SkyScapes – Fine-Grained Semantic Understanding of Aerial Scenes – Supplementary Material –

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

1. Annotation techniques

Several annotators worked on the creation of the ground truth, each focusing on a separate set of classes. To ensure annotation consistency, a list of rules was established and extended as special cases were discovered. These guidelines relate to two aspects of the annotation work: target identification and boundary topology. For the former, the annotators referred to the comprehensive class definitions found in section 2 to assign every object in the image to a semantic category. The vertical ordering of classes (or class overlays) was based on the natural physical ordering found in the real world, and as also considered in transportation systems, *i.e.*, vehicles were put on top of all road-like objects, etc. Some classes were annotated together to ensure that inter-object borders were not overlapping, but only after fixing the vertical class order, similarly to CityScapes [2]: the object boundaries of low-level classes were drawn more coarsely at places where they would be overlaid with the accurate masks of higher-level classes. This sped up the annotation process while still satisfying our quality requirements. Other objects such as vehicles were annotated separately. As a consequence, their borders did not necessarily match the boundaries of other classes in the resulting merged ground truth. In the final verification step, these seams were corrected pixel by pixel by the annotators.

2. Semantic classes

In table 10, we provide detailed definitions of the 31 annotated classes, including a typical visual example per class.

3. Further details on SkyScapesNet

In SkyScapesNet, we use the same number of pooling and unpooling steps as in the FC-DenseNet [3] baseline, *i.e.*, 5 pooling and 5 unpooling steps. Between the encoder and decoder we use an extra *fully dense block (FDB)* module similar to the DenseBlock (DB) module in the baseline together with *concatenated reverse ASPP (CRASPP)*. The number of Separable Layers (SL) is similar to the baseline: 4, 5, 7, 10, 12, 15, 12, 10, 7, 5, 4. However, for the majority of the ablation studies we used the SL sequence 1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1 due to limited GPU memory in Titan XPs. The experiments marked with '*' in the ablation study table were carried out with the same number of SL modules as in the baseline.

We use HeUniform to initialize our model and train it with ADAM using a constant learning rate of 0.0001. We

Table 1. Architecture details of SkyScapesNet. The abbreviations stand for: FDB: Fully DenseBlock, DoS: Down-sampling, UpS: Up-sampling, SL: separable layer, and fm: number of feature maps. Note that skip-connections and LKBR modules have not been illustrated for simplicity.

Network Architecture
Input, fm=3
Convolution (3x3), fm:48
FDB (4 SLs), MaxPool→FRSR
Concatenation→DoS→Concatenation
FDB (5 SLs), Conv(3x3) + MaxPool→FRSR
Concatenation→DoS→Concatenation
FDB (7 SLs), Conv(3x3) + MaxPool→FRSR
Concatenation→DoS→Concatenation
FDB (10 SLs), Conv(3x3) + MaxPool→FRSR
Concatenation→DoS→Concatenation
FDB (12 SLs), Conv(3x3) + MaxPool→FRSR
Concatenation→DoS→Concatenation
FDB (15 SLs)
CRASPP repeated in parallel for each task UpS + FDB (12 SLs)
UpS + FDB (10 SLs)
UpS + FDB (7 SLs)
UpS + FDB (5 SLs)
UpS + FDB (4 SLs)
Convolution (1x1), fm=No. of classes
Softmax

did not use any learning rate scheduler for the sake of fair benchmarking of several architectures. We train all models on the augmented data with horizontal and vertical flips. We use current batch statistics for batch normalization in all three phases: training, validation, and test. The number of features in SL modules is the multiplication of the number of SL modules and the growth-rate. We used the same growth-rate of 16 as the baseline. The number of feature maps in separable-convolutions is the same as in the standard convolution layers. We use a stride of 1 in separable convolutions. MaxPooling is done with a kernel size of 2×2 with a stride of 2. For convolutions, we use a kernel size of 3×3 throughout the network. In the *full-resolution separable residual (FRSR)* module, the number of feature maps in the first convolution and in the separable convolution is twice as many as the number of feature maps in FDB at the same step. The last convolution has equal number of feature maps as the corresponding FDB.

The input convolution of the FRSR modules (except the first one) is 1×1 and the number of feature maps is equal to $growth\ rate * number\ of\ SL\ modules$. We use 21 feature maps in the *large-kernels with boundary refinements (LKBRs)* modules.

In our experiments, we combine the Soft-IoU loss [4] as well as the Soft-Dice loss [5] with the cross-entropy loss function. For the multi-class segmentation task, cross-entropy is defined as

$$L_{cross-entropy} = -\frac{1}{C}\sum_{c=1}^{C}\sum_{N} y_{nc}\log\hat{y}_{nc} \quad (1)$$

where $y_{nc} \in \{0, 19\}$ is the ground-truth value for class c at location n, $\hat{y}_{nc} \in [0, 19]$ is the prediction probability, C stands for the total number of classes, N is the total number of pixel locations and L stands for the loss function. The Soft-IoU loss is computed as:

$$\mathrm{L}_{soft-IOU} = -\frac{1}{C}\sum_{c=1}^{C}\frac{\sum_N y_{nc}*\hat{y}_{nc}}{\sum_N y_{nc}+\hat{y}_{nc}-y_{nc}*\hat{y}_{nc}} \quad (2)$$

The total loss is then defined as

$$\mathrm{L}_{total} = L_{soft-IOU} + L_{cross-entropy} \quad (3)$$

When the Soft-Dice loss is used, we compute the following:

$$\mathrm{L}_{soft-Dice} = -\frac{1}{C}\sum_{c=1}^{C}\frac{2*|\sum_N y_{nc}*\hat{y}_{nc}|}{|\sum_N y_{nc}|^2+|\sum_N \hat{y}_{nc}|^2} \quad (4)$$

In table 2, we evaluate the above losses on SkyScapes-Dense, both separately and in combination, and show that the combination of soft-IoU loss with cross-entropy is more beneficial than soft-Dice with cross-entropy.

