IEEE T-ITS (1524-9050) Supplementary Material

BoxCars: Improving Fine-Grained Recognition of Vehicles using 3D Bounding Boxes in Traffic Surveillance

Jakub Sochor, Jakub Špaňhel, Adam Herout

1 Additional BoxCars116k Dataset Statistics

# tracks	27 496
# samples	116 286
# cameras	137
# make	45
# make & model	341
# make & model & submodel	421
# make & model & submodel & model year	693

	hard	medium
# classes	107	79
# train+val cameras	81	81
# test cameras	56	56
# training tracks	11 653	12 084
# training samples	51 691	54 653
# validation tracks	637	611
# validation samples	2 763	2 802
# test tracks	11 125	11 456
# test samples	39 149	40 842

Table 1: **Left:** Statistics of our new *BoxCars116k* dataset. **Right:** Statistics about splits with different difficulty (*hard* and *medium*).

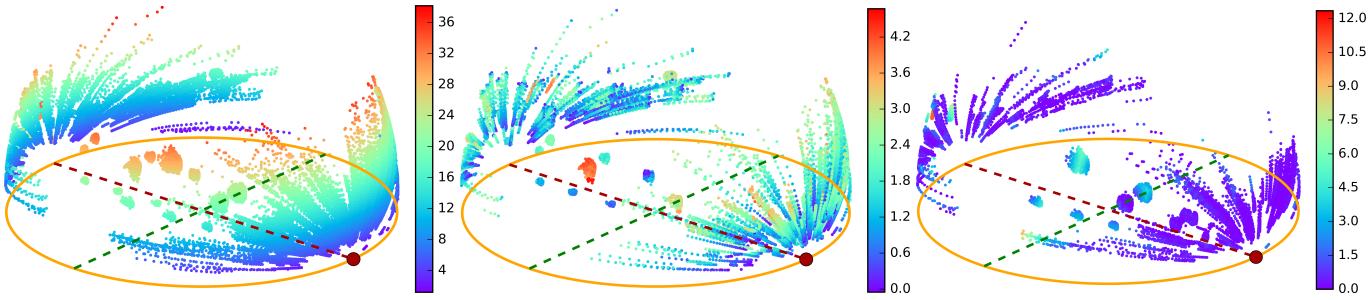

Figure 1: Viewpoints to dataset samples (horizontal flips are not included). Red dot on the unit circle denotes the frontal viewpoint. **Left:** all samples with elevation color coding (in degrees), **center:** training samples for hard split with color coded by 2D BB area (in thousands of pixels), **right:** test samples for hard split color coded by angle to the nearest training viewpoint sample (in degrees).

2 Additional Experimental Data

Due to page limit restrictions, we present some of the raw experimental data and results in this supplementary document.

2.1 Vehicle Types Resisting to Fine-Grained Recognition

net	accuracy [%] all types	merged types
AlexNet + ALL	77.79/88.60	79.08/89.70
VGG16 + ALL	84.13/92.27	85.42/**93.28**
VGG16+CBL + ALL	75.06/83.42	76.82/85.07
VGG19 + ALL	84.12/92.00	**85.51**/92.97
VGG19+CBL + ALL	75.62/83.76	78.56/86.62
ResNet50 + IMAGE	82.27/90.79	83.51/91.79
ResNet101 + IMAGE	83.41/91.59	84.65/92.55
ResNet152 + IMAGE	83.74/91.71	85.10/92.84

Table 2: Comparison of accuracy with all types and 8 merged types into supertypes.

Figure 2: Example of vehicle types merged into one supertype. **Left:** Renault Traffic, **right:** Opel Vivaro.

As possible applications of the fine-grained recognition may vary, we merged pairs of fine-grained classes during testing into one supertype. The merge was done for vehicles which are made by the same concern, have the same dimensions and shape, and which are only differentiated by subtle branding details on the mask. This merge can be beneficial if the task is for example determining the dimensions of the vehicle.
We merged 8 pairs of vehicle types (see Figure 2 for an example) affecting 1 034 tracks and 5 567 image samples. We show the results in Table 2; the accuracy improves only slightly – by ~ 1 percent point.

