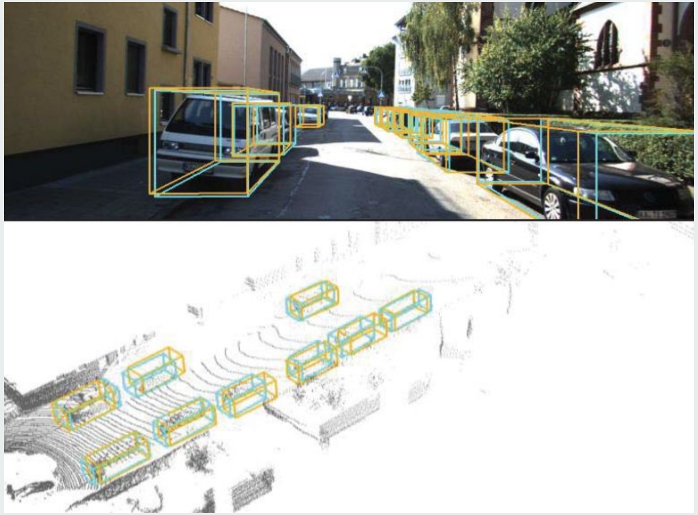
IDA-3D:
Instance-Depth-Aware 3D Object Detection from Stereo Vision for Autonomous Driving



Wanli Peng, Hao Pan, He Liu, Yi Sun
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020

Presented by:
Maria Isabel Saludares
7 October 2020

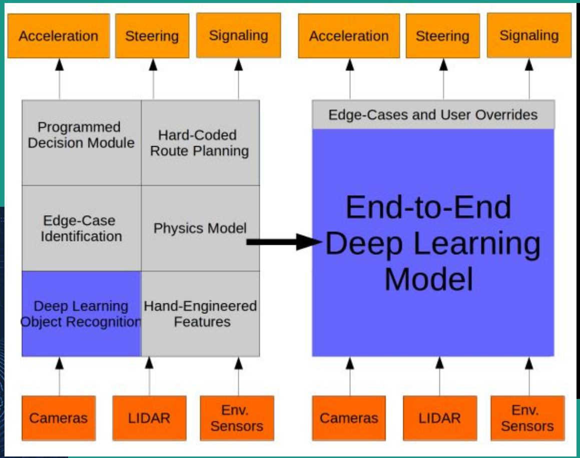
IDA-3D: Instance-Depth-Aware 3D Object Detection from Stereo Vision for Autonomous Driving

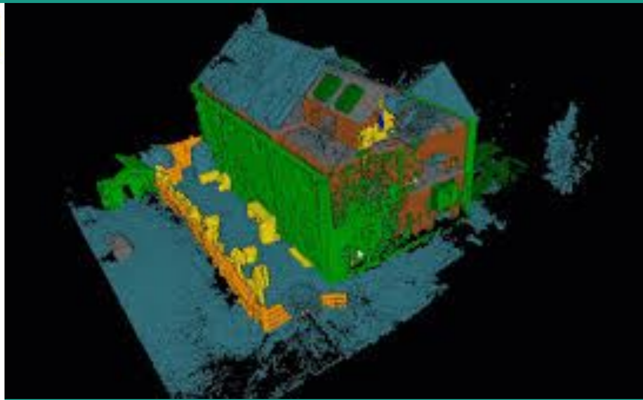
3D object detection is an important scene understanding task in autonomous driving and virtual reality. Approaches based on LiDAR technology have high performance, but LiDAR is expensive. Considering more general scenes, where there is no LiDAR data in the 3D datasets, we propose a 3D object detection 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 post-processing algorithm. We provide detailed experiments on KITTI benchmark and achieve impressive improvements compared with the existing image-based methods.

https://github.com/swords123/IDA-3D

Applications

- Scene understanding for Autonomous driving
- Virtual Reality and Augmented Reality





Rosenzweig, J., & Bartl, M. (2015). A review and analysis of literature on autonomous driving. E-Journal Making-of Innovation.
https://towardsdatascience.com/reinforcement-learning-from-grid-world-to-self-driving-cars-52bd3e647bc4

Object detection

Cloud-based methods [6, 5, 11, 21, 30, 24, 16, 15, 28, 13]

- LiDAR — one of the best performances, but expensive
- some datasets do not provide LiDAR data, such as PASCAL 3D+ [26]

Monocular image-based methods [3, 20, 19, 27, 12, 22, 1, 25, 18]

- cheapest, most convenient to install
- lacks reliable depth information

Binocular image-based methods [4, 14, 23, 25] (Stereo-based)

- not expensive
- can provide denser information for smaller objects in distance (compared to LiDAR)
- inherently provide absolute depth information

Motivation

Goal: Estimate the oriented 3D bounding boxes of objects from stereo vision for autonomous driving

➔ LiDAR is expensive -- propose a stereo-based 3D object detection approach which does not rely on depth data as input

➔ Stereo based depth estimated for an object -- especially far-away objects, m key factor affecting the performance of the detector

➔ Reduce depth estimation error by instance-depth awareness and improve the performance of the detector

Main Contributions

- ➔ 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 accu- racy 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.

[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.

