The Cityscapes Dataset

Marius Cordts[1,2] Mohamed Omran[3] Sebastian Ramos[1,4] Timo Scharwächter[1,2] Markus Enzweiler[1]
Rodrigo Benenson[3] Uwe Franke[1] Stefan Roth[2] Bernt Schiele[3]

[1]Daimler AG R&D, [2]TU Darmstadt, [3]MPI Informatics, [4]TU Dresden
mail@cityscapes-dataset.net www.cityscapes-dataset.net

Abstract

Semantic understanding of urban street scenes through visual perception has been widely studied due to many possible practical applications. Key challenges arise from the high visual complexity of such scenes. In this paper, we present ongoing work on a new large-scale dataset for (1) assessing the performance of vision algorithms for different tasks of semantic urban scene understanding, including scene labeling, instance-level scene labeling, and object detection; (2) supporting research that aims to exploit large volumes of (weakly) annotated data, e.g. *for training deep neural networks. We aim to provide a large and diverse set of stereo video sequences recorded in street scenes from 50 different cities, with high quality pixel-level annotations of 5000 frames in addition to a larger set of weakly annotated frames. The dataset is thus an order of magnitude larger than similar previous attempts.*

Several aspects are still up for discussion, and timely feedback from the community would be greatly appreciated. Details on annotated classes and examples will be available at www.cityscapes-dataset.net*. Moreover, we will use this website to collect remarks and suggestions.*

1. Introduction

Over the last decade, the problem of scene understanding has increasingly gained attention in the computer vision community [11]. Research on this topic has been done from different points of view (*e.g.*, semantic [25] or holistic [7]) and in different application scenarios (*e.g.*, outdoors [26] and indoors [9]). In particular, understanding outdoor scenarios through visual data has been widely researched due to its high visual complexity and many possible practical applications. Challenges arise from a large number of different objects in the scene, both static and dynamic, with a large variation in scale. Despite this complexity, the ability to correctly analyze the environment is critical for the development of autonomous systems such as self-driving cars [5, 6].

In the last few years, we have experienced great advances toward visually understanding outdoor scenarios. Increasingly sophisticated approaches have been developed (*e.g.*, [7, 18, 22]) and more challenging datasets have been created to push forward the development of these approaches. In particular, the *KITTI Vision Benchmark Suite* [8] supports and assesses vision algorithms in the context of autonomous driving. This benchmark includes datasets for stereo vision, optical flow, visual odometry, 3D object recognition, tracking, and road estimation. Furthermore, the task of a semantic understanding of urban scenarios is addressed by several datasets, such as *CamVid* [1], *Leuven* [13], and *Daimler Urban Segmentation* [21].

Figure 1. Example images (2 MP, color, HDR) with annotations from our dataset. The labels are encoded in different colors. Note that instances of traffic participants are annotated individually.

Progress in computer vision in recent years has also been driven by datasets that pushed the boundaries in terms of scale. Examples include *ImageNet* [2, 19], *PASCAL VOC* [4] and its extensions (*e.g.*, *PASCAL-Context* [16]), *Caltech Pedestrians* [3], *LabelMe* [20], *Microsoft COCO* [14, 15], *Yahoo Flickr CC100m* [24],

and *Places* [27]. Such large-scale datasets have paved the way for the success of vision algorithms that can leverage large amounts of data, *e.g*. deep learning methods [12,23]. To the best of our knowledge, there is no large-scale dataset available for semantic urban scene understanding, yet would enable further significant progress in this research area.

To fill this gap, we present ongoing work on creating a novel dataset for semantic urban scene understanding, along with a benchmark of different challenges. Annotations of a large set of classes and object instances, high variability of the urban scenes, a large number of annotated images, and various metadata are some of the highlights of the presented dataset. Examples of our annotations can be seen in Fig. 1.

