A RUGD Dataset for Autonomous Navigation and Visual Perception in Unstructured Outdoor Environments

Maggie Wigness[1], Sungmin Eum[1,2], John G. Rogers III[1], David Han[1] and Heesung Kwon[1]

***Abstract*— Research in autonomous driving has benefited from a number of visual datasets collected from mobile platforms, leading to improved visual perception, greater scene understanding, and ultimately higher intelligence. However, this set of existing data collectively represents only highly structured, urban environments. Operation in unstructured environments, e.g., humanitarian assistance and disaster relief or off-road navigation, bears little resemblance to these existing data. To address this gap, we introduce the Robot Unstructured Ground Driving (RUGD) dataset with video sequences captured from a small, unmanned mobile robot traversing in unstructured environments. Most notably, this data differs from existing autonomous driving benchmark data in that it contains significantly more terrain types, irregular class boundaries, minimal structured markings, and presents challenging visual properties often experienced in off road navigation, e.g., blurred frames. Over $7,000$ frames of pixel-wise annotation are included with this dataset, and we perform an initial benchmark using state-of-the-art semantic segmentation architectures to demonstrate the unique challenges this data introduces as it relates to navigation tasks.**

I. Introduction

The curation of large labeled visual benchmark datasets [30], [52], [5], [19], [53] has helped advance state-of-the-art in recognition [8], [37], detection [28] and semantic segmentation [51], [47]. The success of these techniques has made visual perception onboard autonomous vehicles a primary source of information when making navigation decisions. So much so, that numerous benchmark datasets tailored specifically to autonomous navigation have emerged [2], [4], [6], [9], [20], [50]. This data features real-world scenarios where vehicles must interact with pedestrians, bicycles, and other vehicles while negotiating the roadways in urban cities. These datasets are one of many essential tools to help evaluate progress towards releasing a safe and reliable product.

While substantial developments have been made in autonomous driving technologies, current state-of-the-art is limited to driving on well paved and clearly outlined roadways. However, environments encountered in applications such as Humanitarian Assistance and Disaster Relief (HADR) [21], [22], agricultural robotics [32], [16], environmental surveying in hazardous areas, and humanitarian demining [14], lack the structure and well identifiable features commonly representative of urban cities. Operation in these unstructured, off-road driving scenarios requires a visual perception system to semantically segment scenes with highly irregular class boundaries, as seen in Figure 1, to precisely localize and recognize terrain, vegetation, manmade structures, debris and other hazards to assess traversability and safety.

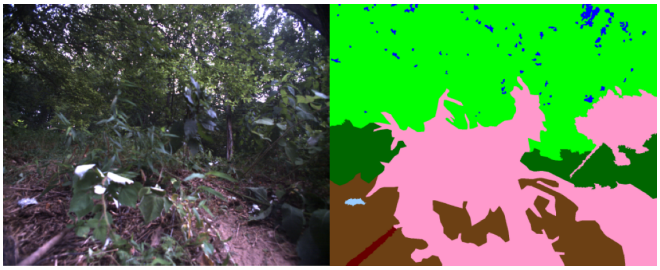


Fig. 1. Irregular boundaries are common in unstructured environments as seen in this example image and annotation. This challenging property exists throughout all sequences in the RUGD dataset and arises because of the natural growth and pourous-like texture of vegetation and occlusion. The color legend from Figure 2 is used for this annotation.

Although there have been some efforts in vision-based autonomous navigation in unstructured environments, the scope of these works is quite narrow. For example, many works define "unstructured" environments simply as a roadway without markings [39], [13], [24]. Defense Advanced Research Projects Agency (DARPA) funded projects like BigDog [46] and the LAGR program [10] operated in highly unstructured environments, but the navigation task focused only on binary classification of terrain, i.e., traversable vs. non-traversable. For more complex learning tasks, finer-grained semantic understanding of the environment is necessary such as learning traversal costs for multiple semantic concepts using inverse reinforcement learning [36], [43].

Because existing benchmark datasets for self-driving vehicles are collected from urban cities, they seldom contain elements that may be present in unstructured environments. As such, there is a need for an entirely new dataset relevant to these driving conditions, where structured cues cannot be readily relied on by the autonomous system. This paper presents the ***Robot Unstructured Ground Driving (RUGD)*** dataset, which is composed of a set of video sequences collected from a small unmanned ground robot performing an exploration task in a variety of natural, unstructured environments and semi-urban areas. In addition to the raw frame sequences, RUGD contains dense pixel-wise annotations (examples seen in Figure 2) for every fifth frame in a sequence, providing a total of $7,453$ labeled images for learning and evaluation in this new driving scenario. The RUGD data is

[1]CCDC Army Research Laboratory (ARL), Adelphi, MD 20783, USA. {maggie.b.wigness.civ,john.g.rogers59.civ}@mail.mil, {david.k.han.civ,heesung.kwon.civ}@mail.mil

[2]Booz Allen Hamilton Inc., McLean, VA 22102, USA. Eum_Sungmin@bah.com

publicly available for download at rugd.vision.

The unique characteristics, and thus, major contributions, of the RUGD dataset can be summarized as follows: 1) The majority of scenes contain no discernible geometric edges or vanishing points, and semantic boundaries are highly irregular. 2) Robot exploration traversal creates irregular routes that do not adhere to a single terrain type; eight distinct terrains are traversed. 3) Unique frame viewpoints are encountered due to sloped terrains and vegetation occlusion. 4) Off-road traversal of rough terrain results in challenging frame focus. These characteristics provide realistic off-road driving scenarios that present a variety of new challenges to the research community in addition to existing ones, e.g., illumination changes and harsh shadows.

