Perception Characteristics Distance: Measuring Stability and Robustness of Perception System in Dynamic Conditions under a Certain Decision Rule

Supplementary Material

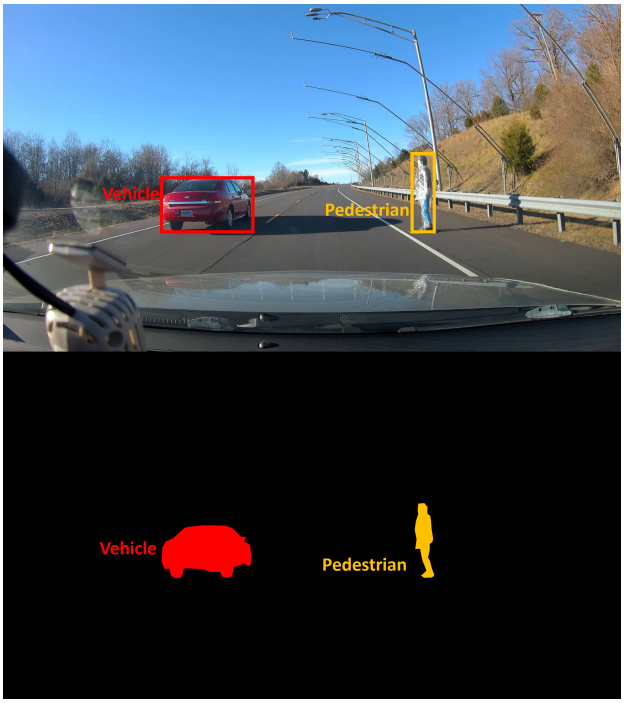

Figure 10. Example of ground-truth bounding box and instance mask under clear daylight (vehicle: 9.751 m, pedestrian: 9.817 m from ego vehicle).

7. SensorRainFall Dataset Description

The SensorRainFall dataset offers a uniquely high-fidelity benchmark for evaluating perception robustness under precipitation. This dataset enables modeling of perception variability across distance, rainfall intensity, and lighting conditions, making it particularly well suited for uncertainty-aware metrics such as PCD.

The SensorRainFall dataset was developed under realistic yet highly controlled conditions. It was collected on the Virginia Smart Roads facility, which supports precise simulation of weather and lighting scenarios. The rainfall was generated using weather simulation towers spaced at 10 meters, producing two nominal rainfall intensities—20 mm/h (heavy rain) and 40 mm/h (excessive rain)—as verified using a Weather Characterizer system based on OTT Parsivel2 laser disdrometry.

To ensure safety and repeatability, target objects were positioned one lane offset from the ego vehicle's path—an approach validated during pilot testing to produce comparable sensor responses to in-lane positioning. Testing was conducted across three ambient lighting settings (daylight, night, and night with overhead 3000 K LED streetlights) and four sensor modalities (camera, lidar, radar, and an ADAS-specific object detection camera), with vehicle speed ranging between 10–55 mph.

Target objects consisted of a red sedan and an ISO 19206-compliant surrogate pedestrian, with the latter wrapped in a plastic poncho to prevent rain damage. Pilot testing confirmed that the poncho created a "halo" effect that improved visibility; thus, it was consistently applied across both rainy and non-rainy runs to isolate rainfall effects. Target positions were recorded using differential GPS (DGPS), and all sensor outputs were time-synchronized using a Robot Operating System (ROS)–based architecture. A summary of the SensorRainFall dataset, along with descriptive statistics of the distance between the ego vehicle and target objects across all environmental conditions, is provided in Table 2.

Table 2. SensorRainfall dataset overview.

