LED: LOCALIZATION-QUALITY ESTIMATION EMBEDDED DETECTOR

Shiquan Zhang, Xu Zhao, Liangji Fang, Haiping Fei, Haitao Song

Computer Vision Lab, SJTU, 2018.10

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Contents

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

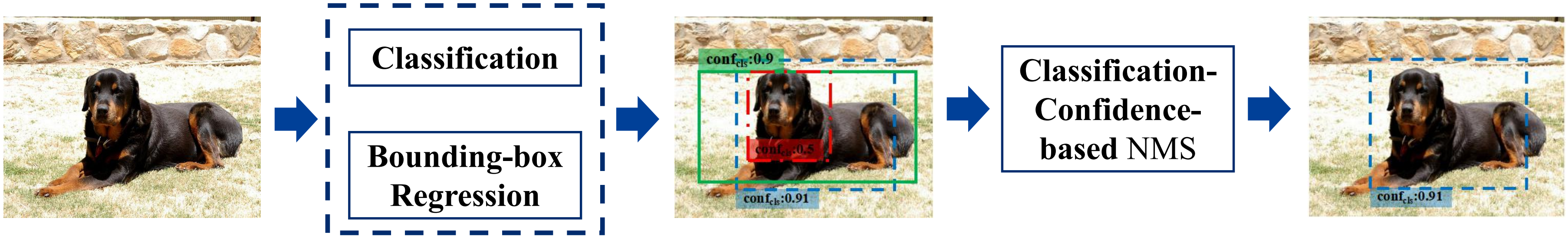
Motivation

Mainstream Object Detection Methods

- **Two-stage detectors** (Faster-RCNN R-FCN FPN etc.)
- **One-stage detectors** (SSD YOLO RetinaNet etc.)