Additionally, we provide an initial semantic segmentation benchmark using several current state-of-the-art semantic segmentation architectures [51], [47] originally developed for indoor and urban scene segmentation. Our evaluation results indicate that RUGD presents many challenges for current technologies, highlighting the need for additional research and computer vision algorithms capable of understanding scenes in unstructured environments for autonomous navigation.

II. Related Work

There are a number of visual benchmark datasets available to the autonomous driving community. We discuss the major differences between what is currently available and the novel conditions that RUGD provides. We also further outline relevant real-world applications and tasks that could be advanced with the use of the RUGD dataset.

A. Existing Datasets

Advances in semantic scene parsing have become particularly important for autonomous driving applications, where visual perception of the scene is required to make correct and safe navigation decisions. This has ultimately led to video benchmark datasets including KITTI [6], CityScapes [4], DeepDrive [50], ApolloScape [9], Oxford RobotCar [20] and CamVid [2] that represent urban environments that a self-driving vehicle may be expected to operate within. These datasets are collected from camera sensors mounted on a moving vehicle as it travels through city scenes. The Mapillary Vistas dataset [23] also captures visual data from street scenes, but differentiates itself in that data is captured from different devices, e.g., mobile phones and tablets. This provides greater viewing perspectives compared to previous datasets. However, all of these datasets focus specifically on environments that exhibit highly structured scenes, and the emphasis placed on annotating structured cues such as road signs and lane markings make these datasets extremely useful for city-like autonomous driving technologies.

Less structured than the previously mentioned datasets is the IDD [39] dataset of street scenes. The visual data is collected mostly in India, where street scenes have a higher volume of pedestrian traffic, include a more diverse set of vehicle types, and roads are much less structured. Nevertheless, this data is still designed for autonomous navigation tasks focused on driving on roads, while avoiding other vehicles and a large population of pedestrians.

There is a clear gap in the existing datasets for applications that require visual perception in unstructured, off-road environments. The challenges of debris, e.g., logs or rocks, water hazards, and lack of structural cues are virtually non-existant in these datasets, making them less reliable for applications that often exhibit these off-road navigation scenarios. RUGD provides these challenging scenario characteristics, and with the densely annotated ground truth can help push the state of autonomous driving to more complex, unstructured environments.

B. Relevant Applications

There is a large research interest in determining terrain traversability from camera sensors onboard robots [29], [25], [34], [12], [38], [46], [26]. Most of these works focus on establishing a binary classification of the environment, traversable vs non-traversable. The large set of ground truth annotation associated with RUGD would provide greater higher level semantic understanding of the environment, which would allow different platforms to make more sophisticated traversal decisions given their capabilities.

The continuous video sequences of the ground vehicle motion in RUGD also serves well as training data for developing autonomous vehicle control algorithms. Since RUGD video footage was captured while the vehicle was remotely controlled by a human operator, it can serve as training data in the context of imitation learning or inverse reinforcement learning. The temporal consistency of frames in the video sequences also makes RUGD highly relevant to both supervised [15], [11] and unsupervised [48], [42] semantic video segmentation.

We focus solely on providing an initial benchmark for supervised semantic segmentation on RUGD, but there are a number of other relevant learning approaches and tasks that could benefit from our dataset. The unique qualities of the RUGD dataset obviously leads to a data sparsity issue, in that the unstructured nature of the environments may not represent a large number of consistent or reoccurring patterns. This is particularly relevant and challenging if segmentation algorithms rely on deep learning techniques. A discussion specific to the class sparsity that exists in RUGD is provided later in Section IV.

The small sample set problem that exists in RUGD will further promote in-depth research investigation into alleviating the data sparsity issue. Techniques such as data augmentation by synthetic images using simulations [27], [31], [41], [45] or Generative Adversarial Networks (GANs) for the domain transfer between synthetic and real images [7], [35] could be enhanced and evaluated using RUGD. One/few-shot learning could also be used to address the data scarcity issue, using approaches based on features that are invariant between previously seen categories and novel classes [1], [40], or considering structural composition of objects and their similarities between previously seen versus

