# A Benchmark for Cross-Weather Traffic Scene Understanding

Shuai Di[1,2], Honggang Zhang[1], Xue Mei[3], Danil Prokhorov[3], and Haibin Ling[2]

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[2]Center for Data Analytics & Biomedical Informatics, Dept. of Computer & Information Sciences, Temple University, Philadelphia, USA
[3]Toyota Research Institute, North America, Ann Arbor, USA

***Abstract*— Understanding traffic scene images taken from vehicle mounted cameras provides important information for high level tasks such as autonomous driving and advanced driver assistance. The problem is far from trivial especially due to challenges from weather and illumination variation. To facilitate the research against such challenges, in this paper we present a new benchmark for cross-weather traffic scene understanding[1]. The dataset consists of 1,356 traffic scene images collected at 226 different locations. For each location, there are six images taken by a vehicle mounted camera under different weather/illumination conditions including sunny day, night, snowy day, rainy night, cloudy day I and cloudy day II. We manually annotated each image with scene understanding labels such as road, sky, building, etc. To the best of our knowledge, this is the first carefully collected benchmark for cross-weather traffic scenes. In addition, we also provide results from two popular scene parsing systems as the baselines. We expect the benchmark to help boost research in improving the robustness of traffic scene understanding algorithms.**

## I. INTRODUCTION

With the development of autonomous driving and driver assistance systems [21], [22], [23], [24], visual sensors play an increasingly important role in related tasks. In particular, reliable and efficient understanding of road scene images, which are taken from vehicle mounted cameras, often serves as an important basis for high level situation awareness tasks. While traffic scene understanding in the condition of well-structured roads is already available in modern vehicles, it remains an unsolved problem in terms of system reliability e.g., traffic scene understanding influenced by various illumination and weather conditions.

In this paper, we propose a traffic scene dataset, which includes different illumination and weather scenarios of the same road route. In short, 1,356 traffic scene images of 226 different locations are included in our proposed dataset. At each location, images of 6 different scenarios were captured respectively i.e., sunny day, night, snowy day, rainy night, cloudy day I and cloudy day II. All images are manually annotated by LabelMe [1]. Note that the illumination is different between cloudy day I and cloudy day II. For our proposed dataset, we intend to understand traffic scene by using label transfer e.g., understanding traffic scene in the challenging scenarios with the help of scenes of the same location, but different time periods. More specifically, given a test scene image, the similar images named as nearest neighbor set are first found by using similarity measure in the Convolutional Neural Network (CNN) based feature space. Then, dense correspondences between the test image and each of the nearest neighbors are computed. Finally, labels of images in the nearest neighbor set are transferred to the test image based on the dense correspondences between them.

The contributions of this paper can be concluded as follows. We first propose a new traffic scene dataset. To our best knowledge, this is the first traffic scene dataset which contains various illumination and weather conditions of the same road route. In addition, we benchmark the state-of-the-art dense correspondence methods on our proposed dataset in order to understand traffic scenes via label transfer.

The rest of the paper is organized as follows. After summarizing related work in Section 2, the description of our proposed dataset is given in Section 3. Then we compare our dataset with existing traffic scene datasets in Section 4. The evaluation for baseline methods on our dataset is presented in Section 5, followed by the conclusion in Section 6.

## II. RELATED WORK

### A. Traffic Scene Datasets

Existing traffic scene related datasets can be roughly divided into two categories i.e., datasets used for traffic scene understanding and datasets for road/lane marking detection. The former one usually contains several video sequences taken by a driving car, such as the CamVid dataset [13], [14], daimler urban segmentation dataset [15], [16], etc. Several densely labeled frames are also included in these datasets for scene understanding algorithms evaluation. As for the latter category, besides the traffic scene videos, the annotated road areas or lane markings are included for evaluation e.g., road/lane detection evaluation [18] and dataset for road area estimation [19].

The proposed dataset in this paper belongs to the former category. However, different scenarios of the same road route are included in our dataset, which is different from existing traffic scene related datasets.

### B. Scene Understanding via Label Transfer

Scene understanding based on label transfer is the non-parametric method such as [12], [25], [26], [27], [3], [4], [5]. Given a test image, a set of nearest neighbors that share similar scene configuration are retrieved using scene retrieval techniques. Then, classification confidence maps are obtained

[1]The dataset is available for research exploration at .


