Appendix for SODA10M: Towards Large-Scale Object Detection Benchmark for Autonomous Driving

A SODA10M dataset

We publish the SODA10M dataset, benchmark, data format and annotation instructions at our website https://soda-2d.github.io. It is our priority to protect the privacy of third parties. We bear all responsibility in case of violation of rights, etc., and confirmation of the data license.

License, Terms of use, Terms of privacy. The SODA10M dataset is published under CC BY-NC-SA 4.0 license, which means everyone can use this dataset for non-commercial research purpose. Find more details in Appendix F.

Dataset documentation. https://soda-2d.github.io/documentation.html shows the dataset documentation and intended uses.

Data maintenance. https://soda-2d.github.io/download.html provides data download links for users (in Google Drive & Baidu YunPan). We will maintain the data for a long time and check the data accessibility in a regular basis. We also plan to extend the scale of unlabeled data to a 100-million level to help develop robust models.

Benchmark and code. The codebases used in our benchmark are open-source. More details of the reproduction code and experiment settings are illustrated in Appendix B.

Data format & Evaluation metrics. Annotation files in SODA10M are stored in standard COCO format, which can be easily accessible to most object detection codebases. We follow the popular COCO API [14] to utilize the Average Precision metric for evaluating detection performance.

Limitations. The major limitation of SODA10M is that some domains exist in the unlabeled set but not in labeled sets, which raises the problem that we can not verify the domain adaptation abilities in these domains. To overcome the limitation, we plan to provide more labeled data in these domains in the future.

B Implement Details

In order to make the experiment results reproducible, we further provide detailed experimental settings for each method in this paper. The basic differences compared with default setting is that class number is set to 6, syncBN is on and multi-scale training (in supervised and semi-supervised methods) is utilized with scale 1920×(864, 907.2, 950.4, 993.6, 1036.8, 1080). Without specifying, all self-supervised and semi-supervised methods adopt ResNet50 [12] backbone. Our models are trained on servers with 8 Nvidia V100 GPU(32GB) cards with Intel Xeon Platinum 8168 CPU(2.70GHz).

B.1 Open-Source Codebase

Table 1: The codebases used in our benchmark.

Codebase	link
Detectron2 [21]	https://github.com/facebookresearch/detectron2
PVT [19]	https://github.com/whai362/PVT
OpenSelfSup	https://github.com/open-mmlab/OpenSelfSup
VINCE [9]	https://github.com/danielgordon10/vince
STAC [17]	https://github.com/google-research/ssl_detection
Unbiased Teacher [15]	https://github.com/facebookresearch/unbiased-teacher

B.2 Supervised Methods

For the 1x schedule, the learning rate is set to 0.02, decreased by a factor of 10 at 8th, 11th epoch of total 12 epochs. Random crop is used as the only data augmentation method and SGD optimizer is adopted with momentum set as 0.9.

Table 2: Implement details for supervised learning benchmark on SODA10M with 8 Tesla V100.

Model	Train split	Default Setting	Difference	GPU hours
RetinaNet [13] 1x	train set	Detectron2	1. backbone no freeze. 2. turn on precise_bn.	0.78×8
Faster RCNN [16] 1x	train set	Detectron2	same as above	0.83×8
Cascaded RCNN [2] 1x	train set	Detectron2	same as above	0.92×8

B.3 Self-supervised Methods

For self-supervised methods, considering time limit, 5-million unlabeled images (split 0, 2, 4, 6, 8) of SODA10M are used, while we also make full use of the other 5-million subset in a sequential training manner (mentioned in †). We follow the standard data augmentation pipeline adopted by MoCov2 [5], which consists of random resized crop, color jitter, gaussian blur and random horizontal flip, for all considered models except MoCov1 [11], which implements all augmentations except gaussian blur. Auto Augmentation [7] is further used in DetCo [22] following the original paper. Cosine learning rate decay is adopted for all models except MoCov1, which uses step-wise learning rate decay to decrease learning rate by 10x at 40 and 50 epochs. The base learning rates are 0.03, 0.03, 0.3, 4.8, 0.06, 0.03 sequentially for the models in Table 3. LARS optimizer [23] is used in SimCLR [4], while all other models implement SGD optimizer with momentum set as 0.9.

