BMD-45: A Large-Scale CCTV Vehicle Detection Dataset for Urban Traffic in Developing Cities

Supplementary Material

A. Dataset Details

We present additional details about the BMD-45 dataset and its annotation process in this section.

A.1. BMD-45 Vehicle Classes

As mentioned in the main text, we focus on 14 fine-grained vehicle classes that reflect the diversity of India's vehicle fleet. Detailed descriptions of each vehicle class are provided in Table 1 of the Appendix, and cropped samples of the vehicles from each class are shown in Figure 1.

A.2. Additional Dataset Statistics

We provide additional statistics to further characterize the BMD-45 dataset. Figure 2 illustrates the distribution of image timestamps across daylight hours, complementing the temporal properties discussed in the main paper. Figure 3 shows the distribution of bounding-box areas across all vehicle classes, summarizing the range of object scales present in our fixed-camera CCTV views. These plots offer a more comprehensive view of the dataset's diversity and scene characteristics.

B. Disagreement and Image Difficulty

Notation. Let M be the number of detectors/models (or annotators) used for comparison and C the number of classes. For a given image, let $c_{m,i}$ denote the count of bounding boxes of class i predicted by model m ($m \in \{1,\dots,M\}$, $i \in \{1,\dots,C\}$). Define $B_m = \sum_{i=1}^{C} c_{m,i}$ as the total bounding-box count produced by model m. We use i as an image index where required; when ambiguity is possible we write $c_{m,i}^{(img)}$ or D_i for the image-level disagreement score of image i.

B.1. Disagreement Score

The disagreement score captures four complementary aspects of inter-model variability.

- **Per-class Count Disagreement.** For each class i compute the standard deviation of counts across models:

$$\sigma(c_i) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(c_{m,i} - \overline{c_i}\right)^2} \qquad \overline{c_i} = \frac{1}{M}\sum_{m=1}^{M} c_{m,i} \tag{1}$$

  Summing across classes yields the per-image class-count disagreement:

$$N_{dci} = \sum_{i=1}^{C} \sigma(c_i) \tag{2}$$

  This term measures how much the models disagree on counts for each class (e.g., some models see three three-wheelers while others see one).
- **Maximum Pairwise Class-count Disagreements.** For each class i count how many model pairs disagree in their class counts:

$$D_i = \sum_{m=1}^{M-1}\sum_{n=m+1}^{M} \mathbb{I}\left(c_{m,i} \neq c_{n,i}\right) \tag{3}$$

  Then take the worst (maximum) across classes:

$$M_{mdi} = \max_{i\in\{1,\dots,C\}} D_i \tag{4}$$

  M_{mdi} highlights the single class with the largest pairwise disagreement and emphasizes hard, contested categories.

We combine the main components into a compact per-image disagreement score:

$$D_i = N_{dci} + M_{mdi} \tag{5}$$

which balances aggregate count variance with the worst-case per-class pairwise disagreement. To make scores comparable across the dataset, we normalize:

$$D_i^{\text{norm}} = \frac{D_i - D_{\min}}{D_{\max} - D_{\min}} \times 100 \tag{6}$$

where $D_{\min}$ and $D_{\max}$ are the observed minimum and maximum D_i values. All component quantities used in selection (e.g., V_{uci}, V_{nbi}, N_{dci}, M_{mdi}) are retained for analysis and can be inspected individually when diagnosing why a particular image is contentious.

B.2. Difficulty Score

While disagreement measures *inter-model uncertainty* (useful to prioritize images where detectors disagree), it does not by itself quantify visual complexity. To ensure annotator workload was balanced and to construct zone-wise difficulty progression in the gamified challenge, we computed a complementary image-level *difficulty score* that captures intrinsic visual factors (object count, scale, density, overlap) together with model disagreement.

Figure 1. Example cropped images of each of 14 classes in the BMD-45 dataset.

Definitions. For a given image of resolution $H \times W$ with N_{bboxes} detected or annotated boxes $B_1, \ldots, B_{N_{\text{bboxes}}}$ (each box B_j has width w_j and height h_j), we compute the following normalized components:

- **Bounding-box Count:**

$$M_{\text{bbox\_count}} = N_{\text{bboxes}} \tag{7}$$

normalized by a dataset maximum $M_{\text{bb\_max}}$:

$$\tilde{C} = \frac{M_{\text{bbox\_count}}}{M_{\text{bb\_max}}} \in [0, 1] \tag{8}$$

