The BDD-Nexar Collective: A Large-Scale, Crowsourced, Dataset of Driving Scenes

Vashisht Madhavan
Trevor Darrell
Fisher Yu, Ed.

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2017-113
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-113.html

May 29, 2017

The BDD-Nexar Collective: A Large-Scale, Crowsourced, Dataset of Driving Scenes

by Vashisht Madhavan

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Trevor Darrell
Research Advisor

(Date)

* * * * * * *

Professor Sergey Levine
Second Reader

(Date)

Abstract

In recent years, the meteoric rise of deep learning has catalyzed significant progress in algorithms for visual perception tasks. One particularly popular application of deep neural networks is in the domain of autonomous driving, which is now at the forefront of many research efforts in industry and academia. Despite the plethora of work devoted to learning from unlabeled data, deep networks still require large amounts of labeled examples to learn effective representations. Additionally, deep learning research for autonomous driving is hamstrung by the lack of large, publicly available datasets. To this end, we present the BDD-Nexar dataset, a large-scale collection of urban driving scenes. Our dataset is comprised of high-quality video sequences taken from multiple vehicles, across three major cities in the United States: San Francisco, New York, and Los Angeles. Along with the GPS, IMU, and time data attached to each video sequence, we present 5,000 images with pixel-level and instance-level semantic labels, as well as 10,000 images with bounding box annotations. Although the size of our dataset is comparable to that of contemporary datasets, we exceed previous efforts in terms of scene variability, content, and complexity. Through our in-depth analysis, we verify our dataset as a challenging and extensive benchmark for computer vision research for autonomous driving.

1 Introduction

Perception for autonomous vehicles is now at the forefront of many computer vision research efforts. With the automobile industry investing heavily in the autonomous driving market, computational systems for understanding road scenes are more commercially relevant than ever. Two of the most promising directions for vehicle perception are semantic segmentation, or dense pixel-level prediction, and object detection, or rough instance localization. As these tasks require both prediction and localization of objects in a scene, they are a tall order for vehicles operating in dense urban environments. With the emergence of deep convolutional neural networks and large labeled datasets[1, 2], however, numerous advances in image classification[3, 4, 5] and object detection[6, 7, 8] have been made in recent years. In a similar manner, fully convolutional networks (FCNs)[9, 10] have proven successful for semantic segmentation tasks. Despite these advancements, however, deep learning approaches for image localization tasks require large amounts of densely annotated data to learn effective representations. In the domain of autonomous driving, it is especially cumbersome and expensive to collect and annotate large amounts of data.

Unlike image classification tasks, which only require image-level tags, dense prediction tasks require numerous instance-level tags per image, making it quite tough to collect a large set of annotated images. For autonomous driving research, the high object density in urban street scenes, along with the cost required to set up vehicles for data collection only aggravate the issue. Recent work has focused on improving the annotation process[11] and learning from various other modalities[12, 13], yet it is insufficient to replace the value of large, labeled datasets. Existing public datasets, such as Cityscapes[14], CamVid[15], and KITTI[16], have tackled this problem, yet only contain a relatively limited number of labeled images. Moreover, these datasets fail to capture the complexity and variation of real driving scenes, as models trained on this data are prone to overfitting and exhibit poor performance when applied to real-world data[17]. Variation in weather, traffic density, road structure, and ego-vehicle are all very important factors for consideration when deploying vision systems in cars, yet they are all things that contemporary datasets fail to capture. These factors culminate to severely limit the progress of computer vision research in autonomous driving; a problem we wish to alleviate with our new dataset.

We propose a highly diverse collection of driving video sequences along with pixel-level, instance-level, and bounding box annotations. By collecting data from three very diverse cities in the United States: San Francisco, New York, and Los Angeles, across many different times in the day, weather conditions, and driving scenarios(i.e. suburban, urban, highway, etc.), we are able to create a diversity of annotated images unseen previously. Unlike any previous dataset, we crowdsource the collection our data, resulting in images from many different car models, over many regions of the United States. The contents of our dataset include:

- 5,000 images with pixel-level and instance-level annotations
- 10,000 images with bounding-box annotations
- 1000 hours of HD video sequences along with:
  - GPS locations
  - IMU data - calibrated orientation and acceleration
  - Timestamps

Although we are only releasing 1000 hours of video data, our image data draws from over 100,000 hours of video data, allowing us to capture a high variation in scene content (i.e.rare scenarios) and object density. In the proceeding sections, we discuss the contents of the dataset in further detail, as well as present analysis to validate this dataset as a more challenging and exhaustive benchmark than that of other contemporaries.

2 Dataset

Our dataset was designed with the goal of being the standard benchmark for object detection and semantic segmentation of driving scenes. As a result, we optimized the image dataset in terms of density, content, and geographic diveristy. Additionally, we wanted to enable research in unsupervised learning from videos, dense 3D reconstruction, and reinforcement learning, by providing a large set of safe driving sequences. Naturally, our video dataset was built with these applications in mind.