2. Dataset Specifications

Designing a new large-scale dataset involves a multitude of decisions, including the type of annotations, images, the metadata, and an estimation of a reasonable volume of annotated data to achieve. In the following, we describe our design choices in order to create a dataset focusing on semantic understanding of urban environments.

Annotation type. We chose to focus on semantic, instance-wise, dense pixel annotations. Datasets with this type of annotations are the scarcest among existing ones, mainly because annotating images in this way is an expensive and tedious process. Nevertheless, we believe these annotations are the most informative ones for training scene understanding algorithms, and the ones that enable the richest set of evaluations. We also think that these annotations easily allow for future extensions of the dataset; *e.g*., adding finer-grained categories to already segmented instance regions is easier than the initial region boundary annotation.

Our annotations enable out-of-the-box evaluations for scene labeling at the image and instance level, as well as object detection. We discuss the benchmark suites and evaluation metrics for these tasks in detail in Sec. 3.

Labeled classes. We annotate 25 different labels, selected by striking a balance between common classes, different applications, and covering a large part of the image with "non-void" classes, while trying to limit the annotation effort. Table 1 lists the specific labels we consider.

Image diversity. The high variability of outdoor urban streets makes accurate scene understanding very challenging. Unfortunately, most available datasets fall behind in capturing this high variance. For instance, the KITTI dataset contains 6 h of video material, all recorded in the Karlsruhe, Germany, metropolitan area. Out of these recordings, only 25 % are publicly available and roughly 430 images have pixel-level semantic annotations, provided by different independent research groups. With our novel dataset, we tackle this issue by focusing on a high diversity between annotated images in order to capture a wider range of street scenes than any previous dataset. To this end, we record in approximately 50 different cities to reduce city-specific overfitting. We further acquire images during the span of several months, covering spring, summer, and autumn. Recordings are restricted to good weather conditions, which already pose a significant challenge for computer vision. We believe that more challenging and diverse weather conditions should be addressed in specialized datasets such as used in [17]. To achieve a high diversity, each annotated frame is manually selected, in order to focus on "non-empty" scenes, *i.e*. with a large number of dynamic objects on the street, and simultaneously achieving a high variety in scene layout and background.

Table 1. Overview of annotated classes

Group	Class	Description
ground (2)	road	where cars drive or stand
	sidewalk	where people go or rest
human (2)	person[1]	walking or sitting
	rider[1]	people on means of transport
vehicle (7)	car[1]	
	truck[1]	
	bus[1]	
	on rails[1]	trams, trains
	motorcycle[1]	also scooters *etc*.
	bicycle[1]	
	license plate[2]	on vehicles
infra-structure (8)	building	houses, skyscrapers
	wall	not part of a building
	fence	
	traffic sign	
	traffic light	
	pole	
	bridge[2]	
	tunnel[2]	
nature (2)	tree	any vertical vegetation
	terrain	grass, soil, sand
sky (1)	sky	
void (3)	ground	any other horizontal surface
	dynamic	*e.g*. baby strollers, animals
	static	other/unrecognizable objects

[1] Single instance annotation available.
[2] Not included in fine label set challenges, see Sec. 3.

Metadata. In addition to the annotated frame, our dataset will contain preceding and trailing video frames. Approximately half of the annotated images are extracted from long video sequences, while the remaining are the 20$^{\text{th}}$ images from 30 frame video snippets (1.8 s). The surrounding frames are provided as context for methods exploiting optical flow, tracking, or structure-from-motion. In addition to video, we provide corresponding right stereo views, precomputed depth maps, GPS coordinates, and ego-motion data from the vehicle odometry.

Volume. Deciding upon a specific volume is a difficult task. Our general mindset is to aim for a tenfold increase over existing comparable datasets. We currently finished roughly 800 images with high quality annotations and are aiming to reach about 5 000 for the first release of the dataset. These images will be divided up for training, validation, and testing. Further, we plan to acquire up to 20 000 additional images with coarse annotations. We believe this volume would serve research on exploiting large amounts of (weakly annotated) data.