Table 3: Implement details for semi-supervised learning benchmark on SODA10M with 8 Tesla V100.

Model	Train split	Default Setting	Difference	GPU days
MoCov1 [11], MoCov2 [5], SimCLR [4], SwAV [3], DenseCL [20]	5-million unlabeled	OpenSelfSup	60 epochs	8.40×8
DetCo [22]	5-million unlabeled	OpenSelfSup	same as above	14.1×8
MoCov1 [11]†, MoCov2 [5]†, SimCLR [4]†	10-million unlabeled	OpenSelfSup	same as above	16.8×8
MoCov1 [11], MoCov2 [5], SimCLR [4], BYOL [10] on **PVT**	5-million unlabeled	OpenSelfSup	same as above	8.60×8
MoCov1 [11], MoCov2 [5], VINCE [9], VINCE+Jigsaw [9] on **VIDEO**	5-million unlabeled to 90k videos	VINCE	1. 800 epochs 2. VINCE augmentations	2.80×8

When finetuning on SODA10M labeled set and other datasets (Cityscape [6] & BDD100K [24]), the setting keeps consistent with the supervised methods and MoCo [11] respectively.

B.4 Semi-supervised Methods

For semi-supervised methods, considering the time limit, only 1-million unlabeled images (split 0) of SODA10M are used. Compared with self-supervised methods, semi-supervised methods are much more efficient. The learning rate of pseudo labeling, STAC [17] and unbiased teacher [15] is set to 0.02, 0.01 and 0.02 respectively. Random crop is used as the only data augmentation method and SGD optimizer is adopted with momentum set as 0.9.

Table 4: Implement details for semi-supervised learning benchmark on SODA10M with 8 Tesla V100.

Model	Train split	Default Setting	Difference	GPU days
Pseudo Labeling(50K)	50K unlabeled	Detectron2	1. backbone no freeze. 2. turn on precise_bn.	0.21×8
Pseudo Labeling(100K)	100K unlabeled	Detectron2	same as above	0.39×8
Pseudo Labeling(500K)	500K unlabeled	Detectron2	same as above	2.00×8
STAC [17]	1-million unlabeled	STAC	same as default	2.50×8
Unbiased Teacher [15]	1-million unlabeled	Unbiased Teacher	same as default	2.80×8

C More Experiments

We also compare SODA10M with other pre-training datasets (i.e., nuScenes [1], Waymo [18] and BDD100K [24]) on object detection task (BDD100K [24] dataset) and instance segmentation task (Cityscapes[6] dataset). As shown in Table 5, SODA10M performs better than other pre-training datasets for the same self-supervised learning method.

Table 5: Detection results(%) of self-supervised models evaluated on BDD100K (B) dataset (with object detection task) and Cityscapes (C) dataset (with instance segmentation task).

		Faster-RCNN 1x (B)			RetinaNet 1x (B)			Mask-RCNN 1x (C)	
Pre-train Dataset	Method	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50
	random init	27.8	53.9	24.8	24.6	47.4	21.6	25.4	51.1
	super. IN	31.8	59.3	29.5	30.6	56.7	28.0	32.9	59.6
BDD100K [24]	MoCo-v1 [11]	31.4	58.6	29.0	29.9	55.6	27.4	31.8	58.7
	MoCo-v2 [5]	31.3	58.5	28.5	30.5	56.7	27.9	32.0	58.8
nuScenes [1]	MoCo-v1 [11]	31.1	58.5	28.6	29.5	55.3	26.9	31.4	58.0
	MoCo-v2 [5]	30.9	58.0	28.3	30.0	55.9	27.5	31.5	58.9
Waymo [18]	MoCo-v1 [11]	31.2	58.3	28.4	29.8	55.6	27.2	31.4	58.5
	MoCo-v2 [5]	31.1	58.3	28.4	30.1	56.0	27.8	31.8	58.7
SODA10M	MoCo-v1 [11]	31.5	58.7	28.7	30.6	57.0	28.2	33.9	60.6
	MoCo-v2 [5]	31.4	58.5	28.9	30.9	56.9	28.6	33.7	61.0

Table 6 shows the performance of existing self-supervised methods evaluated on SODA10M labeled set with a longer schedule and instance segmentation result on Cityscape [6] dataset. We observe that dense contrastive methods (Detco, DenseCL) show excellent results when pre-trained on ImageNet [8], but relatively poor pre-trained on SODA10M unlabeled set. For semantic segmentation performance on Cityscapes with MoCov1, the model pre-trained on SODA10M even surpasses the one pre-trained in ImageNet by 1.6%, further verifying the generalization ability of pre-training on SODA10M.

