Self-Supervised Pretraining for Object Detection in Autonomous Driving

Aytaç Kanacı* , **Izzeddin Teeti** , **Andrew Bradley** , **Fabio Cuzzolin**
[1]Oxford Brookes University
{aytac.kanaci, iteeti, abradley, fabio.cuzzolin}@brookes.ac.uk,

Abstract

The detection of road agents, such as vehicles and pedestrians are central in autonomous driving. Self-Supervised Learning (SSL) has been proven to be an effective technique for learning discriminative feature representations for image classification, alleviating the need for labels, a remarkable advancement considering how time-consuming and expensive labeling can be in autonomous driving. In this paper, we investigate the effectiveness of contrastive SSL techniques such as BYOL and MOCO on the object (agent) detection task using the ROad event Awareness Dataset (ROAD) and BDD100K benchmarks. Our experiments show that using self-supervised pretraining, we can achieve a 3.96 and 0.78 percentage points improvement on the AP_{50} metric on the ROAD and BDD100K benchmarks for the object detection task compared to supervised pretraining. Extensive comparisons and evaluations of current state-of-the-art SSL methods (namely MOCO, BYOL, SCRL) are conducted and reported for the object detection task.

1 Introduction

Autonomous driving as a subject has been steadily rising in popularity in recent times due to important advances in computer vision and machine learning. As completely autonomous self-driving cars appear imminent, numerous publicly available datasets aimed at evaluating various aspects of the problem have been released [Singh *et al.*, 2022; Sun *et al.*, 2019; Wilson *et al.*, 2021]. Analyzing and making sense of this ever-increasing wealth of data becomes a must both for companies active in this space and for researchers, with a particular focus on public safety.

As cameras are the most commonly used sensors, computer vision plays an essential role in continuously studying new ways of addressing fundamental perception problems relevant to autonomous driving, such as object detection and segmentation. The cost of manually labeling the necessary data, however, seriously limits the range of conditions available for studying in the self-driving context. In this regard, *contrastive self-supervised learning* (SSL) has been successfully shown to be a promising solution, as it achieves a performance comparable to that of supervised learning while having at the same time the ability to mitigate the cost of data labeling, on benchmarks such as ImageNet [Deng *et al.*, 2009] for image classification and COCO [Lin *et al.*, 2014] object detection.

In our work, we provide an extensive analysis of contrastive self-supervised learning methods for object detection in the autonomous driving setting. We show competitive results on object detection accuracy (*e.g.* vehicle, pedestrian...) using state-of-the-art self-supervised pretraining methods such as MOCO [He *et al.*, 2020], BYOL [Grill *et al.*, 2020] and SCRL [Roh *et al.*, 2021] compared to supervised pretraining on ImageNet large-scale benchmark. We test the performance of object detection using ROAD: The ROad event Awareness Dataset [Singh *et al.*, 2022] and BDD100K [Yu *et al.*, 2020] autonomous driving benchmarks. Both datasets provide data with diverse weather and illumination conditions to test robustness of methods.

2 Related Work

2.1 Representation Learning

Supervised and Unsupervised Learning. Discriminative approaches to learning representations learn a representation by directly modeling the conditional distribution $p(y|\mathbf{x})$ with a parametrised model that takes as input the data sample $\mathbf{x}$ and outputs the label variable y. Discriminative modeling consists of an inference step that infers the values of the latent variables $p(\mathbf{v}|\mathbf{x})$, and then directly makes downstream decisions from those inferred variables $p(y|\mathbf{v})$.

Since a self-supervised discriminative model does not have labels corresponding to the inputs like its supervised counterparts, the success of self-supervised methods come from the elegant design of the pretext tasks to generate a pseudo-label $\hat{y}$ from part of the input data itself [Misra and van der Maaten, ; Dosovitskiy *et al.*, 2014; Zhang *et al.*, 2017]

Contrastive Representation Learning. Contrastive representation learning can be considered as learning by comparing. Unlike a discriminative model that learns a mapping

*Contact Author

to some (pseudo-)labels and a generative model that reconstructs input samples, in contrastive learning, a representation is learned by comparing the input samples.

