CATS 2: Color And Thermal Stereo Scenes with Semantic Labels

Wayne Treible Philip Saponaro Yi Liu Agnijit Das Gupta
Vinit Veerendraveer Scott Sorensen Chandra Kambhamettu
University of Delaware, Newark, DE
{wtreible, saponaro, yliu, dasgupta, vinitvs, sorensen, chandrak}@udel.edu

Abstract

The CATS dataset introduced a new set of diverse indoor and outdoor scenes with ground truth disparity information for testing stereo matching algorithms in color and thermal imagery. These scenes included nighttime, foggy, low light, and complex lighting in scenes. To extend the usefulness of the CATS dataset we add pixel- and instance-level semantic labels. This includes labels for both color and thermal imagery, and the labels also apply to 3D point clouds as a result of the existing 2D-3D alignment. We compare the new CATS 2.0 dataset against other similar datasets and show it is similar in scope to the KITTI-360 and WildDash datasets, but with the addition of both thermal and 3D information. Additionally, we run a benchmark pedestrian detection algorithm on a set of scenes containing pedestrians.

1. Introduction

The CATS dataset [14] is a stereo matching dataset that includes stereo color, stereo thermal, and ground truth disparity via a LIDAR. CATS includes tabletop, indoor, and outdoor scenes captured in a variety of conditions including low light, twilight, nighttime, light fog, and heavy fog. The addition of the stereo thermal cameras when compared to typical stereo datasets can allow algorithms to work in conditions when a regular camera would fail. Even in normal daytime conditions, some objects can be hidden in shadow which only the thermal camera can see.

Despite having hundreds of objects, the original CATS dataset does not provide object labels. These labels are essential for common computer vision tasks such as recognition, detection, and segmentation. Thus, the CATS dataset is limited in its applicability. In this work, we extend the CATS dataset to include instance- and pixel-level semantic label segmentation in both color and thermal modalities.

Our contributions are as follows:

1. Manually labeled at a pixel and instance level, which includes both color and thermal images.
2. Organized and provided all (image, pixel) locations for each class and each instance. At each pixel location, available information includes color, thermal, depth, and semantic labels.
3. Evaluated a state-of-the-art multispectral pedestrian detector on a subset of CATS 2.0

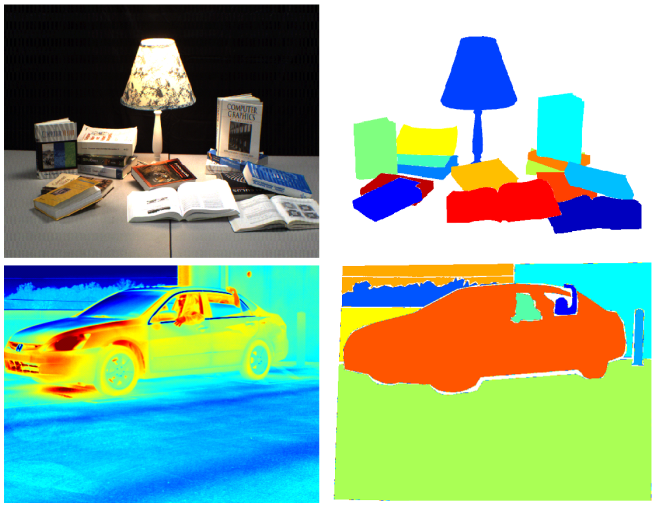

Figure 1. Example indoor and outdoor scenes with pixel- and instance-level semantic labels. Top-left: Color image of a tabletop scene of books. Bottom-left: False color thermal image of an outdoor scene of a car with people inside. Right: Corresponding pixel-level instance labels.

2. The Color and Thermal Stereo Dataset 2.0

In this section we will briefly summarize what was included in the original dataset and then detail the additions that were made for CATS 2.0. For a more detailed view of the original dataset, see [14].

