TSBOW: Traffic Surveillance Benchmark for Occluded Vehicles Under Various Weather Conditions

Ngoc Doan-Minh Huynh, Duong Nguyen-Ngoc Tran, Long Hoang Pham, Tai Huu-Phuong Tran, Hyung-Joon Jeon, Huy-Hung Nguyen, Duong Khac Vu, Hyung-Min Jeon, Son Hong Phan, Quoc Pham-Nam Ho, Chi Dai Tran, Trinh Le Ba Khanh, Jae Wook Jeon

Automation Lab, Department of Electrical and Computer Engineering
Sungkyunkwan University, Suwon, South Korea
{ngochdm, jwjeon}@skku.edu

Abstract

Global warming has intensified the frequency and severity of extreme weather events, which degrade CCTV signal and video quality while disrupting traffic flow, thereby increasing traffic accident rates. Existing datasets, often limited to light haze, rain, and snow, fail to capture extreme weather conditions. To address this gap, this study introduces the **T**raffic **S**urveillance **B**enchmark for **O**ccluded vehicles under various **W**eather conditions (**TSBOW**), a comprehensive dataset designed to enhance occluded vehicle detection across diverse annual weather scenarios. Comprising over **32 hours** of real-world traffic data from densely populated urban areas, TSBOW includes more than **48,000 manually annotated** and **3.2 million semi-labeled frames**; bounding boxes spanning eight traffic participant classes from large vehicles to micromobility devices and pedestrians. We establish an object detection benchmark for TSBOW, highlighting challenges posed by occlusions and adverse weather. With its varied road types, scales, and viewpoints, TSBOW serves as a critical resource for advancing Intelligent Transportation Systems. Our findings underscore the potential of CCTV-based traffic monitoring, pave the way for new research and applications. The TSBOW dataset is publicly available at the following link.

Code — https://github.com/SKKUAutoLab/TSBOW

Introduction

Climate change has escalated the frequency and intensity of extreme weather events, significantly challenging computer vision tasks by degrading connection and image quality. These conditions disrupt traffic flow, increase traffic congestion and accident rates.

Analyzing traffic flow is essential for understanding traffic surveillance systems, enhancing transportation infrastructure through various applications. Public datasets and benchmarks have significantly advanced machine perception tasks, including image classification (Lu and Weng; Deng et al.), object detection (Zou et al.; Lin et al.), object tracking (Yilmaz, Javed, and Shah; Zhou et al.), semantic segmentation (Hafiz and Bhat; Gupta, Dollár, and Girshick).

Recent developments introduce specialized traffic datasets such as Dataset Quantization (Zhou et al.), compressed large datasets into smaller subsets; PointOdyssey (Zheng et al.), designed for long-term point tracking; TrafficCAM (Deng et al.), focused on traffic flow segmentation; and TUMTraf Video QA (Zhou et al.), targets unified spatio-temporal video understanding.

For object detection task, most existing traffic surveillance benchmarks rely on offline data captured using individual cameras, maintaining video quality in light rain or snow but proving inadequate under extreme weather conditions, such as heavy winds. Established object detection benchmarks, including UAVDT (Yu et al.) and UA-DETRAC (Wen et al.), cover sunny and light rainy conditions but exclude severe weather scenarios.

To advance traffic surveillance research, we introduce the **T**raffic **S**urveillance **B**enchmark for **O**ccluded vehicles under various **W**eather conditions (**TSBOW**), a comprehensive dataset derived from CCTV footage across diverse urban and highway routes. TSBOW encompasses a range of road types—urban streets, standard roads, and boulevards—with objects at fine, medium, and coarse scales, presenting significant challenges for detection models. Spanning a full year, the dataset captures a wide array of weather conditions, from clear skies to heavy snowfall, surpassing existing benchmarks in weather diversity (fig. 1).

