Abstract

This technical report summarizes and provides detailed information about the MIT DriveSeg (Semi-auto) dataset, including technical aspects in data collection, annotation, and potential research directions.

1. Introduction

Solving the external perception problem for autonomous vehicles and driver-assistance systems requires accurate and robust driving scene perception in both regularly-occurring driving scenarios (termed “common cases”) and rare outlier driving scenarios (termed “edge cases”). In order to develop and evaluate driving scene perception models at scale, and more importantly, covering potential edge cases from the real world, we take advantage of the MIT Advanced Vehicle Technology Consortium’s MIT-A VT Clustered Driving Scene Dataset and build a subset for the semantic scene segmentation task. We hereby present the MIT DriveSeg (Semi-auto) Dataset: a large-scale video driving scene dataset, which contains 20,100 video frames with pixel-wise semantic annotation. We propose semi-automatic annotation approaches leveraging both manual and computational efforts to annotate the data more efficiently and at lower cost than manual annotation.

2. Background

2.1. Semantic Scene Segmentation Datasets

Recent progress in deep learning has significantly pushed forward the benchmarks of many popular computer vision problems, such as image classification and segmentation. The improved solutions and algorithms make it possible for people to explore more challenging problems and real-world applications. Scene segmentation, based on semantic image segmentation, is one of the challenging computer vision problems. The goal is to partition the full image into semantically meaningful parts.

However, training neural networks require large-scale data. Datasets for scene segmentation are not as widely available as those for traditional image classification. Some popular challenges along with benchmarks for semantic segmentation and scene parsing have been released in recent years. These include PASCAL VOC [3], MS COCO [7], MIT Scene Parsing [9], and CityScapes [1]. DAVIS [8] is another dataset for video object segmentation. Although some of the datasets have a relatively large number of annotated images (20k images for both MIT Scene Parsing [9] and CityScapes [1]), the cost and effort in manual annotation investigated in these datasets is difficult to deploy at scale. For example, in [1] annotation and quality control required more than 1.5 hours on average for a single image. In our earlier work proposing the MIT DriveSeg (Manual) Dataset, we were able to optimize the manual annotation process with enhanced tooling and frameworks. While that gave us a $10\times$ cost reduction for annotation compared to existing work, it is still too expensive to annotate at scale, especially for high-frequency videos. In addition to solving the actual scene segmentation problem, it is also important to develop efficient mechanisms for the annotation process, which has been a preliminary requirement for obtaining large-scale datasets.

2.2. MIT-AVT Dataset

The MIT-AVT study [4] as a whole seeks to gather a large variety of driver, scene, telemetry and vehicle state data from different vehicles equipped with varying types of advanced vehicle technologies. Cameras include dash (internal cabin of the car), face, front and sometimes instrument cluster-facing orientations recording at 720p resolution, 30 frames per second (fps), for the full duration of any trip taken in the vehicle. Telemetry data gathered includes IMU, GPS and selected signals from the vehicles CAN bus.
2.3. MIT-AVT Clustered Driving Scene Dataset

Driving encompasses a wide variety of external scenes, many of which are periodic (exits and intersections) or infrequent (unique bridges). Weather and light conditions among other factors may further augment those scenes. Approaches that go beyond random sampling large-scale naturalistic driving dataset may include manually implemented filters for conditions such as weather, season, road type, or time of day based on date and GPS data. However, these filters do not account for differences in the visual characteristics of the scene, since the imagery data is not utilized in the sampling process. For example, bridges and overpasses may cause more varying illuminance conditions at the same location in similar weather conditions, compared to normal highways. Moreover, in some cases, the visual sensor may be partially occluded or blurred by the presence of obscuring materials such as snow, ice, rain, fog, dirt, etc.

In order to overcome these challenges, a vision-based approach to obtaining both representative samples and a variety of edge cases from the AVT data was proposed. The method entails extracting 10-second clips from the MIT AVT dataset. 5 evenly spaced images are selected from each clip and are run through a Resnet-50 [6] model pre-trained on ImageNet [2] to obtain feature embeddings. These embeddings are then reduced in dimension using PCA, combined chronologically and then clustered using Mini-Batch K-means. Clusters are then manually annotated for basic, high-level scene characteristics to assign contextual classes to each cluster. These algorithms were selected to minimize processing time, cost and local memory use in order to locally process terabytes of driving data while still effectively differentiating vision-based scene characteristics.