Fig. 1. Example images from the proposed benchmark. Six traffic scene images of the same location were captured via a vehicle mounted camera under varying weather and illumination conditions.

by matching the test image with the images in the nearest neighbor set. Finally, the final semantic labels of the test image are obtained by solving a MRF model.

In this paper, our focus is not on the state-of-the-art traffic scene understanding methods but the application on our proposed dataset. We have noticed that image retrieval plays an important role in these nonparametric methods. CNN based models have been the top performers in scene recognition [11], [29], [32], [31]. In particular, as shown in recent works [6], [7], [8], [9], deep CNN features learned on a large dataset, such as ImageNet (ILSVRC) [10] and Places [29], [30], can be used as the powerful descriptors applicable to other datasets.

## III. DATASET DESCRIPTION

### A. Data Collection

A GoPro HERO3+ camera was mounted forward facing on the car dashboard for recording videos. All six videos share the same road route, but different time periods i.e., sunny day, night, snowy day, rainy night, cloudy day I and cloudy day II. Both urban and highway traffic scenes are included in this road route through the city center and suburb of Philadelphia in Pennsylvania, USA. Images in our dataset are from these six videos. More Specifically, we select one image for each of the 226 different locations in each of the six videos. Therefore, for each location, six different images are included, as illustrated in Figure 1. Finally, we obtain totally 1,356 images for 226 different locations. All images were cropped to remove the dashboard i.e., $856 \times 270$ pixels.

### B. Data Annotation

Most object categories were manually annotated in the scenarios of sunny day, night, snowy day, rainy night, cloudy day I and cloudy day II by LabelMe [1]. The remaining objects in all these scenarios were annotated as undefined class because of indeterminate objects or poor visibility. More specifically, 13 object categories are included i.e., sky, building, tree, car, road, median strip, bridge, wiper (device that wipes rain from vehicle's windshield), vegetation, traffic sign, pole, traffic lights and pedestrian. Any other object categories are classified as the 14th category: undefined. The statistics of the annotated object categories for each scenario are shown in Figure 2. As can be seen in Figure 2, for the night scenario, undefined category occupies larger percentage comparing to other scenarios. The reason is that the visibility of night condition is very low and categories in some images are very difficult to be seen. Hence, the undefined class is assigned to them. In addition, undefined class in the rainy night scenario occupies a little larger percentage comparing with daytime scenario but smaller percentage than night scenario. The reason is that some categories in a few images of the rainy night scenario is also hard to be recognized because of the low visibility and the undefined class is assigned to them. What is more, the visibility of the rainy night is better than the night scenario. Therefore, more object categories can be annotated in the rainy night condition. Note that, to improve the annotation accuracy, images of the same location in the other four scenarios are used as reference guide when annotating images in the night and rainy night scenarios.

## IV. DATASET ANALYSIS

In this section, we compare our dataset with existing traffic scene related datasets. As introduced in related work, we divide the existing traffic scene related datasets into two classes: datasets used for scene understanding and datasets for road/lane marking detection.

### A. Traffic Scene Understanding Datasets

The CamVid dataset [13], [14] consists of daytime and dusk videos taken from a car driving through Cambridge of England. A total of 701 densely labeled frames (11 class labels) are included in the dataset. The daimler urban segmentation dataset [15], [16] consists of video sequences recorded in urban traffic. There are 500 labeled frames (5 class labels) in this dataset. The street scenes dataset [20] were taken from a camera at Boston of USA. Some object categories (totally 9 classes) were manually labeled for each image in the dataset.

Our dataset also consists of video sequences taken from a car. However, as can be seen in Figure 1, six different scenarios of the same location are included in our dataset i.e., different sequences of the same road route not just several different sequences. There are 1,356 labeled traffic scene images (13 class labels) in our dataset. We mainly focus on understanding one scenario of the traffic scene by