2.1 Data Collection

Instead of setting up our own data collection rig, we partnered with Nexar, an AI-dashcam company, to streamline the crowdsourced collection of driving data. Although this does not contain the wide array of sensory modalities of standard collection vehicles, it is a suitable alternative since our contributions focus on visual recognition tasks. Nexar's phone application requires users to leave their phone's camera pointed towards the road while driving, enabling the application to provide safety warnings in case of an emergency or potential crash. By nature of the application agreement, Nexar is able to collect the data recorded while driving for further processing and analysis. Our partnership with Nexar allowed us to freely access and annotate this data.

Since these applications run on a smart phone, sending a constant stream of HD video data would consume a large amount of power and memory. As a result, Nexar is only able to collect 1 min of HD video at 30Hz for every $\sim$ 2 hrs of driving. The rest of the time, the application collects snapshots at 5 Hz, which are much easier to process and store. Since Nexar collects thousands of hours of video from various locations in the US weekly, we had a very large and diverse collection of videos to compile our dataset from. In this release we focused on three major cities with both dense urban environments and extensive highway systems, allowing us to ensure visual diversity and object class coverage in our dataset. Another upside of this collection method is that we can gather data from different vehicle models, increasing the variety of road scenes encountered. The downside, however, is that video quality relies heavily on a user's mobile camera and the proper setup for recording dashcam videos. This resulted in much of the data being of poor quality and, thus, useless. In the next section, we outline a method for automatic filtration of low quality videos, which we used to select videos for release.

Figure 1: An example image taken from the collected Nexar videos

2.2 Data Filtration Process

Inspection of our initial data collected helped us identify 4 main sources of error:

1. Too Dark - videos where it is too dark for the camera to perceive relevant objects

Figure 2: An example of a scene that is too dark

2. Too Blurry - videos where the boundaries between various objects is unclear or muddled due to rain, snow, or other inclement conditions

Figure 3: An example of a scene that is too blurry

3. Non Road Scenes - videos where the point of focus is not the road ahead. Usually occurs when the phone camera is inverted to face the driver or the driver forgets to turn off the recording when finished driving.

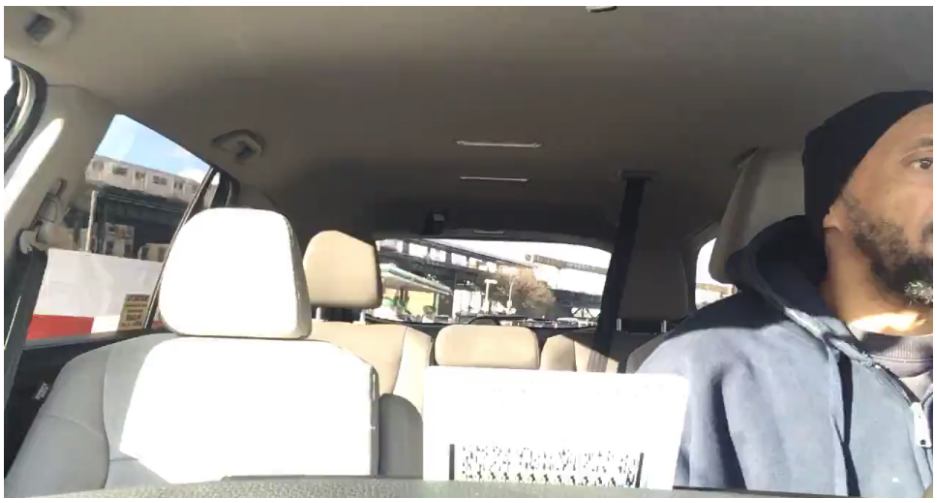


Figure 4: An example of a scene that is not of the road

4. Heavily Occluded Scenes - videos where the dashboard and windshield ornaments of the car heavily occlude most of the scene's contents.

Figure 5: An example of a scene that is too heavily occluded

For the filtration process, we were able to automatically filter out images with the first 3 types of error using off-the-shelf detection methods. For detecting dark images, we gathered representative frames from all HD videos and

calculated the pixel magnitude of each. We then filtered out the everything below the 10th percentile of pixel magnitudes and above the 90th percentile to filter out images that were too bright. Empirically, this filter was effective in filtering out images beyond the camera's dynamic range. To handle blurry images, we calculated the variation of the Laplacian image [18] as a measure of blur. Then we filtered everything with a blur level below a fixed variance, in this case 400. Finally, to filter out the non-road scenes, we used a Caffe[19] model for face detection outlined in [20]. If the bounding box outlining the face was larger than a 100x100 pixel region we filtered it out. This prevented the removal of images with pedestrians' faces. In addition, we ran a dilated fully convolutional network[10] to segment roads and cars in the representative images, filtering out those that contained neither. Although these automated methods are subject to error, we set the parameters for each algorithm to minimize the False Negative (FN) rate. This significantly increased the number of false positives, but since our video corpus is quite large, we were still left with many hours of video.

The final issue of heavy occlusion was a bit more difficult to detect. Since very few automation methods exist for this task, we used manual, in-house filtering. Once these videos were filtered, the representative frames were sampled for annotation.

2.3 Image Dataset Specifications

For each of the good videos remaining from the filtration process, we took a few representative frames from each video. Each frame was of 720x1280 8-bit RGB resolution. Because the frame rate is very high, there is a high visual and geographic consistency between consecutive frames in a video. As a result, annotating each frame from a video would result in low geographic diversity with high redundancy. To circumvent this issue, we sampled every 20 seconds from each HD video, which varied in length from 40 sec - 80 sec. This decorrelated the content of sampled frames and improved the coverage of object classes in the dataset. To enforce geographic diversity, however, a location based sampling method was necessary.

