MIT-AVT Clustered Driving Scene Dataset: Evaluating Perception Systems in Real-World Naturalistic Driving Scenarios

Li Ding[1], Michael Glazer[1], Meng Wang[1], Bruce Mehler[1], Bryan Reimer[1], and Lex Fridman[1]

***Abstract*— Solving the driving scene perception problem for driver-assistance systems and autonomous vehicles requires accurate and robust performance in both regularly-occurring driving scenarios (termed "common cases") and rare outlier driving scenarios (termed "edge cases"). We propose an automated method for clustering common cases and detecting edge cases based on the visual characteristics of the external scene. We apply this approach to develop a large-scale real-world video driving scene dataset of edge cases and common cases. This dataset consists of 1,156,592 10-second video clips, including 450 clusters of common cases, and 5,601 edge cases. We assign human-interpretable metadata labels (*e.g.*, weather, lighting conditions) to the clusters through manual annotation. We further propose two automated methods for large-scale evaluation of scene segmentation models on naturalistic driving datasets that can capture potential system failures without human inspection. Video illustrations of select clusters will be made available to help with future research.**

***Index Terms*— clustering and outlier detection, naturalist driving, scene perception evaluation, large-scale dataset.**

I. Introduction

Large-scale naturalistic driving data collection is an important part of developing autonomous navigation and perception systems for both validating systems and improving them through machine learning and deep learning pipelines. However, such data is not always uniformly useful. The wide diversity of external scenes native to the driving task requires large-scale data collection to capture various environmental conditions, including different road types, weather, seasons, illuminance, visual appearance of scenes and objects, etc.

Existing approaches that go beyond random sampling take advantage of manually-designed filters to obtain reasonable data distributions based on metadata from date, time, IMU, CAN, and GPS. While such filters may succeed in utilizing high-level information in addition to the driving scene that improves the variability of the dataset, what they do not take into account is the actual visual appearance of the scene, which is most likely to affect the performance of perception systems. For example, bridges and overpasses may cause more varying illuminance conditions at the same location in similar weathers, comparing to normal highways. Moreover, in some cases, the visual sensor may fail to maintain expected performance when being partially occluded or blurred by the presence of obscuring materials such as snow, ice, rain, fog, dirt, etc.

In order to overcome those challenges, an automated processing and sampling step for large-scale naturalistic driving data collection is necessary in order to obtain adequate samples of different visual appearances of the external driving scenes. In this work, we propose a vision-based approach that determines two kinds of driving scenes in the data based on their frequency and visual characteristics: 1) representative common cases and 2) uncommon edge cases. The former is important for understanding what kind of visual scene appears frequently in naturalistic driving, ensuring the perception system performing reliably well in the majority of cases. The latter aims at improving the robustness of the system to outliers that exist in the long tail of rarely-occurred driving scenarios.

First, we extract 10-second video clips from the *MIT-AVT* large-scale naturalistic driving dataset [1]. We use a pre-trained deep convolutional neural network (ResNet-50 [2]) to obtain visual feature embeddings on a small set of uniformly sampled images from each clip. These embeddings are further processed and combined to represent each clip, and then modeled for clustering using the Mini-Batch K-means algorithm [3]. The resulting clusters are labeled for basic, high-level scene characteristics and metadata through manual and automatic annotations. We use the Local Outlier Factor algorithm [4] to further detect and validate the edge cases. The algorithms are optimized in order to save processing time, cost and local memory use for large-scale efficient processing. The resulting dataset, named *MIT-AVT Clustered Driving Scene Dataset*, consists of 1,156,592 10-second video clips of the front driving scene, including 450 clusters of common cases, and 5,601 edge cases.

For evaluation of scene perception systems, existing methods require pixel-level ground truth annotation, which is too costly and time-consuming for large-scale datasets. We propose two methods for automated evaluation of scene perception on large-scale naturalist driving datasets (*e.g.*, the proposed dataset). Experiments show that existing deep learning segmentation models fail on some edge cases. Those system failures are automatically captured by the proposed methods without human inspection.

The proposed *MIT-AVT Clustered Driving Scene Dataset* has been used for multiple ongoing research on driving scene perception and naturalistic driving study. Video illustrations of select clusters will be released to help with future research.

In summary, the main contributions of this work are:

- A vision-based clustering approach that determines representative common cases and uncommon edge cases in driving scenes.