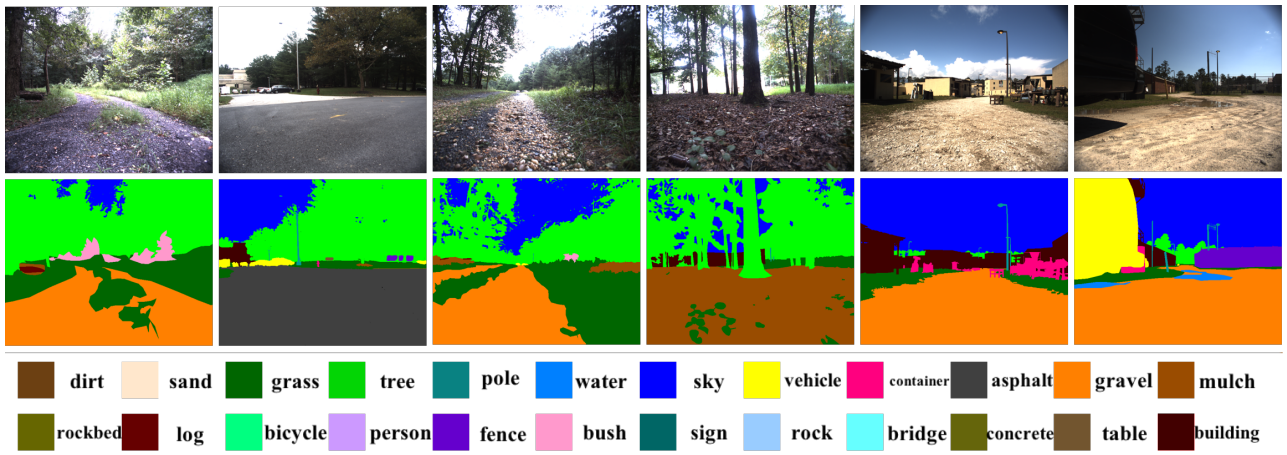


Fig. 2. Example ground truth annotations provided in the RUGD dataset. Frames from the video sequences are densely annotated with pixel-wise labels from 24 different visual classes.

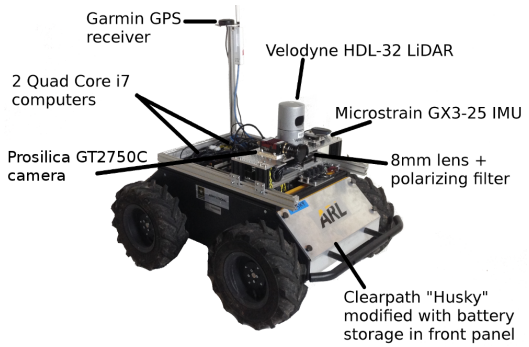


Fig. 3. Robot used to collect the RUGD dataset.

novel categories [17], [44]. Results from these approaches have shown some promising results for dealing with small training set problems.

III. Data Collection

A. Sensors and Robot

The robot used to collect the RUGD dataset is based on a Clearpath 'Husky' platform, seen in Figure 3. The robot chassis is equipped with a sensor payload consisting of a Velodyne HDL-32 LiDAR, a Garmin GPS receiver, a Microstrain GX3-25 IMU, and a Prosilica GT2750C camera. The camera is capable of 6.1 megapixels at 19.8 frames per second, but most sequences are collected at half resolution and approximately 15 frames per second. This reduction in resolution and frame rate provides a compromise between file size and quality. The images are collected through an 8mm lens with a wide depth of field for outdoor lighting conditions, with exposure and gain settings for minimized motion blur. The robot typically operated at its top speed of 1.0 m/s.

This platform provides a unique camera viewpoint compared to existing data. The Husky is significantly smaller than vehicles commonly used in urban environments, with external dimensions of 990 x 670 x 390 mm. The camera sensor is mounted on the front of the platform just above the external height of the robot, resulting in an environment viewpoint from less than 25 centimeters off the ground.

B. RUGD Video Overview

RUGD is a collection of robot exploration video sequences captured as a human teleoperates the robot in the environment. The exploration task is defined such that the human operator maneuvers the robot to mimic autonomous behavior aimed at trying to visually observe different regions of the environment. Given this definition, the sequences depict the robot traversing not only on what may be commonly defined as a road, but also through vegetation, over small obstacles, and other terrain present in the area. The average duration of a video sequence is just over 3 minutes.

Exploration traversals are performed in areas that represent four general environment categories:

- creek - areas near a body of water with some vegetation
- park - woodsy areas with buildings and paved roads
- trail - areas representing non-paved, gravel terrain in woods
- village - areas with buildings and limited paved roads

Example images from these environment categories can be seen in Figure 4. Videos of each traversal are captured from the robot's onboard camera sensor at a frame rate of 15Hz except for *village*, which is captured at only 1Hz, with frame resolution 1376x1110. In addition to showing the numerous terrain types captured in this dataset, notice that Figure 4 also highlights many of the challenging visual properties in this dataset. This includes harsh shadows, blurred frames, illumination changes, washed out regions, orientation changes caused by sloped planes, and occluded perspectives caused by traversal through regions of tall vegetation. Inclusion of such scenes ensures that the dataset closely reflects realistic conditions observed in these environments.

Currently, the RUGD dataset is skewed to represent the trail environment. This skew is representative of data collection challenges that are often faced in real-world robot

Fig. 4. Example frames from robot exploration traversal videos from each of the four environment categories.

exploration applications. That is, it is difficult to capture and annotate data a priori from all possible environments a mobile robot may be deployed. Thus, available environments must be used to collect as diverse a dataset as possible to prepare for whatever the robot may encounter during operation. As discussed later in Section V, we withhold the creek environment video for testing only to represent the scenario when a novel environment is encountered.

IV. Annotation Ontology and Statistics

For an autonomous robot to traverse in an unknown unstructured environment, it needs to detect, recognize, and avoid obstacles while also steering itself toward desirable terrains without getting damaged or bogged down. It also needs to remain stable and has to avoid potential collisions with dynamic entities such as humans, animals or other vehicles. With these objectives in mind, the ontology for the RUGD dataset focuses on terrain and objects that will dictate autonomous agent maneuver. Terrain and ground cover observed in the RUGD dataset are divided into ten categories including dirt, sand, grass, water, asphalt, gravel, mulch, rock bed, bush, and concrete. Factors such as transit impedance, vehicle stability, and water hazards should all be considered as various autonomous platforms may have different sensitivity to these ground conditions. Tracked vehicles may traverse through most of the categories, while wheeled vehicles may have difficulties in sand or wet soil. Although the image segmentations may be useful for binary decisions of trafficability, more nuanced valuation of terrain may be required in the context of disaster relief or military operation. Given several choices of terrain ahead, the vehicle may take into account speed, energy, exposure to enemy fire, slippage, incline, roughness, etc. in assigning overall values in its trajectory planning. Due to these different concerns, we need to differentiate as much of the elements in the field of view to allow development of appropriate autonomy.

