. Since the 3D depth of object center has a large dynamic range and its deviation accounts for the majority of the difference in 3D object detection, we separately design the IDA module to obtain the depth of a 3D bound-

| Name | Layer Setting | Output Dimension |
|---|---|---|
| input | | $D \times 16 \times 16 \times 96$ |
| conv0 | $3 \times 3 \times 3, 64$<br>$3 \times 3 \times 3, 128$ | $D \times 16 \times 16 \times 128$ |
| maxpool0 | maxpooling<br>stride=(1,2,2) | $D \times 8 \times 8 \times 128$ |
| conv1 | $3 \times 3 \times 3, 128$<br>$3 \times 3 \times 3, 128$ | $D \times 8 \times 8 \times 128$ |
| maxpool1 | maxpooling<br>stride=(1,2,2) | $D \times 4 \times 4 \times 128$ |
| conv2 | $3 \times 3 \times 3, 64$<br>$3 \times 3 \times 3, 1$ | $D \times 4 \times 4$ |
| avgpool | avgpooling<br>stride=(4, 4) | $D \times 1 \times 1$ |

Table 1. Parameters of the proposed IDA model. D denotes the number of depth levels.

ing box center which is also called instance depth in this paper. In stereo regression network, we also predict the 2D bounding box as the input of IDA module during inference.

### 3.2. Instance Disparity (Depth) Estimation

Unlike previous stereo networks that regress the disparity of each pixel between a rectified stereo images, we are specifically interested in computing the disparity of each instance to locate its position. Instead of computing the correspondence of each pixel between two images, we measure the correspondence of the same instance between two images, paying more attention to the global spatial information of the object. Therefore, after forming a cost volume of dimensionality $disparity \times height \times width \times feature\ size$ by concatenating the left and right feature maps across each disparity level, we employ two consecutive 3D convolution layers, each followed by a 3D max-pooling layer, to learn and perform down sampling on feature representations from the cost volume. Since disparity is inversely proportional to depth and both represent the position of an object, we transform the disparity into depth representation after formulating cost volume. Relying on the networks regularization, the down sampled features by 3D CNN are finally merged into depth probability of the 3D box center. By performing the sum of each depth $z$ weighted by its normalized probability, the depth of a 3D box center is finally obtained, as shown in Eq. 1, where N denotes the number of depth levels and $P(i)$ the normalized probability.

$$\hat{z} = \sum_{i=0}^{N} z_i \times P(i) \tag{1}$$

We train our model with supervised learning using ground truth depth of 3D box center, where supervised regression loss is defined using the error between the ground truth depth($z$) and the model's predicted depth ($\hat{z}$) in Eq. 2:

$$L(z, \hat{z}) = smooth_{L1}(\hat{z} - z) \tag{2}$$

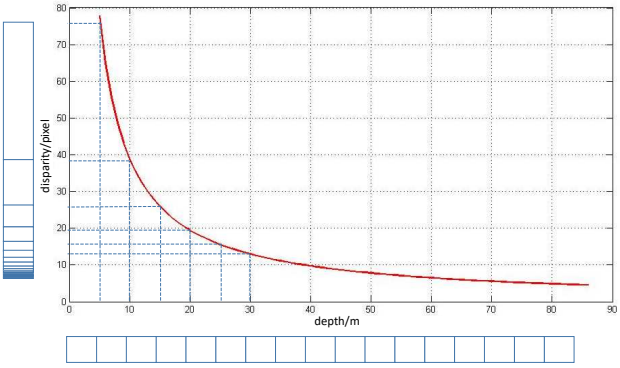



Figure 2. Relation between disparity and depth.

As shown in IDA module in Fig. 1, if the disparity level of a car is $d_i$, when its left and right feature maps shift to the opposite direction, the two feature maps match exactly at $d_i$, as shown by the red mark in 4D cost volume. The IDA module output the maximum probability at $Z_i$, where $Z_i$ is the depth value corresponding to $d_i$.

