Automated Model Evaluation for Object Detection via Prediction Consistency and Reliability

Supplementary Material

A. Experimental Details

A.1. Detectors

We evaluate both one-stage and two-stage detection frameworks employing diverse backbones and detection heads, including RetinaNet+R50 [17, 31], RetinaNet+Swin [31, 32], Faster R-CNN+R50 [17, 39], and Faster R-CNN+Swin [32, 39]. The backbones are initialized with pre-trained classification weights, and the models are trained on a $3\times$ schedule (36 epochs). Model training was performed using the MMDetection framework [4], with training conducted via the stochastic gradient descent (SGD) optimizer using an initial learning rate of 10^{-3}, weight decay of 10^{-4}, and momentum of 0.9 [45]. Synchronized SGD was employed over 4 GPUs with a total batch size of 8 images (2 images per GPU). For training, the input images are resized such that the shorter side is at most 800 pixels and the longer side does not exceed 1333 pixels, with horizontal image flipping as the only data augmentation technique. During testing, the input images are resized to a fixed dimension of 800×1333.

A.2. Computational Resources

For all experiments, an Intel(R) Xeon(R) Gold 6226R CPU, clocked at 2.90 GHz, was used. Regarding GPUs, detector training was executed on four NVIDIA RTX A5000 GPUs, whereas testing (including AutoEval experiments) was carried out on a single NVIDIA RTX A5000 GPU with PyTorch 1.13.1 and CUDA 11.6.

A.3. Replication of BoS Results

All BoS [45] results presented in our paper are based on our replication, with modifications made to their publicly released code. We found that the experimental settings described in the paper slightly differ from those implemented in the code, and that certain parts of the experiments are not directly reproducible. Therefore, we modified the code to match the reported performance as closely as possible.

B. Additional Results

B.1. When Test Set is Extremely Difficult

Table B.1 shows the performance on extremely challenging test sets, where mAP is as low as 1–2%. To construct such a test set, we selected the bottom 10% of images (in terms of mAP) from the original test set consisting of 250 images. We evaluate the trained linear regression model not only on the real-world dataset but also on its bottom 10% subset as a test set. This experiment demonstrates that our PCR method enables AutoEval to perform reliably, outperforming other methods even under such extreme conditions.

Meta-dataset	Corruption-based (Ours)				Avg. RMSE
Model	**RetinaNet**		**Faster R-CNN**		
	ResNet-50	Swin-T	ResNet-50	Swin-T	
BoS [45]	15.17	13.55	11.24	13.20	13.29
PCR (Ours)	**9.87**	**10.65**	**10.88**	**10.00**	**10.35**

Table B.1. Performance on the real-world dataset and its bottom 10% subset as a test set in vehicle detection with our corruption-based meta-dataset.

B.2. Impact of Set Size

Figure B.1 reports RMSE under varying sizes of the meta-dataset, sample set, and test set. Increasing the meta-dataset and sample set sizes provides richer information, generally resulting in lower RMSE. For smaller test set size, it remains robust with > 150 samples.

B.3. Hyperparameter Tuning

Confidence Threshold. Figure B.2 (a) analyzes how changing the confidence threshold impacts mAP estimation accuracy. The results presented correspond to the mAP estimation for vehicle detection and pedestrian detection. Based on the confidence threshold, the evaluation determines for each final prediction box whether to assess the consistency between that box and its corresponding pre-NMS predictions or to focus on reliability.

Reliability Measurement. Figure B.2 (b) shows the RMSE for different reliability measurement methods. Reliability measures the trustworthiness of classification and localization by considering all pre-NMS predictions; therefore, relying solely on predictions with high confidence can distort this metric when numerous predictions with low confidence are omitted. Additionally, a standard sigmoid function forces low confidence to produce near-zero values, effectively rendering them negligible in summation (in the case of $\alpha = 0$). To address these issues, our proposed method incorporates all pre-NMS predictions using a function that preserves the influence of those with low confidence, resulting in better performance than the method with $\alpha = 0$. Further, we apply a lower bound of 0.2 to the sigmoid function to prevent the impact of low-confidence score from being lost during the summation.

