

IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments

Girish Varma¹ Anbumani Subramanian² Anoop Nambodiri¹

Manmohan Chandraker³ C V Jawahar¹

¹IIT Hyderabad

²Intel Bangalore

³University of California, San Diego

<http://idd.insaan.iit.ac.in/>

Abstract

While several datasets for autonomous navigation have become available in recent years, they tend to focus on structured driving environments. This usually corresponds to well-delineated infrastructure such as lanes, a small number of well-defined categories for traffic participants, low variation in object or background appearance and strict adherence to traffic rules. We propose IDD, a novel dataset for road scene understanding in unstructured environments where the above assumptions are largely not satisfied. It consists of 10,004 images, finely annotated with 34 classes collected from 182 drive sequences on Indian roads. The label set is expanded in comparison to popular benchmarks such as Cityscapes, to account for new classes. It also reflects label distributions of road scenes significantly different from existing datasets, with most classes displaying greater within-class diversity. Consistent with real driving behaviors, it also identifies new classes such as drivable areas besides the road. We propose a new four-level label hierarchy, which allows varying degrees of complexity and opens up possibilities for new training methods. Our empirical study provides an in-depth analysis of the label characteristics. State-of-the-art methods for semantic segmentation achieve much lower accuracies on our dataset, demonstrating its distinction compared to Cityscapes. Finally, we propose that our dataset is an ideal opportunity for new problems such as domain adaptation, few-shot learning and behavior prediction in road scenes.

1. Introduction

Autonomous navigation is rapidly maturing towards becoming a mainstream technology, with even consumer deployment by major automobile manufacturers. A significant contributor to this progress has been the availability of large-scale datasets for sensing and scene understanding. Yet, several challenges remain in enabling self-driving across



Figure 1. Some examples of the diverse and unstructured conditions that is covered by the dataset.

diverse geographies. A key challenge is to achieve data scale and diversity large enough to ensure safety and reliability in extreme corner cases. Even more importantly, algorithms are largely untested in their ability to generalize to road conditions that are significantly more diverse and unstructured.

In this paper, we propose IDD, a dataset that takes the first steps towards addressing the above concerns. Our dataset shares several traits such as scale, annotation and tasks with similar ones in structured environments, namely KITTI [15] or Cityscapes [5]. But it also intends to significantly expand the scope of the autonomous navigation problem, along each of those dimensions. Similar to Cityscapes [5], we provide large-scale raw data with multiple cameras and sensors across cities and lighting conditions. But the scale of our data is larger, consisting of 10,004 labeled images with fine instance-level boundaries. Next, while the annotation type is similar for our dataset and Cityscapes [5], the number of object classes and within-class diversity of appearance are higher for IDD. Finally, while we also initially propose instance segmentation as the task of interest, our label diversity and novel hierarchy might allow novel machine learning techniques and computer vision algorithms.

The singular defining aspect of our dataset is that it corresponds to driving in less structured environments. We argue that this is a better reflection of the needs for autonomous

