OFFSED: Off-Road Semantic Segmentation Dataset

Peter Neigel [1,2], Jason Rambach [1] and Didier Stricker [1,2]
[1] *German Research Center for Artificial Intelligence, Kaiserslautern, Germany*
[2] *University of Kaiserslautern, Germany*
peter.neigel@dfki.de *jason.rambach@dfki.de* *didier.stricker@dfki.de*

Keywords: Semantic Segmentation, ADAS, Outdoor, Industrial Vehicles

Abstract: Over the last decade, improvements in neural networks have facilitated substantial advancements in automated driver assistance systems. In order to manage navigating its surroundings reliably and autonomously, self-driving vehicles need to be able to infer semantic information of the environment. Large parts of the research corpus focus on private passenger cars and cargo trucks, which share the common environment of paved roads, highways and cities. Industrial vehicles like tractors or excavators however make up a substantial share of the total number of motorized vehicles globally while operating in fundamentally different environments. In this paper, we present an extension to our previous Off-Road Pedestrian Detection Dataset (OPEDD) that extends the ground truth data of 203 images to full image semantic segmentation masks which assign one of 19 classes to every pixel. The selection of images was done in a way that captures the whole range of environments and human poses depicted in the original dataset. In addition to pixel labels, a few selected countable classes also come with instance identifiers. This allows for the use of the dataset in instance and panoptic segmentation tasks.

1 INTRODUCTION

Advanced Driver Assistance Systems (ADAS) for private passenger cars, trucks, mobile working machines and other vehicles need to be able to navigate their often complex environments in an intelligent manner in order to provide a benefit to the human driver while maintaining all required safety standards. To this effect a suitable understanding of the surroundings is required. In computer vision terms, the system must be able to semantically understand the content of images received by attached cameras in its entirety: Pedestrians must be recognized to support emergency breaks, passable ground and marked lanes need to be understood in order to maneuver the vehicle and so forth. Current state-of-the-art approaches to semantic full image segmentation mainly rely on convolutional neural networks that take in a monocular RGB-image as input and produce a class label for every pixel in the image. To train these networks in a supervised manner large image datasets are needed that come with ground truth semantic labels for every pixel in the image. Due to the focus of current research on private passenger cars and cargo trucks, most available datasets including semantic pixel labels

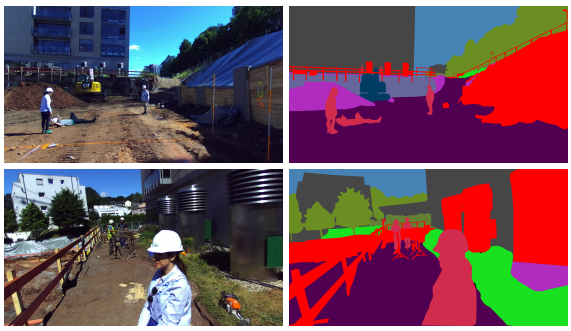

Figure 1: Full image segmentation masks (right) and their corresponding images (left). Construction sites are common environments for mobile working machines, but are underrepresented in most deep learning datasets.

depict urban environments or highway roads. These environments are not suitable as training images for the assistance systems deployed on industrial vehicles and working machines, since they often operate in semantically and visually distinct surroundings, i.e. the passable ground for a passenger car and an excavator differ greatly and excavators encounter different objects than cars in a city. Since Tabor et al. Tabor et al., 2015 have demonstrated that visual gradient orientation alignment is distinctly specific to the environment, it is unclear whether neural

Figure 2: Exemplary frames taken from different datasets. Left: Cityscapes dataset Cordts et al., 2016. Center: KITTI Stereo 2015 Dataset Geiger, Lenz, and Urtasun, 2012. Right: OPEDD. Urban scenes differ substantially from off-road environments in how they are made up visually, including colour spectrum, gradient orientations and human poses.

networks are able to generalize from one surrounding to the other. Additionally, even classes found in both environments can differ between them: Persons in urban settings often display more constrained poses (upright standing/walking) than in working environments (often crouching, lying on the floor). Although these industrial vehicles are used in dozens of industries, from coarse earthwork to constructions and from plowers to harvesters, they are neglected by the currently available datasets.