3. Benchmark Suite

We will setup a benchmark suite together with an evaluation server, such that authors can upload their results and get a ranking

regarding the different tasks. Our evaluation concept is designed such that a single algorithm can contribute to multiple challenges.

For each challenge, we aim for scalar scores to allow for an intuitive ranking of the approaches compared. We use different metrics to assess various aspects of these challenges, but for each such aspect, we restrict ourselves to a single score to avoid redundancy in the evaluation. Further, we evaluate all measures on a coarse label set, *i.e.* label groups such as vehicle or infrastructure, and on a fine label set, *i.e.* classes such as car, truck, bus, building, pole, or fence. For details on the label groups and classes we defined see Table 1.

In the following, we discuss some challenges we plan to set up and the corresponding evaluation metrics we are planning to use. However, the exact definition of challenges and metrics is still open for discussion.

Scene labeling task. The task is to assign a single label to each pixel in the image, *i.e.* the required output of an algorithm are images where a pixel value represents its class. To analyze the performance on the *pixel level* we utilize the PASCAL VOC intersection-over-union metric (IU) [4], which is also known as Jaccard Index (JI). This score is one of the standard measures for scene labeling. For each label, the number of true positive (TP), false positive (FP), and false negative (FN) pixels are determined over the whole test set. Then, the IU metric is defined as

$$\mathrm{IU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \,. \tag{1}$$

The final score on the coarse label set is denoted by $\mathrm{mIU}_{\mathrm{coarse}}$ and computed as the mean of the individual group scores. Analogously, we compute the score on the fine label set $\mathrm{mIU}_{\mathrm{fine}}$. In both cases, pixels with void label do not contribute to the score.

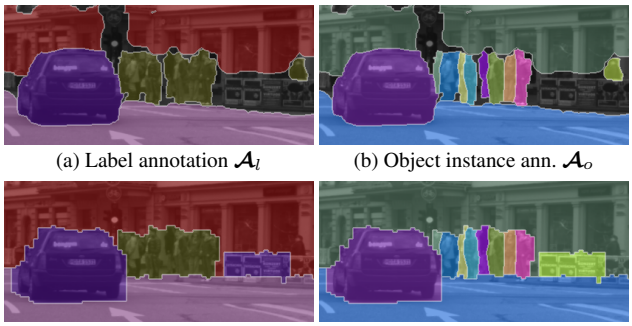
One shortcoming of the global IU measure is the strong focus on object instances that cover many pixels in the image. In the context of street scenes with their large variation in object scale, this problem is particularly pronounced. Thus, a normalization regarding this issue in terms of a metric that assesses the performance of a method also on an *instance level* is needed. To account for this fact, we propose an additional score for the scene labeling task that aims to answer the question of how well a certain approach represents the individual traffic participants in the scene.

The proposed measure aims to assess both detection and segmentation capabilities of a method at the same time. To this end, we compute a curve where the x-axis is the minimum pixel-level overlap between a predicted and a ground truth region for a true positive and the y-axis is the resulting F-score

$$F = 2\,\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \,. \tag{2}$$

While a higher F-score stands for a better detection performance, reaching a certain F-score with a larger pixel overlap means better segmentation accuracy.

We define the area under the curve as the average F-score (AF) per label. The mean over all classes is then used for the final scores, $\mathrm{mAF}_{\mathrm{coarse}}$ and $\mathrm{mAF}_{\mathrm{fine}}$, respectively. Note that the average could also be computed in a smaller range of the overlap, *e.g.* 25 % to 50 % for a focus on the detection part, or 50 % to 100 % to focus on the segmentation aspect.



(a) Label annotation $\boldsymbol{\mathcal{A}}_l$ (b) Object instance ann. $\boldsymbol{\mathcal{A}}_o$

(c) Label prediction $\boldsymbol{\mathcal{P}}_l$ (d) Object instance pred. $\boldsymbol{\mathcal{P}}_o$

Figure 2. Overview of different annotation and prediction formats. The object instance prediction (d) is generated using Alg. 1. Note the false positive *vehicle* prediction to the right of the pedestrians.