2.1. CATS 1.0 Summary

The CATS imaging setup consists of 4 cameras – two visible-band (color), two long wave infrared (thermal) – and a LiDAR. Specifically, this consists of two Point Grey

	# Imgs	Scn Types	Lbl Types	# Labels	# Classes	Modalities	3D?	Nightime?	Fog?	Year
WildDash [16]	1800	Outdoor	Pxl Inst	***	28	C	No	Yes	Yes	'18
CityScape [3]	3,475**	Outdoor	Pxl Inst	78651	30	C	No	No	No	'16
BDD [15]	100,000	Outdoor	Pxl Inst	1,841,435	10	C	No	Yes	Yes	'18
KAIST [10]	95,000	Outdoor	Box Inst	103,128	3	C + T	No	Yes	No	'15
FLIR ADAS [1]	9,214	Outdoor	Box Inst	37,723	5	C + T †	No	Yes	No	'18
KITTI-360 [2]	400	Outdoor	Pxl Inst	4147	31	C	Yes*	No	No	'18
CATS 2.0 (Ours)	686	In+Outdoor	Pxl Inst	3400	9/96 ‡	C + T	Yes	Yes	Yes	'19

Table 1. Comparison of several object label datasets.
*The main KITTI dataset does provide 3D information via disparity, but KITTI-360 which contains the semantic labels does not directly have this information when downloaded. **The number of images in the "fine annotations" data. ***Unlisted † Unaligned, but synchronized pair. ‡ There are 9 coarse classes and 96 fine classes.

Table 2. Dataset information for labeled images

Attribute	# of Imgs
None	336
Dark	159
Low Light	159
Light Fog	16
Heavy Fog	16

Flea2 cameras capturing at 1280 x 960 resolution, two Xenics Gobi-640-GigEs long wave infrared cameras capturing at 640 x 480 resolution at 50 mK thermal sensitivity, and a Trimble GX Advanced TLS LiDAR.

The dataset is split into two main groups: indoor and outdoor scenes, with the indoor scenery comprised primarily of tabletop and room scenes. There are 10 indoor scenes that rougly correspond to 10 different categories: electronics, plants, books, statues, toys, tools, materials, spooky decorations, miscellaneous objects, and objects in a storage room. Each scene was captured with the following lighting conditions: low light, dark, and normal lighting. Some scenes include fog to simulate fire and suspended particulate conditions with people hidden under collapsing objects.

The outdoor scenery comprises of scenes from the following locations: a backyard, a courtyard, a parking garage, a forest, a garden, a house, a tool shed, and a university campus building. Scenes were captured either during the day, during twilight, or at nighttime. For the very visibly dark scenes, the thermal images – due to the invariance of thermal imagery to visible light differences – are essentially unchanged, while some or most objects become impossible to see in the color imagery.

2.2. CATS 2.0 Additions

CATS 2.0 contains the same imagery and 3D point clouds as CATS 1.0, but with the addition of pixel-level instance and semantic labels, such as those seen in Figure 1. The labels were manually generated by the authors. The following labeling policies were used:

- Foreground objects are to be each labeled separately, with each given a unique identifier. Objects are finely annotated as precisely as possible.
- Background objects are to be labeled separately if possible, such as individual windows, trees, or buildings.
- Uniform background objects such as grass or very complex objects such as foliage were given a single label. For example, individual leaves on a tree and blades of grass are not given separate label identifiers. Additionally, the complex objects are labeled more coarsely.
- When finished, the labeled image is cross-checked between both the color and thermal images to verify objects that are hard to see in one modaility of image are not missed in the other.

There are a total of 686 labeled images. The dataset includes 3400 labels across 9 coarse/96 fine classes, as well as the corresponding 3D points.

2.3. Dataset Comparison

Table 1 gives a summary comparison between recent datasets involving semantic labeling. This is not an exhaustive list, but tries to list the most recent and largest datasets in order to give ours some context.

Due to the rise of self-driving application, more and more cityscape datasets have been released containing labels for mostly cars, pedestrians, and signage. The "Cityscapes Dataset" released in 2016 [3, 4] is a recent addition with high quality, good weather/lighting condition videos and pixel-level instance labels across 50 cities. The Oakland 3D dataset [8] contains labeled point clouds of urban environments. The KITTI-ROAD dataset [5] is subset