Our model parameters are shown in Table 1. Fig. 3 takes the depth estimation of a car as an example to visualize this process, where the bright yellow and dark blue color in the feature maps indicate stronger activation and the lower activation respectively. Given the features of cost volume as input, it can be seen that the feature maps extracted by our network are gradually changed from low-level features of the car to high-level global features of its center depth probability. Meanwhile, the depth level that is unique for the car results in the highest probability as shown at the bottom of the figure. This phenomenon illustrates that the proposed model is effective to learn the probability of correct depth level for an object instance.

### 3.3. Instance Disparity (Depth) Adaptation

Most previous works optimize the accuracy of disparity estimation. However, for the same disparity error, the error in depth increases quadratically with distance. It means that the influence of the disparity error in depth estimation of a far-away object is greater than a nearby one. This is the key factor that leads to poor 3D object detection. In order to adapt the model and loss function to lay more emphasis on a far-away object, we change the disparity level in cost volume from uniform quantization to nonuniform quantization where the farther the object is, the less the partition cell between two consecutive disparity levels. In this way, the depth of a distant object can be more precisely estimated. The nonuniform quantization or disparity is shown in Fig. 2. We take advantage of nonuniform disparity quantization

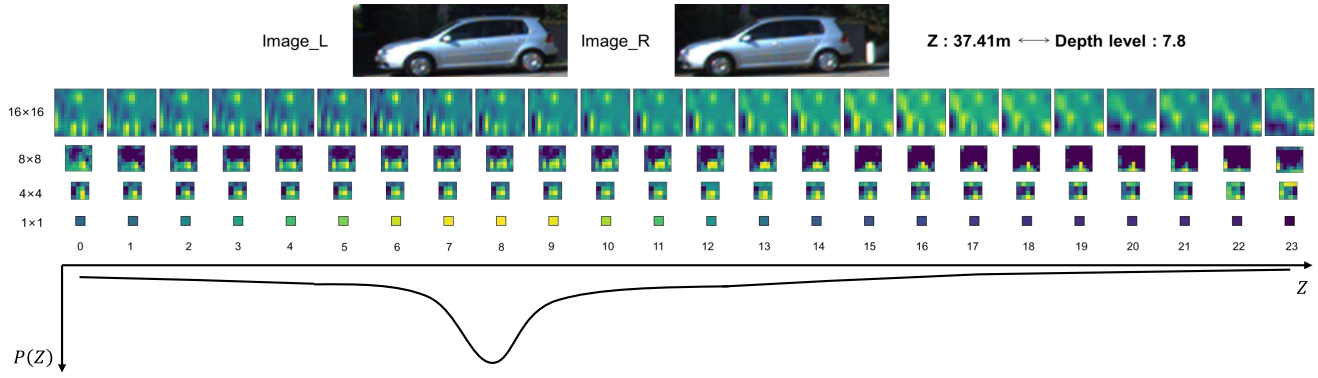



Figure 3. Global spatial information extraction process. Feature maps are sampled at a channel and sorted by the depth level. The bright yellow color in the feature map indicates stronger activation, while dark blue indicates the lower activation.

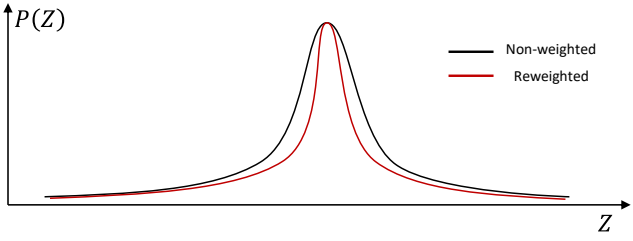



Figure 4. A graphical depiction of the matching cost reweighting which we propose in this work. The two cost curves along each depth level before and after reweighting are respectively shown in black and red.