Table 2. Evaluation of the different losses and their combinations on the SkyScapes-Dense benchmark. mIoU numbers are in [%]. Higher value is better. SSNet stands for SkyScapesNet.

Network	cross-entropy	soft-IoU	soft-Dice	mIoU [%]
Baseline [3]	✓			36.88
SSNet		✓		36.95
SSNet			✓	36.93
SSNet	✓	✓		37.08
SSNet	✓		✓	37.01

Table 3. The class merging policy we used to make the results of our model comparable with the ground-truth labels in Potsdam.

SkyScapes-Dense	Potsdam
low-vegetation	low-vegetation
paved-road	impervious-surface
non-paved-road	impervious-surface
paved-parking-place	impervious-surface
non paved-parking-place	impervious-surface
bikeways	impervious-surface
sidewalks	impervious-surface
entrance-exit	impervious-surface
danger-area	impervious-surface
lane-markings	impervious-surface
danger-area	impervious-surface
car	vehicle
trailer	clutter
van	vehicle
truck	vehicle
large-truck	vehicle
bus	vehicle
clutter	clutter
impervious-surface	impervious-surface
tree	tree

4. Class merging policy for the Potsdam and GRSS_DFC_2018 datasets

In order to be able to evaluate the performance of our method trained on SkyScapes on the Potsdam and GRSS_DFC_2018 datasets with different class definitions, we adopt the class merging policy shown in table 3 on the SkyScapes-Dense prediction task. For the GRSS_DFC_2018 dataset, we applied a similar policy.

5. Further quantitative results

In table 4, we present an extensive benchmark on SkyScapes-Dense using several different methods ranging from the initial FCN8, as the first semantic segmentation method that uses fully convolutional neural networks, to the very recent DenseASPP, BiSeNet, and DeepLabv3+ algorithms. Table 5 shows the IoU_{class}, *i.e.*, the IoU for each of the 20 classes separately. Similarly, table 6 and table 7 show the benchmark results on SkyScapes-Lane (overall and for each class separately). Finally, results for the merged dense classes (the SkyScapes-Dense-Category task) are given in table 8 and table 9.

Table 4. Benchmark of the state-of-the-art methods on the SkyScapes-Dense dataset considering the performance over all 20 classes as a whole. '-' means no specific backbone network is used. 'IoU' and 'f.w.' represent intersection over union and frequency weighted IoU. Models: **best** and **second best**.

method scheme	base modularities	pixel accuracy [%]	IoU [%] mean	IoU [%] f.w.	average [%] recall	average [%] precision
FCN-8s	VGG19	76.95	32.11	63.45	40.73	50.63
FCN-8s	ResNet50	79.19	33.06	67.02	40.78	**65.01**
Dilation	–	72.41	25.65	58.65	34.49	38.48
SegNet	–	74.24	23.14	61.32	29.21	59.56
U-Net	–	52.74	14.15	36.33	21.88	22.87
AdapNet	–	74.52	30.23	61.09	38.38	47.73
BiSeNet	ResNet50	73.25	30.82	59.62	40.25	49.42
BiSeNet	ResNet101	74.62	29.98	61.27	39.21	46.44
BiSeNet	ResNet152	75.41	29.84	62.17	39.30	45.08
DeepLabv3	Res50	68.43	23.36	53.60	30.76	43.98
DeepLabv3	Res101	71.32	25.30	57.30	33.29	41.92
DeepLabv3	Res152	70.27	26.38	56.11	34.39	46.84
DeepLabv3	InceptionV4	26.58	2.44	11.38	5.61	28.83
DenseASPP	MobileNetV2	19.67	2.17	9.01	4.86	19.57
DenseASPP	ResNet50	70.96	24.70	56.60	32.35	39.46
DenseASPP	ResNet101	71.27	24.73	56.58	32.21	40.82
DenseASPP	ResNet152	67.67	24.53	52.58	32.49	40.11
Encoder-Decoder	–	77.83	30.35	65.65	39.91	43.28
Encoder-Decoder-Skip	–	79.08	37.16	67.18	**48.26**	50.16
FC-DenseNet-56	–	77.28	33.22	64.86	42.92	46.98
FC-DenseNet-67	–	78.45	34.67	66.26	44.38	47.71
FC-DenseNet-103	–	79.21	37.78	67.44	46.66	53.89
FRRNA	–	77.59	37.20	65.10	46.44	53.22
FRRNB	–	76.78	32.49	64.10	40.85	49.07
GCN	Res50	77.88	32.88	65.82	43.26	46.99
GCN	Res101	77.57	32.80	65.55	42.14	48.06
GCN	Res152	77.50	32.92	65.12	41.60	49.65
Mobile-U-Net	–	75.25	26.01	62.35	34.01	39.70
Mobile-U-Net-Skip	–	77.56	34.96	65.26	44.52	49.49
PSPNet	Res50	74.49	30.31	61.45	40.02	44.51
PSPNet	Res101	74.62	30.44	61.62	40.48	43.63
PSPNet	Res152	74.09	30.20	60.95	39.76	43.91
RefineNet	Res50	77.02	34.23	64.68	44.15	49.54
RefineNet	Res101	77.08	33.27	64.66	42.23	48.46
RefineNet	Res152	77.75	36.39	65.52	46.12	52.17
DeepLabv3+	Res50	75.88	31.95	63.00	40.20	49.76
DeepLabv3+	Res101	75.94	31.95	63.25	41.48	48.61
DeepLabv3+	Res152	76.14	31.91	63.29	42.48	46.85
DeepLabv3+	Xception65	**80.25**	**38.20**	**68.81**	**47.97**	55.34
SkyScapesNet	–	**83.56**	**40.13**	**72.67**	47.85	**65.93**

6. Further qualitative results

We also provide more qualitative results to demonstrate the generalization capability of our method. Figure 1 shows the satellite image of the whole area of Munich, Germany. This image was taken by the WorldView4 satellite with a *ground sampling distance (GSD)* of 30 cm.

