SkyScapes – Fine-Grained Semantic Understanding of Aerial Scenes

Seyed Majid Azimi[1] Corentin Henry[1] Lars Sommer[2] Arne Schumann[2] Eleonora Vig[1]

[1]German Aerospace Center (DLR), Wessling, Germany [2]Fraunhofer IOSB, Karlsruhe, Germany

https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-12760

Aerial image with overlaid annotation: dense (19 classes) and lane markings (12 classes); the dataset covers 5.7 km^2.

Abstract

Understanding the complex urban infrastructure with centimeter-level accuracy is essential for many applications from autonomous driving to mapping, infrastructure monitoring, and urban management. Aerial images provide valuable information over a large area instantaneously; nevertheless, no current dataset captures the complexity of aerial scenes at the level of granularity required by real-world applications. To address this, we introduce SkyScapes, an aerial image dataset with highly-accurate, fine-grained annotations for pixel-level semantic labeling. SkyScapes provides annotations for 31 semantic categories ranging from large structures, such as buildings, roads and vegetation, to fine details, such as 12 (sub-)categories of lane markings. We have defined two main tasks on this dataset: dense semantic segmentation and multi-class lane-marking prediction. We carry out extensive experiments to evaluate state-of-the-art segmentation methods on SkyScapes. Existing methods struggle to deal with the wide range of classes, object sizes, scales, and fine details present. We therefore propose a novel multi-task model, which incorporates semantic edge detection and is better tuned for feature extraction from a wide range of scales. This model achieves notable improvements over the baselines in region outlines and level of detail on both tasks.

1. Introduction

Automated methods for creating maps of today's urban and rural infrastructures with *centimeter-level (cm-level)* accuracy are of great aid in handling their growing complexity. Applications of such accurate maps include urban management, city planning, and infrastructure monitoring/maintenance. Another prominent example is the creation of *high definition (HD)* maps for autonomous driving. Applications here include the use of a general road network for navigation and more advanced automation tasks in *Advanced driver assistance systems (ADAS)*, such as lane departure warnings, which rely on precise information about lane boundaries, sidewalks, etc. [37, 40, 33, 51, 31].

Currently, the data collection process to generate HD maps is mainly carried out by so-called mobile mapping systems, which comprise of a vehicle equipped with a broad range of sensors (*e.g.*, Radar, LiDAR, cameras) followed by automated analysis of the collected data [17, 18, 5, 24]. The limited field-of-view and occlusions due to the oblique sensor angle make this automated analysis complicated. In addition, mapping large urban areas in this way requires a lot of time and resources. An aerial perspective can alleviate many of these problems and simultaneously allow for processing of much larger areas of cm-level georeferenced data in a short time. Existing aerial semantic segmentation datasets, however, are limited in the range of their annotations. They either focus on a few individual classes, such as roads or building footprints in the INRIA [30], Massachusetts [35], SpaceNet [43], or DeepGlobe [11] datasets, or they provide very coarse classes, such as the GRSS_DFC_2018 [1], or the ISPRS Vaihingen and Potsdam datasets [20]. Other datasets are recorded at sensor angles and at flight heights unsuitable for HD mapping [29, 15] or contain potentially inaccurate annotations generated automatically [44]. In addition, only few works tackle lane-marking extraction in aerial imagery, and they either rely on third-party sources such as OpenStreetMap, or only provide a binary extraction in Azimi et al. [2].

Ground imagery has greatly benefited from large-scale datasets, such as ImageNet [12], Pascal VOC [13], MS-COCO [26], but in aerial imagery the annotation is scarce and more tedious to obtain. In this work, we propose a new aerial image dataset, called SkyScapes, which closes this gap by providing detailed annotations of urban scenes for established classes, such as buildings, vegetation, and roads, as well as fine-grained classes, such as various types of lane markings, vehicle entrance/exit zones, danger areas, etc. Fig. 1 shows sample annotations offered by SkyScapes.

The dataset contains 31 classes and a rigorous annotation process was established to provide a high degree of annotation accuracy. SkyScapes uniquely combines the fine-grained annotation of road infrastructure with an overhead viewing angle and coverage of large areas, thus enabling the generation of HD maps for various applications. We evaluate several state-of-the-art semantic segmentation models as baselines on SkyScapes. Existing models achieve a significantly lower accuracy on our dataset than on established benchmarks with either ground-views or a much coarser set of classes. Our analysis of the most common errors hints at many merged regions and inaccurate boundaries. We therefore propose a novel segmentation model, which incorporates semantic edge detection as an auxiliary task. The secondary loss function emphasizes edges more strongly during the learning process, leading to a clear reduction of the prominent error cases. Furthermore, the proposed architecture takes both large- and small-scale objects into account.

In summary: i) we provide a new aerial dataset for semantic segmentation with highly accurate annotations and fine-grained classes, thus enabling the development of models for previously unsupported tasks, such as aerial HD-mapping; ii) we carry out extensive evaluations of current state-of-the-art models and show that existing approaches struggle to handle the large number of classes and level of detail in the dataset; iii) hence, we propose a new multi-task model, which combines semantic segmentation with edge detection, yielding more precise region outlines.

2. The SkyScapes Dataset

The data collection was carried out with a helicopter flying over the greater area of Munich, Germany. A low-cost camera system [23, 16] consisting of three standard DSLR cameras and mounted on a flexible platform was used for recording the data, with only the nadir-looking capturing images. In total, 16 non-overlapping RGB images of size 5616×3744 pixels were chosen. The flight altitude of about 1000 m above ground led to a *ground sampling distance (GSD)* of approximately 13 cm/pixel. The images represent urban and partly rural areas with highways, first/second order roads, and complex traffic situations, such as crossings and congestion, as exemplified in fig. 1.