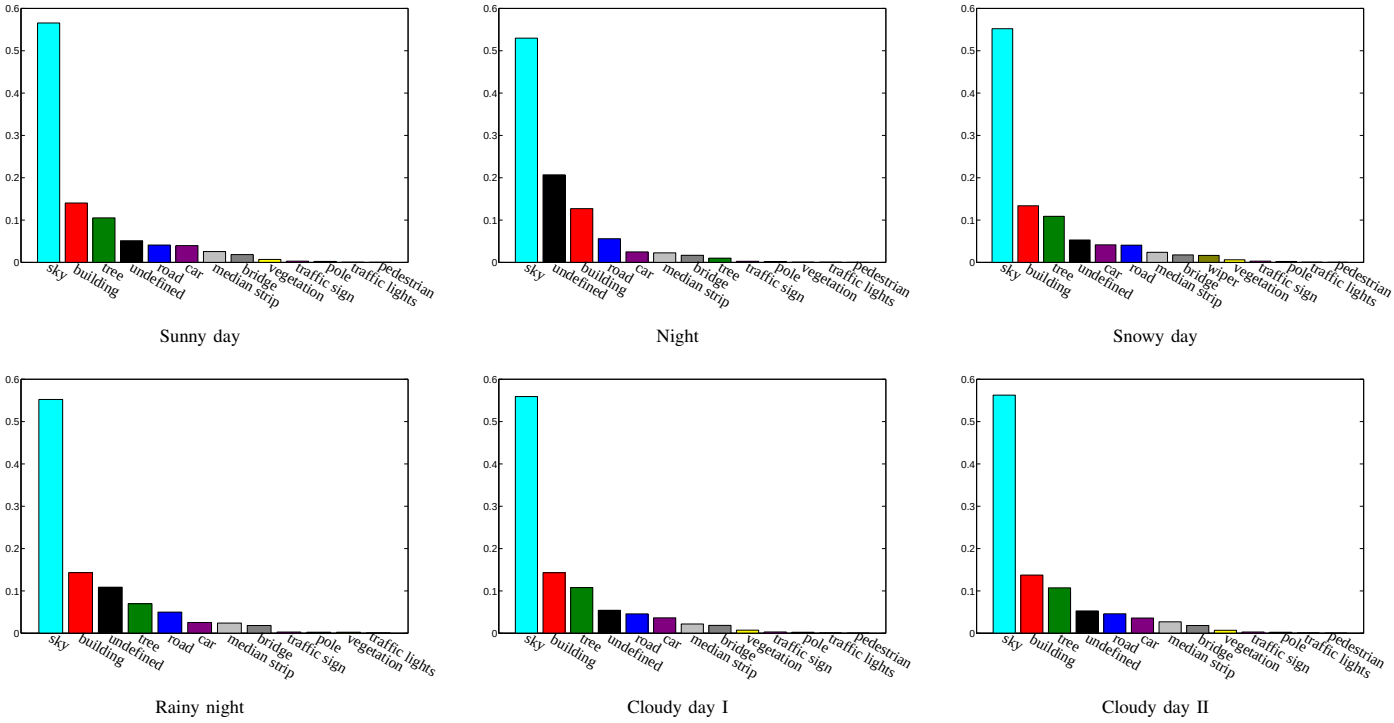


Fig. 2. Statistics for the annotation results of our proposed traffic scene dataset with 13 object categories (sky, building, tree, car, road, median strip, bridge, wiper, vegetation, traffic sign, pole, traffic lights and pedestrian). Any other things are annotated as undefined.


Fig. 3. Example images from [19]. There are only two different scenarios in this dataset i.e., sunny day and daytime after raining. Unlike it, more challenging scenarios are included in our dataset.

transferring information from other scenarios of the same road route.

### B. Road/Lane Marking Detection Datasets

There are also datasets for performance evaluation of road marking extraction algorithms [17], road/lane detection evaluation [18] and dataset for road area estimation [19]. Images of the dataset introduced in [19] look similar to ours, but only have two different scenarios of the same road route, i.e., sunny day and daytime after raining, as illustrated in Figure 3. Our dataset includes more challenging scenarios of the same road route. In addition, our focus is on scene understanding not road area estimation e.g., most objects being annotated for each image in our dataset not just road area. For comparison, the basic statistics of these traffic scene related datasets are concluded in Table 1.

## V. BENCHMARK EVALUATION

We intend to understand the traffic scene by transfer i.e., understanding one scenario via label transfer from the other scenarios. Given one test image, we first need to find similar images as nearest neighbor set whose labels would be transferred to the test image. Then, the dense correspondences should be established between the test image and images in the nearest neighbor set. Our focus is on the cross-weather traffic scene understanding. Scenarios of cloudy day I and cloudy day II are integrated into one weather condition i.e., cloudy in the evaluation, even though the illumination is different between them. Therefore, five different scenarios (sunny day, snowy day, cloudy day, night and rainy night) are used in the evaluation.

### A. Evaluation Metrics

For evaluation, we use the average pixel-wise and per-class recognition rate, which are commonly used as measures for scene understanding system [3], [4], [5].