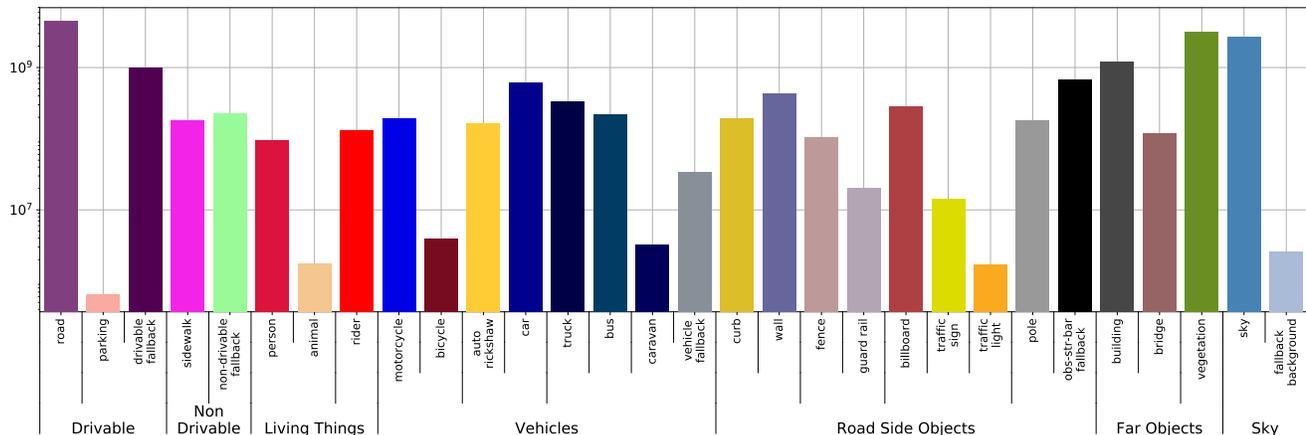


Figure 2. Label distribution in our dataset. The following information is shown here: (i) pixel counts of individual labels on the y-axis (ii) four-level label hierarchy used by the dataset at the bottom, (iv) the color legend for the predicted and ground truth masks shown in the paper is used for the corresponding bars. We define metrics at 4 levels of the hierarchy with 30 (level 4), 26 (level 3), 16 (level 2) and 7 (level 1) labels, respectively, giving different complexity levels for training models.

navigation in large portions of the world, including Asia, South America and Africa. Accordingly, we collect our data in India where road scenes differ markedly from those in Europe or North America. The variety of traffic participants in Indian roads is larger, including novel classes such as autorickshaws or animals. The within-class diversity is also higher, for example, since vehicles span a larger range of manufacturing years and ply with larger variation in wear. Even the distribution of classes that overlap with Cityscapes is significantly different, for instance, the proportion of motorcycles is far higher, as is that of multiple riders on two-wheeled vehicles. Background classes also display greater diversity, such as city scenes rich in novel classes such as billboards. Besides variations in weather and lighting, other ambient factors such as air quality and dust also span greater ranges in our dataset. Such greater complexity of road scenes necessitates a larger scale of data. Thus, we provide high-quality annotation at a scale significantly larger than available for other contemporary datasets such as KITTI or Cityscapes.¹

We provide a detailed analysis of the label distributions in the IDD dataset, while highlighting some of the above differences. We also showcase those differences through quantitative evaluation of state-of-the-art algorithms on our dataset. We consistently observe that semantic segmentation performances are far lower on IDD as compared to Cityscapes, using identical models and with larger-scale training data for IDD. Firstly, this highlights that conventional semantic segmentation datasets such as Cityscapes are getting saturated and the next set of challenges lie in more complex datasets

¹Datasets such as Berkeley Deep Driving [26] and Apolloscape [10] have recently been released with labels at a similar scale. However, they are contemporaneous with our work, which precludes a detailed comparison. In any case, we note that they are in structured environments, which makes our dataset clearly different.

like IDD. Secondly, this highlights the need for ever-larger training data as we expand the scope of the autonomous navigation problem to newer geographies.

Besides segmentation, the nature of our dataset also enables novel problems for vision and learning. This is already reflected in some of our annotation choices. For instance, while the notion of a drivable area in Europe is largely defined by classes such as roads or lanes that have distinct appearances, it is more ambiguous in our dataset and likely also informed by semantic cues such as presence of dynamic traffic participants. Thus, we include labels for safely drivable and non-drivable areas. Our label hierarchy is attuned to the autonomous navigation problem and we postulate that exploiting it might lead to semantic segmentation more suited to subsequent applications such as collision avoidance or path planning. We label classes that are rare but important for navigation (such as animals), or classes that exhibit large within-class variance (such as autorickshaw), which motivates problems such as few-shot learning.

The contrast of our dataset with structured ones also suggests interesting directions of future research. For instance, domain adaptation between Cityscapes and IDD is clearly a need given the large performance drops encountered in cross-dataset settings. Classes that are unique to our dataset also encourage consideration of domain adaptation with non-overlapping label spaces. Even higher-level reasoning problems such as behavior prediction pose new challenges in IDD, since traffic participants have lower adherence to traffic rules, motions can be sudden, complex obstructions might be present, drivable areas can be ambiguous and traffic lanes need not correspond to lane markings on the road. While not considered in this paper, we highlight that these novel problems do arise in unstructured environments such as ours.



Figure 3. Cityscape models do not distinguish between the road and possible unsafe drivable area on both sides of the road.

2. Challenges in Unstructured Environments

We collect data from Indian roads and analyze the shortcomings of models trained on existing datasets. As illustration, we describe some of the qualitative issues observed when using predicted outputs of a model that obtains 70% mean IoU on the Cityscapes validation set.