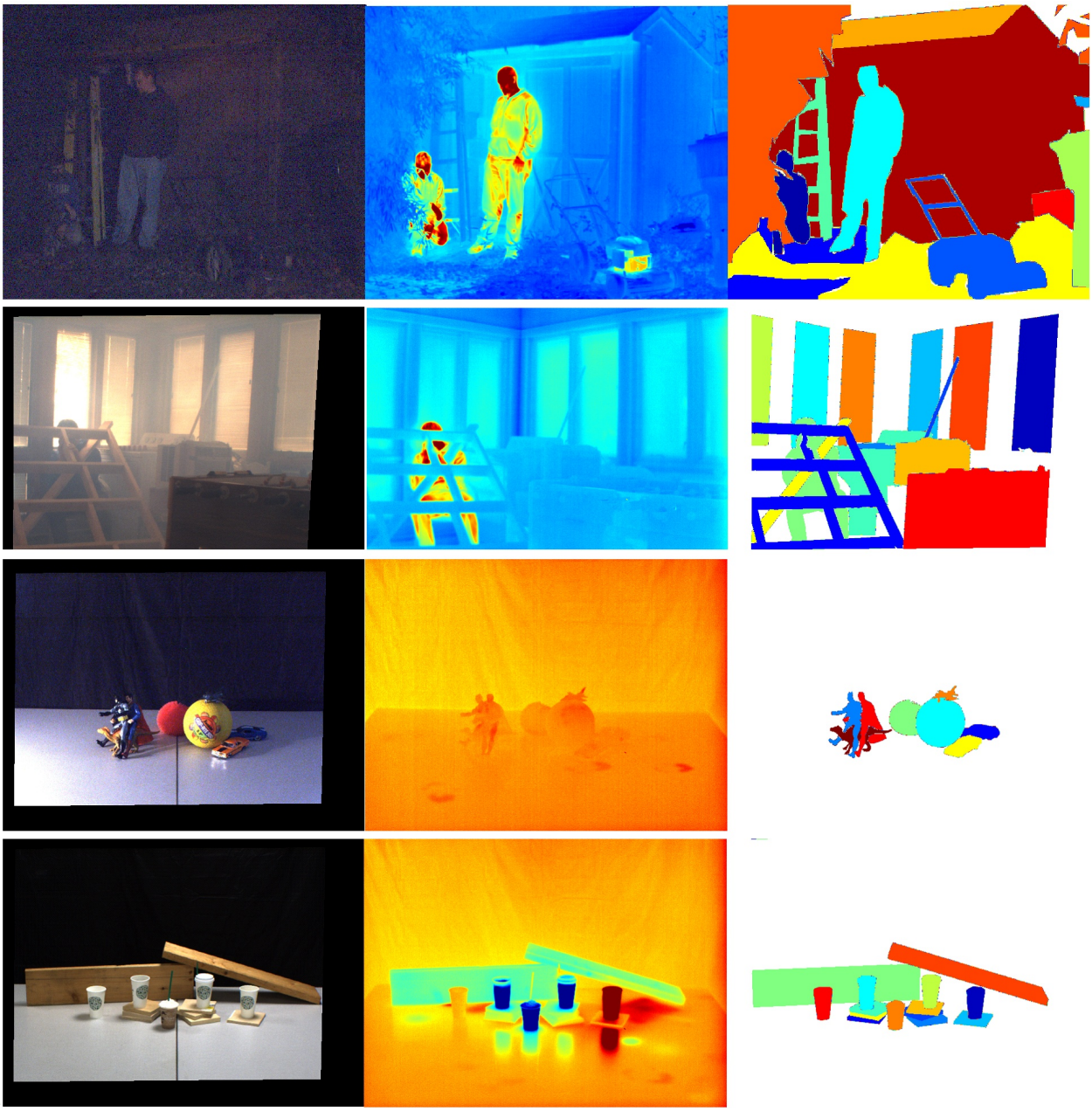

Figure 2. Example environmental conditions in the CATS 2.0 Dataset including outdoor, inside, and tabletop scenes. These include nighttime, low light, complex lighting, and foggy conditions, as well as hidden objects only visible in one modality.

of the KITTI dataset [7, 6] containing 600 frames with the road and the ego-lane annotated. Additionally, the KITTI dataset [2] contains semantic labels for 400 images conforming to the Cityscape dataset format.

Recently there has been a push for varied lightning and environmental conditions in datasets. The WildDash benchmark [16] is a recent dataset containing outdoor road scenes with realistic conditions, including fog and rain. The Berkeley Deep Drive (BDD) dataset [15] contains 100,000 HD videos with both bounding boxes and fully pixel-labeled images in diverse driving conditions. This dataset is the currently largest and most diverse color-camera dataset for self-driving car applications.

Additionally, there are a wide variety of datasets that contain semantic labels in more general scenes. The COCO dataset [12] is one of the largest and well-known pixel-level semantically labeled dataset for general scenes, both indoor and outdoor, with 80 object categories. The Semantic3D dataset [9] which contains 3D labels over a diverse set of outdoor scenes. The ISPRS dataset [13] gives semantic la-

Figure 3. Example detections using the MSDS pedestrian detector. Green is ground truth, red is from the MSDS detector.

bels from satelite views.