The patches in fig. 2 highlight binary lane-marking segmentation results on the satellite image, the feasibility of which is, to our knowledge, demonstrated here for the first time. In this work, we expanded the work of Azimi et al. [1] on binary lane-marking extraction. It is thus feasible to extract whole-city lane-marking maps from a single satellite image.

Figure 3, fig. 4, and fig. 5 show further qualitative re-

Table 5. Evaluation of the state-of-the-art methods on the SkyScapes-Dense dataset for each class separately. '-' means no specific backbone network is used. 'IoU' represents intersection over union. LV, PR, nPR, PPC, nPPC, BW, SW, EE, DA, LM, B, Ca, TR, V, TK, LT, Bu, Cl, IS, and T represent low-vegetation, paved-road, non-paved-road, paved-parking-place, non-paved-parking-place, bikeway, sidewalk, entrance-exit, danger area, lane-marking, building, car, trailer, van, truck, long truck, bus, clutter, impervious surface, and tree. Models: **best** and **second best**.

method	base	IoU [%]																				
		mean	LV	PR	nPR	PPC	nPPC	BW	SW	EE	DA	LM	B	C	TR	V	TK	LT	Bu	Cl	IS	T
FCN-8s	VGG19	32.11	67.11	63.74	**6.82**	29.11	0.12	25.9	32.64	7.14	43.99	36.46	81.2	64.09	0.08	32.67	7.86	0.0	2.01	50.47	17.24	73.53
FCN-8s	ResNet50	33.06	68.45	67.71	**6.41**	34.71	0.0	32.08	40.72	17.8	36.53	8.31	86.7	67.88	0.0	29.87	8.65	**5.27**	0.0	**50.64**	23.75	75.65
Dilation	–	25.65	58.11	58.84	1.78	25.74	0.02	19.74	31.87	17.15	0.0	1.49	80.55	47.5	0.0	21.87	15.1	4.62	1.21	40.24	19.64	67.48
SegNet	–	23.14	63.96	61.9	0.94	27.5	**1.19**	7.7	30.65	0.72	0.0	4.99	81.92	43.94	0.0	0.0	0.0	0.0	0.0	44.73	21.7	70.86
U-Net	–	14.15	46.68	37.17	1.6	14.89	0.07	0.07	8.81	0.0	0.0	37.66	49.63	23.0	0.44	2.34	0.91	0.11	0.0	15.84	6.83	36.87
AdapNet	–	30.23	59.99	65.28	1.49	27.4	0.19	28.7	36.86	19.08	34.08	21.49	80.74	54.7	3.07	26.04	11.5	0.92	11.4	31.27	19.95	70.5
BiSeNet	ResNet50	30.82	59.68	65.43	2.14	25.25	0.95	25.9	38.5	15.2	47.01	22.93	82.76	60.9	**3.99**	31.34	12.85	0.71	8.42	27.26	22.07	63.0
BiSeNet	ResNet101	29.98	61.55	65.39	0.62	21.99	0.52	24.39	37.71	13.12	23.59	20.62	82.7	63.84	**4.07**	32.16	17.5	0.68	2.7	34.52	23.85	68.13
BiSeNet	ResNet152	29.84	63.02	65.87	1.99	25.5	0.05	27.4	38.77	17.65	8.58	19.93	84.19	62.79	1.74	32.81	15.57	0.01	10.03	28.79	24.13	67.89
DeepLabv3	Res50	23.36	57.12	55.25	1.7	20.86	0.64	14.41	27.7	10.49	3.49	4.27	75.17	52.43	1.24	25.14	7.07	0.0	8.24	26.28	18.56	57.06
DeepLabv3	Res101	25.30	59.69	57.28	0.85	22.39	0.31	14.24	29.28	9.85	9.43	6.91	78.65	53.57	0.25	26.66	6.43	1.63	14.73	30.02	19.14	64.61
DeepLabv3	Res152	26.38	56.96	60.2	2.86	20.61	0.42	17.76	31.76	10.55	19.21	8.85	80.38	56.38	1.43	27.78	8.77	6.47	7.57	29.7	20.75	59.15
DeepLabv3	InceptionV4	2.44	5.13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19.72	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23.93
DenseASPP	MobileNetV2	2.17	17.64	0.06	0.21	1.11	0.19	0.0	1.03	0.04	0.0	0.0	13.95	0.0	0.0	0.09	0.0	0.0	0.0	0.73	1.08	7.35
DenseASPP	ResNet50	24.70	58.19	60.79	1.44	23.31	0.06	16.88	31.85	11.96	0.2	20.65	76.57	57.3	0.44	19.4	5.32	0.0	1.98	24.22	21.18	62.23
DenseASPP	ResNet101	24.73	59.05	60.85	1.47	23.07	0.09	16.92	31.51	12.85	4.56	18.96	76.59	56.12	0.55	17.84	4.81	0.01	8.92	18.22	20.28	61.87
DenseASPP	ResNet152	24.53	51.53	59.83	2.47	22.74	0.04	19.55	31.92	11.66	2.64	22.44	77.94	56.86	0.66	18.32	9.84	0.96	2.27	27.92	18.96	52.08
Encoder-Decoder	–	30.35	67.6	65.69	2.28	31.27	0.05	30.7	40.71	19.72	0.5	23.87	84.59	55.64	0.75	20.72	15.84	2.67	8.3	34.55	26.44	75.02