[1] All the authors are with Center for Transportation and Logistics, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA. Email: {liding, glazermi, mengw, bmehler, reimer, fridman}@mit.edu

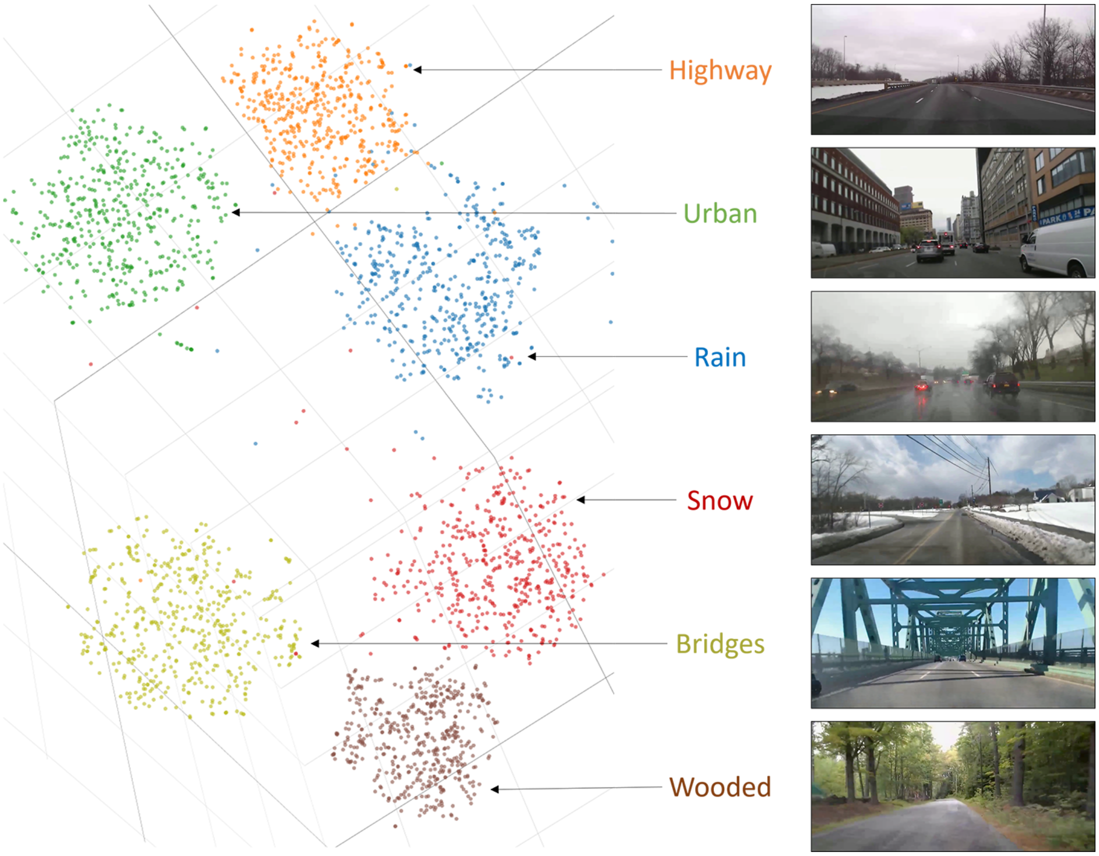


Fig. 1: Visualization via t-SNE 3D embedding of 500 clips (each clip is a point in the plot) from 6 sample clusters in the *MIT-AVT Clustered Driving Scene Dataset*. Each cluster is associated with a human-assigned semantic label, in this case related to road type, weather, and other categories.

- *MIT-AVT Clustered Driving Scene Dataset*: a large-scale naturalistic driving dataset of 1,156,592 video clips for common cases and edge cases, with manually annotated semantic labels for each cluster.
- Two automated methods for the evaluation of scene perception on naturalist driving datasets.

II. Related Work

A. Image Feature Embedding and Clustering

Using pre-trained deep ConvNets as image feature extractors has become an essential component in many computer vision applications [5], [6], [7], [8]. By training on large-scale image datasets (*e.g.*, ImageNet [9]) for recognition tasks, deep ConvNets learn to extract general-purpose visual characteristics that can be directly used for various computer vision tasks [10].

Taking high-dimensional image feature as input for dimensionality reduction, clustering, and density estimation have been widely studied in machine learning [11] and regularly used in computer vision applications [12]. A recent example is [13] that uses deep convnets feature as embeddings for face recognition and clustering. Our work extends on this to explore effective and efficient methods to cluster driving scene video clips at large scale.

