RGB Road Scene Material Segmentation
Supplementary Document

Sudong Cai[0000−0002−5446−5618], Ryosuke Wakaki[0000−0003−3917−9012], Shohei Nobuhara[0000−0002−3204−8696], and Ko Nishino[0000−0002−3534−3447]

Graduate School of Informatics, Kyoto University, Kyoto, Japan
https://vision.ist.i.kyoto-u.ac.jp/

1 KITTI-Materials Dataset

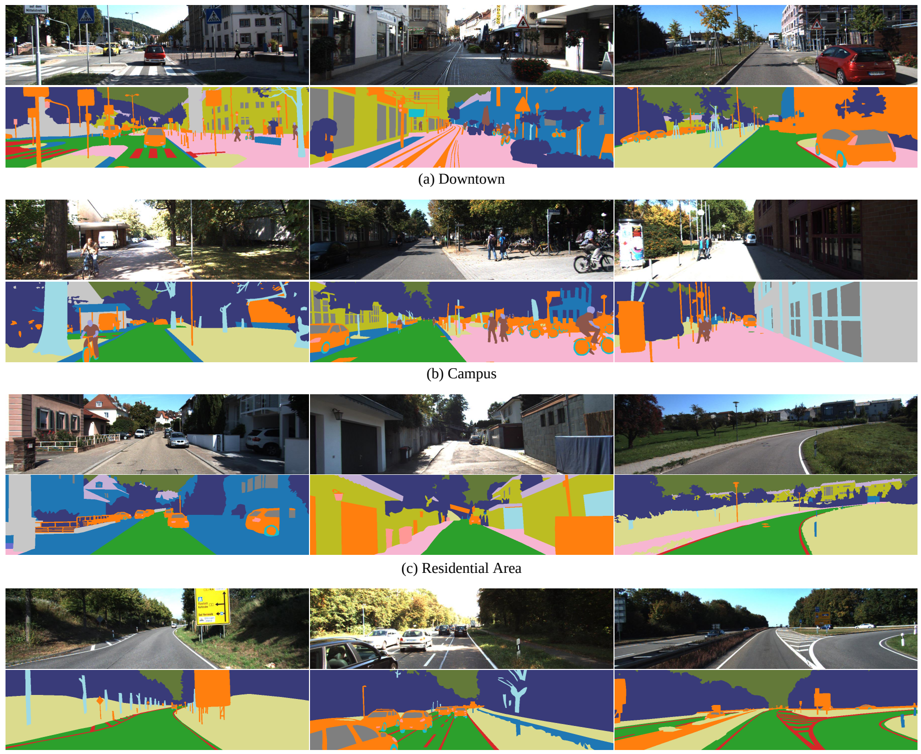
KITTI-Materials dataset consists of 1000 images covering 24 different road scenes of downtown, campus, residential area, highway, and other cityscapes. Figure 1 shows an example of the various types of road scene images with their corresponding material annotations. Table 1 reports the detailed per-class pixel statistics of each scene (with scene IDs), where material categories "sand," "gravel," and "water" only show up in very few scenes and are comparatively rare due to the natural long-tail distribution of road scene materials.

For evaluation on KITTI-materials, we define two training-test data splits (*i.e.*, Split-1 and Split-2), where the test set of Split-1 (consists of scenes 0926019, 0926086, 0930034, and 1003047) contains more scenes with highways and rural areas while Split-2 (consists of scenes 0926064, 0926095, 0929004, and 0930016) is biased to city scenes. Both splits contain all 1000 images of KITTI-Materials, where 800 images are for training and 200 images are for testing, but with different combinations of scenes. Figure 2 shows the different characteristics of Split-1 and -2 with example images, and Table 1 reports the per-class statistics of them. Both training and test sets of these two splits show very strong imbalance in the material categories. Note that researchers can define their tailored split policies for their experiments after we publicly disseminate the KITTI-Materials dataset. That said, as some of the materials only appear in images of a few scenes, splits with all material categories in both train and test sets are hard to realize when based on scenes except for the suggested Split-1 and -2.

Table 2 shows the statistical properties of KITTI-Materials and other related street-view material (*i.e.*, MCubeS [4]) and semantic segmentation datasets (*i.e.*, Cityscapes [3], Daimler Urban Segmentation (DUS) [5], and KITTI Semantic Segmentation (KITTI-SS) [1]), where only KITTI-Materials and MCubeS provide dense material annotations. In contrast to MCubeS which comprises rare suburban scenes, our KITTI-Materials is comprised of diverse images of city and suburban landscapes with a broader city-scale sampling range. KITTI-Materials has higher annotation density compared to road scene semantic segmentation datasets (Cityscapes, DUS, and KITTI-SS), which ensures high quality material annotations for realistic driving view images.