IDA-3D Object Detector

Data

- ❑ **Pair images**
- ❑ **Position (x,y,z), Size(l,w,h) (6D)**

Output

- ❑ **Depth estimation (z)** -- IDA-3D Module
- ❑ **Position estimation --** horizontal and vertical coordinates (x,y) of the object center
- ❑ **Orientation estimation --** view point angle
- ❑ **Dimension estimation --** object stereo bounding boxes

IDA-3D

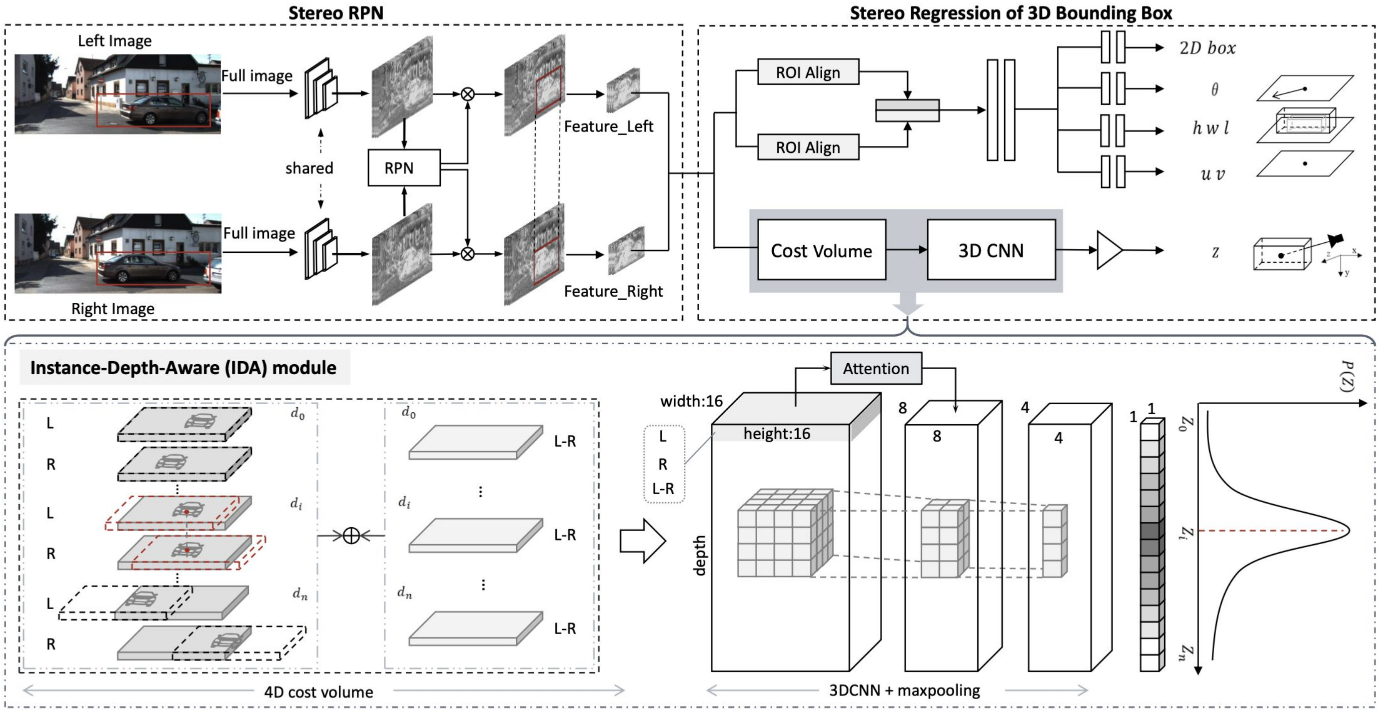


Figure 1. Overview of the proposed IDA-3D. Top: Stereo RPN takes a pair of left and right images as input and outputs corresponding left-right proposal pairs. After stereo RPN, we predict position, dimensions and orientation of 3D bounding box. Bottom: Instance-depth-aware module builds a 4D cost volume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

IDA-3D

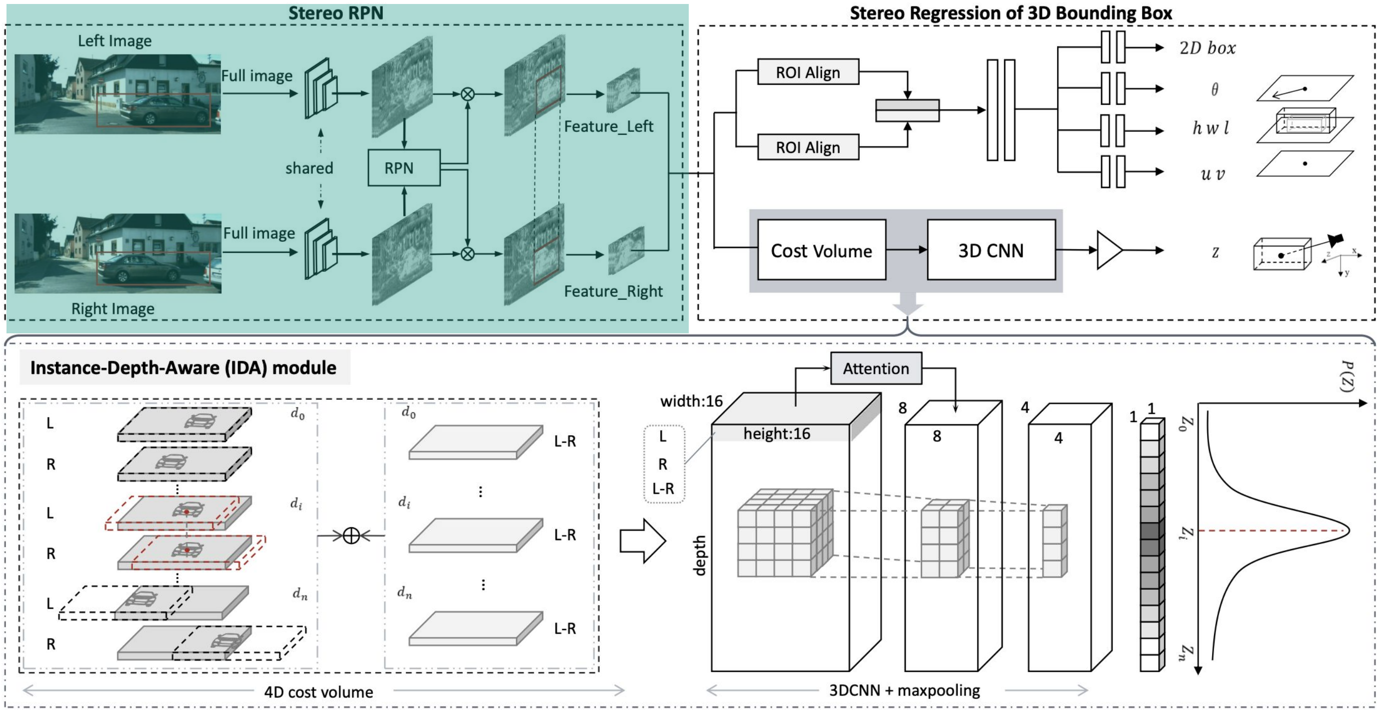
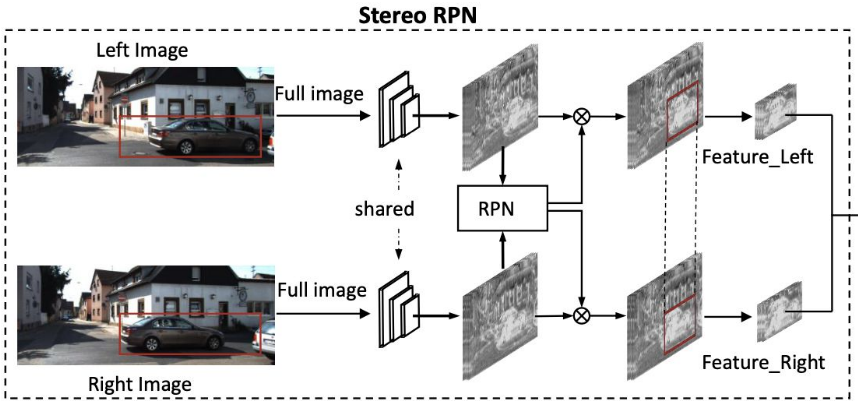


