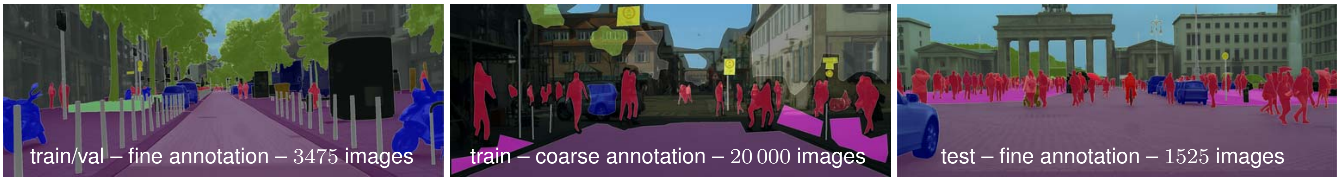
2016 IEEE Conference on Computer Vision and Pattern Recognition

The Cityscapes Dataset for Semantic Urban Scene Understanding

Marius Cordts[1,2] Mohamed Omran[3] Sebastian Ramos[1,4] Timo Rehfeld[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
www.cityscapes-dataset.net


train/val – fine annotation – 3475 images
train – coarse annotation – 20 000 images
test – fine annotation – 1525 images

Abstract

Visual understanding of complex urban street scenes is an enabling factor for a wide range of applications. Object detection has benefited enormously from large-scale datasets, especially in the context of deep learning. For semantic urban scene understanding, however, no current dataset adequately captures the complexity of real-world urban scenes. To address this, we introduce Cityscapes*, a benchmark suite and large-scale dataset to train and test approaches for pixel-level and instance-level semantic labeling. Cityscapes is comprised of a large, diverse set of stereo video sequences recorded in streets from* 50 *different cities.* 5000 *of these images have high quality pixel-level annotations;* 20 000 *additional images have coarse annotations to enable methods that leverage large volumes of weakly-labeled data. Crucially, our effort exceeds previous attempts in terms of dataset size, annotation richness, scene variability, and complexity. Our accompanying empirical study provides an in-depth analysis of the dataset characteristics, as well as a performance evaluation of several state-of-the-art approaches based on our benchmark.*

1. Introduction

Visual scene understanding has moved from an elusive goal to a focus of much recent research in computer vision [27]. Semantic reasoning about the contents of a scene is thereby done on several levels of abstraction. Scene recognition aims to determine the overall scene category by putting emphasis on understanding its global properties, *e.g.* [46, 82]. Scene labeling methods, on the other hand, seek to identify the individual constituent parts of a whole scene as well as their interrelations on a more local pixel- and instance-level, *e.g.* [41, 71]. Specialized object-centric methods fall somewhere in between by focusing on detecting a certain subset of (mostly dynamic) scene constituents, *e.g.* [6,12,13,15]. Despite significant advances, visual scene understanding remains challenging, particularly when taking human performance as a reference.

The resurrection of deep learning [34] has had a major impact on the current state-of-the-art in machine learning and computer vision. Many top-performing methods in a variety of applications are nowadays built around deep neural networks [30, 41, 66]. A major contributing factor to their success is the availability of large-scale, publicly available datasets such as *ImageNet* [59], *PASCAL VOC* [14], *PASCAL-Context* [45], and *Microsoft COCO* [38] that allow deep neural networks to develop their full potential.

Despite the existing gap to human performance, scene understanding approaches have started to become essential components of advanced real-world systems. A particularly popular and challenging application involves self-driving cars, which make extreme demands on system performance and reliability. Consequently, significant research efforts have gone into new vision technologies for understanding complex traffic scenes and driving scenarios [4, 16–18, 58, 62]. Also in this area, research progress can be heavily linked to the existence of datasets such as the *KITTI Vision Benchmark Suite* [19], *CamVid* [7], *Leuven* [35], and *Daimler Urban Segmentation* [61] datasets. These urban scene datasets are often much smaller than datasets addressing more general settings. Moreover, we argue that they do not fully capture the variability and complexity of real-world inner-city traffic scenes. Both shortcomings currently inhibit further progress in visual understanding of street scenes. To this end, we propose the *Cityscapes* benchmark suite and a corresponding dataset, specifically

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.350

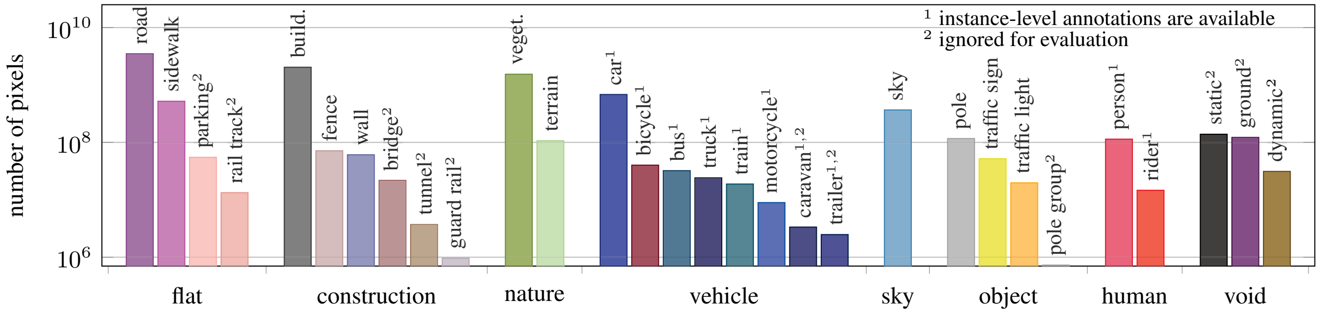


Figure 1. Number of finely annotated pixels (y-axis) per class and their associated categories (x-axis).

tailored for autonomous driving in an urban environment and involving a much wider range of highly complex inner-city street scenes that were recorded in 50 different cities. Cityscapes significantly exceeds previous efforts in terms of size, annotation richness, and, more importantly, regarding scene complexity and variability. We go beyond pixel-level semantic labeling by also considering instance-level semantic labeling in both our annotations and evaluation metrics. To facilitate research on 3D scene understanding, we also provide depth information through stereo vision.