	Clear daylight	Rainy daylight	Rainy night	Rainy night under streetlights
Target: Vehicle				
Count of images	278	317	354	281
Mean*	107.989	122.251	119.798	107.761
Std*	61.852	69.800	72.988	61.088
Min*	5.553	5.476	4.412	4.888
Median*	105.976	121.208	114.734	107.240
Max*	214.247	240.669	248.937	214.637
Target: Pedestrian				
Count of images	279	230	352	280
Mean*	107.661	152.615	120.919	108.593
Std*	61.947	55.303	72.654	60.889
Min*	4.442	4.596	4.639	4.501
Median*	105.847	156.005	115.997	107.946
Max*	214.081	240.042	249.275	215.011

**Descriptive statistics of the distance (m) between the ego vehicle and the target.*

Each image in the dataset was carefully examined by trained researchers and manually annotated with ground-truth bounding boxes and pixel-level segmentation masks for both the vehicle and the pedestrian, as illustrated in Fig-

ure 10. The annotations include precise bounding box coordinates (four corners) and detailed instance masks that capture the full spatial extent of each object. These high-resolution annotations enable both object detection and instance segmentation tasks and provide reliable ground truth for evaluating spatial accuracy across perception models.

8. Complete benchmark results

The complete results of all 16 experiments are presented in this section. Tables 3, 4, 5, and 6 report the detected variance change points at $\alpha = 0.05$ for all benchmarks, including both the count and corresponding distance locations (m). Tables 7, 8, 9, and 10 provide comprehensive evaluation results using the proposed aPCD metric and conventional metrics ($AP_{50:95}$, AP_{50}, AP_{75}, AP_S, AP_M, AP_L, AR, and $F1_{50}$).

9. Benchmark setup and configuration

Deformable DETR employs a ResNet-50 backbone and uses a two-stage refinement mechanism to improve region proposal quality and localization. In our configuration, we build on the deformable-detr-refine_r50 COCO recipe and enable the two-stage pipeline explicitly by setting as_two_stage=True.

Grounding DINO utilizes a Swin-B backbone initialized with a pretraining image size of 384×384, uses 128 embedding dimensions, and follows a $[2, 2, 18, 2]$ depth configuration with a window size of 12 and a drop-path rate of 0.3. The multi-head self-attention stages use $[4, 8, 16, 32]$ heads, matching the Swin-B architecture. The neck is adapted to process the high-dimensional feature maps (256/512/1024 channels) produced by the backbone.

DyHead integrates an ATSS detection head with a Swin Transformer backbone and a hybrid FPN–DyHead neck, following the official COCO training recipe. The model uses a Swin-Large backbone pretrained on ImageNet-22K at a resolution of 384×384, producing multi-scale feature maps with 384, 768, and 1536 channels. These features are fed into a five-level Feature Pyramid Network (FPN) with 256-channel outputs and subsequently refined using a six-block DyHead module that applies dynamic, scale-aware attention to strengthen spatial and semantic feature fusion. The ATSS head operates on 256-channel inputs with a 1×1 prediction kernel, a single-ratio anchor generator, ΔXYWH bounding-box coding, and employs Focal Loss for classification, GIoU Loss for regression, and a centerness term for improved localization. Training follows the ATSSAssigner with top-k=9, and uses a data pipeline that includes multi-scale random resizing between 2000×480 and 2000×1200, random horizontal flipping, and dataset repetition to stabilize optimization for the large Transformer backbone. Optimization is performed using AdamW with a learning rate of 5×10^{-5}, along with parameter-wise weight-decay handling for positional embeddings, relative-bias tables, and normalization layers. During inference, images are resized to 2000×1200, and detections are generated using non-maximum suppression with an IoU threshold of 0.6, keeping up to 100 predictions per image.

YOLOX uses a CSPDarknet backbone (deepen = 0.33, widen = 0.50), a YOLOX-PAFPN neck with 128 output channels, and a decoupled YOLOX head trained on COCO with strong augmentation. In our configuration, the network is scaled up substantially: the backbone deepen and widen factors are increased to 1.33 and 1.25, producing a deeper and wider CSPDarknet. The neck is also enlarged, taking 320/640/1280-channel inputs from the backbone and outputting 320-channel features with four CSP blocks. The YOLOX head is widened accordingly to 320 channels.