Figure 1. Overview of the proposed IDA-3D. Top: Stereo RPN takes a pair of left and right images as input and outputs corresponding left-right proposal pairs. After stereo RPN, we predict position, dimensions and orientation of 3D bounding box. Bottom: Instance-depth-aware module builds a 4D cost volume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

Stereo RPN

Purpose

- *avoid the complex matching of all pixels between left and right images*
- *Eliminate the adverse effect of background on object detection*



Output

creates an union RoI for each object whose size, location are the same on the left and right -- ensures starting points of each pair of RoIs

IDA-3D

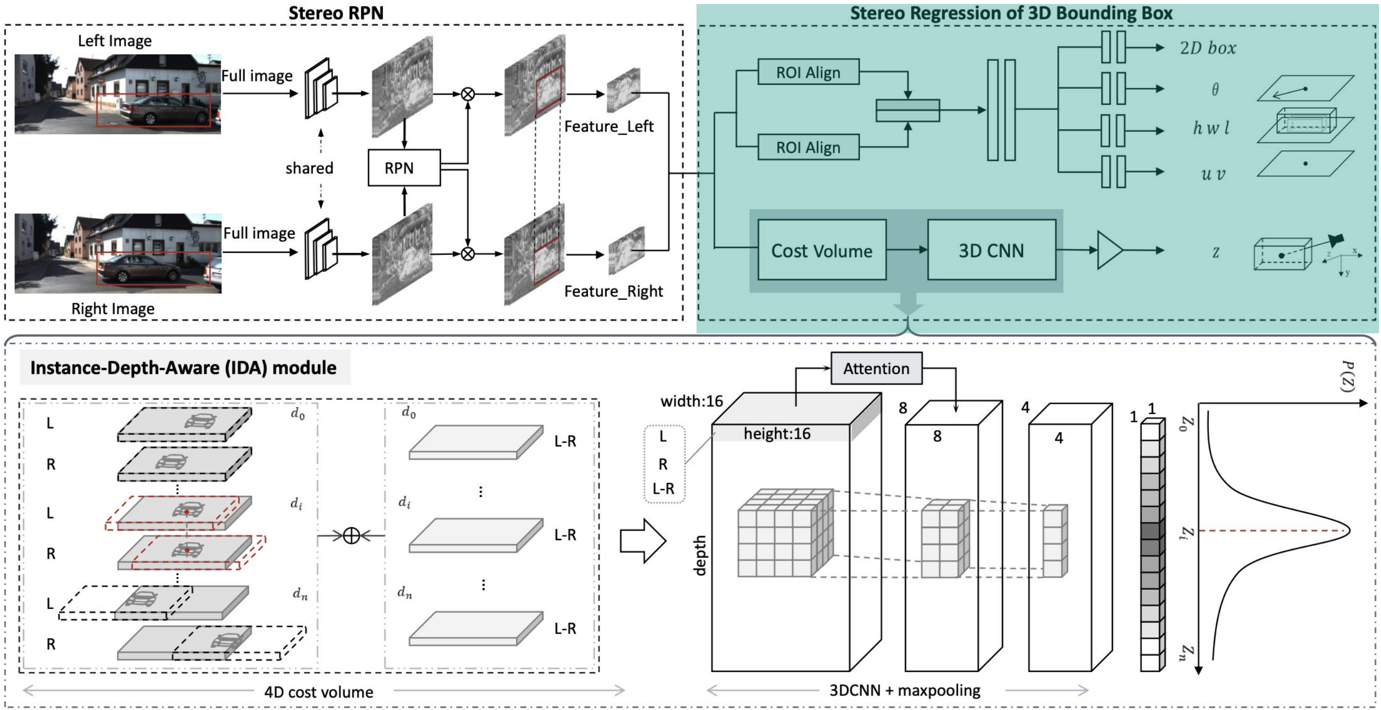
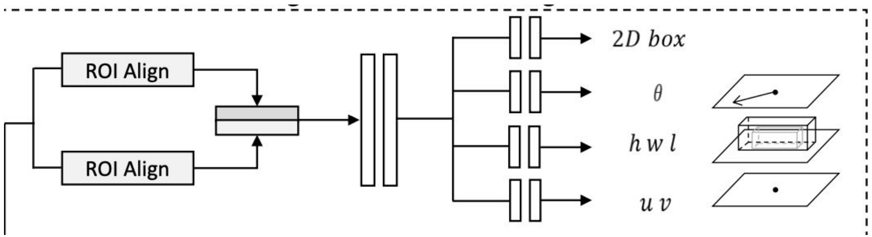