Very recently, [75] announced a new semantic scene labeling dataset for suburban traffic scenes. It provides temporally consistent 3D semantic instance annotations with 2D annotations obtained through back-projection. We consider our efforts to be complementary given the differences in the way that semantic annotations are obtained, and in the type of scenes considered, *i.e.* suburban *vs.* inner-city traffic. To maximize synergies between both datasets, a common label definition that allows for cross-dataset evaluation has been mutually agreed upon and implemented.

2. Dataset

Designing a large-scale dataset requires a multitude of decisions, *e.g.* on the modalities of data recording, data preparation, and the annotation protocol. Our choices were guided by the ultimate goal of enabling significant progress in the field of semantic urban scene understanding.

2.1. Data specifications

Our data recording and annotation methodology was carefully designed to capture the high variability of outdoor street scenes. Several hundreds of thousands of frames were acquired from a moving vehicle during the span of several months, covering spring, summer, and fall in 50 cities, primarily in Germany but also in neighboring countries. We deliberately did not record in adverse weather conditions, such as heavy rain or snow, as we believe such conditions to require specialized techniques and datasets [51].

Our camera system and post-processing reflect the current state-of-the-art in the automotive domain. Images were recorded with an automotive-grade 22 cm baseline stereo camera using 1/3 in CMOS 2 MP sensors (OnSemi AR0331) with rolling shutters at a frame-rate of 17 Hz. The sensors were mounted behind the windshield and yield high dynamic-range (HDR) images with 16 bits linear color depth. Each 16 bit stereo image pair was subsequently debayered and rectified. We relied on [31] for extrinsic and intrinsic calibration. To ensure calibration accuracy we recalibrated on-site before each recording session.

For comparability and compatibility with existing datasets we also provide low dynamic-range (LDR) 8 bit RGB images that are obtained by applying a logarithmic compression curve. Such tone mappings are common in automotive vision, since they can be computed efficiently and independently for each pixel. To facilitate highest annotation quality, we applied a separate tone mapping to each image. The resulting images are less realistic, but visually more pleasing and proved easier to annotate. 5000 images were manually selected from 27 cities for dense pixel-level annotation, aiming for high diversity of foreground objects, background, and overall scene layout. The annotations (see Sec. 2.2) were done on the 20th frame of a 30-frame video snippet, which we provide in full to supply context information. For the remaining 23 cities, a single image every 20 s or 20 m driving distance (whatever comes first) was selected for coarse annotation, yielding 20 000 images in total.

In addition to the rectified 16 bit HDR and 8 bit LDR stereo image pairs and corresponding annotations, our dataset includes vehicle odometry obtained from in-vehicle sensors, outside temperature, and GPS tracks.

2.2. Classes and annotations

We provide coarse and fine annotations at pixel level including instance-level labels for humans and vehicles.

Our 5000 fine pixel-level annotations consist of layered polygons (à la LabelMe [60]) and were realized in-house to guarantee highest quality levels. Annotation and quality control required more than 1.5 h on average for a single image. Annotators were asked to label the image from back to front such that no object boundary was marked more than once. Each annotation thus implicitly provides a depth ordering of the objects in the scene. Given our label scheme,

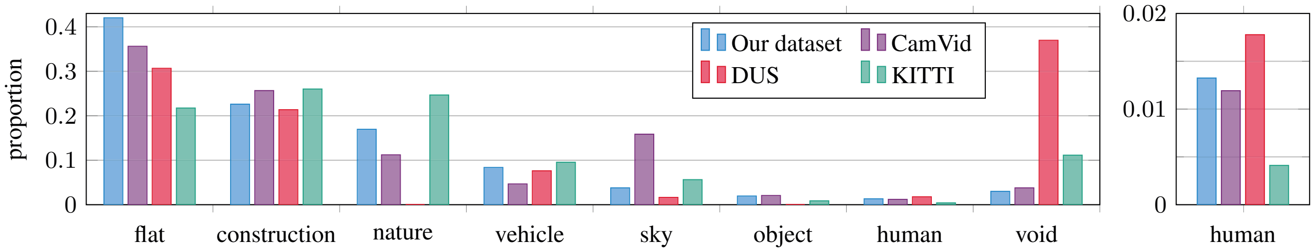


Figure 2. Proportion of annotated pixels (y-axis) per category (x-axis) for Cityscapes, CamVid [7], DUS [61], and KITTI [19].

annotations can be easily extended to cover additional or more fine-grained classes.

For our 20 000 coarse pixel-level annotations, accuracy on object boundaries was traded off for annotation speed. We aimed to correctly annotate as many pixels as possible within a given span of less than 7 min of annotation time per image. This was achieved by labeling coarse polygons under the sole constraint that each polygon must only include pixels belonging to a single object class.

In two experiments we assessed the quality of our labeling. First, 30 images were finely annotated twice by different annotators and passed the same quality control. It turned out that 96 % of all pixels were assigned to the same label. Since our annotators were instructed to choose a *void* label if unclear (such that the region is ignored in training and evaluation), we exclude pixels having at least one *void* label and recount, yielding 98 % agreement. Second, all our fine annotations were additionally coarsely annotated such that we can enable research on densifying coarse labels. We found that 97 % of all labeled pixels in the coarse annotations were assigned the same class as in the fine annotations.

We defined 30 visual classes for annotation, which are grouped into eight categories: flat, construction, nature, vehicle, sky, object, human, and void. Classes were selected based on their frequency, relevance from an application standpoint, practical considerations regarding the annotation effort, as well as to facilitate compatibility with existing datasets, *e.g.* [7, 19, 75]. Classes that are too rare are excluded from our benchmark, leaving 19 classes for evaluation, see Fig. 1 for details. We plan to release our annotation tool upon publication of the dataset.