### B. Scene Retrieval

Given an image $I$, scene retrieval is to retrieve a set of traffic scene images in an archived dataset that are visually similar to $I$. Denoting the resulting image set by $R$, it is created according to a similarity measure $m(I, I_d)$ between query image $I$ and the database images $I_d$ and retrieve top-N most similar images $R = \{I_d^1, I_d^2, ..., I_d^N\}$. Here, given an image in one scenario, we intend to find the other images of the same location. However, as can be seen in Figure 1, images of the same location undergo drastic changes

**Table 1. Statistics of the traffic scene related datasets**

| Datasets | Intention | Original data | Number of labeled images (number of classes) |
|---|---|---|---|
| Ours | scene understanding | video sequences of the same road route | 1,356 (13 classes) |
| [13], [14] | scene understanding | video sequences captured through a city | 701 (11 classes) |
| [15], [16] | scene understanding | video sequences recorded in urban traffic | 500 (5 classes) |
| [20] | scene understanding | images taken in a city | 3,547 (9 classes) |
| [17] | road marking extraction evaluation | images taken in various sites | 116 (2 classes) |
| [18] | road/lane detection evaluation | video sequences captured through a city | 289 (2 classes) |
| [19] | road area estimation | images taken in the same road route | 1,005 (1 classes) |

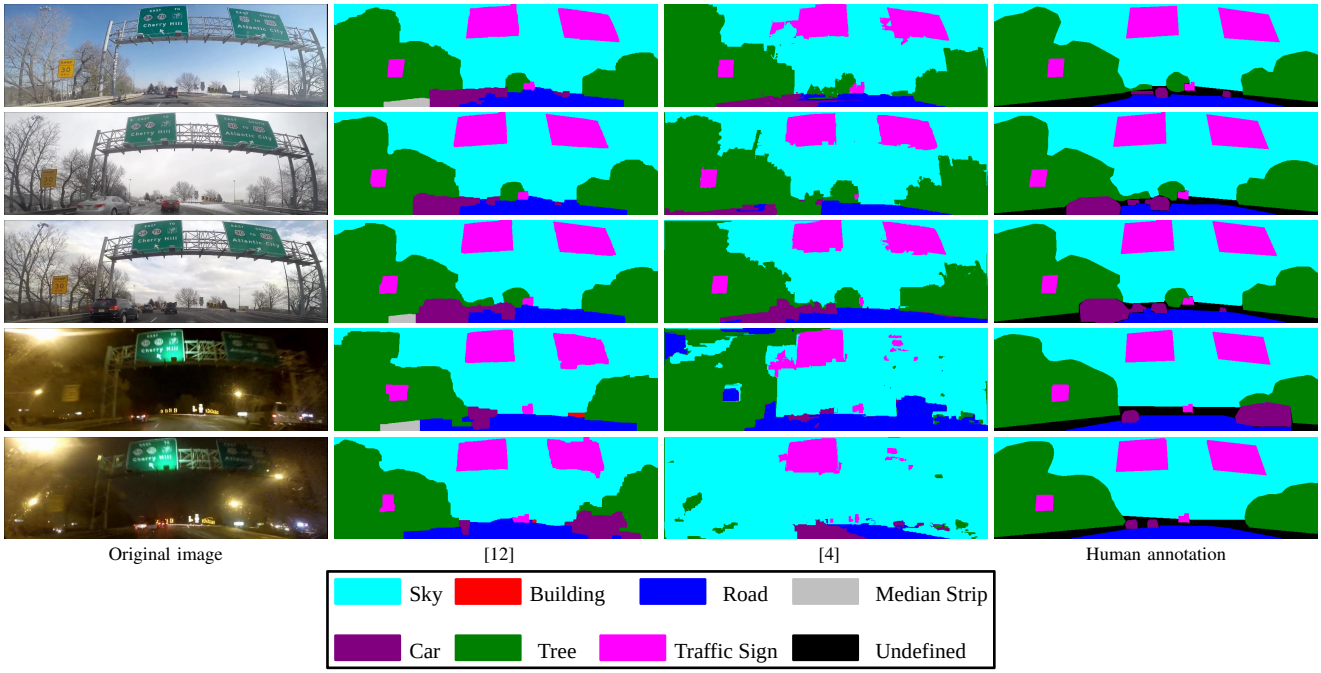


Fig. 4. Some representative scene understanding results of the highway (I).

