The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, Peter Kontschieder
Mapillary Research
research@mapillary.com

Abstract

The Mapillary Vistas Dataset is a novel, large-scale street-level image dataset containing 25 000 high-resolution images annotated into 66 object categories with additional, instance-specific labels for 37 classes. Annotation is performed in a dense and fine-grained style by using polygons for delineating individual objects. Our dataset is 5× larger than the total amount of fine annotations for Cityscapes and contains images from all around the world, captured at various conditions regarding weather, season and daytime. Images come from different imaging devices (mobile phones, tablets, action cameras, professional capturing rigs) and differently experienced photographers. In such a way, our dataset has been designed and compiled to cover diversity, richness of detail and geographic extent. As default benchmark tasks, we define semantic image segmentation and instance-specific image segmentation, aiming to significantly further the development of state-of-the-art methods for visual road-scene understanding.

1. Introduction

Visual scene understanding [19] belongs to the most fundamental and challenging goals in computer vision. It comprises many different levels of abstraction and a large body of research groups have recently contributed to significantly pushing the state of the art in the field. Scene *recognition* puts the emphasis on providing a global description of a scene, typically summarized at a single scene category level [41, 64]. *Object detection* focuses on finding object instances and their categories within a scene, typically localized using bounding-boxes [13, 14, 34, 43, 44]. *Semantic segmentation* emphasizes on providing a finer-grained prediction of the semantic category each pixels belongs to [5, 28, 36, 37, 50, 62], while *instance-specific seman-tic segmentation* adds the difficulty of identifying the pixels that compose each object instance, thus integrating semantic segmentation with fine-grained object detection [9, 25].

In the last few years the computer vision field has been undergoing a revolution, mainly driven by the big successes of deep learning [22]. Indeed, most top-performing algorithms for visual scene understanding are nowadays developed using deep learning. However, it is well-known that training of deep learning models requires a substantial amount of (annotated) data and computational resources. Accordingly, the availability of large-scale datasets such as *ImageNet* [48], *Places* [63], *PASCAL VOC* [10], *PASCAL-Context* [39], *Microsoft COCO* [30], *ADE20K* [65] and *YouTube-8M* [1] is of major importance for both, the scientific and industrial communities. Another recent trend involves integrating synthetically-rendered data from sources like the *Grand Theft Auto V engine* [45], *Synthia* [46] or semantic instance labeling via 3D to 2D label transfer [56].

An application field that currently attracts a lot of interest from both, the industrial and scientific community is that of self-driving vehicles, where the decision-making component is mostly based on visual data analysis and requires reliable, real-time semantic image understanding. *E.g.*, autonomously driving vehicles or navigating robots need to comprehend street-level images in terms of relevant object categories, being able to precisely locate and enumerate them. To gain deeper understanding of the complex interactions between *e.g.* traffic participants in street-level images, a lot of research efforts and investments have recently gone into designing and creating datasets which were used to train deep models for this specific task, such as *CamVid* [3], the *KITTI Vision Benchmark Suite* [12], *Leuven* [23], *Daimler Urban Segmentation* [49] and *Cityscapes* [6].

As opposed to datasets addressing more general tasks, datasets for semantic street scene understanding are often

limited in their total number of fine-grained annotated images, the overall number of object categories, are restricted to specific capturing areas, or urban scenes, thus strongly biasing the appearance of the elements to be analysed. In addition, such datasets may suffer from a bias towards specific capturing modalities due to the usage of a sole imaging sensor and therefore not properly covering the spectrum of available sensor noise or overly strict specifications on camera mounting. In essence, neglecting such real world conditions impairs the overall amount of a datasets expressiveness and restrains the diversity.

Contributions. Our contribution is a new dataset for semantic segmentation of urban, countryside and off-road scenes comprising 25 000 densely-annotated street level images into 66 object categories, featuring locations from all around the world, and taken from a diverse source of image capturing devices. Images are annoted in a fine-grained style by using polygons for annotation and contain instance-specific object annotations for 37 object categories. This dataset is $5\times$ larger than Cityscapes [6] (in terms of fine-grained annotations) and exhibits a significantly larger variability in terms of geographic origins and number of object categories. The image data is extracted from *Mapillary*[1] and visually covers parts of Europe, North and South America, Asia, Africa and Oceania (see, Fig. 4), consequently addressing the global spectrum of possible object appearances. In addition, we are proposing a statistical protocol for quality assurance (QA), such that targeted annotation accuracies for precision and recall can be monitored and verified. The dataset is also available as commercial edition with annotations for 100 object classes (with instance-specific annotations for 58 classes), providing more detailed semantics for autonomous driving specific categories.

2. The Mapillary Vistas Dataset

The proposed dataset is built upon images extracted from www.mapillary.com . Mapillary is a community-led service for people who collaboratively want to visualize the world with street-level photos. Anyone can contribute with photos of any place and the data is available for anyone to explore under a CC-BY-SA license agreement.

The proposed dataset is designed to capture the broad range of outdoor scenes available around the world. While such design choices primarily affect the image content, a broader interpretation of sampling data from around the world also includes the data recording modality and the sensing equipment. In what follows, we try to characterize our dataset according to the targeted distributions for geographical and seasonal distribution, weather conditions, viewing perspectives, capturing time, image resolution, *i.e.* the diversity of images taken in the wild.

The following characteristics describe the dataset properties for the complete dataset in terms of training and validation data (18 000 + 2 000 images, respectively). The remaining 5 000 test images are sampled proportionally to the geographic training data distribution and their annotation characteristics remain undisclosed for fair comparisons between novel methods on our benchmark server[2].