2.3. Dataset splits

We split our densely annotated images into separate training, validation, and test sets. The coarsely annotated images serve as additional training data only. We chose not to split the data randomly, but rather in a way that ensures each split to be representative of the variability of different street scene scenarios. The underlying split criteria involve a balanced distribution of geographic location and population size of the individual cities, as well as regarding the time of year when recordings took place. Specifically, each of the three split sets is comprised of data recorded with the following properties in equal shares: (i) in large, medium, and small cities; (ii) in the geographic west, center, and east; (iii) in the geographic north, center, and south; (iv) at the beginning, middle, and end of the year. Note that the data is split at the city level, *i.e.* a city is completely within a single split. Following this scheme, we arrive at a unique split consisting of 2975 training and 500 validation images with publicly available annotations, as well as 1525 test images with annotations withheld for benchmarking purposes.

	#pixels $[10^9]$	annot. density [%]
Ours (fine)	9.43	**97.1**
Ours (coarse)	**26.0**	67.5
CamVid	0.62	96.2
DUS	0.14	63.0
KITTI	0.23	88.9

Table 1. Absolute number and density of annotated pixels for Cityscapes, DUS, KITTI, and CamVid (upscaled to 1280×720 pixels to maintain the original aspect ratio).

In order to assess how uniform (representative) the splits are regarding the four split characteristics, we trained a fully convolutional network [41] on the 500 images in our validation set. This model was then evaluated on the whole test set, as well as eight subsets thereof that reflect the extreme values of the four characteristics. With the exception of the time of year, the performance is very homogeneous, varying less than 1.5 % points (often much less). Interestingly, the performance on the *end of the year* subset is 3.8 % points better than on the whole test set. We hypothesize that this is due to softer lighting conditions in the frequently cloudy fall. To verify this hypothesis, we additionally tested on images taken in low- or high-temperature conditions, finding a 4.5 % point increase in low temperatures (cloudy) and a 0.9 % point decrease in warm (sunny) weather. Moreover, specifically training for either condition leads to an improvement on the respective test set, but not on the balanced set. These findings support our hypothesis and underline the importance of a dataset covering a wide range of conditions encountered in the real world in a balanced way.

2.4. Statistical analysis

We compare Cityscapes to other datasets in terms of (i) annotation volume and density, (ii) the distribution of visual

	#humans $[10^3]$	#vehicles $[10^3]$	#h/image	#v/image
Ours (fine)	24.4	**41.0**	**7.0**	**11.8**
KITTI	6.1	30.3	0.8	4.1
Caltech	**192**[1]	-	1.5	-

Table 2. Absolute and average number of instances for Cityscapes, KITTI, and Caltech ([1] via interpolation) on the respective training and validation datasets.

classes, and (iii) scene complexity. Regarding the first two aspects, we compare Cityscapes to other datasets with semantic pixel-wise annotations, *i.e.* CamVid [7], DUS [62], and KITTI [19]. Note that there are many other datasets with dense semantic annotations, *e.g.* [2, 56, 65, 69, 70]. However, we restrict this part of the analysis to those with a focus on autonomous driving.

CamVid consists of ten minutes of video footage with pixel-wise annotations for over 700 frames. DUS consists of a video sequence of 5000 images from which 500 have been annotated. KITTI addresses several different tasks including semantic labeling and object detection. As no official pixel-wise annotations exist for KITTI, several independent groups have annotated approximately 700 frames [22, 29, 32, 33, 58, 64, 77, 80]. We map those labels to our high-level categories and analyze this consolidated set. In comparison, Cityscapes provides significantly more annotated images, *i.e.* 5000 fine and 20 000 coarse annotations. Moreover, the annotation quality and richness is notably better. As Cityscapes provides recordings from 50 different cities, it also covers a significantly larger area than previous datasets that contain images from a single city only, *e.g.* Cambridge (CamVid), Heidelberg (DUS), and Karlsruhe (KITTI). In terms of absolute and relative numbers of semantically annotated pixels (training, validation, and test data), Cityscapes compares favorably to CamVid, DUS, and KITTI with up to two orders of magnitude more annotated pixels, *c.f.* Tab. 1. The majority of all annotated pixels in Cityscapes belong to the coarse annotations, providing many individual (but correlated) training samples, but missing information close to object boundaries.

Figures 1 and 2 compare the distribution of annotations across individual classes and their associated higher-level categories. Notable differences stem from the inherently different configurations of the datasets. Cityscapes involves dense inner-city traffic with wide roads and large intersections, whereas KITTI is composed of less busy suburban traffic scenes. As a result, KITTI exhibits significantly fewer *flat* ground structures, fewer *humans*, and more *nature*. In terms of overall composition, DUS and CamVid seem more aligned with Cityscapes. Exceptions are an abundance of *sky* pixels in CamVid due to cameras with a comparably large vertical field-of-view and the absence of certain categories in DUS, *i.e.* *nature* and *object*.

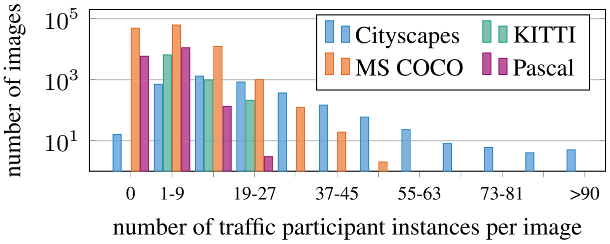


Figure 3. Dataset statistics regarding scene complexity. Only MS COCO and Cityscapes provide instance segmentation masks.

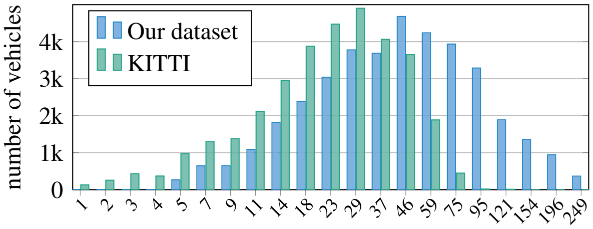


Figure 4. Histogram of object distances in meters for class *vehicle*.