Table 1. Per-class pixel statistics for each scene in KITTI-Materials dataset. “Scn ID” and “Imgs” denote “Scene ID” and “Images,” respectively; “road mk,” “fab, lthr,” “rubr, vl,” “cob,” and “hum bd” denote “road marking,” “fabric, leather,” “rubber, vinyl,” “cobblestone,” and “human body,” respectively. Note that scene-0926095 includes an invalid pixel. “Trn-1, -2” and “Tst-1, -2” denote training and test sets of Split-1 and -2, respectively.

Scn ID	Imgs	asphalt	concrete	metal	road mk	fab, lthr	glass	plaster	plastic	rubr, vl	sand	gravel	ceramic	cob	brick	grass	wood	leaf	water	hum bd	sky	Pixels
0926002	1	83K	58K	13K	4125	17K	2314	0	0	3656	0	0	1018	0	20K	8919	566	137K	0	1035	39K	389K
0926019	50	2575K	501K	391K	23K	794	38K	0	3165	4850	0	0	0	0	207K	5579K	71K	9375K	308	95	686K	19M
0926039	50	2478K	472K	3039K	0	5297	1186K	6814K	85K	128K	0	0	411K	1158K	643K	375	16K	2107K	0	1954	911K	19M
0926048	5	0	162K	326K	0	0	258K	768K	8725	11K	0	0	9232	0	274K	5945	905	29K	0	0	93K	1946K
0926056	50	4340K	1237K	1517K	298K	11K	369K	188K	24K	26K	0	22K	114K	0	149K	2025K	723K	6934K	0	3170	1475K	19M
0926059	50	2635K	744K	2715K	341K	18K	821K	3175K	107K	115K	0	0	264K	1390K	701K	1481K	147K	3731K	0	4795	1066K	19M
0926064	50	2228K	4846K	2390K	7939	22K	745K	0	65K	88K	0	0	363K	1175	707K	44K	117K	7234K	714	7907	590K	19M
0926070	50	3251K	1576K	595K	264K	2611	87K	477K	25K	11K	0	0	75K	0	34K	4441K	416K	7144K	0	1489	1054K	19M
0926079	20	947K	316K	357K	0	0	97K	888K	4578	3582	0	0	39K	676K	0	343K	81K	3924K	0	0	106K	7782K
0926084	50	4444K	3761K	2519K	263K	13K	702K	0	55K	111K	0	0	67K	1190K	73K	576K	441K	4679K	0	3659	557K	19M
0926086	50	2203K	1478K	3582K	7604	25K	87K	2417K	119K	17K	0	1797	238K	508K	376K	1193K	253K	5785K	0	3874	1163K	19M
0926091	50	0	3967K	2603K	68K	519K	1260K	2399K	104K	139K	6789	0	298K	4781K	298K	41K	59K	2366K	0	80044	468K	19M
0926095	50	2802K	500K	2808K	62K	89K	1115K	4797K	59K	120K	5604	0	155K	1584K	2589K	124K	281K	1930K	0	31440	404K	19M
0926117	50	2808K	1270K	2641K	4468	57K	771K	0	54K	97K	25K	0	3580	643K	556K	1953K	544K	7699K	0	4311	326K	19M
0928037	15	366K	311K	794K	1668	192K	81K	125K	5065	63K	0	0	1282	1254K	217K	37K	126K	2107K	0	47515	108K	5837K
0928045	9	374	9552	154K	0	85K	415K	0	633	916	0	0	0	782K	521K	82K	511K	894K	0	11971	34K	3502K
0929004	50	2689K	602K	1220K	251K	0	487K	0	33K	54K	0	0	0	0	0	3400K	376K	9648K	0	0	696K	19M
0930016	50	4053K	638K	854K	593K	0	50K	190K	15K	11K	0	3497	175K	24K	1048K	1608K	74K	8762K	0	0	1356K	19M
0930020	50	2487K	1751K	3302K	20K	26K	177K	486K	30K	70K	269K	0	389K	1202K	48K	2866K	455K	3489K	0	4965	2385K	19M
0930033	50	2912K	168K	287K	103K	9962	12K	117K	2083	5772	387	0	61K	847K	5901	2499K	90K	10562K	0	1541	1772K	19M
0930034	50	742K	756K	373K	1545	0	45K	845K	32K	7065	17K	0	215K	1313K	275K	2365K	795K	10558K	0	0	1116K	19M
1003034	50	3951K	777K	660K	166K	0	89K	95K	12K	17K	0	36K	13842	1377K	40K	2570K	594K	8145K	0	0	912K	19M
1003042	50	3957K	719K	1585K	455K	0	61K	9931	6069	15K	0	0	0	0	0	2646K	2508	7321K	0	0	2677K	19M
1003047	50	3902K	4854	2488K	184K	0	481K	0	71K	134K	0	0	0	0	0	670K	3498	8398K	0	0	3120K	19M
Total	1K	56M	27M	37M	3121K	1092K	9437K	24M	920K	1253K	323K	63K	2891K	19M	8782K	37M	6178K	133M	1022	210K	23M	389M
Trn-1	800	46M	24M	30M	2904K	1066K	8786K	20530K	695K	1090K	306K	61K	2438K	16910K	7924K	26751M	5055K	99M	714	206K	17M	311M
Tst-1	200	9422K	2739K	6833K	217K	26K	651K	3262K	225K	163K	17K	1797	453K	1820K	859K	9807M	1123K	34M	308	3969	6085K	78M
Trn-2	800	44M	20M	29942K	2206K	981K	7040K	19M	748K	980K	318K	60K	2198K	17M	4439K	31M	5330K	105M	308	170K	20M	311M
Tst-2	200	12M	6587K	7271K	914K	111K	2397K	4987K	172K	273K	5604	3497	693K	1609K	435K	5175K	848K	28M	714	39K	3046K	78M