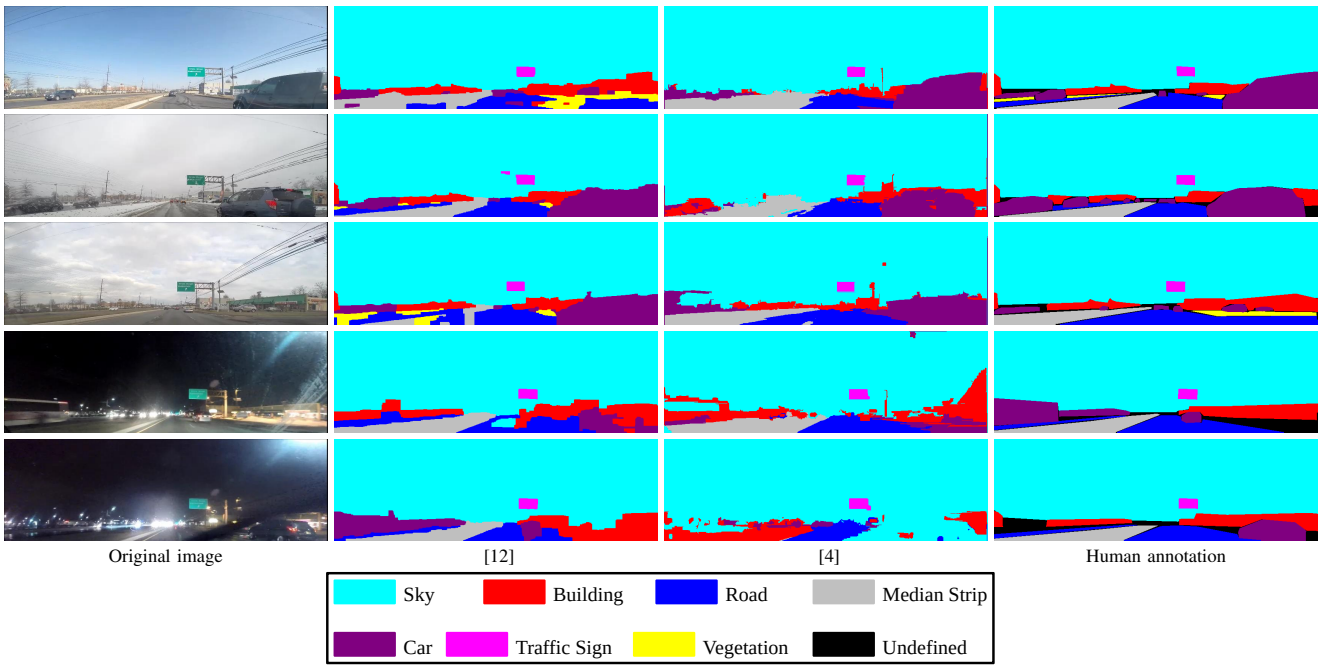


Fig. 5. Some representative scene understanding results of the highway (II).