Our data collection process centered around three dense, urban environments: San Francisco, Los Angeles, and New York, as well as their surrounding areas. Naturally, many of the frames extracted were very close in geographic

vicinity (i.e. airports, main streets, etc.).Dense annotation is both expensive and time consuming, so we aimed to maximize the geographic coverage of our image dataset, while annotating a minimal number of images. This improved the variety of scenes captured reduced the likelihood of models to overfit to certain city streets. The algorithm in 1 outlines our selection process.

Algorithm 1 Geographic Selection Algorithm

I = set of images and corresponding GPS coordinates
procedure GEOSELECT(I, N)
    $S = \{\}$ ▷ set of sampled images
    while $|S| < N$ **do**
        randomly pick $(s, l) \in I$ where s = image and l = GPS
        $I = \{(s', l') : l' \neq l, l' \in dist(l', l) \leq 0.1\}$ ▷ $dist$ is in miles
        $S = S \cup \{s\}$
    return S

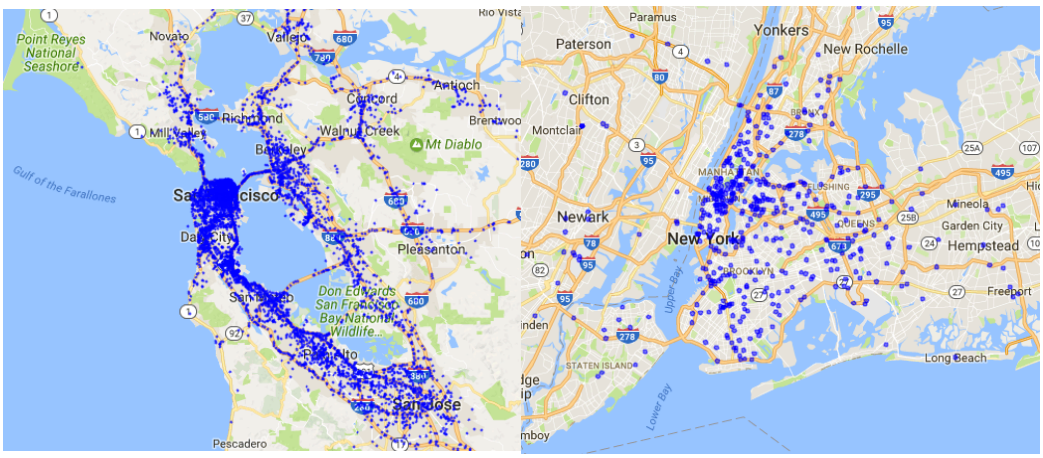


Figure 6: Sample GPS point cloud for SF and NY

As a final step to ensure data diversity and difficulty, we imposed constraints on the types of scenes annotated and the object instance density. In order for an image to be informative for object detection and semantic segmentation, it must contain high-quality examples of the object to be predicted. By the same logic, the more instances of objects an image contains, the more informative it will be for these models. Consequently, we imposed a

constraint on the number of traffic participants in each image annotated. If the image fell below the predefined threshold of instances, we would discard it from the set. This process removed images of empty roads and uninformative highway scenes that are not necessarily useful for learning. From the remaining set of images, we took even proportions from sparse and dense urban areas, to ensure a good mix of highway, urban, and suburban road scenes. Figure 6 shows the location point cloud of our annotated images for San Francisco and New York. Although this method of selection was a good baseline for choosing informative sample, future methods centered around active learning would better automate thin process, while better identifying failure modes.

2.4 Object Classes

The object classes in our dataset are very similar to those outlined Cityscapes, sans some slight extensions. This motivates various cross-dataset transfer research and encourages the combination of both types of data. The formal list of categories is as follows:

- Flat
  - Road, Sidewalk, Parking Spot, Rail Track
- Human
  - Pedestrian, Rider
- Vehicle
  - Car, Truck, Bus, Train, Motorcycle, Bicycle, Caravan, Trailer
- Construction
  - Building, Wall, Fence, Guard Rail, Bridge, Tunnel, Garage
- Object
  - Pole, Traffic Sign, Traffic Light, Banner, Billboard, Street Light , Traffic Device, Lane Divider, Highway Frame, Parking Sign, Traffic Cone, Pedestrian Crossing Signal
- Nature

  - Vegetation, Terrain
- Sky
- Void
  - Ground, Dynamic, Static

Since bounding boxes do not require each image pixel to be labeled, we compress the label set to only consider dynamic categories that are relevant for driving. Usually these are objects like traffic lights, vehicles, pedestrians, etc. This enabled us to 1) annotate images at a much quicker pace, and 2) amass a larger set of images. By retaining similar label sets to contemporary datasets, such as KITTI, we enable straightforward evaluation for previous object detection research in autonomous driving. The labels used for the bounding box component of the dataset are as follows:

- Vehicle
  - Car, Truck, Bus, Train, Motorcycle, Bicycle, Caravan, Trailer
- Humans
  - Pedestrian, Rider
- Traffic Light
- Traffic Sign