2.1. Dataset Compilation Process

The prime motivations for designing and introducing this dataset were *diversity*, *richness of detail* and *geographic extent* of street-level data. Given these requirements, the goal was to compile a dataset with reduced bias towards highly developed countries and instead reflect the heterogeneous compositions in appearance that can be found on a street level perspective around the globe. To achieve this goal, we deployed and followed a three-fold process:

1. From a large pool of available images (we were able to browse from a repository of $\approx$140 million images), the initial selection process was made according to criteria defined in Sect. 2.1.1.
2. An image approved for annotation was segmented by one of 69 annotators in a dense, polygon-based and instance-specific manner as described in Sect. 2.1.2
3. Each annotated image was followed by a single-stage quality assurance (QA) process. In order to guarantee high annotation accuracy, an external annotation party was stochastically used for a second QA phase in order to maximize both, precision and recall rates. The QA processes are discussed in Sect. 2.1.3.

2.1.1 Image and Category Selection

In order to accept an image for the semantic annotation process, several criteria have to be met. First, the image resolution has to be at least full HD, *i.e.* a minimum width/height of 1920$\times$ 1080 was imposed. Additionally, $\approx 90\%$ of the images should be selected from road/sidewalk views in urban areas and the remaining ones from highways, rural areas[3] and off-road. Given these constraints, the database was queried in a way to randomly present potential candidates to a human for further evaluation and selection as follows.

Images with strong wide-angle view (focal length below 10mm) or 360° images were removed. Degraded images exhibiting strong motion blur, rolling shutter artifacts, interlacing artifacts, major windshield reflections or containing dominant image parts from the capturing vehicle/device (like car hood, camera mount or wipers) were removed as well. However, a small amount of distortion for motion blur was accepted, as long as individual objects could still be

recognized and precisely annotated. Since Mapillary images can belong to a sequence (*i.e.* a series of images taken with either manual, pre-fixed or distance-depending capture rate), we restrict to selecting images with unique views on a scene to avoid redundant labeling of image content.

In order to satisfy diversity criteria, images were selected to cover seasonal variability on both, northern and southern hemispheres and different weather showing sunny, rainy, cloudy, foggy or snowy conditions as well as varying lighting conditions like at dawn, dusk or even at night (see, Fig. 6). Moreover, images are taken from all continents (North and South America, Europe, Africa, Asia, Wider Geographic Oceania) except for Antarctica. An image was counted to belong to a city area when their population number was $\geq 50\,000$ and to a rural area otherwise.

Images are preferably selected for annotation in case they contain i) multiple object categories, ii) multiple instances of particular object categories and/or iii) rarely appearing object categories (according to histogram visualization in Fig. 1. In addition, objects appearing in the images should be reasonably close to the camera center instead of being only far away. Another way to select images for annotation is to run a segmentation model on a large number of previously unseen data (from the global image pool), after training it on the already available data and manually check for images with qualitatively lowest performance. In principle, we however tried to select images without introducing a bias towards specific machine annotations.

Categories. We distinguish between 66 visual object categories, mostly pertaining to a street-level scene domain. The categories are organized into 7 root-level groups (see, Fig. 1), namely *object*, *construction*, *human*, *marking*, *nature*, *void* and *animal*. Each root-level group is organized into different macro-groups. A subset of $37/66$ categories are additionally annotated in an instance-specific way. *E.g.*, classes like cars, pedestrians, cyclists are labeled on the basis of individual objects and are therefore individually accessible within an image.

The criteria to add categories to the annotation process were driven by several factors. Inspired by earlier works like Cityscapes [6] and SIFT Flow [31], we used the street-level and nature categories therein. In addition, taking a closer look at open map initiatives like www.openstreetmap.org inspired many of class selections in root- and macro-categories *object*, *barrier* and *flat*. To help with recognition tasks in research for autonomous driving, special emphasis was put on properly annotating different classes within macro-groups *vehicle* or *traffic sign*.

2.1.2 Image Annotation

Image annotation was conducted by a team of 69 professional image annotators, delivering an average rate of ≈ 5.1 images per annotator and day. Consequently, the average annotation time is at around 94 minutes per image, which is in line with what is reported in [6]. From the pool of annotators, 11 people are performing an internal round of first level quality assurance (QA, see description in next subsection; takes approximately 15 minutes per image and is included in 94 minutes), *i.e.* for each team of 5 annotators, one person is performing a 100% check of annotations.

The image annotation process was inspired by the one reported for Cityscapes [6]. Each annotator is encouraged to start with annotating the images with object categories from back to front, *i.e.* annotation is typically started with sky (the most distant category to the camera center) and gradually works closer to the camera. In this way, a *z*-ordering for individual object layers can be imposed, allowing to rasterize the final label images. Consequently, sky might be annotated with a single polygon despite showing several areas in the image. Another positive aspect of such an annotation protocol is that areas eventually getting overruled from objects closer to the camera can be drawn faster and not necessarily have to be corrected in case foreground objects labels need to be refined.

Designing the annotation protocol for each object category requires a well-defined object taxonomy and systematic instructions including fallback annotation solutions for each object category (see, supplementary material for category descriptions). In such a way and via regular feedback sessions with the annotation specialists, a commonly-agreed and jointly-consistent protocol was developed.

For annotation, we implemented a tool with graphical user interface, allowing annotators to seamlessly zoom to the required level of detail, change the level of opacity, quickly provide category selection and necessary toolboxes for simple drawing and modifications functions for polygons. In addition, the tool allows to sort objects in correct *z*-order and provides a visualization bar, indicating minimal object sizes to be annotated for a given zoom level.