Table 6: Detection results(%) of self-supervised models evaluated on SODA10M labeled dataset (with object detection task) and Cityscapes (C) dataset (with instance segmentation task).

		Faster-RCNN 2x			RetinaNet 2x			Mask-RCNN 1x	
Pre-train Dataset	Method	mAP	AP50	AP75	mAP	AP50	AP75	mAP (C)	AP50 (C)
	random init	29.6	49.8	31.2	20.9	35.4	21.4	25.4	51.1
	super. IN	38.7	61.0	41.5	35.0	57.0	36.0	32.9	59.6
ImageNet [8]	MoCo-v1 [11]	39.3	60.9	42.5	35.9	57.4	37.3	32.3	59.3
	MoCo-v2 [5]	40.4	62.7	43.6	37.4	59.1	39.3	33.9	60.8
	SimCLR [4]	37.9	61.0	40.4	32.7	53.3	33.8	32.8	59.4
	SwAV [3]	38.2	61.9	40.9	32.6	53.4	33.8	33.9	62.4
	DetCo [22]	39.8	62.1	43.3	35.8	57.5	37.5	34.7	63.2
	DenseCL [20]	40.6	62.9	43.8	37.5	59.4	39.2	34.3	62.5
SODA10M	MoCo-v1 [11]	38.7	60.9	41.1	33.4	56.2	34.3	33.9	60.6
	MoCo-v2 [5]	39.1	60.8	42.6	33.6	56.2	34.8	33.7	61.0
	SimCLR [4]	36.7	59.6	39.1	31.6	53.8	32.3	30.2	57.0
	SwAV [3]	36.0	59.8	37.9	29.7	50.0	30.4	29.4	57.7
	DetCo [22]	37.2	58.9	39.8	31.2	53.5	31.3	32.5	59.8
	DenseCL [20]	38.9	61.0	41.9	33.2	55.4	33.7	33.1	60.7

D Domain Illustration & Diversity Comparison

The distribution of each fine-grained domain in the validation set, testing set and unlabeled set is shown in the Table 7, Table 8 and Table 9.

Table 7: The number of images in each domain in validation set.

	Daytime			Night		
	City street	Highway	Country road	City street	Highway	Country road
Clear	383	961	63	137	167	312
Overcast	597	517	240	288	627	59
Rainy	177	406	0	0	66	0

Table 8: The number of images in each domain in testing set.

	Daytime			Night		
	City street	Highway	Country road	City street	Highway	Country road
Clear	490	2024	250	1591	498	146
Overcast	1216	1103	481	361	237	133
Rainy	917	520	24	9	0	0

Table 9: The number of images in each domain in unlabeled set.

	Daytime				Night				Dawn/Dusk			
	Clear	Overcast	Rainy	Snowy	Clear	Overcast	Rainy	Snowy	Clear	Overcast	Rainy	Snowy
City street	2247K	1483K	458K	140K	1274K	582K	157K	71K	325K	215K	62K	22K
Highway	506K	311K	114K	4K	186K	58K	24K	0.70K	37K	27K	9K	0.17K
Country road	499K	333K	69K	7K	170K	40K	12K	1K	38K	23K	5K	0.50K
Residential	146K	154K	29K	20K	61K	40K	7K	10K	9K	1K	2K	2K

More images in each domain are shown in Fig. 1.

E Acknowledgements

We thank our two data suppliers, named Testin[1] and Speechocean[2] (collected from King-IM-055), for helping us collect and annotate SODA10M dataset.

F Terms of Use and Licenses

Description. Huawei Technologies Co. Ltd (the 'Organizers' ,we", "us", and "our",) provides public access to and use of data that it collects and publishes. The data are organized in datasets (the "Datasets") may be accessed at https://sslad2021.github.io/index.html. Any individual or entity (hereinafter You" or "Your") with access to the Datasets free of charge subject to the terms of this agreement (hereinafter "Dataset Terms"). By using or downloading the Datasets, you are agreeing to comply with the Dataset Terms and any licensing terms referenced below. Use of any data derived from the Datasets, which may appear in any format such as tables and charts, is also subject to these Dataset Terms.