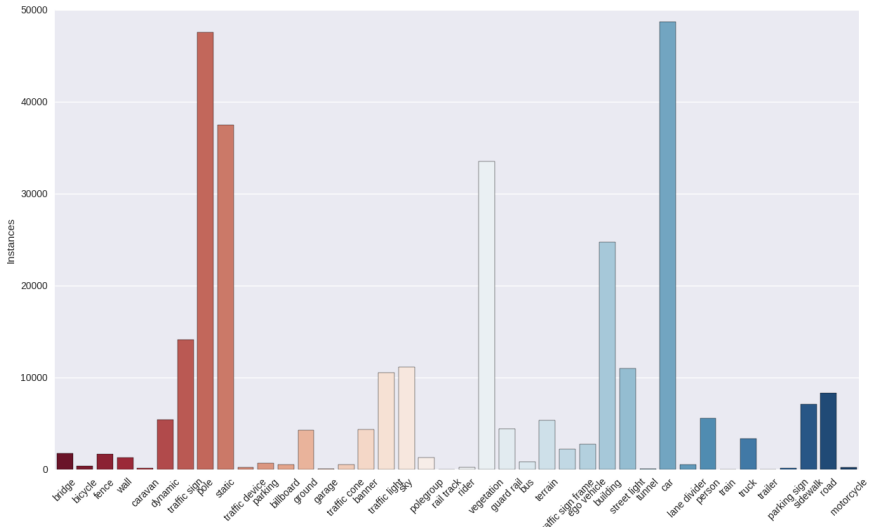


Figure 7: Number of instances for each object class in the dataset

2.5 Annotation Process

As mentioned before, dense labeling is expensive. To minimize costs, we made sure every dollar we spent went towards high quality, informative image annotations. Although open-source tools, such as LabelMe[21] exist, they not only require in-house annotation, but also do not possess the ability to share work across multiple annotators. More sophisticated web tools, like the ones presented by Bell et. al [22] are also readily available, but rely on annotation through Amazon Mechanical Turk, which provides no guarantee on quality and often produces inconsistent results for dense annotation tasks. After considering the cost and benefit of various approaches, we decided to outsource our annotation to professional annotation services.

We relied on three main annotation companies: Samasource, Deepomatic, and Datatang. With these services, dedicated workers are assigned to both pixel-labeling and bounding box tasks, along with a project manager, who works to ensure quality and enforce our outlined annotation standard. At rates between $8 - $12 an hour, these services provide high-quality annotations at fair prices. Additionally, parallelizing the annotation process be-

tween three different services allowed us to achieve high throughput in a short amount of time. For our bounding box annotation task, the time required to annotate each image is low and the label set is quite small. As a result, it was relatively cheap to generate a large corpus of bounding box images using annotation services. On the other hand, pixel-level labeling tasks require not only a much more extensive label set, but also more time and detail to annotate an image. Consequently, the cost to annotate a large dataset skyrockets and renders our current method monetarily infeasible. In our exploration we discovered annotation services that would reduce the cost dramatically if annotation tools and review platforms were provided on the web. As a result, we developed our own open-source annotation tool, loosely based on the platform from OpenSurfaces [22].

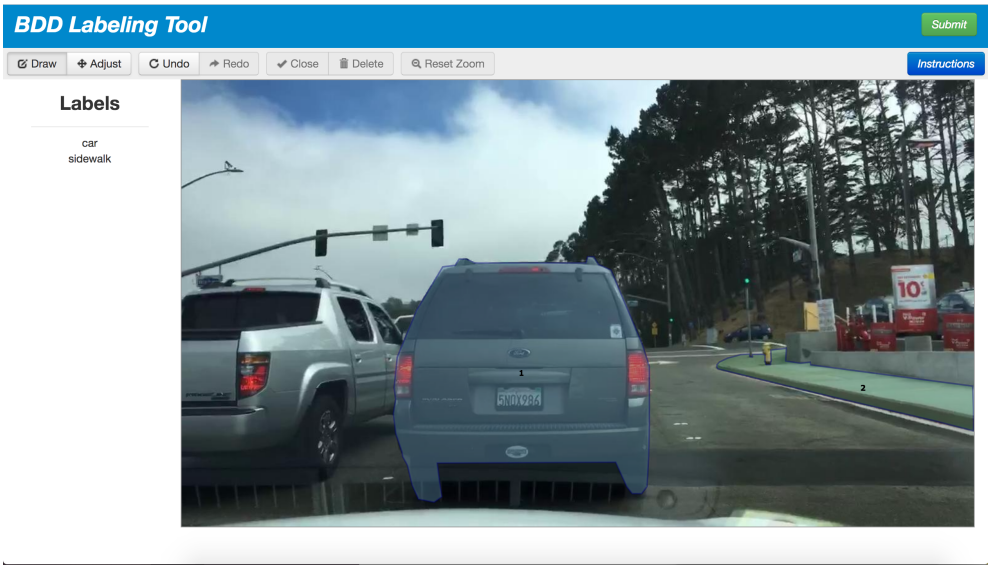


Figure 8: A snapshot of our annotation tool

The platform, found at https://github.com/VashishtMadhavan/BDDLabeler, is a lightweight web application that seeks to minimize the time to deployment. Despite its simplicity, the tool contains extensive annotation and review interfaces. Exposing the platform as a hosted web application also facilitates direct transfer from annotators/reviewers to our central image database.