Table 1. Vehicle Classes and Descriptions

Class	Description
Cycle	Non-motorized, manually pedaled vehicles including geared, non-geared, women's, and children's cycles. Bounding boxes include both the vehicle and rider.
2-Wheeler	Motorbikes and scooters for single or double riders. Bounding boxes include both vehicle and rider.
3-Wheeler	(i.e., Auto-rickshaw) Compact vehicles with one front wheel and two rear wheels, featuring a covered passenger cabin.
Hatch-back	Small passenger cars without a protruding rear boot/trunk.
Sedan	Passenger cars with a low-slung design and a separate protruding rear trunk/boot.
MUV	(i.e., Multi-Utility Vehicle) Large vehicles with three seating rows, combining passenger and cargo functionality.
SUV	(i.e., Sport Utility Vehicle) Car-like vehicles with high ground clearance, a sturdy body, and no protruding boot.
Van	Medium-sized vehicles for transporting goods or people, typically with a flat front and sliding side doors. Smaller than Tempo Travellers.
T. Traveller	(i.e., Tempo Traveller) Medium-sized passenger vans with tall roofs and side windows. Larger than vans but smaller than minibuses, with a protruding front.
M. Bus	(i.e., Mini Bus) Shorter, compact buses with fewer seats. Larger than a Tempo Traveller, often featuring a flat front.
Bus	Large passenger vehicles used for public or private transport, including office shuttles and intercity buses.
LCV	(i.e., Light Commercial Vehicle) Lightweight goods carriers used for short to medium distance transport.
Truck	Heavy goods carriers with a front cabin and a rear cargo compartment.
Other	Vehicles not covered in other classes, including agricultural, specialized, or unconventional designs.

- **Average Box Size:** the mean relative area

$$M_{\text{bbox\_size}} = \frac{1}{HW} \cdot \frac{1}{N_{\text{bboxes}}} \sum_{j=1}^{N_{\text{bboxes}}} w_j h_j \tag{9}$$

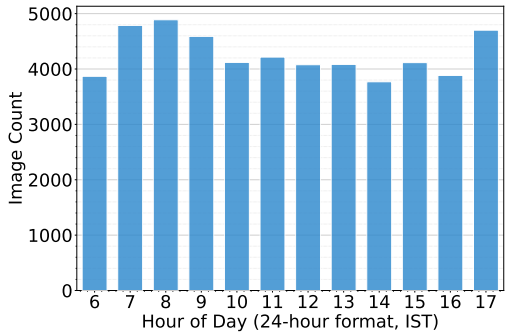


Figure 2. Time of day distribution of images in BMD-45 across 25 days.

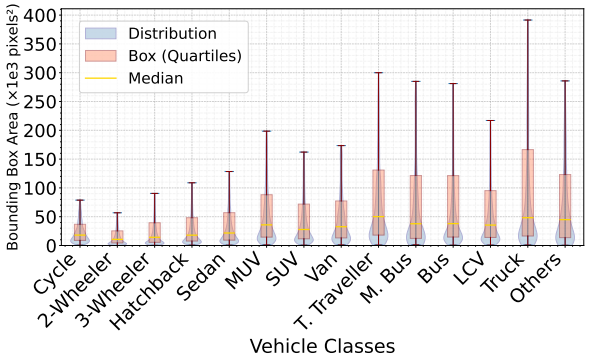


Figure 3. Distribution of bounding box area across all the classes in BMD-45

and we use its complement

$$(1 - M_{\text{bbox\_size}}) \in [0, 1] \tag{10}$$

so that smaller average objects increase difficulty.

- **Bounding-box Density:** total box area fraction

$$M_{\text{bbox\_density}} = \frac{1}{HW} \sum_{j=1}^{N_{\text{bboxes}}} w_j h_j \tag{11}$$

clipped or normalized into $[0, 1]$ (we use $\min(1, M_{\text{bbox\_density}})$)

- **Class Diversity:**

$$M_{\text{class\_count}} = |\{\text{unique classes in image}\}| \tag{12}$$

normalized by the maximum number of classes $M_{\text{max\_classes}}$:

$$\tilde{K} = \frac{M_{\text{class\_count}}}{M_{\text{max\_classes}}} \in [0, 1] \tag{13}$$

- **Average IoU Overlap:** for each unordered box pair (B_p, B_q) define

$$\text{IoU}(B_p, B_q) = \frac{|B_p \cap B_q|}{|B_p \cup B_q|} \tag{14}$$

and the mean pairwise overlap

$$M_{\text{iou\_overlap}} = \frac{1}{N_{\text{pairs}}} \sum_{p<q} \text{IoU}(B_p, B_q) \qquad N_{\text{pairs}} = \binom{N_{\text{bboxes}}}{2} \tag{15}$$

High $M_{\text{iou\_overlap}}$ indicates occlusion and crowding.