Instead of learning a signal from individual data samples one at a time, contrastive learning *learns by comparing* among similar/dissimilar data samples. Contrastive learning approaches only need to define the similarity distribution in order to sample a positive input $\mathbf{x}^+ \sim p^+(\cdot|\mathbf{x})$, and data distribution for a negative input $\mathbf{x}^- \sim p^-(\cdot|\mathbf{x})$, with respect to an input sample $\mathbf{x}$.

In the self-supervised setting, (*i.e.* contrastive self-supervised learning), instead of deriving a pseudo-label from the pretext task, contrastive learning methods learn a discriminative model on multiple-input pairs, according to some notion of similarity. Methods such as SimCLR provided a basic framework for contrastive SSL using siamese networks. Follow-up work MOCO [He *et al.*, 2020] used an out-of-batch list of negative examples rather than utilizing a large batch size as in SimCLR to learn from samples. BYOL [Grill *et al.*, 2020] was the first method not to require negative samples with a siamese SSL framework.

2.2 Object Detection

Compared to classification, the notion of a negative sample definition needs careful thought in object detection as images can contain multiple subjects. For this reason, recent detection focused methods such as SCRL [Roh *et al.*, 2021] and MultiSiam [Chen *et al.*, 2021] improve upon BYOL with detection specific modifications as it doesn't require negative examples during training.

3 Method

3.1 Contrastive Self Supervised Learning

Contrastive Self-Supervised Learning can be formulated as a dictionary look-up problem [He *et al.*, 2019], where a given reference image $\mathcal{I}$ is augmented into two views, query and key. Tthe query token q should match its designated key k^+ over a set of sampled negative keys $\{k^-\}$ from other images. Generally, the framework can be summarized as the following components: (i) A data augmentation module $\mathcal{T}$ constituting n atomic augmentation operators, such as random cropping, color jittering, and random flipping. We denote a predefined atomic augmentation as a random variable X_i. Each time the atomic augmentation is executed by sampling a specific augmentation parameter from the random variable, i.e., $x_i \sim X_i$. One sampled data augmentation module transforms image $\mathcal{I}$ into a random view $\widetilde{\mathcal{I}}$, denoted as $\widetilde{\mathcal{I}} = \mathcal{T}[x_1, x_2, \ldots, x_n](\mathcal{I})$. Positive pair (q, k^+) is generated by applying two randomly sampled data augmentations on the same reference image. (ii) An encoder network f which extracts the feature $\boldsymbol{v}$ of an image $\mathcal{I}$ by mapping it into a d-dimensional space $\mathbb{R}^d$. (iii) A projection head h which further maps extracted representations into a hyper-spherical (normalized) embedding space. This space is subsequently used for a specific pretext task, i.e., contrastive loss objective for a batch of positive/negative pairs. A common choice is InfoNCE [van den Oord *et al.*, 2018]:

$$\mathcal{L}_q \triangleq -\log \frac{\exp(q{\cdot}k^+/\tau)}{\exp\left(q{\cdot}k^+/\tau\right) + \sum_{k^-} \exp(q{\cdot}k^-/\tau)}, \qquad (1)$$

where τ is a temperature hyper-parameter scaling the distribution of distances.

Bootstrap Your Own Latent

We follow BYOL [Grill *et al.*, 2020] for learning contrastive representations. The online network is appended with a projector g_θ, and a predictor q_θ to obtain latent embeddings. Both g_θ and q_θ are two-layer MLPs. The target network is only appended with the projector g_ξ to avoid trivial solutions. The target network provides the regression target to train the online network while the target network's parameter set ξ follows the online network's parameter set θ, by using an exponential moving average(EMA) with a decay parameter τ, *i.e.* , $\xi \leftarrow \tau\xi + (1-\tau)\theta$.

$$\mathcal{L}_{\theta,\xi} \triangleq \left\|\overline{q_\theta}(z_\theta) - \overline{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi\rangle}{\left\|q_\theta(z_\theta)\right\|_2 \cdot \left\|z'_\xi\right\|_2}. \qquad (2)$$

We symmetrize the loss $\mathcal{L}_{\theta,\xi}$ in 2 by separately feeding v' to the online network and v to the target network to compute $\widetilde{\mathcal{L}}_{\theta,\xi}$.