Licenses. Unless specifically labeled otherwise, these Datasets are provided to You under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License ("CC BY-NC-SA 4.0"), with the additional terms included herein. The CC BY-NC-SA 4.0 may be accessed at https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode. When You download or use the Datasets from the Website or elsewhere, You are agreeing to comply with the terms of CC BY-NC-SA 4.0, and also agreeing to the Dataset Terms. Where these Dataset Terms conflict with the terms of CC BY-NC-SA 4.0, these Dataset Terms shall prevail. We reiterate once again that this dataset is used only for non-commercial purposes such as academic research, teaching, or scientific publications. We prohibits You from using the dataset or any derivative works for commercial purposes, such as selling data or using it for commercial gain.

Sharing. We prohibits You from distributing this dataset or modified versions. It is permissible to distribute derivative works in as far as they are abstract representations of this dataset (such as models trained on it or additional annotations that do not directly include any of our data).

[1] http://www.testin.cn

[2] http://en.speechocean.com

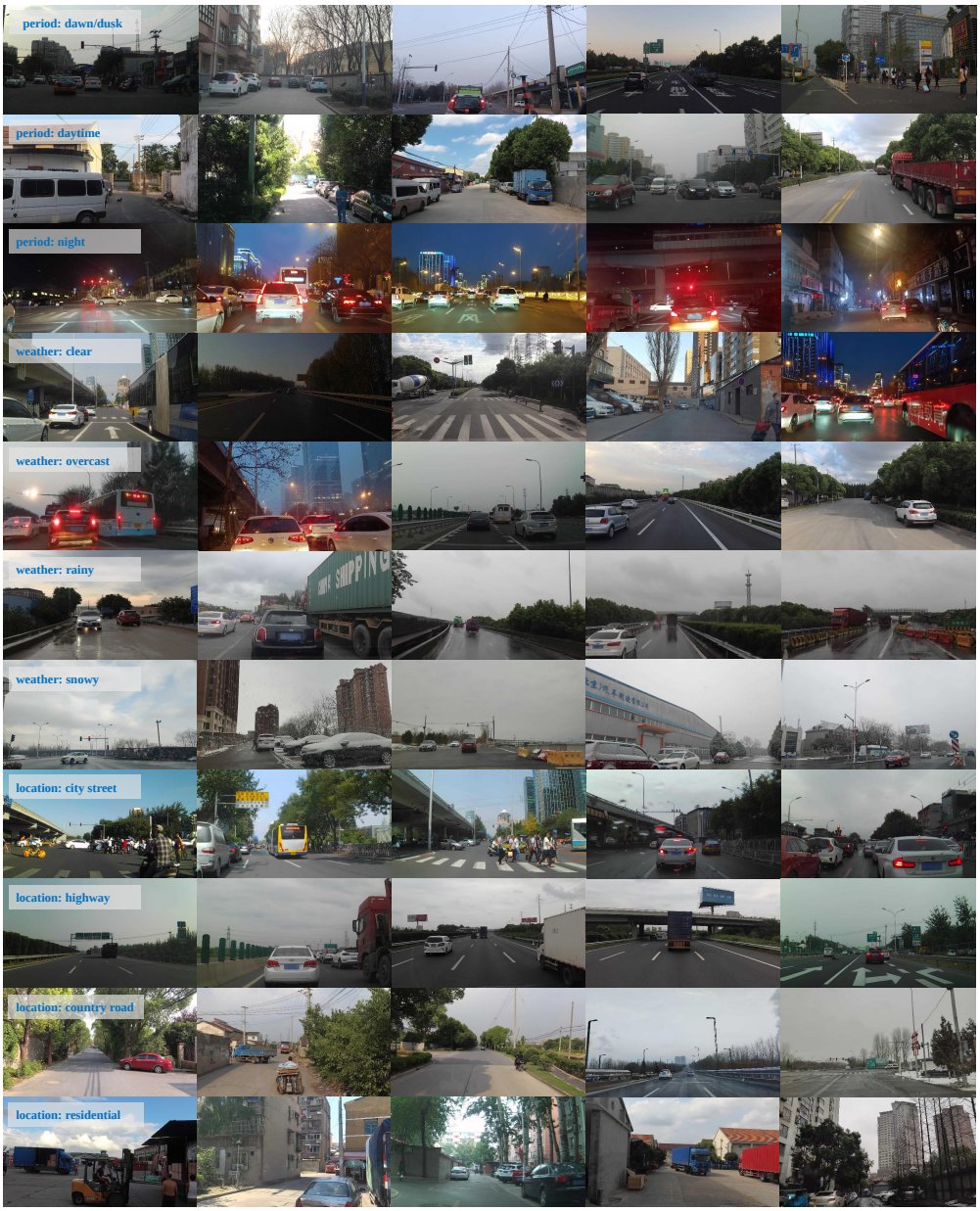

Figure 1: More examples of challenging environments in our dataset.

Trademark. All logos and trademarks used on this website are the properties of us or other third parties as stated if applicable. No content provided on the website shall be deemed as granting approval or the right to use any trademark or logo aforesaid by implication, lack of objection, or other means without the prior written consent of us or any third party which may own the mark. No individual shall use the name, trademark, or logo of us by any means without the prior written consent of us.

Privacy. We will take reasonable care to remove or scrub personally identifiable information (PII) including, but not limited to, faces of people and license plates of vehicles. Furthermore, We prohibits You from using the Datasets in any manner to identify or invade the privacy of any person even when such use is otherwise legal. If You have any privacy concerns, including to remove your name or other PII from the Dataset, please contact us by sending an e-mail to xu.hang@huawei.com.

Warranties. The datasets and the website (including, without limitation, all content and modifications of original datasets posted on the website) are provided "as is" and "as available" and without warranty of any kind, express or implied, including, but not limited to, the implied warranties of title, non-infringement, merchantability and fitness for a particular purpose, and any warranties implied by any course of performance or usage of trade, all of which are expressly disclaimed. Without limiting the foregoing, motional does not warrant that: (a) the content or modifications to the dataset are timely, accurate, complete, reliable or correct in their posted forms at the website; (b) the website will be secure; (c) the website will be available at any particular time or location; (d) any defects or errors will be corrected; (e) the