Our primary contributions are outlined as follows:

- Development of a semi-automatic iterative annotation pipeline for efficient and accurate labeling (fig. 2).
- Introduction of TSBOW, a novel large-scale traffic surveillance dataset comprising 198 videos and over 3.2 million extracted frames across 145 regions of interest. Collected over four seasons, TSBOW includes diverse weather conditions, notably heavy haze and snow, and covers varied road types, including straight roads, intersections, shared lanes, overpasses, and constructions.
- Compilation of frames from densely populated areas with numerous, often occluded objects, posing significant challenges for object detection. The dataset includes diverse road types hosting eight object categories, including vehicles and pedestrians, with a balanced class distribution. Traffic lights and signs partially obscuring vehicles are also annotated, enhancing the differentiation of

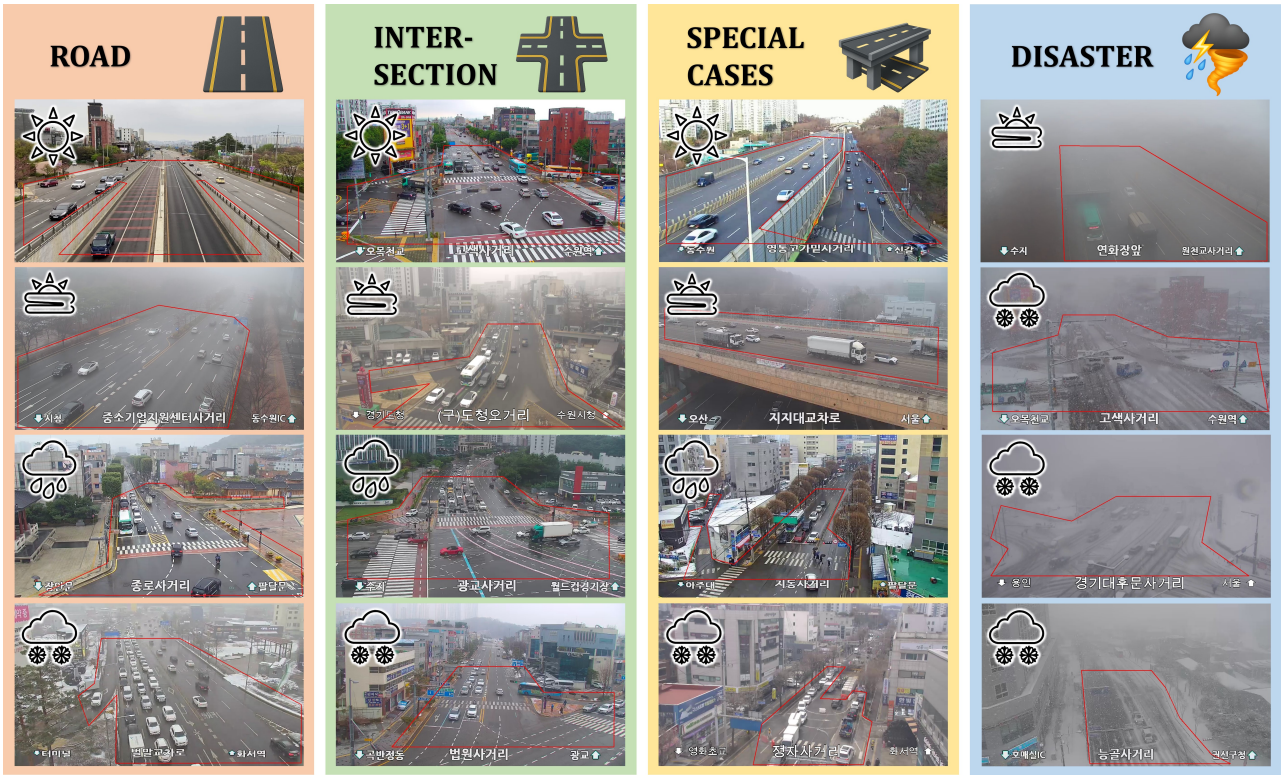

Figure 1: Scenes from the TSBOW dataset, comprising 198 videos recorded across four distinct scenarios spanning all seasons (sunny/cloudy, haze/fog, rain, snow) over a year. The dataset emphasizes adverse weather conditions and densely populated urban areas with heavy traffic, addressing significant challenges in image degradation and vehicle occlusion.

vehicle features from backgrounds.