Previous works (Halevy, Norvig, and Pereira, 2009; Zhu et al., 2015) hint to the fact that data may be more important to neural network detection performance than algorithms or architecture. While our previous work, OPEDD (Neigel et al., 2020), tried to address the problem of person detection, this new paper intends to fill the gap and add an off-road variant to the collection of semantic segmentation datasets. Our dataset is a subset of 203 images from OPPED and therefore depicts the same off-road environments including Meadows, Woods, Construction Sites, Farmland and Paddocks. The pedestrians are portrayed in varying poses, of which many are highly unusual in the ADAS context, including crouching, lying down or handstands to offer some extremes. To our knowledge, there is no other dataset besides ours including real construction sites complete with semantic segmentation. Our work comes with stereo-images, where the left images come with manually created ground truth segmentation masks for the full image. In addition to per-pixel semantic labels, our work offers instance labels for objects where instances can be defined (*thing* classes) and depth from stereo. These ground truth labels allow for the tasks of semantic segmentation, object and instance detection with bounding boxes and object masks, as well as panoptic segmentation.

Figure 3: Countable classes like persons can be separated by instance IDs.

2 RELATED WORK

For the last decade, Neural-network-based approaches have been dominating many computer vision tasks in terms of segmentation quality. This fact coupled with the need of those networks for data has propelled the creation of many scene segmentation datasets for ADAS. Due to the commercial nature of said systems, most publicly available datasets show urban surroundings or highway roads. This chapter gives an incomplete overview of popular datasets for different tasks and surroundings.

Cityscapes (Cordts et al., 2016) is one of the most used image datasets for dense urban environments. The images were obtained with the help of a stereo-camera mounted onto a car which was driven through 50 German cities. The annotations include pixel-level semantic labels for 30 classes including persons, cars, roads, buildings, traffic lights, of which 17 classes are used for evanulation in their benchmark. Out of a total of 25,000 images, 5,000 come with finely annotated pixel masks and the remaining 20,000 are coarsely annotated, meaning that the pixel masks often don't coincide with the true object borders. In addition to semantic labels, Cityscapes also includes instance IDs allowing for the differentiation of countable objects. All annotations in total allow for the use of the dataset in object-detection, semantic-, instance- and panoptic-segmentation procedures.

Kitti (Geiger, Lenz, and Urtasun, 2012) is another widely used dataset for driving scenes. It displays

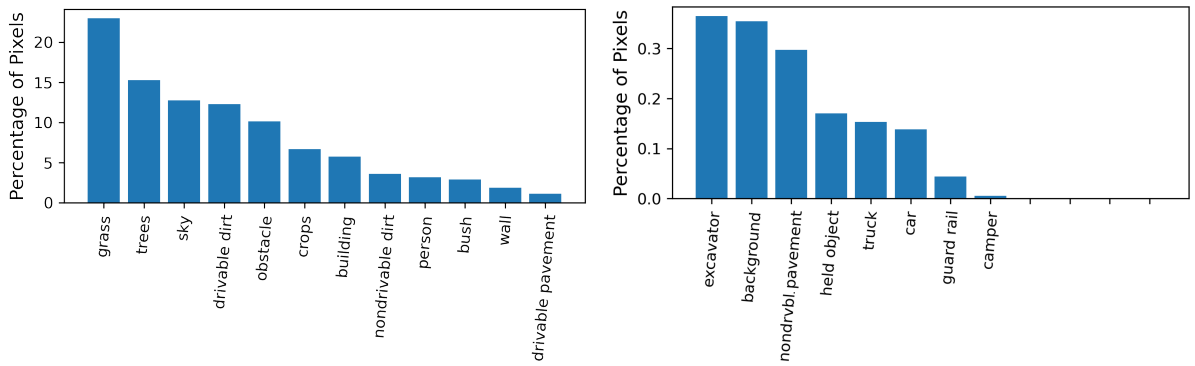

Figure 4: Percentage of pixels belonging to each class over the whole dataset. Left and right plots have different y-axis scales for better visibility.