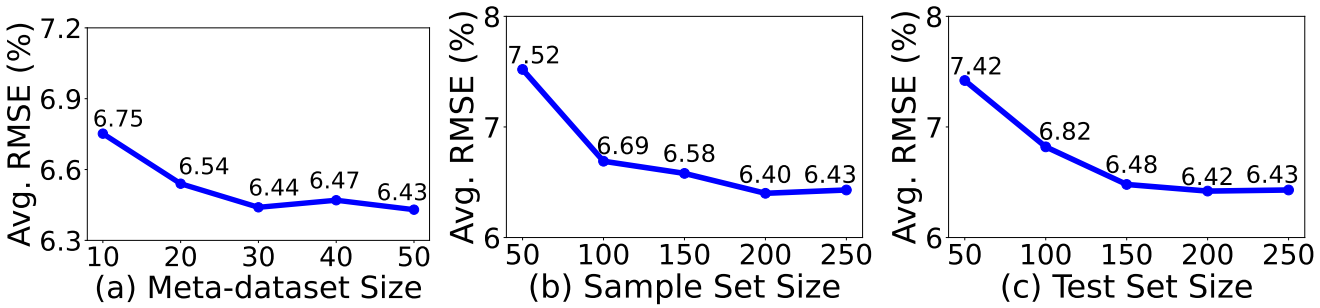

Figure B.1. Effect of set size on vehicle detection with our corruption-based meta-dataset: (a) meta-dataset, (b) sample set, and (c) test set.

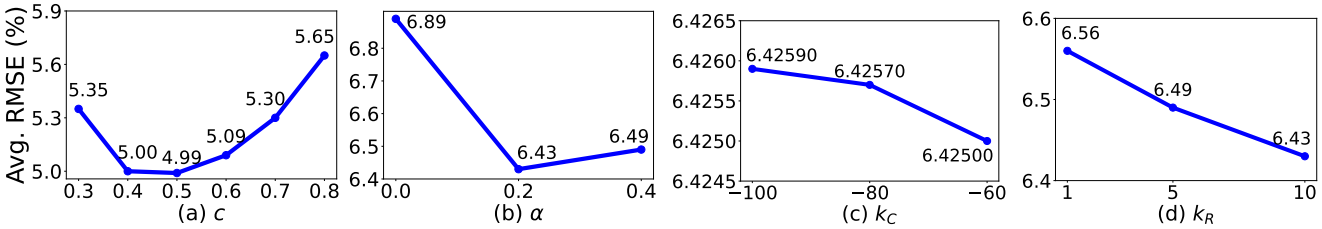

Figure B.2. (a) Effect of confidence threshold c in Eq. (4) and Eq. (5) (b) Effect of α in σ_R. (c–d) Effect of slope parameters k_C and k_R in σ_C and σ_R, respectively.

Slope parameters. We identify the optimal values of k_C and k_R, the slope parameters for the sigmoid functions used in the definitions of consistency and reliability scores, respectively. Figure B.2 (middle and right) shows the experimental results for consistency and reliability, respectively. The role of this function varies between the two scores. For consistency, a large negative scaling factor of k is used to effectively filter out final predictions with low confidence. In contrast, reliability employs a smaller positive scaling factor because it considers all confidence scores of pre-NMS predictions rather than focusing only on high-confidence ones.

B.4. Multi-Class Detection

We conducted experiments not only on single-class detection but also on multi-class detection. The detector is trained on a COCO training set. Also, we evaluate our method on five datasets: Cityscapes, COCO, ExDark, KITTI, and Self-driving. Multi-class detection requires considering the relationships between classes, which necessitates more detailed evaluation metrics for AutoEval. PCR demonstrates stable performance not only for the car and person classes but also for multi-class detection, exceeding BoS [45], the top-performing existing method.

B.5. Vehicle Detection

Tables B.3 to B.10 present the mAP estimation results for vehicle detectors, categorized by the dataset used for testing the mAP estimation model and the detector architecture. Tables B.3 to B.6 present the results of a linear regression model trained on meta-datasets generated using the transformations from BoS [45]. Our proposed PCR method, when evaluated under this setting, achieves performance estimation comparable to existing methods. And Tables B.7 to B.10 display results on meta-datasets constructed using the ImageNet-C [19] transformations. Specifically, when considering the meta-dataset derived from the ImageNet-C transformation, our approach achieved the lowest RMSE—averaged over all datasets—across every evaluated detector architecture. This observation suggests that our method generally exhibits robust estimation performance, but it performs particularly well under realistic transformation conditions, further emphasizing its potential effectiveness in a wide range of practical applications.

Meta-dataset	Corruption-based (Ours)				Avg. RMSE
Model	**RetinaNet**		**Faster R-CNN**		
	ResNet-50	Swin-T	ResNet-50	Swin-T	
BoS [45]	10.04	**5.07**	9.38	**7.34**	7.96
PCR (Ours)	**7.09**	7.06	**8.35**	7.80	**7.58**

Table B.2. Performance for multi-class detection with corruption-based meta-dataset (Ours).