2.1. Classes and Annotations

Thirty-one semantic categories were annotated: low vegetation, paved road, non-paved road, paved parking place, non-paved parking place, bike-way, sidewalk, entrance/exit, danger area, building, car, trailer, van, truck, large truck, bus, clutter, impervious surface, tree, and 12 lane-marking types. The considered lane-markings are the following: dash-line, long-line, small dash-line, turn sign, plus sign, other signs, crosswalk, stop-line, zebra zone, no parking zone, parking zone, other lane-markings. The selection of classes was influenced by their relevance to real-world applications, hence, road-like objects dominate. Class definitions and visual examples for each class are given in the supplementary materials, class statistics can be found in Fig. 2.

The SkyScapes dataset was manually annotated using tools adapted to each object class and following a strict annotation policy. Annotating aerial images requires considerable time and effort, especially when dealing with many small objects, such as lane-markings. Shadows, occlusion, and unclear object boundaries also add to the difficulty. Due to the size and shape complexity, and to the large number of classes/instances, annotation required considerably more work than for ground-view benchmarks (such as CityScapes [10]), also limiting the dataset size. To ensure high quality, the annotation process was performed iteratively with a three-level quality check over each class, overall taking about 200 man-hours per image. We show one such annotated image in Fig. 1.

In SkyScapes, we enforce pixel-accurate annotations, as even small offsets lead to large localization errors in aerial images (*e.g.*, a 1-pixel offset in SkyScapes would lead to a 13 cm error). As autonomous vehicles require a min. accuracy of 20 cm for on-map localization [52], we chose the highly accurate annotation of a smaller set of images over coarser annotations of a much larger set. In fact, in section 6, we show high generalization of our model when trained on SkyScapes and tested on third-party data.

2.2. Dataset Splits and Tasks

We split the dataset into training, validation, and test sets with 50%, 12.5%, and 37.5% portions respectively. We chose this particular split due to the class imbalance and to avoid splitting larger images. The training and validation sets will be publicly available. Test images will be released as an online benchmark with undisclosed ground-truth.

Lane-markings and the rest of the scene elements (such as buildings, roads, vegetation, and vehicles) present different challenges, with lane-markings operating on much finer scales and requiring a fine-grained differentiation,

Figure 1: SkyScapes image with overlaid annotation and zoomed-in samples (×2: solid line, ×4: dashed line). Top to bottom: RGB, dense annotation (20 classes), lane markings annotation (12 classes), multi-class edges. Class colors as in fig. 2.

whereas other scene elements are represented on a much wider scale. Having considered these challenges, we defined five different tasks: 1) **SkyScapes-Dense** with 20 classes as the lane-markings were merged into a single class, 2) **SkyScapes-Lane** with 13 classes comprising 12 lane-marking classes and a non-lane-marking one, 3) **SkyScapes-Dense-Category** with 11 merged classes comprising nature (low-vegetation, tree), driving-area (paved, non-paved), parking-area (paved, non-paved), human-area (bikeway, sidewalk, danger area), shared human and vehicle area (entrance/exit), road-feature (lane-marking), residential area (building), dynamic-vehicle (car, van, truck, large-truck, bus), static-vehicle (trailer), man-made surface (impervious surface), and others objects (clutter), 4) **SkyScapes-Dense-Edge-Binary**, and 5) **SkyScapes-Dense-Edge-Multi**. The two latter tasks are binary and multi-class edge detection, respectively. Defining separate tasks allows for more fine-grained control to fit the model to the dense object regions, their boundaries, and their classes. This is especially helpful when object boundary accuracy is paramount and difficult to extract, *e.g.*, for multi-class lane-markings.

2.3. Statistical Properties

SkyScapes is comprised of more than 70K annotated instances that are divided into 31 classes. The number of annotated pixels and instances per class for SkyScapes-Dense and SkyScapes-Lane are given in fig. 2. The majority of pixels are annotated as low vegetation, tree, or building, whereas the most common classes are lane markings, tree, low vegetation, and car. This illustrates the wide range from classes with fewer large regions to those with many small regions. A similar range can be observed among the lane markings within the more fine-grained SkyScapes-Lane task. With an average pixel area of about 9 pixels,