- **Model Disagreement (Normalized):** we reuse the disagreement score above and scale it to $[0, 1]$:

$$\widetilde{D}_i = \frac{D_i^{\text{norm}}}{100} \in [0, 1] \tag{16}$$

B.2.1. Composite Difficulty Score

The image difficulty Δ_i is a weighted sum of normalized components:

$$\Delta_i = \tilde{C} \; + \; (1 - M_{\text{bbox\_size}}) \; + \; \widetilde{D}_i \; + \; M_{\text{iou\_overlap}} \tag{17}$$

here all components are pre-normalized to $[0, 1]$. Optionally, one can include $\tilde{K}$ (class diversity) or $M_{\text{bbox\_density}}$ as additional terms if finer control is required. Finally, as with disagreement, Δ_i can be rescaled to $[0, 100]$ for presentation.

C. Crowdsourcing Annotations

Large-scale expert annotation of CCTV imagery is prohibitively expensive; therefore, we adopted a controlled crowdsourcing strategy to obtain high-quality labels at scale. A total of 568 volunteer participants contributed annotations through a custom web platform during a 5-week online challenge.

As described in the main text , to reduce workload and improve consistency, each image was presented with *pre-annotations* produced by the RT-DETRv2-X model fine-tuned on our Gold Dataset. Participants verified and corrected these predictions by adjusting bounding boxes, editing class labels, and adding or removing instances. Images were selected for annotation based on model disagreement scores to focus human effort on diverse and informative cases, while difficulty was varied to mitigate fatigue.

Quality was monitored using *Gold* images inserted uniformly throughout the workflow. Participants were presented a mixture of known ground-truth (gold) and unknown (non-gold) images in randomized order and they were not informed of the presence of the gold images. These gold images provided per-participant accuracy estimates and enabled automatic detection of low-quality submissions.

Each image received between 3 and 9 independent annotations (mean $\approx$ 5), allowing both coverage and redundancy. Annotation assignment ensured that participants did not see repeated images and that the overall distribution balanced load with dataset breadth. As detailed in §**??**, the resulting multi-annotator submissions were subsequently aggregated into a single consensus using IoU-based box matching for localization and majority voting for class labels.

D. Comparison with Other Datasets

D.1. Vehicle Perception Datasets

To provide a comprehensive view of the datasets commonly used in autonomous driving, surveillance, and aerial vehicle perception research, Table 2 summarizes major moving-camera and fixed-camera benchmarks beyond those directly evaluated in the main paper. These datasets remain influential in computer vision, but differ substantially from our problem setting in terms of viewpoint, task formulation, annotation scope, or geographic context. While they are not used for quantitative comparison due to these mismatches, we include them here to acknowledge their importance and to situate our dataset within the broader ecosystem of vehicle perception resources.

D.2. Model Training

All detectors are trained under a unified protocol with model-specific optimizer settings and augmentation pipelines. For reproducibility, we report all hyperparameters, including learning rates, warmup strategies, augmentation policies, and per-model training schedules, in Table 3.

D.3. Cross-Domain Evaluations

D.3.1. IDD vs. BMD-45

We analyze cross-dataset transfer between IDD and BMD-45 over their shared vehicle categories. When evaluating on the IDD validation split, models trained on IDD consistently outperform those trained on BMD-45, as shown in Figure 4b. For example, D-FINE trained on IDD achieves 0.444 mAP@0.50:0.95 on the IDD validation set, whereas the same model trained on BMD-45 attains 0.261; similar gaps appear for other models as well.

In the reverse direction, evaluating on the BMD-45 validation split, models trained on BMD-45 substantially outperform those trained on IDD, as illustrated in Figure 4a. D-FINE improves from 0.468 (trained on IDD) to 0.823 (trained on BMD-45). YOLOv12 variants show the largest differences, with YOLOv12-S rising from 0.174 to 0.613 and YOLOv12-X from 0.189 to 0.341.

Together, these two-way results (Figure 4 and Table 4) highlight the strong viewpoint mismatch between egocentric dashcam footage (IDD) and fixed CCTV imagery (BMD-45), and show that each dataset best supports models trained within its respective domain.

Table 2. Comparison of major vehicle detection and tracking datasets.