	AlexNet	VGG16+CBL	VGG19+CBL	VGG16	VGG19	**mean**	**best**
Unpack	+3.47/+4.37	+0.69/+1.06	+1.02/+1.31	+2.07/+2.51	+3.29/+3.48	+2.11/+2.55	+3.47/+4.37
View	−0.96/−1.20	−0.19/−0.19	+0.19/+0.31	−0.46/−0.93	−0.19/+0.26	−0.32/−0.35	+0.19/+0.31
Rast	−0.80/−1.18	+0.30/+0.27	+0.28/+0.72	−0.20/−0.08	+0.28/+0.09	−0.03/−0.04	+0.30/+0.72
Color	+4.80/+3.60	+2.08/+0.97	+2.47/+1.65	+2.72/+1.38	+3.79/+2.55	+3.17/+2.03	+4.80/+3.60
ImageDrop	+0.05/−0.47	+0.29/−0.43	+1.53/+0.96	+0.63/+0.07	+1.00/+0.84	+0.70/+0.20	+1.53/+0.96

Table 3: **Raw data for Table IV of the main document.** Improvements for different nets and modifications computed as $[\textit{base net} + \textit{modification}] - [\textit{base net}]$, where $[\ldots]$ stands for the accuracy of the classifier described by its contents.

	AlexNet	VGG16+CBL	VGG19+CBL	VGG16	VGG19	**mean**	**best**
Unpack	+6.93/+7.60	+2.18/+2.22	+2.06/+2.32	+2.82/+2.46	+3.07/+2.82	+3.41/+3.48	+6.93/+7.60
View	+0.09/+0.18	−0.41/−0.19	−0.78/−0.64	+0.36/+0.15	+0.05/−0.27	−0.14/−0.15	+0.36/+0.18
Rast	+0.22/+0.17	+0.11/−0.08	−0.76/−0.58	+0.30/+0.20	−0.01/−0.11	−0.03/−0.08	+0.30/+0.20
Color	+6.34/+6.18	+2.54/+1.28	+2.21/+1.31	+3.08/+1.73	+2.92/+1.67	+3.42/+2.43	+6.34/+6.18
ImageDrop	+1.07/+0.79	+4.24/+3.54	−0.79/−1.21	+0.89/+0.05	+1.19/+0.68	+1.32/+0.77	+4.24/+3.54

Table 4: **Raw data for Table V of the main document.** Improvements for different nets and modifications computed as [*base net* + *all*] − [*base net* + *all* − *modification*], where [...] stands for the accuracy of the classifier described by its contents.