Algorithm 1 Generate object instance from label prediction

Input: Label prediction $\boldsymbol{\mathcal{P}}_l$, Object instance annotation $\boldsymbol{\mathcal{A}}_o$
  Allocate searched object instance prediction $\boldsymbol{\mathcal{P}}_o$
  $\mathcal{R}_p \leftarrow \mathrm{connectedRegions}(\boldsymbol{\mathcal{P}}_l)$
  for all regions $R \in \mathcal{R}_p$ **do**
    $l \leftarrow \mathrm{label}(R)$
    // Find all candidate object instances for the region:
    $\mathcal{C}_o \leftarrow \{o \,|\, o \in R(\boldsymbol{\mathcal{A}}_o) \wedge \mathrm{label}(o) = l\}$
    if $\mathcal{C}_o = \emptyset$ **then**
      $\boldsymbol{\mathcal{P}}_o(R) \leftarrow$ new object instance (false positive)
    else
      for all pixels $p \in R$ **do**
        // Find closest candidate object instance for p:
        $\boldsymbol{\mathcal{P}}_o(p) \leftarrow \underset{o \in \mathcal{C}_o}{\mathrm{argmin}} \underset{\substack{\text{pixel } q \in \boldsymbol{\mathcal{A}}_o, \text{ with} \\ \mathrm{label}(q) = \mathrm{label}(o)}}{\min} \mathrm{dist}(p, q)$
      end for
    end if
  end for
Output: $\boldsymbol{\mathcal{P}}_o$

Since we want to evaluate standard scene labeling methods within this challenge, we do not require the method to provide instance-level predictions. However, to still allow for this type of evaluation, we artificially generate instance predictions from the regular label predictions using ground truth annotations. See Fig. 2 for an example and Algorithm 1 for details on the generation of these predictions.

Instance-level scene labeling task. This challenge assesses algorithms that aim at predicting the individual instances in the scene. Such an algorithm is expected to provide instance predictions associated with a class, a segmentation, and a confidence score. As evaluation metric, we compute the average precision on the region level (APR) for each label, which corresponds to the AP^r_{vol} value from [10]. This metric is the volume under a graph with the three axes recall, precision, and overlap. The latter is computed on the region level and equals the IU of a single instance. For each such overlap value, precision-recall curves are computed in a standard fashion, where multiple predictions of the same ground truth instance count as false positives. To obtain a single scalar score, we average all APR volumes over the coarse label set, yielding $\mathrm{mAPR}_{\mathrm{coarse}}$, and over the fine label set, giving $\mathrm{mAPR}_{\mathrm{fine}}$.

Object detection task. Algorithms are expected to deliver bounding box predictions of the object instances in the scene, associated with a confidence score. This task is evaluated as in the PASCAL VOC challenge [4] via the average precision score APB. This metric equals the APR metric, except for an overlap computed on the box level and a fixed threshold of 50 %. We denote the average result over coarse labels as $\text{mAPB}_{\text{coarse}}$ and over fine labels as $\text{mAPB}_{\text{fine}}$.

Meta information. In addition to the previously introduced measures, we plan to report timings for each method and a video of the inferred labeling result on a designated sequence of the dataset to allow for a qualitative assessment of the approach. Further, we list the kind of information each algorithm is using, *e.g.* stereo or video.

4. Conclusion and Outlook

In this work, we presented our current plans for a novel, densely annotated large-scale dataset for semantic understanding of urban street scenes. We discussed the scale of the dataset, the kind of data recorded, the annotations we plan to provide, as well as a preliminary list of evaluation metrics we are considering for the tasks of image- and instance-level semantic labeling, as well as object detection.