GLIP features a Swin-S backbone with 192 embedding dimensions, stage depths of $[2, 2, 18, 2]$, and a window size of 12, together with an increased drop-path rate of 0.4 for stronger regularization. The FPN neck is adapted to process multi-scale features of 384, 768, and 1536 channels from the backbone. The ATSS-style bounding box head is enhanced through early fusion of text and visual features and expanded to eight DyHead blocks, with gradient checkpointing enabled to reduce memory consumption. The model is initialized from the publicly available GLIP-L pretrained weights.

Mask R–CNN is configured with a ResNeXt-101–64×4d backbone and a Feature Pyramid Network (FPN) neck. The backbone uses 101 layers, 64 groups, a base width of 4, four stages with outputs at all stages, and frozen parameters in the first stage. Batch normalization is applied with trainable parameters, and the PyTorch-style configuration is used. The backbone is initialized from the official ResNeXt-101–64×4d pretrained checkpoint. Training follows a multi-scale polynomial learning rate schedule and a 3× COCO schedule.

ConvNeXt-V2 is implemented in a Mask R–CNN framework using a ConvNeXt-V2-B backbone pretrained with FCMAE. The backbone employs the `base` configuration with `out_indices=[0,1,2,3]`, a drop-path rate of 0.4, GRN enabled, and initialization from the FCMAE checkpoint. An FPN neck processes feature maps of 128, 256, 512, and 1024 channels. Training uses large-scale jittering at a resolution of 1024×1024, random cropping, and a 36-epoch schedule with linear warm-up and multi-step learning rate decay. Optimization is performed with AdamW and layer-wise learning rate decay, and testing applies NMS in the RPN and soft-NMS in the RCNN.

SOLOv2 is configured with a ResNeXt-101-64×4d backbone in which DCNv2 deformable convolutions are enabled for stages 2–4, and the backbone is initialized from the official pretrained ResNeXt-101 checkpoint. An FPN

Table 3. Detected variance change points ($\alpha = 0.05$) across benchmarks: Object detection, Vehicle.

Model	Number of variance change points	x_{τ_1} (m)	x_{τ_2} (m)	x_{τ_3} (m)	x_{τ_4} (m)	x_{τ_5} (m)
		Clear daylight				
Deformable DETR	0	–	–	–	–	–
Grounding DINO	2	29.407	154.556	–	–	–
DyHead	2	109.377	198.193	–	–	–
YOLOX	**3**	36.513	133.427	170.522	–	–
GLIP	2	81.399	154.556	–	–	–
		Rainy daylight				
Deformable DETR	3	130.054	189.140	198.106	–	–
Grounding DINO	4	81.681	87.833	190.611	217.114	–
DyHead	4	16.037	84.795	195.101	208.407	–
YOLOX	3	119.157	138.339	153.980	–	–
GLIP	3	41.514	110.335	158.641	–	–
		Rainy night				
Deformable DETR	5	31.169	51.627	122.244	211.178	218.465
Grounding DINO	4	51.627	164.393	209.696	215.572	–
DyHead	3	82.943	117.468	211.178	–	–
YOLOX	4	70.742	87.990	104.899	186.902	–
GLIP	2	148.341	240.375	–	–	–
		Rainy night under streetlights				
Deformable DETR	3	46.315	54.084	146.579	–	–
Grounding DINO	3	49.268	107.240	177.282	–	–
DyHead	3	22.784	39.309	148.470	–	–
YOLOX	3	65.933	77.431	165.402	–	–
GLIP	2	92.664	142.363	–	–	–

Table 4. Detected variance change points ($\alpha = 0.05$) across benchmarks: Object detection, Pedestrian.