(a) SkyScapes-Dense

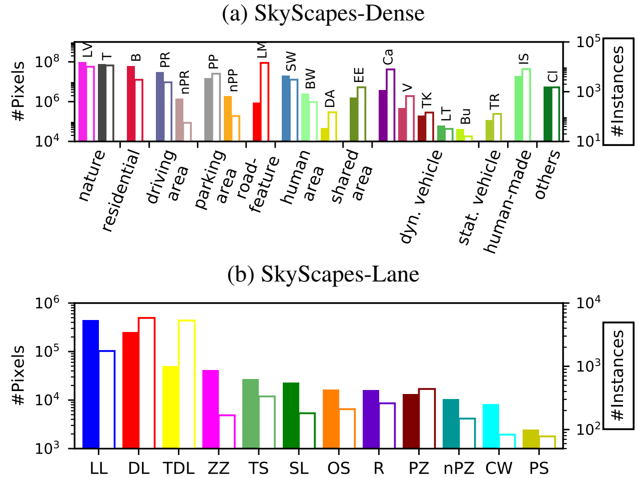


Figure 2: Number of annotated pixels (filled) and instances (non-filled) per class in SkyScapes-Dense and SkyScapes-Lane for low-vegetation (LV), tree (T), building (B), paved-road (PR), paved-parking-place (PP), non-paved-parking-place (nPP), non-paved-road (nPR), lane-marking (LM), sidewalk (SW), bikeway (BW), danger-area (DA), entrance-exit (EE), car (Ca), van (V), truck (TK), trailer (TR), long-truck (LT), bus (Bu), impervious-surface (IS), clutter (Cl), long line (LL), dash line (DL), tiny dash line (TDL), zebra zone (ZZ), turn sign (TS), stop line (SL), other signs (OS), the rest of lane-markings (R), parking zone (PZ), no parking zone (nPZ), crosswalk (CW), and plus sign (PS).

'tiny dash lines' are the smallest instances.

A quantitative comparison of SkyScapes against existing aerial segmentation datasets is provided in table 1. Existing datasets lack the high detail level and annotation quality of SkyScapes. Potsdam contains fewer classes (6 vs 31), less accurate labels, and image distortions due to suboptimal orthorectification. TorontoCity focuses on quantity: its wider spatial coverage requires (a less precise) automated labeling. SkyScapes offers the largest number of classes including various fine-structures (*e.g.*, lane markings). In absolute terms, SkyScapes contains also notably more region instances, which emphasizes the higher complexity of SkyScapes. Handling this range of classes and variety of object instance sizes is one of the main challenges. The capability of state-of-the-art segmentation methods to address these challenges has not yet been thoroughly explored.

3. Semantic Benchmarks

In the following, we review several state-of-the-art segmentation methods and benchmark these on SkyScapes.

3.1. Metrics

To assess the segmentation performance, we use the Jaccard Index, known as the PASCAL VOC *Intersection over Union (IoU)* metric: $\frac{TP}{TP+FP+FN}$ [13], where TP, FP, and FN stand for the numbers of true positive, false positive, and false negative pixels for each class, determined over the test set. We also report other metrics, such as frequency weighted IoU, pixel accuracy, average recall/precision, and mean IoU, *i.e.*, the average of IoUs over all classes as defined in [28]. In the supplementary material, we report IoU_{class} for SkyScapes-Dense and $IoU_{category}$ for the best baseline on SkyScapes-Dense-Category. Unlike in the street scenes of CityScapes [10], in aerial scenes the objects can be as long as the image size (roads or long-line lane-markings). Therefore, we do not report $IoU_{instance}$.

3.2. State of the Art in Semantic Segmentation

As detection results have matured, reaching around 80% mean AP on Pascal VOC [22] and on the DOTA aerial object detection dataset [45, 3], the interest has shifted to pixel-level segmentation, which yields a more detailed localization of an object and handles occlusion better than bounding boxes. In recent years, *fully-convolutional neural networks (FCNs)* [28, 41] achieved remarkable performance on several semantic segmentation benchmarks. Current state-of-the-art methods include Auto-Deeplab [27], DenseASPP [46], BiSeNet [47], Context-Encoding [49], and OcNet [48]. While specific architecture choices offer a good baseline performance, the integration of a multi-scale context aggregation module is key to competitive performance. Indeed, context information is crucial in pixel labeling tasks. It is best leveraged by so-called "pyramid pooling modules", using either stacks of input images at different scales, as in PSPNet [50], or stacks of convolutional layers with different dilation rates, as in DeepLab [6]. However, context aggregation is often performed at the expense of fine-grained details. As a remedy, FRRN [38] implements an architecture comprising a full-resolution stream for segmenting the details and a separate pooling stream for analyzing the context. Similarly, GridNet [14] uses multiple interconnected streams working at several resolutions. For our benchmark, in addition to the aforementioned models, we train several other popular segmentation networks: FCN [28], U-Net [39], MobileNet [19], SegNet [4], RefineNet [25], Deeplabv3+ [9], AdapNet [42], and FC-DenseNet [21], as well as a custom U-Net-like MobileNet and custom Decoder-Encoder with skip-connections.

In tables 2 and 4, we report our benchmarking results for the above methods. As anticipated, all methods struggle on SkyScapes due to the significant differences between ground and aerial imagery exposed in the introduction. On the SkyScapes-Dense task (table 2), classification mistakes are for the most part found around the inter-class boundaries. We observe the same inter-class misclassification on the SkyScapes-Lane task (table 4), and furthermore notice that many lane-markings are entirely missed and classified as background, certainly due to their few-pixel size. Both tasks hence represent a new type of challenge. This is reinforced by the fact that the performance of the networks remained consistent from one task to the other, showing that none are specialized enough to obtain a significant advantage on either task. In our method, we tackled this challenge by focusing on object boundaries.

4. Method

Thirty-one highly similar classes and small complex objects in SkyScapes necessitate a specialized architecture that unifies latest architectural improvements (FC-DenseNet [21], auxiliary tasks, etc.) and proves more effective than the state of the art. Motivated by the major errors from our benchmarking analysis, we propose a multi-task method that tackles both dense prediction and edge detection to improve performance on boundary regions. In the case of multi-class lane-markings, we modify the method to enable both multi-class and binary lane-marking segmentation to decrease the number of false positives in non-lane areas. We consider FC-DenseNet [21] as the main baseline. SkyScapesNet, illustrated in fig. 3, can be seen as a modified case of FC-DenseNet, but more generally as a multi-task ensemble-model network, encapsulating units from [21, 38, 7, 36]. Thus, it also shares their advantages, such as alleviating the gradient-vanishing problem. Figure 4 illustrates the building blocks, which are explained below.