The set of objects used in the ontology are both common object categories found in other benchmark datasets considered to be relevant to autonomous navigation decision making, e.g., fence, pole, sign, and tree, and also objects such as vehicles, bicycles, tables, and buildings that may

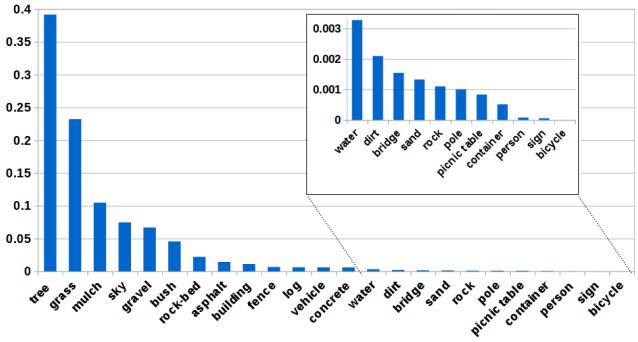


Fig. 5. Overall percentage breakdown of class annotations in the 18 video sequences from the RUGD dataset.

enable the robot to predict human or animal activities nearby to avoid collisions. Although other object categories are sparsely seen in this dataset, label annotation is restricted to the label set outlined in Figure 2.

As this is the first dataset of unstructured, off-road driving scenarios, we seek to collect an initial set of high quality ground truth annotations for the community. Pixel-wise labels are collected with the training data annotation platform provided by Figure Eight, an Appen company[1]. Annotators are asked to label regions in a frame from a video sequence according to the defined ontology. To minimize the effect of inconsistencies between annotators, a verification stage was employed, where labeled frames from a surrounding time window in the video could be viewed to determine if similar regions in the environment were labeled consistently.

RUGD includes a dense set of ground truth labels for every fifth frame of a video sequence, yielding a total of $7,456$ annotated frames. Thus, the full large scale dataset includes over $37,000$ images, where $\sim 20\%$ of those images include ground truth annotations. Example ground truth annotations can be seen in Figure 2.

Further, Figure 5 provides the overall class annotation statistics for the entire RUGD dataset. It is clear from the distribution breakdown that many of the classes are sparsely

[1]https://www.figure-eight.com/

	C	P1	P2	P8	T	T3	T4	T5	T6	T7	T9	T10	T11	T12	T13	T14	T15	V	Total	%
train			x		x	x	x		x		x	x	x	x		x	x	x	4759	63.83
val				x				x											733	9.83
test	x	x								x					x				1964	26.34

TABLE I

Training, validation and testing splits. VIDEO SEQUENCES (C: CREEK, P: PARK, T: TRAIL, V: VILLAGE) THAT FALL INTO EACH SPLIT ARE DENOTED WITH AN X. THE TOTAL NUMBER OF FRAMES AND PERCENTAGE OF EACH SPLIT WITHIN THE WHOLE DATASET ARE SEEN IN THE LAST TWO COLUMNS.

represented in the ground truth annotation (as noted by the need for an inset to better visualize some of the classes). This is in part due to the complete absence of labels in many videos for labels such as rock bed, bicycle, bridge and picnic table. In other cases, visual concepts may simply be sparsely represented in all videos. These skewed annotation statistics further support the discussion in II-B regarding RUGD providing interesting scenarios for research focused on learning with sparse data.

V. BENCHMARKS

Baseline Approaches. We have selected a number of state-of-the-art semantic segmentation approaches and trained/evaluated them on the RUGD dataset. For each semantic segmentation approach we use a fixed encoder, ResNet50 [8], with and without the dilated convolution of size 8. Thus, the baseline approaches differ in their decoder portion of the network. We have selected Pyramid Scene Parsing Network (PSPNet) [51] (Winner of ImageNet Scene Parsing Challenge 2016) and Unified Perceptual Parsing Network (UPerNet) [47] as the main baseline approaches which demonstrate some of the best performance on other large-scale semantic segmentation benchmark datasets such as ADE20K [53] or CityScapes [4]. On top of different variations of PSPNet and UPerNet, we have also included an additional approach (denoted as m1 in Tables II, III) where the decoder does not use either the pyramid pooling module (used in m2, m3) but contains other beneficial components such as bilinear upsampling and deep supervision. 'Deep supervision' is a technique proposed by [51] where an auxiliary loss is imposed in addition to the original softmax loss which is to train the final classifier. This technique decomposes the network optimization into two easier problems to solve, eventually improving the performance.

PSPNet. Zhao et al.[51] mainly focused on incorporating effective global priors on top of the local cues, and introduced the "pyramid pooling module (PPM)", noting that Fully Convolutional Network (FCN) [33] based approaches do not contain suitable techniques for utilizing global clues. PPM pools features in several different pyramid scales, upsamples them using bilinear interpolation to match the original feature map. These maps are then concatenated to be fed into a convolutional layer which makes the final prediction map. In our experiments, we follow [51] and make use of 4 different scales for PPM. We use down-sampling rate of 8. We set 8 as the dilated convolution scale.