After training phase, we only keep the encoder f_θ to generate image features; as in [He *et al.*, 2019].

Data Preprocessing

We use the same set of image augmentations in SimCLR [Chen *et al.*, 2020] and BYOL [Grill *et al.*, 2020] without modification, *i.e.* random crops are resized to 224×224, followed by random horizontal flip, color jitter, Gaussian blur and solarization. We flip back the projected 2D feature map before feature alignment if the horizontal flip is applied previously. All the augmentation parameters are kept the same with BYOL.

SimCLR established that data augmentations with drastic color shifts are beneficial for learning invariant representations. For the autonomous driving scenario we keep the same procedure.

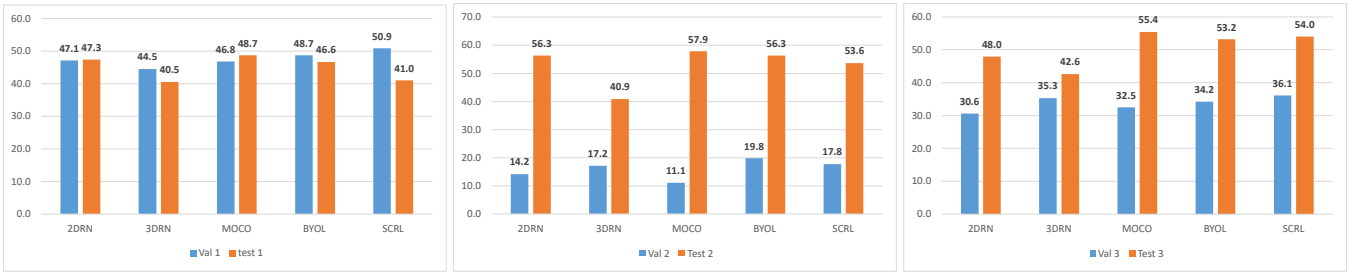
3.2 Object Detection

For object detection in autonomous driving setting, we decided to utilize single-shot detection framework RetinaNet [Lin *et al.*, 2017b] with feature pyramid network (FPN) [Lin *et al.*, 2017a] to demonstrate our key design principles as it has been a common benchmark for both ROAD and BDD100K benchmarks as well as recent state-of-the-art self-supervised methods such as BYOL and SCRL evaluation using the COCO [Lin *et al.*, 2014] benchmark.

4 Experiments

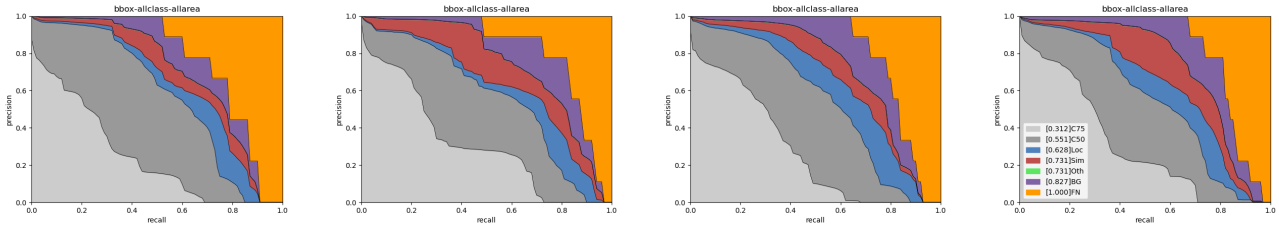
4.1 Datasets

ImageNet. The default setting for visual representation learning in a self-supervised manner takes uses the ImageNet [Deng *et al.*, 2009] benchmark as training data. ImageNet contains 1.2M labeled images with 1000 classes.



(a) `Split-1` (b) `Split-2` (c) `Split-3`

Figure 1: Agent detection results on ROAD benchmark. Each method has been tested on all 3 splits in ROAD. Accuracies reported are `AP@50` metric and mean of 3 different training runs. 2DRN: Supervised RetinaNet, 3DRN: Supervised 3DRetinaNet; where as MOCO, BYOL, SCRL are Self-Supervised RetinaNet accuracies.