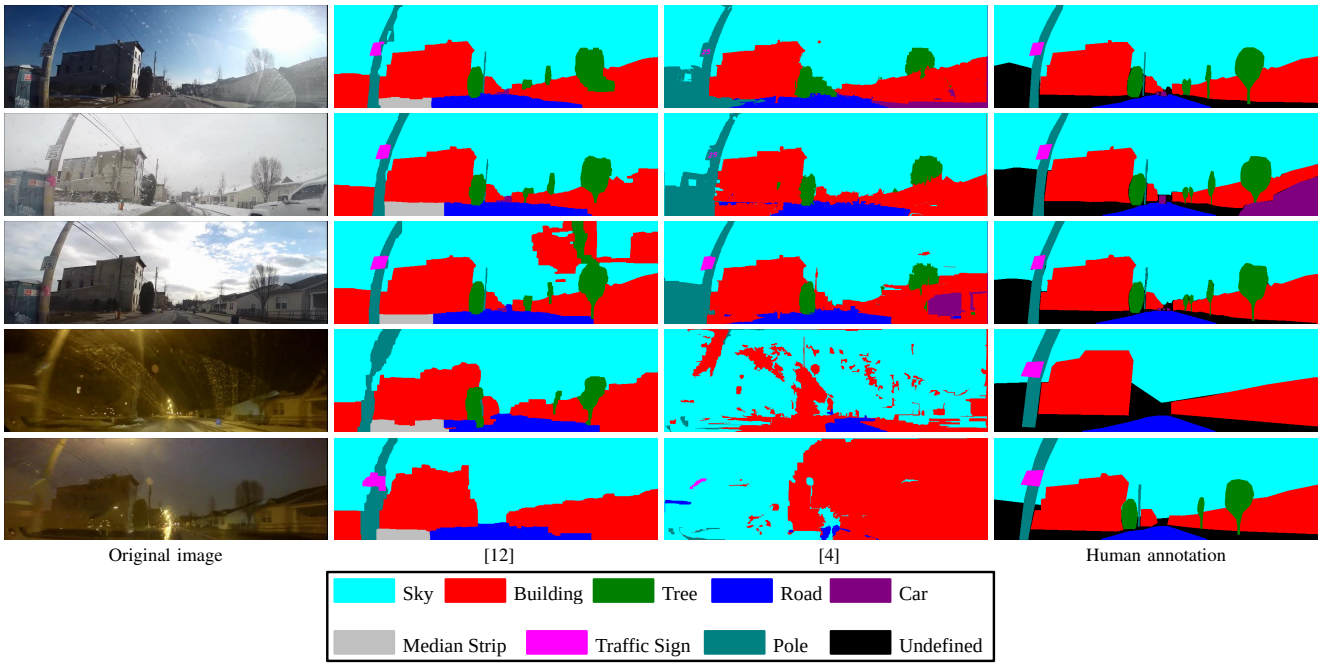


Fig. 6. Some representative scene understanding results of the urban traffic.

in appearance due to illumination, rain, snow, etc. Hence, robust feature representation will be helpful for computing the similarity measure. Deep CNN features learned on a large dataset can be used as the powerful descriptors applicable to other datasets. In this paper, all the traffic scene image representations transfer from the model pre-trained on the large dataset. Specifically, all the traffic scene image representations transfer from the output of the last convolutional layer of AlexNet [11] pre-trained on ILSVRC-2012. Therefore, given a test image in one scenario, the similarity between the test image and the database images is computed respectively in the deep feature space.

### C. Label Transfer

Our aim is to transfer the labels of the retrieved similar images to the input image. It is important to find the dense correspondences between them. As shown in [2], [12], SIFT flow is able to establish semantically meaningful correspondences between two images by matching local SIFT representations. In order to reduce the computational burden for large amount of pixel-level inferences. An alternative strategy, as explored in [27], [3], [4], [5], is to work on superpixels. In this paper, representative systems of the two kinds of methods are validated on our dataset i.e., SIFT flow [2] and superpixel matching [4].

After similar images are retrieved and the dense correspondences between them are available. The labels of the retrieved similar images can be transferred to the input image by using the dense correspondences.

### D. Results

To make full use of traffic scene images in our dataset, cross-validation method is applied for all the experiments in this paper. Specifically, for the five scenarios of our dataset, images in one scenario are taken as test set and images in another four scenarios as training set. We retrieve $N = 4$ images for each test image. The average per-class and per-pixel rates for each scenario are shown in Table 2 and Table 3. Note that Sunny Day in the Table means this scenario is taken as test set and other four scenarios are taken as training set, and vice versa. We intend to verify the performance of traffic scene understanding via label transfer. It would be meaningful to compute the recognition rates for categories which are jointly owned by different scenarios of the road route. For example, as can be seen in Figure 2, the wiper class is only included in snowy day and wiper in this scenario may not find labels to be transferred in other scenarios. Hence, the recognition rates of 11 classes owned by all scenarios are reported in Table 2 and Table 3 i.e., the recognition rates for pedestrian and wiper not being included. A major challenge for traffic scene understanding is induced by the non-uniform statistics of object categories in the scene. Specifically, "stuff" classes, such as Sky, Road, Building, Tree, etc, occupy a large percentage of the image pixels, which have no consistent shape but consistent texture. Whereas, "thing" categories which are characterized by overall shape constitute the minority of all image pixels e.g., Car, Traffic Lights and Traffic Sign. As can be seen in Table 2, for the daytime scenario i.e., sunny day, snowy day and cloudy day, both stuff categories (Sky, Building, Tree, Road and Vegetation) and thing classes (Median Strip, Bridge, Traffic Sign, Pole and Traffic Lights) have good recognition performance. However, for the rainy night scenario, both the average per-class and per-pixel recognition rates are lower with respect to the daytime scenario. The main reason is that the visibility of night scenario is very low and the