Figure 1. Overview of the proposed IDA-3D. Top: Stereo RPN takes a pair of left and right images as input and outputs corresponding left-right proposal pairs. After stereo RPN, we predict position, dimensions and orientation of 3D bounding box. Bottom: Instance-depth-aware module builds a 4D cost volume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

Stereo Regression of 3D Bounding Box

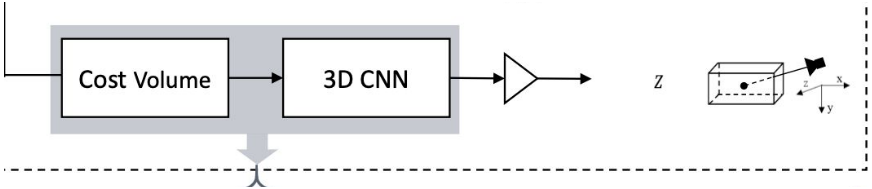
Dimension, Position, Orientation

- **Position estimation (x,y)** -- horizontal and vertical coordinate of the object center
- **Orientation estimation (θ)**-- view point angle
- **Dimension estimation (2D box)** -- object stereo bounding boxes



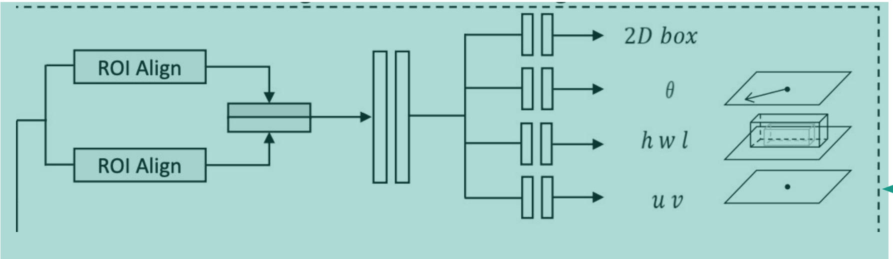
IDA-3D Module - Depth

- **Depth estimation (z)** -- IDA-3D Module



Stereo Rregression of 3D Bounding Box

- ❑ Depth estimation (z)
- ❑ Position estimation (x,y)
- ❑ Orientation estimation
- ❑ Dimension estimation

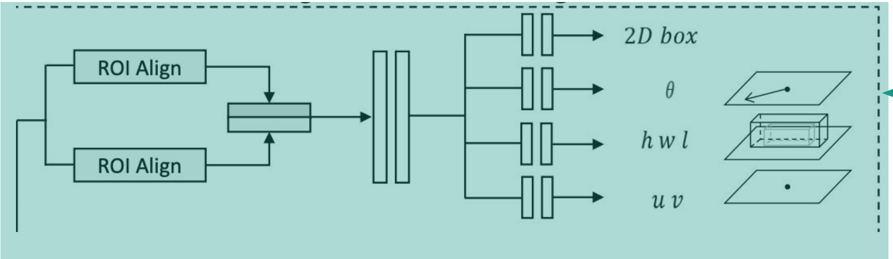


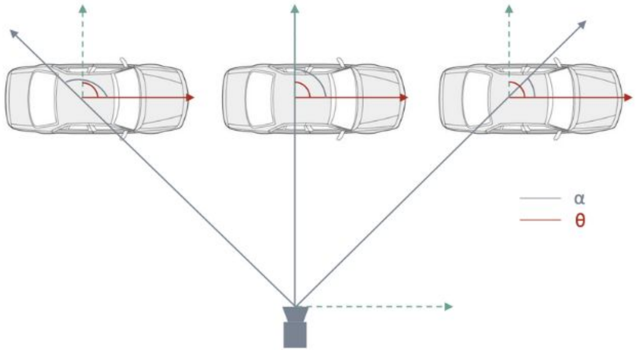
$$x = \frac{(u - c_u) \times z}{f_u} \qquad y = \frac{(v - c_u) \times z}{f_v}$$

(c_u, c_v) : camera center
f_u, f_v : horizontal and vertical focal length

Stereo Regression of 3D Bounding Box

- ❑ Depth estimation (z)
- ❑ Position estimation (x,y)
- ❑ **Orientation estimation**
- ❑ Dimension estimation





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

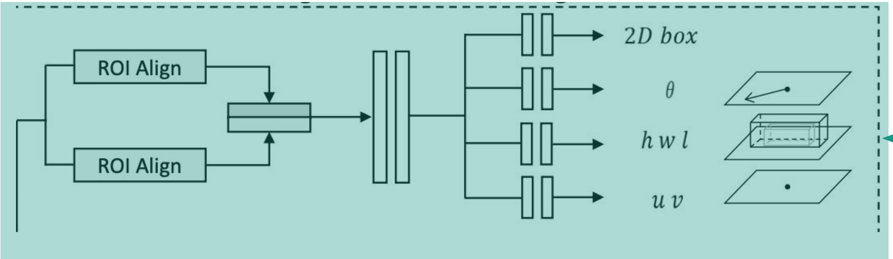
θ : orientation angle

x, z : horizontal position, depth

Figure 5. Relation between object orientation θ and the viewpoint angle α.