(a) Supervised (b) MOCO (c) BYOL (d) SCRL

Figure 2: Evaluation of RetinaNet results on trained on Split-3 of ROAD. Plots show precision-recall curves computed from AP_{50} metric on the `testset`. AP_{50} metric plotted with dark gray. Blue: Localization error; Red: Classification error; Purple: False positives; Orange: False negatives.

	ROAD		BDD100K
Self-Supervised	**Val**	**Test**	**Val**
2D RetinaNet [Lin *et al.*, 2017b]	30.77	50.04	53.40
3D RetinaNet† [Singh *et al.*, 2022]	32.33	39.17	-
YOLOv5† [Singh *et al.*, 2022]	57.9	56.9	-
Self-Supervised	**Val**	**Test**	**Val**
MOCOv1 [He *et al.*, 2019]	30.13	54.00	54.42
BYOL [Grill *et al.*, 2020]	34.24	52.03	49.78
SCRL [Roh *et al.*, 2021]	34.92	49.55	51.57

Table 1: Evaluation of RetinaNet object detection using ROAD and BDD100K datasets. The accuracies reported here are AP_{50} bounding box metric For ROAD benchmark we report evaluation using the mean of three splits. For BDD100K we report results on the validation set. †: Results from citation.

ROAD Benchmark. ROAD benchmark is split into two sets. 18 videos for `trainval` and 4 videos for `testset` for equally representing four types of weather conditions *e.g.* sunny, overcast, snow and night. There are 3 evaluation splits, each containing the same images for `trainval`. For each split, a single type of weather condition is selected for that validation set, while `testset` uniformly contains 1 video from each weather condition. `Split-1` is designed to use an overcast condition in the validation set where the training set is a balanced mix of weather conditions. `Split-2` has all nighttime videos in the validation set, hence the most challenging for drastic illumination shift setting. `Split-3` training set contains all conditions similar to `Split-1`, while `val-3` includes only sunny condition. There are 10 different classes of agents for object detection in ROAD *e.g. Ped, Car, Tl, Cyc, Medveh, OthTL, Bus, LarVeh, Mobike, EmVeh.* The dataset contains 22 videos with 122K annotated video frames, for a total of 560K detection bounding boxes. More details of can be found in [Singh *et al.*, 2022].

BDD100K. BDD100K, a large-scale driving video dataset with extensive annotations for heterogeneous tasks. BDD100K provides 100K images with diverse weather conditions. For object detection with 10 class annotations are provided. Images are split into `train`(70K), `val`(10K) and `test`(20K) sets. Further details can be found in [Yu *et al.*, 2020]. BDD100K has object detection annotation for the following classes: *Car, Sign, Light, Person, Truck, Bus, Bike, Rider, Motor, Train.*

4.2 Evaluation

Pretraining. All supervised and self-supervised pretraining before transferring to detection downstream task is trained on ImageNet benchmark following state-of-the-art methods BYOL, MOCO and SCRL[Grill *et al.*, 2020; He *et al.*, 2020; Roh *et al.*, 2021].

Downstream object detection training. After pretraining we train a RetinaNet detector on the ROAD and BDD100K object detection task with their respective labels in a supervised setting. ROAD benchmarks are evaluated with AP_{50} metric as per [Singh *et al.*, 2022]. For both BDD100K and

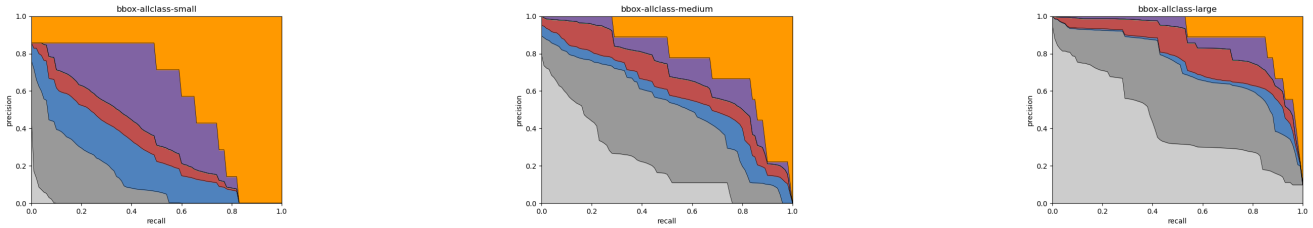


Figure 3: Evaluation of RetinaNet results on trained on `Split-3` of ROAD using MOCO. Different bounding box size results are aggregated in different plots. Small(left), medium(middle), large(right). Plots show precision-recall curves computed from AP_{50} metric on the `testset`. AP_{50} metric plotted with dark gray. Blue: Localization error; Red: Classification error; Purple: False positives; Orange: False negatives.

