TABLE I

DATA AUGMENTATION CONFIGURATIONS

Augmentation	Magnitude	Probability
Mosica	-	0.5
Mixup	-	0.5
Random Translation	± 0.1	1.0
Random Scale	± 0.1	1.0
Horizontal Flip	-	0.5

TABLE II

PERFORMANCE COMPARISON OF EVENT REPRESENTATIONS

Method	Representation	AP@50	AP@50:95
YOLOv5s	Event Histogram	<u>0.822</u>	<u>0.690</u>
	Time Surface	0.819	0.682
	Event Volume	**0.850**	**0.711**
YOLOv5m	Event Histogram	<u>0.856</u>	<u>0.727</u>
	Time Surface	0.848	0.715
	Event Volume	**0.861**	**0.728**
YOLOv8s	Event Histogram	**0.833**	**0.701**
	Time Surface	0.813	0.680
	Event Volume	<u>0.831</u>	<u>0.699</u>
YOLOv8m	Event Histogram	<u>0.868</u>	<u>0.737</u>
	Time Surface	0.843	0.714
	Event Volume	**0.871**	**0.743**
YOLO11s	Event Histogram	**0.863**	**0.726**
	Time Surface	0.839	0.708
	Event Volume	<u>0.859</u>	<u>0.725</u>
YOLO11m	Event Histogram	**0.873**	**0.742**
	Time Surface	0.856	0.725
	Event Volume	<u>0.865</u>	<u>0.739</u>
RT-DETR-R18	Event Histogram	**0.826**	**0.680**
	Time Surface	0.801	0.663
	Event Volume	<u>0.809</u>	0.659

Note: Bold indicates the highest score among representations for each model, while underline indicates the second-highest score.

TABLE III
DETECTION PERFORMANCE ON EVENT HISTOGRAM INPUTS

Method	AP@50	AP@50:95	APs
YOLOv5s	0.822	0.690	0.458
YOLOv5m	0.856	0.727	0.508
YOLOv8s	0.833	0.701	0.453
YOLOv8m	0.868	0.737	<u>0.512</u>
YOLO11s	0.863	0.726	0.490
YOLO11m	<u>0.873</u>	<u>0.742</u>	<u>0.512</u>
RT-DETR-R18	<u>0.826</u>	<u>0.680</u>	<u>0.499</u>
PoolFormer-S12	0.768	0.610	0.420
RepViT-M0.9	0.783	0.624	0.447
MobileNetV2	0.768	0.623	0.431
M2Former	0.822	0.663	0.466
RT-DETR-R18 + AAL	0.833	0.694	0.516
RT-DETR-R18 + AAL + Aug1	0.869	0.710	0.543
RT-DETR-R18 + AAL + Aug2	0.883	0.729	0.553
M2Former + AAL	0.826	0.677	0.511
M2Former + AAL + Aug1	0.850	0.713	0.559
M2Former + AAL + Aug2	**<u>0.903</u>**	**<u>0.743</u>**	**<u>0.580</u>**

Note: Underline indicates the best score within each framework, and bold highlights the best score across all methods. AAL denotes the proposed area-aware loss. Aug1 refers to the default data augmentation strategy used in RT-DETR, while Aug2 denotes the improved data augmentation strategy proposed in this work.

TABLE IV

MODEL EFFICIENCY COMPARISON

Method	Params (M)	GFLOPs	Speed (FPS)	Energy (mJ)
YOLOv5s	**9.1**	23.6	**195.1**	21.5
YOLOv5m	20.9	63.7	149.2	67.5
YOLOv8s	11.1	28.6	188.3	**20.3**
YOLOv8m	25.9	78.9	144.9	82.1
YOLO11s	9.4	**21.5**	188.7	28.2
YOLO11m	20.1	68.0	141.9	80.1
RT-DETR-R18	20.1	58.2	98.7	58.1
PoolFormer-S12	20.3	53.7	91.4	89.0
RepViT-M0.9	13.6	37.9	96.4	54.3
MobileNetV2	10.6	28.8	99.5	40.0
M2Former	9.70	27.5	104.3	31.1

Note: Bold denotes the best performance in each column. Inference speed (FPS) and energy consumption (mJ/image) are measured on an NVIDIA RTX 4090 GPU.

TABLE V

ABLATION STUDY ON M^{2}FORMER COMPONENTS

Method	AP@50	AP@50:95
M2Former (baseline)	0.822	0.663
w/o Res2Net	0.795 (-0.027)	0.644 (-0.019)
w/o Spatial Attention	0.809 (-0.013)	0.658 (-0.005)
w/o Channel Attention	0.805 (-0.017)	0.650 (-0.013)
w/o SPD-Conv	0.814 (-0.008)	0.651 (-0.012)

TABLE VI

ABLATION STUDY ON DATA AUGMENTATION

Method	AP@50	AP@50:95
M2Former (baseline)	0.822	0.663
with Mosica	0.851 (+0.029)	0.698 (+0.026)
with Mixup	0.847 (+0.025)	0.688 (+0.025)
with Transformations	0.837 (+0.015)	0.680 (+0.017)

TABLE VII

DETECTION PERFORMANCE ON EVENT HISTOGRAM

UNDER LOWER RESOLUTION INPUT

Method	AP@50	AP@50:95	APs
YOLOv5s	0.719	0.556	0.414
YOLOv5m	<u>0.796</u>	<u>0.617</u>	<u>0.506</u>
YOLOv8s	0.771	0.598	0.482
YOLOv8m	0.790	<u>0.617</u>	0.503
YOLO11s	0.726	0.564	0.448
YOLO11m	0.728	0.572	0.423
RT-DETR-R18	<u>0.733</u>	<u>0.562</u>	<u>0.492</u>
PoolFormer-S12	0.621	0.448	0.392
RepViT-M0.9	0.659	0.487	0.421
MobileNetV2	0.589	0.434	0.376
M2Former	0.699	0.524	0.472
RT-DETR-R18 + AAL	0.752	0.580	0.524
RT-DETR-R18 + AAL + Aug1	0.760	0.593	0.526
RT-DETR-R18 + AAL + Aug2	0.785	0.609	0.534
M2Former + AAL	0.721	0.546	0.488
M2Former + AAL + Aug1	0.744	0.568	0.469
M2Former + AAL + Aug2	**<u>0.809</u>**	**<u>0.622</u>**	**<u>0.546</u>**

Note: Underline indicates the best score within each framework, and bold highlights the best score across all methods. AAL denotes the proposed area-aware loss. Aug1 refers to the default data augmentation strategy used in RT-DETR, while Aug2 denotes the improved data augmentation strategy proposed in this work.

TABLE VIII

ZERO-SHOT DETECTION PERFORMANCE FROM

SYNTHETIC DOMAIN TO REAL DOMAIN

Method	Modality	Light	AP@50	AP@50:95
YOLOv8s	RGB	NE	0.014	0.003
		OE	0.001	0.000
		UE	0.000	0.000
		Average	0.005	0.001
YOLOv8s	Event	NE	0.184	0.117
		OE	0.143	0.104
		UE	0.258	0.139
		Average	0.195	0.120
RT-DETR-R18	Event	NE	0.132	0.059
		OE	0.029	0.009
		UE	0.578	0.286
		Average	0.246	0.118
M2Former	Event	NE	0.332	0.170
		OE	0.137	0.079
		UE	0.624	0.259
		Average	0.364	0.169

Note: NE denotes Normal Exposure, OE denotes Overexposure, and UE denotes Underexposure
lighting conditions. Avg. denotes the average performance across the three lighting conditions.