- Manual annotation and verification of a substantial portion of the data by trained personnel for training, validation, and testing. Bounding boxes were independently labeled, cross-checked for consistency. The SOTA detection model is fine-tuned to annotate remaining frames.
- An object detection baseline for TSBOW provides a benchmark for real-time detection applications.

Related Works

Traffic Surveillance Dataset Traffic surveillance systems depend on high-quality, diverse datasets to optimize performance (Sun et al. 2024; Li et al. 2024; He et al. 2024). Several public traffic datasets support this development, such as Waymo (Sun et al. 2020), which provides 2D and 3D bounding boxes; TrafficMOT (Liu et al. 2024), focused on multi-object tracking; eTram (Verma et al. 2024), offering 2D bounding boxes for event-based cameras; STEP (Weber et al. 2021), providing object segmentation and tracking; and OVT-B (Liang and Han 2024), a benchmark for vocabulary multi-object tracking.

Recent advancements in autonomous driving frequently combine data from color cameras and LiDAR sensors. Color cameras capture visual details, such as color, texture, and semantic information, facilitating the creation of 2D bounding boxes. In contrast, LiDAR generates 3D spatial data, including distance, depth, and point clouds. Notable autonomous driving datasets include Ithaca365 (Diaz-Ruiz et al. 2022) and SODA10M (Han et al. 2021), which provide 3D bounding boxes; HoloVIC (Ma et al. 2024), offering 3D bounding boxes and multi-object tracking; and TAP-Vid (Doersch et al. 2022), which includes object tracking points in videos. Despite their complementary strengths, the integration of camera and LiDAR data presents notable limitations. First, LiDAR's performance is limited by its height and coverage range, particularly when positioned at elevated locations, resulting in unreliable data. Second, LiDAR struggles to detect small objects, such as pedestrians and bicycles, at long distances due to sparse point clouds, which provide limited actionable information. Third, many existing traffic surveillance systems rely exclusively on color cameras, as incorporating LiDAR is often cost-prohibitive and impractical. Consequently, this research leverages existing government CCTV systems to analyze traffic flow across diverse weather conditions over a year, with a particular focus on disasters that significantly disrupt traffic.

The UAVDT dataset (Yu et al. 2020) comprises 10 hours of UAV-captured video across urban areas under sunny and rainy conditions. This dataset presents detection challenges, including water puddle reflections, shadows, and camera motion blur, exacerbated by UAV altitudes ranging from low to high (above 70 meters), rendering it less