Input Image

Raw Detections

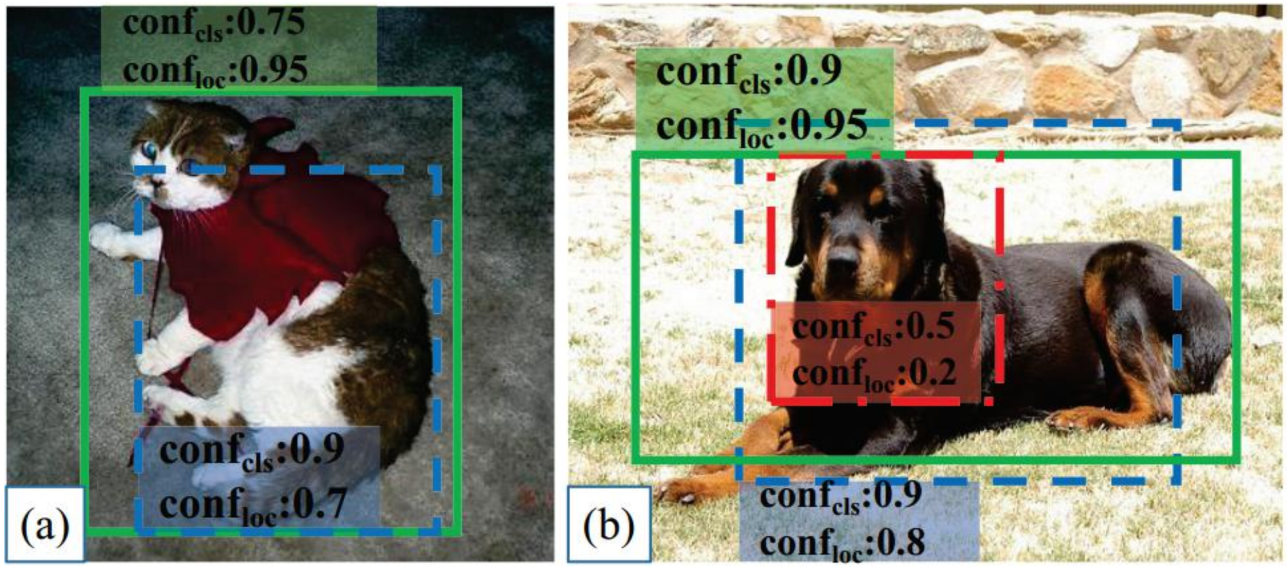


Detection Network

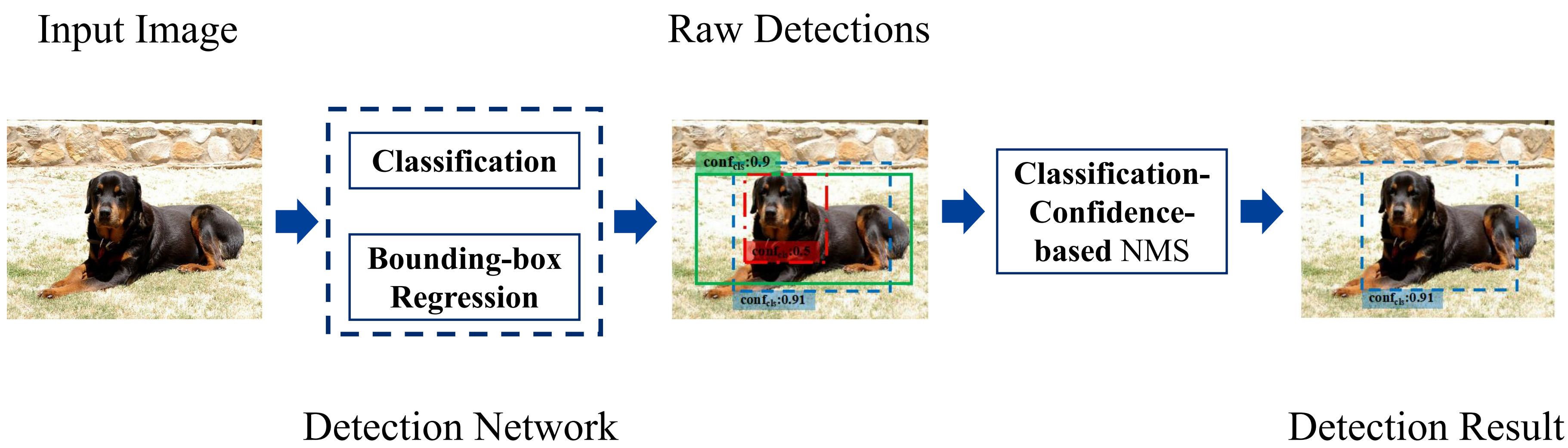
Detection Result

Contradiction

- Some **better** localized detections do not correspond to **higher** classification confidences
- Classification confidences can not fully reflect the **localization-quality (loc-quality)** of each detection