Going forward, we hope to utilize this interface to collect much larger volumes of fine-grained annotations at a much lower cost. A snapshot of the interface is shown in Figure 8.

2.6 Dataset Splits

As part of our release, we split both the bounding box annotations and pixel-level annotations into training, validation, and test sets. Instead of splitting randomly, each split was sampled to achieve a balanced distribution of geographic locations, object classes, and scene density. Within each split we took an even amount of data from San Francisco, New York, and Los Angeles. To ensure variation between scene content and size in each split, we added an even number of images from each object category to each split. We also balanced the number of traffic object instances between splits, with each containing a large, medium, and small amount of vehicles, pedestrians, and traffic lights/traffic signs. This resulted in a split of 3815 training examples, 500 validation examples, and 585 test examples for the pixel-annotation dataset. For the bounding box dataset, the splits generated were 8640 train, 475 validation, and 885 test examples.

2.7 Pixel Annotation Analysis

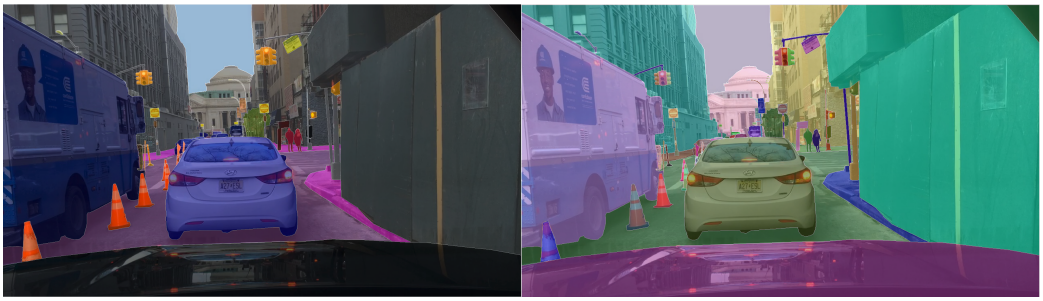


Figure 9: Some samples of class-level(left) and instance-level(right) fine-grained annotations overlaid on images

We compare the BDD-Nexar dataset to other existing driving datasets in terms of (i) annotation volume, (ii) scene content, and (iii) scene complexity. Since Cityscapes [14] is the leading dataset for pixel and instance-level annotations, we use it as a reference point for our analysis.

As a brief overview, Cityscapes contains 5000 images at a 720x1280 pixel resolution, each coming with pixel and instance-level annotations. These image are taken throughout 50 cities in Germany, all with the same vehicle sensor suite. Our dataset contains the same amount of pixel-annotated images, so by a direct comparison of dataset size, there is no advantage to using our data. For a more nuanced analysis, we compared the number of total instances across categories, as shown in Figure 10. Despite having the same number of total images, our dataset contains higher instance density almost across the board. This not only means that object categories are well represented, but also that there is more information, per sample, to learn from. This suggests that our data is more fruitful for learning algorithms, especially since the number of labeled images is not very large. In the future, we plan to increase the number of images with fine-grained annotations, giving our dataset an edge over Cityscapes in size as well.

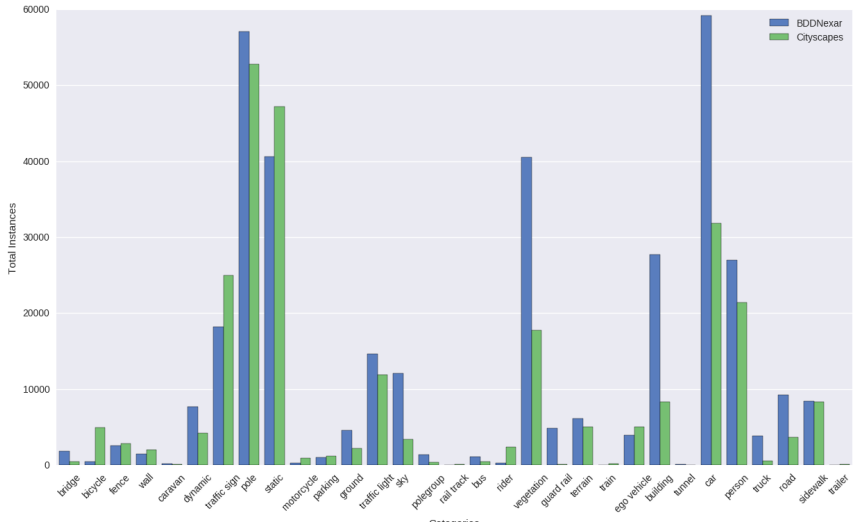


Figure 10: The total number of instances per object class. Our dataset contains higher density of instances for almost every major object class.