Model	Number of variance change points	x_{τ_1} (m)	x_{τ_2} (m)	x_{τ_3} (m)
	Clear daylight			
Deformable DETR	3	60.610	133.429	170.443
Grounding DINO	2	57.731	109.034	–
DyHead	3	81.144	109.438	189.473
YOLOX	2	66.391	125.249	–
GLIP	2	67.841	116.894	–
	Rainy daylight			
Deformable DETR	3	71.048	137.862	173.446
Grounding DINO	2	60.182	159.687	–
DyHead	3	66.157	111.520	168.912
YOLOX	3	42.607	128.127	182.582
GLIP	3	107.546	142.961	171.961
	Rainy night			
Deformable DETR	2	49.710	105.067	–
Grounding DINO	2	66.170	109.308	–
DyHead	2	76.553	109.308	–
YOLOX	2	57.243	87.212	–
GLIP	2	76.552	118.000	–
	Rainy night under streetlights			
Deformable DETR	2	83.928	134.777	–
Grounding DINO	2	57.603	121.279	–
DyHead	2	40.958	122.783	–
YOLOX	1	77.980	–	–
GLIP	2	64.637	124.293	–

neck processes multi-scale feature maps for instance segmentation. The mask head employs DCNv2 for all convolutional layers through both the mask feature head and the main mask head configuration. Training follows the multi-scale strategy defined in the base configuration and uses a $3\times$ schedule on the COCO dataset.

Table 5. Detected variance change points ($\alpha = 0.05$) across benchmarks: Instance segmentation, Vehicle.

Model	Number of variance change points	x_{τ_1} (m)	x_{τ_2} (m)	x_{τ_3} (m)	x_{τ_4} (m)
		Clear daylight			
Mask R-CNN	3	33.670	60.446	152.926	–
ConvNeXt-V2	3	75.422	131.801	175.657	–
SOLOv2	2	72.465	121.949	–	–
Mask2Former	4	54.697	69.518	167.761	173.679
RTMDet	3	46.122	150.112	186.139	–
		Rainy daylight			
Mask R-CNN	2	141.458	159.004	–	–
ConvNeXt-V2	2	63.918	153.980	–	–
SOLOv2	3	52.074	141.458	163.257	–
Mask2Former	4	41.514	84.795	187.994	202.521
RTMDet	4	16.037	122.468	158.641	181.621
		Rainy night			
Mask R-CNN	4	70.742	141.67	178.982	199.641
ConvNeXt-V2	3	17.985	88.327	186.902	–
SOLOv2	3	60.461	193.126	213.022	–
Mask2Former	3	62.057	108.764	200.74	–
RTMDet	4	50.379	97.558	191.583	210.083
		Rainy night under streetlights			
Mask R-CNN	2	95.259	157.346	–	–
ConvNeXt-V2	3	91.167	107.612	186.633	–
SOLOv2	3	56.666	74.449	174.194	–
Mask2Former	2	68.520	165.020	–	–
RTMDet	3	95.259	194.742	211.892	–

Table 6. Detected variance change points ($\alpha = 0.05$) across benchmarks: Instance segmentation, Pedestrian.

Model	Number of variance change points	x_{τ_1} (m)	x_{τ_2} (m)	x_{τ_3} (m)
		Clear daylight		
Mask R-CNN	2	84.149	118.716	–
ConvNeXt-V2	2	60.610	116.894	–
SOLOv2	2	53.436	107.423	–
Mask2Former	2	77.037	130.156	–
RTMDet	3	56.314	104.262	172.027
		Rainy daylight		
Mask R-CNN	2	92.918	158.135	–
ConvNeXt-V2	2	73.693	136.679	–
SOLOv2	2	87.209	150.370	–
Mask2Former	3	79.714	113.804	158.135
RTMDet	2	66.157	113.449	–
		Rainy night		
Mask R-CNN	0	–	–	–
ConvNeXt-V2	0	–	–	–
SOLOv2	0	–	–	–
Mask2Former	0	–	–	–
RTMDet	0	–	–	–
		Rainy night under streetlights		
Mask R-CNN	0	–	–	–
ConvNeXt-V2	0	–	–	–
SOLOv2	0	–	–	–
Mask2Former	0	–	–	–
RTMDet	0	–	–	–

Mask2Former is configured with a Swin-S backbone using a patch size of 4, a window size of 7, and stage depths of $[2,\ 2,\ 18,\ 2]$, initialized from the official pretrained weights. The architecture includes a pixel decoder and a transformer decoder head for mask prediction. Training follows a large-scale jittering strategy over 50 epochs with an 8×2 batch

Table 7. Evaluation results of benchmarks: Object detection, Vehicle.