Ambiguous Road Boundaries. Road boundaries in Cityscapes are very well defined and usually flanked on both sides by barriers or sidewalks. However, this is not the case in our setting. Road sides can have muddy terrain, while also being drivable to some extent. Roads themselves can be covered by dirt or mud, making the boundaries very ambiguous. On the other hand, Cityscapes models often recognize flat areas beside the road which need not be safe for driving as road, as seen in Figure 3.

Diversity of Vehicles and Pedestrians. Indian roads have a variety of unique vehicles like auto-rickshaws, which behave very differently than other vehicles like cars. Even for standard categories like cars, the appearance variations are higher due to greater wear and tear. Further, the frequency and variety of trucks and buses are also high. Another distinction is the large number of motorbikes with multiple persons riding it. Pedestrians often cross the road at arbitrary locations, rather than crosswalks. Bikes and autorickshaws are also less likely to follow traffic discipline, thus, there are fewer correlation between traffic participants and road signage such as lanes or traffic lights.

Extensive Use of Information Boards. Information displays such as billboards appear extensively in our dataset. They can be significant for localization and mapping problems by indicating buildings or landmarks. Sometimes they also indicate special vehicles, such as advertisements attached to a driving school car, or a delivery vehicle.

Diversity of Ambient Conditions. Lighting variation in our dataset is high since we acquire images at various times of the day, including mid-day, dawn and dusk. Also, some of



Figure 4. (Top) A herd of buffaloes on the road at dusk. (Bottom) Many motorbikes with multiple riders, not necessarily following



Figure 5. (Left) An array of billboards indicating the shops. (Right) A vehicle with a billboard of a driving school.

the images have heavy shadows, which are common during a long summer season. Our dataset also contains scenes with heavily clouded skies. The greater variation in particulate matter due to fog, dust or smog also leads to significant appearance variations. Cityscapes pretrained models yield lower accuracies in these settings, as seen in Figure 6.

3. Dataset

3.1. Acquisition

The data was collected from Bangalore and Hyderabad cities in India and their outskirts. The locations have a mix

Dataset	Calibration	Nearby frames / Video	Distortion /Night	#Images/ #Sequences	#Labels Train/Total	Average Resolution
Cityscapes [5]	✓	✓		5K / 50	19/34	2048x1024
IDD	✓	✓		10K / 180	30/34	1678x968
BDD100K [26]		✓	✓	10K / 10K	19/30	1280x720
MVD [16]				25K / -	65/66	>1920x1080

Table 1. Comparison of semantic segmentation datasets for autonomous navigation.



Figure 6. Heavy shadows in the image (top) or low light conditions (bottom) can greatly degrade the quality of predictions using models trained on Cityscapes.

of urban and rural areas, highway, single lane and double lane roads with a variety of traffic. The driving conditions in these localities are highly unstructured due to multiple reasons: (i) these cities are rapidly growing and have a lot of construction near the roads, (ii) road boundaries are not well defined, (iii) pedestrians and jaywalkers are aplenty in these road images, and (iv) high density of motorbikes and trucks on the road. The variety of vehicle models are also very large. A total of 182 drive sequences were used for the preparation of the dataset.

3.2. Frame Selection

We chose images from one of the forward facing cameras of a stereo pair, for fine annotation. Images were sampled at varying rates from the video sequence, with denser sampling around crowded and special interest places like traffic junctions. These images were annotated very finely, by layered polygon masks similar to Cityscapes. Since the road conditions are highly unstructured, we need a wider variety of labels. We annotated a total of 10,004 frames.

3.3. Label Hierarchy & Annotation

We used a total of 34 labels in the fine annotations. The labels were given definition by means of a textual description as well as example images. However, we found that it is difficult to completely avoid ambiguity between some labels. For example labels like parking, caravan or trailer cannot

be precisely defined due to the diversity of the scenes and vehicles in the data collected. For resolving this issue, we designed a 4 level label hierarchy having 7 (level 1), 16 (level 2), 26 (level 3) and 30 (level 4) labels (see Figure 2). Each level defines a category as the union of labels in the succeeding level, which are chosen such that they are ambiguous. Since we take unions of the most ambiguous labels while designing the hierarchy, the lower levels have lesser ambiguity. We have a set of new labels not available in Cityscapes [5] like auto rickshaw, billboards, animal, curb. We also have separate labels for road, drivable fall-back and non-drivable fall-back indicating safe, unsafe and non-drivable flat surfaces. We have added fall-back labels whenever appropriate so that highly ambiguous objects can be given labels.