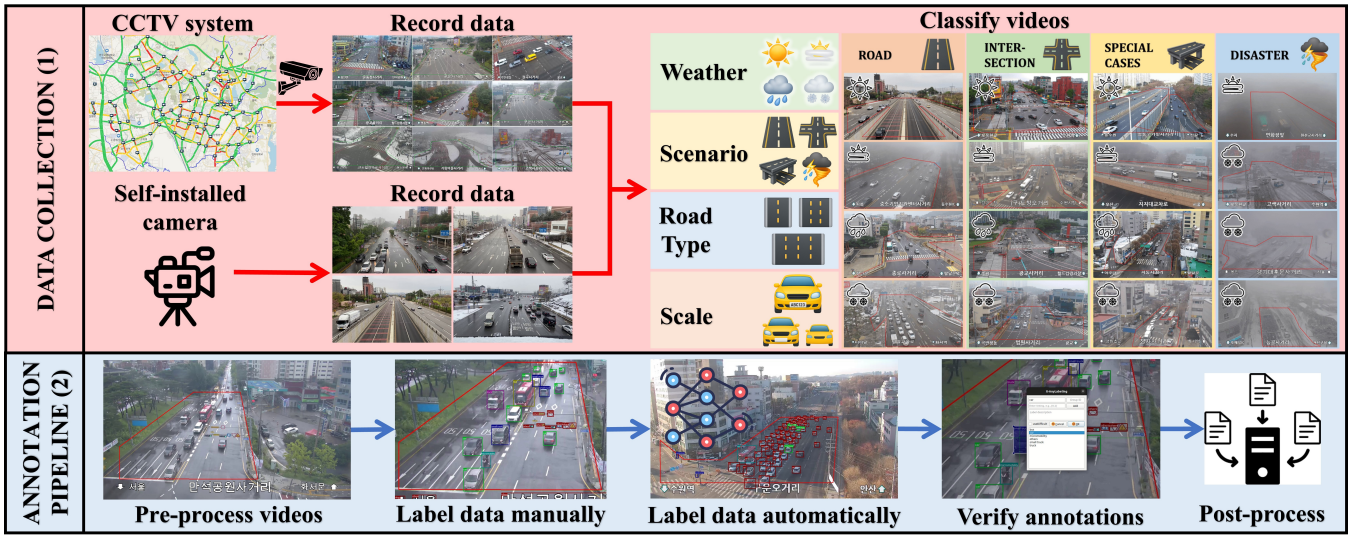

Figure 2: Detailed overview of the data collection and annotation pipeline. The process commences with the recording and categorization of videos during the data collection phase. Subsequently, the videos are preprocessed and allocated to a team of annotators for manual labeling. Next, a state-of-the-art model is fine-tuned to automatically annotate the remaining frames. The resulting annotations are then verified against predefined labeling criteria. Finally, the annotated instances are aggregated and undergo post-processing to finalize the dataset.

Dataset	Camera	Total Duration	Manually Labeled Frames	Total Frames	Labeled Classes	Resolution	FPS			Unique Scenes			
							20	25	30	Sunny	Haze	Rain	Snow
UAVDT	UAVs	10 hrs	40,409	80K	3	1080 × 540	x			42	1	7	
UA-DETRAC	CCTV	10 hrs	–	140K	5	960 × 540		x		44		12	
AAURainSnow	CCTV	1.83 hrs	2,200	132K	-	640 × 480			x		1	5	16
TSBOW (ours)	CCTV	**32.36 hrs**	**48,061**	**3.2M**	**8**	**1280 × 720**	**x**	**x**	**x**	**52**	**15**	**46**	**85**

Table 1: Comparison of traffic surveillance datasets.

applicable to ground-based surveillance like CCTV systems. Conversely, UA-DETRAC (Wen et al. 2020) includes 10 hours of video recorded with a Canon camera (Canon 2010) across 24 locations in China under four weather conditions (cloudy, nighttime, sunny, rainy). While sharing similar challenges—water puddles, shadows, and motion blur—UA-DETRAC's ground-proximate camera setup enhances model training performance but reduces complexity and real-world representativeness compared to UAVDT.

Both UAVDT and UA-DETRAC datasets are limited to sunny and rainy conditions, overlooking snowfall, a critical factor affecting video quality. Addressing this gap, the AAU RainSnow Traffic Surveillance Dataset (Bahnsen and Moeslund 2019) is introduced, captured using both a conventional RGB color camera and a thermal infrared camera. Comprising 22 five-minute videos, this dataset documents rainfall and snowfall across seven intersections in Denmark, providing segmentation for 13,297 objects under four weather conditions: rain, snow, haze, and fog, though it omits bounding boxes for vehicles. The AAU RainSnow dataset exhibits shared challenges with UAVDT and UA-DETRAC, including puddle reflections, raindrops on the lens, and camera variations. Like UA-DETRAC, its ground-proximate camera positioning reduces complexity, limiting its applicability to real-world traffic surveillance systems such as CCTV-based monitoring.