Stereo Regression of 3D Bounding Box

- ❏ Depth estimation (z)
- ❏ Position estimation (x,y)
- ❏ Orientation estimation
- ❏ Dimension estimation

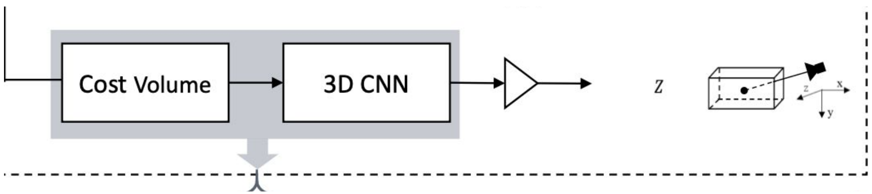


produce dimension offsets $(\Delta h, \Delta w, \Delta l)$ to the mean class $(\bar{h}, \bar{w}, \bar{l})$

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

Stereo Regression of 3D Bounding Box

- Depth estimation (z)
- Position estimation (x,y)
- Orientation estimation
- Dimension estimation



IDA-3D

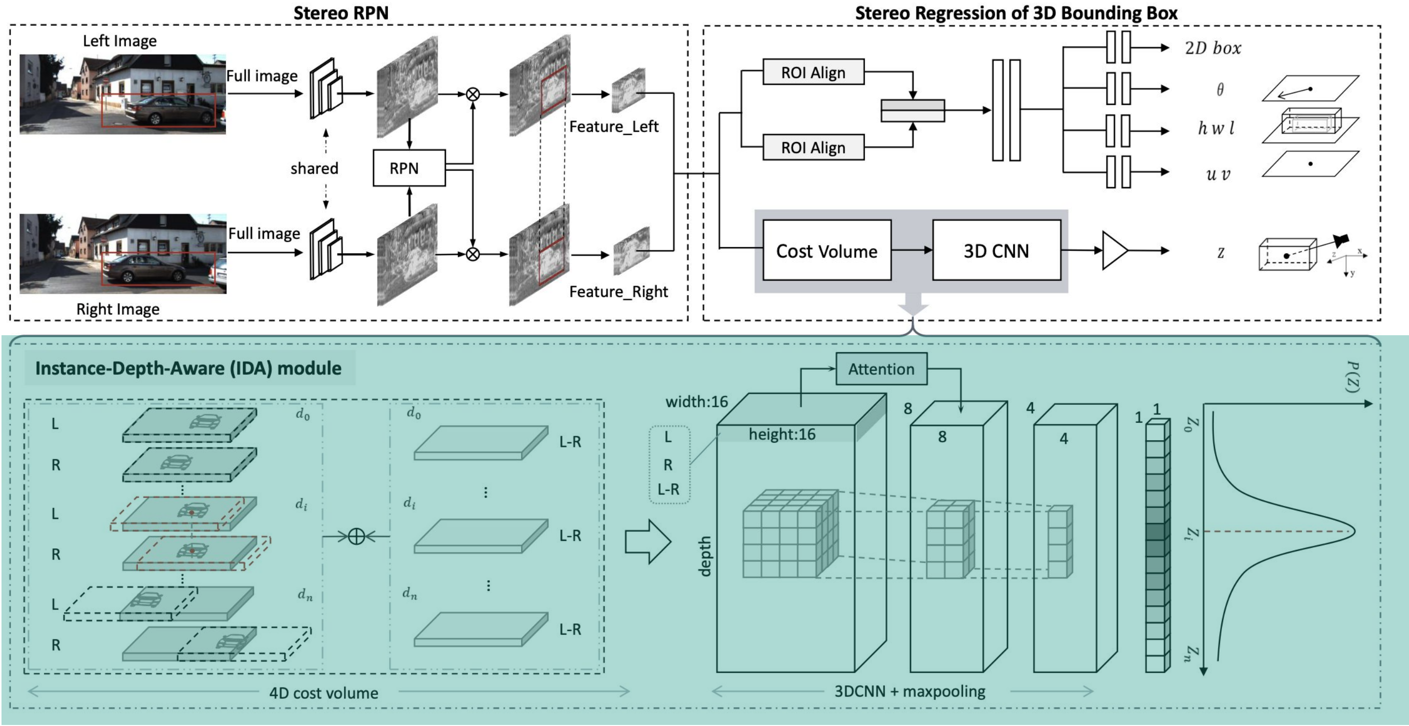
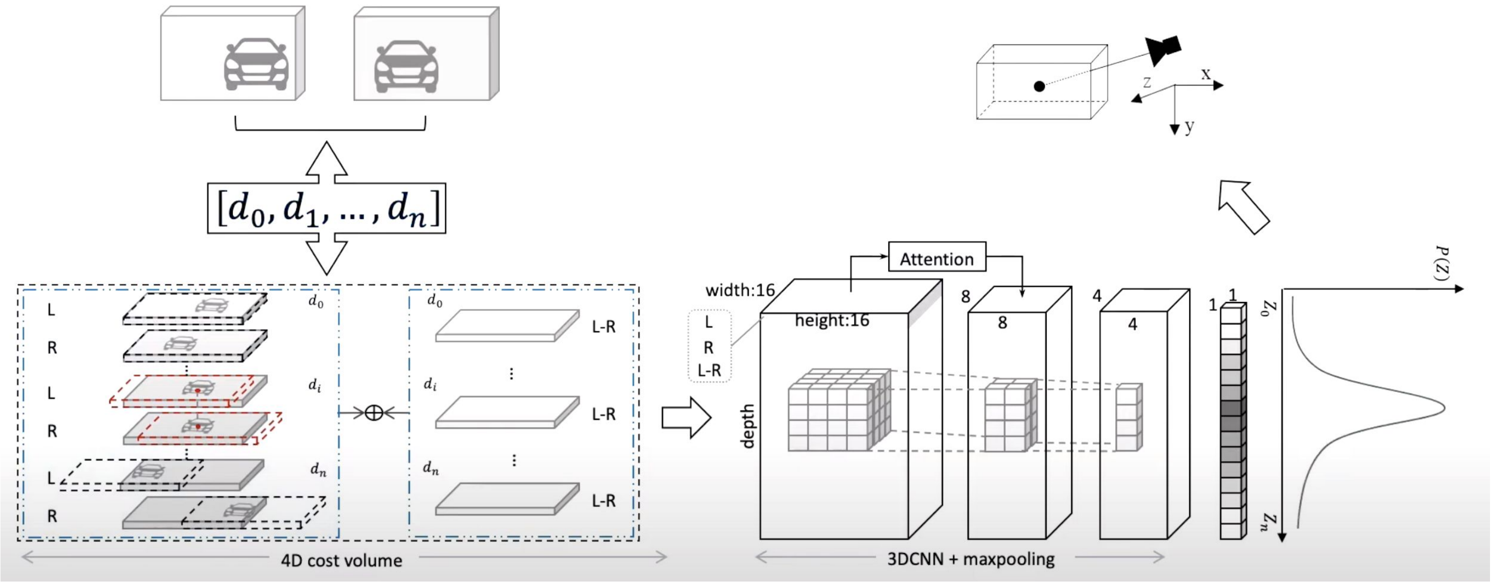