B. Clustering-based Anomaly Detection

Anomaly (outlier) detection is broadly studied in data mining [14], aiming to identify rare items, events or observations which raise suspicions by differing significantly from the majority of the data. One popular technique is clustering-based anomaly detection [15], [16], [4], which detects outliers based on local data distributions with distance measurements. The outliers provide insights about the global data distribution, which is essential to large-scale data collection and sampling.

C. Driving Scene Datasets

A number of driving-related datasets [17], [18], [19], [20], [21], [22] have been developed and made publicly available in recent years. Some of those also provide pixel-level semantic annotation of front scenes. However, most of the datasets are still at small scale relative to the variability of naturalistic driving, and sampled either randomly or based on simple characteristics (the largest one [17] has 100,000 video clips sampled for a variety of weather, time conditions).

On the other hand, the *MIT-AVT Clustered Driving Scene Dataset* utilizes clustering methods to separate common cases and edge cases at a larger scale than prior work.

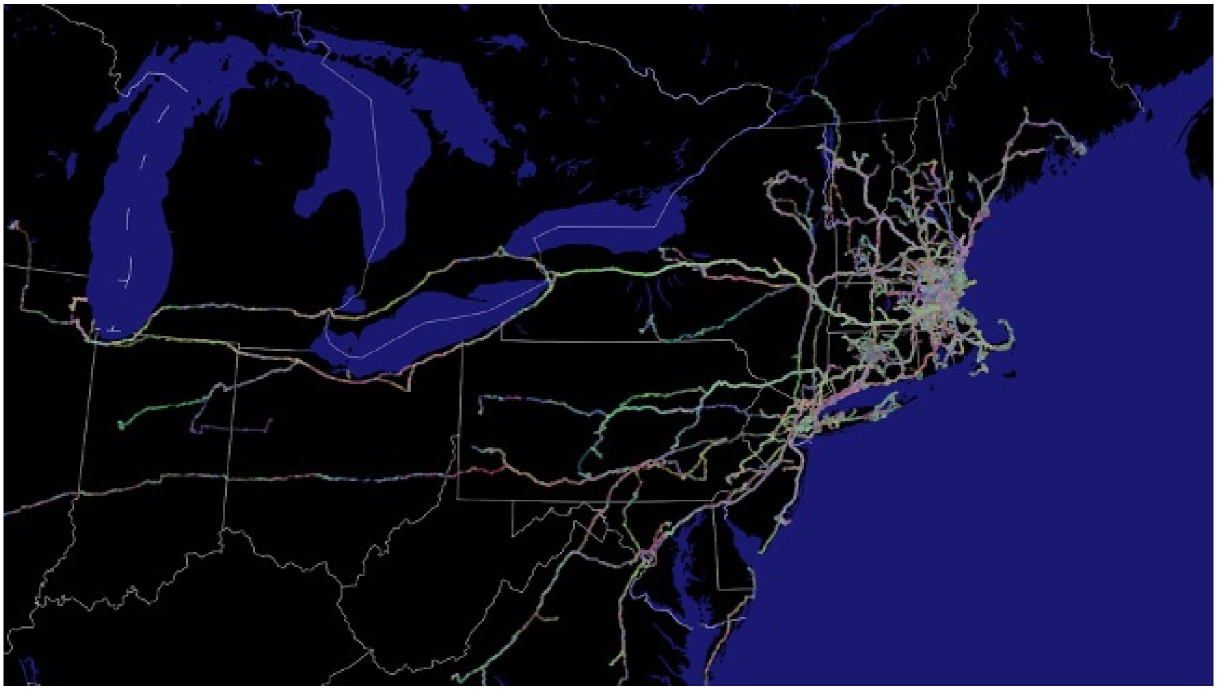


Fig. 2: Visualization of all the GPS locations of clips in *MIT-AVT Clustered Driving Scene Dataset*. Each point represents one clip. All clips from the same cluster are associated with the same color.

III. Visual Embeddings and Clustering

A. Naturalistic Driving Dataset

The *MIT-AVT* study [1] 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. Subjects were enrolled for either medium-term (1 month) or long-term (over a year) study (see [1] for detail). Recorded data includes 720p 30fps videos of: 1) the forward roadway; 2) the driver's face; and 3) the instrument cluster. Telemetry data gathered includes IMU and GPS. The state of certain vehicle subsystems may be obtained from CAN data.