Also recently, thermal datasets have become more common. The FLIR ADAS dataset [1] contains 10,000 bounding box labels of pedestrians and cars in thermal cityscape scenes, with unaligned reference RGB imagery additionally given. The KAIST Multispectral pedestrian dataset [10] contains aligned and dually annotated color-thermal pairs of pedestrians with bounding boxes.

However, to the best of our knowledge, there does not exist a dataset that contains semantic labels with thermal imagery, nor one that combines color, thermal, depth, and semantic labels until CATS 2.0. Due to the amount of classes and data, our dataset is useful for testing algorithms but larger datasets such as BDD can be used for training.

3. Baseline Experiment

In this section we provide a baseline experiment that utilizes CATS 2.0 object labels to evaluate the MSDS multispectral pedestrian network outlined in [11]. The MSDS architecture is composed of a fusion of a multispectral proposal network to generate pedestrian proposals, and a multispectral classification network to help distinguish those proposals. The network is trained by optimizing both pedestrian detection and semantic segmentation jointly and, at the time of publication, was the top-performing method on the KAIST multispectral pedestrian dataset [10]. This work also outlined some inaccuracies in the annotations in the KAIST dataset, while providing a sanitized version.

We used pre-trained weights from the MSDS project page [1] which were learned by training the network on the sanitized KAIST dataset and ran the network on a set of pixel-aligned pairs of color and thermal images from CATS.

Due to CATS images having a higher resolution, images were resized to (640, 512) to match the resolution of KAIST thermal imagery. Because pedestrians generally appear larger in CATS images, we also used three additional scales to simulate different pedestrian sizes with images resized to $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4}$, or image sizes of (320, 256), (213, 170), and (160, 128), respectively.

Each color and thermal image pair was then run through the network to detect pedestrians and generate bounding boxes with a confidence threshold set to 0.35 and a non-maximum suppression threshold set to 0.3. The generated pedestrian bounding boxes from MSDS were then compared to the manually annotated CATS 2.0 bounding boxes using intersection over union (IoU). Out of 100 images containing pedestrians (25 images at 4 scales), the MSDS network was able to detect pedestrians in 29 with the average IoU score for the detections being 0.428. We present some of the better detection results in Fig 3, where the green boxes indicate CATS 2.0 ground truth labels and red boxes are the MSDS pedestrian detections. Qualitative analysis of the results indicates that pedestrians closer to the camera and pedestrians in very dark scenes were not detected by the network.

4. Conclusion

In this work we have introduced pixel-level instance-level semantic labels to the CATS dataset. We labeled 686 images including both color and thermal images across a variety of indoor and outdoor scenes containing varied lighting and environmental conditions. We compared the new CATS 2.0 semantically labeled dataset against other similar datasets and found there has been a push for large-scale, diverse lighting, and environmental conditions, but the scope of current multispectral datasets is limited. We benchmarked a pedestrian detector and found that there is still room for improvement for multispectral detection. In the future, we will build on this dataset by applying the same methodology on a larger scale across seasons.

[1] https://github.com/Li-Chengyang/MSDS-RCNN

References

[1] Flir thermal dataset for algorithm training. https://www.flir.com/oem/adas/adas-dataset-form/. Accessed: 2019-04-12. 2, 4

[2] H. Alhaija, S. Mustikovela, L. Mescheder, A. Geiger, and C. Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision (IJCV)*, 2018. 2, 3

[3] 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. 2

[4] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015. 2

[5] J. Fritsch, T. Kuehnl, and A. Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013. 2

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 3

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361, June 2012. 3

[8] Y. Guo, J. Zhang, M. Lu, J. Wan, and Y. Ma. Benchmark datasets for 3d computer vision. 06 2014. 2

[9] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys. Semantic3d.net: A new large-scale point cloud classification benchmark. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-1-W1, pages 91–98, 2017. 3

[10] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, June 2015. 2, 4

[11] C. Li, D. Song, R. Tong, and M. Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference (BMVC)*, September. 4

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 3

[13] F. Rottensteiner. Isprs test project on urban classification and 3d building reconstruction: Evaluation of building reconstruction results. 2009. 3

[14] S. S. A. K. M. O. B. P. K. S. C. K. Wayne Treible, Philip Saponaro. Cats: A color and thermal stereo benchmark. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[15] 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*, abs/1805.04687, 2018. 2, 3

[16] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, and G. Fernandez Dominguez. Wilddash - creating hazard-aware benchmarks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 3