2.1.3 Quality Assurance

We adopt a two-stage quality assurance (QA) process, which is targeting instance-specific annotation accuracies with $\geq 97\%$ for both, precision and recall. The first round of QA is applied to *each* image and is conducted as follow-up step after annotation to correct potential mislabeling in terms of precision and recall. The person conducting QA is different from the annotator but reports back major issues to the annotator for improving the initial annotation quality.

In order to further improve the final annotation quality, we install a second QA process guided by a modified variant of the four-eyes principle. To this end, a second (and therefore different) annotation provider revisits selected images and is incentivized to spot errors measured by the weighted intersection over union criterion as typically used to assess the performance for instance-specific semantic segmenta-
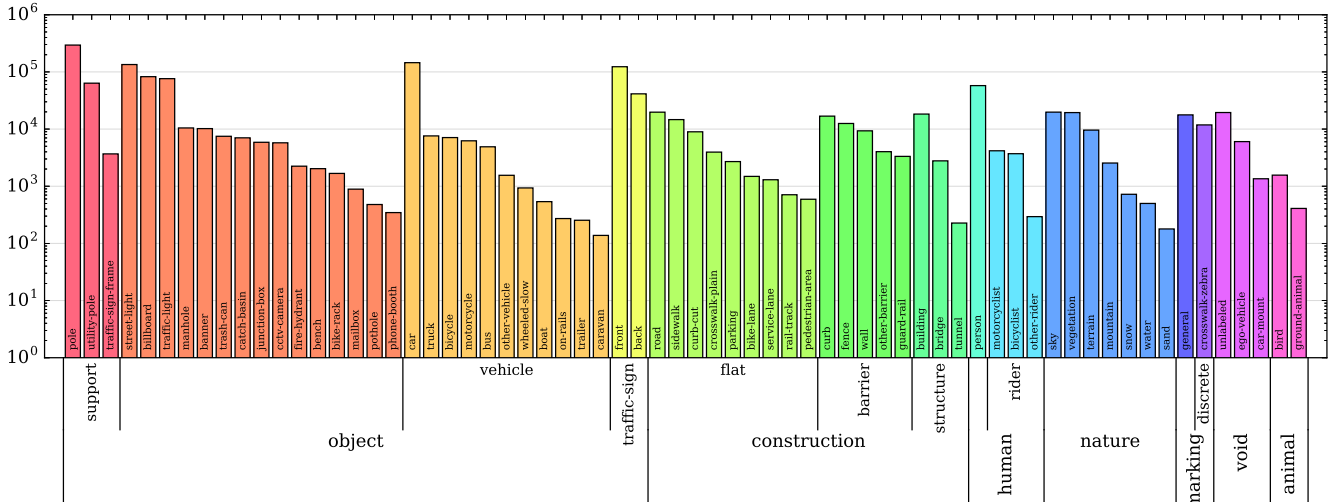
Figure 1. Illustration of number of labeled instances per category and corresponding macro- and root-level class.

tion [15]. Next, we describe how we determine the sample size for the second round of QA in order to probabilistically validate the anticipated annotation accuracy.

QA Protocol. Let $\mathcal{X}$ be the set of true object instances of a specific category among the dataset and let $q : \mathcal{X} \to \{0, 1\}$ be a deterministic function that checks if the quality of the segmentation provided by the annotator for the specific object instance is good-enough (*i.e.* $q(x) = 1$) or not (*i.e.* $q(x) = 0$). *E.g.*, the quality function in our case is $q(x) = 1$ if and only if both precision and recall of the segmentation are $\geq 97\%$. Let $\mathcal{S}_n = \{x_1, \ldots, x_n\}$ be a uniform random sample from $\mathcal{X}$ and let $q_i = q(x_i)$ be the binary outcome from the segmentation quality assessment on instance x_i. Also, let $s = \sum_{x \in \mathcal{X}} q(x)$ be the unknown number of correctly annotated images over the entire dataset, and let $N = |\mathcal{X}|$ be the population size. Following a Bayesian setting, the posterior distribution of s after observing quality assessments for the n object instances in $\mathcal{S}_n$ under Jeffrey's, non-informative prior[4] is *beta-binomial*(a, b, N) with beta parameters $a = s_n + \frac{1}{2}$ and $b = n - s_n + \frac{1}{2}$ and number of trials N, where $s_n = \sum_{i=1}^{n} q(x_i)$ is the empirical number of correctly labelled instances that we observe for instances in $\mathcal{S}_n$. We fix an accepted tolerance s_0 about the annotators' accuracy to $s_0 = 0.99N$ correctly annotated images, *i.e.* the annotator should correctly segment at least 99% of the instances in $\mathcal{X}$ according to the quality criterion q or, in other terms, we want $s \geq s_0$ to be satisfied. Next, we want to determine the smallest interval $l_n \leq s_0 \leq u_n$ around the tolerance s_0 ensuring the following two conditions: *(i)* $p(s > s_0 | s_n \leq l_n) \leq \alpha$ and *(ii)* $p(s < s_0 | s_n \geq u_n) \leq \alpha$, where $1 - \alpha = 99\%$ is our confidence level. The values of l_n and u_n depend on the sample size n and the larger n the smaller this interval will be. If no value of l_n satisfies

property *(i)* then we set $l_n = -\infty$, while if no value of u_n satisfies *(ii)* then we set $u_n = \infty$. The goal here is to determine, with confidence level $1 - \alpha$, whether the annotator succeeded, or failed, in annotating the entire set with accuracy $\geq 99\%$. We want to take this decision based on the observation of the number of successful segmentations s_n that were assessed by the quality checker from the small sample set $\mathcal{S}_n$. By definition of l_n and u_n, we achieve this goal by simply checking whether $s_n \geq u_n$ or $s_n \leq l_n$, respectively. If none of the conditions is satisfied, *i.e.* if $l_n < s_n < u_n$, then we need more evidence to draw a conclusion with the desired level of confidence from the Bayesian perspective.