B.6. Pedestrian Detection

Tables B.11 to B.18 report the mAP estimation results for pedestrian detectors, categorized by the dataset used for

Method	COCO 34.6	BDD 32.8	Cityscapes 40.5	DETRAC 40.4	ExDark 28.4	Kitti 42.4	Self-driving 32.0	Roboflow 33.2	Udacity 29.4	Traffic 26.2	Avg. RMSE
PS [20]	1.70	7.67	3.84	12.23	6.47	11.60	2.58	2.18	**0.08**	0.19	4.85
ES [40]	6.39	3.83	15.28	4.25	**0.77**	18.85	4.85	5.30	2.02	3.59	6.51
AC [15]	6.88	5.04	10.55	44.09	2.89	15.37	4.23	2.31	2.02	1.57	9.50
ATC [10]	4.53	7.19	11.48	**0.84**	4.67	15.79	6.06	**0.86**	3.71	**0.05**	5.52
BoS [45]	**1.20**	**0.47**	**2.06**	4.53	4.57	7.71	2.80	2.85	4.06	0.82	3.11
PCR	6.07	1.93	2.85	1.39	2.08	**7.28**	**1.47**	3.00	1.99	2.08	**3.01**

Table B.3. Comparison of AutoEval methods for vehicle detection with augmentation-based meta-dataset from [45]. RetinaNet with ResNet-50 trained on the COCO dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 35.8	BDD 33.5	Cityscapes 40.2	DETRAC 39.9	ExDark 25.8	Kitti 42.1	Self-driving 30.6	Roboflow 27.3	Udacity 28.7	Traffic 25.5	Avg. RMSE
PS [20]	9.29	4.84	5.93	22.51	2.61	17.66	5.05	0.42	2.16	3.38	7.38
ES [40]	10.85	**1.21**	14.04	**4.25**	**0.91**	16.91	1.59	1.19	1.60	3.27	5.58
AC [15]	9.55	3.49	11.59	31.93	1.84	16.28	2.27	0.42	**0.87**	3.49	8.17
ATC [10]	8.85	4.70	11.20	62.18	2.22	15.54	3.06	**0.06**	1.04	**3.13**	11.20
BoS [45]	**2.78**	5.57	**2.45**	8.88	10.29	**4.23**	10.30	13.87	10.59	7.90	7.69
PCR	9.50	3.97	5.70	10.50	2.87	7.80	**1.20**	8.34	1.34	3.47	**5.47**

Table B.4. Comparison of AutoEval methods for vehicle detection with augmentation-based meta-dataset from [45]. RetinaNet with Swin Transformer trained on the COCO dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

testing the mAP estimation model and the detector architecture. Tables B.11 to B.14 resent the results of a linear regression model trained on meta-datasets generated using the transformations from BoS [45]. Our proposed PCR appears to exhibit outstanding performance estimation relative to existing methods in most cases. And Tables B.15 to B.18 display results on meta-datasets constructed using the ImageNet-C [19] transformations. In this setting, our proposed PCR shows superior performance estimation relative to existing methods. Consistent with the vehicle detection case, it recorded the lowest RMSE across all detector architectures.

B.7. Correlation Results

In addition to RMSE, correlation is also an important evaluation metric. The related experimental results are shown in Figure B.3 and Figure B.4, based on meta-dataset built using the ImageNet-C [19] transformations. The correlation is computed between the mAP of each sample set in the meta-dataset and its corresponding AutoEval score. PCR shows a higher correlation than BoS [45] indicating that the applied transformations effectively differentiate characteristics of different sample sets. Therefore, PCR not only achieves the best performance on the test sets but also shows superior consistency across the training meta-dataset.

Method	COCO 36.0	BDD 32.2	Cityscapes 40.5	DETRAC 36.2	ExDark 25.9	Kitti 38.9	Self-driving 29.4	Roboflow 29.4	Udacity 27.8	Traffic 27.7	Avg. RMSE
PS [20]	8.85	5.96	8.48	2.36	2.99	4.14	2.63	8.02	5.01	1.10	4.95
ES [40]	13.57	4.28	12.64	**0.52**	**0.26**	8.67	**0.16**	12.44	3.32	1.58	5.74
AC [15]	11.79	5.63	11.15	7.83	0.50	8.53	1.40	6.47	**2.06**	1.42	5.68
ATC [10]	8.80	7.66	10.60	4.15	2.66	5.63	2.16	3.30	3.64	**0.28**	4.89
BoS [45]	**6.93**	7.23	**5.01**	2.32	4.11	**0.66**	0.95	2.31	3.50	0.30	**3.33**
PCR	7.15	**3.98**	8.36	0.90	3.33	6.40	3.41	**1.00**	4.53	0.97	4.00