converted from uniform depth quantization via the following transform ($f_u$: horizontal focal length, $b$: baseline of binocular camera),

$$D = \frac{f_u \times b}{z} \tag{3}$$

In addition to the nonuniform quantization, we don't have to estimate the depth in the range of 0-80m since the depth of a car is inversely proportional to its size in the image. Given camera intrinsic parameters, we can roughly calculate the range according to the width of the union box in the image. We therefore reduce the searching range in depth estimation to $[z_{\min}, z_{\max}]$, where $z_{\min}$ and $z_{\max}$ represent the minimum and maximum depth values of each object respectively. Such depth adaption minimizes the average partition cell of quantization for a fixed number of disparity levels, thus improving depth estimation.

### 3.4. Matching Cost Reweighting

As Eq. 1 indicates that the depth of a 3D box center is a weighted average of all depth levels, not the most likely, which may lead to non-discriminative depth estimation. To penalize the depth levels that are not unique for an object instance and promote the depth levels that have high probabilities, we reweight the matching cost. The reweighting is split into two parts, with the first part (shown in the 4D cost volume of Fig. 1) in 4D volume packing a difference feature map between the left and right feature maps across each disparity level and second part (shown in the 3DCNN in Fig. 1) in 3DCNN employing attention mechanism on depth. The 4D volume with these residual feature maps will make the subsequent 3D CNN take into account the difference between left and right feature maps in a certain depth level and refine depth estimation, while disparity attention mechanism sets the weight $r_i$ for each channel. The correlation score $r_i$ that is obtained by calculating the correlation between left and right feature maps on each disparity is defined as:

$$r_i = \cos < F_i^l, F_i^r >= \frac{F_i^l \cdot F_i^r}{\|F_i^l\| \cdot \|F_i^r\|} \tag{4}$$

where $r_i$ is the weight for the $i^{th}$ channel, cos is the cosine similarity function, and $F_i^l$, $F_i^r$ are the $i^{th}$ pair of feature maps in the cost volume. The two cost curves along each depth level before and after reweighting are respectively shown in black color and red color in Fig. 4. We can see that the gradient of the reweighted curve is steeper than that of the non-weighted curve which shows the increased probability of correct instance depth.

### 3.5. 3D Object Detection

In addition to the instance depth estimation, we also need to estimate the horizontal and vertical coordinates $(x, y)$ of the object center, object stereo bounding boxes, dimensions and viewpoint angle to complete the task of 3D object detection. We design a six-parallel fully-connected network with the concatenated left and right RoI features as input. After the depth of the instance is determined, the coordinates $(x, y)$ of the object center in the left cameras coordinate system can be calculated according to its projections

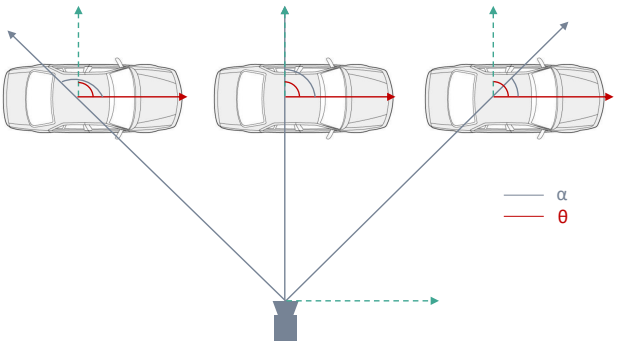



Figure 5. Relation between object orientation $\theta$ and the viewpoint angle $\alpha$.

$(u, v)$ on the left and right image pairs as follows,

$$x = \frac{(u - c_u) \times z}{f_u} \qquad y = \frac{(v - c_v) \times z}{f_v} \tag{5}$$

where $(c_u, c_v)$ denotes the camera center and $f_u$, $f_v$ are the horizontal and vertical focal length. From Eq. 5, we can observe that the estimation of the horizontal and vertical position of the 3D box center is affected by the result of depth estimation, which indicates the significant role of depth estimation in object detection. Because of no translation in the vertical coordinate $(y)$ between the left and right center of an object instance, this branch estimates the offset $(\Delta u^l, \Delta v, \Delta u^r)$ to the groundtruth directly.