For labeling the dataset, the annotation team was first asked to re-annotate images from the Cityscapes [5] dataset. The difference between the annotations were subsequently shown to the annotators. This process was done until the annotators were achieving greater than 95% accuracy with respect to the Cityscapes ground truth labels.

3.4. Statistical Analysis and Dataset Splits

The pixel statistics among the labels can be seen in Figure 2. The labels in level 4 have high class imbalance. Labels like parking, animal, caravan or traffic light have much fewer pixels. The annotated dataset also has labels for trailer and rail track, which were combined with vehicle fallback and nondrivable fallback in level 4, since they have very few pixels that mostly fell within a few drive sequences. The lower levels are designed such that the imbalance is lesser.

Class imbalance at level 4, creates a problem while splitting the dataset in train, test and validation sets. Since the splitting is done at the level of drive sequences (that is, all images within a drive sequence are moved to the same split), we need to be careful that the few drive sequences that contains a label are rightly split. We roughly divide the dataset in to 70% train, 10% validation and 20% test splits. The split was done by randomly assigning the drive sequences with the said distribution. We did the splitting multiple times to come up with a split where all the 30 labels in level 4 have approximately 70, 10 and 20 percentage of pixels in the train, validation and test sets, respectively.

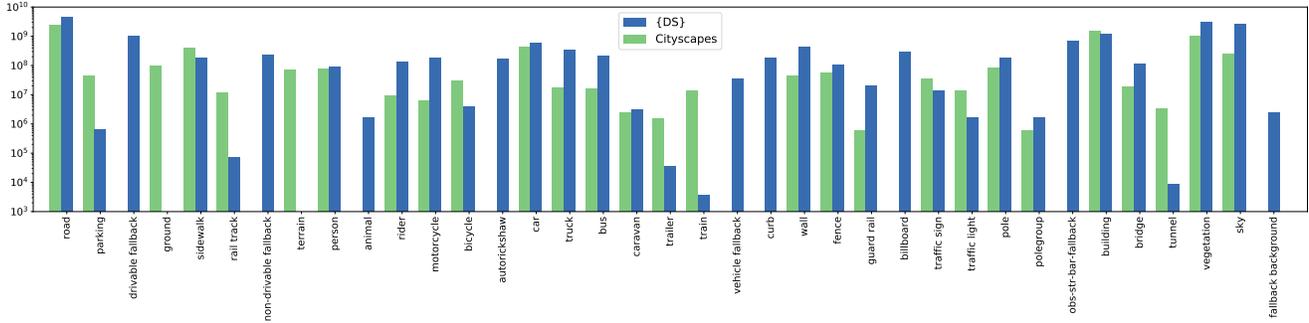


Figure 7. Comparison of the pixel count in our dataset with Cityscapes. The y axis is plotted in log-scale. Note that for most classes of vehicles, the number of pixels are 5-10 times more than Cityscapes. Moreover our dataset has newer labels like autorickshaw, billboard, drivable/nondrivable fallback which also have significant number of labeled pixels.

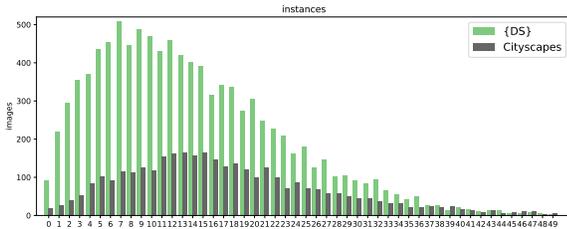


Figure 8. Comparison of traffic participants in our dataset with Cityscapes.

3.5. Comparison with Other Datasets

Various datasets have been proposed for studying the semantic and instance segmentation problems like Pascal VOC [7], MS COCO [13], SUN [23, 22]. The datasets proposed for semantic segmentation that focus on autonomous navigation are Cityscapes [5], KITTI [15], Camvid [2, 1], Leuven [12] and the Daimler Dataset [21]. ADE20K [29] is a recent dataset which focuses on the general scene parsing problem. More recently Mapillary Vistas dataset [16] (which focuses on street view imagery) and the Berkeley Deep Drive Dataset [26] (for autonomous navigation) was released. A comparison of the metadata available in these datasets can be seen in Table 1. As can be seen, our dataset is more similar to Cityscape in the sense that we collect data from calibrated cameras without any distortions. BDD100K uses dashboard cameras kept inside the car and hence often has internal reflections from the glass as well as rain distortions. Moreover a good fraction of the dataset consists of night images. Mapillary Vistas dataset consists of images taken using a variety of cameras (including smart phones) having varying perspectives of the road and the road side. They do not have video data or images of near by frames.