Finally, we assess scene complexity, where density and scale of traffic participants (humans and vehicles) serve as proxy measures. Out of the previously discussed datasets, only Cityscapes and KITTI provide instance-level annotations for humans and vehicles. We additionally compare to the Caltech Pedestrian Dataset [12], which only contains annotations for humans, but none for vehicles. Furthermore, KITTI and Caltech only provide instance-level annotations in terms of axis-aligned bounding boxes. We use the respective training and validation splits for our analysis, since test set annotations are not publicly available for all datasets. In absolute terms, Cityscapes contains significantly more object instance annotations than KITTI, see Tab. 2. Being a specialized benchmark, the Caltech dataset provides the most annotations for humans by a margin. The major share of those labels was obtained, however, by interpolation between a sparse set of manual annotations resulting in significantly degraded label quality. The relative statistics emphasize the much higher complexity of Cityscapes, as the average numbers of object instances per image notably exceed those of KITTI and Caltech. We extend our analysis to MS COCO [38] and PASCAL VOC [14] that also contain street scenes while not being specific for them. We analyze the frequency of scenes with a certain number of traffic participant instances, see Fig. 3. We find our dataset to cover a greater variety of scene complexity and to have a higher portion of highly complex scenes than previous datasets. Using stereo data, we analyze the distribution of vehicle distances to the camera. From Fig. 4 we observe, that in comparison to KITTI, Cityscapes covers a larger distance range. We attribute this to both our higher-resolution imagery and the careful annotation procedure. As a consequence, algorithms need to take a larger range of scales and object sizes into account to score well in our benchmark.

3. Semantic Labeling

The first Cityscapes task involves predicting a per-pixel *semantic labeling* of the image without considering higher-level object instance or boundary information.

3.1. Tasks and metrics

To assess labeling performance, we rely on a standard and a novel metric. The first is the standard Jaccard Index, commonly known as the PASCAL VOC intersection-over-union metric $\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}}$ [14], where TP, FP, and FN are the numbers of true positive, false positive, and false negative pixels, respectively, determined over the whole test set. Owing to the two semantic granularities, *i.e.* classes and categories, we report two separate mean performance scores: $\text{IoU}_{\text{category}}$ and $\text{IoU}_{\text{class}}$. In either case, pixels labeled as void do not contribute to the score.

The global IoU measure is biased toward object instances that cover a large image area. In street scenes with their strong scale variation this can be problematic. Specifically for traffic participants, which are the key classes in our scenario, we aim to evaluate how well the individual instances in the scene are represented in the labeling. To address this, we *additionally* evaluate the semantic labeling using an instance-level intersection-over-union metric $\text{iIoU} = \frac{\text{iTP}}{\text{iTP+FP+iFN}}$. Here, iTP, and iFN denote weighted counts of true positive and false negative pixels, respectively. In contrast to the standard IoU measure, the contribution of each pixel is weighted by the ratio of the class' average instance size to the size of the respective ground truth instance. As before, FP is the number of false positive pixels. It is important to note here that unlike the instance-level task in Sec. 4, we assume that the methods only yield a standard per-pixel semantic class labeling as output. Therefore, the false positive pixels are not associated with any instance and thus do not require normalization. The final scores, $\text{iIoU}_{\text{category}}$ and $\text{iIoU}_{\text{class}}$, are obtained as the means for the two semantic granularities, while only classes with instance annotations are included.

3.2. Control experiments

We conduct several control experiments to put our baseline results below into perspective. First, we count the relative frequency of every class label at each pixel location of the fine (coarse) training annotations. Using the most frequent label at each pixel as a constant prediction irrespective of the test image (called *static fine*, SF, and *static coarse*, SC) results in roughly 10 % $\text{IoU}_{\text{class}}$, as shown in Tab. 3. These low scores emphasize the high diversity of our data. SC and SF having similar performance indicates the value of our additional coarse annotations. Even if the ground truth (GT) segments are re-classified using the most frequent training label (SF or SC) within each segment mask, the performance does not notably increase.

Secondly, we re-classify each ground truth segment using FCN-8s [41], *c.f.* Sec. 3.4. We compute the average scores within each segment and assign the maximizing label. The performance is significantly better than the static predictors but still far from 100 %. We conclude that it is necessary to optimize both classification and segmentation quality at the same time.

Thirdly, we evaluate the performance of subsampled ground truth annotations as predictors. Subsampling was done by majority voting of neighboring pixels, followed by resampling back to full resolution. This yields an upper bound on the performance at a fixed output resolution and is particularly relevant for deep learning approaches that often apply downscaling due to constraints on time, memory, or the network architecture itself. Downsampling factors 2 and 4 correspond to the most common setting of our 3rd-party baselines (Sec. 3.4). Note that while subsampling by a factor of 2 hardly affects the IoU score, it clearly decreases the iIoU score given its comparatively large impact on small, but nevertheless important objects. This underlines the importance of the separate instance-normalized evaluation. The downsampling factors of 8, 16, and 32 are motivated by the corresponding strides of the FCN model. The performance of a GT downsampling by a factor of 64 is comparable to the current state of the art, while downsampling by a factor of 128 is the smallest (power of 2) downsampling for which all images have a distinct labeling.

Lastly, we employ 128-times subsampled annotations and retrieve the nearest training annotation in terms of the Hamming distance. The full resolution version of this training annotation is then used as prediction, resulting in 21 % $\text{IoU}_{\text{class}}$. While outperforming the static predictions, the poor result demonstrates the high variability of our dataset and its demand for approaches that generalize well.

3.3. State of the art

Drawing on the success of deep learning algorithms, a number of semantic labeling approaches have shown very promising results and significantly advanced the state of the art. These new approaches take enormous advantage from recently introduced large-scale datasets, *e.g. PASCAL-Context* [45] and *Microsoft COCO* [38]. Cityscapes aims to complement these, particularly in the context of understanding complex urban scenarios, in order to enable further research in this area.