B. Candidate Clip Selection

To reach the goal of creating a representative and edge case sample of the *MIT-AVT* dataset, a method for reducing the trip recordings into unique video segments was implemented using kinematic data. This reduction is necessary for minimizing the computational cost of processing the entire subset of *MIT-AVT* dataset leveraged for this effort, which has over 20,000 hours of video data. A clip length of 10 seconds was chosen in order to provide enough context in a video segment. Clips were sampled from all daylight trips with the first and last 3 minutes removed to reduce footage of static parking scenes and origin / destination information. Each 10-second clip was separated by 0.1 mile, which further reduces the total considered data and contributes to scene diversity. The final number of clips after the above sampling process is 1,156,861.

C. Visual Embeddings

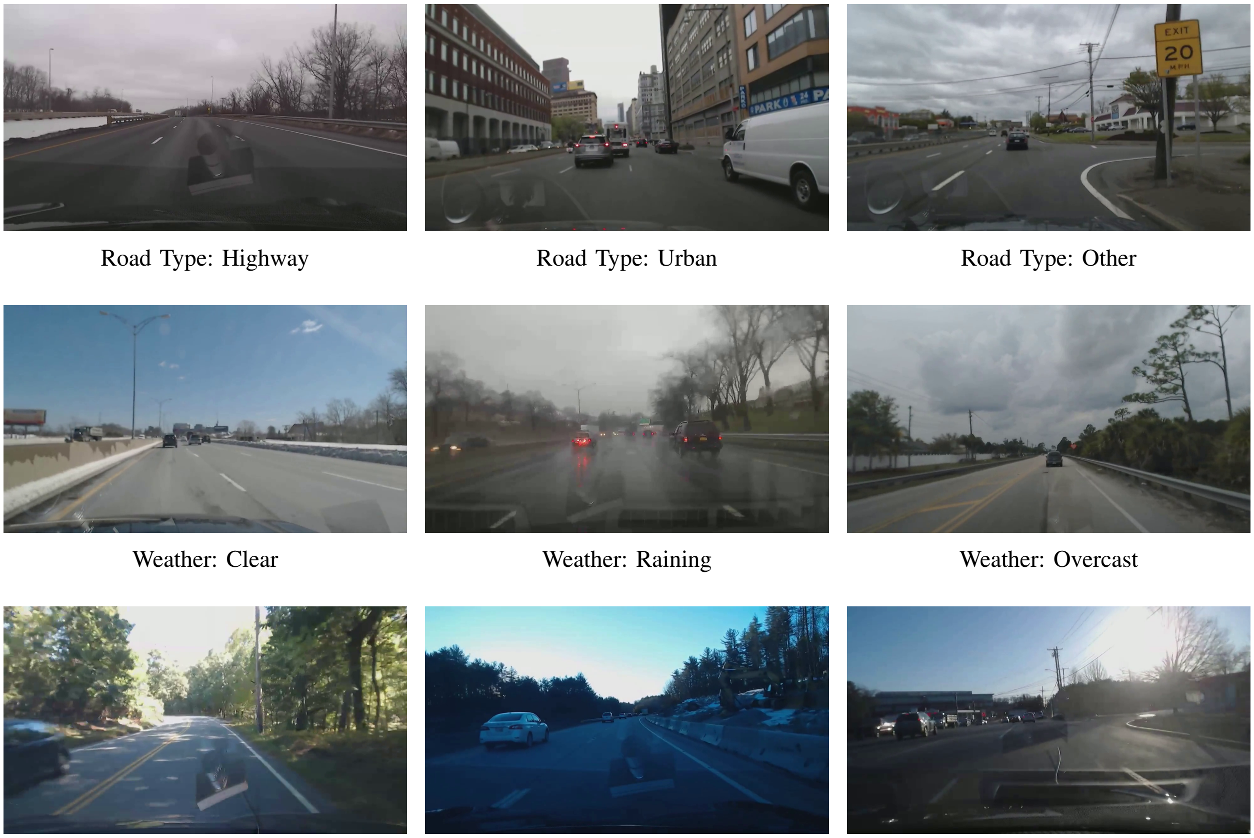
For each 10-second clip, 5 evenly spaced frames from the entire clip were selected. We used the Resnet-50 [2] pre-trained on ImageNet [9] to obtain feature embeddings of 2,048 dimensions for each selected frame and saved to disk, totaling about 107GB. Principal Component Analysis (PCA) was run on disk-mapped memory arrays of the ResNet output, reducing dimensions from 2,048 to 100 in order to minimize the computational cost of further processing. Embeddings for each clip were then combined in chronological order to form final clip feature vectors of 500 dimensions. The final set of clip embeddings occupied a disk size of about 4.5GB.