We compare the label statistics of our dataset with Cityscapes (since it is more similar to our dataset as described above) in terms of pixel counts (Figure 7) and the number of instances of traffic participants (Figure 8). We have more pixels of truck, bus, motorcycle, guard rail, bridge and rider (see Figure 7). The pixel counts for new

Method	% mIoU at Hierarchy Levels		
	1	2	3
GT subsampled by 2	99	97	97
GT subsampled by 4	98	96	95
GT subsampled by 8	96	92	90
GT subsampled by 16	92	87	84
GT subsampled by 32	86	78	74
GT subsampled by 64	77	66	61
GT subsampled by 128	65	53	48

Table 2. Control experiments to estimate upper bounds for semantic segmentation results, assessed by Intersection-over-Union (IoU, in %) scores for different levels of the hierarchy.

labels (auto rickshaw, billboard, curb, drivable-fallback, nondrivable-fallback) are also high. In terms of instances of traffic participants, we have almost double the counts, with a distribution similar to Cityscapes (see Figure 8).

4. Benchmarks

4.1. Control Experiments

In Table 2, we provide the results of some control experiments which provide upper bounds for IoU scores for models giving predictions at a given factor of the input resolution. We first downsample the ground truth by a given factor and then upsample it to the original image size for evaluation of average IoU at original scale. We provide the mean IoU scores of different levels of the hierarchy, confirming that low-resolution processing contributes significantly to overall degradation of segmentation results.

4.2. Domain Discrepancy

Domain discrepancy studies the quantitative shift in data distributions between datasets. To understand it, we train a DRN-D-38 (Dilated Residual Networks [25]) model in Cityscapes [5], Mapillary [16], BDD100K [26] and our dataset. We compare the IoU scores in a set of 16 com-

Train	Test	road	sidewalk	person	motorcycle	bicycle	car	truck	bus	wall	fence	traffic sign	traffic light	pole	building	vegetation	sky	mIoU of common labels
CS	DS	72	22	30	47	10	58	30	19	17	13	19	8	23	32	76	68	34
DS	CS	81	26	74	34	55	85	16	17	21	24	25	21	47	77	90	88	49
BD	ID	83	0	38	44	2	52	21	13	0	0	0	0	36	42	83	94	32
ID	BD	84	16	57	34	44	77	14	24	10	33	18	13	41	68	82	87	44
CS	CS	98	84	81	60	76	94	56	78	49	58	77	67	62	92	92	94	76
MV	MV	85	58	73	55	61	90	61	65	45	58	72	67	50	86	90	98	70
ID	ID	92	68	73	80	42	89	79	78	64	45	60	38	58	75	90	97	70
BD	BD	95	62	61	32	22	90	52	57	25	45	52	58	49	85	87	97	60

Table 3. The domain discrepancy between Cityscapes (CS) [5], Mapillary Vistas (MV) [16], Berkeley Deepdrive (BD) [26] Dataset and IDD (ID) using the DRN-D-38 Model [25]. Performance for only the common labels between the four datasets are used. First two rows compares the accuracy of a model trained on one of IDD or Cityscapes and tested on the other dataset. As can be seen, IDD trained model can predict CS and BD labels, better than predictions of trained models of the corresponding datasets on IDD. The bottom four rows gives the performance of models in each of the datasets. IDD dataset is harder than CS dataset and similar in hardness to MV on these 16 labels. BD is harder because i.) it has night scenes ii.) the images are take from a dash board cam, hence has reflections from inside the car as well as distortions like rain drops on the mirror.

mon labels between the four datasets in Table 3. As seen from the last four rows, our dataset is harder than Cityscapes while having a similar level of hardness compared to the Mapillary Vistas Dataset. We also report IoU scores of predictions given by models trained on one dataset and tested on the other. A pretrained model trained on our dataset performs better when tested on Cityscapes and BDD100K, as compared to the converse experiment.