For the orientation regression as shown in Fig. 5, we estimate the viewpoint angle which accounts for the change in appearance using the method similar to $Multibin$ in [20]. The orientation angle $\theta$ can be calculated by Eq. 6, which illustrates that the result of depth estimation affects the orientation as well.

$$\theta = \alpha + \tan^{-1} \frac{x}{z} \tag{6}$$

For the dimension regression, we produce dimension offsets $(\Delta h, \Delta w, \Delta l)$ to the mean class sizes $(\bar{h}, \bar{w}, \bar{l})$, which are the mean dimensions over all objects of the given class. The dimensions of 3D bounding box can be calculated via the following transformation,

$$h = \bar{h} e^{\Delta h} \qquad w = \bar{w} e^{\Delta w} \qquad l = \bar{l} e^{\Delta l} \tag{7}$$

### 3.6. Implementation Details

The whole multi-task loss can be formulated as:

$$\begin{aligned} L = w_1 L_{rpn} + w_2 L_{2Dbox} + w_3 L_{3D}^{(u,v)} + w_4 L_{3D}^{z} \\ + w_5 L_{dim} + w_6 L_{\alpha} \end{aligned} \tag{8}$$

where the $L_{rpn}$ and $L_{2Dbox}$ denote the loss of 2D boxes on stereo RPN module and stereo regression module respectively, $L_{3D}^{(u,v)}$ is the regression loss for the projection of object instance centers and $L_{3D}^{z}$ the instance depth of objects, $L_{dim}$ the offset regression loss for the 3D bounding box dimensions, $L_{\alpha}$ the orientation loss which includes a classification loss for the discrete angle bins and a regression loss for the angle bin offsets, $w_1$-$w_6$ the trade-off parameters to balance the whole multi-task loss.

We employ two weight-share ResNet50 [9] with FPN [17] architecture as the feature extractor. In the training stage, we flip the images in the training set, exchange the left and right image and mirror the 2D boxes annotation, viewpoint angle and 2D projection of centroid at the same time for data augmentation. For the IDA module, we divide the depth between $z_{\max}$ and $z_{\min}$ into 24 levels for the estimation of the object center depth. We use the 2D boxes obtained from the RPN module as the input to the IDA module, as this module can provide more samples for training. While during inference, we use the 2D boxes obtained from the 2D regression module as input of IDA module because it provides fewer boxes with more precision which can reduce the computation cost. We train our network using SGD optimizer with the initial learning rate 0.02, the momentum 0.9 and the weight decay 0.0005. Meanwhile, we set the batch size to 4 on two NVIDIA 2080Ti GPUs and train 80000 iterations with about 26 hours.

## 4. Experiment

We evaluate our approach on the KITTI 3D object detection dataset which contains 7481 training images and 7581 testing images. We follow the same training and validation splits as [4], each contains 3712 and 3769 images respectively. We report 3D average precision ($AP_{3D}$) and birds-eye-view average precision ($AP_{bev}$) on car category with the IoU thresholds at 0.5 and 0.7, where each category is divided into easy, moderate, and hard case according to the 2D box height, occlusion and truncation levels. In Sec. 4.1, we give our results and make comparisons to monocular-based [3, 1, 18] and stereo-based methods [4, 23, 14] quantitively. Qualitive results are given in Sec. 4.2.

