Efficient and Low-Footprint Object Classification using Spatial Contrast

MATTHEW BELDING, Electrical and Computer Engineering, University of Pittsburgh, USA
DANIEL C. STUMPP, NSF-SHREC Center, Electrical and Computer Engineering, University of Pittsburgh, USA
RAJKUMAR KUBENDRAN, Electrical and Computer Engineering, University of Pittsburgh, USA

ACM Reference Format:
Matthew Belding, Daniel C. Stumpp, and Rajkumar Kubendran. 2023. Efficient and Low-Footprint Object Classification using Spatial Contrast. In . ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

This document contains supplemental information for object detection using spatial contrast. It is organized as follows:

- Overview of Argoverse-HD dataset
- Network architectures for detection
- Setup for training
- Performance results of localization/detection

2 OVERVIEW OF ARGOVERSE-HD DATASET

In the main paper, we evaluated the traffic sign classification performance on a simulated output of a neuromorphic spatial contrast (SC) sensor. This was done across multiple thresholds with two different SC techniques: absolute and relative. Here, we evaluated the localization and detection on the Argoverse-HD dataset to further investigate the efficacy of SC in autonomous driving. We chose Argoverse-HD as opposed to the german traffic sign dataset (GTSDB) because only Argoverse-HD dataset provides annotations on traffic signs.

The Argoverse-HD dataset was introduced as part of the 2021 Streaming Perception Challenge at CVPR [5]. It acts as an extension of the video dataset Argoverse 1.1 and contains urban scenes from two U.S cities. The center RGB ring camera was used as the source of video and operated at 30 FPS. As previously described in the methods section of the main paper, traditional RGB images were converted to an SC sensor output representation for comparison. A subset of the MS COCO [7] class set is used for simplicity in evaluating off-the-shelf models. This subset contained 8 classes that are related to autonomous driving application: stop sign, traffic light, car, person, bicycle, motorcycle, bus, and truck. All evaluation is performed on the validation set which contains 15000 total frames and 24 videos.

3 NETWORK ARCHITECTURES FOR DETECTION

To properly evaluate SC in a detection setting, several object detectors were considered that were released over the last few years. These detectors include DDOD [3], TOOD [4], Cascade R-CNN [1], and RetinaNet [6].

Object detectors can be divided into two categories: one-stage and two-stage architectures. Two-stage architectures separate the localization from classification and generate the proposal first before classifying. The advantage of this method is high detection accuracy but impacts inference speed. Cascade R-CNN is one example of a two-stage network that combines a sequence of detectors with increasing IoU thresholds to prevent overfitting and inference-time quality mismatch [1]. The Cascade R-CNN is shown in Fig. 1.

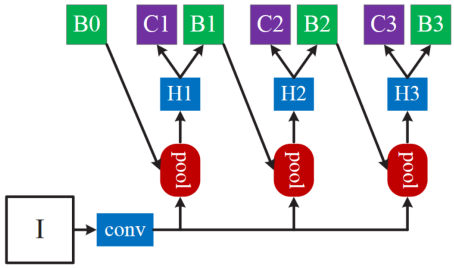


Fig. 1. Cascade R-CNN Architecture [1].

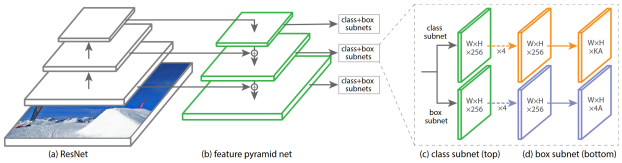


Fig. 2. RetinaNet [6].

One-stage detectors generate class probabilities and location coordinates in a single stage without the need for a separate regional proposal operation. This enables high detection speed at a cost of accuracy typically. Arguably one of the most well-known detectors, RetinaNet introduced the focal loss to address the accuracy difference between one and two-stage detectors. It was determined that extreme class imbalance between the background and foreground was the significant cause [6]. The design for this network is shown in Fig. 2. Another detector, Task-aligned One-stage Object Detection (TOOD), optimizes the parallel classification and localization heads of the one-stage detector by introducing a task-aligned head and learning. This addresses the spatial misalignment that can occur in predictions when optimizing a one-stage detector [4]. Fig. 3 provides a comparison of a traditional one-stage head and the task-aligned head. DDOD or Disentangled Dense Object Detector [3] proposed a training paradigm that sought to decompose conjunctions in label assignment, spatial feature alignment, and pyramid supervision for one-stage detectors. Chen [3] found that disentanglement can improve the performance of several state-of-the-art detectors by more than 2 mAP.