4.3. Semantic Segmentation Benchmark

The semantic segmentation benchmark on our dataset quantifies the mean Intersection over Union (mIoU) scores at the four levels of the hierarchy. There are some labels in level 4 like traffic light, parking or animal for which the number of labeled pixels are very few. Hence, this serves as an excellent benchmark for transfer learning or domain adaptation problems. We also have level 1 and level 2 mIoU scores, which are a useful benchmark for real-time models, since they might not be able give good results on the fine grained classification task at level 3 and 4. The level 1 benchmark still has classes for most of the essential labels for autonomous navigation.

We benchmarked our dataset using the DRN-D-38 ([25]) and the ERFNet (real-time model [19]) model. We also conducted a challenge and evaluated submissions which use some of the state-of-the-art models. The results are shown in Table 4.

4.4. Class IoUs and Confusion Matrix

The IoUs for every class can be found in Figure 9. We observe that the IoUs are lower than 25% for bicycle, traffic light, vehicle fall-back and fence labels. The low scores for

Method	% mIoU at Levels		
	L1	L2	L3
ERFNet	-	-	55.4
DRN-D-38	85.9	72.6	66.6
*DeeplabV3+ [4]	89.8	78.0	74.0
*PSPNet [27]	89.9	78.0	74.1
*Wider Resnet-38, DeeplabV3 Decoder, Inplace ABN [20], Ensemble of 4	89.7	77.9	74.3

Table 4. The mIoU scores of models at 3 level of the hierarchy. The performance numbers of * models are obtained from the submissions of a AutoNUE challenge [11] conducted based on the dataset.

bicycle and traffic light can be explained by the low pixel counts. We also plot the confusion matrix between labels in Figure 10. Note that there is significant confusion between:

- motorcycle and bicycle.
- billboard and traffic sign.
- obs-str-bar-fallback, vegetation and traffic light.
- building and billboard.
- vegetation and wall, pole, fence.
- drivable, non-drivable, vegetation.

We analyze some examples of predictions in Figure 11. As can be seen, model trained on our dataset gives prediction of much better quality in unstructured setting. It identifies the muddy areas which can be driven. New labels like au-

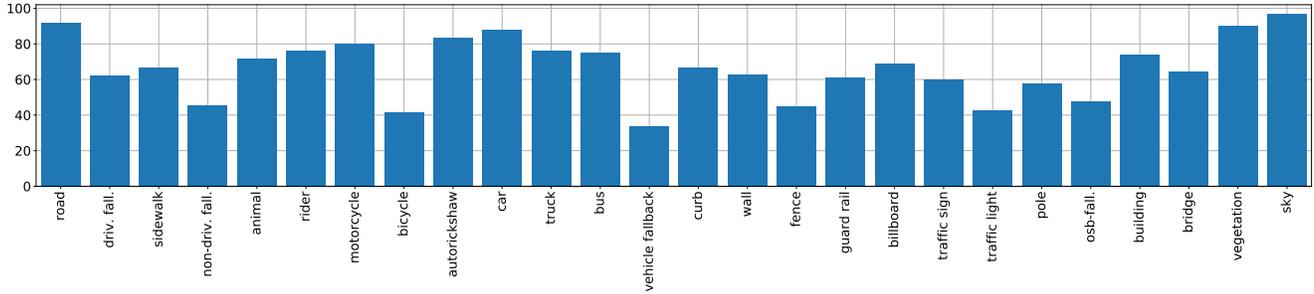


Figure 9. The IoUs for every class for the DRN D 38 model trained on IDDwith mIoU of 66.5%.

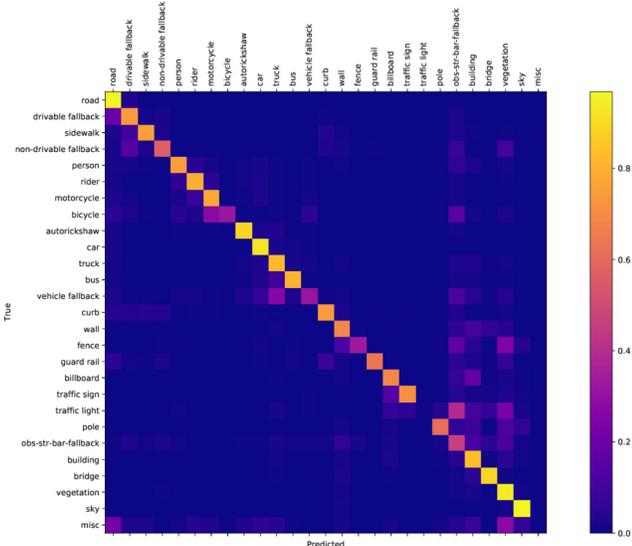


Figure 10. The confusion matrix of the trained model.

