LOCO: Logistics Objects in Context

Christopher Mayershofer, Dimitrij-Marian Holm, Benjamin Molter, Johannes Fottner
Chair of Materials Handling, Material Flow, Logistics
Technical University of Munich
Garching, Germany
{christopher.mayershofer, dimitrij-marian.holm, benjamin.molter, j.fottner}@tum.de

***Abstract*—Machine perception is a key challenge towards autonomous systems. Especially in the field of computer vision, numerous novel approaches have been introduced in recent years. This trend is based on the availability of public datasets. Logistics is one domain that could benefit from such innovations. Yet, there are no public datasets available. Accordingly, we create the first public dataset for scene understanding in logistics. The Logistics Objects in COntext (*LOCO*) dataset contains 39,101 images. In its first release there are 5,593 bounding-box annotated images. In total 151,428 instances of pallets, small load carriers, stillages, forklifts and pallet trucks were annotated. We also present and discuss our data acquisition approach which features enhanced privacy protection for workers. Finally, we provide an in-depth analysis of *LOCO*, compare it to other datasets (i.e. *OpenImages* and *MS COCO*) and show that it has far more annotations per image and also a considerably smaller annotation size. The dataset and future extensions will be available on our website (https://github.com/tum-fml/loco).**

***Index Terms*—Dataset, Object Detection, Logistics, Perception**

I. Introduction

Since *AlexNet*'s [1] winning entry in the *ImageNet* [2] Large Scale Visual Recognition Challenge 2012, the field of computer vision has continued to make great strides. This is due to the fact that cutting-edge computer vision approaches such as object classification, object detection or panoptic segmentation are considered to be an enabling technology for new applications throughout a multitude of sectors. The availability of large, public datasets is perhaps the most important factor in this success of machine learning in computer vision.

Current datasets mostly focus on common scenes and objects [2]–[4] or task specific use-cases (e.g. autonomous driving [5], remote sensing [6]). To date, the industrial sector in general and logistics in particular has not displayed any interest in this respect, although enhanced environment perception capabilities are deemed necessary to enable intelligent material flow. As a result of increased digitalization and autonomization, current research in logistics is confronted with similar problems as in the field of robotics or autonomous driving. However, the environment, as well as the hardware (i.e. sensors and compute) in industry is fundamentally different compared to previously mentioned application areas.

In order to overcome these problems, we are releasing the first publicly available dataset that depicts logistics objects in realistic logistics scenes. In its first release, the Logistics Objects in Context (*LOCO*) dataset considers pallets, small load carriers, stillages (also known as lattice boxes), forklifts and pallet trucks (illustrated in Fig. 1). Our dataset reflects previously mentioned challenges and offers a first basis for combining current computer vision research with industrial environments. In addition, *LOCO* might also point out new directions for basic research in the field of computer vision as a result of increased requirements (e.g. class imbalance, small object size, etc.) as further discussed in Sec. IV.

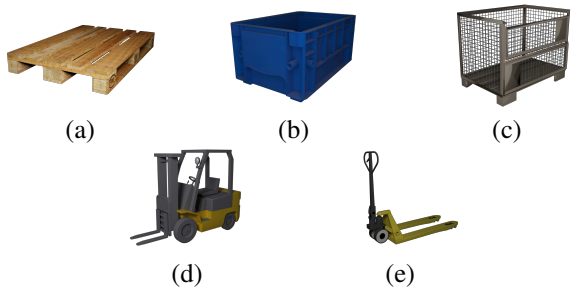

Fig. 1. Logistics-specific objects. Pallets (a), small load carriers (b), stillages (c), forklifts (d) and pallet trucks (e) are common objects within the logistics domain.

To date, generating a scene understanding dataset for logistics has failed most probably due to one of the following problems: *No images on the web.* In contrast to common images, only a few logistics images are available online which could be used to create a dataset. Furthermore, those images that are available are often press images which have little to do with real logistics environments. *Economic pressure.* Logistics environments can almost always be found in companies that pursue an economic goal, and therefore image recordings are viewed critically, as they might deliver information to competition. *Personal privacy.* Furthermore, capturing images of workers is needed in order to create a realistic dataset. At the same time, however, it must be ensured that there is no invasion of a worker's privacy in taking and publishing pictures.