The popular work of Long *et al.* [41] showed how a top-performing Convolutional Neural Network (CNN) for image classification can be successfully adapted for the task of semantic labeling. Following this line, [9, 37, 40, 63, 81] propose different approaches that combine the strengths of CNNs and Conditional Random Fields (CRFs).

Other work takes advantage of deep learning for explicitly integrating global scene context in the prediction

Average over	Classes		Categories	
Metric [%]	IoU	iIoU	IoU	iIoU
static fine (SF)	10.1	4.7	26.3	19.9
static coarse (SC)	10.3	5.0	27.5	21.7
GT segmentation with SF	10.1	6.3	26.5	25.0
GT segmentation with SC	10.9	6.3	29.6	27.0
GT segmentation with [41]	79.4	52.6	93.3	80.9
GT subsampled by 2	97.2	92.6	97.6	93.3
GT subsampled by 4	95.2	90.4	96.0	91.2
GT subsampled by 8	90.7	82.8	92.1	83.9
GT subsampled by 16	84.6	70.8	87.4	72.9
GT subsampled by 32	75.4	53.7	80.2	58.1
GT subsampled by 64	63.8	35.1	71.0	39.6
GT subsampled by 128	50.6	21.1	60.6	29.9
nearest training neighbor	21.3	5.9	39.7	18.6

Table 3. Quantitative results of control experiments for semantic labeling using the metrics presented in Sec. 3.1.

	train	val	coarse	sub	Classes		Categories	
					IoU	iIoU	IoU	iIoU
FCN-32s	✓	✓			61.3	38.2	82.2	65.4
FCN-16s	✓	✓			64.3	41.1	84.5	69.2
FCN-8s	✓	✓			65.3	41.7	85.7	70.1
FCN-8s	✓	✓		2	61.9	33.6	81.6	60.9
FCN-8s		✓			58.3	37.4	83.4	67.2
FCN-8s			✓		58.0	31.8	78.2	58.4
[4] extended	✓			4	56.1	34.2	79.8	66.4
[4] basic	✓			4	57.0	32.0	79.1	61.9
[40]	✓	✓	✓	3	59.1	28.1	79.5	57.9
[81]	✓			2	62.5	34.4	82.7	66.0
[9]	✓	✓		2	63.1	34.5	81.2	58.7
[48]	✓	✓	✓	2	64.8	34.9	81.3	58.7
[37]	✓				66.4	**46.7**	82.8	67.4
[79]	✓				**67.1**	42.0	**86.5**	**71.1**

Table 4. Quantitative results of baselines for semantic labeling using the metrics presented in Sec. 3.1. The first block lists results from our own experiments, the second from those provided by 3rd parties. All numbers are given in percent and we indicate the used training data for each method, *i.e.* train fine, val fine, coarse extra as well as a potential downscaling factor (sub) of the input image.

of pixel-wise semantic labels, in particular through CNNs [4,39,44,67] or Recurrent Neural Networks (RNNs) [8,52]. Furthermore, a novel CNN architecture explicitly designed for dense prediction has been proposed recently by [79].

Last but not least, several studies [5,11,48–50,53,74,76] lately have explored different forms of weak supervision, such as bounding boxes or image-level labels, for training CNNs for pixel-level semantic labeling. We hope our coarse annotations can further advance this area.

3.4. Baselines

Our own baseline experiments (Tab. 4, top) rely on fully convolutional networks (FCNs), as they are central to most state-of-the-art methods [9, 37, 41, 63, 81]. We adopted VGG16 [68] and utilize the PASCAL-context setup [41] with a modified learning rate to match our image resolution under an unnormalized loss. According to the notation in [41], we denote the different models as *FCN-32s*, *FCN-16s*, and *FCN-8s*, where the numbers are the stride of the finest heatmap. Since VGG16 training on 2 MP images exceeds even the largest GPU memory available, we split each image into two halves with sufficiently large overlap. Additionally, we trained a model on images downscaled by a factor of 2. We first train on our training set (*train*) until the performance on our validation set (*val*) saturates, and then retrain on *train+val* with the same number of epochs.

To obtain further baseline results, we asked selected groups that have proposed state-of-the-art semantic labeling approaches to optimize their methods on our dataset and evaluated their predictions on our test set. The resulting scores are given in Tab. 4 (bottom) and qualitative examples of three selected methods are shown in Fig. 5. Interestingly enough, the performance ranking in terms of the main $\text{IoU}_{\text{class}}$ score on Cityscapes is highly different from PASCAL VOC [14]. While DPN [40] is the 2nd best method on PASCAL, it is only the 6th best on Cityscapes. FCN-8s [41] is last on PASCAL, but 3rd best on Cityscapes. Adelaide [37] performs consistently well on both datasets with rank 1 on PASCAL and 2 on Cityscapes.

From studying these results, we draw several conclusions: (1) The amount of downscaling applied during training and testing has a strong and consistent negative influence on performance (*c.f*. *FCN-8s vs. FCN-8s* at half resolution, as well as the 2nd half of the table). The ranking according to $\text{IoU}_{\text{class}}$ is strictly consistent with the degree of downscaling. We attribute this to the large scale variation present in our dataset, *c.f*. Fig. 4. This observation clearly indicates the demand for additional research in the direction of memory and computationally efficient CNNs when facing such a large-scale dataset with high-resolution images. (2) Our novel iIoU metric treats instances of any size equally and is therefore more sensitive to errors in predicting small objects compared to the IoU. Methods that leverage a CRF for regularization [9, 40, 48, 81] tend to over smooth small objects, *c.f*. Fig. 5, hence show a larger drop from IoU to iIoU than [4] or FCN-8s [41]. [37] is the only exception; its specific FCN-derived pairwise terms apparently allow for a more selective regularization. (3) When considering $\text{IoU}_{\text{category}}$, Dilated10 [79] and FCN-8s [41] perform particularly well, indicating that these approaches produce comparatively many confusions between the classes within the same category, *c.f*. the buses in Fig. 5 (top). (4) Training FCN-8s [41] with 500 densely annotated