D. Visual Clusters and Outliers

Mini Batch K-Means clustering [3] was run on all the clip embeddings to obtain 500 clusters. Mini batches were favored over normal K-Means or similar clustering approaches in order to reduce the cost of processing the entire dataset. Local Outlier Factor [4] was then run on the resulting dataset to find outliers in approximately 1% of the data.

E. Automated Metadata Annotation

For each clip, various metadata is extracted from the driving data based on the mean latitude, longitude, speed, and time of day. This metadata can then be combined on a per-cluster basis to find the ranges of speeds, geographical regions and times that are represented by that cluster. Fig. 2 visualizes all the GPS locations of clips associated with each cluster. Major highways have more consistent color, suggesting that most of the clips on a highway are of similar visual appearances and thus grouped into one cluster.



Road Type: Highway

Road Type: Urban

Road Type: Other

Weather: Clear

Weather: Raining

Weather: Overcast

Illuminance: Dappled

Illuminance: Shadowed

Illuminance: Direct

Fig. 3: Snapshots from clusters that are strongly associated with a particular condition (road type, weather, illuminance).

F. Manual Metadata Annotation

In order to validate that the final set of clusters represent a variety of conditions, manual single-pass annotations were performed on 10 randomly selected clips from each cluster. The annotation includes the following categories:

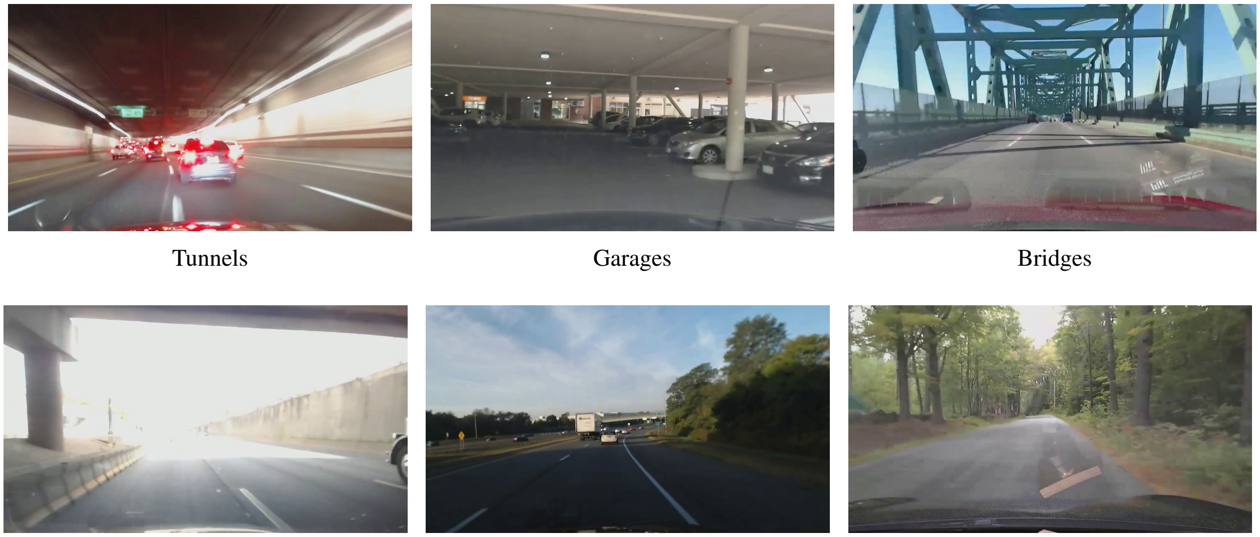
- **Seasons**: summer, fall, winter, varied
- **Weather**: clear, overcast, raining, varied
- **Lanes**: two-lane, varied
- **Simplified Road Type**: highway, other, urban, varied
- **Illuminance**: sunny, overcast, direct, dappled, shadowed, varied
- **Other**: tunnels, bridges, underpasses, garages, trucks, mountains, various other elements of the scene shared across clips of this cluster
- **Unique**: 1 if this cluster represents an uncommon quality (enumerated in 'Other'), 0 otherwise

G. MIT-AVT Clustered Driving Scene Dataset

The final dataset contains 1,156,592 10-second clips, among which 5,601 are outliers (as defined in §III-D). The total size of the dataset is approximately 4TB, with an average clip size of 3.5 Mb. The average number of clips per cluster is about 2,313 with a standard deviation of 1,414. Table I shows counts associated with clusters for which more than 90% randomly sampled clips conformed to a particular manually annotated class.