Our contributions are two-fold and can be summarized as follows:

a) *Logistics Objects in COntext (LOCO) dataset.* This is the first publicly available, annotated dataset focusing on object and scene understanding within the logistics domain.
b) *Image acquisition approach with privacy protection.* In order to meet the aforementioned requirements for image

acquisition, we introduce a new method to ensure that workers' privacy is always protected.

II. Related Work

Wide-ranging datasets. One key element for successful machine-learning-based computer vision is a representative dataset. Current datasets try to cover different tasks like image classification, object detection or semantic segmentation. One approach for datasets is to cover as many different classes and tasks as possible, allowing a more general neural network to be trained [2]–[4], [7]. As implied by the name, the Common Objects in Context (*COCO*) datasets goal for example is to cover everyday scenes, contributing over 200,000 labeled images, including 80 object categories [7]. The Developers of the Open Image dataset [4] collected their data by using Flickr as an image source, downloading all accessible images within the Creative Commons Attribution (CC-BY) license, and therefore the dataset was not designed for a specific purpose, it instead tries to cover as many different tasks, scenes and classes as possible.

Specialized datasets. Furthermore, more specialized datasets are available, covering specific tasks often related to certain challenges within diverse use-cases. *ScanNet* [8] provides 12.5 million images created from 1513 3D-scanned indoor environments used for 3D object classification and semantic voxel labeling. Moving from single rooms to outdoor environments, *CityScapes* [9] was created to train models to semantically understand urban street scenes, containing 30 different classes captured in 50 different cities. To address the field of autonomous driving and enhance driver-assistant systems, several datasets exist: The *India Driving Dataset* [10] contains 46,588 images for object detection and 10,003 for segmentation, captured from a camera mounted on a car in India. The *KITTI* dataset [5] was also introduced for autonomous-driving-related challenges, like scene flow estimation [11], and road-area and ego-lane detection [12]. In industrial environments datasets mostly focus a specific tasks rather than scene understanding: *MVTec Industrial 3D Object Detection Dataset* (MVTec ITODD) [13] was created to allow object recognition for different industrial objects, containing 38 different objects like cylinders, clamps and screws. The dataset contains images as well as information about the object, such as pose, diameter and symmetry. Due to the manifold industrial environment, it only covers a small range of objects.

Despite the vast amount of wide-ranging and specialized datasets available, most industrial areas have still not been covered. To tackle the lack of data in the logistical environment, we created the *LOCO* dataset in order to accelerate and improve research based on scene understanding in logistic environments.

III. Data Acquisition Approach

In contrast to common datasets, it is not sufficient to create a realistic logistics dataset by crawling images from the web, due to the amount of available images and the fact that their content does not represent realistic environments. Consequently, we decided to record images using a mobile platform in various logistics warehouses, automatically preprocessing and subsequently annotating them. The following chapter gives an insight into the process and describes its key features.

A. Image Collection

Data acquisition forms the basis for our dataset and is decisive for its quality. From a sensor-technology point of view, five different cameras were chosen to record the dataset, in order to increase the variance of the captured images. On the one hand, low-cost consumer hardware was selected, and on the other hand established computer vision cameras such as the Microsoft Kinect 2 are used. However, industry trends also influenced the selection: Intel Realsense cameras for example are already being used on a large number of mobile robots in logistics. High-level camera information can be found in Table I.

In order to ensure good portability, the cameras were mounted on a mobile unit (see Fig. 2 (a)). Special fixings (see Fig. 2 (b)) hold the cameras in place while driving, but also allow easy re-orientation. The cameras were mounted both in and perpendicular to the direction of travel at different heights. These measures ensure a high variance of the camera perspective throughout the dataset. The cameras were not intrinsically calibrated. We did not do an extrinsic calibration either, as we changed the positions and angles of the cameras on the mobile unit during the recording in order to increase the image variance by introducing novel viewpoints. The effort for an extrinsic calibration of all cameras after each re-orientation would have been too high. Moreover, time delays during the recordings in productive environments at the companies had to be minimized. In addition to the cameras, the mobile unit is equipped with a mobile power supply (i.e. battery and converter) and two recording computers. The recording process is as follows: the mobile unit is moved through the logistics environment and the computers store images at a frequency of one hertz. During recording, the position of individual cameras is changed to further increase the perspective change.