FDB: in *fully dense block (FDB)*, we use more residual connections compared to the existing Dense Blocks (DBs)

Table 1: Statistics of SkyScapes and other aerial datasets. To date, TorontoCity is not publicly available.

	SkyScapes	Potsdam [20]	Vaihingen [20]	Aerial KITTI [32]	TorontoCity [44]
Classes	31	6	6	4	2+8
Images	16	38	33	20	N/A
Image dimension (px)	5616×3744	6000×6000	2493×2063 (avg)	variable	N/A
GSD (cm/pixel)	13	5	9	9	10
Aerial coverage (km^2)	5.69 (urban&rural)	3.42	1.36	3.23	712
Instances	70,346	42,389	10,700	2,814	N/A

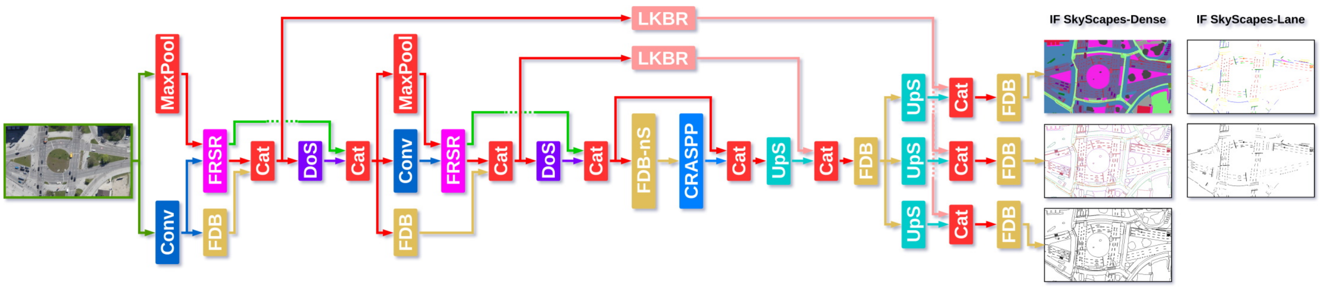


Figure 3: The architecture of SkyScapesNet. Three branches are used to predict dense semantics and multi-class/binary edges. For multi-class lane-marking prediction, two branches are used to predict multi-class and binary lane-markings.

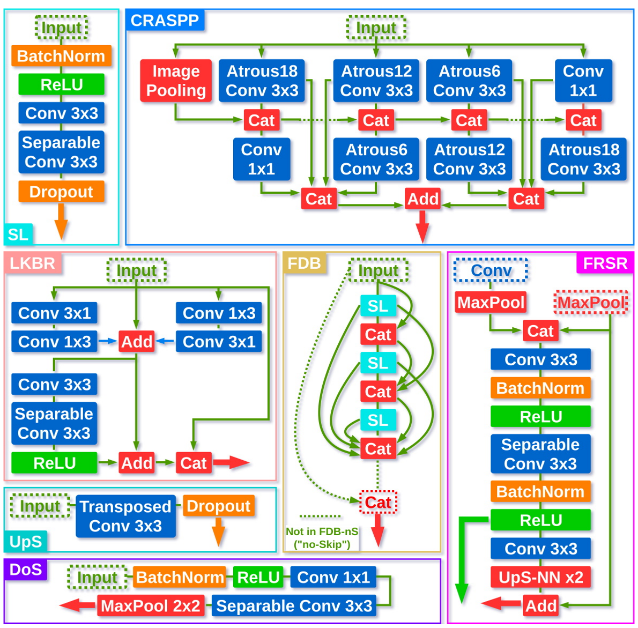


Figure 4: Configuration of SkyScapesNet building blocks. SL, DoS, and UpS are Separable, Downsampling, and Upsampling blocks, UpS-NN is a Nearest-Neighbor Upsampling layer. Add/Cat are addition/concatenation operators.

in the baseline, as inspired by DenseASPP [46]. However, instead of using atrous convolutions, we add separable-convolutions due to their recent success [7]. Moreover, as SkyScapes contains large scale variation, making receptive fields larger by using larger atrous rates deteriorates the feature extraction from very small objects such as lane-markings. The number of sub-blocks, referred to as Separable Layer (**SL**), is the same as in the DBs from the baseline.

FRSR: inspired by [38] and the comparable performance of this model with DenseNet, we add a residual-pooling stream (similar to the full-resolution residual unit – FRRU from [38]) as *full-resolution separable residual (FRSR)* unit to the main stream. Similar to FDB, we utilize separable convolutions. As the original FRRU, FRSR has two processing streams: a residual stream (for better localization) and a pooling stream (for better recognition). Inside the pooling stream, the downsampled results go through several depth-wise separable convolutions, batch-normalization, and ReLU layers and, after applying a 1×1 convolution, the output is upsampled and added to FDB. We limit the number of downsamplings in FRSR to one as the main stream applies consecutive downsampling.

CRASPP: inspired by the success of *atrous spatial pyramid pooling block (ASPP)* [46, 9], after five downsampling steps, we add the *concatenated reverse ASPP (CRASPP)* to enhance the feature extraction of large objects. In CRASPP, we 'reverse' the original ASPP (*i.e.*, the order of atrous rates) and concatenate it with the original ASPP, so as to obtain receptive fields optimal for both small/large objects.