3. MIT DriveSeg (Semi-auto) Dataset

For the purpose of training and evaluating driving scene perception models at scale and more importantly, covering potential edge cases from the real world, we take advantage of the MIT-AVT Clustered Driving Scene Dataset for
3.1. Dataset Specifications

In order to keep camera positioning and reflection patterns constant, a single vehicle (Range Rover Evoque) was sampled from the MIT-AVT Clustered Driving Scene Dataset. This sub-sampling was carried out to focus efforts on the variability of the external scene independent of vehicle factors. Data from this vehicle spread across different clusters was used, equally obtaining common cases and edge cases in the naturalistic driving scenarios.

The dataset contains 20,100 video frames (67 video clips, 10 seconds each at 30 frames per second) at 720P resolution (1280×720). We semi-automatically annotate pixel-wise semantic labels for each frame with 12 classes in our dataset: vehicle, pedestrian, road, sidewalk, bicycle, motorcycle, building, terrain (horizontal vegetation), vegetation (vertical vegetation), pole, traffic light, and traffic sign. The annotation is done through semi-automatic approaches leveraging both manual and computational efforts.

3.2. Semi-automatic Annotation

Existing full scene segmentation datasets [1, 9] rely on hiring a very small group of professional annotators to do full frame annotation, which are usually highly costly and time-consuming. This is reasonable because the pixel-wise annotation is a task of high-complexity that reach the limits of human perception and motor function. However, the driving task requires annotated data in potentially much larger scale in order to cover the long tails of edge cases. For this purpose, we explore and develop several approaches for automatic and semi-automatic annotation that can largely improve the annotation efficiency and at the same time maintain similar level of noise on the annotated data as manual annotation.

3.2.1 Drivable Area Projection From Vehicle Trajectory

Naturalistic driving data contains sequential information of visual scene and vehicle motion. Since steering commands result in future vehicle motion, we can use the sequential vehicle control data to infer the drivable area, by projecting future vehicle trajectory onto the current front scene. The inferred drivable area is of high precision such that it only consists of road and on-road fast-moving objects (vehicles), because it is actually the to-be-driven area according to the future information.

To obtain the drivable area, we first warp the front scene into the bird’s eye view with calibrated visual perspective transformation. We then calculate the future vehicle trajectory with the steering, speed, and other vehicle specifications. We draw the vehicle trajectory on a bird’s eye view of front scene, and project the single-lane area back onto the front scene image. Fig. 2 shows examples of the inferred drivable area overlaid on front camera image. We use the projected drivable area for the further relabeling as described in the next section.

3.2.2 Automatic Segmentation Relabeling with Information Fusion

Previous work [5] shows that the disagreement between different deep learning models can be viewed as a strong signal that is sufficient to improve the accuracy of the overall system given human supervision. In addition to using a full pre-trained full scene segmentation model for automatic labeling, we introduce another detection model that can detect and predict the bounding box over certain objects, including vehicles and pedestrians. Theoretically, a pixel being predicted as vehicle by the segmentation model should also...
be within the bounding box of a vehicle predicted by the detection model.

Model fusion, or ensembling, is a common method in machine learning that averages predictions from different sources in order to get a more accurate, robust result. In the scene segmentation task, this idea can also be utilized, but in a more structural and reasonable way, given the prior knowledge of the nature of the driving scene. We propose an information fusion method that leverages different source of perception to combine the advantages of each and automatically fix some obvious errors introduced by the main scene segmentation model. The sources and their descriptions are listed as below:

- **Full Scene Segmentation** - Dense and precise labeling of every pixel in the whole scene, similar to the format of annotation, not robust to edge cases (lighting condition, unseen objects).

- **Object Detection** - Bounding box prediction for a few object classes (e.g. car, person), accurate and robust because the model is trained on much larger scale of data compared to scene segmentation.

- **Drivable Area** - Road prediction from steering commands described in Sec. 3.2.1, high precision but low recall as it is only available for part of the scene.