Figure 1. Overview of the proposed IDA-3D. Top: Stereo RPN takes a pair of left and right images as input and outputs corresponding left-right proposal pairs. After stereo RPN, we predict position, dimensions and orientation of 3D bounding box. Bottom: Instance-depth-aware module builds a 4D cost volume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

IDA-3D Module

Input: Left ROI Right ROI

Output: Depth of the 3D bounding box's center



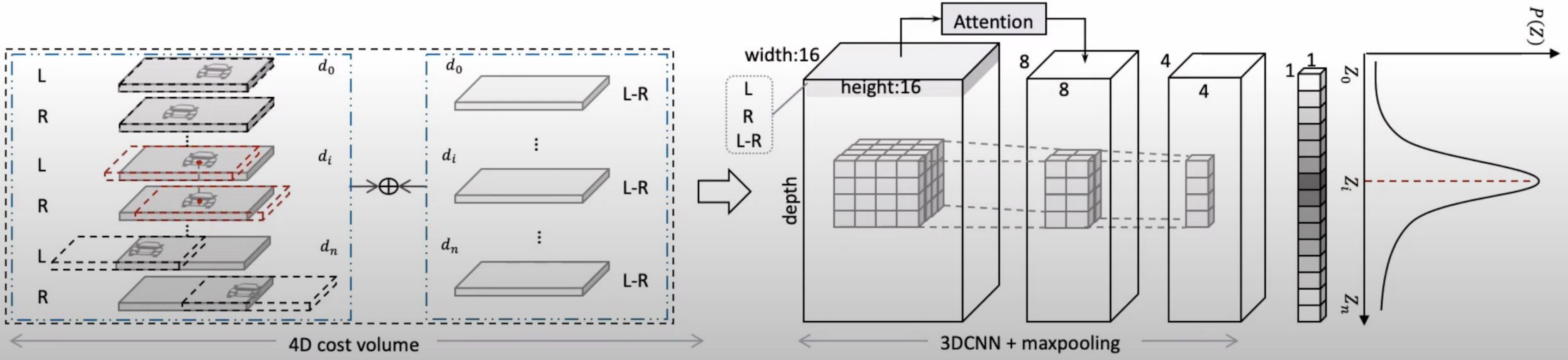
[1] Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." Computer8 (2009): 30-37.
[2] He, Xiangnan, et al. "Neural collaborative filtering." *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017.

IDA-3D Module: depth estimation

Instance depth awareness: measuring the correspondence of the same instance between two images and paying attention to the global spatial information of the object

Disparity Adaptation Strategy: changing the disparity level in cost volume from uniform quantization to non-uniform quantization

Match Cost Reweighting: penalizing the depth levels that are not unique for an object instance and promoting the depth level that have high probabilities



Instance Disparity (depth) estimation

conv0-maxpool0-conv1-maxpool1

- learn and perform downsampling 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

conv2-avgpool

- down sampled features by 3D CNN are finally merged into depth probability of the 3D box center

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

N = number of depth levels

P_i = normalized probability

cost volume of dimensionality:
disparity × height × width × feature size

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

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

Instance Disparity (depth) estimation

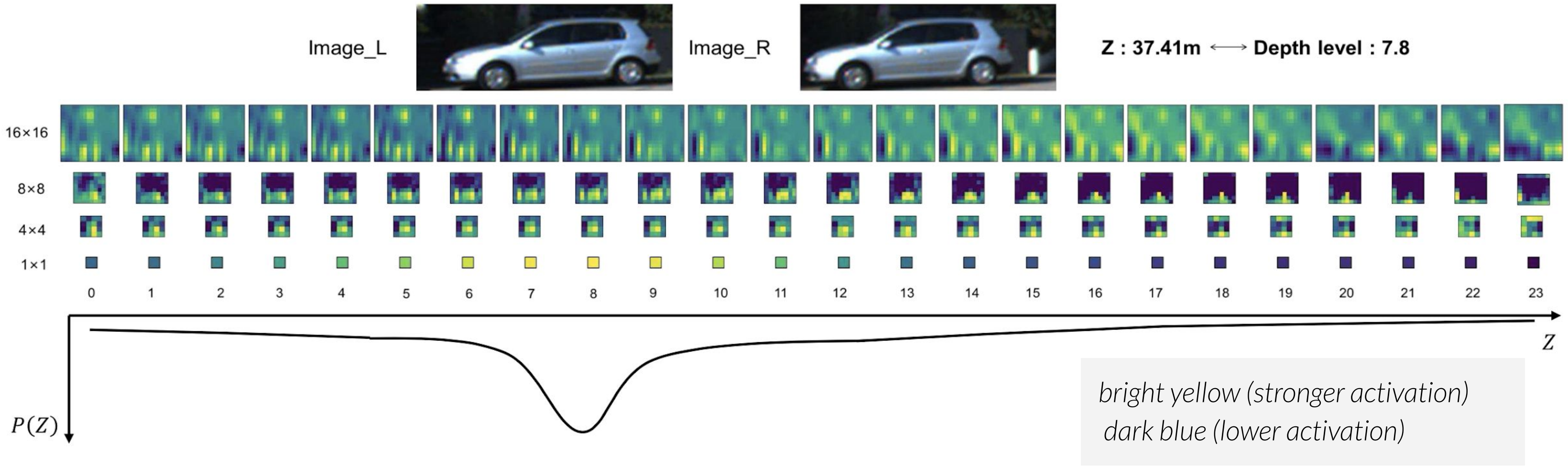


Figure 3. Global spatial information extraction process. Feature maps are sampled at a channel and sorted by the depth level. The bright yellow color in the feature map indicates stronger activation, while dark blue indicates the lower activation.