LKBR: for boundary refinement and to improve the extraction of tiny objects, we apply – in addition to five skip-connections – *large-kernels with boundary refinements (LKBRs)*. LKBR [36] is composed of two streams including a boundary refinement module. Unlike [21], we apply a residual path from the output of the last downsampling module to the input of the first upsampling module.

Multi-task learning: we use three separate branches to predict dense semantics and multi-class and binary edges simultaneously. The streams are separated from each other after the second upsampling layer. The motivation is to allow the auxiliary tasks to modify the shared weights so as to augment the network performance on boundary regions. For multi-class lane-marking segmentation, we consider two streams with similar configuration.

Loss functions: instead of relying only on cross-entropy, we propose to add either the Soft-IoU-loss [31] or the Soft-Dice-loss [34] to it (taking the sum of indiv. losses).

By the direct application of the cost-aware cross-entropy loss, the network tries to fill in lane-marking areas which leads to a high TP rate for the lane-marking classes, but also high FP for the non-lane class. However, due to the very high number of non-lane pixels, the resulting FP does not have much effect on the overall accuracy. To alleviate this, we propose the scheduled weighting mechanism in which the costs of corresponding classes gradually move towards the final weighted coefficients as the training process evolves. Further details about the architecture as well as loss formulas are included in the supplementary material.

5. Evaluation

For our experiments, we crop the images into 512×512 patches, as the original 21 MP images would not fit into GPUs. As data augmentation, we carry out horizontal and vertical flipping, and use 50% overlap between neighboring crops both in vertical and horizontal directions. During inference we use 10% overlap as a partial solution to the lower performance at image boundaries. We use Titan XP and Quadro P6000 GPUs for training. The learning rate was 0.0001 and a batch size of 1 was chosen. We trained the algorithms for 60 epochs to make the comparison fair (the majority of the methods converged at this step). In total, there are 8820 training images. Our model has 137 M parameters. As we deal with offline mapping, inference at 355 ms per 512×512 image patch is of little concern.

SkyScapes-Dense – 20 main classes: The benchmarking results reported in table 2 demonstrate the complexity of the task. Our method described above achieves 1.93% mIoU improvement over the best benchmark. Qualitative examples of the best baselines and our proposed algorithm are depicted in fig. 5. Our algorithm exhibits the best trade-off between accurately segmented coarse and fine structures. Ablation studies in table 3 quantifying the effect of several components show that the main improvement is achieved by including both binary and multi-class edge detection.

SkyScapes-Lane – multi-class lane prediction: Here, a further challenge is the highly imbalanced dataset. Results in table 4 show that despite the tiny object sizes, our algorithm achieves 51.93% mIoU, outperforming the state of the art by 3.06%. Qualitative examples in fig. 6 highlight that our algorithm generates fewer decomposed segments.

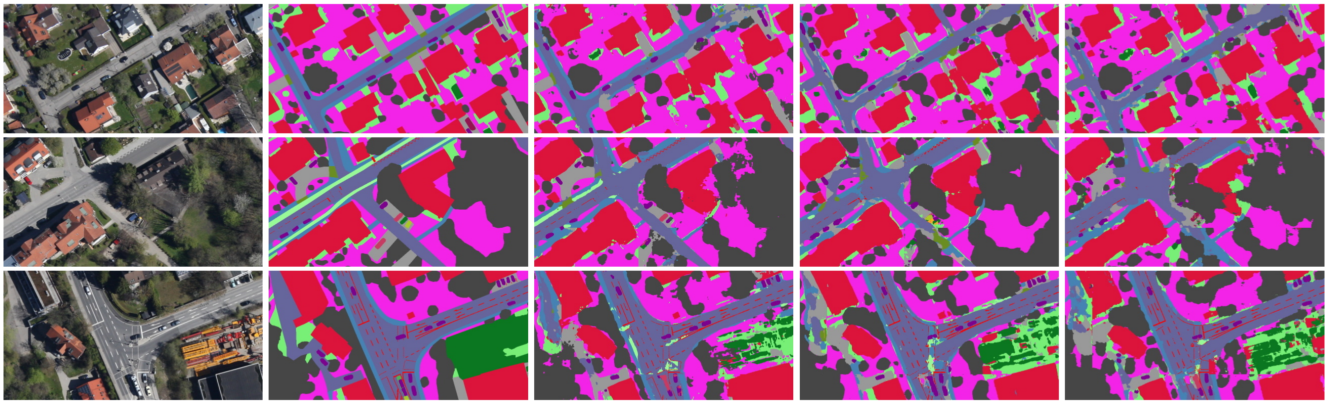
Table 2: Benchmark of the state of the art on the SkyScapes-Dense task over all 20 classes; '-' means no specific backbone; 'f.w.' is frequency weighted IoU; * skip connections.