website, content or any modifications are free of viruses or other harmful components; or (f) the results of using the website will meet your requirements. Your use of the website, the datasets, and any content is solely at your own risk. Any entity or individual who suspects that the content on the website (including but not limited to the datasets posted on the website) infringes upon legal rights or interests shall notify our contact xu.hang@huawei.com in written form and provide the identity, ownership certification, associated link (url), and proof of infringement. We will remove the content related to the alleged infringement by law upon receiving the foregoing legal documents.

Limitation of liability. In no event shall motional and its affiliates, or their directors, employees, agents, partners, or suppliers, be liable under contract, tort, strict liability, negligence or any other legal theory with respect to the website, the datasets, or any content or user submissions (i) for any direct damages, or (ii) for any lost profits or special, indirect, incidental, punitive, or consequential damages of any kind whatsoever.

Applicable Law and Dispute Resolution. Access and all related activities on or through the website shall be governed by, construed, and interpreted in accordance with the laws of the People's Republic of China. You agree that any dispute between the parties arising out of or in connection with this legal notice or your access and all related activities on or through this website shall be governed by a court with jurisdiction in Shenzhen, Guangdong Province of the People's Republic of China.

References

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[2] Z. Cai and N. Vasconcelos. Cascade R-CNN: delving into high quality object detection. In *CVPR*, 2018.

[3] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICLR*, 2020.

[5] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297*, 2020.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[7] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] D. Gordon, K. Ehsani, D. Fox, and A. Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv:2003.07990*, 2020.

[10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: a new approach to self-supervised learning. In *NeurIPS*, 2020.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

[15] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021.

[16] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[17] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister. A simple semi-supervised learning framework for object detection. *arXiv:2005.04757*, 2020.

[18] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020.

[19] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv:2102.12122*, 2021.

[20] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.

[21] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[22] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, Z. Li, and P. Luo. DetCo: Unsupervised contrastive learning for object detection. *arXiv:2102.04803*, 2021.

[23] Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[24] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.