Contradiction

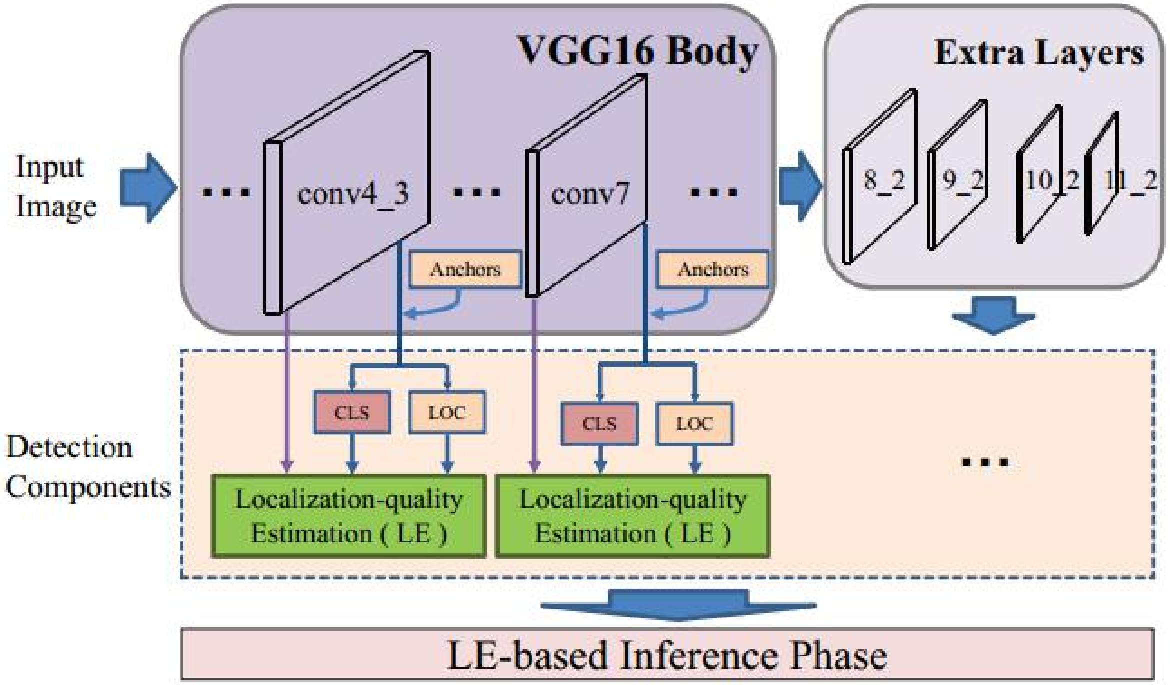


Classification subnet → Translation-invariant

Bounding-box Regression subnet → Translation-sensitive

Overall Architecture

Framework of LED



CLS **Classification branch**

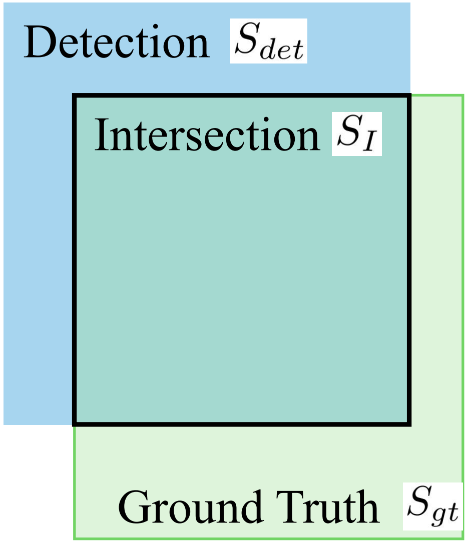
LOC **BBox Regression branch**

- For efficiency, LED is designed as an **one-stage** network.
- Following SSD, anchors are empirically set on each selected layer with **multiple sizes** based on the receptive field, and with **multiple aspect ratios**.

Localization-quanlity Estimation (LE)

Loc-quality Estimation (LE)

- **Model**



- We model the loc-quality of a detection by several spatial cues

overall-quality $$IoU = \frac{S_I}{S_{det}+S_{gt}-S_I}$$

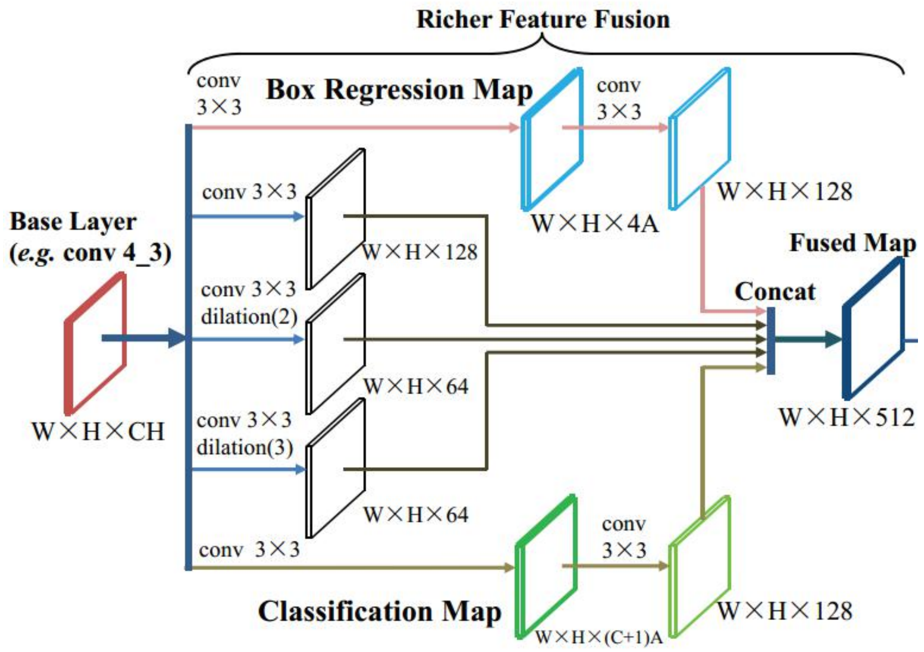
objectiveness-quality $$IoD = \frac{S_I}{S_{det}}$$