Dataset	Venue (Year)	Task†	PoV‡	#Frames	#Annotations	#Veh. classes	#Cameras◇	Location
Moving-camera								
KITTI [9]	CVPR 2012	D, M	E	15k	80k	3	-	DE
Cityscapes [3]	CVPR 2016	S	E	25k	65.4k	8	-	DE
UAV-DT [6]	ECCV 2018	D, M	T	80k	$\approx 841.5k$	3	-	CN
VisDrone [7]	ECCV 2018	D, M	T	262k	2.6M	8	-	CN
IDD [16]	WACV 2019	D, S	E	10k	$111.3k$	9	-	IN
BDD100K [19]	CVPR 2020	D, S, M	E	10k	3.3M	5	-	US
Fixed-camera								
CityCam [20]	CVPR 2017	D	FC	60k	900k	10	212	US
CityFlow [14]	CVPR 2019	D, M	F	117k	230k	9	40	US
UA-DETRAC [18]	CVIU 2020	D, M	F	140k	$1.21M$	4	24*	CN
TrafficCAM [4]	T-ITS 2025	S	FC	4.3k	$\approx 84.2k$	9	NA	IN
BMD-45 (Our)	–	D	FC	45k	$481.9k$	14	3679	IN

†**Task Abbreviations:** D - Detection, S - Segmentation, M - Multi-object tracking. ‡**Point of View Abbreviations:** E - Ego-centric; F - Fixed camera (non-CCTV); FC - Fixed CCTV cameras. *__Camera count:__ UA-DETRAC contains sequences recorded at 24 distinct locations, which we treat as separate fixed-camera viewpoints. ◇**Camera count note:** A value of "–" indicates that the dataset was captured using moving cameras (ego-view or drone), for which the notion of a fixed camera count is not meaningful.

Table 3. Training hyperparameter and architectural settings used for all detectors.

Settings	YOLOv12-S	YOLOv12-X	RT-DETRv2-X	D-FINE-X	RF-DETR-X
Batch Size	16	16	16	16	16
Epochs	100	100	100	100	100
Learning Rate	0.01	0.01	1×10^{-4}	2.5×10^{-4}	1×10^{-4}
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Weight Decay	5×10^{-4}	5×10^{-4}	1×10^{-4}	1.25×10^{-4}	1×10^{-4}
AdamW Betas	(0.937, 0.999)	(0.937, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
LR Policy	Cosine	Cosine	MultiStep	MultiStep	Step LR
Warmup	3 epochs	3 epochs	2000-iteration linear warmup	500-step linear warmup	none
Warmup Details	momentum=0.8; bias LR=0.1	momentum=0.8; bias LR=0.1	momentum untouched; uniform LR ramp	no bias/momentum overrides	warmup disabled
Augmentation Summary	HSV, translate=0.1, scale=0.5, flip=0.5, erase=0.4; no mosaic/mixup	HSV, translate=0.1, scale=0.5, flip=0.5, erase=0.4; no mosaic/mixup	Photometric, ZoomOut, IoU crop; ops disabled after epoch 151	Photometric, ZoomOut, IoU crop, flip, sanitize, resize	Flip + multi-scale RandomResize/Crop + normalize

D.3.2. UA-DETRAC vs. BMD-45

UA-DETRAC exhibits a strong domain mismatch relative to BMD-45, and this asymmetry is clearly reflected in the two-way transfer results (Figure 6 and Table 5). When evaluated on the BMD-45 validation split (Figure 5a), models trained on UA-DETRAC experience a severe drop in accuracy: RT-DETRv2 X falls from 0.838 mAP@0.50:0.95 (trained on BMD-45) to 0.336 (trained on UA-DETRAC), with similar declines for other models as well.

In contrast, the reverse direction, evaluating on the UA-DETRAC test split (Figure 5b), shows a much smaller difference. For RT-DETRv2 X, training on UA-DETRAC yields 0.674 mAP@0.50:0.95, only moderately higher than the 0.578 obtained when trained on BMD-45. Other models also follow this pattern.