Table B.5. Comparison of AutoEval methods for vehicle detection with augmentation-based meta-dataset from [45]. Faster R-CNN with ResNet-50 trained on the COCO dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 38.7	BDD 34.4	Cityscapes 42.6	DETRAC 36.4	ExDark 31.6	Kitti 44.4	Self-driving 31.4	Roboflow 31.9	Udacity 29.2	Traffic 27.3	Avg. RMSE
PS [20]	7.16	4.39	6.06	18.93	2.15	7.42	2.76	4.64	6.95	2.33	6.28
ES [40]	12.75	7.17	9.34	7.27	1.32	8.97	**0.39**	7.58	7.90	**1.88**	6.46
AC [15]	8.34	3.79	8.56	19.68	2.68	11.26	2.45	4.40	**4.49**	4.05	6.97
ATC [10]	7.03	3.63	7.62	23.44	1.61	9.86	3.30	**3.18**	5.10	3.78	6.86
BoS [45]	**0.42**	**1.04**	5.55	**3.65**	2.32	7.40	6.66	13.30	9.19	6.19	5.57
PCR	6.78	1.68	**4.56**	14.31	**0.27**	**7.05**	2.67	5.24	5.03	3.17	**5.08**

Table B.6. Comparison of AutoEval methods for vehicle detection with augmentation-based meta-dataset from [45]. Faster R-CNN with Swin Transformer trained on the COCO dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 34.6	BDD 32.8	Cityscapes 40.5	DETRAC 40.4	ExDark 28.4	Kitti 42.4	Self-driving 32.0	Roboflow 33.2	Udacity 29.4	Traffic 26.2	Avg. RMSE
PS [20]	11.71	14.90	14.48	9.31	**3.97**	21.11	10.68	11.59	8.03	7.18	11.30
ES [40]	14.23	12.59	21.40	31.70	7.24	23.48	11.13	12.83	8.74	5.39	14.87
AC [15]	15.11	13.45	20.87	20.82	7.22	24.02	11.88	12.72	9.78	6.44	14.23
ATC [10]	14.05	14.59	20.65	18.41	6.51	24.19	12.75	11.56	10.78	7.46	14.10
BoS [45]	13.77	12.16	21.94	20.59	7.01	23.06	10.49	12.46	8.16	5.32	13.50
PCR	**10.37**	**6.18**	**8.14**	**6.30**	4.30	**13.20**	**4.98**	**7.96**	**1.48**	**1.34**	**6.43**

Table B.7. Comparison of AutoEval methods for vehicle detection with corruption-based meta-dataset (Ours). RetinaNet with ResNet-50 trained on the COCO dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 35.8	BDD 33.5	Cityscapes 40.2	DETRAC 39.9	ExDark 25.8	Kitti 42.1	Self-driving 30.6	Roboflow 27.3	Udacity 28.7	Traffic 25.5	Avg. RMSE
PS [20]	12.79	7.87	4.23	38.03	**1.11**	21.21	8.71	**4.32**	5.96	0.12	10.44
ES [40]	13.94	**2.91**	16.53	**1.85**	3.54	19.70	3.52	4.51	4.57	0.93	7.20
AC [15]	13.68	6.21	13.87	34.55	2.92	20.12	5.49	4.92	4.68	**0.08**	10.65
ATC [10]	12.76	7.92	13.00	53.19	2.42	19.06	6.70	4.58	4.97	0.55	12.52
BoS [45]	**6.14**	4.02	9.78	8.04	3.17	13.51	**0.12**	4.92	**1.11**	0.97	5.18
PCR	9.67	3.52	**3.09**	8.68	4.62	**5.27**	0.87	7.42	1.73	3.14	**4.80**

Table B.8. Comparison of AutoEval methods for vehicle detection with corruption-based meta-dataset (Ours). RetinaNet with Swin Transformer trained on the COCO dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 36.0	BDD 32.2	Cityscapes 40.5	DETRAC 36.2	ExDark 25.9	Kitti 38.9	Self-driving 29.4	Roboflow 29.4	Udacity 27.8	Traffic 27.7	Avg. RMSE
PS [20]	16.36	13.26	16.39	**9.57**	4.89	13.86	5.17	13.81	2.62	8.15	10.41
ES [40]	19.41	12.04	20.07	12.84	7.02	17.44	7.49	15.59	4.79	8.64	12.53
AC [15]	18.37	13.16	19.15	9.61	6.86	17.37	6.57	12.44	5.64	8.49	11.77
ATC [10]	17.14	15.01	20.02	12.20	6.01	17.21	7.23	11.05	5.55	8.34	11.98
BoS [45]	14.64	13.04	16.96	13.47	3.93	14.51	7.63	6.59	4.61	7.81	10.32
PCR	**11.65**	**5.51**	**13.47**	12.83	**1.92**	**13.20**	**4.49**	**1.94**	**1.99**	**2.00**	**6.90**