As shown in Tab. 1, compared to others, our dataset features longer video recordings and higher-quality traffic videos with higher-resolution frames. Additionally, it offers greater diversity in FPS, weather conditions, and scenarios compared to other benchmark datasets. Specifically, we include special scenarios and disaster cases that have not been covered in previous datasets. Because of limiting resources, we first focus on different weathers in day time, the night time will be updated in subsequent versions of our dataset.

Object detection (Li et al. 2024) is a machine learning task that involves image localization and object classification. Several high-accuracy object detection models have been developed, such as Faster R-CNN (Girshick 2015), CenterNet (Duan et al. 2019), DETR (DEtection TRansformer) (Carion et al. 2020). However, when balancing processing speed and accuracy, the YOLO family of models emerges as a promising choice due to its exceptional performance and real-time processing capabilities. Various versions of YOLO have been proposed (Hussain 2023; Hidayatullah et al. 2025) are introduced, such as YOLOv3 (Zhao and Li 2020), YOLOv5 (Olorunshola, Irhebhude, and Evwiekpaefe 2023), YOLOv8 (Sohan et al. 2024), YOLOX (He et al. 2023), YOLOv11 (Alkhammash 2025), and YOLOv12 (Alif and Hussain 2025). Compared

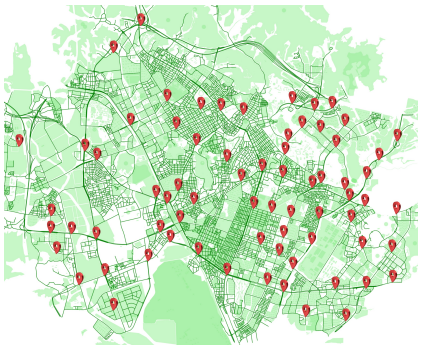

Figure 3: Suwon recording locations in TSBOW dataset.

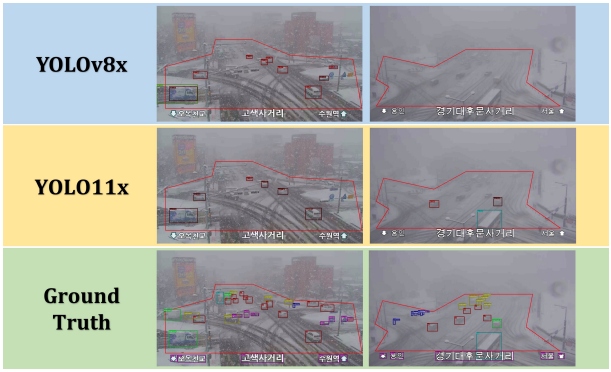

Figure 4: An example of detecting vehicles in heavy snow by using the YOLOv8x and YOLO11x models

to earlier versions, YOLOv8 features an enhanced backbone and employs a path aggregation network, achieving high accuracy while maintaining real-time processing. The YOLOv11 model further integrates vision transformers (ViTs) (Khan et al. 2022) to enhance contextual understanding, yielding the highest accuracy in the mean Average Precision (mAP) metric, albeit with a slight reduction in processing speed. YOLOv12 combines FlashAttention and an R-ELAN backbone to achieve higher accuracy without compromising real-time detection. RT-DETR (Zhao et al. 2024), leveraging a transformer architecture, excels in dense and complex scene understanding. In this study, we employ YOLOv8, YOLOv11, YOLOv12, and RT-DETR, specifically the x-large variants, to establish experimental baselines for the **T**raffic **S**urveillance **B**enchmark for **O**ccluded vehicles under various **W**eather conditions (**TSBOW**).

CCTV Traffic Surveillance Benchmark

The Traffic Surveillance Benchmark for Occluded vehicles under various Weather conditions (TSBOW) dataset is specifically engineered to capture traffic flow within the diverse road networks of Suwon city, Gyeonggi, Korea.