While recording, it is possible to take pictures of employees at work. On the one hand, this is an intentional feature in order to create a representative dataset. On the other hand, this would endanger the employee's privacy. To resolve this discrepancy, neural networks are deployed for automated face recognition. Even before an image is saved on the hard drive, faces are recognized and made unrecognizable via pixelation (see Fig. 2 (c)). The automated process step is supplemented by a manual annotation step, where possible false positives are manually annotated and pixelated.

Lastly, potential companies in different sectors were identified, contacted and images were recorded. Again, emphasis was put on the variance of the selected warehouses in order to keep the information entropy of the dataset high.

TABLE I
CAMERAS IN USE. IMAGES WERE CAPTURED USING DIFFERENT CAMERAS (I.E. ACTION CAMERA, GAME CONSOLE CAMERA, WEBCAM, ETC.) IN ORDER TO ENSURE A DIVERSE DATASET

Camera	Data	Resolution in pixel	Field of View (HxV) in degrees
MS Kinect v2	RGBD	1920 x 1080	84.1 x 53.8
Intel Realsense D435	RGBD	1920 x 1080	91 x 65
SJCAM SJ-4000	RGB	1920 x 1080	170 x N/A
MS LifeCam HD-3000	RGB	1280 x 800	68.5 x N/A
Logitech C310	RGB	1280 x 800	60 x N/A

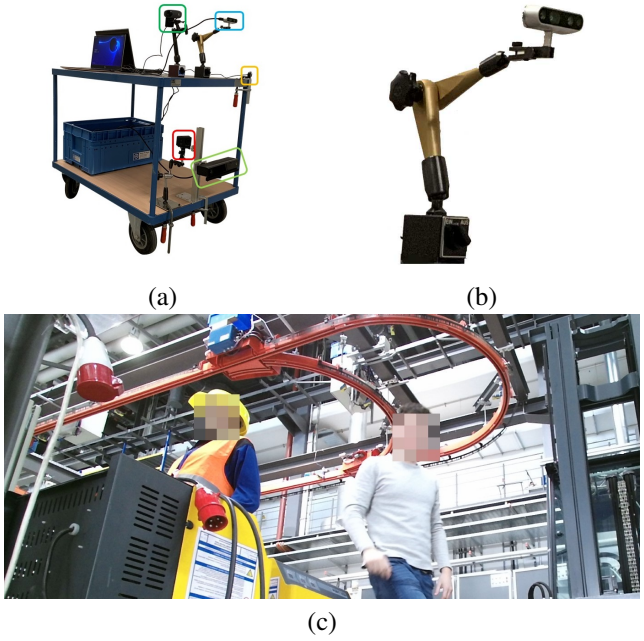

Fig. 2. Data acquisition with privacy protection. (a) We mount five cameras on our mobile unit using special fixings (b) to be able to re-adjust the camera perspective. (c) In order to ensure workers' privacy is protected, we deploy a two-step blurring approach.

B. Image Preprocessing

In total, 64,993 color images were captured. Before annotation all images underwent a three-step preprocessing procedure. These steps were chosen in order to enable high-quality annotations and maintain high information entropy of the dataset per image.

Step 1: Removing blurred images. This step eliminates images that are too blurry due to extensive motion, making them unrecognizable and therefore un-annotatable, even for experts. Variance of Laplacian was chosen as a measure for blurriness. Images exceeding an empirically determined threshold were excluded. In total, 17,109 images were removed due to blurriness.

Step 2: Removing similar images. Here we aim to increase the information entropy of the dataset per image by removing images with the same structural content from the dataset. This simplifies the annotation effort and at the same time ensures the datasets balance. In order to analyze the similarity of the images, they were sorted by acquisition time, and the structural similarity [14] of two consecutive images was calculated. The exclusion criterion is again an empirically determined threshold. In total, 8,783 images were sorted out due to structural similarity.

Step 3: Random sampling. Finally, 15 % of the remaining images were randomly selected for annotation. Here, too, our goal is to maximize the information entropy of the dataset per image while minimizing the annotation effort.

C. Image Annotation

Annotations were generated using the *COCO-Annotator*[1]. The images were annotated using bounding boxes. We extended the annotator in the form of a dedicated bounding box tool, new hotkeys and additional automation to make labeling more ergonomic, efficient and smooth. The majority of changes were fed back into the project. Lastly we included a blurring mechanism which allows the automatic blurring of certain bounding boxes. This is used to blur faces which were not detected by the neural network.