UPerNet. This approach is based on the Feature Pyramid Network (FPN) [18] which is devised to efficiently learn and make use of semantically strong, multi-scale features. This top-down architecture with lateral connections creates high-level feature maps across all scales. In UperNet, FPN and PPM function together as PPM is applied on the last layer of the backbone network (in our case ResNet50), before the feature is fed into the top-down hierarchy in FPN. Note that the dilated convolution [49] which has become the *de facto* model for semantic segmentation is omitted in UperNet as it presents several drawbacks including computational complexity. For PPM in our UPerNet, we use down-sampling rate of 4.

Experimental Setup. Videos from the RUGD dataset are partitioned into train/val/test splits for our benchmark experimental evaluation. Table I lists which videos belong to each data split, with $\sim$64% of the total annotated frames used for training, $\sim$10% for validation, and the remaining $\sim$26% for testing. While the selection of videos for each split was decided to roughly produce specific sizes of each split, two videos were specifically placed in the test split to test realistic challenges faced in many motivating applications. First, the *creek* sequence is the only example with significant *rock bed* terrain. Reserving this as a testing video demonstrates how existing architectures are able to learn from sparse label instances. This is a highly realistic scenario in unstructured environment applications as it is difficult to collect large amounts of training data for all possible terrain a priori. Second, the *train-7* sequence represents significantly more off-road jitter than others, producing many frames that appear to be quite blurry. This property is also present in training sequences, but again we reserve the difficult video to determine how well the techniques are able to perform under these harsh conditions.

Training for all baseline models was performed using a PC with 4 NVIDIA Pascal Titan Xp GPUs, on the train set. The val set is used to evaluate the model performance as training progresses. Image resolution has been downsampled by half to 688x555. Batch size is fixed as 8 (2 per GPU) and we use the "poly" learning rate policy following [51], [3]. Base learning rate and power is set to 0.01 and 0.9, respectively. The iteration number has been set to 80K. Momentum and weight decay are set as 0.9 and 0.001, respectively.

Performance of the models are measured using recognized standard semantic segmentation metrics [33]: mean Intersection-over-Union (Mean IoU) and pixel-wise classification accuracy. The IoU for each class is computed as *TP/(TP+FP+FN)*, where *TP, FP, FN* each represent true positive, false positive, and false negative, respectively. Mean IoU is obtained by averaging the IoUs over all 24 semantic

Architecture			Mean	Pixel	Mean Pix.
Method	Encoder	Decoder	IoU [%]	Acc. [%]	Acc. [%]
m1	ResNet50+Dilated conv(8)	1conv, bilinear upsample,deep supervision	35.66	89.93	52.36
m2	ResNet50+Dilated conv(8)	PSPNet [51]	36.27	90.05	53.27
m3	ResNet50+Dilated conv(8)	PSPNet [51] + deep supervision	36.67	90.15	54.64
m4	ResNet50	UperNet [47]	35.87	90.36	52.06

TABLE II

Performance Evaluation on Validation Set. OVERALL SCORE IS THE AVERAGE OF MEAN IOU AND PIXEL ACCURACY.

Architecture			Mean	Pixel	Mean Pix.
Method	Encoder	Decoder	IoU [%]	Acc. [%]	Acc. [%]
m1	ResNet50+Dilated conv(8)	1conv, bilinear upsample,deep supervision	32.24	75.36	52.62
m2	ResNet50+Dilated conv(8)	PSPNet [51]	32.07	75.74	52.14
m3	ResNet50+Dilated conv(8)	PSPNet [51] + deep supervision	31.78	75.03	52.96
m4	ResNet50	UperNet [47]	31.95	75.85	50.72

TABLE III

Performance Evaluation on Test Set. OVERALL SCORE IS THE AVERAGE OF MEAN IOU AND PIXEL ACCURACY.

categories. Pixel-wise classification accuracy (denoted as pixel accuracy in the Tables) is the percentage of correctly classified pixels across all categories. We also present mean per class pixel accuracy (denoted as mean pixel accuracy in the Tables), which is the average pixel accuracy for each semantic category. Because the class distribution (Figure 5) in the RUGD data is unbalanced, the mean pixel accuracy provides an evaluation criteria that evenly weights each class.

Experimental Evaluation.

Semantic segmentation performances on the val and test sets for all methods are reported in Table II and Table III, respectively. On the whole, the baseline benchmark provides a great deal of opportunity for improvements on the RUGD dataset. Although the pixel accuracy on the validation split appears to be quite high, we see a significant performance degrade on the mean per-class accuracy. This suggests that the baseline techniques do a reasonable job learning visual classes that are highly represented in the dataset. This is further supported by the per class pixel accuracy results that are shown in Table IV, where tree and grass (the two most common classes in RUGD) are among the classes that are learned by each of the techniques. Again, the sparsity of some classes in the RUGD dataset is a difficult challenge, yet the identification of these classes can be of significant importance for autonomous driving applications.

Notice that the pixel accuracy achieved on the test set in Table III is also significantly lower than that observed on the validation split. As mentioned previously, we thought this would be the case because of the deliberate inclusion of some challenging video sequences in the test set, i.e., blurry frames from *trail-7* and rare terrain from *creek*. However, each method correctly classifies most of the rock bed pixels in the test set, as seen in Table IV. Thus, it is likely that visual properties associated with off-road navigation are a major challenge faced by these existing architectures.

Although the pixel accuracy of the rock bed class in the test set was very high, this class also had some of the lowest IoU scores produced by each of the methods. The detailed breakdown of IoU per class for the testing split is shown in Table V. IoU is particularly a critical performance metric in applications toward autonomous driving since it may translate to a steering error. For a simple case of a blob of terrain being segmented as a traversable region, the Mean IoU values listed in Table 4 means a large directional error.