**Table 2. Recognition results on our dataset (%) [12]**

| | Sky | Building | Tree | Road | Car | Median Strip | Bridge | Vegetation | Traffic Sign | Pole | Traffic Lights | Per-class | Per-pixel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sunny Day | 98.6 | 93.1 | 92.2 | 84.7 | 38.5 | 88.3 | 97.6 | 80.5 | 94.7 | 82.9 | 78.3 | 84.5 | 93.0 |
| Snowy Day | 96.3 | 92.1 | 90.6 | 83.1 | 37.3 | 82.1 | 96.5 | 78.2 | 91.9 | 79.2 | 75.5 | 82.1 | 90.0 |
| Rainy Night | 95.7 | 83.7 | 12.8 | 78.3 | 47.6 | 76.0 | 84.0 | 20.5 | 85.8 | 54.6 | 75.8 | 65.0 | 83.6 |
| Night | 95.4 | 87.7 | 77.8 | 78.8 | 44.0 | 81.9 | 92.1 | 88.9 | 85.4 | 57.2 | 80.4 | 79.1 | 88.8 |
| Cloudy Day | 92.8 | 92.0 | 93.2 | 82.6 | 44.9 | 75.2 | 95.9 | 74.2 | 88.3 | 78.9 | 75.7 | 81.2 | 89.3 |

**Table 3. Recognition results on our dataset (%) [4]**

| | Sky | Building | Tree | Road | Car | Median Strip | Bridge | Vegetation | Traffic Sign | Pole | Traffic Lights | Per-class | Per-pixel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sunny Day | 98.2 | 85.5 | 90.8 | 74.1 | 33.0 | 72.5 | 92.7 | 51.6 | 81.7 | 19.2 | 23.3 | 65.7 | 89.4 |
| Snowy Day | 97.3 | 86.7 | 85.0 | 73.2 | 40.5 | 62.6 | 88.1 | 44.5 | 76.1 | 14.6 | 19.7 | 62.6 | 85.2 |
| Rainy Night | 93.4 | 60.7 | 8.1 | 39.3 | 37.8 | 33.3 | 74.0 | 11.2 | 46.2 | 9.5 | 21.7 | 39.6 | 73.1 |
| Night | 88.5 | 63.6 | 29.0 | 45.1 | 31.5 | 56.9 | 82.8 | 54.0 | 54.3 | 5.8 | 35.5 | 49.7 | 74.5 |
| Cloudy Day | 95.3 | 90.2 | 90.3 | 80.2 | 47.8 | 75.3 | 96.1 | 32.7 | 79.4 | 17.0 | 47.5 | 68.3 | 90.1 |

undefined class is assigned to many objects. Hence, there are not enough labels to be transferred in this scenario. More specifically, given a test image in the rainy night scenario, it is more easier to find the similar image in the night scenario comparing with daytime scenario. However, there is not enough label information to be transferred from the night scenario. Unlike the rainy night scenario, the night scenario has good recognition performance. The reason is that many undefined categories are included in the night scenario and less instances are counted in computing the recognition rate.

The similar situation can be found in Table 3. Note that, the recognition rates in Table 3 are lower than Table 2. The main reason is that appearances of traffic scene images in our dataset undergo large changes and it is hard for superpixel segmentation [28]. Hence, superpixels tend to fragment objects specifically for the thing classes. For example, as can be seen in Table 3, the recognition rates of pole and traffic lights have a sharp decline with respect to rates in Table 2. We present some qualitative results of the two baseline methods for each scenario in Figure 4-6.

## VI. CONCLUSIONS

In this paper we proposed a benchmark for cross-weather traffic scene understanding. Our dataset includes six different weather/illumination scenarios for each of 226 locations. In addition, we ran two popular scene understanding algorithms on the dataset and provided baseline results. The proposed benchmark is expected to help future studies on traffic scene understanding as well related problems such as cross-weather image matching and image retrieval.

## REFERENCES


[1] B. Russell and A. Torralba and K. Murphy and W. T. Freeman, LabelMe: a Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision (IJCV), 77(1):157-173, 2008.

[2] C. Liu, J. Yuen, and A. Torralba, SIFT Flow: Dense Correspondence across Scenes and its Applications. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 33(5): 978-994, 2011.

[3] J. Tighe and S. Lazebnik. Finding Things: Image Parsing with Regions and Per-Exemplar Detectors. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[4] J.M. Yang, B. Price, S. Cohen and M.H. Yang. Context Driven Scene Parsing with Attention to Rare Classes. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[5] J. Tighe, M. Niethammer and S. Lazebnik. Scene Parsing with Object Instances and Occlusion Ordering. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[6] J. Donahue, Y.Q. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. International Conference on Machine Learning (ICML), 2014.