Model	aPCD (m)	Mean of IoU × Confidence score	$AP_{50:95}$	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	$F1_{50}$
			Clear daylight							
Deformable DETR	**104.855**	**0.487**	0.511	**0.947**	0.439	**0.921**	**0.982**	**0.956**	0.552	**0.974**
Grounding DINO	92.685	0.425	**0.553**	0.777	0.594	0.684	**0.982**	**0.956**	**0.568**	0.867
DyHead	70.990	0.337	0.533	0.719	**0.616**	0.585	**0.982**	**0.956**	0.564	0.832
YOLOX	73.105	0.336	0.401	0.548	0.427	0.257	**0.982**	**0.956**	0.452	0.716
GLIP	60.164	0.257	0.421	0.517	0.469	0.310	**0.982**	**0.956**	0.442	0.655
			Rainy daylight							
Deformable DETR	57.224	0.249	0.367	0.583	0.372	0.491	0.777	0.846	0.370	0.739
Grounding DINO	**86.593**	**0.354**	**0.484**	**0.684**	**0.526**	**0.565**	**0.984**	**0.923**	**0.487**	**0.814**
DyHead	66.010	0.274	0.468	0.634	0.533	0.500	**0.984**	0.884	0.471	0.778
YOLOX	52.426	0.208	0.323	0.413	0.353	0.223	0.857	**0.923**	0.326	0.587
GLIP	42.591	0.166	0.288	0.328	0.324	0.083	0.936	**0.923**	0.292	0.497
			Rainy night							
Deformable DETR	3.846	0.021	0.056	0.133	0.04	0.029	0.220	0.645	0.058	0.239
Grounding DINO	29.583	0.156	0.125	0.297	0.105	0.087	0.634	**0.968**	0.128	0.461
DyHead	21.546	0.121	**0.144**	**0.362**	0.107	**0.154**	**0.720**	**0.968**	**0.146**	**0.534**
YOLOX	23.772	0.102	0.106	0.212	0.107	0.017	0.476	**0.968**	0.109	0.353
GLIP	**37.299**	**0.182**	0.133	0.288	**0.119**	0.050	0.707	**0.968**	0.136	0.451
			Rainy night under streetlights							
Deformable DETR	4.829	0.032	0.049	0.121	0.032	0.026	0.197	0.560	0.052	0.222
Grounding DINO	**34.287**	**0.182**	**0.235**	0.505	0.185	0.311	**0.864**	**0.960**	**0.238**	0.675
DyHead	37.115	0.180	0.226	**0.527**	**0.185**	**0.379**	0.848	0.720	0.23	**0.693**
YOLOX	32.213	0.157	0.187	0.391	0.157	0.195	0.712	**0.960**	0.19	0.566
GLIP	34.283	0.145	0.150	0.327	0.128	0.058	0.848	0.920	0.154	0.497

Table 8. Evaluation results of benchmarks: Object detection, Pedestrian.