To give a practical example, which is the one we implement, assume that we provide the quality checker with 400 images to check, which corresponds on average to about $n = 344$ instances per category. Assume also that the population size corresponds to the dataset size, *i.e.* $N = 25\,000$, and that our targets in terms of minimum accuracy and confidence level are as described above, *i.e.* 99%. Given this setting, we obtain $l_n = 335$ (*i.e.* $\approx 97.38\%n$) and $u_n = n$. The quality check assesses that s_n images out of n are correctly labelled. Now, if $s_n \leq l_n$ then with a level of confidence of 99% the annotator will fail to deliver an annotation accuracy $\geq 99\%$ for the entire dataset. If this happens we inform the annotator by showing typical error scenarios in order to induce an improvement in the quality of the annotations. Viceversa, if $s_n \geq u_n$ then under the same level of confidence we can assume that the entire dataset will be annotated with the desired quality target and we can interrupt further quality assessments. In the specific numerical example, this is true only if the annotator committed no error in annotating the instances in $\mathcal{S}_n$, for $u_n = n$. Otherwise, if $l_n < s_n < u_n$, we need to continue the quality assessment procedure on other sampled instances, because we can draw no conclusion under the given confidence level.

---

[4]If we had prior knowledge about the annotator's performance, this could be encoded into the prior.

To conclude, with the given QA protocol, we will guarantee with a confidence level of 99% that at least 99% of the instances in the dataset will have an instance-level segmentation with precision and recall $\geq 97\%$, under the assumption that the quality assessment will stop in the positive case $s_n \geq u_n$, given an available budget for QA.

2.2. Dataset Splitting

Similar to Cityscapes [6], PASCAL VOC [10], Microsoft COCO [30] or ADE20K [65], we decided to have a fixed dataset splitting into training, validation and test sets. We provide the labels for training and validation data and withhold the labels for the test data. Training and validation data comprise 18 000 and 2 000 images, respectively and the remaining 5 000 images form the test set. Each of the sets is compiled in a way to represent the characteristics described in Sect. 2.1.1. Each of the sets is additionally grouped by geographical regions. In such a way, segmentation assessment can be performed on more regional levels, potentially revealing classifier discrepancies *e.g.* when comparing results on images from North America to Asia.

3. Statistical Analysis

In this section we provide statistical analyses about the Mapillary Vistas dataset and put some aspects on the distribution of classes, volume of annotation, and complexity of the scene in relation to other standard benchmark datasets.

3.1. Image Varieties

The set of images that populate the dataset are diverse in terms of image resolution, focal length, and camera model. In Fig. 2, top-left, we report the distribution of the image resolution across the dataset. All images are at least FullHD, but we have images with more than 22 Mpixels. Most pictures are in landscape orientation. The dominant aspect ratio is 4:3, followed by 16:9, but also other ratios are represented (mean ratio is 1.38). In Fig. 2, top-right, we provide the distribution of the focal length, which is mostly concentrated in the 25-35mm range, but we find also images taken with a wide-angle with focal length ranging between 15-20mm. Finally, in Fig. 2, bottom, we report the distribution of the camera brands used to capture the images. We see that there is a predominance of mobile devices from Apple and Samsung, but in general the dataset spans a wide range of different camera types, also head- or car-mounted ones like Garmin and GoPro. All those statistics show that the Mapillary Vistas dataset is rich in terms of variability of the image acquisition sensors and settings. As opposed to benchmark datasets like Cityscapes [6] or CamVid [3] having a unique, ad-hoc acquisition system, images in our dataset better match the real-world distribution of possible capturing scenarios.
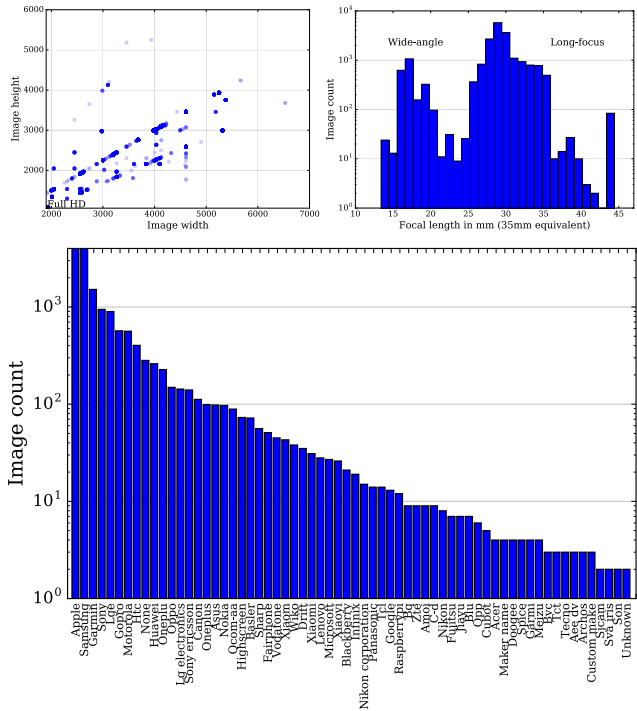

Figure 2. Top-left. Distribution of image resolution. Minimum size is fixed at Full HD (bottom left) and maximum image resolution is >22MPixels. Top-right. Focal length distribution. Bottom. Distribution of camera sensors/devices used for image capturing.