Encoder-Decoder-Skip	–	37.16	67.48	69.7	3.68	33.54	0.37	**36.83**	42.88	23.14	33.77	**65.13**	86.47	69.86	1.09	33.33	**22.81**	3.27	4.47	44.3	25.87	75.26
FC-DenseNet-56	–	33.22	66.47	65.47	1.74	29.89	0.34	26.26	38.25	17.01	26.48	61.56	83.99	63.51	0.13	24.95	11.07	0.31	11.37	39.21	21.88	74.54
FC-DenseNet-67	–	34.67	68.4	66.71	2.28	29.84	0.06	30.78	41.28	18.14	25.15	64.93	84.65	66.33	0.42	29.04	18.87	1.12	0.01	45.53	24.24	75.52
FC-DenseNet-103	–	37.78	69.18	68.19	0.79	33.4	0.01	31.97	42.67	20.28	**56.5**	66.69	85.53	66.94	1.21	31.81	20.51	3.61	4.26	49.84	25.88	76.42
FRRN-A	–	37.20	61.59	67.23	3.61	19.17	0.7	32.28	38.65	11.53	8.55	63.45	83.28	68.83	1.99	32.92	**20.74**	4.03	7.74	37.39	23.66	64.66
FRRN-B	–	32.49	65.53	67.04	1.62	27.86	0.0	31.94	39.27	18.82	15.38	61.62	82.4	62.3	1.95	26.28	11.05	1.61	13.01	24.03	24.92	73.09
GCN	Res50	32.88	67.28	67.24	1.08	31.87	0.08	22.75	38.84	14.16	20.32	55.47	85.12	66.68	0.1	29.67	13.25	0.18	6.23	37.04	25.59	74.75
GCN	Res101	32.80	66.95	66.47	4.97	25.36	0.52	24.43	40.04	17.04	18.48	52.98	85.85	67.39	2.1	30.41	13.24	2.07	2.41	34.25	26.26	74.77
GCN	Res152	32.92	66.44	64.86	2.27	25.81	0.0	28.21	39.48	16.4	19.67	54.41	85.38	66.72	2.39	30.8	8.63	0.87	4.16	42.28	25.19	74.4
Mobile-U-Net	–	26.01	64.3	63.87	2.93	27.31	0.37	23.36	36.18	18.84	0.0	5.68	80.98	43.9	0.04	15.67	15.5	0.98	6.48	18.11	22.61	73.02
Mobile-U-Net-Skip	–	34.96	66.49	67.49	2.5	30.94	0.5	26.26	38.46	19.95	38.01	62.15	84.5	64.75	3.67	31.05	15.67	0.41	10.4	37.88	24.06	74.01
PSPNet	Res50	30.31	64.11	60.03	1.01	21.88	0.91	17.46	31.74	10.6	16.83	50.08	80.36	63.94	1.68	28.76	18.09	1.29	7.65	36.66	20.74	72.34
PSPNet	Res101	30.44	64.2	59.72	0.79	22.61	0.28	19.53	32.42	10.52	31.29	50.22	80.53	62.78	0.8	27.48	15.42	3.09	0.04	34.09	20.13	72.81
PSPNet	Res152	30.20	64.04	57.95	3.75	22.14	0.79	19.91	31.45	10.62	27.78	51.29	79.96	63.45	0.95	27.87	13.23	0.35	4.0	31.49	21.47	71.53
RefineNet	Res50	34.23	66.78	63.34	3.58	29.77	0.07	26.41	36.11	14.97	32.31	41.48	83.62	69.29	2.07	37.82	15.91	2.43	13.61	46.32	24.59	74.12
RefineNet	Res101	33.27	66.19	64.63	4.28	29.91	0.21	28.68	35.6	14.3	17.41	41.92	84.08	69.41	0.57	38.12	17.31	3.02	7.21	44.08	24.79	73.72
RefineNet	Res152	36.39	67.05	65.41	3.26	32.83	0.3	32.19	38.08	17.19	**56.6**	44.79	84.23	69.06	2.39	37.4	16.77	3.45	**15.85**	42.31	23.86	74.75
DeepLabv3+	Res50	31.95	64.36	63.69	2.68	29.05	0.56	25.32	35.69	15.12	31.4	42.54	81.97	65.27	1.22	31.69	13.97	4.4	1.78	34.82	21.02	72.56
DeepLabv3+	Res101	31.95	64.61	63.7	1.58	29.5	0.59	23.85	35.22	15.34	27.76	41.3	82.25	65.01	2.93	29.81	11.07	0.0	13.58	35.38	22.73	72.83
DeepLabv3+	Res152	31.91	64.78	63.88	2.42	27.86	0.23	24.8	36.55	14.22	17.19	45.27	83.0	66.59	2.37	33.24	16.28	2.74	4.87	36.98	23.04	71.97
DeepLabv3+	Xception65	**38.20**	**69.92**	**69.79**	2.62	**34.85**	0.67	28.72	**43.98**	**25.84**	46.43	46.73	**88.12**	**70.73**	2.44	**39.25**	15.99	**5.33**	**16.64**	50.38	**28.45**	**77.16**
SkyScapesNet	–	**40.13**	**72.33**	**78.48**	5.86	**52.04**	**4.13**	**51.39**	**52.9**	**27.24**	4.33	**65.26**	**89.16**	**72.01**	1.03	**38.33**	19.33	0.0	0.0	**56.02**	**35.39**	**77.41**

sults on three aerial images with different scales, GSD, illumination conditions, and from different geographical areas. These figures show the whole-image dense prediction and zoomed-in sample areas with dense, multi-class lane-marking, and multi-class edge segmentations.