This asymmetric behavior indicates that models trained on the structured, low-diversity highway scenes of UA-

Table 4. Cross-dataset evaluation – IDD and BMD-45 on adapted classes

Trained On	Evaluated On	Model	mAP@ 0.50:0.95	mAP@ 0.75	mAP@ 0.50
IDD	IDD	RT-DETRv2 X	0.431	0.449	0.603
		D-FINE X	**0.444**	**0.466**	**0.612**
		RF-DETR X	0.403	0.416	0.580
		YOLOv12 S	0.358	0.376	0.513
		YOLOv12 X	0.352	0.370	0.495
BMD-45	IDD	RT-DETRv2 X	0.258	0.279	0.376
		D-FINE X	0.261	0.280	0.379
		RF-DETR X	0.235	0.250	0.354
		YOLOv12 S	0.141	0.155	0.208
		YOLOv12 X	0.083	0.073	0.165
IDD	BMD-45	RT-DETRv2 X	0.463	0.514	0.585
		D-FINE X	0.468	0.520	0.595
		RF-DETR X	0.416	0.464	0.562
		YOLOv12 S	0.174	0.190	0.260
		YOLOv12 X	0.189	0.210	0.263
BMD-45	BMD-45	RT-DETRv2 X	**0.833**	**0.881**	**0.906**
		D-FINE X	0.823	0.878	**0.906**
		RF-DETR X	0.787	0.856	0.899
		YOLOv12 S	0.613	0.673	0.748
		YOLOv12 X	0.341	0.349	0.554

Table 5. Cross-dataset evaluation – UA-DETRAC and BMD-45 on adapted classes

Trained On	Evaluated On	Model	mAP@ 0.50:0.95	mAP@ 0.75	mAP@ 0.50
UA-DETRAC	UA-DETRAC	RT-DETRv2 X	**0.674**	**0.778**	**0.849**
		D-FINE X	0.649	0.754	0.821
		RF-DETR X	0.656	0.757	0.840
		YOLOv12 S	0.520	0.604	0.718
		YOLOv12 X	0.463	0.528	0.630
BMD-45	UA-DETRAC	RT-DETRv2 X	0.559	0.681	0.798
		D-FINE X	0.577	0.704	0.814
		RF-DETR X	0.578	0.702	0.814
		YOLOv12 S	0.340	0.388	0.540
		YOLOv12 X	0.282	0.291	0.519
UA-DETRAC	BMD-45	RT-DETRv2 X	0.336	0.402	0.445
		D-FINE X	0.284	0.333	0.380
		RF-DETR X	0.232	0.2730	0.308
		YOLOv12 S	0.087	0.099	0.134
		YOLOv12 X	0.087	0.100	0.116
BMD-45	BMD-45	RT-DETRv2 X	**0.838**	**0.881**	**0.902**
		D-FINE X	0.828	0.877	0.900
		RF-DETR X	0.795	0.859	0.893
		YOLOv12 S	0.599	0.656	0.708
		YOLOv12 X	0.301	0.309	0.493

DETRAC generalize poorly to the broader camera viewpoints and richer taxonomy of BMD-45, whereas models trained on BMD-45 retain partial transferability back to UA-DETRAC's narrower domain.

D.3.3. TrafficCAM vs. BMD-45

TrafficCAM is the closest to BMD-45 in viewpoint and geography, but differences in scale, camera diversity, and annotation policy lead to asymmetric transfer performance

Table 6. Cross-dataset evaluation – TrafficCAM and BMD-45 on adapted classes

Trained On	Evaluated On	Model	mAP@ 0.50:0.95	mAP@ 0.75	mAP@ 0.50
TrafficCAM	TrafficCAM	RT-DETRv2 X	0.5214	0.561	**0.724**
		D-FINE X	**0.532**	**0.577**	0.716
		RF-DETR X	0.422	0.463	0.599
		YOLOv12 S	0.394	0.429	0.566
		YOLOv12 X	0.312	0.332	0.439
BMD-45	TrafficCAM	RT-DETRv2 X	0.323	0.355	0.489
		D-FINE X	0.329	0.364	0.496
		RF-DETR X	0.300	0.323	0.470
		YOLOv12 S	0.163	0.161	0.286
		YOLOv12 X	0.133	0.117	0.265
TrafficCAM	BMD-45	RT-DETRv2 X	0.474	0.536	0.626
		D-FINE X	0.469	0.53	0.625
		RF-DETR X	0.418	0.468	0.582
		YOLOv12 S	0.120	0.1290	0.192
		YOLOv12 X	0.122	0.137	0.173
BMD-45	BMD-45	RT-DETRv2 X	**0.798**	**0.855**	0.888
		D-FINE X	0.786	0.849	**0.889**
		RF-DETR X	0.746	0.823	0.880
		YOLOv12 S	0.569	0.625	0.728
		YOLOv12 X	0.356	0.363	0.586

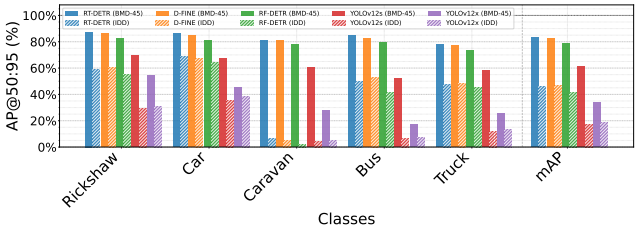
(Figure 6 and Table 6). When evaluated on the BMD-45 validation split (Figure 6a), models trained on TrafficCAM perform noticeably worse than those trained on BMD-45. RT-DETRv2 X drops from 0.798 mAP@0.50:0.95 (trained on BMD-45) to 0.474 (trained on TrafficCAM); other models follow the same trend.