In Fig. 3 we characterize the level of expertise of users that indirectly contributed images to our dataset, in terms of their activity on Mapillary. Specifically, we show for a given number of contributed images to Mapillary, how many users have also contributed at least one image used in Mapillary Vistas dataset. As we can see, the distribution ranges from very active users, who uploaded up to 10M images, to users that uploaded around 30 images. The mode of the distribution is at about 35 000 uploaded images with 50 users. This is a weak indication that the majority of the pictures that form the Mapillary Vistas dataset were probably taken by experienced users and, in general, there is also a large variability of humans taking pictures.
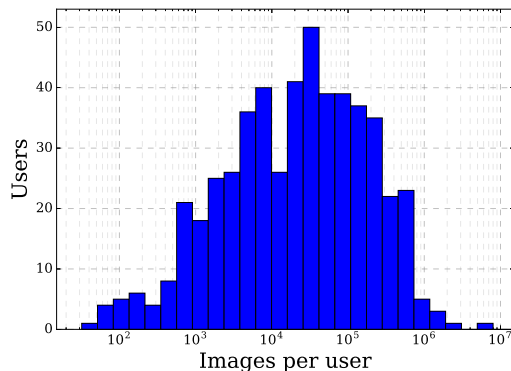

Figure 3. Histogram of totally committed images to Mapillary for user IDs behind images selected for annotation.

Finally, in Fig. 4 we characterize also the distribution of images in terms of their position in the world. To this end, we provide a geo-localized histogram superimposed on a map of the world. As we can see the Mapillary Vistas dataset spans all the continents excepting the Antarctic. This is a strong indication that the images in the Mapillary Vistas dataset are very diverse in terms of appearance of the different categories that are annotated, which is a distinctive feature of the proposed benchmark.
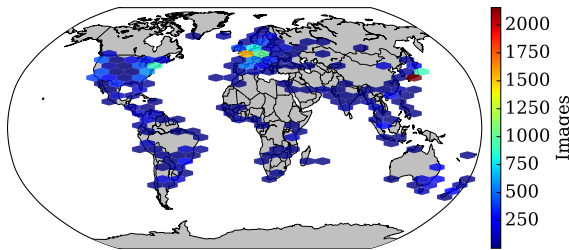

Figure 4. Distribution of image recordings.

3.2. Distribution of Instances and Comparisons

We report in Fig. 1 the total number of actual instances per class that have been segmented by an annotator. Each group is sorted in decreasing order of number of instances per class and coded with a different color. The root- and macro-level class *object* encompasses the majority of the instances and illustrates the broad number of separately annotated categories. Given such fine-grained annotations, a large number of additional attributes can be inferred to hopefully strengthen the research community towards development of better models for street scene understanding.

In Fig. 5, left, we compare the frequency of images with a fixed number of annotated objects in the Mapillary Vistas dataset and Cityscapes [6]. Cityscapes [6] is the largest dataset focused on urban street views, which contains 5 000 finely-annotated and 20 000 coarsely-annotated images, with instance segments covering 30 object categories (19 publicly available). It is the dataset that most resembles ours in terms of scene domain. As we can see, the Mapillary Vistas dataset surpasses Cityscapes [6] in terms of density of object instances per image, with some images exhibiting over 280 instances, and in general the distribution of images covers mostly dense scenes. Cityscapes [6] covers a smaller spectrum of object densities and is slightly biased towards small-density images, with at most 20 object instances. It has to be mentioned that the statistics for Cityscapes [6] cover the finely-annotated images.

A more specific analysis is given in Fig. 5, middle, where we report the frequency of images with a certain number of *traffic regulation objects* (*i.e.* traffic signs, traffic lights, poles, pole groups, and guardrail). Also in this restricted group of typical traffic elements our dataset is in general richer in terms of object instances than Cityscapes [6], the

Downsampling factor	category level	<	macro-level	<	root-level
2	98.00		98.23		98.54
4	96.24		96.67		97.28
8	92.90		93.71		94.87
16	87.12		88.43		90.54
32	77.80		80.11		83.69
64	65.29		69.10		74.48
128	50.67		56.53		63.85

Table 1. Control experiments to estimate upper bounds for semantic segmentation results, assessed by Intersection-over-Union (IoU, in %) scores for different grouping levels.

latter exhibiting a slight bias in terms of image frequency towards images where no such elements are present.

In Fig. 5, right, we report a comparison about the distribution of images with respect to the density of traffic participants. In addition to Cityscapes [6], this comparison has been extended also to other datasets like Microsoft COCO [30] and PASCAL VOC [10], which contain street scenes despite not begin focused on them, and to KITTI [12], which addresses tasks including semantic segmentation and object detection.[5] We can see that both Cityscapes [6] and the Mapillary Vistas dataset cover a larger variety of scene complexities in terms of traffic participants compared to the other datasets, and both exhibit a larger portion of densely-populated images.

4. Semantic Segmentation

Semantic segmentation is the first task defined on the Mapillary Vistas dataset and consists in assigning a semantic category to each pixel of the image. If multiple instances from the same category, *e.g.* cars, are present in the scene, their pixels will be assigned the same label *car*.