Since the object classes are logistically specific, annotators had to be trained prior to the start. During this training, object classes were explained and exemplified using images. Additionally a logistics-specific compendium and experts were available for further exchange. To ensure consistent annotation results, only a single class was selected at a time and annotated throughout a subset, before the annotation of the next class was started. As far as possible, annotators were only assigned to one object class. Finally, samples were taken from each subset for validation purposes.

IV. DATASET STATISTICS, BENCHMARKS AND PRELIMINARY ANALYSIS

LOCO consists of 39,101 images grouped into five image subsets. There are 5,593 labeled images, totalling in 151,428 human-labeled annotations over five different logistics-specific object categories: Small load carrier, pallet, stillage, forklift and pallet truck. Each subset represents a specific warehouse and contains images acquired using our previously described approach. Images are stored in JPEG format. Annotations are provided in *COCO* format. *LOCO* and future extensions will be available on our website (https://github.com/tum-fml/loco).

A. Statistics

To better illustrate the dataset and the possible challenges for computer vision applications in logistics, we analyzed *LOCO* in its current release with respect to object class distribution, number of annotations per image and object size distribution, and compared it to *COCO* and *OpenImages*. The object class distribution is illustrated in Fig. 4. This shows the unbalanced character of the application specific dataset. In particular, classes of the super-category *load carrier* (i.e. pallet, small load carrier and stillage) are represented 43 times more often than objects of the super-category *transportation vehicles* (i.e. pallet truck, forklift).

[1]https://github.com/jsbroks/coco-annotator

Fig. 3. *LOCO* sample images. The figure shows different object categories labeled throughout different subsets. For illustration purposes, only labels of specified classes are shown. Note the object difference within classes (e.g. pallet truck), image quality due to different cameras, and the contextual difference within subsets. In addition, subset four and five do not contain stillage instances, as they are not common within distribution centers.

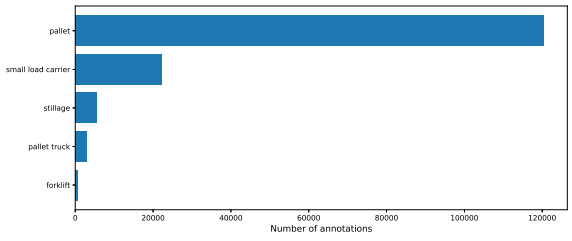

Fig. 4. Number of annotations per class in the *LOCO* dataset.

Moreover, the number of annotations per image in *LOCO*'s subsets was evaluated and compared. A cumulative density histogram in Fig. 5 highlights the difference between the datasets under consideration. The domain difference between logistics and other common datasets is apparent: While half of the images in common datasets have fewer than five annotations per image, our dataset provides on average 31.1 annotations per image across the subsets. Finally, the instance size of the different classes was examined, compared and plotted in Fig. 5. The graph shows the relative bounding box size cumulated over all annotations within a dataset. Once again, a difference can be observed: 90 % of *LOCO*'s annotations have a relative bounding box size of less than 2 %. In comparison, less than 25 % of annotations in *OpenImages* training set and approximately 70 % of annotations in the *COCO* dataset cover the same relative size, meaning that there are considerably more small annotations in the *LOCO* dataset. In summary, analysis and comparison of the *LOCO* dataset showed that its classes are unevenly distributed, and that, on average, *LOCO* has more annotations per image than the *COCO* and *OpenImage* datasets. Additionally these annotations are much smaller.

B. Benchmark

To be able to compare the detection performance of different models, we also define the *LOCO* dataset benchmark. We specify a training and evaluation split which, unlike other datasets, is not random. On the contrary, we use subsets