Model	aPCD (m)	Mean of IoU × Confidence score	$AP_{50:95}$	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	$F1_{50}$
			Clear daylight							
Deformable DETR	25.308	**0.132**	**0.135**	**0.373**	**0.065**	**0.222**	0.875	0.800	**0.138**	**0.547**
Grounding DINO	22.430	0.095	0.071	0.215	0.022	0.028	0.813	0.867	0.074	0.359
DyHead	22.407	0.110	0.093	0.308	0.032	0.130	0.917	0.800	0.096	0.475
YOLOX	13.534	0.049	0.035	0.107	0.010	0.0	0.312	**0.933**	0.038	0.200
GLIP	**27.392**	**0.132**	0.099	0.269	0.036	0.065	**0.979**	0.800	0.102	0.428
			Rainy daylight							
Deformable DETR	18.724	0.094	0.085	0.245	0.041	0.164	0.558	0.474	0.087	0.394
Grounding DINO	12.596	0.063	0.103	0.242	0.079	0.129	0.605	0.842	0.106	0.394
DyHead	20.846	0.098	**0.147**	**0.336**	0.113	**0.195**	0.907	0.842	**0.149**	**0.507**
YOLOX	7.991	0.029	0.042	0.101	0.028	0.039	0.116	0.789	0.044	0.188
GLIP	**34.512**	**0.144**	0.137	0.302	**0.116**	0.145	**0.930**	**0.895**	0.139	0.467
			Rainy night							
Deformable DETR	4.563	0.022	0.043	0.077	0.043	0.0	0.488	0.294	0.046	0.147
Grounding DINO	14.107	0.060	0.105	0.190	0.102	0.048	0.878	**0.882**	0.107	**0.324**
DyHead	10.078	0.044	0.097	0.185	0.08	**0.068**	0.805	0.588	0.099	0.316
YOLOX	6.016	0.023	0.059	0.111	0.057	0.014	0.512	0.706	0.061	0.204
GLIP	**24.178**	**0.099**	**0.106**	**0.190**	**0.108**	0.041	**0.927**	**0.882**	**0.109**	**0.324**
			Rainy night under streetlights							
Deformable DETR	4.129	0.016	0.027	0.046	0.032	0.0	0.208	0.583	0.03	0.095
Grounding DINO	10.167	0.033	0.044	0.071	0.054	0.0	0.333	**0.917**	0.047	0.140
DyHead	5.542	0.023	0.048	**0.082**	0.050	**0.012**	0.417	0.667	0.051	**0.158**
YOLOX	2.962	0.012	0.020	0.039	0.021	0.0	0.083	0.667	0.024	0.082
GLIP	**13.588**	**0.046**	**0.049**	**0.082**	**0.057**	0.0	**0.458**	**0.917**	**0.052**	**0.158**

configuration. Parameter-wise optimization is applied by assigning reduced learning rates to all backbone layers (`lr_mult=0.1`) and zero decay to normalization, positional embedding, query embedding, and level embedding parameters, as defined in the custom parameter rules.

RTMDet is configured with a CSPNeXt-X backbone us-

Table 9. Evaluation results of benchmarks: Instance segmentation, Vehicle.

Model	aPCD (m)	Mean of IoU × Confidence score	$AP_{50:95}$	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	$F1_{50}$
Clear daylight										
Mask R-CNN ok	89.750	0.405	0.376	0.579	0.410	0.444	0.942	**0.947**	0.381	0.736
ConvNeXt-V2	89.454	0.403	0.395	0.553	0.453	0.401	**0.980**	**0.947**	0.399	0.715
SOLOv2	36.572	0.165	0.233	0.276	0.276	0.028	**0.980**	**0.947**	0.237	0.438
Mask2Former	**107.095**	**0.479**	**0.423**	**0.633**	**0.492**	**0.507**	**0.980**	**0.947**	**0.427**	**0.778**
RTMDet	43.471	0.210	0.349	0.593	0.338	0.454	**0.980**	**0.947**	0.353	0.747
Rainy daylight										
Mask R-CNN	67.237	0.250	0.249	0.381	0.246	0.292	0.563	0.894	0.252	0.556
ConvNeXt-V2	78.806	0.316	0.352	0.504	0.410	0.353	**0.981**	**0.947**	0.355	0.673
SOLOv2	24.957	0.099	0.144	0.201	0.157	0.090	0.399	**0.947**	0.147	0.340
Mask2Former	**107.037**	**0.421**	**0.380**	**0.570**	**0.429**	**0.452**	0.927	**0.947**	**0.383**	**0.729**
RTMDet	35.252	0.152	0.212	0.365	0.201	0.185	0.945	0.894	0.215	0.539
Rainy night										
Mask R-CNN	19.096	0.090	0.075	0.153	0.072	0.0	0.500	0.778	0.078	0.269
ConvNeXt-V2	**24.870**	0.128	**0.094**	**0.225**	0.072	**0.026**	**0.827**	0.806	**0.096**	**0.371**
SOLOv2	11.661	0.069	0.090	0.186	0.075	0.0	0.673	0.833	0.092	0.318
Mask2Former	22.170	0.108	0.090	0.186	0.075	0.011	0.673	0.750	0.092	0.318
RTMDet	14.079	0.080	0.089	0.178	**0.081**	0.0	0.577	**0.917**	0.091	0.306
Rainy night under streetlights										
Mask R-CNN	23.587	0.095	0.073	0.152	0.068	0.032	0.500	0.600	0.076	0.268
ConvNeXt-V2	**51.942**	**0.223**	**0.160**	**0.319**	**0.152**	**0.151**	**0.929**	0.800	**0.163**	**0.487**
SOLOv2	12.976	0.062	0.069	0.121	0.071	0.0	0.357	0.767	0.072	0.220
Mask2Former	30.648	0.136	0.104	0.186	0.115	0.012	0.786	0.733	0.106	0.318
RTMDet	21.757	0.098	0.110	0.207	0.121	0.020	0.833	**0.833**	0.113	0.348