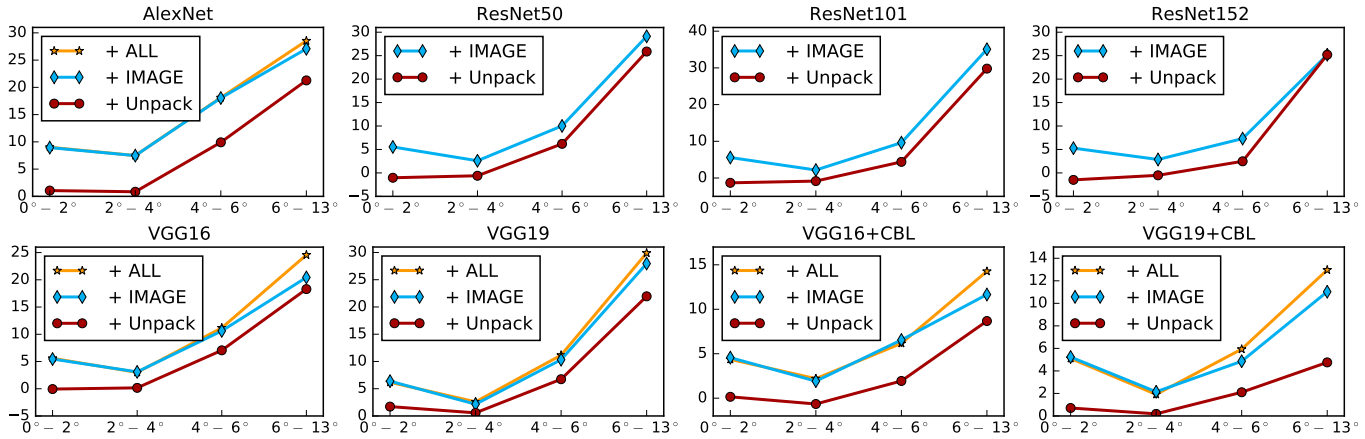

Figure 3: **All results for Figure 10 of the main document.** Correlation of improvement relative to CNNs without modification with respect to train-test viewpoint difference. The x-axis contains bins viewpoint difference bins (in degrees), and the y-axis denotes improvement compared to base net in percent points. The graphs show that with increasing viewpoint difference, the accuracy improvement of our method increases.

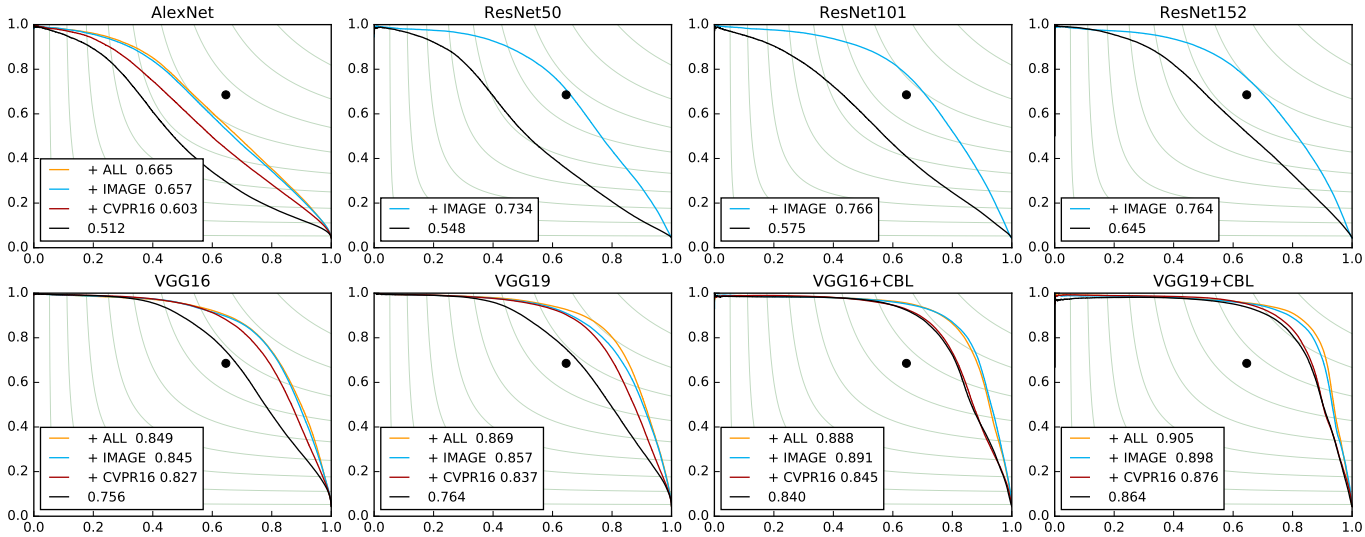

Figure 4: **All results for Figure 12 of the main document.** Precision-Recall curves for verification of fine-grained types. Black dots represent the human performance.