Method	Base	IoU [%] mean	IoU [%] f.w.	average [%] recall	average [%] prec.
FCN-8s [28]	ResNet50	33.06	67.02	40.78	**<u>65.01</u>**
SegNet [4]	–	23.14	61.32	29.21	59.56
U-Net [39]	–	14.15	36.33	21.88	22.87
BiSeNet [47]	ResNet50	30.82	59.62	40.25	49.42
DenseASPP [46]	ResNet101	24.73	56.58	32.21	40.82
Encoder-Decoder*	–	37.16	67.18	**<u>48.26</u>**	50.16
FC-DenseNet-103 [21]	–	37.78	67.44	46.66	53.89
FRRNA [38]	–	37.20	65.10	46.44	53.22
GCN [36]	ResNet152	32.92	65.12	41.60	49.65
Mobile-U-Net*	–	34.96	65.26	44.52	49.49
PSPNet [50]	ResNet101	30.44	61.62	40.48	43.63
RefineNet [25]	ResNet152	36.39	65.52	46.12	52.17
DeepLabv3+ [7]	Xception65	**38.20**	**68.81**	47.97	55.34
SkyScapesNet	–	**<u>40.13</u>**	**<u>72.67</u>**	**47.85**	**65.93**

Table 3: Evaluation of different parts of SkyScapesNet. 'Baseline' was trained only with cross-entropy (*i.e.*, no IoU loss added). Max stride is 32 pixels. * using original number of sub-sampling as in the baseline in SkyScapesNet.

Network	loss IoU	sep. branch.	FDB	FSRRB	CRASPP	LKBR	mIoU [%]
Baseline* [21]							37.78
Baseline							36.88
SkyScapesNet	✓						37.08
SkyScapesNet	✓	✓					38.55
SkyScapesNet	✓	✓	✓				38.77
SkyScapesNet	✓	✓	✓	✓			38.90
SkyScapesNet	✓	✓	✓	✓	✓		39.09
SkyScapesNet	✓	✓	✓	✓	✓	✓	39.30
SkyScapesNet*	✓	✓	✓	✓	✓	✓	**40.13**

SkyScapes-Dense – auxiliary tasks: We further provide results for the three auxiliary tasks **SkyScapes-Dense-Category**, **SkyScapes-Dense-Edge-Binary**, and **SkyScapes-Dense-Edge-Multi** in table 5 (cf. sec. 2.2 for task definitions). As multiple categories are merged into a single category, *e.g.*, low vegetation and tree into nature, the mIoU for SkyScapes-Dense-Category is notably higher than for the more challenging SkyScapes-Dense. For the edge detection branches, used to enforce the learning of more accurate boundaries, high mIoU is obtained for SkyScapes-Dense-Edge-Binary, while still a low one for the more challenging multi-class edge detection.

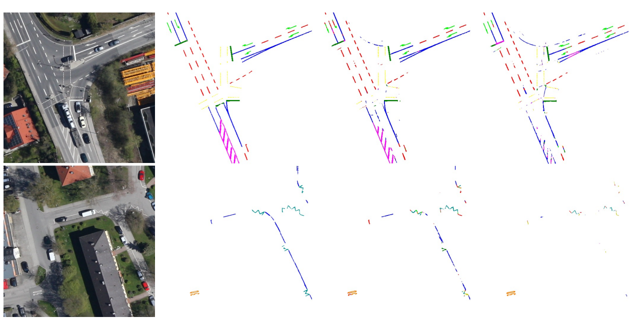


(a) RGB Image (b) Ground Truth (c) SkyScapesNet (d) DeepLabv3+ (e) FC-DenseNet103

Figure 5: Result samples for SkyScapes-Dense task by SkyScapesNet and the two best baselines. For class colors, cf. fig. 2.

Table 4: Benchmark of the state of the art on the SkyScapes-Lane task over all 13 classes. Cf. table 2 for abbreviations.

Method	Base	IoU [%]		average [%]	
		mean	f.w.	recall	precision
FCN-8s [28]	ResNet50	13.74	99.69	15.23	77.96
U-Net [39]	–	8.97	99.62	12.73	**88.26**
AdapNet [42]	–	20.20	99.67	22.21	53.60
BiSeNet [47]	ResNet50	23.77	99.66	28.71	51.42
DeepLabv3 [8]	Res50	16.15	99.62	18.94	55.44
DenseASPP [46]	ResNet101	17.00	99.65	18.74	46.02
FC-DenseNet-103 [21]	–	48.42	99.85	**55.32**	69.01
FRRN-B [38]	–	47.02	99.85	54.72	66.19
GCN [36]	Res50	35.65	99.82	43.09	55.65
Mobile-U-Net*	–	41.21	99.84	47.48	64.60
PSPNet [50]	Res101	35.85	99.82	42.64	58.23
DeepLabv3+ [7]	Xception65	37.14	99.77	43.14	62.07
Encoder-Decoder*	–	**48.87**	**99.85**	55.31	70.63
SkyScapesNet	–	**51.93**	**99.87**	**60.53**	**72.29**



(a) Image (b) GT (c) Ours (d) Enc-Dec*

Figure 6: Result samples for the SkyScapes-Lane task by SkyScapesNet and the best baseline. Class colors: cf. fig. 2.

6. Generalization

Our aim in this paper is to promote aerial imagery (in its widest sense) as a means to create HD-maps. Hence,

Table 5: Results on SkyScapes-Dense-Category, multi-class edge, and binary edge prediction tasks.

Method	Task	IoU [%]		average [%]	
		mean	f.w.	recall	prec.
SkyScapesNet	Category	52.27	77.77	63.49	65.65
SkyScapesNet	Multi-class Edge	13.00	88.74	16.82	22.74
SkyScapesNet	Binary Edge	58.72	89.52	64.81	71.99

Table 6: Generalization of our model trained on SkyScapes-Dense and evaluated on Potsdam and DFC2018.

training data	test data	IoU [%]		average [%]	
		mean	f.w.	recall	prec.
SkyScapes	Potsdam	47.46	70.58	62.28	66.09
SkyScapes	Data Fusion Contest 2018	26.42	47.58	55.67	37.64

our method is not restricted to aerial images captured by a helicopter, but would work for satellites and lower-flying drones, too. To demonstrate the good generalization capability of our method, here we show results on four additional data types covering a wide range of sensors (camera and platform), spatial resolutions, and geographic locations.