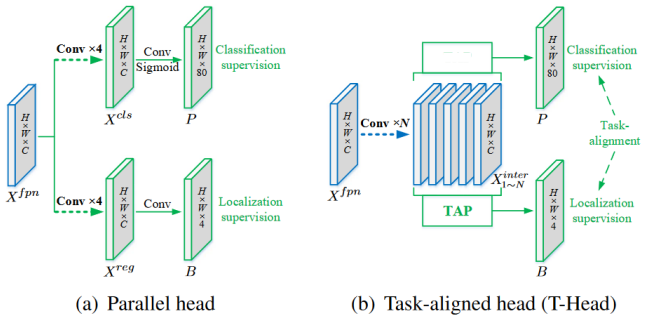
ICONS '23, August 01–03, 2023, Santa Fe, NM

Table 1. Summary of basic parameters for training. Parameter names are kept consistent with the nomenclature of MMdetection.

	Cascade R-CNN [1]	RetinaNet [6]	TOOD [4]	DDOD [3]
Backbone	DetectoRS_ResNet-50	SwinTransformer	ResNet-101	ResNet-50
Neck	RFP	FPN	FPN	FPN
Head	RPNHead + CascadeRoIHead	RetinaHead	TOODHead	DDODHead
Loss BBox	SmoothL1Loss	L1Loss	GIoULoss	GIoULoss
Loss Cls	CrossEntropyLoss	FocalLoss	QualityFocalLoss	FocalLoss
Loss IoU	-	-	-	CrossEntropyLoss
Optimizer-LR	SGD-0.0025	Adam-0.0001	SGD-0.01	SGD-0.01
Resolution	1333×800	1333×800	1333×800	1333×800

Table 2. Comparison of Detector Performance on traditional RGB images against absolute-0.0025 SC. Results show SC is roughly comparable to RGB results but is unable to match the performance of RGB.

Detector	Input	AP	AP_L	AP_M	AP_S	AP_{50}	AP_{75}
Cascade R-CNN [1]	RGB	26.0	47.6	30.8	8.5	43.6	27
	SC	19.6	37.7	22.3	7.5	32.6	19.5
RetinaNet [6]	RGB	23.3	39.4	28.7	8.1	42.6	21.5
	SC	15.6	33.4	19.2	2.1	30.6	13.5
TOOD [4]	RGB	30.4	53.5	32.9	13.2	49.0	30.9
	SC	20.6	43.3	22.0	5.6	34.1	21.1
DOOD [3]	RGB	29.7	51.1	33.5	12.3	48.4	30.1
	SC	23.1	46.8	24.4	8.2	38.3	24.2



(a) Parallel head (b) Task-aligned head (T-Head)

Fig. 3. Task-Aligned One-Stage Object Detection (TOOD) [4]. (a) Traditional one-stage head (b) task-aligned head.

4 SETUP FOR TRAINING

As mentioned in Section 2, Argoverse-HD enables the use the out-of-the-box detectors due to the use of MS COCO format. MMdetection is an open-source toolbox for object detection that supports numerous state-of-the-art frameworks [2]. The networks discussed previously are used with modifications to the heads to accommodate for 8 classes instead of the 80 classes from MS COCO. Due to models being pretrained on MS COCO from the start, very limited fine tuning was performed before finalizing results. A summary of some of the training parameters can be seen for each model in Table 1.

5 PERFORMANCE RESULTS OF LOCALIZATION/DETECTION

In Table 2, localization and detection results on the absolute SC technique under threshold 0.0025 can be seen. This threshold stays within 10 mAP of the baseline RGB, getting as low as within almost 4 AP on large objects under DDOD and 1 AP on small objects with Cascade R-CNN. 0.0025 is a very low threshold and therefore introduces a considerable number of events and noise. Thus it is not surprising under direct comparison to a traditional RGB output, SC is unable to match performance. Upon observation, it's worth noting that recent frameworks (TOOD and DOOD) using SC show nearly identical or even surpass the RGB performance on other networks. In particular, DDOD and TOOD surpass the RGB RetinaNet by more than 3 AP for AP_L when they use SC and is within 1 AP_L on Cascade R-CNN for RGB. Hence, given the growth in performance using more recent frameworks, SC should be able to moderately scale it's performance compared with RGB.

REFERENCES

[1] Zhaowei Cai and Nuno Vasconcelos. 2019. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1. https://doi.org/10.1109/tpami.2019.2956516

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155* (2019).

[3] Zehui Chen, Chenhongyi Yang, Qiaofei Li, Feng Zhao, Zheng-Jun Zha, and Feng Wu. 2021. Disentangle Your Dense Object Detector. In *Proceedings of the 29th ACM*

International Conference on Multimedia. 4939–4948.
[4] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. 2021. TOOD: Task-aligned One-stage Object Detection. In *ICCV*.
[5] Mengtian Li et al. 2020. Towards Streaming Perception. In *ECCV*.
[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*.
[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.