In the reverse direction, evaluating on the TrafficCAM test split (Figure 6b), models trained on TrafficCAM outperform those trained on BMD-45, but the margin is smaller.D-FINE achieves 0.532 mAP@0.50:0.95 when trained on TrafficCAM versus 0.329 when trained on BMD-45; this trend is seen in other models as well.

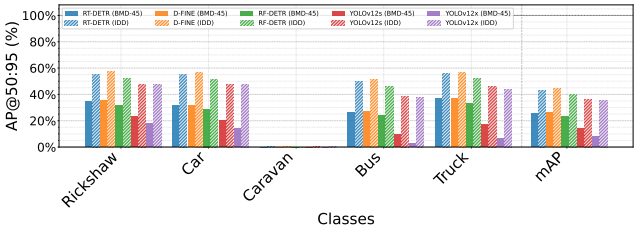
This asymmetry stems from differences in annotation granularity, label policies, and viewpoint diversity. Together, these factors explain why models trained on BMD-45 transfer only partially to TrafficCAM, and likewise why TrafficCAM-trained models underperform on BMD-45, despite the geographic and viewpoint alignment between the two datasets.

D.4. Evaluation on the Held-Out Test Split

Following standard practice in benchmark dataset releases (e.g., COCO, Cityscapes), BMD-45 publicly provides annotations only for the train and validation splits, while testset annotations remain private [3, 12]. Accordingly, all main paper experiments are reported on the validation split. For completeness, we additionally evaluate models trained on the BMD-45 train split on the held-out test split using the private ground truth. As shown in Figure 7, D-FINE achieves 0.712 mAP@0.50:0.95, RF-DETR reaches 0.612, and RT-DETR attains 0.689. YOLOv12 models perform

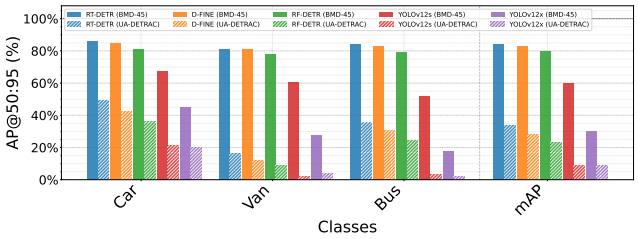


(a) AP@50:95 distribution for selected models trained on BMD-45 dataset and IDD dataset and evaluated on BMD-45 validation split.

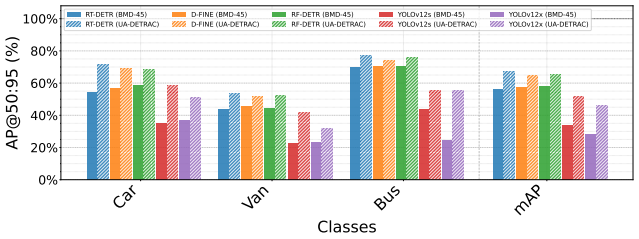


(b) AP@50:95 distribution for selected models trained on BMD-45 dataset and IDD dataset and evaluated on IDD validation split. *The Caravan results are not visible as they are near zero due to a low sample count and inconsistent labels.*

Figure 4. Cross-dataset transfer results between IDD and BMD-45.

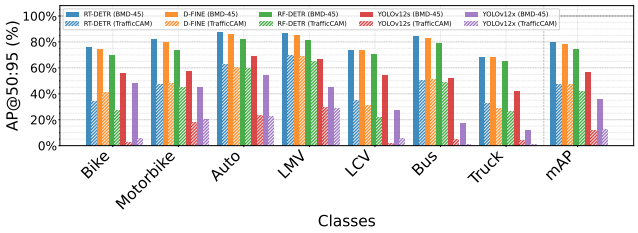


(a) AP@50:95 distribution for selected models trained on BMD-45 dataset and UA-DETRAC dataset and evaluated on BMD-45 validation split.