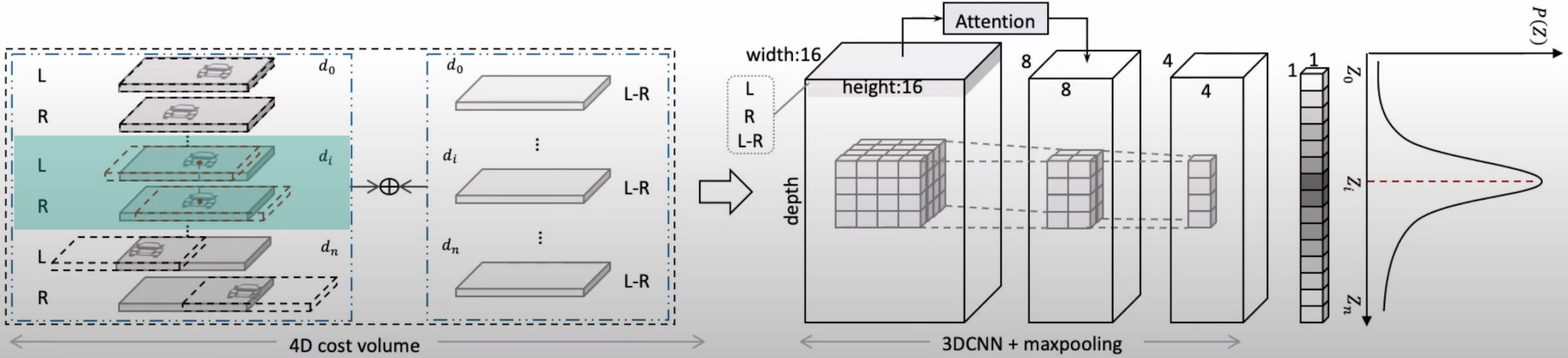
Note from figure: feature maps are gradually changed from low-level features to high-level global features of its center depth probability

IDA-3D Module: depth estimation

Instance depth awareness: measuring the correspondence of the same instance between two images and paying attention to the global spatial information of the object

Disparity Adaptation Strategy: changing the disparity level in cost volume from uniform quantization to non-uniform quantization

Match Cost Reweighting: penalizing the depth levels that are not unique for an object instance and promoting the depth level that have high probabilities



IDA-3D Module: depth estimation

Disparity Adaptation Strategy: changing the disparity level in cost volume from uniform quantization to non-uniform quantization

- Previous works: optimize the accuracy of disparity estimation
- Error in depth increases quadratically with distance — influence of error of further objects is larger, leading to poor 3D object detection
- Disparity level in cost volume: uniform to non-uniform

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

f_u : **horizontal focal length,**

b : **baseline of binocular camera**

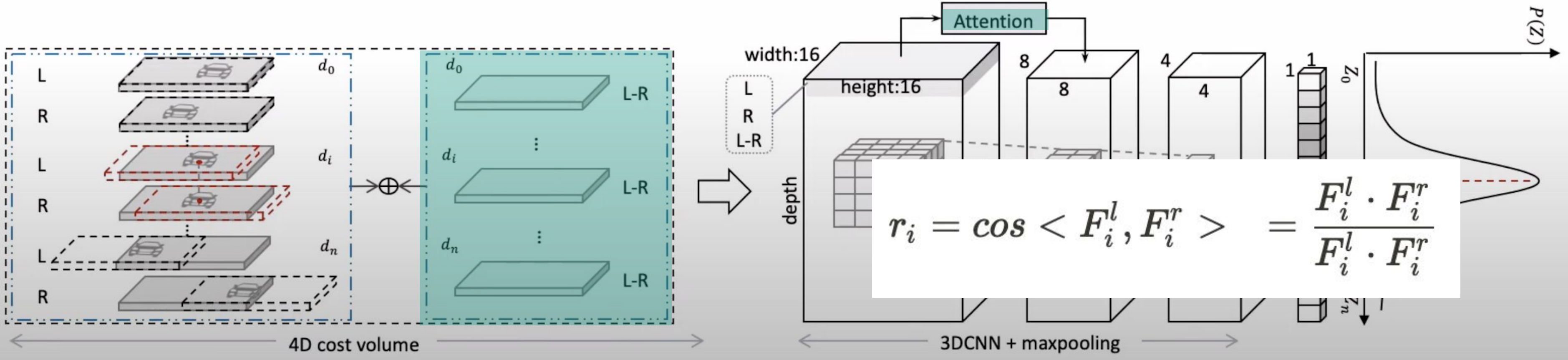
- calculate range according to the width of the union box of the image $[z_{min}, z_{max}]$, minimum and maximum depth values of each object respectively
    - use intrinsic camera parameters
    - minimizes the average partition cell quantization for a fixed number of disparity levels

IDA-3D Module: depth estimation

Instance depth awareness: measuring the correspondence of the same instance between two images and paying attention to the global spatial information of the object

Disparity Adaptation Strategy: changing the disparity level in cost volume from uniform quantization to non-uniform quantization

Match Cost Reweighting: penalizing the depth levels that are not unique for an object instance and promoting the depth level that have high probabilities



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

Loss Function

Whole multi-task loss as formulated is:

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

L_{rpn}, L_{2Dbox} : loss of 2D boxes on stereo RPN module, stereo regression module
$L_{3D}^{(u,v)}$: loss of projection of object instance centers
L_{3D}^{z} : loss of instance depth of objects
L_{dim} : offset regression loss for the 3D box dimension
L_{α} : orientation loss - classification loss for discrete angle bins and angle bin offsets

Implementation

- Feature extractor: ResNet50 + FPN
- Training data: flip images in training set, exchange the left and right image, mirror 2D boxes annotation, viewpoint angle and 2D projection of centroid (data augmentation)
- IDA Module: divide depth between z_{max} , z_{min} into 24 levels
- During inference: use 2D boxes obtained from 2D regression as input to IDA module
- Optimizer: SGD with initial learning rate 0.02, momentum 0.9, weight decay 0.0005
- Batch size: 4, 80 000 iterations, 26 Hrs
- Computer: two (2) NVIDIA 2080Ti GPUs
- KITTI 3D object detection dataset: 7481 training images, 7581 testing images – 3712 and 3769 images respectively

Results: overall

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.