Table 6. Benchmark of the state-of-the-art methods on the SkyScapes-Lane dataset considering the performance over all 13 classes as a whole. '-' means no specific backbone network is used. 'IoU' and 'f.w.' represent intersection over union and frequency weighted IoU. Models: **best** and **second best**.

method scheme	base modularities	pixel accuracy [%]	IoU [%] mean	IoU [%] f.w.	average [%] recall	average [%] precision
FCN-8s	VGG19	99.81	10.86	99.66	11.66	**92.84**
FCN-8s	ResNet50	99.83	13.74	99.69	15.23	77.96
Dilation	–	99.77	8.56	99.57	8.90	50.80
SegNet	–	99.80	9.02	99.64	10.11	**94.45**
U-Net	–	99.73	8.97	99.62	12.73	88.26
AdapNet	–	99.82	20.20	99.67	22.21	53.60
BiSeNet	ResNet50	99.81	23.77	99.66	28.71	51.42
BiSeNet	ResNet101	99.81	18.30	99.64	20.22	52.66
BiSeNet	ResNet152	99.81	17.85	99.65	19.78	49.54
DeepLabv3	Res50	99.80	16.15	99.62	18.94	55.44
DeepLabv3	Res101	99.80	13.27	99.61	14.35	45.67
DeepLabv3	Res152	99.80	12.64	99.61	13.42	60.52
DeepLabv3	InceptionV4	58.60	4.51	58.54	5.47	23.06
DenseASPP	MobileNetV2	99.80	7.68	99.60	7.69	69.22
DenseASPP	ResNet50	99.81	16.16	99.65	17.50	52.98
DenseASPP	ResNet101	99.81	17.00	99.65	18.74	46.02
Encoder-Decoder	–	99.85	21.87	99.74	25.51	40.27
Encoder-Decoder-Skip	–	**99.92**	**48.87**	**99.85**	55.31	70.63
FRRN-A	–	**99.92**	46.85	**99.85**	55.06	67.11
FRRN-B	–	**99.92**	47.02	**99.85**	54.72	66.19
GCN	Res50	99.90	35.65	99.82	43.09	55.65
GCN	Res101	99.90	34.71	99.82	41.42	56.49
GCN	Res152	99.90	33.43	99.82	39.88	56.61
Mobile-U-Net-Skip	–	99.91	41.21	99.84	47.48	64.60
PSPNet	Res50	99.90	35.44	99.82	42.80	57.15
PSPNet	Res101	99.90	35.85	99.82	42.64	58.23
PSPNet	Res152	99.90	34.09	99.82	40.56	56.32
RefineNet	Res152	99.80	7.68	99.60	7.69	99.98
DeepLabv3+	Res50	99.86	27.68	99.75	31.82	55.81
DeepLabv3+	Res101	99.86	27.36	99.74	32.61	50.54
DeepLabv3+	Res152	99.86	31.88	99.75	36.82	59.16
DeepLabv3+	Xception65	99.87	37.14	99.77	43.14	62.07
FC-DenseNet-56	–	**99.92**	44.91	**99.85**	52.47	65.67
FC-DenseNet-67	–	**99.92**	47.35	**99.85**	54.83	69.01
FC-DenseNet-103	–	**99.92**	48.42	**99.85**	**55.32**	69.01
SkyScapesNet	–	**99.93**	**51.93**	**99.87**	**60.53**	72.29

Table 7. Evaluation of the state-of-the-art methods on the SkyScapes-Lane dataset for each class separately. '-' means no specific back-bone network is used. 'IoU' represents intersection over union. NL, DL, LL, TDL, TS, OS, PS, CW, SL, ZZ, nPZ, PZ, and R represent non lane-marking, dash line, long line, tiny dash line, turn sign, other signs, plus sign, crosswalk, stop line, zebra zone, no parking zone, parking zone, and the rest of lane-markings.