Dataset	Best reported result	Our result
Camvid [7]	62.9 [4]	72.6
KITTI [58]	61.6 [4]	70.9
KITTI [64]	82.2 [73]	81.2

Table 5. Quantitative results (avg. recall in percent) of our half-resolution FCN-8s model trained on Cityscapes images and tested on Camvid and KITTI.

images (750 h of annotation) yields comparable IoU performance to a model trained on 20 000 weakly annotated images (1300 h annot.), *c.f.* rows 5 & 6 in Tab. 4. However, in both cases the performance is significantly lower than *FCN-8s* trained on all 3475 densely annotated images. Many fine labels are thus important for training standard methods as well as for testing, but the performance using coarse annotations only does not collapse and presents a viable option. (5) Since the coarse annotations do not include small or distant instances, their iIoU performance is worse. (6) Coarse labels can complement the dense labels if applying appropriate methods as evidenced by [48] outperforming [9], which it extends by exploiting both dense and weak annotations (*e.g.* bounding boxes). Our dataset will hopefully stimulate research on exploiting the coarse labels further, especially given the interest in this area, *e.g.* [25, 43, 47].

Overall, we believe that the unique characteristics of our dataset (*e.g.* scale variation, amount of small objects, focus on urban street scenes) allow for more such novel insights.

3.5. Cross-dataset evaluation

In order to show the compatibility and complementarity of Cityscapes regarding related datasets, we applied an FCN model trained on our data to Camvid [7] and two subsets of KITTI [58, 64]. We use the half-resolution model (*c.f.* 4th row in Tab. 4) to better match the target datasets, but we do not apply any specific training or fine-tuning. In all cases, we follow the evaluation of the respective dataset to be able to compare to previously reported results [4, 73]. The obtained results in Tab. 5 show that our large-scale dataset enables us to train models that are on a par with or even outperforming methods that are specifically trained on another benchmark and specialized for its test data. Further, our analysis shows that our new dataset integrates well with existing ones and allows for cross-dataset research.

4. Instance-Level Semantic Labeling

The pixel-level task, *c.f.* Sec. 3, does not aim to segment individual object instances. In contrast, in the instance-level semantic labeling task, we focus on simultaneously detecting objects and segmenting them. This is an extension to both traditional object detection, since per-instance segments must be provided, and semantic labeling, since each instance is treated as a separate label.

4.1. Tasks and metrics

For instance-level semantic labeling, algorithms are required to deliver a set of detections of traffic participants in the scene, each associated with a confidence score and a per-instance segmentation mask. To assess instance-level performance, we compute the average precision on the region level (AP [23]) for each class and average it across a range of overlap thresholds to avoid a bias towards a specific value. Specifically, we follow [38] and use 10 different overlaps ranging from 0.5 to 0.95 in steps of 0.05. The overlap is computed at the region level, making it equivalent to the IoU of a single instance. We penalize multiple predictions of the same ground truth instance as false positives. To obtain a single, easy to compare compound score, we report the mean average precision AP, obtained by also averaging over the class label set. As minor scores, we add $\text{AP}^{50\%}$ for an overlap value of 50 %, as well as $\text{AP}^{100\text{m}}$ and $\text{AP}^{50\text{m}}$ where the evaluation is restricted to objects within 100 m and 50 m distance, respectively.

4.2. State of the art

As detection results have matured (70 % mean AP on PASCAL [14, 55]), the last years have seen a rising interest in more difficult settings. Detections with pixel-level segments rather than traditional bounding boxes provide a richer output and allow (in principle) for better occlusion handling. We group existing methods into three categories.

The first encompasses *segmentation, then detection* and most prominently the R-CNN detection framework [21], relying on object proposals for generating detections. Many of the commonly used bounding box proposal methods [28, 54] first generate a set of overlapping segments, *e.g.* Selective Search [72] or MCG [1]. In R-CNN, bounding boxes of each segment are then scored using a CNN-based classifier, while each segment is treated independently.

The second category encompasses *detection, then segmentation*, where bounding-box detections are refined to instance specific segmentations. Either CNNs [23, 24] or non-parametric methods [10] are typically used, however, in both cases without coupling between individual predictions.

Third, simultaneous *detection and segmentation* is significantly more delicate. Earlier methods relied on Hough voting [36, 57]. More recent works formulate a joint inference problem on pixel and instance level using CRFs [11, 26, 42, 71, 78, 80]. Differences lie in the generation of proposals (exemplars, average class shape, direct regression), the cues considered (pixel-level labeling, depth ordering), and the inference method (probabilistic, heuristics).

4.3. Lower bounds, oracles & baselines

In Tab. 6, we provide lower-bounds that any sensible method should improve upon, as well as oracle-case results

