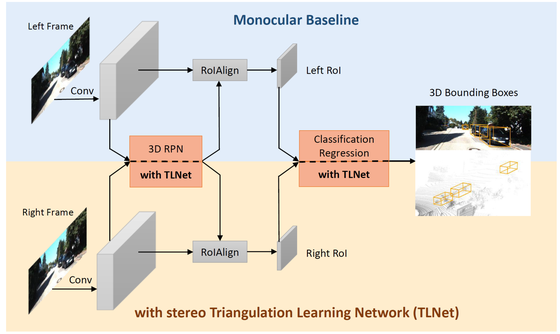
Triangulation Learning Network: from Monocular to Stereo 3D Object Detection

Zengyi Qin (THU) Jinglu Wang (MSRA) Yan Lu (MSRA)

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Overview

We propose the stereo **T**riangulation **L**earning **Net**work (TLNet) for 3D object detection from stereo images, which is free of computing pixel-level depth maps and can be easily integrated into the baseline monocular detector.

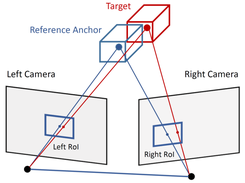


Our contributions are three-fold:

- A solid monocular baseline 3D detector with comparable performance to its state-of-the-art stereo counterpart.
- A triangulation learning network that leverages the geometric correlations of stereo images to localize targeted 3D objects, which outperforms the baseline model by a significant margin.
- A feature reweighting strategy that enhances informative channels of view-specific RoI features, benefiting triangulation learning by biasing the network attention towards the key parts of an object.

Motivation

The key idea is to use a 3D anchor box to explicitly construct object-level geometric correspondences of its two projections on a pair of stereo images, from which the network learns to triangulate a targeted object near the anchor.

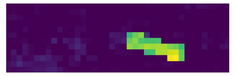


Anchor triangulation. By projecting the 3D anchor box to stereo images, we obtain a pair of RoIs. The left RoI establishes a geometric correspondence with the right one via the anchor box. Our network takes the RoI pair as input and utilizes the 3D anchor as reference to localize the targeted object.

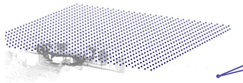
Approach

Baseline Monocular Network. The baseline network taking a monocular image as the input is composed of a backbone and three subsequent modules, i.e., the front view anchor generation, the 3D box proposal and refinement.

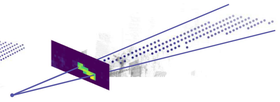
(a) Input image



(b) 2D objectness confidence map



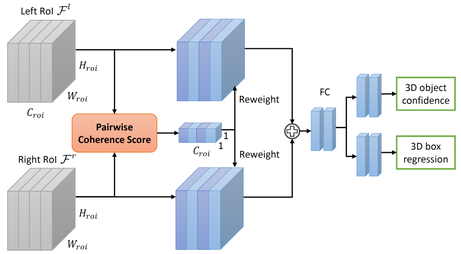
(c) 3D anchor candidates



(d) Potential anchors

Front view anchor generation. Potential anchors are of high objectiveness in the front view. Only the potential anchors are fed into RPN to reduce searching space and save computational cost.

Triangulation Learning Network. The stereo 3D detection is performed by integrating a triangulation learning network into the baseline model. the mechanism of the TLNet focuses on object-level triangulation rather than pixel-level.



Coherence score.

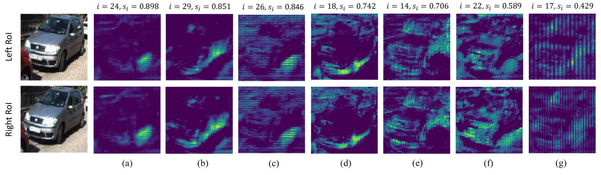
$$s_i = cos < \mathcal{F}_i^l, \mathcal{F}_i^r >= \frac{\mathcal{F}_i^l \cdot \mathcal{F}_i^r}{||\mathcal{F}_i^l|| \cdot ||\mathcal{F}_i^r||}$$

Channel reweighting.

$$\mathcal{F}_i^{l,re} = s_i \mathcal{F}_i^l, \quad \mathcal{F}_i^{r,re} = s_i \mathcal{F}_i^r$$

The TLNet takes as input a pair of left-right RoI features which are obtained using RoIAlign by projecting the same 3D anchor to the left and right frames. The reweighted features are fused by element-wise addition and passed to fully-connected layers to predict 3D bounding box.

The efficient feature reweighting strengthens informative feature channels by left-right coherence, filtering out the signals from noisy and mismatched channels, enabling our network to focus more on the key parts.



Activations of different channels. Coherence score s_i is calculated for channel i. Our objective is enhancing channels like (a) and weaken those like (g) and (f).