Method	AP	AP@50
*MaskRCNN [8] with ResNet101	0.268	0.499
*PANet [14]	0.376	0.661

Table 5. The AP scores of models for instance labels. The performance numbers of * models are obtained from the submissions of the AutoNUE challenge [11] conducted based on the dataset.

torickshaw, curb, billboard etc are getting identified.

4.5. Instance Segmentation Benchmark

Similar to other datasets, we also specified an instance segmentation benchmark, where individual instances of the same label need to be segmented separately. The algorithms are required to predict a set of detections of traffic participants in the frame, with a confidence score and a per-instance binary segmentation mask. To assess the performance, the average precision on the region level for each class and average it across a range of overlap thresholds ranging from 0.5 to 0.95 in steps of 0.05, similar to [13].

The results of some best performing submissions from the challenge are given in Table 5.

5. Conclusion

We present a novel dataset for studying problems of autonomous navigations in unstructured driving conditions. We identify several drawbacks of existing datasets, such as distinguishing safe or unsafe drivable areas beside the road, additional labels required for vehicles and a label hierarchy that reduces ambiguity. We analyze the label statistics and the class imbalance present in the dataset. We also examine the domain discrepancy properties with respect to other semantic segmentation datasets. In contrast to existing datasets on semantic segmentation, ours is acquired in India, which leads to greater diversity due to variations in appearance of traffic participants as well as background categories. Not only does this pose interesting challenges for the state-of-the-art in semantic segmentation, it is also the first effort in our knowledge to focus on problems related to autonomous driving in geographies outside North America or Europe with relatively less developed road infrastructure.

In the future, we plan to extend the benchmark to computer vision problems beyond semantic segmentation. In particular, the unconstrained nature of the dataset provides a uniquely novel setting for higher-level reasoning problems such as scene understanding [9, 17, 28] and path planning [6]. Motions in the dataset are less constrained due to greater freedom in traffic participant behavior and less adherence to traffic rules. The possible absence of visual cues such as lanes that constrain traffic participant behavior poses further challenges. Besides the presence of rare categories [24], even common categories have diverse attributes or appearance variations. Besides, the ambient conditions differ greatly across weather, time of day and air quality. This also motivates interesting new problems for few shot learning [3] and domain adaptation [18], which our future work will study in greater detail.

Acknowledgements. The authors would like to thank Intel, specially the Intel India team for the efforts in capture of the data and coordination. Authors would like to specially acknowledge the support and helps from Bharat Kaul, Prabhavathy Adavikolanu and Silar Shaik in making this possible. We would also like to thank Governments of Telangana and Karnataka for the permissions, encouragement and enabling this effort.