In order to make the most use of the advantages of the above methodologies, we design the automatic annotation framework to use the more robust and accurate algorithm to guide the labeling in certain parts. This can be briefly described as two principles:

1. A pixel can be labeled as car or person by the segmentation model only if it is within the bounding box of the object predicted by the detection model;

2. A pixel can only be labeled as road or car (fast moving object that can exist in the way of driving) if the pixel is within the drivable area projected from steering commands.
3.2.3 Semi-automated Annotation with Confidence-level Adjustment

State-of-the-art semantic segmentation methods utilize deep convolutional architectures that are good at detecting image features, such as shapes, edges, colors. While models may not generalize well to unseen or edge case situations existing in some parts of a full scene image, most other parts are predicted with reasonably good performance. In order to combine the strength of human perception and machine efficiency, we propose a semi-automated annotation framework that can perform image annotation by adjusting the prediction from strong deep neural networks.

For a deep convolutional neural network doing semantic image segmentation, a common format of prediction is

\[ Y_{h,w} = \arg\max_c y_{i,j,c} \quad \forall i \in h, j \in w \]

where \( h, w \) are height and width of the image, and \( c \) is the number of classes. The network predicts the probability of each class on each pixel in the image, and take the class of top probability as the prediction. In this case, for a wrong prediction usually due to domain shift or bad generalization, the predicted class is very likely to have lower probability than other correctly predicted ones. Based on this observation, we design a semi-automatic annotation framework in which the human user can remove the prediction of a single class with simple keyboard entry, and adjust the confidence level to remove less-confident predictions, resulting a coarsely annotated image with high precision. An example is shown in Fig. 4.

3.2.4 Annotation Process

We first sample 67 video clips across different scene clusters from the MIT-AVT Clustered Driving Scene Dataset. The video clips are validated through a manual process to make sure the data quality is reasonable after the Automatic Segmentation Relabeling process. We then perform the Semi-automatic Interactive Segmentation process to further improve the annotation.

4. Research Directions

There are many potential research directions that can be pursued with this large-scale video scene dataset. We provide a few open research questions where people may find the dataset helpful:

**Spatio-temporal semantic segmentation.** We are interested in further research finding a novel way to utilize temporal data, such as optical flow and driving state, to improve perception from using static images only.

**Predictive modeling.** Can we know ahead what is going to happen on the road? Predictive power is an important component of human intelligence, and can be crucial to the safety of autonomous driving. The dataset we provide is consistent in time and therefore can be used for predictive perception research.

**Deep learning with video encoding.** Most of the current deep learning systems are based on RGB encoded images. However, preserving exact RGB values for each single frame in the video is too expensive for computation and storage.

**Solving redundancy of video frames.** How can we efficiently find useful data from visually similar frames? What shall be the best fps for a good perception system? One of the most important problems of real-time applications is trade-off between efficiency and accuracy.

5. Conclusion

The MIT DriveSeg (Semi-auto) dataset has been presented to train and evaluate driving scene perception models at large scale, covering a wide range of real-world driving scenarios. The semi-automatic annotation methods proposed in this work can be further explored and extended to other computer vision tasks.
Acknowledgments

This work was in part supported by the Toyota Collaborative Safety Research Center. The views and conclusions being expressed are those of the authors and do not necessarily reflect those of Toyota.

References


Appendix

A. License Agreement

The MIT DriveSeg (Semi-auto) Dataset is made freely available to academic and non-academic entities for non-commercial purposes such as academic research, teaching, scientific publications. Permission is granted to use the data given that you agree:

1. That the dataset comes “AS IS”, without express or implied warranty. Although every effort has been made to ensure accuracy, we (MIT, Toyota) do not accept any responsibility for errors or omissions.

2. That you include a reference to the MIT DriveSeg (Semi-auto) Dataset in any work that makes use of the dataset. For research papers, cite our preferred publication as listed on our website; for other media cite our preferred publication as listed on our website or link to the website.

3. That you do not distribute this dataset or modified versions. It is permissible to distribute derivative works in as far as they are abstract representations of this dataset (such as models trained on it or additional annotations that do not directly include any of our data) and do not allow to recover the dataset or something similar in character.

4. That you may not use the dataset or any derivative work for commercial purposes as, for example, licensing or selling the data, or using the data with a purpose to procure a commercial gain.

5. That all rights not expressly granted to you are reserved by us (MIT, Toyota).