ROAD benchmarks we report detailed results using AP metric using COCOAPI[Lin *et al.*, 2014] in Appendix A for completion and bounding box size results.

4.3 Results

We compare 3 self-supervised methods as well as provide 3 supervised pretraining baselines for comparison. 2 of the SSL methods, MOCO and BYOL, are only trained for general representation learning for image classification while SCRL builds up on BYOL with detection specific data augmentation and feature and alignment components after the backbone with modified Rio loss that increases performance on COCO benchmark [Roh *et al.*, 2021].

Table 1 summarizes the evaluation accuracies for both supervised and self-supervised methods for agent detection on ROAD and BDD100K benchmarks for object detection. All experiments use RetinaNet [Lin *et al.*, 2017b] single-shot detection with Resent [He *et al.*, 2016] convolution neural network as backbone, exception are 3DRetinaNet, YOLOv5 results from [Singh *et al.*, 2022] for ROAD dataset. Supervised 2D RetinaNet baseline surpassed 3DRetinaNet accuracy as 3DRetinaNet was designed for object detection (agents) as well as action labels in the ROAD dataset. We refer to [Singh *et al.*, 2022] for YOLOv5 results to better put our evaluations in state-of-the-art object detection context.

Results on ROAD benchmark

Between self-supervised methods, MOCO has the best accuracy, that is 3.96 percentage points($p.p.$) better than supervised baseline using AP_{50} metric, on the `testset` which evaluates all weather conditions, while detection-specific SCRL has the best average score in three validation splits that evaluates generalization to different weather conditions.

Looking at performance on individual splits, MOCO is the best across the broad for the testset, while for validation splits, there is no clear winner. It is noteworthy to report that contrastive SSL methods either match or surpass supervised pretraining in all splits and testset. Figure 1 shows evaluation results for all splits in detail.

Figure 2 shows detailed plots of Precision-Recall(PR) curves of different methods, also, it highlights the amount of inaccuracies for the evaluations such as localization errors, class confusion, and misdirection's such as false-negatives and false-positives. PR-curves follow a similar trend across all experiments as the detector is a RetinaNet; however, there are noticeable differences in results as the initial training weights are from different pretraining methods. MOCO makes the least amount of localization errors($8p.p.$), whereas it has the largest($12p.p.$) class confusion. Supervised and BYOL has low class confusion($6p.p. - 6p.p.$) while the localization error is ($6p.p. - 7p.p.$). We show a similar analysis in Figure 3 where we look at accuracies of ground truth boxes categorized by size as it was defined in the COCO benchmark. `Large` bounding boxes have high localization accuracy but the most significant class confusion error. The opposite is true for `Small` bounding boxes. `Medium` bounding box results fall in between in both respects. This behavior is observed across all experiments.

Results on BDD100K benchmark

Here we again see MOCO as the best method by $0.78p.p.$ above supervised baseline for the object detection task. We can also notice other SSL pretraining methods perform lower than the supervised baseline. This points to the superior transfer learning performance of MOCO in both benchmarks and raises the question importance of negative examples in pretraining phase for autonomous driving compared to generic object detection as BYOL and SCRL learn solely from positive examples.

5 Conclusions and Future work

In this work, we have provided one of the first benchmarks for contrastive self-Supervised learning, based on new newly released ROAD benchmark as well as BDD100K for the object detection task. Our results show that contrastive SSL methods can match or outperform supervised pretraining. We also report that object detection focused SSL methods, such as SCRL, which outperform generic SSL methods (such as MOCO or BYOL) on the COCO benchmark do not outperform in the autonomous driving setting. For future work, we have identified areas of weak performance such small-size object localization and big-size object class confusion. We believe performance can be further improved using smarter data augmentations techniques within the contrastive SSL framework. We leave this for future work.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme, under grant agreement No. 964505 (E-pi).