Finally, Figure 6 shows example segmentation output for each of the baseline models. Qualitatively, these results highlight some of the more challenging areas of the RUGD dataset. Specifically, the models have a difficult time identifying the exact boundaries between terrains, where classes like grass, tree, bush and mulch do not exhibit structured edges.

VI. CONCLUSION AND FUTURE WORK

RUGD is a unique dataset presenting realistic imagery captured from a vehicle driven in off-road terrains. Compared to other benchmark data for autonomous driving, RUGD images and video clips contain significantly greater variety of terrain types. Unlike other image datasets, most of the scenes contain no discernible geometrical structures, and semantic boundaries are highly convoluted. Capturing images from a moving vehicle in natural off-road environments resulted in images of unfavorable lighting conditions and focus. Overall, it presents a variety of new challenges to the community and may prompt innovative algorithmic solutions for better understanding of scenes in unstructured or natural environments. With pixel level semantic segmentation and accompanying video clips, the dataset would attract interest particularly from those in autonomous off-road driving and natural scene recognition and understanding. Our future plan includes additional collection of imagery data in a variety of natural environments and disaster areas. Since RUGD currently contains only a small sample of off-road environments, we will also investigate effectiveness of the latest techniques designed for sparse dataset learning.

approaches	dirt	sand	grass	tree	pole	water	sky	vehicle	container	asphalt	gravel	building	mulch	rock bed	log	bicycle	person	fence	bush	sign	rock	bridge	concrete	picnic table
m1	4.54	62.07	78.33	82.60	62.64	95.94	89.11	61.90	2.12	17.85	35.23	84.28	52.58	97.76	71.75	-	0.0	75.30	80.64	72.79	17.64	0.0	93.01	77.47
m2	1.37	61.42	76.37	84.71	61.26	93.21	89.10	68.34	5.68	19.58	35.26	83.86	51.63	97.89	64.29	-	0.25	72.94	83.29	66.00	14.28	0.0	92.57	80.22
m3	10.07	66.29	76.21	81.32	74.22	70.24	90.21	76.34	3.67	7.17	38.79	85.62	52.71	97.53	76.15	-	0.0	72.59	79.47	72.85	19.82	0.0	93.44	79.33
m4	4.96	34.10	76.04	86.09	61.19	80.57	89.49	70.00	6.58	16.95	39.47	87.66	48.11	97.97	73.38	-	0.0	71.77	79.07	63.98	12.14	0.0	90.59	79.78