Although Figure 10 validates our dataset in terms of object coverage, the analysis gives no indication as to the variation between scenes. For example, a set of images with exactly 3 instances of vehicles per image is less useful to a deep network than a dataset with a varying number of instances per image. We desired deep models trained on our dataset to be more robust to real-world traffic conditions. As a proxy for measuring the variation in traffic conditions, we analyze the distribution of traffic participant (vehicles and pedestrians) instances across images. Figure 11 displays a histogram of traffic participant instances for each dataset. The vehicle instance distribution for our dataset better covers the 10-30 instance range much better than Cityscapes does. However, Cityscapes has slightly better coverage for very high instance ranges(40-60 inst/image). In a similar vein, the instance distribution for pedestrians is also better distributed across the range of instance levels, leading us to claim greater variety in terms of scene content. This level of coverage for traffic conditions not only makes our dataset less likely for models to overfit, but also more likely to be useful for models applied in the real-world.

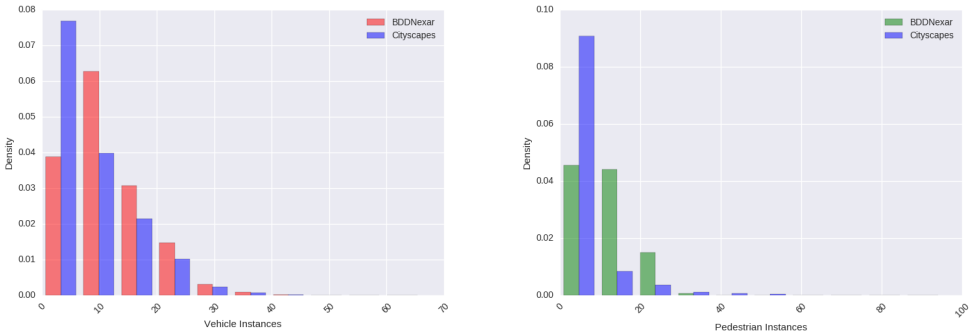


Figure 11: Distribution of # of instances per image for traffic participants. Our dataset shows higher variance and a more even histogram than that of Cityscapes.

The final element in our analysis is with respect to scene complexity. In this instance, we measure scene complexity in terms of the proportions of each image that pedestrians and vehicles take up. In work by Hoeim et al. [23], it is shown that the largest source of error in object detection, and more generally localization tasks, is small differences in object size. Semantic segmentation and object detection models trained on standard datasets tend to show high performance when given large object instances, yet exhibit poor localization performance on smaller, more obscure objects. By increasing the coverage of small object instances in our dataset, we hope to assist deep networks in being more robust to this type of error. The results ins Figure 12 suggest that our dataset contains higher coverage of small objects than Cityscapes, as the distribution of object pixel proportions is skew right. Although our coverage of larger objects is not as extensive, our aim is to address issues that Cityscapes does not. Ideally, our goal is to have an even distribution of object sizes, with good coverage for all sizes, but in this instance, our contribution lies in the a more thorough treatment of small objects, which are more difficult for deep neural networks trained for localization tasks. As mentioned before, we hope that our data is used in conjunction with Cityscapes to improve overall coverage of road scenes, and enable the training of more robust models.

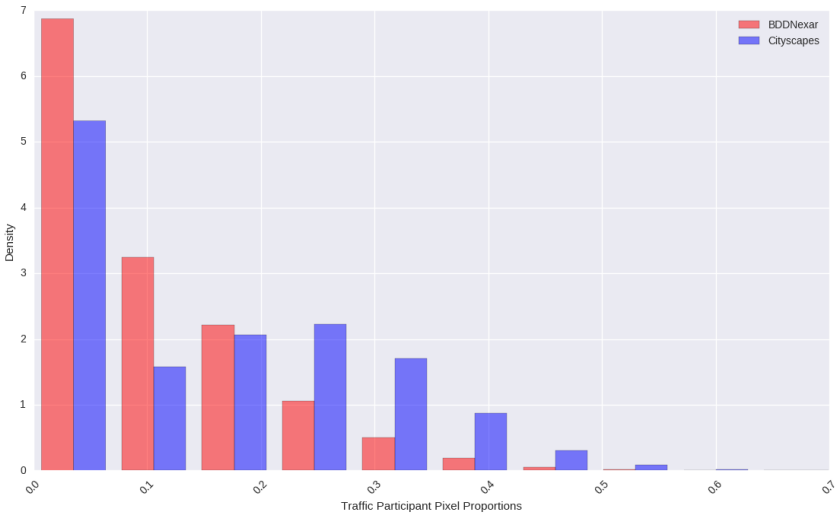
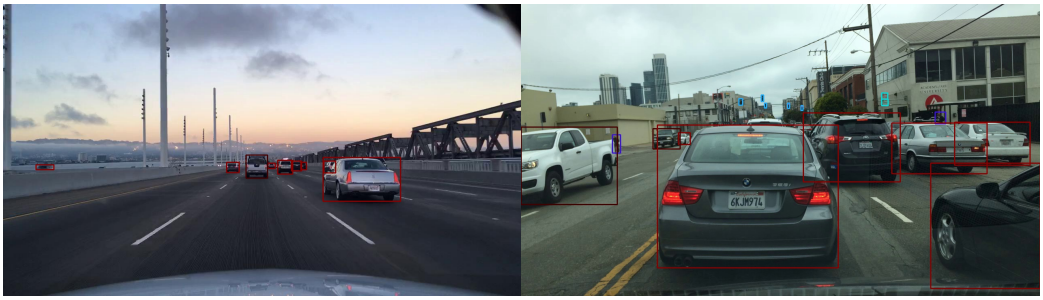