Label Category	Label	Number of Clusters
Seasons	Summer	93
	Winter	15
	Fall	11
Lighting	Sunny	37
	Overcast	37
	Shadowed	23
	Dappled	19
	Direct	6
Road Type	Highway	158
	Rural	28
	Urban	7
Weather	Clear	62
	Overcast	11
	Raining	8
Other	Wooded	6
	Tunnels	5
	Trucks	4
	Overpasses	2
	Bridges	2
	Garages	1

TABLE I: Number of clusters associated with a particular human-assigned label.



Tunnels Garages Bridges

Overpasses Trucks Wooded Areas

Fig. 4: Snapshots from clusters that are strongly associated with a particular rarely-occurred condition (edge cases).

Fig. 5: Snapshots from outlier clips that have unique visual appearances.

IV. Automated Evaluation of Driving Scene Perception

Multiple mathematical metrics (*e.g.*, IoU [23]) can be used to evaluate deep learning models on the task of semantic scene segmentation when the pixel-level ground truth annotation is available. However, such requirement is hard to fulfill regarding evaluation on large-scale datasets, such as the proposed *MIT-AVT Clustered Driving Scene Dataset*.

Considering the fact that manual full-scene annotation becomes too costly and time-consuming at large-scale, we propose two automated methods for evaluation of scene perception on naturalist driving datasets. The automated methods evaluate certain aspects of the scene perception performance, which are essential to the safety of real-world applications towards both common cases and edge cases.

A. Disagreement between Deep Learning Models

Recent work [24] introduces the Arguing Machine framework showing that the disagreement between different deep learning models can be viewed as a strong signal that is able to capture potential system failures. In addition to the scene perception model that predicts pixel-level semantic labels, we introduce another object detection model that can detect and predict the bounding box over certain objects, including vehicle and person. Theoretically, a pixel being predicted as vehicle by segmentation model should also be within the bounding boxes of the same class predicted by detection model. Based on this assumption, we can evaluate the performance of the segmentation model automatically by calculating the agreement of two models (termed as $Acc.^{(Agreement)}$) as:

$$Acc.^{(Agreement)}(C) = \frac{\sum pixel_{i,j} \in \{boxes^{(C)}\}}{\sum pixel_{i,j}} \quad \forall pixel_{i,j} \in C \tag{1}$$

where C is a joint class label for both models, such as vehicle and pedestrian, and $boxes^{(C)}$ refers to all the bounding boxes predicted as class C.

B. 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

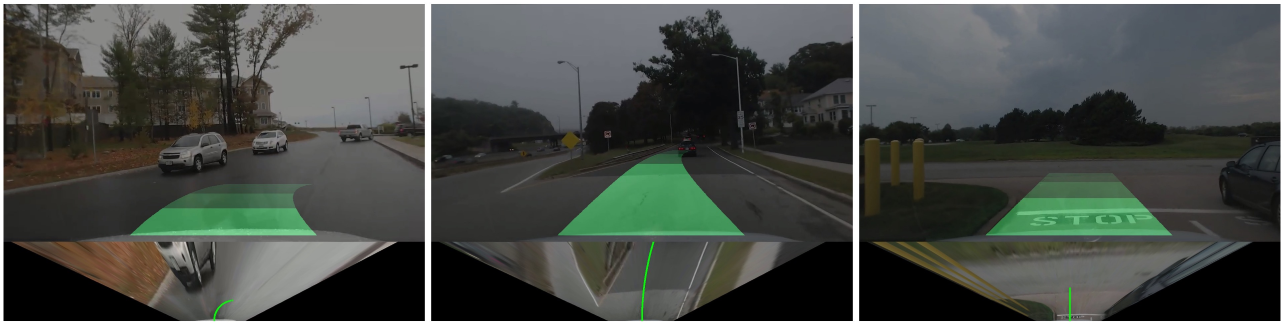


Fig. 6: Examples of inferred drivable area overlaid on the driving scene. The vehicle trajectory is shown on the bottom bird's eye view image, and the drivable area is inferred by perspective transformation. Different levels of the green color represent 1 second, 2 seconds, and 3 seconds of future vehicle trajectory.

transformation. We then calculate the future vehicle trajectory with the steering, speed, and other vehicle specifications, using the following formula:

$$r = \frac{d_{track}}{2} \cdot \frac{1}{\sin(\frac{a_{steer}}{r_{steer}})} \tag{2}$$

where d_{track} is the vehicle track width, a_{steer} is the steering wheel angle, and r_{steer} is the ratio of steering angle to road angle. r is the radius of a curve that shows the future vehicle trajectory. Note that when a_{steer} is close to zero, r will be close to infinite, which simply represents a straight line. We draw the vehicle trajectory on the bird's eye view of front scene, and project the single-lane area back onto the front scene image. Fig. 6 shows the examples of inferred drivable area overlaid on the front camera image.