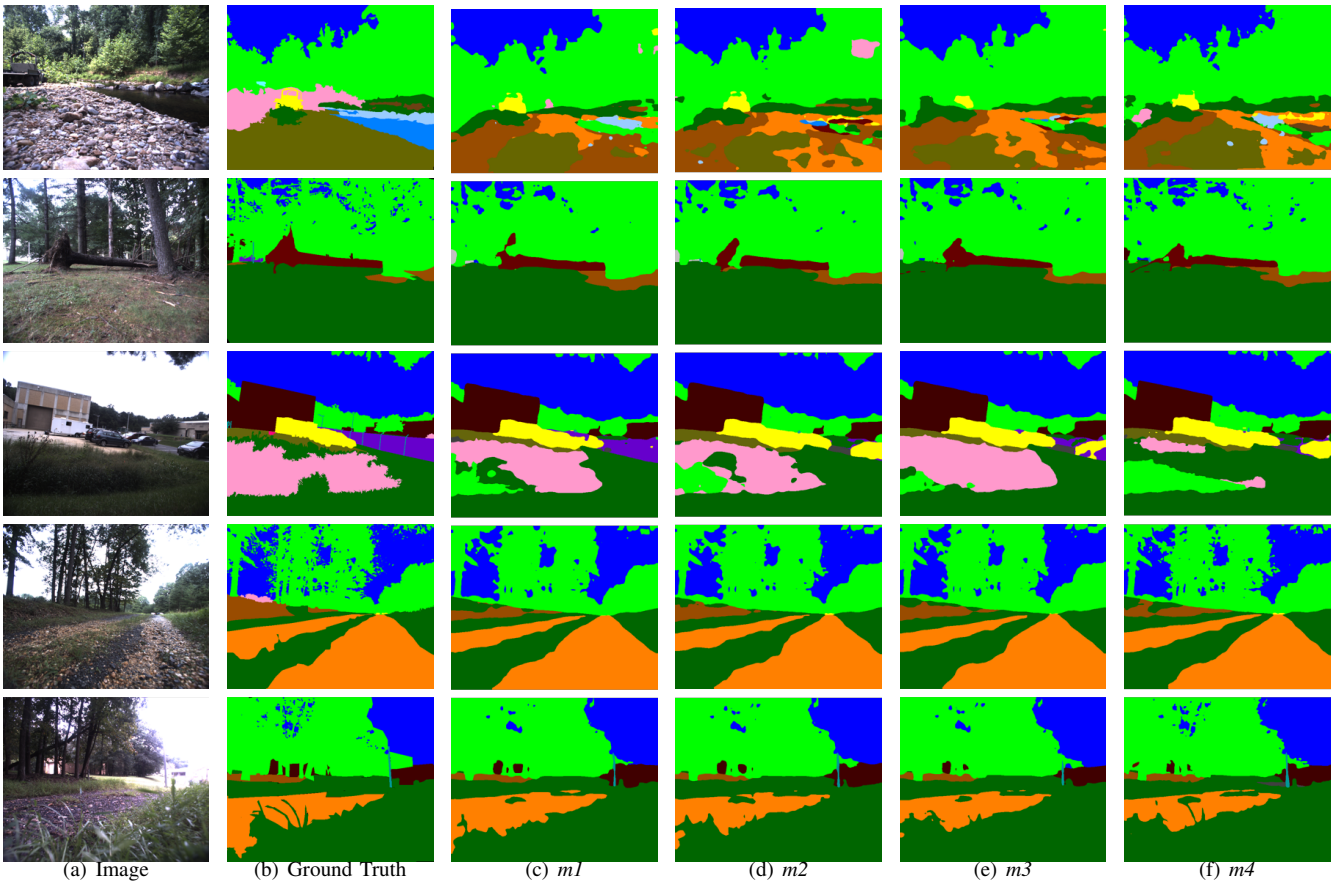
TABLE IV

Per class pixel accuracy [%] on the test split. EACH BASELINE APPROACH IS LABELED WITH *m1-m4* WHERE EACH APPROACH CORRESPONDS TO THE APPROACHES SHOWN IN TABLE II AND III. THERE ARE NO 'BICYCLE' PIXELS IN THE TEST SPLIT, THUS '-' IS USED.

approaches	dirt	sand	grass	tree	pole	water	sky	vehicle	container	asphalt	gravel	building	mulch	rock bed	log	bicycle	person	fence	bush	sign	rock	bridge	concrete	picnic table
m1	0.48	40.17	73.31	79.76	15.57	5.30	79.16	54.47	1.10	12.47	33.94	73.44	49.71	9.77	43.13	-	0.0	52.53	20.93	7.70	12.51	0.0	84.71	55.96
m2	0.03	38.68	71.97	81.60	16.00	4.39	79.27	60.97	1.60	12.71	33.96	73.28	48.64	11.76	40.21	-	0.22	50.41	22.75	6.14	8.89	0.0	84.80	53.38
m3	0.18	43.06	71.93	78.85	15.20	2.47	79.58	64.35	0.59	5.48	37.25	73.62	49.32	8.86	41.50	-	0.0	50.58	19.05	4.60	13.29	0.0	80.11	54.72
m4	0.21	26.80	71.63	82.53	19.18	1.77	81.12	60.29	1.77	12.62	37.70	72.71	45.22	9.61	41.55	-	0.0	45.36	27.02	10.81	8.95	0.0	83.33	58.56

TABLE V

Per class IoU [%] on the test split. EACH BASELINE APPROACH IS LABELED WITH *m1-m4* WHERE EACH APPROACH CORRESPONDS TO THE APPROACHES SHOWN IN TABLE II AND III.



(a) Image (b) Ground Truth (c) *m1* (d) *m2* (e) *m3* (f) *m4*

Fig. 6. **Sample semantic segmentation results using baseline approaches.** The annotation results follow the color code in Table 2. m1-m4 correspond to baseline approaches in Table II and III.

ACKNOWLEDGMENTS

The authors would like to thank Julia Donlon and Matthew Young for their hard work and assistance in the initial data collection of RUGD.

REFERENCES

[1] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 672–679, 2005.

[2] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *Trans. on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[4] 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 *Conf. on Computer Vision and Pattern Recognition*, pages 3213–3223. IEEE, 2016.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and

A. Zisserman. The pascal visual object classes (voc) challenge. *Int. Journal of Computer Vision*, 88(2):303–338, June 2010.

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *Int. Journal of Robotics Research*, 2013.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conf. on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016.

[9] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang. The apolloscape dataset for autonomous driving. In *Conf. on Computer Vision and Pattern Recognition Workshops*, pages 954–960. IEEE, 2018.

[10] L. D. Jackel, E. Krotkov, M. Perschbacher, J. Pippine, and C. Sullivan. The darpa lagr program: Goals, challenges, methodology, and phase i results. *Journal of Field robotics*, 23(11-12):945–973, 2006.

[11] X. Jin, X. Li, H. Xiao, X. Shen, Z. Lin, J. Yang, Y. Chen, J. Dong, L. Liu, Z. Jie, et al. Video scene parsing with predictive feature learning. In *Int. Conf. on Computer Vision*, pages 5581–5589, 2017.

[12] D. Kim, J. Sun, S. M. Oh, J. M. Rehg, and A. F. Bobick. Traversability classification using unsupervised on-line visual learning for outdoor robot navigation. In *Int. Conf. on Robotics and Automation*, pages 518–525. IEEE, 2006.

[13] S. Kolski, D. Ferguson, M. Bellino, and R. Siegwart. Autonomous driving in structured and unstructured environments. In *Intelligent Vehicles Symposium*. IEEE, 2006.

[14] P. Kopacek. Robots for humanitarian demining. In *Advances in Automatic Control*, pages 159–172. Springer, 2004.

[15] A. Kundu, V. Vineet, and V. Koltun. Feature space optimization for semantic video segmentation. In *Conf. on Computer Vision and Pattern Recognition*. IEEE, June 2016.

[16] H. Kushwaha, J. Sinha, T. Khura, D. K. Kushwaha, U. Ekka, M. Purushottam, and N. Singh. Status and scope of robotics in agriculture. In *Int. Conf. on Emerging Technologies in Agricultural and Food Engineering*, volume 12, page 163, 2016.

[17] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum. One shot learning of simple visual concepts. In *Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[18] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the Conf. on Computer Vision and Pattern Recognition*, volume 00, pages 936–944, July 2017.