Collection Routes Our dataset is derived from fixed routes comprising a wide range of scenes (fig. 3), enabling robustness evaluation under diverse weather conditions. It is systematically classified according to distinct attributes, including scenario, weather, road type, and scale. Detailed descriptions are provided below.

Figure 5: An example of road types (RT) and scales (S)

First, the video scenarios are categorized into four distinct types—road, intersection, special case, and disaster—as illustrated in fig. 1.

- **Road** comprises straight roads or those where traffic flow remains unaffected by traffic lights.
- **Intersection** includes three subtypes—T, Y, or crossroad—and extends to scenes where traffic lights or pedestrian crossings influence traffic flow, thereby warranting classification as intersections.
- **Special case** includes videos featuring shared lanes, overpasses, or mid-road construction. Shared lanes, including narrow one-way variants, are characterized by bidirectional traffic and pedestrian activity within a single lane, prevalent in space-constrained, densely populated areas. Overpass footage is subdivided into two groups: scenes solely depicting the overpass and those capturing both the overpass and adjacent or underlying roads. The latter poses greater detection challenges due to significant scale disparities among vehicles within a single frame.
- **Disaster** pertains to scenarios where hostile weather severely degrades video quality, such as heavy snow, rendering vehicle identification exceedingly difficult and presenting the most formidable challenge for detection models. Fig. 4 illustrates the object detection outcomes using YOLOv8x and YOLOv11x.

Second, videos for the TSBOW dataset, covering roads, intersections, and special cases, were recorded in Suwon under diverse weather conditions—normal (sunny), haze, rain, and snow—throughout the year. Unlike prior datasets where "rain" videos resemble sunny conditions due to wet roads without visible raindrops, TSBOW classifies "rain" only when raindrops or active rainfall are evident. Similarly, "snow" videos require visible snowflakes or snowfall, altering object appearances (e.g., vehicles with white pixels from snow or frost). These conditions, combined with unstable connections and camera vibrations from strong winds, degrade video quality and complicate computer vision tasks. Videos with wet roads or residual snow lacking active precipitation are classified as "normal" (sunny). In the disaster scenario, heavy haze and extreme snow significantly impair object detection by obscuring visual features, as shown in fig. 4. Snow-covered vehicles blend into white snowy backgrounds, challenging model performance and object-background differentiation. Termed "disaster" due to its severe impact on traffic flow, this scenario exacerbates congestion and accident rates.

Third, the TSBOW dataset classifies data collection zones by the number of straight lanes per direction, excluding turn-

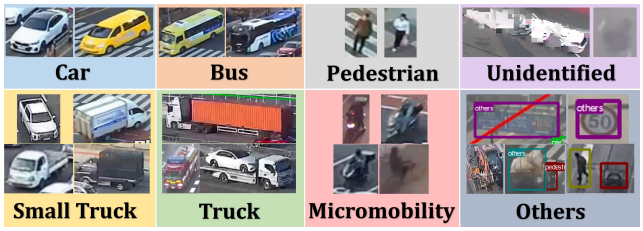

Figure 6: Visualization of annotated instances of different classes in TSBOW dataset.

Scenario/Weather	Normal	Haze	Rain	Snow	Total
Road	12	6	15	28	**61**
Intersection	28	6	25	37	**96**
Special cases	12	2	6	11	**31**
Disaster	–	1	–	9	**10**
Total	**52**	**15**	**46**	**85**	**198**

Table 2: Statistics on scenarios and weathers in TSBOW.

Road Type/Scale	Fine	Medium	Coarse	Total
Urban	14	66	9	**89**
Standard	13	61	4	**78**
Boulevard	2	20	9	**31**
Total	**29**	**147**	**22**	**198**

Table 3: Statistics on scales and road types in the TSBOW.

ing lanes, into three road types: urban (two lanes, primarily cars, small trucks, and pedestrians), standard (four lanes, including larger vehicles like trailer trucks), and boulevard (over six lanes, featuring flatbed trucks and car transporters). Unlike other datasets limited to urban and standard roads, TSBOW includes boulevards, where high vehicle density and occlusion intensify detection challenges due to frequent object overlap.