State Of The Art. The state-of-the-art approaches for semantic segmentation are based on deep learning and benefit from large-scale semantic segmentation datasets (or combinations thereof) such as *Microsoft COCO* [30], *Pascal-Context* [39], *Cityscapes* [6], or *ADE20K* [65]. A line of successful approaches have been inspired by the Fully Convolutional Network (FCN) [37], which has shown that effective semantic segmentation networks can be obtained from state-of-the-art architectures for image classification such as *VGG* [52], *GoogleNet* [53], *ResNet* [18], *Wider ResNet* [58], *etc.*, pre-trained on *ImageNet* [48] and/or *Places2* [63], by turning fully-connected layers into convolutional layers. Some works [5, 28, 36, 50, 62] combine FCNs with Conditional Random Fields (CRFs). Others, shape the architecture to explicitly integrate global context in the pixelwise classifier, or ensemble features at multiple scales, either through CNNs [2, 35, 38, 40, 51, 57, 61] or

---

[5]No pixel-wise annotations are provided in KITTI [12] but several independent groups contributed to the annotation of about 700 frames.
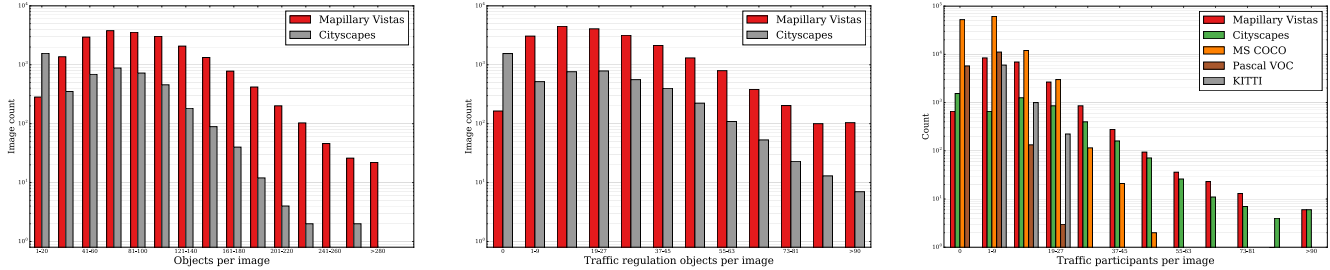
Figure 5. Comparative statistics of the frequency of images with a fixed number of: Left) annotated object instances per image (compared to Cityscapes [6] finer-grained images); Middle) annotated traffic regulation objects, *i.e.* traffic signs, traffic lights, poles, pole group and guard rail (compared to Cityscapes [6] finer-grained images); Right) traffic participants instances (compared to Cityscapes [6] fine-grained, KITTI [12], PASCAL VOC [10] and Microsoft COCO [30]).

Recurrent Neural Networks (RNNs) [4, 42]. Further works have focused on relaxing the level of supervision in FCNs by considering bounding boxes [7], scribbles [27], or image class labels [20, 21] as sources of weak supervision.

Goal and Metrics. The semantic segmentation algorithm provides a category for each pixel in a test image. The learner is trained on all non-void categories and ignores pixels belonging to *void* during inference. To assess the performance of the semantic segmentation we adopt the standard Jaccard Index, *a.k.a.* the PASCAL VOC intersection-over-union IoU metric [11], which is given by $\texttt{IoU} = \frac{\texttt{TP}}{\texttt{TP+FP+FN}}$, where $\texttt{TP}$, $\texttt{FP}$ and $\texttt{FN}$ denote true positive, false positive and false negative pixels, respectively, obtained over the test set. The IoU is computed for each category separately and the outcome is averaged, yielding $\texttt{IoU}_{\text{class}}$.

In Tab. 1, we provide some control experiments where we estimate the upper bounds for the IoU scores for given downsampling factors 2–128. We first downscale a label image by a given factor (*i.e.* simulating processing at reduced size input) and then upsample it to the original image size for evaluation of average IoU at original scale. The reported numbers are an indication of the highest obtainable scores. We provide detailed numbers for category-, macro- and root-level labels, confirming that low-resolution processing is significantly contributing to overall degradation of segmentation results.

Baseline Results. In Tab. 2 we present baseline results using a Wider Network (ResNet38) [55] architecture with cross-entropy loss as well as imbalance correction via loss max-pooling and/or alternative minibatch compilation strategies as described in [47].

Moreover, we list results from the winning submission of the 2017 Large Scale Scene Understanding (LSUN) challenge on semantic segmentation, which was based on our Mapillary Vistas dataset. The winning team PSPNET [60] built upon [61], extending the basic ResNet 101 (pretrained on ImageNet and Cityscapes, though Cityscapes contribution was negligible) architecture with the following features: i) Modifying the *res4b* module according to the *hybrid dilation convolution (HDC)* approach intro-

Method	mean IoU validation	mean IoU test
Single model & test scale		
Wider Network [55]		
uniform sampling, cross entropy loss	41.12	40.79
[55] + loss max-pooling [47]		
uniform sampling	43.78	42.98
balanced class sampling	47.72	44.84
LSUN'17 segmentation challenge winner		
PSPNet [61] (single model & scale)	49.76	–
+ HDC + aux. loss	50.28	–
+ class reweighting	51.50	–
+ Cityscapes pre-training	51.59	–
+ multi-scale test (6 scales)	53.51	–
4 model ensemble	53.85	52.41

Table 2. Semantic segmentation results for different network architectures and extensions.