[19] 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 *European Conf. on Computer Vision*, pages 740–755. Springer, 2014.

[20] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 year, 1000 km: The oxford robotcar dataset. *The Int. Journal of Robotics Research*, 36(1):3–15, 2017.

[21] R. R. Murphy. *Disaster robotics*. MIT press, 2014.

[22] K. Nagatani, S. Kiribayashi, Y. Okada, K. Otake, K. Yoshida, S. Tadokoro, T. Nishimura, T. Yoshida, E. Koyanagi, M. Fukushima, et al. Emergency response to the nuclear accident at the fukushima daiichi nuclear power plants using mobile rescue robots. *Journal of Field Robotics*, 30(1):44–63, 2013.

[23] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Int. Conf. on Computer Vision*, pages 4990–4999. IEEE, 2017.

[24] T. Ort, L. Paull, and D. Rus. Autonomous vehicle navigation in rural environments without detailed prior maps. In *Int. Conf. on Robotics and Automation*, 2018.

[25] P. Papadakis. Terrain traversability analysis methods for unmanned ground vehicles: A survey. *Engineering Applications of Artificial Intelligence*, 26(4):1373–1385, 2013.

[26] M. J. Procopio, J. Mulligan, and G. Grudic. Learning terrain segmentation with classifier ensembles for autonomous robot navigation in unstructured environments. *Journal of Field Robotics*, 26(2):145–175, 2009.

[27] W. Qiu and A. Yuille. Unrealcv: Connecting computer vision to unreal engine. In *European Conf. on Computer Vision*, pages 909–916. Springer, 2016.

[28] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Conf. on Computer Vision and Pattern Recognition*, 2017.

[29] H. Roncancio, M. Becker, A. Broggi, and S. Cattani. Traversability analysis using terrain mapping and online-trained terrain type classifier. In *Intelligent Vehicles Symposium*, pages 1239–1244. IEEE, 2014.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *Int. Journal of Computer Vision*, 115(3):211–252, 2015.

[31] A. Shafaei, J. J. Little, and M. Schmidt. Play and learn: Using video games to train computer vision models. *arXiv preprint arXiv:1608.01745*, 2016.

[32] R. R. Shamshiri, C. Weltzien, I. A. Hameed, I. J. Yule, T. E. Grift, S. K. Balasundram, L. Pitonakova, D. Ahmad, and G. Chowdhary. Research and development in agricultural robotics: A perspective of digital farming. *Int. Journal of Agricultural and Biological Engineering*, 11(4):1–14, 2018.

[33] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *Trans. on Pattern Analysis and Machine Intelligence*, 39(4):640–651, Apr. 2017.

[34] M. Shneier, T. Chang, T. Hong, W. Shackleford, R. Bostelman, and J. S. Albus. Learning traversability models for autonomous mobile vehicles. *Autonomous Robots*, 24(1):69–86, 2008.

[35] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *Conf. on Computer Vision and Pattern Recognition*. IEEE, 2017.

[36] D. Silver, J. Bagnell, and A. Stentz. High performance outdoor navigation from overhead data using imitation learning. In *Robotics: Science and Systems*, 2008.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Conf. on Computer Vision and Pattern Recognition*, pages 2818–2826.

[38] A. Talukder, R. Manduchi, R. Castano, K. Owens, L. Matthies, A. Castano, and R. Hogg. Autonomous terrain characterisation and modelling for dynamic control of unmanned vehicles. In *Int. Conf. on Intelligent Robots and Systems*, pages 708–713. IEEE, 2002.

[39] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *Winter Conf. on Applications of Computer Vision*, pages 1743–1751. IEEE, 2019.

[40] Y.-X. Wang and M. Hebert. Learning from small sample sets by combining unsupervised meta-training with cnns. In *Advances in Neural Information Processing Systems*, pages 244–252, 2016.

[41] Z. Wang, J. Yang, H. Jin, E. Shechtman, A. Agarwala, J. Brandt, and T. S. Huang. Deepfont: Identify your font from an image. In *Int. Conf. on Multimedia*, pages 451–459. ACM, 2015.

[42] M. Wigness and J. G. Rogers. Unsupervised semantic scene labeling for streaming data. In *Conf. on Computer Vision and Pattern Recognition*, pages 4612–4621. IEEE, 2017.

[43] M. Wigness, J. G. Rogers, and L. E. Navarro-Serment. Robot navigation from human demonstration: Learning control behaviors. In *Int. Conf. on Robotics and Automation*, pages –. IEEE, 2018.

[44] A. Wong and A. L. Yuille. One shot learning via compositions of meaningful patches. In *Int. Conf. on Computer Vision*, pages 1197–1205. IEEE, 2015.

[45] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Symposium on Eye Tracking Research & Applications*, pages 131–138. ACM, 2016.

[46] D. Wooden, M. Malchano, K. Blankespoor, A. Howardy, A. A. Rizzi, and M. Raibert. Autonomous navigation for bigdog. In *Int. Conf. on Robotics and Automation*, pages 4736–4741. IEEE, 2010.

[47] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *European Conf. on Computer Vision*. Springer, 2018.

[48] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *European Conf. on Computer Vision*, pages 626–639. Springer, 2012.

[49] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *Int. Conf. on Learning Representations*, 2016.

[50] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

[51] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Conf. on Computer Vision and Pattern Recognition*.

[52] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Conf. on Neural Information Processing Systems*, pages 487–495, 2014.

[53] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Conf. on Computer Vision and Pattern Recognition*. IEEE, 2017.