Final, CCTV cameras along routes and intersections vary in angle and height, producing bounding box sizes categorized into three scales: fine (near road surface, detailed object visualization), medium (elevated cameras, discernible license plates but reduced clarity), and coarse (distant cameras, high vehicle count but partial visibility due to distance and occlusion). Fig. 5 provides an example of road types and scales. While UA-DETRAC covers fine and medium scales, UAVDT focuses on medium and coarse scales.

The diverse attributes of our dataset—spanning locations (fig. 3), weather conditions, road types, and object scales (fig. 5)—render it a robust resource for assessing and enhancing traffic surveillance systems. Furthermore, it incorporates eight labeling classes, surpassing other benchmark datasets in granularity, facilitating detailed classification of vehicles and pedestrians and fostering deeper insights into traffic dynamics for infrastructure improvement. Beyond object detection, the TSBOW dataset supports additional applications, including crowd counting (Li et al. 2021), speed estimation (Fernández Llorca, Hernández Martínez, and García Daza 2021), and object tracking (Soleimanitaleb, Keyvanrad, and Jafari 2019), thereby offering practical utility for real-world traffic management systems.

Labeling Process The annotation process for the TSBOW dataset ensures high-quality ground truth data through five phases: video pre-processing, manual labeling, automatic labeling, annotation verification, and post-processing. Firstly, in pre-processing, regions of interest (ROIs) are defined to capture the main road sections where traffic objects are most visible. After that, frames are then extracted at set intervals and manually annotated using X-Anylabeling (Wang 2023), an open-source tool for precise bounding box creation, focusing on vehicles and pedestrians in high-density urban roads and intersections. Subsequently, a YOLOv12x model trained on region-specific vehicle characteristics (e.g., size, shape, color) in Korea, is used for semi-automatic labeling of remaining frames. Annotations undergo rigorous review and quality control to eliminate substandard entries, ensuring high-quality data. Lastly, the annotated images are compiled and subjected to post-processing to produce the final version of the dataset.

Objects within the TSBOW dataset are classified into eight distinct categories: *car*, *bus*, *truck*, *small truck*, *micromobility*, *pedestrian*, *unidentified*, and *others*. Annotated examples across these categories are depicted in fig. 6.

In the first version of the dataset, which is aimed for public release, 48,061 frames have been manually labeled and verified. The annotations for remaining frame from over 3.2M were generated using the YOLOv12x model. Subsequent versions will substantially increase the proportion of manually annotated frames. License plates and pedestrian faces have been obscured to comply with privacy regulations.

Dataset Statistic and Characteristics This study analyzes data statistics under two conditions: contextual influences and road structures. Tab. 2 details video counts by scenario and weather, with intersections—impacted by traffic lights and pedestrians—exhibiting more occluded objects, prompting increased data collection. Video distribution across scenes remains balanced, comparable to other scenarios. Tab. 3 categorizes videos by scale and road type, with urban and standard roads prevailing over boulevards due to the city-zone focus. CCTV cameras, primarily offering medium-scale perspectives, are strategically placed to capture traffic dynamics, with sufficient fine- and coarse-scale videos to cover diverse scenarios.