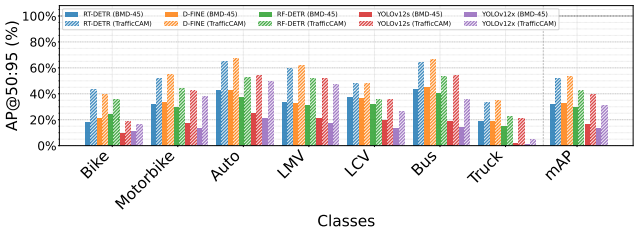


(b) AP@50:95 distribution for selected models trained on BMD-45 dataset and UA-DETRAC dataset and evaluated on UA-DETRAC validation split.

Figure 5. Cross-dataset transfer results between UA-DETRAC and BMD-45.



(a) AP@50:95 distribution for selected models trained on BMD-45 dataset and TrafficCAM dataset and evaluated on BMD-45 validation split.



(b) AP@50:95 distribution for selected models trained on BMD-45 dataset and TrafficCAM dataset and evaluated on TrafficCAM validation split.

Figure 6. Cross-dataset transfer results between TrafficCAM and BMD-45.

Figure 7. AP@50:95 distribution for selected models trained on BMD-45 dataset and evaluated on BMD-45 test set.

comparatively lower, with YOLOv12-S achieving 0.451 and YOLOv12-X reaching 0.31. These results provide a reference for the difficulty of the hidden test set; however, the annotations will not be released to enable future benchmarking protocols.

D.5. Result Stability Across Random Seeds

To verify the stability of our reported results, we re-trained RT-DETRv2-X and D-FINE-X on BMD-45 with 3 independent random seeds and evaluated on the BMD-45 validation set. RT-DETRv2-X achieved a mAP@0.50:0.95 of $73.28 \pm 1.04\%$, and D-FINE-X achieved $72.19 \pm 0.06\%$. These are consistent with the values reported in Figure **??** (72.08% and 72.26%, respectively), confirming that the observed performance differences across architectures are not artifacts of seed selection. RT-DETRv2-X exhibits slightly higher variance across runs, while D-FINE-X remains stable.

E. Privacy and Ethical Considerations

The BMD-45 dataset is derived from fixed-position urban CCTV cameras, which may contain personally identifiable information (PII) such as vehicle license plates, human faces, and on-frame camera metadata. Following established practices in large-scale public computer vision datasets like Street View [8], Cityscapes [2], Waymo Open [17], and Mapillary [13], we apply a structured anonymization protocol before public release. All privacy-preserving transformations were performed *after* the completion of annotation, ensuring that volunteers only interacted with unblurred imagery during labeling to maintain accuracy for small objects and fine-grained vehicle classes. No identifiable annotator information is released with the dataset.

E.1. License Plates

License plates are detected using a YOLO-based one-stage detector trained for road-traffic imagery. Detected regions are blurred with a Gaussian kernel whose size is proportional to the bounding-box dimensions, ensuring complete removal of alphanumeric content across varying viewpoints and resolutions. Multi-scale inference is applied to reliably identify distant or low-resolution plates. This approach maintains local visual appearance while eliminating identifiable text, consistent with redaction strategies widely used in traffic and street-view data [13, 17].

E.2. Faces

Faces are detected using a modern SCRFD-based face detector [10]. To improve robustness under CCTV-specific conditions, like directional lighting, low contrast, and varied skin tones, we apply lightweight pre-processing steps including white-balance correction, contrast-limited adaptive histogram equalization (CLAHE), gamma adjustment, and unsharp masking prior to detection. Multi-scale and tiled inference is used to capture small or partially occluded faces. Each detected region is expanded by 20% and blurred with an adaptive Gaussian kernel, obfuscating identity while preserving overall scene structure. Prior work suggests that such redaction substantially reduces re-identification risk while maintaining utility for downstream scene-understanding tasks [5].

E.3. On-frame Camera Overlays

CCTV feeds occasionally contain textual overlays such as camera IDs, timestamps, and location descriptors. To remove these, we apply an OCR-based redaction pipeline using PP-OCRv3 [11]. Predefined regions (e.g., corners and header bands) are scanned for text; detected polygons are expanded by a small margin to ensure complete coverage. Identified regions are then removed through fast-marching inpainting [15] in OpenCV [1], which reconstructs background content using adjacent pixel information. This removes contextual identifiers without introducing large uniform patches that could bias model training.

E.4. Scope and Limitations

Our anonymization process targets the dominant PII categories present in fixed-view CCTV imagery. While no automated pipeline can guarantee absolute removal of all identifying content, the combination of multi-stage detection, adaptive blurring, and inpainting yields privacy protection comparable to or exceeding that of existing public benchmarks. The final released dataset contains only redacted frames, with all original unblurred images retained offline solely for annotation and evaluation conducted by the authors.