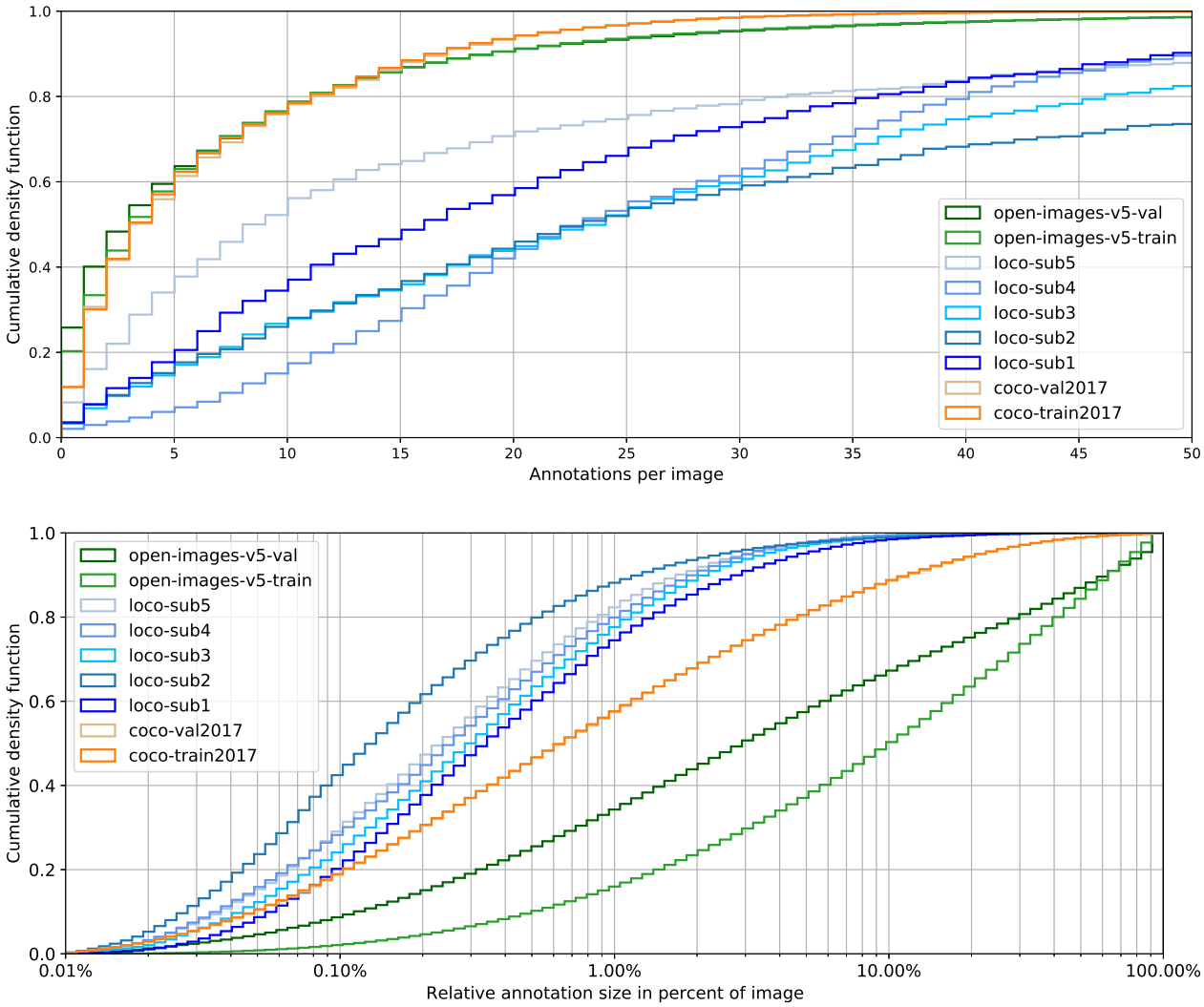

Fig. 5. Statistics of the *LOCO* dataset. Top: Cumulative density function of annotations per image for *LOCO*, *COCO* and *OpenImages*. Bottom: Cumulative density function of object size for *LOCO*, *COCO* and *OpenImages*. Note the logarithmic scale.

to perform the training and evaluation split, since each of them corresponds to one particular logistics environment. This guarantees that the training and evaluation sets are disjoint from each other, which implicitly shifts the focus in machine-learning applications towards generalizable models, since certain conditions (i.e. scene, lighting, color) may have not been encountered in the training set. We divide the five subsets as follows: subsets two, three and five make up the training set, whilst subset one and four serve for evaluation purposes. This results in a ratio of 3/5 training and 2/5 validation split. The subsets were combined to best reflect the class distribution over both, training and validation set. The validation set has 1.96% more pallet truck instances, 8.55% more pallet instances, 0.57% more SLC instances, 7.2% fewer stillage and 48.16% fewer forklift instances. As usual for bounding box detection, results are calculated using mean average precision as the key performance indicator.