completeness-quality $$IoG = \frac{S_I}{S_{gt}}$$

We denote set $V=\{IoD,\ IoG,\ IoU\}$ of each detection

Loc-quality Estimation (LE)

- **Richer Features**



- Features from **classification subnet** and **box regression** subnet are exploited.
- **Dilated convolution** is adopted to encode **context** information

Loc-quality Estimation (LE)

- **Prediction Module**

We intend to predict the value of **each element** in set $V = \{IoD,\ IoG,\ IoU\}$ for each detection

Coarse-to-fine (C2F) prediction module:

Coarse procedure:

Prediction is regarded as a **classification** problem, The value range 0-1 is discretized into four ranges, {0-0.1, 0.1-0.4, 0.4-0.7, 0.7-1.0}, referred as the **background value range, the low value range, the middle value range and the high value range** respectively

Fine procedure:

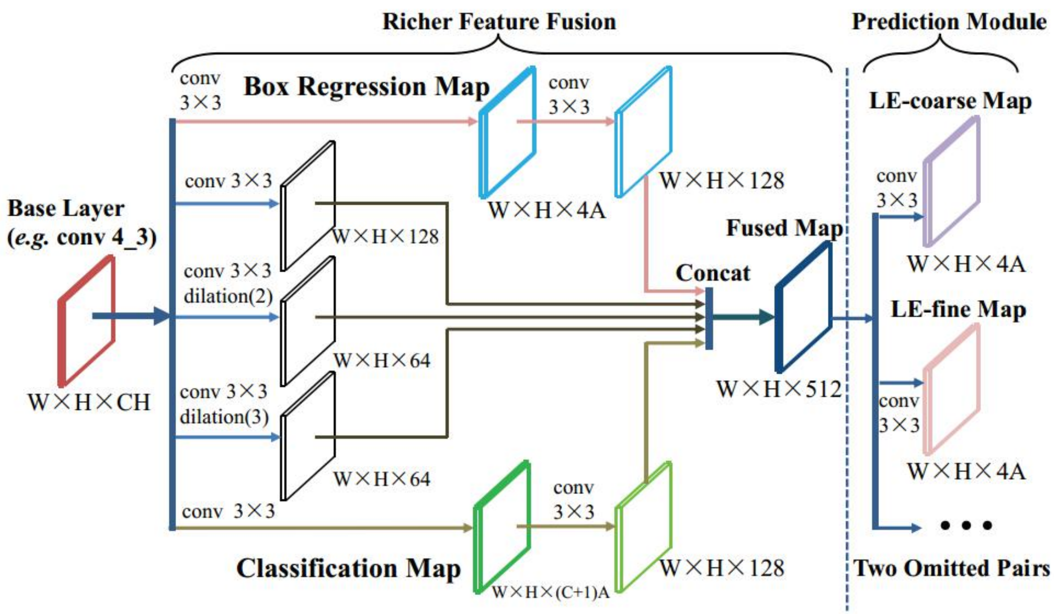
Four independent regressors correspond to the four value ranges respectively, regress continuous values relative to "anchors" in corresponding value ranges. The "anchors" are set to the median of each value range

Loc-quality Estimation (LE)

- **Prediction Module**

Corresponding to the proposed coarse-to-fine (C2F) prediction module, three pairs of coarse-fine feature maps are parallel built for the three elements in V.

$$V=\{IoD,\ IoG,\ IoU\}$$



For each detection, We obtain set V by:

$$v = \sum_{i=1}^{4}(prob_i \cdot val_i), \forall v\ in\ V$$

where v denotes IoU, IoD, or IoG. $prob_i$ denotes the probability of the i-th value range and val_i denotes the finely regressed value of the i-th value range.