(a) Downtown

(b) Campus

(c) Residential Area

(d) Highway

Fig. 1. Example images and their corresponding material annotations from the KITTI-materials dataset. From top to bottom are examples regarding “downtown,” “campus,” “residential area,” and “highway,” respectively.

2 Per-Class Evaluation

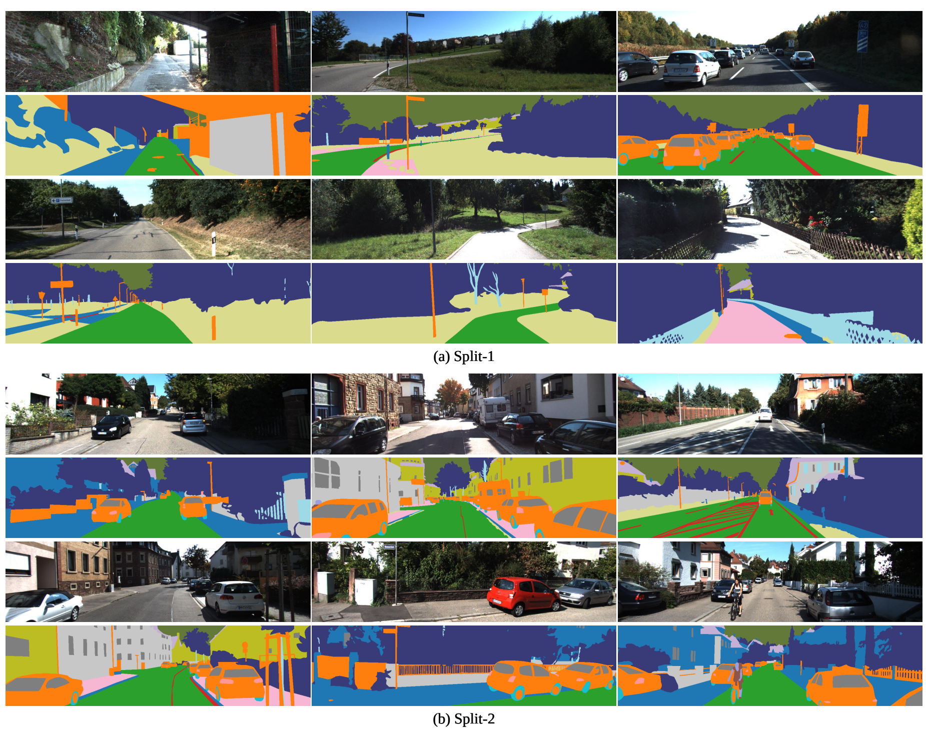
In Table 3, we report detailed per-class results for 20 material categories on Split-1 and Split-2 of KITTI-Materials.

Our results on Split-1 shows that our RMSNet yields the best results on most material classes including “asphalt,” “concrete,” “metal,” “road marking,” “fabric, vinyl,” “glass,” “plaster,” “rubber, vinyl,” “cobblestone,” “brick,” “wood,” “human, body,” and “sky,” where RMSNet introduces clear gains on “asphalt,” “road marking,” “rubber, vinyl,” “plaster,” “metal,” “cobblestone,” “brick,” “wood,” and “human, body,” which are materials critical in road scene understanding.