### 4.1. Results of Instance-Depth-Aware Approach

We conduct experiments both qualitatively and quantitatively. For comparison, we summarize the main results from monocular to binocular methods in Table 2. Our method outperforms previous monocular-based methods across all IoU thresholds by a significant margin in easy, moderate and hard cases. Comparing to stereo-based methods, we obtain the highest $AP_{3D}$ and $AP_{bev}$ at 0.5 IoU and 0.7 IoU. Taking an $AP_{3D}$ of 60.04% using IoU = 0.5 in hard case as an example, our method yields consistent improvement over other methods, 3DOP (30.09%), TLNet (37.99%) and Stereo R-CNN (57.24%). Results in other cases follow similar trends which indicate the consistent detection performance of our approach. Specifically, the result of our method $AP_{bev}$ = 67.3% in hard case at 0.5 IoU outperform-

| Method | Sensor | IoU = 0.5 | | | IoU = 0.7 | | |
|---|---|---|---|---|---|---|---|
| | | Easy | Mode | Hard | Easy | Mode | Hard |
| Mono3D [3] | M | 30.50/25.19 | 22.39/18.20 | 19.16/15.52 | 5.22/2.53 | 5.19/2.31 | 4.13/2.31 |
| M3D-RPN [1] | M | 55.37/48.96 | 42.49/39.57 | 35.29/33.01 | 25.94/20.27 | 21.18/17.06 | 17.90/15.21 |
| Xinzhu et al. [18] | M | 72.64/68.86 | 51.82/49.19 | 44.21/42.24 | 43.75/32.23 | 28.39/21.09 | 23.87/17.26 |
| 3DOP [4] | S | 55.04/46.04 | 41.25/34.63 | 34.55/30.09 | 12.63/6.55 | 9.49/5.07 | 7.59/4.10 |
| TLNet [23] | S | 62.46/59.51 | 45.99/43.71 | 41.92/37.99 | 29.22/18.15 | 21.88/14.26 | 18.83/13.72 |
| Stereo R-CNN [14] | S | 87.13/85.84 | 74.11/66.28 | 58.93/57.24 | 68.50/54.11 | 48.30/36.69 | 41.47/31.07 |
| ours | S | **88.05/87.08** | **76.69/74.57** | **67.29/60.01** | **70.68/54.97** | **50.21/37.45** | **42.93/32.23** |

Table 2. $AP_{bev}$ / $AP_{3D}$ (in %) of the car category on KITTI validation set, where S denotes binocular image pair as input and M denotes monocular image as input.

| Method | $AP_{bev}$ / $AP_{3d}$ (IoU = 0.7) | | |
|---|---|---|---|
| | Easy | Mode | Hard |
| ours | 70.68/54.97 | 50.21/37.45 | 42.93/32.23 |
| PL+FP [25] | 69.7/54.9 | 48.1/36.4 | 41.8/31.1 |
| PL+AVOD [25] | 74.0/56.7 | 54.7/37.9 | 47.3/34.3 |

Table 3. $AP_{bev}$ / $AP_{3D}$ (in %) of the car category on KITTI validation set between Pseudo-LiDAR [25] and our method.

s Stereo R-CNN (58.93%), achieving significant improvement over 8.37%. Similar observation can be seen in $AP_{3D}$ (ours/Stereo R-CNN = 74.57%/66.28%) with 8.29% performance improvement. This may be due to the nonuniform quantization strategy in IDA module which makes our method more robust for the distant objects by reducing the depth estimation error.

We also compare our approach to the Pseudo-LiDAR [25] of stereo version which follows a two-stage network: *i*) depth map estimation via PSMNet [2], *ii*) 3D bounding box regression via F-PointNet [21] or AVOD [11]. However, it is unfair to compare our approach with Pseudo-LiDAR since we do not use the depth map as intermediate supervision. Our method still achieves comparable performance as shown in Table 3. Besides, our method has less complexity because we form an end-to-end network with a light-weight IDA module compared to PSMNet and obtains a high speed of $>$ 12 frames per second on one NVIDIA 2080Ti GPU.