Dataset	*N* Images	Depth	Segm. Masks	Environment	Human Poses
KITTI 2015	200	LIDAR	✓	Urban/Street	Std. Urban
Cityscapes	25,000	Stereo	✓	Urban/Street	Std. Urban
A2D2	41,277	LIDAR	✓	Urban/Street	Std. Urban
Mapillary Vistas	25,000	-	✓	Urban/Street	Mostly Std. Urban
ApolloScape	140,000	LIDAR	✓	Urban/Street	Std. Urban
BDD100K	10,000	-	(✓)	Urban/Street	Std. Urban
NREC	76,662	Stereo	-	Agricultural	Std. Agricultural
KIT MOMA	5663	-	-	Multiple Off-Road	Contruction
OFFSED	**203**	**Stereo**	✓	**Multiple Off-Road**	**Wide Range, Unusual**

Table 1: Comparison of contents of datasets for semantic segmentation. *N* Images indicates how many images come with ground truth annotations.

mostly urban environments with class definitions similar to the Cityscapes dataset. In addition to two stereo-camera pairs - one grayscale, one color - the capturing rig includes an inertial/GPS system as well as a 3D laser scanner. This allows, in addition to per-pixel semantic labels, also the addition of ground truth data for depth, 3D bounding boxes and camera trajectory. For this reason the KITTI dataset offers a large array of benchmarks, from object detection and semantic segmentation over depth estimation up to odometry, tracking and scene flow. For semantic segmentation, only the subset KITTI 2015 offers according segmentation masks.

The Audi Autonomous Driving Dataset (A2D2) (Geyer et al., 2020) makes use of six cameras and five LIDARs. It includes 41,277 frames that come with ground truth annotations for semantic segmentation. Since the environments depicted are highways, country roads and cities in southern Germany, it defines 38 different classes, of which many are similar to KITTI and Cityscapes.. Other provided annotations include 3D point clouds, 3D bounding boxes and instance segmentation in addition to data like steering wheel angle, throttle, and braking.

Mapillary Vistas (Neuhold et al., 2017) is a dataset containing 25,000 images from locations around the world including parts of Europe, North and South America, Asia, Africa and Oceania, therefore covering large diversity and geographic extent. The frames are taken from a wide range of capture devices including mobile phones, tablets, action cameras and professional capturing rigs. While the environments encompass urban, countryside and off-road scenes, most images are still taken from the street or dirt tracks. The ground truth annotations provide pixel-level semantic labels for 66 classes and instance labels for a subset of 37 classes.

One of the largest datasets with a focus on ADAS is ApolloScape (Huang et al., 2020). Captured with a rig including a stereo camera, two 360° laser scanners and an IMU/GNSS system, it consists of 140,000 video frames taken on urban locations in China. It supplies ground truth per-pixel semantic labels as well as 3D point clouds, of which some points are also semantically labelled, for 28 classes. Additionally, instance ID's, 3D car instances, lane markings and camera locations are provided.

The Berkeley Deep Drive Dataset (BDD100K) (Yu et al., 2020) focuses on object detection, supplying 100,000 videos of street and traffic scenes in the USA with ground truth bounding boxes for 10 classes. Semantic segmentation masks are provided