C. Preliminary Analysis

Finally, we present first results of an object detection model trained on the *LOCO* dataset. For this purpose, we used the Darknet[2] and Detectron2[3] framework to train *YOLOv4-608*, *YOLOv4-tiny* and *Faster R-CNN* (R50-FPN-3x). All models were trained and evaluated as described in Section IV-B. Furthermore, we used pretrained weights available in each model zoo and fine-tuned on the *LOCO* dataset with standard training settings provided by each framework.

For evaluation the average precision (AP) metric [3] at an intersection over union (IoU) of 0.50 was chosen. AP results per class are documented in Table II. Across the different classes, this results in a mean AP (mAP) of 41.0%, 22.1% and 20.2% for *YOLOv4-608*, *YOLOv4-tiny* and *Faster R-CNN*, respectively. Looking at mAP@0.50 only, all models perform worse

[2]https://github.com/AlexeyAB/darknet
[3]https://github.com/facebookresearch/detectron2

on the *LOCO* benchmark compared to the *COCO* detection challenge. On average, the fine-tuned models perform 27.5% (mAP@0.50) worse on the *LOCO* challenge when compared to the *COCO* baseline. As quantitative metrics are sometimes hard to grasp, we additionally ran the trained model on a video for illustration purposes. This video was recorded in the chair's laboratory and is disjoint from the *LOCO* dataset. The video is available online (https://github.com/tum-fml/loco).

TABLE II
PRELIMINARY ANALYSIS. TABLE SHOWS EVALUATION RESULTS FOR *YOLOv4-608*, *YOLOv4-tiny* AND *Faster R-CNN* TRAINED ON *LOCO*.

Model	*YOLOv4-608*		*YOLOv4-tiny*		*Faster R-CNN*	
Dataset	*LOCO*	*COCO*	*LOCO*	*COCO*	*LOCO*	*COCO*
mAP@0.50	41.0%	65.7%	22.1%	40.2%	20.2%	60.0%
Small load carrier	27.7%	N/A	18.1%	N/A	28.3%	N/A
Pallet	65.0%	N/A	36.2%	N/A	19.8%	N/A
Stillage	53.1%	N/A	31.3%	N/A	37.6%	N/A
Forklift	31.3%	N/A	11.6%	N/A	2.9%	N/A
Pallet truck	28.1%	N/A	13.3%	N/A	12.5%	N/A

V. FUTURE WORK

We see *LOCO* as the first release towards a bigger objective; a combination of datasets that realistically capture industrial environments, ready to be used for Autonomous Mobile Robot and Computer Vision research for industrial applications. Therefore we are working on extending the dataset with additional data (more subsets), novel data types (e.g. depth data) and annotations (e.g. segmentation). Furthermore, we will provide a synthethic version of the *LOCO* dataset, covering the same object categories.d

VI. CONCLUSION

We presented *LOCO*, the first dataset focusing on scene understanding in logistics environments. To the best of our knowledge, it is the first publicly available dataset in the logistics domain. It currently consists of 39,101 images of which 5,593 were annotated. In total 151,428 pallet, small load carrier, stillage, forklift and pallet truck instances were labeled. Furthermore we presented our data acquisition approach using object detection to automatically blur faces of workers captured. Lastly we thoroughly analyzed our dataset and compared it to the *OpenImages* and *COCO* datasets. The comparison shows that our dataset not only has far more annotations per image but also consists of far smaller instances.

Considering the low number of annotated images, the uneven class distribution, the large amount of annotations per image, the amount of very small annotations (see Sec. IV-A) as well as the preliminary analysis (see Table II), the developed dataset and corresponding industry-oriented benchmark can be regarded as challenging for state-of-the-art object detection approaches.

In the future, we plan to further extend *LOCO* to enable scene understanding for industrial applications.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[4] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari, "The Open Images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *arXiv:1811.00982*, 2018.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[6] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, p. 296307, Jan 2020.

[7] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV) Proceedings*, 2014.

[8] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[10] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar, "IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *Proc. of IEEE Winter Conference on Applications of Computer Vision*, 2019.

[11] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[12] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. of IEEE International Conference on Intelligent Transportation Systems*, 2013.

[13] B. Drost, M. Ulrich, P. Bergmann, P. Hartinger, and C. Steger, "Introducing MVTec ITODD a dataset for 3D object recognition in industry," in *Proc. of IEEE International Conference on Computer Vision Workshops*, 2017.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.