Results: effects of disparity quantization strategy

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**

Results: effects of disparity quantization strategy

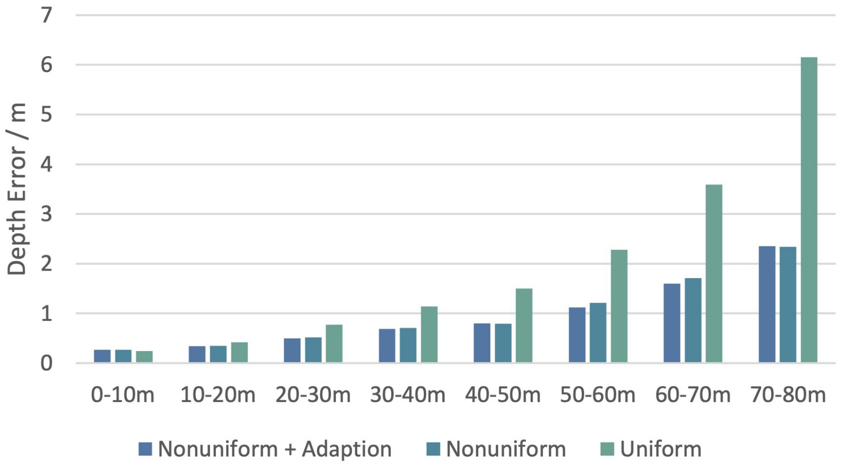


Figure 6. The depth estimation error from different disparity quantization strategies.

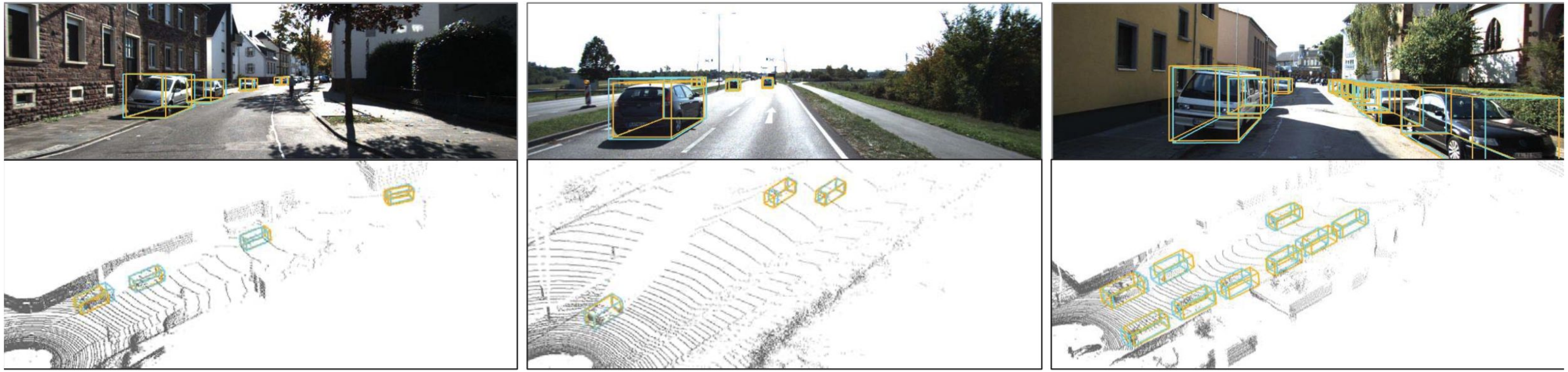
Results: effects of matching cost reweighting

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.

AP_{bev} : average precision (birds-eye view)
AP_{3D} : average precision 3D

Results: sample images



(a) (b) (c)

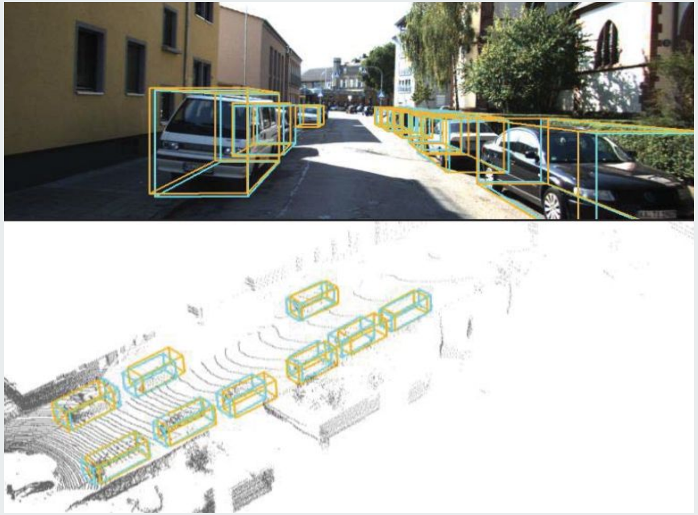
Figure 7. Quantitative results on several scenes in KITTI dataset. At the first row are the ground truth 3D boxes and the predicted 3D boxes projected to the image plane. We also show the detection results on point cloud in order to facilitate observation. The predicted results are shown in yellow and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

Contributions

- Proposed a stereo-based end-to-end learning framework for 3D detection that does not rely on depth images either as input or for training and does not require multi-stage post-processing algorithms
- 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 the 3D object detection
- Provide detailed experiments on the KITTI 3D dataset and achieve state-of-the-art performance compared with the stereo-based methods without depth map supervision

IDA-3D:

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



Wanli Peng, Hao Pan, He Liu, Yi Sun

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020

Presented by:
Maria Isabel Saludares
7 October 2020