method	base	IoU [%]													
		mean	NL	DL	LL	TDL	TS	OS	PS	CW	SL	ZZ	nPZ	PZ	R
FCN-8s	VGG19	10.86	99.83	22.39	18.94	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FCN-8s	ResNet50	13.74	99.84	39.86	27.24	0.0	0.0	0.0	0.0	11.66	0.0	0.02	0.0	0.0	0.0
Dilation	–	8.56	99.77	0.03	5.41	0.65	1.26	2.51	0.0	0.0	1.68	0.0	0.0	0.0	0.0
SegNet	–	9.02	99.83	0.0	17.39	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
U-Net	–	8.97	99.81	0.23	16.56	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AdapNet	–	20.20	99.83	23.78	27.07	12.62	15.08	9.96	2.07	24.44	46.42	0.65	0.16	0.53	0.0
BiSeNet	ResNet50	23.77	99.82	22.62	22.47	13.55	13.72	20.2	1.91	46.1	42.7	16.2	8.81	0.88	0.0
BiSeNet	ResNet101	18.30	99.81	14.5	20.1	9.32	10.71	15.14	0.58	30.65	21.29	13.45	1.86	0.46	0.0
BiSeNet	ResNet152	17.85	99.81	18.1	21.4	8.3	14.3	15.8	0.0	4.26	29.4	18.57	1.78	0.32	0.0
DeepLabv3	Res50	16.15	99.8	6.79	14.64	1.34	2.65	11.9	0.0	49.48	21.44	0.78	1.09	0.0	0.0
DeepLabv3	Res101	13.27	99.8	2.58	10.27	0.26	1.3	8.86	0.0	32.08	17.19	0.09	0.12	0.0	0.0
DeepLabv3	Res152	12.64	99.8	3.1	10.51	1.28	0.35	11.36	0.0	18.44	17.81	1.61	0.05	0.0	0.0
DeepLabv3	InceptionV4	4.51	58.66	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DenseASPP	MobileNetV2	7.68	99.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DenseASPP	ResNet50	16.16	99.82	21.9	21.87	13.03	13.77	0.37	5.9	0.0	32.47	0.17	0.51	0.27	0.0
DenseASPP	ResNet101	17.00	99.82	21.46	21.31	12.7	16.58	0.12	4.5	8.45	34.35	1.43	0.02	0.25	0.0
Encoder-Decoder	–	21.87	99.86	51.2	42.73	13.62	8.02	10.1	11.57	2.13	34.48	6.5	1.97	2.0	0.11
Encoder-Decoder-Skip	–	**48.87**	**99.93**	71.14	53.83	62.16	58.67	**65.75**	28.48	**79.07**	65.75	**22.57**	20.77	6.99	0.22
FRRN-A	InceptionV4	46.85	**99.93**	71.27	**58.89**	60.05	57.74	56.1	31.5	64.2	66.74	13.53	20.06	8.93	0.12
FRRN-B	–	47.02	**99.93**	72.19	58.32	57.25	**61.18**	58.75	31.68	66.36	**69.18**	9.61	22.14	4.65	0.0
GCN	Res50	35.65	99.92	67.16	54.3	47.53	35.22	25.37	18.2	51.71	46.87	5.6	10.05	1.51	0.0
GCN	Res101	34.71	99.91	66.58	50.47	43.64	38.56	20.88	11.13	56.4	47.21	4.05	10.29	2.1	0.0
GCN	Res152	33.43	99.91	65.42	53.32	45.21	28.63	24.47	6.63	51.43	39.34	2.02	15.51	2.71	0.0
Mobile-U-Net	–	19.84	99.84	42.11	39.21	11.6	6.26	16.2	6.83	0.5	32.48	0.92	1.34	0.67	0.0
PSPNet	Res50	35.44	99.91	64.35	52.99	42.44	35.17	22.48	17.5	42.78	56.16	13.41	9.74	3.77	0.06
PSPNet	Res101	35.85	99.91	65.57	52.15	42.23	37.87	18.65	20.86	44.24	58.55	13.84	8.32	3.81	0.11
PSPNet	Res152	34.09	99.91	64.41	53.39	43.07	36.46	11.54	20.59	33.84	56.42	14.46	7.69	1.33	0.0
RefineNet	Res152	7.68	99.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DeepLabv3+	Res50	27.68	99.87	46.04	47.53	27.41	25.31	27.84	8.84	14.53	50.11	6.66	3.67	1.72	**0.33**
DeepLabv3+	Res101	27.36	99.87	42.93	46.32	26.86	26.35	22.04	1.32	34.79	48.02	1.12	4.69	1.41	0.0
DeepLabv3+	Res152	31.88	99.87	42.51	43.16	26.74	29.55	33.12	11.97	49.03	58.63	5.74	9.39	4.69	0.0
DeepLabv3+	Xception65	37.14	99.88	47.75	52.32	31.07	39.88	37.19	12.14	53.6	66.46	17.22	**22.39**	2.04	**0.87**
FC-DenseNet-56	–	44.91	**99.93**	70.01	56.23	63.14	53.86	59.74	**34.86**	51.98	59.75	14.35	13.67	6.32	0.0
FC-DenseNet-67	–	47.35	**99.93**	70.91	56.06	**64.61**	59.9	51.98	30.09	69.29	65.6	13.8	21.16	**12.14**	0.06
FC-DenseNet-103	–	48.42	**99.93**	**72.25**	57.47	64.16	59.9	54.62	**34.89**	**74.34**	66.47	19.04	20.65	5.73	0.0
SkyScapesNet	–	**51.93**	**99.94**	**72.56**	**68.72**	**67.63**	**63.59**	**64.22**	30.97	54.55	**68.48**	**38.53**	**36.88**	**9.01**	0.0

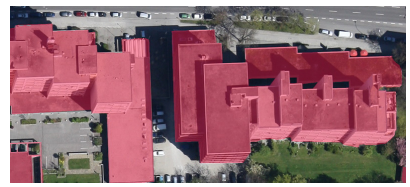
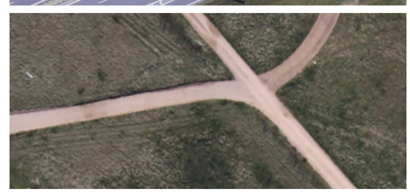
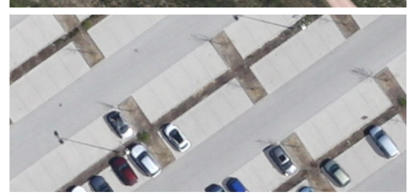
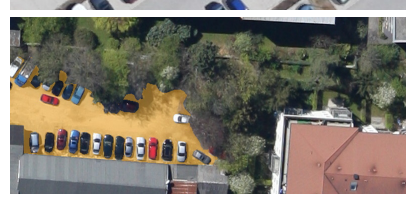
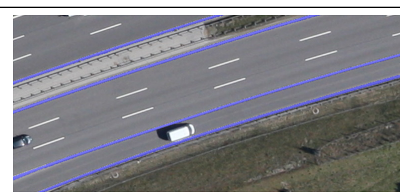
Table 8. Result of SkyScapesNet on the SkyScapes-Dense-Category task over all 11 classes as a whole. '-' means no specific backbone network used. 'IoU' and 'f.w.' represent intersection over union and frequency weighted IoU.