**Ablation Study on the Disparity Adaptation.** In order to verify the benefits of our disparity adaptation strategy, we evaluate the depth estimation error according to the distance using different disparity quantization strategies. The results of depth error are shown in Fig. 6. The error is calculated as the mean differences between the predicted 3D locations and the ground truth for detections with 2D IoU $>$ 0.5. As expected, using nonuniform quantization strategy leads to more reduction of depth estimation error as the distance increase. For the objects 50m away, it can be seen from the histogram that nonuniform quantization has a greater impact on the accuracy of depth estimation. This phenomenon proves our analysis that the distant objects, which have less interval between two consecutive disparity levels, can achieve better results of depth estimation. Since the depth estimation of object instance is the key factor affecting the performance of 3D object detection, our nonuniform quantization strategy performs significant improvements compared to uniform quantization strategy. The detailed statistics can be found in the first two rows of Table 4.

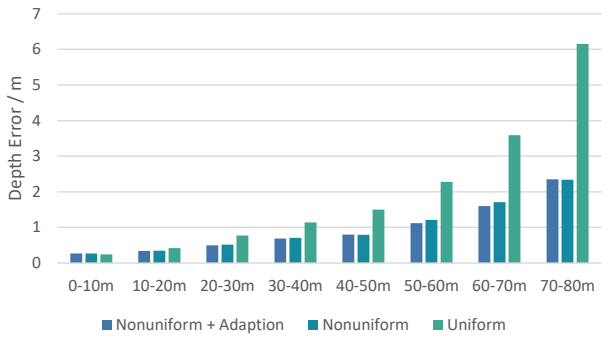
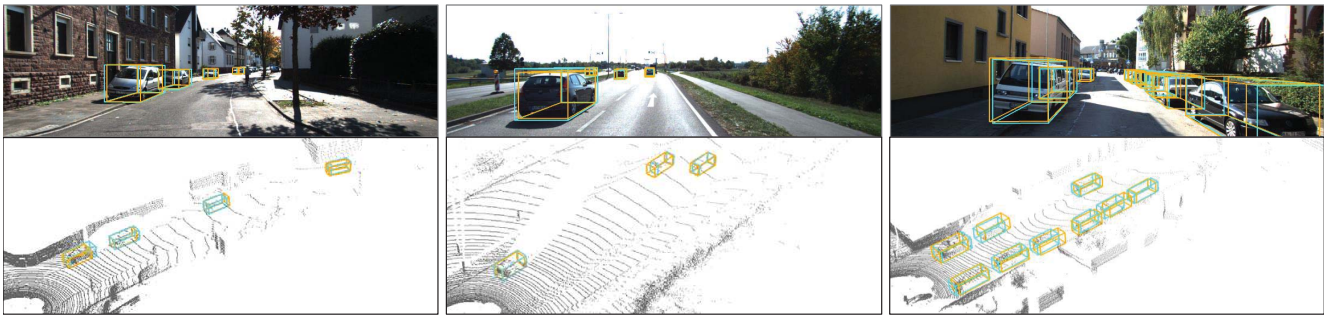



Figure 6. The depth estimation error from different disparity quantization strategies.

| Method | Metric | IoU = 0.7 | | |
|---|---|---|---|---|
| | | Easy | Mode | Hard |
| Uniform | $AP_{bev}$ | 46.59 | 32.35 | 29.58 |
| | $AP_{3D}$ | 34.57 | 23.40 | 21.19 |
| Nonuniform | $AP_{bev}$ | 67.01 | 49.17 | 42.23 |
| | $AP_{3D}$ | 52.16 | 36.40 | 30.93 |
| Nonuniform + Adaption | $AP_{bev}$ | **70.68** | **50.21** | **42.93** |
| | $AP_{3D}$ | **54.97** | **37.45** | **32.23** |

Table 4. The benefits of the disparity quantization strategy.