Loc-quality Estimation (LE)

- **LE Loss**

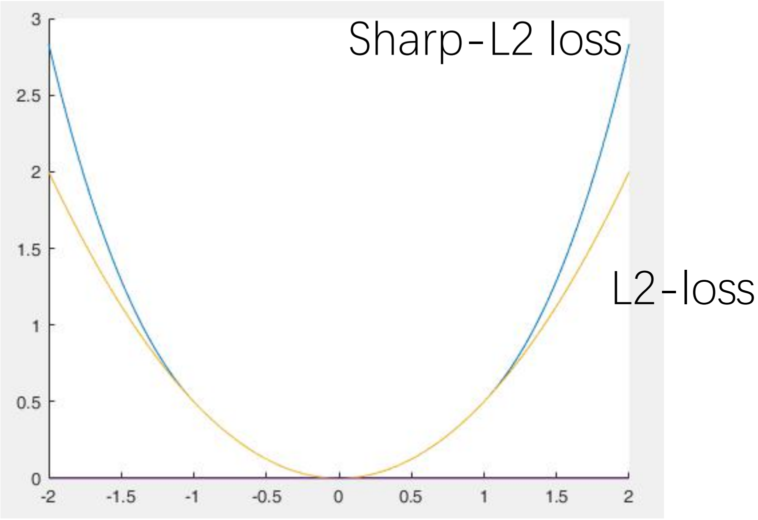
The softmax loss is adopted as the coarse procedure loss L_{coarse}

The **Sharp-L2 loss** is proposed as the fine procedure loss L_{fine}

Each element in V={IoD, IoG, IoU} donates a L_{coarse} and a L_{fine} , thus LE loss L_{LE} is composed of 6 weighted losses from two types.

- The proposed **Sharp-L2 loss**

$$Sharp\text{-}L2(x) = \begin{cases} \frac{1}{2} \cdot x^2 & , |x| < 1 \\ \frac{1}{3} \cdot |x|^3 + \frac{1}{6} & , |x| \geq 1 \end{cases}$$



Embed LE into An One Stage Framework

Loc-quality Estimation Embedded Detector (LED)

- **Training**

Three-step mechanism to optimize LED:

Step 1: Identical to SSD

$$L_1 = L_{cls} + \alpha \cdot L_{reg}$$

Step 2: Freeze all the weights and bias except LE module

$$L_2 = L_{LE}$$

Step 3: Unfreeze all the weights and bias

$$L_3 = L_{cls} + \alpha \cdot L_{reg} + \beta \cdot L_{LE}$$

Loc-quality Estimation Embedded Detector (LED)

- **Training**

Some training strategies are utilized.

- Matching ground truth bounding box with anchors to obtain training samples
- Hard negative mining to balance negative and positive samples for classification and box regression.
- Modified Hard example mining procedure for LE module, based on the L_{LE}
- Data augmentation methods such as expanding, cropping and color distortion to improve the generalization performance

Loc-quality Estimation Embedded Detector (LED)

- **Inference**

In inference phase, we intend to **utilise the estimated loc-quality** (IoD, IoG, IoU)

Based on the definition of IoD, IoG, IoU, We first derive:

$$IoU' = \frac{IoD \cdot IoG}{IoD + IoG - IoD \cdot IoG}$$

Then we obtain the loclization confidence:

$$conf_{loc} = \lambda \cdot IoU + (1 - \lambda) \cdot IoU'$$

The overall confidence which integrates both classification confidence and localization confidence is obtained by gaussian penalty:

$$conf = conf_{cls} \cdot e^{-\frac{(1-conf_{loc})^2}{\sigma}}$$

λ and σ are set to 0.6 and 1 respectively

Finally, **NMS** is applied based on the overall confidence of each detection.