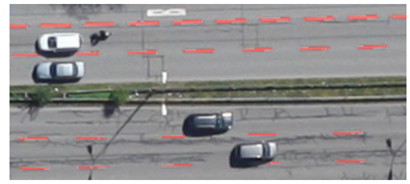
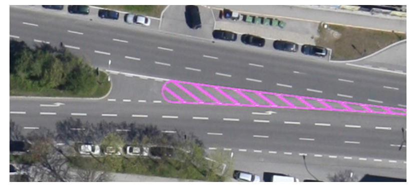
method	base	pixel	IoU [%]		average [%]	
scheme	modularities	accuracy [%]	mean	f.w.	recall	precision
SkyScapesNet	–	86.10	52.27	77.77	63.49	65.65

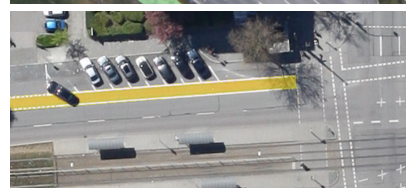
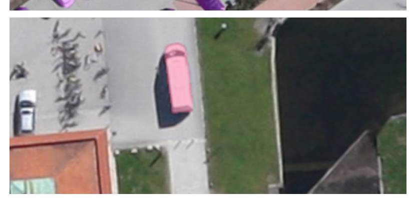
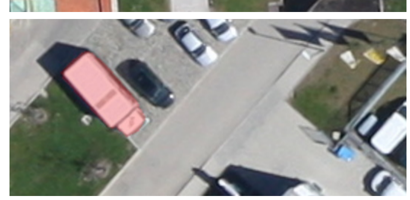
Table 9. Result of SkyScapesNet on SkyScapes-Dense-Category task for each class separately. '-' represents no specific back-bone network used. 'IoU' represents intersection over union. The abbreviations for classes are N: nature, D: driving-area, P: parking-area, H: human-area, SH: shared human and vehicle area, RF: road-feature, R: residential, DV: dynamic-vehicle, SV: static-vehicle, HS: man-made surface, and O: others.

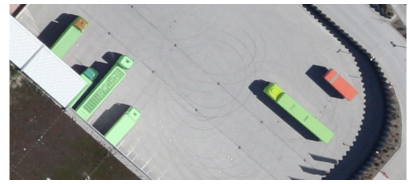
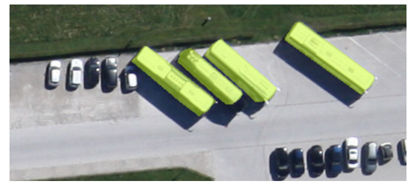
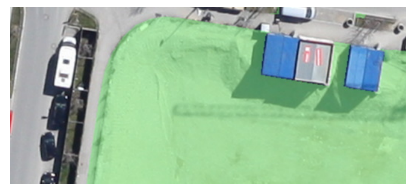
method	base	IoU [%]											
		mean	N	D	P	H	SH	RF	R	DV	SV	HS	O
SkyScapesNet	–	52.27	90.79	68.86	36.8	50.95	25.87	66.09	86.84	72.79	3.45	44.67	27.84

Table 10: List of categories including their definition and a typical example.

Category	Class	Definition	Examples
nature	low vegetation	Includes all natural areas without large plants, *e.g.*, lawns.	
	tree	Areas covered by large plants, such as trees or large bushes.	
residential	building	Structures with walls and a roof, such as houses, factories, and garages.	
vehicle area	paved-road	Includes all roads that are asphalted.	
	non-paved-road	All roads that are not paved, *e.g.*, forest roads, dirt roads, and unsurfaced roads.	
	paved-parking-place	includes all asphalted areas for parking vehicles, such as car parks. The parking area include the vehicle as well which has not been shown in the figure	
	non-paved-parking-place	Unsurfaced areas used for parking. The parking area include the vehicle as well which has not been shown in the figure.	
lane-markings	long line	Thin solid lines, such as no passing lines or roadside markings.	

dash line	Any broken line with long line segments, *e.g.*, lane separators.	
tiny dash line	Any broken line with tiny line segments, *e.g.*, lines enclosing pedestrian crossings.	
zebra zone	Areas with diagonal lines, *e.g.*, restricted zones.	
turn sign	Arrows on the road, such as intersection arrows or merge arrows.	
stop line	Thick solid line across lanes that signal to stop behind the line.	
parking zone	Includes any lines that mark parking spots.	
no parking zone	Zig-zag lines next to the curb mark that indicate that stopping or parking is forbidden.	
crosswalk	Zebra-striped markings across the roadway mark a pedestrian crosswalk.	
plus sign	All crossing tiny lines.	

	other signs	Includes all other signs, *e.g.*, numbers that indicate the speed limit.	
	rest of lane-markings	Any other lane-marking.	
human area	sidewalk	Path with a hard surface on one or both sides of a road for pedestrians.	
	bikeway	Includes all lanes or roads for bikes.	
	danger-area	The intersection of bikeways with road marked with red, blue or green in Germany and some other countries	
shared area	entrance-exit	All entrance and exit areas that are shared with pedestrians.	
vehicle	car	Includes all cars except vans.	
	van	Any vehicles with box-like shapes.	
	truck	Includes all small trucks such as delivery trucks.	

	long-truck	All long trucks such as heavy goods vehicles.	
	trailer	Includes all trailers that can be attached to any vehicle, *e.g.*, trucks or cars.	
	bus	Any buses including tourist coaches, school buses, and public buses.	
other	impervious surface	Includes all other surfaces, such as construction sites, and non-temporary obstacles road users cannot go through (*e.g.*, low wall, rocky terrain, river).	
	clutter	Includes all other human made structures, such as garbage bins, fences, or outdoor furniture.	