Table 10. Evaluation results of benchmarks: Instance segmentation, Pedestrian.

Model	aPCD (m)	Mean of IoU × Confidence score	$AP_{50:95}$	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	$F1_{50}$
Clear daylight										
Mask R-CNN	21.091	0.078	0.048	0.133	0.012	0.094	0.565	0.100	0.050	0.239
ConvNeXt-V2	23.448	0.092	**0.068**	**0.175**	**0.021**	0.101	**0.826**	**0.700**	**0.070**	**0.303**
SOLOv2	9.261	0.039	0.040	0.103	0.015	0.044	0.565	0.600	0.043	0.191
Mask2Former	**23.547**	**0.096**	0.055	0.154	**0.021**	**0.104**	0.652	0.300	0.057	0.272
RTMDet	5.941	0.025	0.033	0.069	0.030	0.017	0.478	0.500	0.035	0.135
Rainy daylight										
Mask R-CNN	2.238	0.010	0.006	0.017	0.003	0.011	0.0	0.091	0.008	0.039
ConvNeXt-V2	**24.364**	**0.108**	**0.079**	**0.212**	**0.020**	**0.147**	**0.538**	**0.818**	**0.082**	**0.354**
SOLOv2	3.479	0.020	0.021	0.056	0.010	0.026	0.115	0.455	0.024	0.112
Mask2Former	15.942	0.072	0.037	0.096	0.017	0.049	0.231	0.727	0.039	0.181
RTMDet	3.162	0.018	0.022	0.060	0.003	0.026	0.269	0.182	0.024	0.118
Rainy night										
Mask R-CNN	0.0	0.001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ConvNeXt-V2	0.0	0.001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SOLOv2	0.0	0.001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mask2Former	0.0	0.002	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RTMDet	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rainy night under streetlights										
Mask R-CNN	0.634	0.017	0.030	0.062	0.031	0.0	0.0	0.417	0.038	0.136
ConvNeXt-V2	2.879	0.028	0.012	0.042	0.0	0.0	0.083	0.167	0.018	0.099
SOLOv2	1.423	0.025	**0.061**	0.094	**0.073**	0.0	**0.250**	0.417	0.070	0.189
Mask2Former	**4.955**	**0.038**	0.030	0.062	0.031	0.0	**0.250**	0.167	0.038	0.136
RTMDet	3.427	0.029	0.059	**0.104**	**0.073**	0.0	0.0	**0.833**	**0.068**	**0.206**

ing a P5 design, an expand ratio of 0.5, and deepen and widen factors of 1.33 and 1.25, respectively, with integrated channel attention. The CSPNeXtPAFPN neck receives feature maps of 320, 640, and 1280 channels and outputs 320-channel features with four CSP blocks. The RTMDetSepBNHead is configured with 320 input and fea-

ture channels and retains the 80-class setting with a DistancePointBBoxCoder. Training uses a base learning rate of 0.002 with a linear warm-up phase for the first 1000 iterations, followed by cosine annealing for the second half of the training schedule.