References

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[Chen *et al.*, 2021] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7546–7554, 2021.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Dosovitskiy *et al.*, 2014] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 766–774, 2014.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*, 2016.

[He *et al.*, 2019] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735. IEEE, 2020.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2017a] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944. IEEE Computer Society, 2017.

[Lin *et al.*, 2017b] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007. IEEE Computer Society, 2017.

[Misra and van der Maaten,] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717.

[Roh *et al.*, 2021] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatilly consistent representation learning. In *CVPR*. IEEE, 2021.

[Singh *et al.*, 2022] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, Tom Nicholson, et al. Road: The road event awareness dataset for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (01):1–1, feb 2022.

[Sun *et al.*, 2019] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. *CoRR*, abs/1912.04838, 2019.

[van den Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[Wilson *et al.*, 2021] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[Yu *et al.*, 2020] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Zhang *et al.*, 2017] Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

A Detailed report of metrics

Here we provide all metric computed by COCOAPI[Lin *et al.*, 2014] for our experiments for completion. All results reported here are mean of 3 runs of the same experiment. Results for ROAD benchmark are provided in Table 2, Table 3 and Table 4. Results for BDD100K are provided in Table 5.

Table 2: Detailed results on ROAD, Split-1

Split-1	Val-1						Test					
	AP	AP_{50}	AP_{75}	AP_{Small}	AP_{Medium}	AP_{Large}	AP	AP_{50}	AP_{75}	AP_{Small}	AP_{Medium}	AP_{Large}
2DRN	23.76	47.15	21.08	3.45	18.59	30.30	23.47	45.81	22.00	5.24	17.23	33.54
MOCO	25.04	46.78	24.81	3.98	21.40	31.16	24.73	48.69	22.63	5.16	18.44	34.48
BYOL	25.98	48.71	25.59	3.01	20.35	33.86	23.93	46.60	22.99	4.56	17.72	33.05
SCRL	27.06	50.86	27.18	3.67	18.66	36.06	20.20	40.98	17.94	3.16	14.67	29.12

Table 3: Detailed results on ROAD, Split-2

Split-2	Val-2						Test					
	AP	AP_{50}	AP_{75}	AP_{Small}	AP_{Medium}	AP_{Large}	AP	AP_{50}	AP_{75}	AP_{Small}	AP_{Medium}	AP_{Large}
2DRN	5.64	14.21	3.14	0.39	3.17	10.80	30.07	56.33	30.68	4.58	23.38	42.72
MOCO	4.69	11.13	3.23	0.22	2.98	8.41	31.31	57.86	31.33	4.79	23.67	44.70
BYOL	8.66	19.83	6.11	0.38	3.51	16.37	28.91	56.30	26.38	3.42	23.23	40.68
SCRL	7.31	17.79	4.67	0.31	3.67	13.70	27.64	53.63	24.18	3.57	21.10	38.59

Table 4: Detailed results on ROAD, Split-3

Split-3	Val-3						Test					
	AP	AP_{50}	AP_{75}	AP_{Small}	AP_{Medium}	AP_{Large}	AP	AP_{50}	AP_{75}	AP_{Small}	AP_{Medium}	AP_{Large}
2DRN	14.90	30.63	13.11	1.92	10.25	27.46	24.43	47.98	22.15	2.93	18.47	36.20
MOCO	15.78	32.49	13.35	1.92	10.57	27.64	29.25	55.43	28.77	4.28	24.01	41.26
BYOL	17.61	34.73	16.77	1.86	10.48	31.07	27.44	54.06	23.80	3.30	22.08	39.15
SCRL	19.48	36.10	19.63	1.77	9.50	34.29	28.60	54.02	28.66	3.11	22.86	39.36

Table 5: Detailed results on BDD100K, Validation Split

	AP	AP_{50}	AP_{75}	AP_{Small}	AP_{Medium}	AP_{Large}
2DRN	28.82	53.40	26.61	11.21	35.28	50.50
MOCO	29.59	54.42	27.63	11.68	36.08	51.69
BYOL	26.33	49.78	24.06	9.83	32.04	47.50
SCRL	27.69	51.57	25.74	10.49	33.38	50.07