Figure 12: Proportion of image pixels that contain traffic participant instances. Our dataset is skew towards lower pixel proportions, which implies smaller objects

2.8 Bounding-Box Analysis



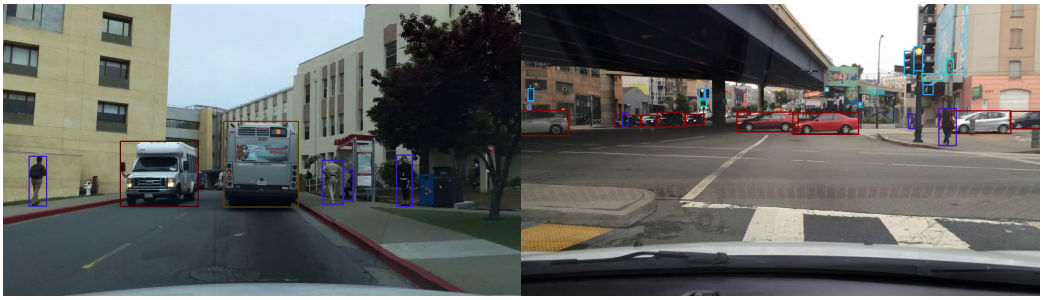


Figure 13: A sample of the bounding box annotations overlaid on images

For the bounding box dataset, we conduct a similar analysis with a leading object detection benchmark, KITTI[24]. KITTI is a collection of road scenes collected by driving around Karlsruhe, Germany. It contains 7,418 images of 375x1242 pixel resolution with tight bounding boxes drawn around vehicles, using the same object categories as Cityscapes, and pedestrians, both sitting and standing. In our dataset, we deliver 10,000 images of 720x1280 pixel resolution across 3 different cities. The label set is similar except for the inclusion of traffic lights and traffic signs. With full overlap with the labelset of KITTI, our dataset enables seamless cross-dataset evaluation. Additionally, having 1) a larger set of images and 2) data collected beyond one city, models trained on our dataset are less prone to overfitting and more robust to real world scenarios, as with the pixel-annotated dataset and Cityscapes. Our analysis of object instance density and object size between our dataset and KITTI verifies this claim.

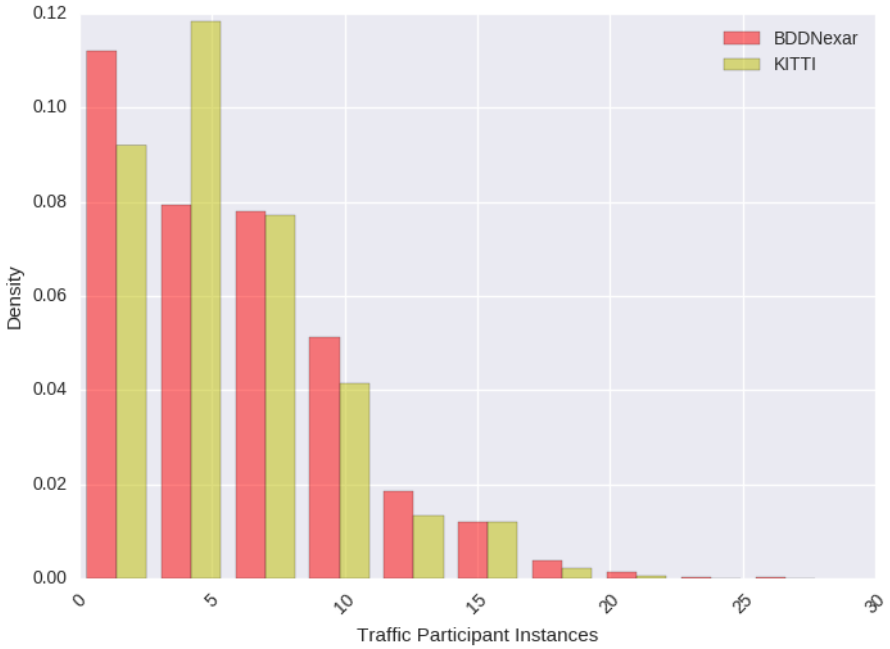


Figure 14: The distribution of traffic participant instances

In the previous section, we argued that instance density distributions is a useful for understanding scene content variety. From the chart in Figure 14 the distribution of instances is fairly similar between our dataset and KITTI, with ours being slightly more balanced. KITTI seems to have a large number of instances in the 4-7 range, while ours has a peak around the 1-2 instance range. However, our distribution covers higher instance scenarios more thoroughly, with some images containing almost 30 instances of pedestrians and vehicles. This is important for high traffic scenarios, where precise driving is necessary. Low traffic scenarios are covered well by both datasets, so their is marginal advantage to our dataset in that respect.