Table B.9. Comparison of AutoEval methods for vehicle detection with corruption-based meta-dataset (Ours). Faster R-CNN with ResNet-50 trained on the COCO dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 38.7	BDD 34.4	Cityscapes 42.6	DETRAC 36.4	ExDark 31.6	Kitti 44.4	Self-driving 31.4	Roboflow 31.9	Udacity 29.2	Traffic 27.3	Avg. RMSE
PS [20]	**9.18**	6.82	**4.37**	15.67	5.04	**6.76**	1.24	7.60	6.00	0.33	6.30
ES [40]	16.25	11.24	17.05	8.66	6.83	18.81	5.92	9.95	**0.42**	5.07	10.02
AC [15]	11.05	6.60	7.00	15.33	6.82	12.27	0.68	8.44	2.69	1.43	7.23
ATC [10]	10.08	6.89	8.50	10.29	5.08	11.86	0.50	6.65	2.28	0.72	6.29
BoS [45]	10.63	7.58	15.20	7.97	4.54	17.46	3.12	**0.54**	0.36	1.08	6.85
PCR	9.52	**5.38**	7.61	**0.77**	**2.64**	10.47	**0.28**	6.08	2.22	**0.15**	**4.51**

Table B.10. Comparison of AutoEval methods for vehicle detection with corruption-based meta-dataset (Ours). Faster R-CNN with Swin Transformer trained on the COCO dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 28.3	Caltech 12.0	citypersons 17.0	cityscapes 21.5	crowdhuman 34.2	ECP 24.1	ExDark 27.0	kitti 15.8	self-driving 17.4	Avg. RMSE
PS [20]	6.61	4.46	**0.13**	4.21	35.67	7.62	5.72	0.60	2.24	7.47
ES [40]	10.97	2.53	2.58	**0.34**	11.29	6.04	9.12	0.97	2.36	5.13
AC [15]	7.68	3.34	2.35	**0.34**	17.50	6.59	5.91	0.61	0.99	5.03
ATC [10]	8.68	3.49	0.81	1.84	28.37	7.36	4.76	**0.12**	**0.46**	6.21
BoS [45]	**4.29**	4.59	1.42	1.14	**9.49**	**4.17**	**0.39**	4.44	3.99	3.77
PCR	6.83	**1.27**	0.46	1.40	10.52	5.41	3.87	0.76	2.03	**3.62**

Table B.11. Comparison of AutoEval methods for pedestrian detection with augmentation-based meta-dataset from [45]. RetinaNet with ResNet-50 trained on the CrowdHuman dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 24.6	Caltech 13.2	citypersons 17.9	cityscapes 21.6	crowdhuman 31.0	ECP 24.8	ExDark 22.0	kitti 16.2	self-driving 17.1	Avg. RMSE
PS [20]	6.15	3.52	0.90	2.63	39.55	5.97	4.80	**0.07**	2.54	7.35
ES [40]	6.40	0.77	0.80	0.97	**5.03**	8.45	4.26	1.92	3.39	3.55
AC [15]	5.42	1.76	0.34	1.97	17.46	8.22	3.20	0.08	1.03	4.39
ATC [10]	6.16	3.10	0.46	3.32	41.66	8.49	2.81	0.82	**0.30**	7.46
BoS [45]	**1.49**	4.03	5.29	1.98	5.78	**2.08**	**2.65**	3.34	1.85	3.17
PCR	4.20	**0.36**	**0.30**	**0.03**	7.92	5.82	4.02	2.62	0.82	**2.90**

Table B.12. Comparison of AutoEval methods for pedestrian detection with augmentation-based meta-dataset from [45]. RetinaNet with Swin Transformer trained on the CrowdHuman dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 28.3	Caltech 11.7	citypersons 17.7	cityscapes 22.6	crowdhuman 36.3	ECP 26.1	ExDark 26.0	kitti 18.1	self-driving 15.9	Avg. RMSE
PS [20]	4.44	6.43	1.16	**2.08**	11.06	8.29	1.24	10.56	10.56	6.20
ES [40]	9.82	8.88	1.45	2.92	20.53	7.93	6.28	2.20	13.42	8.16
AC [15]	7.57	8.79	1.43	2.36	17.45	8.91	2.91	**1.22**	10.55	6.80
ATC [10]	8.44	6.97	**0.91**	4.21	18.23	9.76	0.75	6.02	7.51	6.98
BoS [45]	6.80	6.38	1.47	2.59	14.54	7.85	**0.62**	7.14	5.50	5.88
PCR	**4.06**	**0.08**	3.45	2.59	**3.30**	**4.46**	2.72	1.33	**3.08**	**2.79**