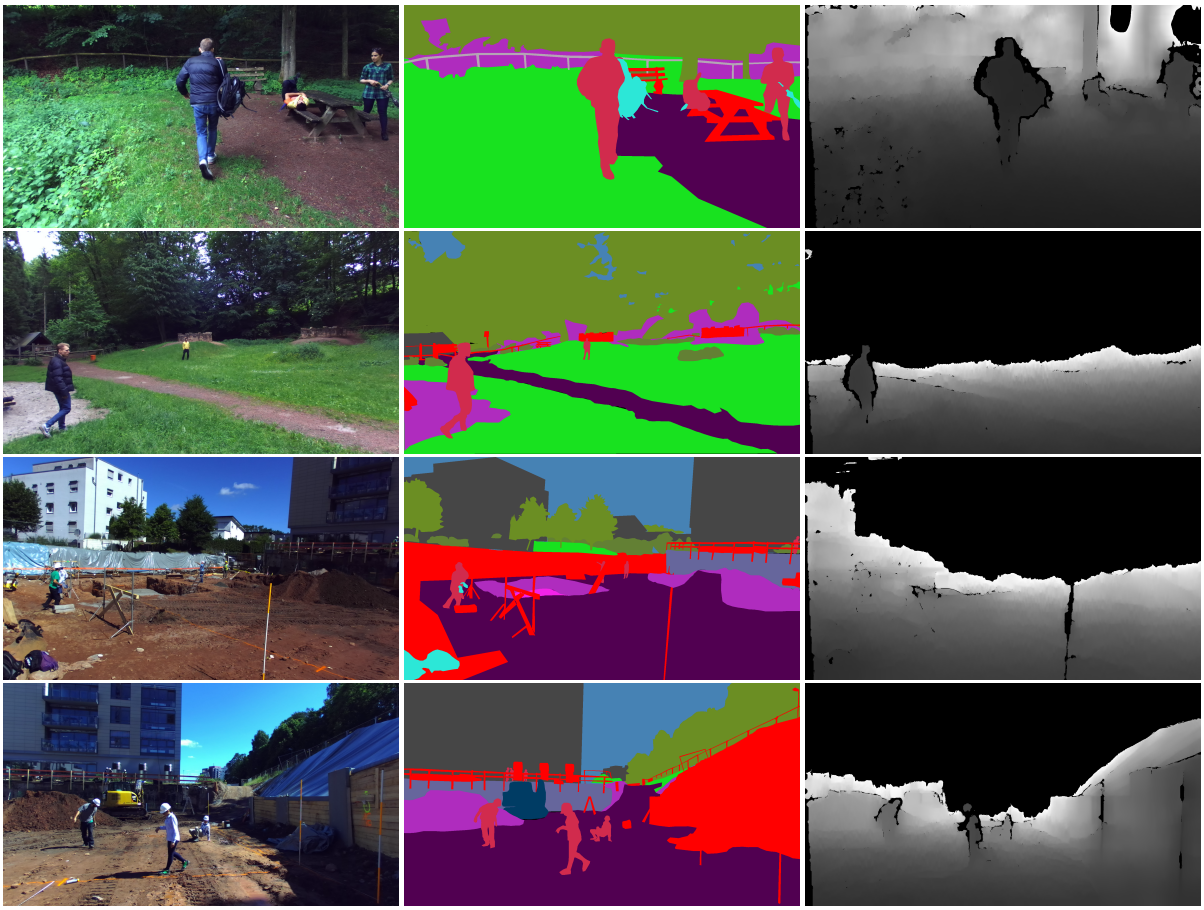

Figure 5: Selection of images from the presented dataset. Left column: Left frame of stereo image. Center column: Corresponding segmentation masks. Right column: Corresponding depth maps. The dataset shows a wide range in environments and unusual human poses.

for 10,000 videos and 40 classes, of which a small subset comes with instance IDs. Additional data provided includes car trajectories from IMU/GPS and lane markings. In contrast to the previous works, the National Robotics Engineering Center Agricultural Person-Detection dataset (Pezzementi et al., 2018) focuses on off-road environments instead of urban traffic scenes. Consisting of 95,000 images taken in apple and orange orchards of which 76,662 have annotations, it provides only bounding boxes for persons, making it unsuitable for complete scene semantic segmentation or instance segmentation. A further distinctive feature of this dataset is the variety of poses that persons are depicted in: Compared to city scenes, workers in the orchards are found in crouching, stretching and otherwise unusual poses more often.

The Karlsruhe Mobile Machines dataset (KIT MOMA) (Xiang et al., 2020) sets a focus on environments where industrial vehicles and mobile machines like excavators, wheel loaders, bulldozers and dumpers operate. The collection includes 5,663 images taken from outside of the vehicles. Eight different vehicle classes are defined and annotated manually with ground truth bounding boxes, resulting in 19,997 object instances.

3 DATA CAPTURING

The images in this dataset are a subset of the data from our previous work, OPEDD (Neigel et al., 2020). OPEDD consists of 1018 stereo images that were captured in five different environments, *meadows, woods, construction sites, farm-land* and *paddocks*. The images were extracted from stereo video sequences recorded with a ZED camera ("ZED", 2020) in lossless compression. The images are rectified and depth from stereo is provided. The full technical details of the data capturing process can be taken from (Neigel et al., 2020).