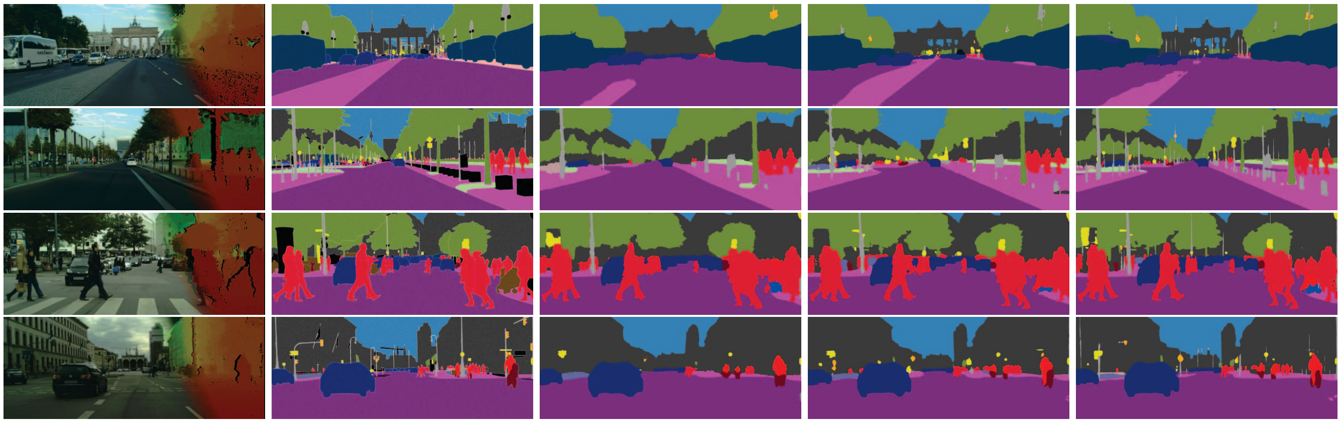


Figure 5. Qualitative examples of selected baselines. From left to right: image with stereo depth maps partially overlayed, annotation, DeepLab [48], Adelaide [37], and Dilated10 [79]. The color coding of the semantic classes matches Fig. 1.

Proposals	Classif.	AP	$\text{AP}^{50\%}$	$\text{AP}^{100\text{m}}$	$\text{AP}^{50\text{m}}$
MCG regions	FRCN	2.6	9.0	4.4	5.5
MCG bboxes	FRCN	3.8	11.3	6.5	8.9
MCG hulls	FRCN	**4.6**	**12.9**	**7.7**	**10.3**
GT bboxes	FRCN	8.2	23.7	12.6	15.2
GT regions	FRCN	41.3	41.3	58.1	64.9
MCG regions	GT	10.5	27.0	16.0	18.7
MCG bboxes	GT	9.9	25.8	15.3	18.9
MCG hulls	GT	11.6	29.1	17.7	21.4

Table 6. Baseline results on instance-level semantic labeling task using the metrics described in Sec. 4. All numbers in %.

(*i.e.* using the test time ground truth). For our experiments, we rely on publicly available implementations. We train a Fast-R-CNN (*FRCN*) detector [20] on our training data in order to score MCG object proposals [1]. Then, we use either its output bounding boxes as (rectangular) segmentations, the associated region proposal, or its convex hull as a per-instance segmentation. The best main score AP is 4.6 %, is obtained with convex hull proposals, and becomes larger when restricting the evaluation to 50 % overlap or close instances. We contribute these rather low scores to our challenging dataset, biased towards busy and cluttered scenes, where many, often highly occluded, objects occur at various scales, *c.f.* Sec. 2. Further, the MCG bottom-up proposals seem to be unsuited for such street scenes and cause extremely low scores when requiring large overlaps.

We confirm this interpretation with oracle experiments, where we replace the proposals at test-time with ground truth segments or replace the FRCN classifier with an oracle. In doing so, the task of object localization is decoupled from the classification task. The results in Tab. 6 show that when bound to MCG proposals, the oracle classifier is only slightly better than FRCN. On the other hand, when the proposals are perfect, FRCN achieves decent results. Overall, these observations unveil that the instance-level performance of our baseline is bound by the region proposals.

5. Conclusion and Outlook

In this work, we presented Cityscapes, a comprehensive benchmark suite that has been carefully designed to spark progress in semantic urban scene understanding by: (i) creating the largest and most diverse dataset of street scenes with high-quality and coarse annotations to date; (ii) developing a sound evaluation methodology for pixel-level and instance-level semantic labeling; (iii) providing an in-depth analysis of the characteristics of our dataset; (iv) evaluating several state-of-the-art approaches on our benchmark. To keep pace with the rapid progress in scene understanding, we plan to adapt Cityscapes to future needs over time.

The significance of Cityscapes is all the more apparent from three observations. First, the relative order of performance for state-of-the-art methods on our dataset is notably different than on more generic datasets such as PASCAL VOC. Our conclusion is that serious progress in urban scene understanding may not be achievable through such generic datasets. Second, the current state-of-the-art in semantic labeling on KITTI and CamVid is easily reached and to some extent even outperformed by applying an off-the-shelf fully-convolutional network [41] trained on Cityscapes only, as demonstrated in Sec. 3.5. This underlines the compatibility and unique benefit of our dataset. Third, Cityscapes will pose a significant new challenge for our field given that it is currently far from being solved. The best performing baseline for pixel-level semantic segmentation obtains an IoU score of 67.1 %, whereas the best current methods on PASCAL VOC and KITTI reach IoU levels of 77.9 % [3] and 72.5 % [73], respectively. In addition, the instance-level task is particularly challenging with an AP score of 4.6 %.

Acknowledgments. S. Roth was supported in part by the European Research Council under the EU's 7th Framework Programme (FP/2007-2013)/ERC Grant agreement no. 307942. The authors acknowledge the support of the Bundesministerium für Wirtschaft und Technologie (BMWi) in the context of the UR:BAN initiative. We thank the 3rd-party authors for their valuable submissions.

References

[1] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marqués, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 7, 8

[2] S. Ardeshir, K. M. Collins-Sibley, and M. Shah. Geo-semantic segmentation. In *CVPR*, 2015. 4

[3] A. Arnab, S. Jayasumana, S. Zheng, and P. H. S. Torr. Higher order conditional random fields in deep neural networks. *arXiv:1511.08119v3 [cs.CV]*, 2015. 8

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv:1511.00561v2 [cs.CV]*, 2015. 1, 6, 7

[5] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. *arXiv:1506.02106v4 [cs.CV]*, 2015. 6

[6] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *CVPR*, 2012. 1

[7] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 1, 3, 4, 7

[8] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene Labeling with LSTM Recurrent Neural Networks. In *CVPR*, 2015. 6

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015. 5, 6, 7

[10] Y.-T. Chen, X. Liu, and M.-H. Yang. Multi-instance object segmentation with occlusion handling. In *CVPR*, 2015. 7

[11] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015. 6, 7

[12] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *Trans. PAMI*, 34(4):743–761, 2012. 1, 4

[13] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. 31(12):2179–2195, 2009. 1

[14] 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):98–136, 2015. 1, 4, 5, 6, 7