The inferred drivable area can be used for automated evaluation of segmentation models. We calculate the accuracy of pixels within the 2 seconds of the future drivable area being classified as either road or vehicle, as:

$$Acc.^{(Drivable)} = \frac{\sum pixel_{i,j} \in \{road, vehicle\}}{\sum pixel_{i,j}} \tag{3}$$
$$\forall pixel_{i,j} \in drivable\_area$$

The $Acc.^{(Drivable)}$ serves as a strong indicator to evaluate the segmentation prediction in edge cases where the segmentation model predicts the drivable area as other classes, suggesting that it is very likely to observe system failure. When further being used with a vehicle detector that can detect and remove vehicles in the current driving path, this method can yield near-perfect precision on road or drivable area detection.

C. Experiments

To further illustrate how the proposed $Acc.^{(Agreement)}$ and $Acc.^{(Drivable)}$ can be used to automatically evaluate a scene perception model with naturalistic driving datasets, we hereby show experiments to evaluate a state-of-the-art scene segmentation model using the *MIT-AVT Clustered Driving Scene Dataset*. We choose DeeplabV3 [25] pre-trained on the Cityscapes [23] dataset in this experiment. The object detector for $Acc.^{(Agreement)}$ calculation is YOLOv3 [26]. For $Acc.^{(Drivable)}$, we sample a single vehicle from the entire dataset and calibrate the parameters for drivable area projection.

Fig. 7 shows illustrative examples of the evaluation. As we see, first, the model performs reasonably well on common cases, but totally fails on some edge cases. The failures can be well captured by the proposed automated evaluation methods by setting thresholds on $Acc.^{(Agreement)}$ and $Acc.^{(Drivable)}$. For further illustration, we generate the pseudo "ground truth" (bottom row) relabeled from segmentation predicitions that have perfect $Acc.^{(Agreement)}$ and $Acc.^{(Drivable)}$. Such ground truth shows what kind of prediction may fulfill both automated metrics at the same time. By comparing this "ground truth" (bottom row) with the prediction (second row), it shows that obvious failures (*e.g.*, predicting most of the scene as vehicle) can be automatically captured by the proposed metrics.

V. Conclusion and Future Work

In this work, we aim at evaluating driving scene perception systems for both representative common cases and uncommon edge cases in real-world naturalistic driving scenarios. We present the *MIT-AVT Clustered Driving Scene Dataset*, a large-scale naturalistic driving dataset of 1,156,592 10-second video clips. The dataset is organized using a vision-based clustering approach that determines common cases and edge cases in driving scenes, and manually annotated with environmental characteristics for each cluster. We further develop two methods for automated evaluation of scene perception systems on naturalist driving datasets.

The proposed dataset has been used for ongoing research on driving scene perception and naturalistic driving study. Video illustrations of select clusters will be released to help with future research.

Acknowledgment

This work is supported by the Toyota Collaborative Safety Research Center (CSRC) and the MIT Advanced Vehicle Technology (AVT) Consortium. The views and conclusions being expressed as those of the authors and may not necessarily represent those of sponsoring organizations.