Table B.13. Comparison of AutoEval methods for pedestrian detection with augmentation-based meta-dataset from [45]. Faster R-CNN with ResNet-50 trained on the CrowdHuman dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 27.6	Caltech 12.9	citypersons 19.7	cityscapes 24.0	crowdhuman 35.2	ECP 27.3	ExDark 27.7	kitti 20.2	self-driving 17.4	Avg. RMSE
PS [20]	3.75	4.97	**1.26**	3.63	3.73	7.03	1.28	13.60	2.52	4.64
ES [40]	9.76	9.11	1.45	5.04	16.96	8.43	7.53	1.91	1.81	6.89
AC [15]	2.82	5.60	1.34	1.54	9.08	6.98	1.89	1.60	5.80	4.07
ATC [10]	4.08	5.70	1.59	3.99	0.82	7.88	**0.90**	7.01	2.99	3.88
BoS [45]	**2.47**	4.49	2.08	**1.51**	**0.75**	**5.57**	1.64	8.71	4.88	3.57
PCR	3.57	**1.65**	2.34	2.07	3.14	5.76	2.19	**0.18**	**0.52**	**2.38**

Table B.14. Comparison of AutoEval methods for pedestrian detection with augmentation-based meta-dataset from [45]. Faster R-CNN with Swin Transformer trained on the CrowdHuman dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 28.3	Caltech 12.0	citypersons 17.0	cityscapes 21.5	crowdhuman 34.2	ECP 24.1	ExDark 27.0	kitti 15.8	self-driving 17.4	Avg. RMSE
PS [20]	9.05	0.83	4.32	8.69	43.82	10.96	8.42	6.00	**1.27**	10.37
ES [40]	14.76	5.59	1.98	3.96	12.29	9.50	12.89	7.34	7.86	8.46
AC [15]	11.90	3.29	2.66	5.57	11.94	11.04	10.20	6.94	4.02	7.51
ATC [10]	12.30	2.39	3.54	6.33	15.24	11.03	7.99	5.61	3.90	7.59
BoS [45]	16.72	**0.10**	5.10	**9.72**	25.11	12.32	14.50	3.60	4.94	10.23
PCR	**7.13**	0.27	**0.05**	1.19	**6.32**	**7.09**	**5.27**	**2.59**	2.15	**3.56**

Table B.15. Comparison of AutoEval methods for pedestrian detection with corruption-based meta-dataset (Ours). RetinaNet with ResNet-50 trained on the CrowdHuman dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 24.6	Caltech 13.2	citypersons 17.9	cityscapes 21.6	crowdhuman 31.0	ECP 24.8	ExDark 22.0	kitti 16.2	self-driving 17.1	Avg. RMSE
PS [20]	8.23	0.57	1.35	4.52	62.54	7.51	7.44	3.60	**0.65**	10.71
ES [40]	9.94	4.78	3.15	5.34	12.68	11.52	7.52	5.03	6.65	7.40
AC [15]	9.35	3.62	3.88	5.97	**0.97**	12.00	7.12	4.74	5.48	5.90
ATC [10]	10.37	2.28	4.96	7.55	4.86	12.55	6.95	4.09	4.96	6.51
BoS [45]	9.25	**0.13**	2.59	6.27	16.62	9.72	**5.51**	2.00	2.68	6.08
PCR	**4.89**	4.27	**0.12**	**0.74**	11.61	**6.90**	5.56	**0.55**	4.05	**4.30**

Table B.16. Comparison of AutoEval methods for pedestrian detection with corruption-based meta-dataset (Ours). RetinaNet with Swin Transformer trained on the CrowdHuman dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 28.3	Caltech 11.7	citypersons 17.7	cityscapes 22.6	crowdhuman 36.3	ECP 26.1	ExDark 26.0	kitti 18.1	self-driving 15.9	Avg. RMSE
PS [20]	9.66	1.46	2.74	6.24	15.07	11.96	**4.82**	4.62	5.00	6.84
ES [40]	16.97	1.63	5.64	10.73	26.25	14.59	13.95	5.82	0.37	10.66
AC [15]	14.47	2.05	4.48	8.60	20.47	14.24	10.65	5.70	1.71	9.15
ATC [10]	14.30	**0.14**	6.19	9.57	20.29	14.64	7.84	0.76	1.24	8.33
BoS [45]	15.88	**0.14**	6.86	10.30	24.67	14.46	11.39	4.15	2.94	10.09
PCR	**9.56**	1.20	**1.45**	**3.73**	**14.29**	**9.18**	6.14	**0.07**	**0.32**	**5.10**

Table B.17. Comparison of AutoEval methods for pedestrian detection with corruption-based meta-dataset (Ours). Faster R-CNN with ResNet-50 trained on the CrowdHuman dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