Experimental Results

Experimental Results

- **Pascal VOC 2007 test results**

PASCAL VOC 2007 test results. **All methods are based on pre-trained VGG16**, and trained with **VOC 2007 trainval and VOC 2012 trainval**. ⋆ indicates our own reproducing of SSD300, slightly higher than the original one. With *Caffe , on a single NVIDIA Titan X (Pascal) GPU*

Approach	FPS	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster R-CNN [4]	–	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
RON384 [18]	–	75.4	78.0	82.4	76.7	67.1	**56.9**	85.3	84.3	86.1	55.5	80.6	71.4	84.7	84.8	82.4	76.2	47.9	75.3	74.1	83.8	74.5
SSD300⋆	**94**	77.6	79.2	84.0	76.1	69.5	50.6	86.9	85.9	88.7	60.4	81.3	**76.8**	86.2	87.4	83.6	79.4	**52.9**	79.2	79.6	87.6	**77.1**
LED300	65	**78.7**	**82.7**	**86.5**	**76.9**	**71.7**	51.7	**87.1**	**88.0**	**89.9**	**60.8**	**84.0**	74.9	**88.2**	**87.9**	**85.1**	**81.3**	52.5	**79.5**	**80.8**	**87.6**	76.8

Experimental Results

- **Ablation studies on Pascal VOC 2007 dataset**

Ablation studies on Pascal VOC 2007. p denotes the setting of corresponding column is employed. Otherwise, base prediction feature map instead of richer features (RF), direct regression instead of coarse-to-fine (C2F), L2 loss instead of Sharp-L2 loss, LE-Product instead of LE-Gaussian. (**Evaluation IoU threshold is set to 0.5**)
With *Caffe , on a single NVIDIA Titan X (Pascal) GPU*

Model	RF	C2F	Sharp-L2	LE-Gaussian	mAP
LED300					77.4
	√				77.9
	√	√			78.3
	√	√	√		78.5
	√	√	√	√	**78.7**
SSD300_{240k}					77.7

Experimental Results

- **KITTI car detection results on validation subset.**

All methods share the same dataset splits.
⋆ indicates that the detection results and inference time are obtained from corresponding references, otherwise from our experiments. Time indicates mean inference time for one image. Mod denotes moderate difficulty and is the metric for ranking.

Approach	Time	Easy	**Mod**	Hard
3DVP [24]⋆	40s	80.48	68.05	57.20
Faster R-CNN [4]⋆	2s	82.91	77.83	66.25
SubCNN [22]⋆	2s	95.77	86.64	74.07
DeepMANTA (GoogLenet) [23]⋆	0.7s	**97.90**	91.01	**83.14**
DeepMANTA (VGG16) [23]⋆	2s	97.45	91.47	81.79
SSD	0.07s	96.50	88.11	77.52
LED (single)	0.11s	97.31	91.32	81.23
LED (ensemble)	0.33s	97.51	**91.93**	83.11

Reference

- [1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in NIPS, 2015, pp. 91–99.
- [2] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in ECCV. Springer, 2016, pp. 354–370
- [3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, “You only look once: Unified, real-time object detection,” in CVPR, 2016, pp. 779–788.
- [4] Joseph Redmon and Ali Farhadi, “Yolo9000: Better,faster, stronger,” in CVPR, July 2017
- [5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, “Ssd: Single shot multibox detector,” in ECCV,2016.
- [6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar, “Focal loss for dense object detection,” in ICCV, Oct 2017.
- [7] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, “R-fcn:Object detection via region-based fully convolutional networks,” in NIPS, 2016, pp. 379–387.
- [8] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser, “Dilated residual networks,” in CVPR, 2017, vol. 1.
- [9] Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Ming Lu, and Yurong Chen, “Ron: Reverse connection with objectness prior networks for object detection,” in CVPR, 2017, vol. 1, p. 2.
- [10] M Everingham, L Van Gool, CKI Williams, J Winn, and A Zisserman, “The pascal visual object classes challenge 2007 (voc 2007) results (2007),” 2008.

Reference

- [11] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in CVPR, 2012, pp. 3354–3361.
- [12] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Celine Teuliere, and Thierry Chateau, "Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image," in CVPR, July 2017.
- [13] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese, "Data-driven 3d voxel patterns for object category recognition," in CVPR, 2015, pp. 1903–1911.
- [14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 675–678.
- [15] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in CVPR, July 2017.

Thanks!