For quantitative evaluation we consider the Potsdam [20] and GRSS_DFC_2018 datasets [1], and show qualitative results also on an aerial images of Perth, Australia. Qualitative results can be seen in figs. 7 to 9. By adjusting the GSD of the test images (through scaling) to match that of our dataset, our model trained on SkyScapes indicates good generalization even without fine-tuning. This is demonstrated also in the quantitative results on Potsdam (see table 6) as the mean IoU is in the range of SkyScapes-Dense-Category. For the quantitative evaluation, we merged our categories according to the Potsdam categories.

Moreover, fig. 10 demonstrates the generalization capability of our algorithm for binary lane-marking extraction at a widely different scale (30 cm/pixel) on a WorldView-4

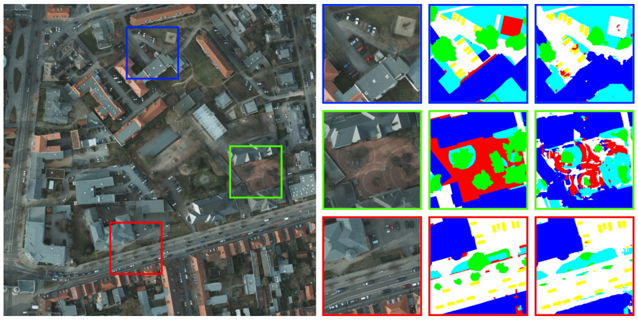


Figure 7: Results of our model trained on SkyScapes and tested on the Potsdam dataset with GSD adjustment and no fine-tuning. Patches from left to right: RGB, ground truth, prediction. Potsdam classes: □ impervious, ■ building, ■ low vegetation, ■ tree, ■ car, ■ clutter.

Figure 8: Results of our model trained on SkyScapes and tested on the GRSS_DFC_2018 dataset (over Houston, USA) with GSD adjustment and without fine-tuning.

satellite image. To the best of our knowledge, satellite images have not been used for lane-marking extraction before.

7. Conclusion

In this paper, we introduced SkyScapes, an image dataset for cm-level semantic labeling of aerial scenes to facilitate the creation of HD maps for autonomous driving, urban management, city planning, and infrastructure monitoring. We presented an extensive evaluation of several state-of-the-art methods on SkyScapes and proposed a novel multi-task network that, thanks to its specialized architecture and auxiliary tasks, proves more effective than all tested baselines. Finally, we demonstrated good generalization of our method on four additional image types ranging from high-resolution aerial images to even satellite images.

Figure 9: Segmentation result samples of our model trained on SkyScapes and tested on an aerial image over Perth, Australia, with GSD adjustment and without fine-tuning.

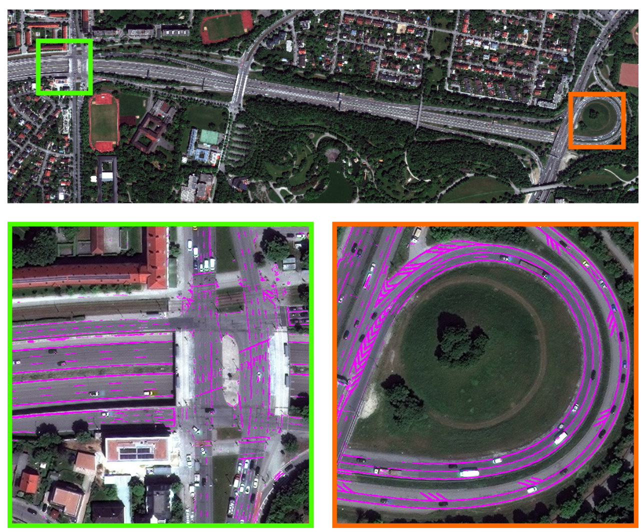


Figure 10: Binary lane segmentation on a Worldview4 satellite image over Munich using our model trained on SkyScapes, and tested on a highway scene with GSD adjustment and no fine-tuning.

Acknowledgements We thank (1) Spookfish/EagleView for the aerial image over Perth; (2) the National Center for Airborne Laser Mapping and the Hyperspectral Image Analysis Lab at the Univ. Houston for acquiring and providing the GRSS_DFC_2018 data in the generalization study, the IEEE GRSS Image Analysis and Data Fusion Technical Committee; (3) Ternow A.I. GmbH for labeling process assistance. E. Vig was funded by a Helmholtz Young Investigators Group grant (VH-NG-1311).