duced in [54]. In this module, 4 blocks are grouped together, setting dilation rates to $1, 2, 5, 9$ for the first 3 blocks and $1, 2, 5$ for the last block. In module *res5b*, dilation rates are set to $5, 9, 17$. ii) Imbalance correction of training data by applying an inverse frequency weighting strategy, thus raising the importance of underrepresented object categories during loss computation. iii) An auxiliary loss layer was added after *res4b22* residual block, weighted by 0.4. iv) Batch normalization parameters were adapted during fine-tuning, using 16 crops of size 713×713. Finally, training images were resized to 1000 pixels at the shorter side before cropping, and random horizontal flipping and resizing between $0.5 - 2$ was applied as training data augmentation. During inference, significant improvements were obtained by applying multi-scale testing at 6 scales $(0.5, 0.75, 1.0, 1.25, 1.5, 1.75)$. Finally, a 4-model ensemble of networks exploiting all of the above yielded the final mean IoU scores of $53.85\%/52.41\%$ on validation/test data as shown in the break-down in Tab. 2[6].

5. Instance-specific Semantic Segmentation

The second task of the Mapillary Vistas dataset introduces an additional degree of difficulty to the semantic segmentation task described in Sect. 4, because it requires to

---

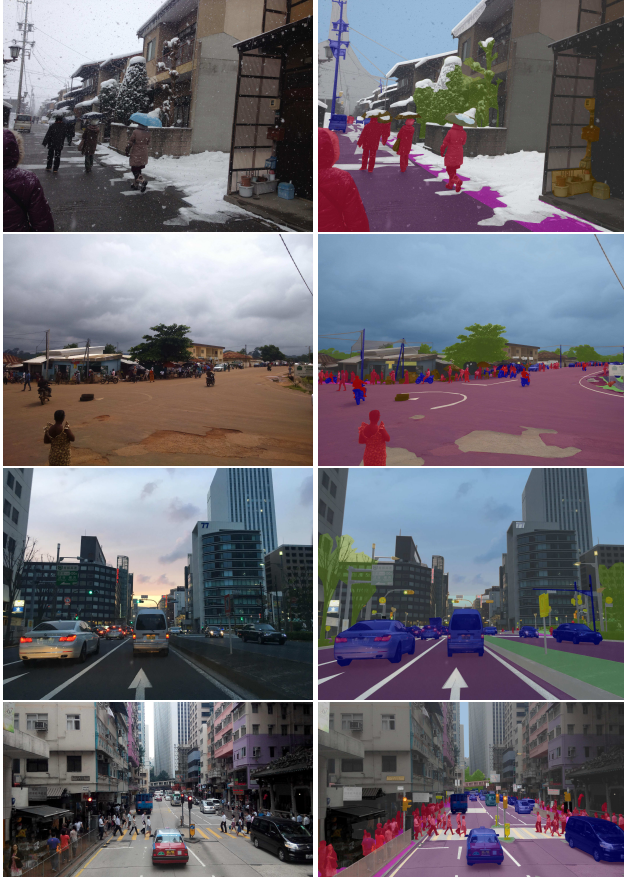[6]Additional information will be made available on `https://github.com/hszhao/PSPNet`

Figure 6. Qualitative labeling examples showing four pairs of original images with corresponding, overlaid color-coded labels. Presented examples show images from different geographical regions and with diverse lighting and weather conditions.

segment also different instances from the same semantic category. In other terms, it combines semantic segmentation and object detection at the pixel level. As a result, each pixel belonging to a category, admitting different instances (*e.g.* cars, persons, but not sky), should be assigned besides the semantic category also an instance identifier.

State Of The Art. The approaches for instance-specific semantic segmentation typically distinguish two sub-tasks, *i.e.* object segmentation and classification. In some works, the two tasks are trained separately: the segmentation task exploits segment proposal methods, which are subsequently classified with region-based methods [13, 14, 44]. Several state-of-the-art approaches fall within this category such as SDS [15], Hypercolumn [16], CFM [8], MNC [9], MultiPathNet [59] and the iterative approach in [24]. In FCIS [25], *i.e.* the winner of the Microsoft COCO [30] 2016 segmentation competition, the two tasks are trained jointly by sharing features between the two sub-tasks. Finally, Mask R-CNN [17] is an extension of Faster R-CNN [44], adding a prediction branch for instance segmentations and obtaining state-of-the-art results on Microsoft COCO [30]

at the time of preparing the final version of this document.

The works in [26, 33] try to adapt FCNs to instance-specific semantic segmentation by applying some sort of clustering to the FCNs' representations. However, those approaches are typically not end-to-end trainable, and rely on hand-crafted post-processing steps.

Goal and Metrics. An instance-specific semantic segmentation algorithm delivers a set of detections of object instances in terms of pixel-level segmentation masks, each with an associated confidence score. We evaluate the quality of instance-specific segmentations akin to [30] by computing for each class the average precision (AP) at the region level [16] for multiple overlap thresholds (from 0.5 to 0.95 with step 0.05), and average the obtained scores to avoid a bias towards a specific value. The overlap coincides with IoU computed on a single instance. Each ground-truth instance is matched to the most confident, suitable prediction, while other potential matches are regarded as false positive. In addition to the class-specific APs, we report the mean AP value obtained over the different classes.