Method	COCO 27.6	Caltech 12.9	citypersons 19.7	cityscapes 24.0	crowdhuman 35.2	ECP 27.3	ExDark 27.7	kitti 20.2	self-driving 17.4	Avg. RMSE
PS [20]	6.80	0.91	5.08	6.93	**6.13**	10.11	**4.00**	11.81	1.56	5.93
ES [40]	14.01	2.05	5.24	10.05	22.81	13.31	13.27	6.33	2.92	10.00
AC [15]	7.92	0.26	3.23	6.08	8.36	11.30	7.01	3.29	**0.85**	5.37
ATC [10]	8.49	**0.02**	6.04	8.23	9.24	11.95	5.24	2.63	1.71	5.95
BoS [45]	10.49	0.58	4.25	8.24	17.48	11.86	7.91	1.14	0.91	6.98
PCR	**6.13**	3.42	**1.80**	**2.93**	7.11	**8.56**	4.21	**0.04**	2.44	**4.07**

Table B.18. Comparison of AutoEval methods for pedestrian detection with corruption-based meta-dataset (Ours). Faster R-CNN with Swin Transformer trained on the CrowdHuman dataset is used, and its ground truth mAP (%) is reported for various unseen test sets (as indicated in the header). RMSE is employed to quantify the accuracy of the mAP estimation. The best result in each dataset is highlighted in **bold**, and the second-best is underlined.

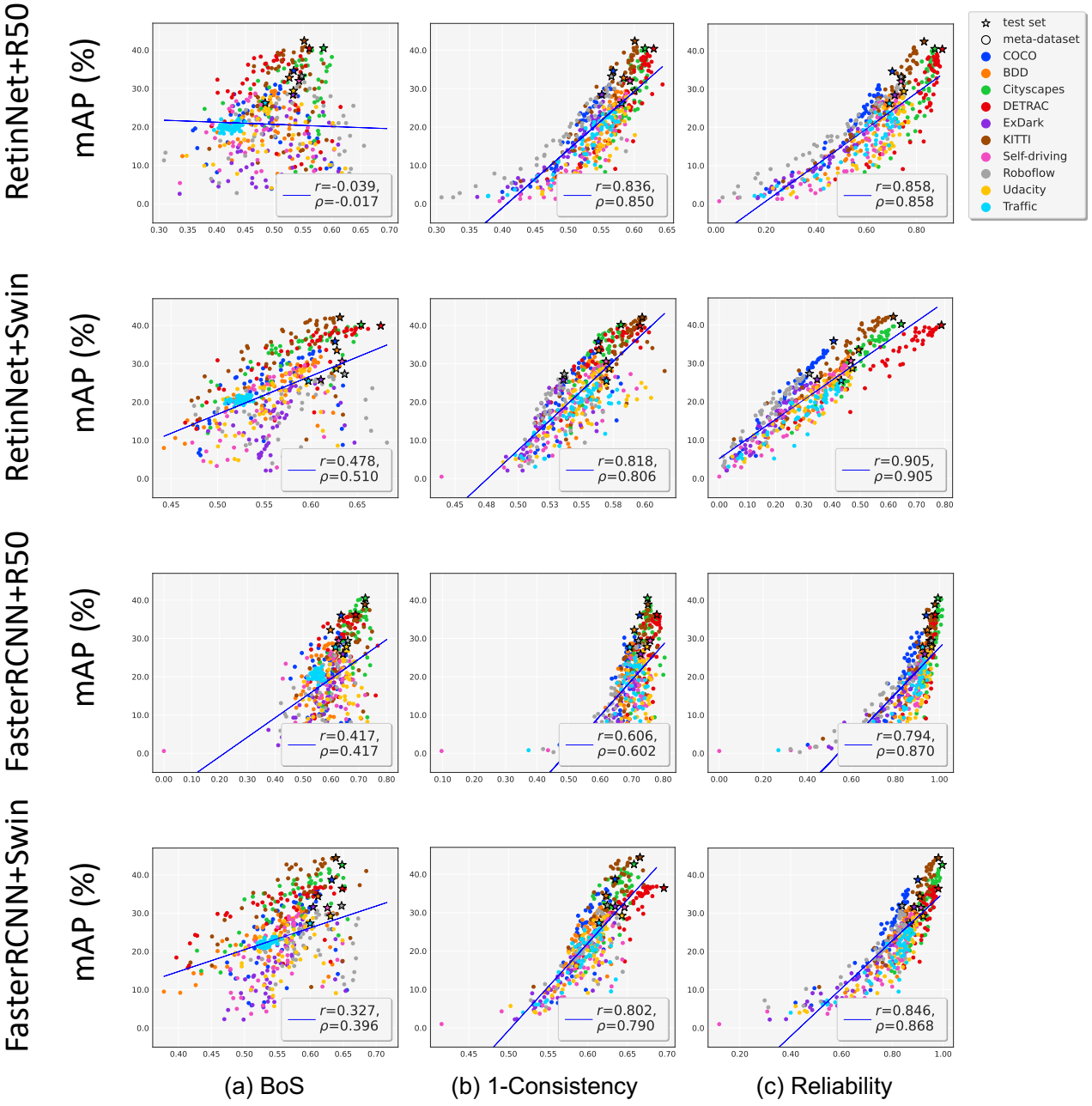

Figure B.3. Correlation results for vehicle detection. Each color represents a different dataset. Markers "⋆" and "○" indicate the test sets (real-world datasets) and the train sets (meta-dataset), respectively. The values in the legend and the blue line show the correlation coefficients between the mAP of the meta-dataset and the AutoEval score (r: Pearson's correlation, ρ: Spearman's rank correlation) and the fitted regression line.