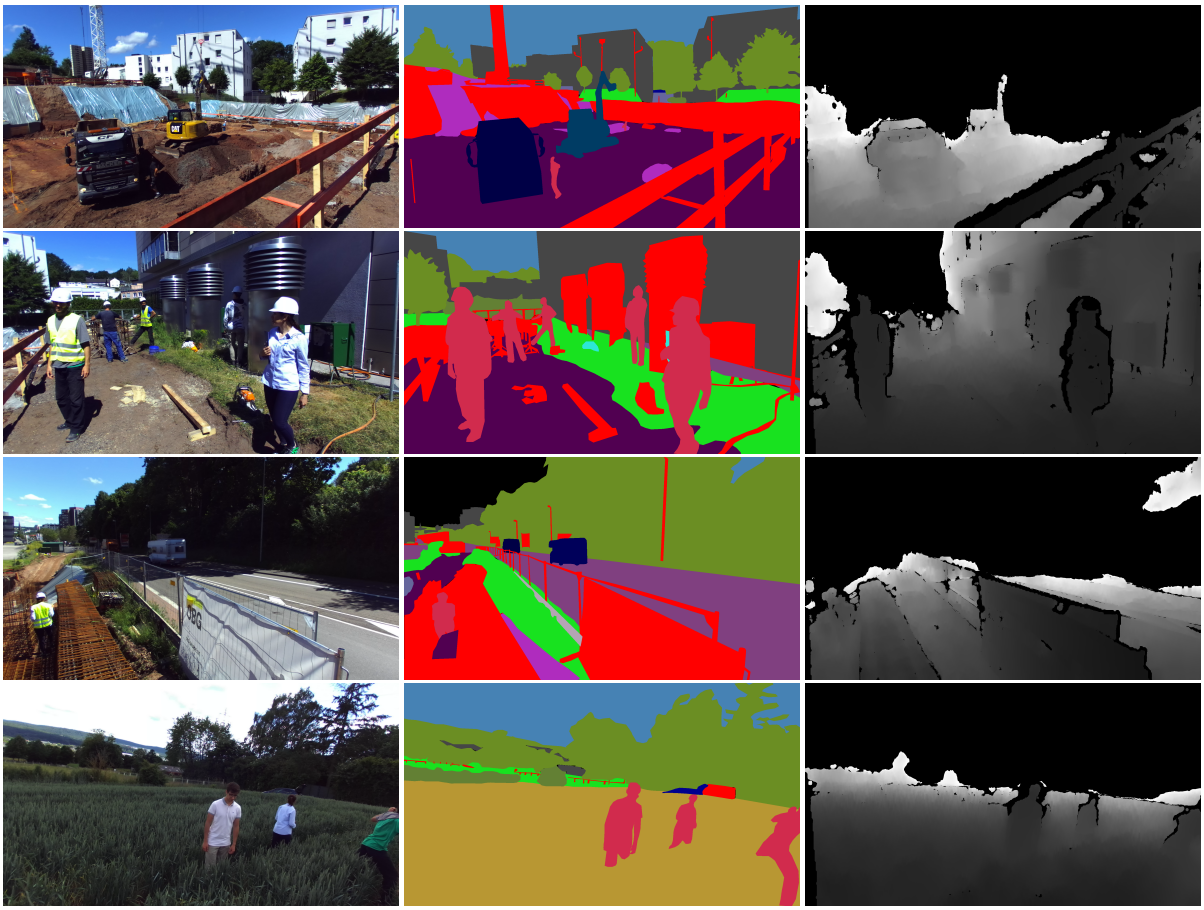

Figure 6: Further selection of environments contained in the presented dataset. Left column: Left frame of stereo image. Center column: Corresponding segmentation masks. Right column: Corresponding depth maps. Among other environments, our dataset delivers segmentation masks for real construction sites.