As seen in results on Split-2, our method shows highest accuracies on most of the material classes, where it achieves clear improvements on “metal,” “glass,” “rubber, vinyl,” “plastic,” “plaster,” “ceramic,” “grass,” “wood,” and “human body,” which are common material categories in city road scenes.

Table 2. Statistical properties of KITTI-Materials and other street-view material/semantic segmentation datasets. "img.," "Ann.," "Mat.," and "cls" denote "image," "annotation," "material," and "classes" respectively; "Suburban" denotes "suburban scenes" and "CS samp.," denotes "city-scale sampling;" "✓" means "yes or applicable" while "–" denotes "no or Non-Applicable (N/A);" "KITTI-SS" and "KITTI-Mats" denote KITTI Semantic Segmentation and our KITTI-Materials datasets, respectively.

Dataset	#Imgs	#Img size	#Ann. den	Mat.	#Mat. cls	Suburban	CS samp.	#Ann. pixels
MCubeS	500	1224×1024	$> 99.0\%$	✓	20	Rare	– (≈ 4KM^2)	627M
Cityscapes	5000	2048×1024	97.1%	–	–	✓	✓	9400M
DUS	500	1024×440	63.0%	–	–	Rare	N/A	140M
KITTI-SS	400	1242×375	88.9%	–	–	✓	✓	230M
KITTI-Mats	1,000	1216×320	$> 99.0\%$	✓	20	✓	✓	389M



(a) Split-1

(b) Split-2

Fig. 2. Visual examples of the test sets of (a) Split-1 and (b) Split-2.

Table 3. Per-class comparative results of our models and other methods on KITTI-Materials. "road mk," "fab, lthr," "rubr, vl," "cob," and "hum bd" denote "road marking," "fabric, leather," "rubber, vinyl," "cobblestone," and "human body," respectively. ""DLv3+" and "SegF" denote raw DeepLabv3+ and SegFormer, respectively.

Method	Split	asphalt	concrete	metal	road mk	fab, lthr	glass	plaster	plastic	rubr, vl	sand	gravel	ceramic	cob	brick	grass	wood	leaf	water	hum bd	sky	mean
DLv3+ [2]	1	79.58	29.29	56.74	53.74	34.03	50.55	44.07	30.88	40.51	0	0	41.60	40.53	26.55	**71.50**	30.29	**85.92**	0	20.24	91.03	41.35
SegF [6]	1	82.67	28.47	57.81	58.59	36.46	60.54	48.36	**43.83**	47.09	0	0	**48.28**	51.85	25.54	65.91	31.32	83.70	0	24.38	94.54	44.47
RMSNet	1	**85.14**	**29.58**	**58.66**	**60.65**	**46.69**	**60.75**	**56.12**	42.91	**48.79**	0	0	45.47	**57.62**	**31.25**	69.62	**35.47**	85.31	0	**27.55**	**94.89**	**46.82**
DLv3+ [2]	2	85.66	15.79	60.24	54.33	48.16	62.82	41.95	40.35	41.06	**0.28**	0	53.01	33.63	**50.12**	77.82	35.21	90.42	0	38.32	92.62	46.09
SegF [6]	2	85.08	**22.87**	60.43	56.99	**55.61**	64.86	38.24	42.48	44.72	0	0	54.24	**52.38**	40.60	76.48	38.30	91.03	0	48.22	**93.80**	48.32
RMSNet	2	**86.51**	22.84	**61.81**	**58.51**	52.56	**67.17**	**48.12**	**48.50**	**47.87**	0	0	**60.80**	51.05	47.72	**79.19**	**40.77**	**91.31**	0	**49.10**	92.93	**50.34**

References

1. Abu Alhaija, H., Mustikovela, S.K., Mescheder, L., Geiger, A., Rother, C.: Augmented Reality Meets Computer Vision: Efficient Data Generation for Urban Driving Scenes. IJCV **126**(9), 961–972 (2018)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Proc. ECCV. pp. 833–851 (2018)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: Proc. CVPR. pp. 3213–3223 (2016)
4. Liang, Y., Wakaki, R., Nobuhara, S., Nishino, K.: Multimodal Material Segmentation. In: Proc. CVPR (2022)
5. Scharwächter, T., Enzweiler, M., Franke, U., Roth, S.: Efficient Multi-cue Scene Segmentation. In: German Conference on Pattern Recognition (2013)
6. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: Proc. NeurIPS (2021)