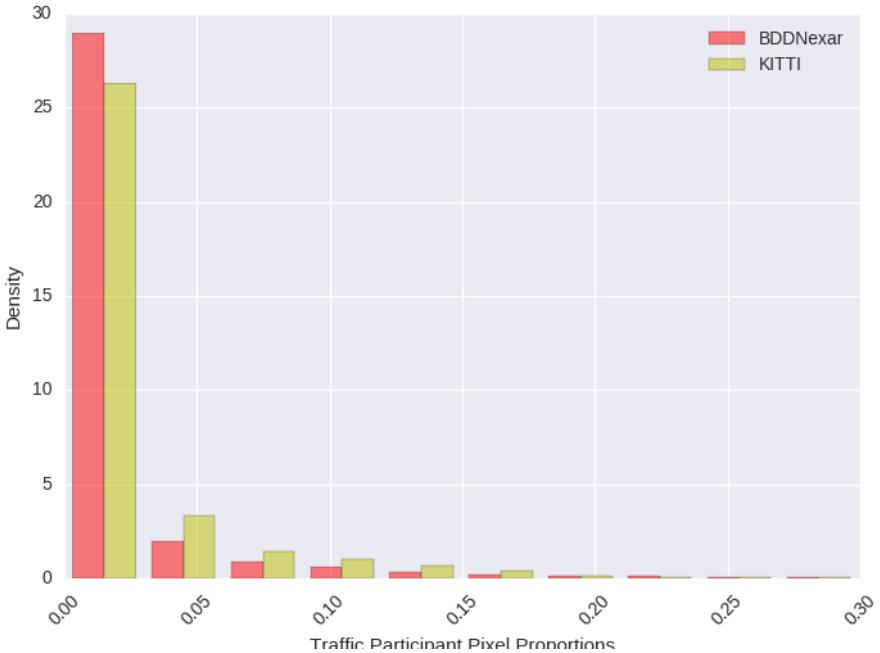


Figure 15: Proportion of pixels that traffic participant boxes take up. The distribution is skew right towards smaller instance sizes.

Similar to the pixel-level annotations, the proportion of pixels occupied by traffic participants is generally lower. Although not significantly so, KITTI provides better coverage of larger objects, which we consider less difficult for object detection models to identify. The better coverage of smaller traffic objects and more even distribution of object instances increase the fidelity of our data in modeling real-world traffic conditions. By collecting a larger corpus over more cities, we are able to outpace KITTI in terms of diversity and difficult, making our dataset a better benchmark for object detection tasks.

3 Conclusion

In this work, we present the BDD-Nexar Dataset, a comprehensive collection of annotated data, designed to spur visual recognition research in the space of autonomous driving. Our contributions, as mentioned previously, include:

(i) a large and comprehensive driving dataset, more diverse and extensive than others in the space; (ii) an analysis of contemporary driving datasets and their limitations, which we address with our new collection and (iii) a large corpus of raw driving video, useful for sequence modeling, 3D motion estimation, and reinforcement learning. As this is the first public data release, we will continue to expand the dataset and adapt our annotations to the needs of specific autonomous driving applications. Although our dataset provides a richer set of images than that of KITTI or Cityscapes, we promote the use of our dataset in conjunction with our contemporaries. We have taken steps to ensure smooth integration with both KITTI and Cityscapes, in hopes of pushing the field forward altogether.

Although the BDD-Nexar dataset has numerous benefits in terms of diversity and complexity, it is one the first datasets collected in the United States. As a result, it captures traffic conditions and road styles not seen previously in other datasets. The variance (i) numerous lighting conditions, (ii) different types of roads(i.e. highways, urban streets,etc.) and (iii) driving situations encountered, make this the most interesting, yet challenging of any dataset available today. As we accrue more data, we hope to understand the failure modes of state-of-the-art semantic segmentation and detection methods, while gaining insight for improving them.

4 Acknowledgements

The authors acknowledge the team from Nexar and their contributions to the data collection effort. We will be partnering with them for future contributions to this dataset

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, June 2010.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[5] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 487–495, Curran Associates, Inc., 2014.

[6] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013.

[7] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. abs/1312.6229, 2013.

[8] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014.

[10] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1511.07122, 2015.

[11] J. Xie, M. Kiefel, M. Sun, and A. Geiger, "Semantic instance annotation of street scenes by 3d to 2d label transfer," *CoRR*, vol. abs/1511.03240, 2015.

[12] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," *CoRR*, vol. abs/1608.02192, 2016.

[14] 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," *CoRR*, vol. abs/1604.01685, 2016.

[15] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. xx, no. x, pp. xx–xx, 2008.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[17] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *CoRR*, vol. abs/1612.02649, 2016.

[18] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martínez, and J. Fernández-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *ICPR*, 2000.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *CoRR*, vol. abs/1408.5093, 2014.

[20] S. S. Farfade, M. J. Saberian, and L. Li, "Multi-view face detection using deep convolutional neural networks," *CoRR*, vol. abs/1502.02766, 2015.

[21] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. J. Comput. Vision*, vol. 77, pp. 157–173, May 2008.

[22] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "OpenSurfaces: A richly annotated catalog of surface appearance," *ACM Trans. on Graphics (SIGGRAPH)*, vol. 32, no. 4, 2013.

[23] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III*, pp. 340–353, 2012.

[24] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J. Rob. Res.*, vol. 32, pp. 1231–1237, Sept. 2013.