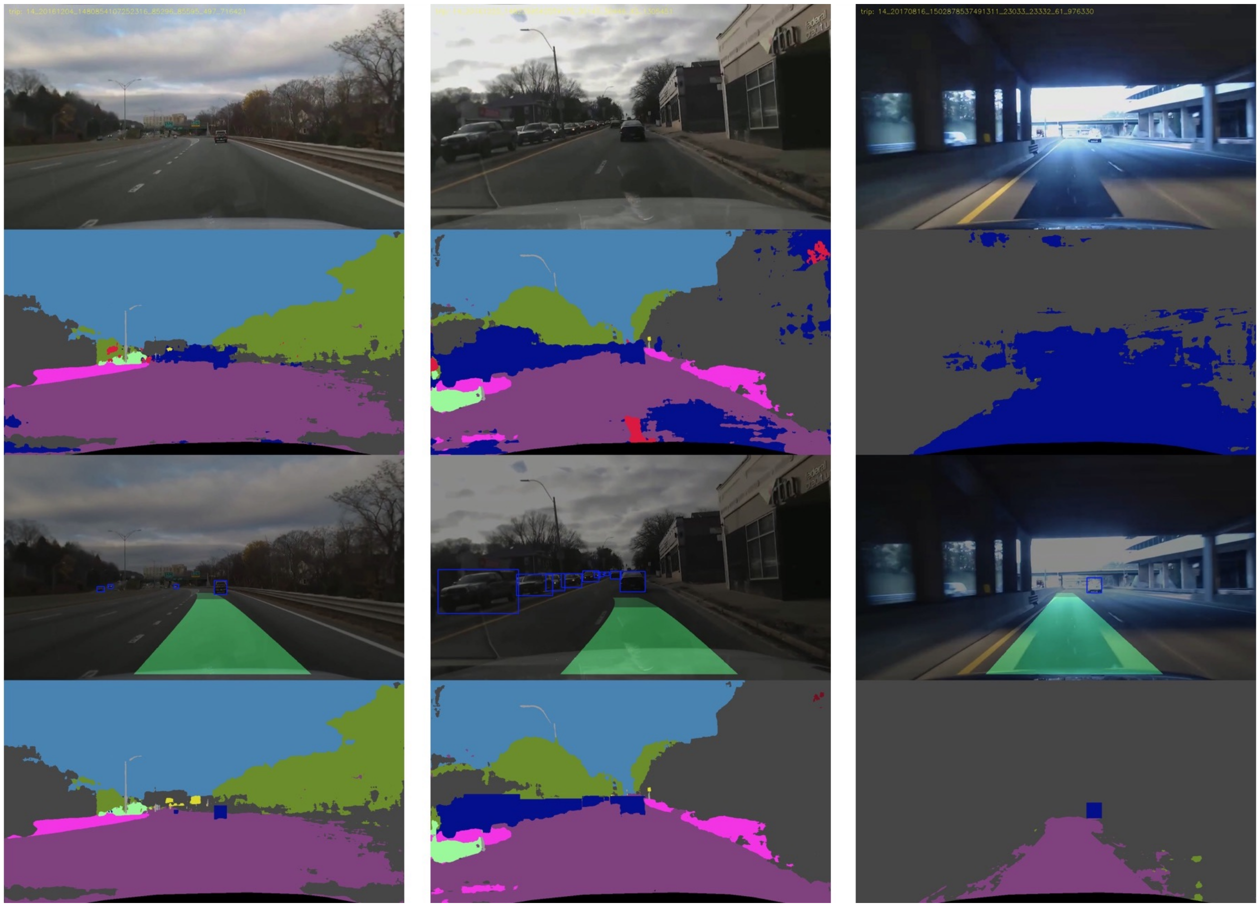

trip: 14_20161204_1480854107252316_85298_85595_497_716421
trip: 14_20170816_1502878537491311_23033_23332_61_976330

Fig. 7: Examples evaluating DeeplabV3 model [25] on the *MIT-AVT Clustered Driving Scene Dataset*. From left to right: common case (highway, clear), less common case (urban, overcast), edge case (overpasses). From top to bottom: front scene input, scene segmentation prediction, detection and drivable area results for $Acc.^{(Agreement)}$ and $Acc.^{(Drivable)}$, relabeled segmentation "ground truth" for perfect $Acc.^{(Agreement)}$ and $Acc.^{(Drivable)}$.

References

[1] L. Fridman, D. E. Brown, M. Glazer, W. Angell, S. Dodd, B. Jenik, J. Terwilliger, A. Patsekin, J. Kindelsberger, L. Ding, *et al.*, "Mit advanced vehicle technology study: Large-scale naturalistic driving study of driver behavior and interaction with automation," *IEEE Access*, vol. 7, pp. 102 021–102 038, 2019.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 1177–1178.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[6] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4733–4742.

[7] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1385–1392.

[8] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[10] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[11] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[12] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1943–1950.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[14] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[15] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1641–1650, 2003.

[16] L. Duan, L. Xu, Y. Liu, and J. Lee, "Cluster-based outlier detection," *Annals of Operations Research*, vol. 168, no. 1, pp. 151–168, 2009.

[17] 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," *CoRR*, vol. abs/1805.04687, 2018. [Online]. Available: http://arxiv.org/abs/1805.04687

[18] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder, "The

mapillary vistas dataset for semantic understanding of street scenes," in *International Conference on Computer Vision (ICCV)*, 2017. [Online]. Available: https://www.mapillary.com/dataset/vistas

[19] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset," 2015.

[20] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[21] 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.

[22] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[23] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] L. Fridman, L. Ding, B. Jenik, and B. Reimer, "Arguing machines: Human supervision of black box ai systems that make life-critical decisions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[25] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[26] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.