SPLIT: MEDIUM	accuracy [%]	improvement [pp]	error reduction [%]
AlexNet + IMAGE	77.77/88.16	+12.09/+11.64	35.21/49.57
AlexNet + ALL	77.52/87.52	+11.84/+10.99	34.49/46.82
AlexNet + CVPR16	70.90/82.18	+5.23/+5.65	15.22/24.06
AlexNet	65.68/76.53	—	—
VGG16 + ALL	83.89/91.75	+7.93/+6.36	32.99/43.55
VGG16 + IMAGE	83.93/91.69	+7.96/+6.30	33.13/43.13
VGG16 + CVPR16	79.50/88.58	+3.54/+3.19	14.71/21.86
VGG16	75.96/85.39	—	—
VGG16+CBL + IMAGE	75.67/83.49	+4.93/+3.27	16.84/16.55
VGG16+CBL + ALL	75.47/83.23	+4.73/+3.01	16.15/15.23
VGG16+CBL + CVPR16	71.07/81.02	+0.33/+0.80	1.12/4.06
VGG16+CBL	70.74/80.22	—	—
VGG19 + ALL	84.43/92.22	+9.03/+7.88	36.70/50.33
VGG19 + IMAGE	83.98/91.71	+8.58/+7.37	34.88/47.05
VGG19 + CVPR16	80.26/89.39	+4.87/+5.05	19.78/32.27
VGG19	75.40/84.34	—	—
VGG19+CBL + IMAGE	76.88/84.63	+5.34/+3.95	18.75/20.46
VGG19+CBL + ALL	75.47/83.88	+3.92/+3.20	13.79/16.58
VGG19+CBL + CVPR16	72.53/81.90	+0.98/+1.22	3.46/6.32
VGG19+CBL	71.54/80.67	—	—
ResNet50 + IMAGE	82.28/90.63	+7.21/+7.09	28.90/43.08
ResNet50	75.07/83.55	—	—
ResNet101 + IMAGE	83.10/90.80	+6.05/+5.19	26.37/36.08
ResNet101	77.05/85.61	—	—
ResNet152 + IMAGE	83.80/91.38	+5.36/+4.40	24.85/33.78
ResNet152	78.44/86.98	—	—

SPLIT: HARD	accuracy [%]	improvement [pp]	error reduction [%]
AlexNet + ALL	77.79/88.60	+11.15/+10.85	33.42/48.77
AlexNet + IMAGE	77.67/88.28	+11.02/+10.53	33.04/47.31
AlexNet + CVPR16	70.21/81.67	+3.56/+3.92	10.68/17.62
AlexNet	66.65/77.75	—	—
VGG16 + ALL	84.13/92.27	+6.88/+5.56	30.24/41.85
VGG16 + IMAGE	83.79/92.23	+6.53/+5.53	28.71/41.58
VGG16 + CVPR16	79.58/89.27	+2.32/+2.56	10.22/19.27
VGG16	77.26/86.71	—	—
VGG16+CBL + ALL	75.06/83.42	+4.67/+3.31	15.78/16.63
VGG16+CBL + IMAGE	75.04/83.16	+4.66/+3.05	15.73/15.32
VGG16+CBL + CVPR16	70.94/81.08	+0.56/+0.97	1.88/4.88
VGG16+CBL	70.38/80.11	—	—
VGG19 + IMAGE	83.91/92.17	+7.17/+6.11	30.83/43.84
VGG19 + ALL	84.12/92.00	+7.38/+5.94	31.74/42.62
VGG19 + CVPR16	79.69/89.42	+2.95/+3.36	12.69/24.11
VGG19	76.74/86.06	—	—
VGG19+CBL + ALL	75.62/83.76	+4.93/+3.50	16.82/17.71
VGG19+CBL + IMAGE	75.47/83.56	+4.78/+3.30	16.31/16.71
VGG19+CBL + CVPR16	71.92/81.64	+1.23/+1.38	4.20/6.97
VGG19+CBL	70.69/80.26	—	—
ResNet50 + IMAGE	82.27/90.79	+6.79/+6.18	27.69/40.13
ResNet50	75.48/84.61	—	—
ResNet101 + IMAGE	83.41/91.59	+6.95/+6.27	29.52/42.72
ResNet101	76.46/85.31	—	—
ResNet152 + IMAGE	83.74/91.71	+6.06/+5.51	27.16/39.93
ResNet152	77.68/86.20	—	—

Table 5: **Raw data for Table I of the main document.** Improvements of our proposed modifications for different CNNs. The accuracy is reported as single sample accuracy/track accuracy. We also present improvement in percentage points and classification error reduction in the same format.