References

[1] 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. 3

[2] 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. 1

[3] 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. 1, 2

[4] 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. 2

[5] 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. 2

Figure 1. A satellite image – acquired by WorldView4 – over the whole area of Munich, Germany. The size of the image is 45386×33753 pixels which is about 173 MP.

Figure 2. Sample patches from the lane-marking map of the whole area of Munich extracted using our SkyScapesNet algorithm applied to a WorldView4 satellite image.

Figure 3. Performance of SkyScapesNet trained on SkyScapes and tested on different images with different timestamps, illumination conditions, camera angle, GSD, and geographical area. The results are without GSD adjustment. This image is from Kitzingen, Germany, taken in 2015. Top images, from left to right: RGB, dense segmentation. Bottom samples, from left to right: RGB, dense segmentation, lane markings segmentation, borders segmentation.

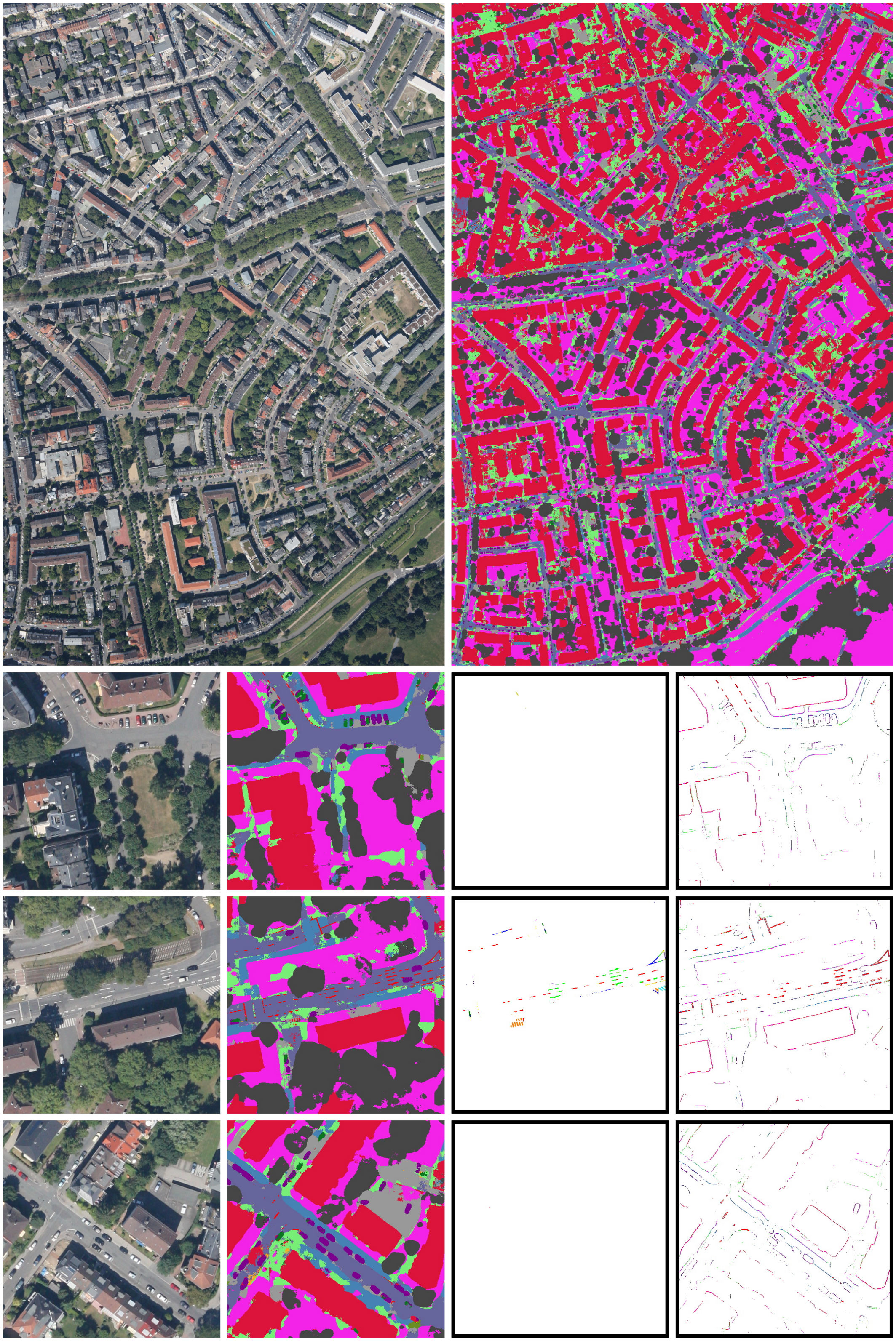


Figure 4. Performance of SkyScapesNet trained on SkyScapes and tested on different images with different timestamps, illumination conditions, camera angle, GSD, and geographical area. The results are without GSD adjustment. This image is from Frankfurt, Germany, taken in 2013. Top images, from left to right: RGB, dense segmentation. Bottom samples, from left to right: RGB, dense segmentation, lane markings segmentation, borders segmentation.

Figure 5. Performance of SkyScapesNet trained on SkyScapes and tested on different images with different timestamps, illumination conditions, camera angle, GSD, and geographical area. The results are without GSD adjustment. This image is from Braunschweig, Germany, taken in 2017. Top images, from left to right: RGB, dense segmentation. Bottom samples, from left to right: RGB, dense segmentation, lane markings segmentation, borders segmentation.