References

[1] Gary Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000. https://docs.opencv.org/. 8

[2] Marius Cordts. Prepare cityscapes dataset (notes on blurred images). https://github.com/mcordts/cityscapesScripts. Cityscapes Scripts, GitHub repository, accessed November 2025. 8

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, Los Alamitos, CA, USA, 2016. IEEE Computer Society. 5, 6

[4] Zhongying Deng, Yanqi Cheng, Lihao Liu, Shujun Wang, Rihuan Ke, Carola-Bibiane Schönlieb, and Angelica I. Aviles-Rivero. Trafficcam: A versatile dataset for traffic flow segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(2):2747–2759, 2025. 5

[5] Julia Dietlmeier, Joseph Antony, Kevin McGuinness, and Noel E. O'Connor. How important are faces for person re-identification? . In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6912–6919, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 8

[6] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, page 375–391, Berlin, Heidelberg, 2018. Springer-Verlag. 5

[7] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, Liefeng Bo, Hailin Shi, Rui Zhu, Aashish Kumar, Aijin Li, Almaz Zinollayev, Anuar Askergaliyev, Arne Schumann, Binjie Mao, Byeongwon Lee, Chang Liu, Changrui Chen, Chunhong Pan, Chunlei Huo, Da Yu, DeChun Cong, Dening Zeng, Dheeraj Reddy Pailla, Di Li, Dong Wang,

Donghyeon Cho, Dongyu Zhang, Furui Bai, George Jose, Guangyu Gao, Guizhong Liu, Haitao Xiong, Hao Qi, Haoran Wang, Heqian Qiu, HongLiang Li, Huchuan Lu, Ildoo Kim, Jaekyum Kim, Jane Shen, Jihoon Lee, Jing Ge, Jingjing Xu, Jingkai Zhou, Jonas Meier, Jun Won Choi, Junhao Hu, Junyi Zhang, Junying Huang, Kaiqi Huang, Keyang Wang, Lars Sommer, Lei Jin, Lei Zhang, Lianghua Huang, Lin Sun, Lucas Steinmann, Meixia Jia, Nuo Xu, Pengyi Zhang, Qiang Chen, Qingxuan Lv, Qiong Liu, Qishang Cheng, Sai Saketh Chennamsetty, Shuhao Chen, Shuo Wei, Srinivas S S Kruthiventi, Sungeun Hong, Sungil Kang, Tong Wu, Tuo Feng, Varghese Alex Kollerathu, Wanqi Li, Wei Dai, Weida Qin, Weiyang Wang, Xiaorui Wang, Xiaoyu Chen, Xin Chen, Xin Sun, Xin Zhang, Xin Zhao, Xindi Zhang, Xinyu Zhang, Xuankun Chen, Xudong Wei, Xuzhang Zhang, Yanchao Li, Yifu Chen, Yu Heng Toh, Yu Zhang, Yu Zhu, Yunxin Zhong, Zexin Wang, Zhikang Wang, Zichen Song, and Ziming Liu. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 213–226, 2019. 5

[8] Andrea Frome, George S. Cheung, Ahmed Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartmut Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in google street view. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009. 8

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 5

[10] Jia Guo, Jiankang Deng, Andreas Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection (scrfd). *arXiv preprint arXiv:2105.04714*, 2021. 8

[11] Chen Li, Wei Liu, Ruoyu Guo, Xiaohui Yin, Kai Jiang, Yuning Du, Yuning Du, Liang Zhu, Bo Lai, Xin Hu, Dian Yu, and Yi Ma. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*, 2022. 8

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 6

[13] Mapillary. Privacy policy and help center: automatic blurring of faces and license plates. https://www.mapillary.com/privacy. Accessed November 2025. See also https://help.mapillary.com/hc/en-us/articles/115001663705-Blurring-images-on-Mapillary. 8

[14] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8798, 2019. 5

[15] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1): 23–34, 2004. 8

[16] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and C.V. Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1743–1751, 2019. 5

[17] Waymo Research. Waymo open dataset faq: "what are you doing to ensure the privacy of people in the images?". https://waymo.com/open/faq/. Accessed November 2025. 8

[18] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193:102907, 2020. 5

[19] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2633–2642. IEEE, 2020. 5

[20] Shanghang Zhang, Guanhang Wu, João P. Costeira, and José M. F. Moura. Understanding traffic density from large-scale web camera data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4264–4273, 2017. 5