(a) (b) (c)

Figure 7. Quantitative results on several scenes in KITTI dataset. At the first row are the ground truth 3D boxes and the predicted 3D boxes projected to the image plane. We also show the detection results on point cloud in order to facilitate observation. The predicted results are shown in yellow and the ground truth are shown in blue.

| Diff. | Att. | $AP_{bev}$ / $AP_{3d}$ (IoU = 0.7) | | |
|---|---|---|---|---|
| | | Easy | Mode | Hard |
| ✓ | ✓ | **70.68/54.97** | **50.21/37.45** | **42.93/32.23** |
| ✓ | × | 67.08/52.17 | 49.90/36.85 | 42.65/31.99 |
| × | ✓ | 67.52/52.03 | 48.51/35.47 | 41.86/29.88 |
| × | × | 66.25/51.82 | 47.41/35.60 | 40.88/30.18 |

Table 5. Improvements of the matching cost reweighting.

Applying the same amount of depth quantization levels, smaller search range means better performance in instance depth estimation. Therefore, we adaptively reduce the search range of disparity via the size of objects in 2D images as described in Sec. 3.3. As expected, the adaptive strategy provides more precise disparity quantization and thus achieves better performance in 3D bounding box estimation as shown in the last row of the Table 4.

**Ablation Study on the Matching Cost Reweighting.** Table 5 shows the effect of the matching cost reweighting strategy. In our approach, we use two strategies to control and promote the peak of depth probability. The first strategy is concatenating a difference feature map between a left and right feature map with original cost volume, which is represented by Diff. in Table 5. And the second strategy is employing attention mechanism in 3DCNN, which is represented by Att. in Table 5. We conduct ablation experiments within our framework to verify the contributions of each strategy. Through Att. and Diff., we penalize the depth levels that are not unique for an object instance and promote the depth levels that have high probability. As a result, our method obtains performance improvements by combining two strategies together.

### 4.2. Qualitative Results

Fig. 7 shows qualitative detection results on several scenes in KITTI dataset. It can be observed that in the common street scenes, our method can accurately detect objects in the scene and the detected 3D boxes are well aligned on both the front-view images and point cloud. Especially when the objects are very far-away from the cameras, our method is still able to obtain accurate detection results as shown in (a) and (b), which benefits from our IDA module. In the cases that too many vehicles are occurred in the scene or heavily occluded by others, our method also has the potential to successfully locate these objects as shown in (c).

## 5. Conclusion

In this work, we propose an end-to-end learning framework for 3D object detection based on stereo images in autonomous driving. It does not rely on depth images either as input or for training and does not require multi-stage or postprocessing algorithms. A stereo RPN module is introduced to produce a pair of union RoI to avoid complex matching of the same object in a left-right image pair and reduce the interference of background on depth estimation. Without dense depth maps as supervision, the specially designed instance-depth-aware (IDA) module focuses on objects and directly performs the instance depth regression. Moreover, our approach pays more attention on far-away objects by disparity adaptation and matching cost reweighting. Our approach has a lightweight network architecture and achieves impressive improvements over the existing image-based performance. Comparing to some methods with depth map supervision, our approach obtains comparable performance as well.

## Acknowledgments

The work was supported by NSFC Grant (U1708263). We also thank Super-computing Center of Dalian University of Technology for providing a high performance computing platform.

## References


[1] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9287–9296, 2019.

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[3] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2156, 2016.

[4] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1259–1272, 2017.

[5] 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.

[6] Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, and Ingmar Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1355–1361. IEEE, 2017.

[7] 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.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[11] 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.

[12] Jason Ku, Alex D Pon, and Steven L Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11867–11876, 2019.

[13] Bo Li. 3d fully convolutional network for vehicle detection in point cloud. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1513–1518. IEEE, 2017.

[14] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019.

[15] 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 Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019.

[16] 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.

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

[18] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6851–6860, 2019.

[19] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2069–2078, 2019.

[20] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017.

[21] 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.

[22] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8851–8858, 2019.

[23] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: From monocular to stereo 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7615–7623, 2019.

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

[25] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019.

[26] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In

*IEEE Winter Conference on Applications of Computer Vision*, pages 75–82. IEEE, 2014.

[27] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2345–2353, 2018.

[28] 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.

[29] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*, 2019.

[30] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.