4 ANNOTATIONS

Manually annotating images for semantic segmentation is a labour intensive and costly task. The ground truth semantic segmentation annotations were created manually with the help of the annotation tool CVAT ("Computer Vision Annotation Tool", 2020). In this tool annotators draw polygons for every object and label them according to predefined classes. Additionally, the polygons are assigned a z-layer level and thereby a depth ordering. This allows for borders between polygons to be drawn only once, saving time in the annotation process. The 19 classes we defined are *grass, trees, sky, drivable and non-drivable dirt, obstacles, crops, building, person, bush, wall, drivable and non-drivable pavement, held/carried object, truck, car, excavator, guard rail* and *camper*. The full image pixel masks are then created from the polygons with the consideration of the polygon depth.

Each polygon can be supplied with additional arbitrary attributes. For a subset of countable *thing*-classes we assign instance identifiers to each object-polygon. Since most classes (e.g. grass, sky, dirt) are not suitable for division into instances, this is done for only a small subset of classes: *person, car, excavator, truck* and *camper*. In some instances single objects are comprised of several polygons. This is necessary when an an object is visually cut in two or more parts by an occluding object.

Because the segmentation masks are created from polygons, the masks can be extended to include more classes easily. This can be desirable if classes should be divided into more fine-grained sub-classes. The classes were chosen in a way that reflects possible environments for mobile working machines. Special interest should be paid to classes like drivable and non-drivable dirt or pavement/concrete since these classes are dependent on the ego-vehicle and are therefore prone to semantic as well as visual ambiguity. A dirt heap that can be driven over by a larger machine may

pose an obstacle for a smaller one for example. During the annotation process, we labelled classes with the background of ADAS for mobile working machines in mind and used that fact to guide decisions e.g. between drivable and non-drivable dirt.

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented an extension to our previous off-road pedestrian detection dataset OPEDD, which adds full image semantic segmentation annotations to 203 images. To this end we defined 19 semantic classes: *grass, trees, sky, drivable and non-drivable dirt, obstacles, crops, building, person, bush, wall, drivable and non-drivable pavement, held/carried object, truck, car, excavator, guard rail* and *camper*. The chosen images were selected in a way that retains the wide range of outdoor environments and special human poses that were depicted in OPEDD. For future work, we intend to completely semantically annotate some of the video sequences the images were taken from, to allow for the use of the dataset in semantic SLAM and tracking tasks.

ACKNOWLEDGEMENTS

We would like to thank Ahmed Elsherif and Mitesh Mittal for their help in annotating and reviewing the quality of the annotations.

References

Computer Vision Annotation Tool (2020). URL: https://software.intel.com/content/www/us/ens/develop/articles/computer-vision-annotation-tool-a-universal-approach-to-data-annotation.html (visited on 12/18/2020).

Cordts, Marius et al. (2016). "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223.

Geiger, Andreas, Philip Lenz, and Raquel Urtasun (2012). "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361.

Geyer, Jakob et al. (2020). "A2D2: Audi Autonomous Driving Dataset". In:

Halevy, Alon, Peter Norvig, and Fernando Pereira (2009). "The unreasonable effectiveness of data". In: *Intelligent Systems. IEEE*.

Huang, Xinyu et al. (2020). "The ApolloScape Open Dataset for Autonomous Driving and Its Application". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.10, pp. 2702–2719.

Neigel, Peter et al. (2020). "OPEDD: Off-road pedestrian detection dataset". In: *Journal of WSCG* 28.1-2, pp. 207–212.

Neuhold, Gerhard et al. (2017). "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes". In: *Proceedings of the IEEE International Conference on Computer Vision* 2017-October, pp. 5000–5009.

Pezzementi, Zachary et al. (2018). "Comparing apples and oranges: Off-road pedestrian detection on the National Robotics Engineering Center agricultural person-detection dataset". In: *Journal of Field Robotics* 35.4, pp. 545–563.

Tabor, T et al. (2015). "People in the weeds: Pedestrian detection goes off-road". In: *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp. 1–7.

Xiang, Yusheng et al. (2020). "KIT MOMA: A Mobile Machines Dataset". In: *ArXiv Preprint*.

Yu, Fisher et al. (2020). "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning". In:

ZED (2020). URL: https://www.stereolabs.com/zed/ (visited on 12/18/2020).

Zhu, Xiangxin et al. (2015). "Do We Need More Training Data?" In: *International Journal of Computer Vision (IJCV)*, pp. 1–17.