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Trans. PAMI*, 32(9):1627–1645, 2010. 1

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

[17] P. Furgale, U. Schwesinger, M. Rufli, W. Derendarz, H. Grimmett, P. Mühlfellner, S. Wonneberger, B. Li, et al. Toward automated driving in cities using close-to-market sensors: An overview of the V-Charge project. In *IV Symposium*, 2013. 1

[18] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3D traffic scene understanding from movable platforms. *Trans. PAMI*, 36(5):1012–1025, 2014. 1

[19] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 32(11), 2013. 1, 3, 4

[20] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 8

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 7

[22] F. Gueney and A. Geiger. Displets: Resolving stereo ambiguities using object knowledge. In *CVPR*, 2015. 4

[23] B. Hariharan, P. Arbeláez, R. B. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 7

[24] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 7

[25] H. Hattori, V. N. Boddeti, K. M. Kitani, and T. Kanade. Learning scene-specific pedestrian detectors without real data. In *CVPR*, 2015. 7

[26] X. He and S. Gould. An exemplar-based CRF for multi-instance object segmentation. In *CVPR*, 2014. 7

[27] D. Hoiem, J. Hays, J. Xiao, and A. Khosla. Guest editorial: Scene understanding. *IJCV*, 2015. 1

[28] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? *Trans. PAMI*, 38(4):814–830, 2015. 7

[29] H. Hu and B. Upcroft. Nonparametric semantic segmentation for 3D street scenes. In *IROS*, 2013. 4

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[31] L. Krüger, C. Wöhler, A. Würz-Wessel, and F. Stein. In-factory calibration of multiocular camera systems. In *SPIE Photonics Europe (Optical Metrology in Production Engineering)*, 2004. 2

[32] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. Rehg. Joint semantic segmentation and 3D reconstruction from monocular video. In *ECCV*, 2014. 4

[33] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *CVPR*, 2014. 4

[34] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 2015. 1

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

[36] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1-3):259–289, 2008. 7

[37] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016, to appear. 5, 6, 8

[38] 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, 4, 5, 7

[39] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv:1506.04579v2 [cs.CV]*, 2015. 6

[40] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *ICCV*, 2015. 5, 6

[41] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 3, 5, 6, 8

[42] M. Maire, S. X. Yu, and P. Perona. Object detection and segmentation from joint embedding of parts and pixels. In *ICCV*, 2011. 7

[43] I. Misra, A. Shrivastava, and M. Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *CVPR*, 2015. 7

[44] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feed-forward semantic segmentation with zoom-out features. In *CVPR*, 2015. 6

[45] 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, 5

[46] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 1

[47] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 7

[48] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. In *ICCV*, 2015. 6, 7, 8

[49] D. Pathak, P. Kraehenbuehl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 6

[50] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR*, 2015. 6

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

[52] P. H. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene parsing. In *ICML*, 2014. 6

[53] P. H. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 6

[54] J. Pont-Tuset and L. Van Gool. Boosting object proposals: From Pascal to COCO. In *ICCV*, 2015. 7

[55] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 7

[56] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool. Learning where to classify in multi-view semantic segmentation. In *ECCV*. 2014. 4

[57] H. Riemenschneider, S. Sternig, M. Donoser, P. M. Roth, and H. Bischof. Hough regions for joining instance localization and segmentation. In *ECCV*, 2012. 7

[58] 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, 4, 7

[59] 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. *IJCV*, 115(3):211–252, 2015. 1

[60] 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):157–173, 2008. 2

[61] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth. Efficient multi-cue scene segmentation. In *GCPR*, 2013. 1, 3

[62] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth. Stixmantics: A medium-level model for real-time semantic scene understanding. In *ECCV*, 2014. 1, 4

[63] A. Schwing and R. Urtasun. Fully connected deep structured networks. *arXiv:1503.02351v1 [cs.CV]*, 2015. 5, 6

[64] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. S. Torr. Semantic modelling of urban scenes. In *ICRA*, 2013. 4, 7

[65] S. Sengupta, P. Sturgess, L. Ladicky, and P. H. S. Torr. Automatic dense visual semantic mapping from street-level imagery. In *IROS*, 2012. 4

[66] 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. 1

[67] A. Sharma, O. Tuzel, and D. W. Jacobs. Deep hierarchical parsing for semantic segmentation. In *CVPR*, 2015. 6

[68] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556v6 [cs.CV]*, 2014. 6

[69] S. Song, S. P. Lichtenberg, and J. Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015. 4

[70] J. Tighe and S. Lazebnik. Superparsing. *IJCV*, 101(2):329–349, 2013. 4

[71] J. Tighe, M. Niethammer, and S. Lazebnik. Scene parsing with object instance inference using regions and per-exemplar detectors. *IJCV*, 112(2):150–171, 2015. 1, 7

[72] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 7

[73] V. Vineet, O. Miksik, M. Lidegaard, M. Niessner, S. Golodetz, V. A. Prisacariu, O. Kahler, D. W. Murray, S. Izadi, P. Perez, and P. H. S. Torr. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *ICRA*, 2015. 7, 8

[74] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, Y. Zhao, and S. Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. *arXiv:1509.03150v1 [cs.CV]*, 2015. 6

[75] J. Xie, M. Kiefel, M.-T. Sun, and A. Geiger. Semantic instance annotation of street scenes by 3D to 2D label transfer. In *CVPR*, 2016, to appear. 2, 3

[76] J. Xu, A. G. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *CVPR*, 2015. 6

[77] P. Xu, F. Davoine, J.-B. Bordes, H. Zhao, and T. Denoeux. Information fusion on oversegmented images: An application for urban scene understanding. In *MVA*, 2013. 4

[78] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 7

[79] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016, to appear. 6, 8

[80] Z. Zhang, A. Schwing, S. Fidler, and R. Urtasun. Monocular object instance segmentation and depth ordering with CNNs. In *ICCV*, 2015. 4, 7

[81] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 5, 6

[82] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 1