Several aspects are still up for discussion, and timely feedback from the community would be greatly appreciated. Some currently open questions are: *(1)* Are there other sensible evaluation metrics for the tasks of scene labeling and object detection that should be considered? *(2)* Which other problems related to street scene understanding would benefit from such a dataset, and what extensions to the dataset, *e.g.* in terms of annotations, would be necessary to support these additional challenges? *(3)* Are changes to the labeling protocol required?

Concurrently with this workshop, we have published a website (`www.cityscapes-dataset.net`) that provides a detailed list of the annotated classes, describes the data formats, and presents annotation examples. Further, we aim to use the website to collect remarks and suggestions.

Due to the rapid progress in our field, we intend for our efforts not to merely culminate in a static one-off release. Instead, we view this dataset as a dynamic entity, which we plan to be growing over time. Thus, subsequent releases will be such that they expand the initial version and not render it obsolete. We plan to release an initial version of the dataset and the benchmark in autumn 2015, and to organize a challenge based on the benchmark at CVPR 2016.

References

[1] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2), 2009. 1

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *Trans. PAMI*, 34(4), 2012. 1

[4] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes challenge: A retrospective. *IJCV*, 111(1), 2014. 1, 3, 4

[5] U. Franke, D. Pfeiffer, C. Rabe, C. Knöppel, M. Enzweiler, F. Stein, and R. G. Herrtwich. Making Bertha see. In *ICCV Workshops*, 2013. 1

[6] P. Furgale, U. Schwesinger, M. Rufli, W. Derendarz, H. Grimmett, P. Mühlfellner, S. Wonneberger, B. Li, B. Schmidt, T. N. Nguyen, E. Cardarelli, S. Cattani, S. Brüning, S. Horstmann, M. Stellmacher, S. Rottmann, H. Mielenz, K. Köser, J. Timpner, M. Beermann, C. Häne, L. Heng, G. H. Lee, F. Fraundorfer, R. Iser, R. Triebel, I. Posner, P. Newman, L. Wolf, M. Pollefeys, S. Brosig, J. Effertz, C. Pradalier, and R. Siegwart. Toward automated driving in cities using close-to-market sensors: An overview of the V-Charge project. In *IV Symposium*, 2013. 1

[7] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3D traffic scene understanding from movable platforms. *Trans. PAMI*, 36(5), 2014. 1

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 32(11), 2013. 1

[9] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation. *IJCV*, 112(2), 2015. 1

[10] B. Hariharan, P. Arbeláez, R. B. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 3

[11] D. Hoiem, J. Hays, J. Xiao, and A. Khosla. Guest editorial: Scene understanding. *IJCV*, 112(2), 2015. 1

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[13] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3D scene analysis from a moving vehicle. In *CVPR*, 2007. 1

[14] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common objects in context. *arXiv:1405.0312 [cs.CV]*, 2014. 1

[15] 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 *ECCV*, 2014. 1

[16] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 1

[17] D. Pfeiffer, S. K. Gehrig, and N. Schneider. Exploiting the power of stereo confidences. In *CVPR*, 2013. 2

[18] G. Ros, S. Ramos, M. Granados, D. Vazquez, and A. M. Lopez. Vision-based offline-online perception paradigm for autonomous driving. In *WACV*, 2015. 1

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *arXiv:1409.0575 [cs.CV]*, 2014. 1

[20] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: A database and web-based tool for image annotation. *IJCV*, 77(1-3), 2008. 1

[21] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth. Efficient multi-cue scene segmentation. In *GCPR*, 2013. 1

[22] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth. Stixmantics: A medium-level model for real-time semantic scene understanding. In *ECCV*, 2014. 1

[23] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014. 2

[24] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv:1503.01817 [cs.CV]*, 2015. 1

[25] J. Tighe, M. Niethammer, and S. Lazebnik. Scene parsing with object instance inference using regions and per-exemplar detectors. *IJCV*, 112(2), 2015. 1

[26] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele. Monocular visual scene understanding: Understanding multi-object traffic scenes. *Trans. PAMI*, 35(4), 2013. 1

[27] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 2