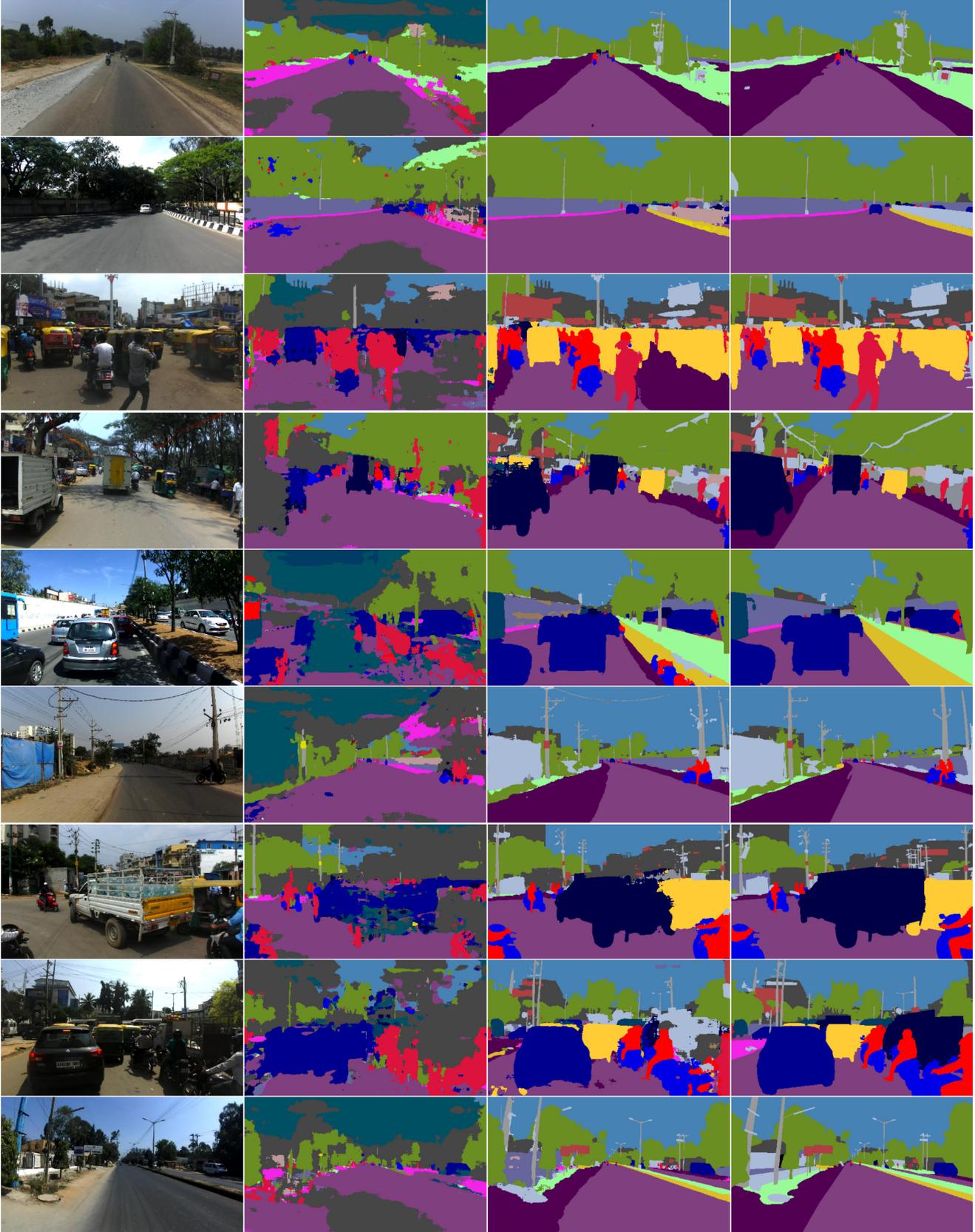


Figure 11. We give many qualitative example with: input image from validation set, predictions from Cityscape pretrained model, prediction from model trained on our training dataset and the ground truth in our dataset in the order of columns.

References

- [1] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2008.
- [2] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008.
- [3] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017.
- [4] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018.
- [5] 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.
- [6] D. Dolgov, S. Thrun, M. Montemerlo, and J. Diebel. Practical search techniques in path planning for autonomous driving. *AAAI*, 2008.
- [7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [8] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.
- [9] D. Hoiem, J. Hays, J. Xiao, and A. Khosla. Guest editorial: Scene understanding. *International Journal of Computer Vision*, 112(2):131–132, Apr 2015.
- [10] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang. The apolloscape dataset for autonomous driving. *arXiv: 1803.06184*, 2018.
- [11] C. Jawahar, A. Subramanian, A. Namboodiri, M. Chandrakar, and S. Ramalingam. Autonomous navigation in unconstrained environments (AutoNUE) workshop and challenge at ECCV’18. <http://cvit.iiit.ac.in/scene-understanding-challenge-2018/>.
- [12] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool. Dynamic 3d scene analysis from a moving vehicle. In *CVPR*, 2007.
- [13] 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 D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *ECCV*, 2014.
- [14] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.
- [15] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015.
- [16] G. Neuhold, T. Ollmann, S. R. Bul, and P. Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.
- [17] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.
- [18] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, May 2015.
- [19] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, Jan 2018.
- [20] S. Rota Bulò, L. Porzi, and P. Kotschieder. In-place activated batchnorm for memory-optimized training of dnns. In *CVPR*, 2018.
- [21] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth. Efficient multi-cue scene segmentation. In J. Weickert, M. Hein, and B. Schiele, editors, *Pattern Recognition*, pages 435–445, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- [22] S. Song, S. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015.
- [23] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.
- [24] J. Yang, B. Price, S. Cohen, and M. Yang. Context driven scene parsing with attention to rare classes. In *CVPR*, 2014.
- [25] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *CVPR*, 2017.
- [26] 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, 2018.
- [27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.
- [28] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.
- [29] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.