Baseline Results. We present the winning submission of the LSUN challenge on instance-specific semantic segmentation, which was based on an extension of Mask R-CNN [17] and submitted from team *UCenter* [32]. Their approach extended the original architecture as follows: i) Using two pre-trained ResNet50 models with Feature Pyramid Network [29] structures for the region proposal network, pretrained on COCO data, using additional and smaller anchors, ii) Using two Inception ResNet 50 models, pretrained on ImageNet. The ResNet 50 models are used due to preference of larger image inputs (max. image size 1900) over deeper feature extractors. Also, all models were trained using step-based learning rate updates instead of polynomial decay. The final mask AP (mAP @ 0.5) results on validation data are $22.8\%(42.5\%)$ and $23.7\%(43.5\%)$ for single model RPN-based model and ensemble of above models, respectively. Finally, the LSUN winner's performance of the ensemble on test data is $\mathbf{24.8\%}(44.2\%)$.

6. Conclusion and Outlook

We have introduced the Mapillary Vistas Dataset - a novel, large-scale dataset of street-level images for the tasks of semantic segmentation and instance-specific semantic segmentation. Given the significantly raised interest for autonomously acting cars and robotic agents in general, we hope that our dataset can help to significantly push the state-of-the-art. The evaluation server for the Mapillary Vistas Dataset will remain live and accept submissions of results from novel algorithms, which will then be ranked in a leaderboard (visit `research.mapillary.com`).

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016. 1

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015. 6

[3] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *(ECCV)*, pages 44–57. 2008. 1, 5

[4] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with LSTM recurrent neural networks. In *(CVPR)*, pages 3547–3555, 2015. 7

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *CoRR*, abs/1606.00915, 2016. 1, 6

[6] 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 *(CVPR)*, 2016. 1, 2, 3, 5, 6, 7

[7] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *(ICCV)*, 2015. 7

[8] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015. 8

[9] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 1, 8

[10] 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. *International Journal of Computer Vision*, 111(1):98–136, 2015. 1, 5, 6, 7

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *(IJCV)*, 88(2):303–338, 2010. 7

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *(IJRR)*, 2013. 1, 6, 7

[13] R. Girshick. Fast R-CNN. In *(ICCV)*, 2015. 1, 8

[14] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *(CVPR)*, 2014. 1, 8

[15] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *(ECCV)*, pages 297–312, 2014. 4, 8

[16] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *(CVPR)*, 2015. 8

[17] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 8

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 6

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

[20] S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *(NIPS)*, 2015. 7

[21] S. Hong, J. Oh, B. Han, and H. Lee. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *(CVPR)*, 2016. 7

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

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

[24] K. Li, B. Hariharan, and J. Malik. Iterative instance segmentation. 2016. 8

[25] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. *CoRR*, abs/1611.07709, 2016. 1, 8

[26] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, and S. Yan. Proposal-free network for instance-level object segmentation. *CoRR*, abs/1509.02636, 2015. 8

[27] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *(CVPR)*, 2016. 7

[28] G. Lin, C. Shen, I. Reid, and A. van den Hengel. Efficient piecewise training of deep structured models for semantic segmentation. *CoRR*, abs/1504.01013, 2015. 1, 6

[29] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016. 8

[30] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. *CoRR*, abs/1405.0312, 2014. 1, 5, 6, 7, 8

[31] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *(PAMI)*, 33(5):978–994, 2011. 3

[32] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. LSUN2017 instance segmentation challenge winning team UCenter, July 2017. 8

[33] S. Liu, X. Qi, J. Shi, H. Zhang, and J. Jia. Multi-scale patch aggregation (MPA) for simultaneous detection and segmentation. In *(CVPR)*, pages 3141–3149, 2016. 8

[34] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *(ECCV)*, 2016. 1

[35] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *CoRR*, abs/1506.04579, 2015. 6

[36] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *(ICCV)*, 2015. 1, 6

[37] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *(CVPR)*, pages 3431–3440, 2015. 1, 6

[38] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *(CVPR)*, June 2015. 6

[39] 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)*, pages 891–898, 2014. 1, 6

[40] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. *CoRR*, abs/1505.04366, 2015. 6

[41] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *(IJCV)*, 2001. 1

[42] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *(ICML)*, pages 82–90, 2014. 7

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *(CVPR)*, June 2016. 1

[44] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *(NIPS)*, 2015. 1, 8

[45] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *(ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 1

[46] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *(CVPR)*, 2016. 1

[47] S. Rota Bulo, G. Neuhold, and P. Kontschieder. Loss max-pooling for semantic image segmentation. In *(CVPR)*, July 2017. 7

[48] 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)*, 2015. 1, 6

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

[50] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. *CoRR*, abs/1503.02351, 2015. 1, 6

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

[52] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 6

[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. 6

[54] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. W. Cottrell. Understanding convolution for semantic segmentation. *CoRR*, abs/1702.08502, 2017. 7

[55] Z. Wu, C. Shen, and A. van den Hengel. High-performance semantic segmentation using very deep fully convolutional networks. *CoRR*, abs/1604.04339, 2016. 7

[56] J. Xie, M. Kiefel, M.-T. Sun, and A. Geiger. Semantic instance annotation of street scenes by 3d to 2d label transfer. In *(CVPR)*, June 2016. 1

[57] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *Int. Conf. on Learning Representations (ICLR)*, 2016. 6

[58] S. Zagoruyko and N. Komodakis. Wide residual networks. In *(BMVC)*, 2016. 6

[59] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár. A multipath network for object detection. In *BMVC*, 2016. 8

[60] Y. Zhang, H. Zhao, and J. Shi. LSUN2017 segmentation challenge winning team PSPNet, July 2017. 7, 8

[61] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *CoRR*, abs/1612.01105, 2016. 6, 7

[62] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 6

[63] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2016. 1, 6

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

[65] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. *CoRR*, abs/1608.05442, 2016. 1, 5, 6