[7] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. European Conference on Computer Vision (ECCV), 2014.

[8] K. Chatfield, K. Simonyan, A. Vedaldi and A. Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Nets. British Machine Vision Conference (BMVC), 2014.

[9] A. S. Razavian, H. Azizpour, J. Sullivan and S. Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. IEEE Conference on Computer Vision and Pattern Recognition DeepVision Workshop (CVPR Workshop), 2014.

[10] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and F.F. Li, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[11] A. Krizhevsky, I. Sutskever and G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. Conference on Neural Information Processing Systems (NIPS), 2012.

[12] C. Liu, J. Yuen, and A. Torralba, Nonparametric Scene Parsing via Label Transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 33(12):2368-2382, 2011.

[13] G. J. Brostow, J. Shotton, J. Fauqueur and R. Cipolla. Segmentation and Recognition Using Structure from Motion Point Clouds. European Conference on Computer Vision (ECCV), 2008.

[14] G.J. Brostow, J. Fauqueur, R. Cipolla. Semantic Object Classes in Video: A High-Definition Ground Truth Database. Pattern Recognition Letters 30(2):88-97, 2009.

[15] T. Scharwachter, M. Enzweiler, S. Roth, and U. Franke. Stixmantics: A Medium-Level Model for Real-Time Semantic Scene Understanding. European Conference on Computer Vision (ECCV), 2014.

[16] T. Scharwachter, M. Enzweiler, S. Roth, and U. Franke. Efficient Multi-Cue Scene Segmentation, Conference of the German Conference on Pattern Recognition (GCPR), 2013.

[17] T. Veit, J.P. Tarel, P. Nicolle and P. Charbonnier. Evaluation of Road Marking Feature Extraction. IEEE International Conference on Intelligent Transportation Systems (ITSC), 2008.

[18] J. Fritsch, T. Kuehnl and A. Geiger. A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms. IEEE International Conference on Intelligent Transportation Systems (ITSC), 2013.

[19] J. M. Alvarez and A. M. Lopez. Road Detection Based on Illuminant Invariance. IEEE Transactions on Intelligent Transportation Systems (TITS), 12(1):184-193, 2011.

[20] S.M. Bileschi. Streetscenes: Towards Scene Understanding in Still Images. Doctoral Dissertation, Massachusetts Institute of Technology (MIT), 2006.

[21] M. Buehler, K. Iagnemma and S. Singh. The 2005 DARPA Grand Challenge:the Great Robot Race. Springer Tracts in Advanced Robotics. Springer, 2007

[22] M. Buehler, K. Iagnemma and S. Singh, editors. The DARPA Urban Challenge: Autonomous Vehicles in City Traffic. Springer Tracts in Advanced Robotics. Springer, 2009

[23] J. C. McCall and M. M. Trivedi. Video-Based Lane Estimation and Tracking for Driver Assistance: Survey, System, and Evaluation. IEEE

Transactions on Intelligent Transportation Systems (TITS), 7(1):20-37,2006.

[24] Q. Zou, H.B. Ling, S.W. Luo, Y.P. Huang and M. Tian. Robust Nighttime Vehicle Detection by Tracking and Grouping Headlights. IEEE Transactions on Intelligent Transportation Systems (TITS), 16(5):2838-2849, 2015.

[25] D. Eigen and R. Fergus. Nonparametric Image Parsing Using Adaptive Neighbor Sets. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[26] G. Singh and J. Kosecka. Nonparametric Scene Parsing with Adaptive Feature Relevance and Semantic Context. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[27] J. Tighe and S. Lazebnik. Superparsing: Scalable Nonparametric Image Parsing with Superpixels. International Journal of Computer Vision (IJCV), 101:329-349, 2013.

[28] P.F. Felzenszwalb, D.P. Huttenlocher. Efficient Graph Based Image Segmentation. International Journal of Computer Vision (IJCV) 2(2): 1-26, 2004.

[29] B.l. Zhou, A. Lapedriza, J.X. Xiao, A. Torralba and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. Conference on Neural Information Processing Systems (NIPS), 2014.

[30] B.L. Zhou, A. Khosla, A. Lapedriza, A. Torralba and A. Oliva. Places2: A Large-scale Database for Scene Understanding. In Arxiv, 2015.

[31] C. Szegedy, W. Liu, Y.Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich. Going Deeper with Convolutions. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[32] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (ICLR), 2015.