References

[1] 2018 IEEE GRSS. Data Fusion Contest. http://www.grss-ieee.org/community/technical-committees/data-fusion/2018-ieee-grss-data-fusion-contest/. [Online; accessed 22-March-2019]. 1, 7

[2] Seyed Majid Azimi, Peter Fischer, Marco Körner, and Peter Reinartz. Aerial LaneNet: lane marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. *arXiv preprint arXiv:1803.06904*, 2018. 2

[3] Seyed Majid Azimi, Eleonora Vig, Reza Bahmanyar, Marco Körner, and Peter Reinartz. Towards multi-class object detection in unconstrained remote sensing imagery. In *Proceedings of the Asian Conference of Computer Vision (ACCV)*, 2018. 4

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 4, 6

[5] Raphael V Carneiro, Rafael C Nascimento, Rânik Guidolini, Vinicius B Cardoso, Thiago Oliveira-Santos, Claudine Badue, and Alberto F De Souza. Mapping road lanes using laser remission and deep neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018. 1

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic Image Segmentation With Deep Convolutional Nets And Fully Connected CRFs. *arXiv preprint arXiv:1412.7062*, 2014. 4

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 4, 5, 6, 7

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 4, 5

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 2, 4

[11] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raska. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209. IEEE, 2018. 1

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 4

[14] Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Alain Trémeau, and Christian Wolf. Residual conv-deconv grid network for semantic segmentation. In *Proceedings of the British Machine Vision Conference, 2017*, 2017. 4

[15] ICGV TU Graz. Semantic Drone Dataset. http://dronedataset.icg.tugraz.at/. [Online; accessed 01-March-2019]. 1

[16] Veronika Gstaiger, Hannes Römer, Dominik Rosenbaum, and Fabian Henkel. Airborne camera system for real-time applications-support of a national civil protection exercise. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(7):1189, 2015. 2

[17] Chunzhao Guo, Kiyosumi Kidono, Junichi Meguro, Yoshiko Kojima, Masaru Ogawa, and Takashi Naito. A low-cost solution for automatic lane-level map generation using conventional in-car sensors. *IEEE Transactions on Intelligent Transportation Systems*, 17(8):2355–2366, 2016. 1

[18] Gi-Poong Gwon, Woo-Sol Hur, Seong-Woo Kim, and Seung-Woo Seo. Generation of a precise and efficient lane-level road map for intelligent vehicle systems. *IEEE Transactions on Vehicular Technology*, 66(6):4517–4533, 2017. 1

[19] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 4

[20] ISPRS. 2D Semantic Labeling Dataset. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html. [Online; accessed 01-March-2019]. 1, 5, 7

[21] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017. 4, 5, 6, 7

[22] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. Parallel feature pyramid network for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018. 4

[23] Franz Kurz, Dominik Rosenbaum, Jens Leitloff, Oliver Meynberg, and Peter Reinartz. Real time camera system for disaster and traffic monitoring. In *International Confer-*

ence on Sensors and Models in Photogrammetry and Remote Sensing, 2011. 2

[24] Pierre Lamon, Cyrill Stachniss, Rudolph Triebel, Patrick Pfaff, Christian Plagemann, Giorgio Grisetti, Sascha Kolski, Wolfram Burgard, and Roland Siegwart. Mapping with an autonomous car. In *IEEE/RSJ IROS Workshop: Safe Navigation in Open and Dynamic Environments*, volume 26, 2006. 1

[25] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4, 6

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[27] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L. Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. *arXiv preprint arXiv:1901.02985*, 2019. 4

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4, 6, 7

[29] Ye Lyu, George Vosselman, Guisong Xia, Alper Yilmaz, and Michael Ying Yang. The uavid dataset for video semantic segmentation. *arXiv preprint arXiv:1810.10438*, 2018. 1

[30] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017. 1

[31] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3438–3446, 2017. 1, 6

[32] Gellért Máttyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Enhancing road maps by parsing aerial images around the world. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2015. 5

[33] Gellért Máttyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. HD Maps: Fine-Grained Road Segmentation by Parsing Ground and Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3611–3619, 2016. 1

[34] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *4th Inter. Conf. on 3D Vision*, 2016. 6

[35] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013. 1

[36] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017. 4, 5, 6, 7

[37] Fabian Poggenhans, Jan-Hendrik Pauls, Johannes Janosovits, Stefan Orf, Maximilian Naumann, Florian Kuhnt, and Matthias Mayr. Lanelet2: A high-definition map framework for the future of automated driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1672–1679. IEEE, 2018. 1

[38] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3309–3318, July 2017. 4, 5, 6, 7

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, Cham, 2015. Springer International Publishing. 4, 6, 7

[40] Heiko G Seif and Xiaolong Hu. Autonomous driving in the iCityHD maps as a key challenge of the automotive industry. *Engineering*, 2(2):159–162, 2016. 1

[41] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 4

[42] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4644–4651, May 2017. 4, 7

[43] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 1

[44] Shenlong Wang, Min Bai, Gellért Máttyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. TorontoCity: Seeing the World with a Million Eyes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 5

[45] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 4

[46] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang Deepmotion. DenseASPP for Semantic Segmentation in Street Scenes. In *CVPR*, pages 3684–3692, Salt Lake City, 2018. 4, 5, 6, 7

[47] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–341, September 2018. 4, 6, 7

[48] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 4

[49] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *CVPR*, Honolulu, 2017. 4, 6, 7

[51] Ling Zheng, Bijun Li, Hongjuan Zhang, Yunxiao Shan, and Jian Zhou. A high-definition road-network model for self-driving vehicles. *ISPRS International Journal of Geo-Information*, 7(11):417, 2018. 1

[52] Julius Ziegler, Philipp Bender, Markus Schreiber, Henning Lategahn, Tobias Strau, Christoph Stiller, Thao Dang, Uwe Franke, Nils Appenrodt, Christoph Keller, Eberhard Kaus, Ralf Herrtwich, Clemens Rabe, David Pfeiffer, Frank Lindner, Fridtjof Stein, Friedrich Erbs, Markus Enzweiler, Carsten Knoeppel, and Eberhard Zeeb. Making Bertha Drive – An Autonomous Journey on a Historic Route. *IEEE Intell. Transp. Syst.*, 2014. 2