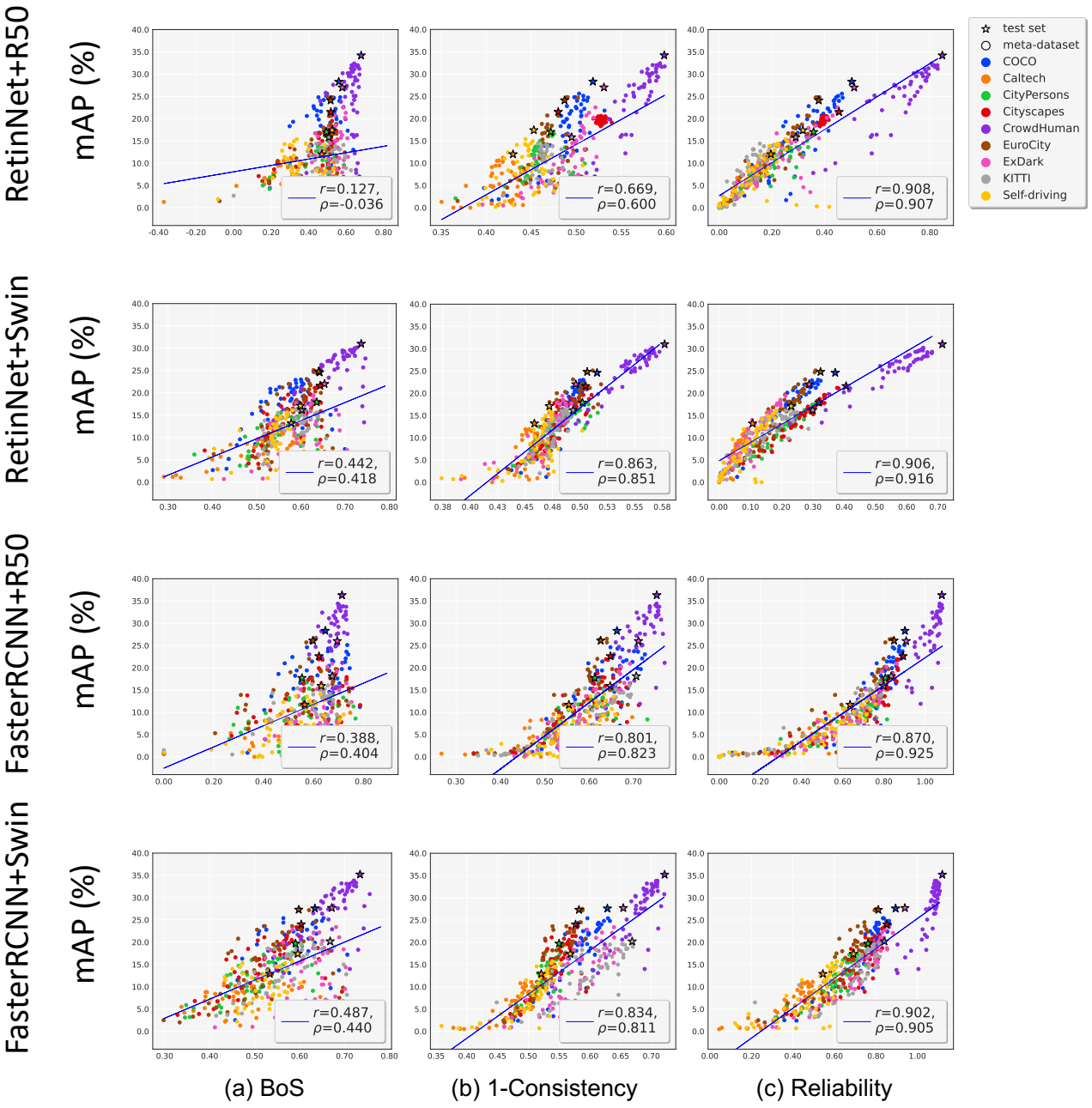

Figure B.4. Correlation results for pedestrian detection. Each color represents a different dataset. Markers "⋆" and "○" indicate the test sets (real-world datasets) and the train sets (meta-dataset), respectively. The values in the legend and the blue line show the correlation coefficients between the mAP of the meta-dataset and the AutoEval score (r: Pearson's correlation, ρ: Spearman's rank correlation) and the fitted regression line.