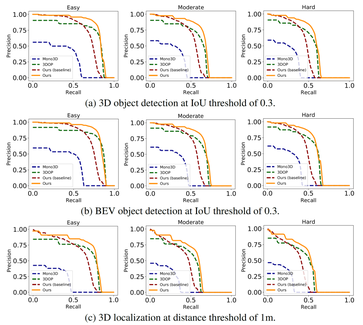
Experiment

Method	Data	AP_{3D}(IoU=0.3)			AP_{3D}(IoU=0.5)			AP_{3D}(IoU=0.7)		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
VeloFCN	LiDAR	/	/	/	67.92	57.57	52.56	15.20	13.66	15.98
Mono3D	Mono	28.29	23.21	19.49	25.19	18.20	15.22	2.53	2.31	2.31
MF3D	Mono	/	/	/	47.88	29.48	26.44	10.53	5.69	5.39
MonoGRNet	Mono	72.17	59.57	46.08	50.51	36.97	30.82	13.88	10.19	7.62
3DOP	Stereo	69.79	52.22	49.64	46.04	34.63	30.09	6.55	5.07	4.10
Ours (baseline)	Mono	72.91	55.72	49.19	48.34	33.98	28.67	13.77	9.72	9.29
Ours	Stereo	**78.26**	**63.36**	**57.10**	**59.51**	**43.71**	**37.99**	**18.15**	**14.26**	**13.72**

Table 1. 3D detection performance.

Method	Data	AP_{BEV}(IoU=0.3)			AP_{BEV}(IoU=0.5)			AP_{BEV}(IoU=0.7)		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
VeloFCN	LiDAR	/	/	/	79.68	63.82	62.80	40.14	32.08	30.47
Mono3D	Mono	32.76	25.15	23.65	30.50	22.39	19.16	5.22	5.19	4.13
MF3D	Mono	/	/	/	55.02	36.73	31.27	22.03	13.63	11.60
MonoGRNet	Mono	73.10	60.66	46.86	54.21	39.69	33.06	24.97	19.44	16.30
3DOP	Stereo	71.41	57.78	51.91	55.04	41.25	34.55	12.63	9.49	7.59
Ours (baseline)	Mono	74.18	57.04	50.17	52.72	37.22	32.16	21.91	15.72	14.32
Ours	Stereo	**81.11**	**65.25**	**58.15**	**62.46**	**45.99**	**41.92**	**29.22**	**21.88**	**18.83**

Table 2. BEV detection performance.



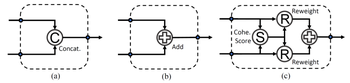
(a) 3D object detection at IoU threshold of 0.3.

(b) BEV object detection at IoU threshold of 0.3.

(c) 3D localization at distance threshold of 1m.

Recall-precision curves.

Ablation study for feature fusion

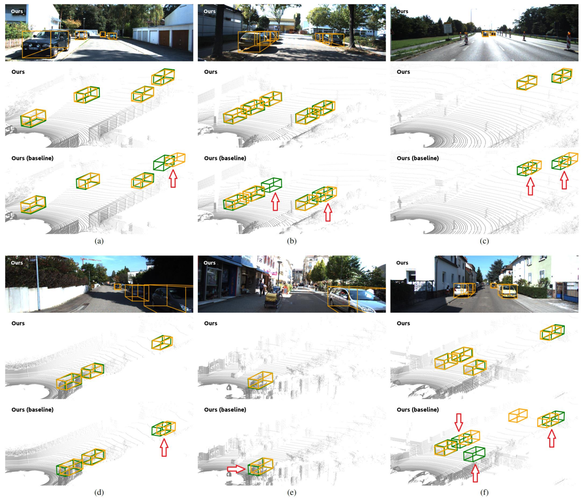


Feature fusion. (a) Concatenation (b) direct addition (c) the proposed strategy.

IoU	Method	AP_{3D}		
		Easy	Moderate	Hard
0.3	Concat.	76.87	60.81	54.70
	Add	75.74	59.89	53.57
	Reweight	**78.26**	**63.36**	**57.10**
0.5	Concat.	56.32	41.20	36.41
	Add	53.41	41.61	36.37
	Reweight	**59.51**	**43.71**	**37.99**
0.7	Concat.	13.97	11.63	9.67
	Add	16.60	13.59	11.20
	Reweight	**18.15**	**14.26**	**13.72**

Effect of reweighting on AP_{3D}.

Qualitative results for person.



Qualitative comparison. Orange bounding boxes are detection results, while the green boxes are ground truths. The lidar point clouds are visualized for reference but not used in both training and evaluation. The triangulation learning method can reduce missed detections and improve the performance of depth prediction at distant regions.