As previously noted, our dataset comprises over 3.2 million frames across 198 videos. Tab. 4 details the bounding box counts for each class in three datasets: UAVDT, UA-DETRAC, and our TSBOW. Of the 1.1 million manually annotated objects, cars constitute 69%, with camera information and traffic signs contributing to the proportion of other objects. Unlike benchmark datasets, where cars exceed 83%, our dataset achieves a more balanced distribution across eight classes. The high prevalence of pedestrians and micromobility devices indicates recordings from densely populated urban areas. Frames average 24 objects, with a maximum of 122. The highest number of semi-labeled bounding boxes in a single video reaches 1,233,828 across 17,789 frames, reflecting frequent object occurrences. Bounding boxes are classified by occlusion level, measured as the percentage of area occluded: no occlusion ($<$15% IoU), light occlusion (15–$<$40% IoU), and heavy occlusion ($\geq$40%

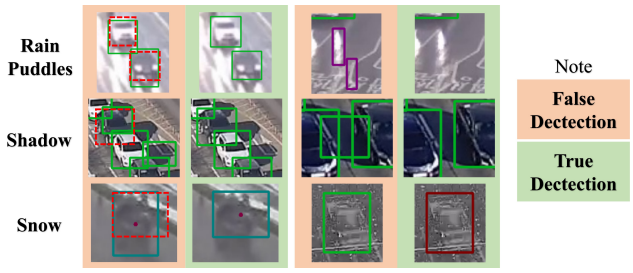

Figure 7: Challenges by Weather Conditions in TSBOW.

Instances	UAVDT Ground Truth	UA-DETRAC Ground Truth	TSBOW Ground Truth	Manually Labeled
Car	756,166	1,052,408	50,269,953	783,203
Bus	0	95,570	2,252,616	35,715
Truck	0	20,641	565,932	10,657
Small Truck	-	105,436	3,258,031	58,272
Pedestrian	-	-	2,935,102	53,414
Micromobility	-	-	1,691,744	29,959
Unidentified	-	-	515,447	8,379
Others	0	-	9,648,223	151,556
Total	**756,166**	**1,274,055**	**71,137,048**	**1,131,155**

Table 4: Instance statistics across UAVDT, UA-DETRAC, and TSBOW datasets.

IoU). Their distribution is 721,684 (no occlusion), 266,420 (light occlusion), and 143,051 (heavy occlusion). Accordingly, traffic flow is categorized into light (44 videos), moderate (98 videos), and heavy (56 videos). This substantial instance count and balanced class distribution enhance the dataset's reliability for traffic surveillance research.

In our TSBOW dataset, numerous factors challenge detection models in accurately identifying and localizing objects, as outlined below:

- Weather conditions: Normal weather scenarios—cloudy and sunny—impact detection performance. Cloudy conditions eliminate vehicle shadows on road surfaces, simplifying detection. In contrast, sunny conditions introduce shadows, leading to less precise bounding boxes that encompass both vehicles and their shadows, reducing model accuracy (fig. 7). Haze degrades image quality, obscuring object features and further complicating detection. Rainy conditions create water puddles that distort bounding box sizes, while strong winds and unstable connections during rain or snow induce camera movement, causing motion blur.
- Scenarios: Continuous traffic flow on roads and overpasses enhances detection model performance due to minimal vehicle occlusions. However, simultaneous recording of overpasses and underlying or adjacent roads complicates detection due to varying vehicle scales and viewing angles. Conversely, traffic lights disrupt flow, causing significant occlusions where only partial objects are visible, posing substantial challenges for models in localizing vehicles and often failing to detect highly occluded instances. Road construction further exacerbates occlusions by reducing lanes, leading to traffic congestion. These issues, compounded by disaster scenarios, underscore the detection difficulties discussed.

Subset	Total	Training	Validation	Test
Manually labeled frames	48,061	10,881	7,559	29,621
Total frames	3,267,598	973,877	658,856	1,634,865

Table 5: Statistics of manually labeled and total video frames across training, validation, and test subsets.

Method	Precision	Recall	mAP50	mAP50–95
YOLOv8x	0.783	0.705	0.733	0.609
YOLO11x	0.786	0.696	0.734	0.614
YOLOv12x	**0.806**	0.662	**0.744**	**0.615**
RT-DETR-x	0.731	**0.740**	0.718	0.552

Table 6: Model performances after training 100 epochs and validating with imgsz=1280 **on manually labeled test set**.

These characteristics collectively impede the precision and reliability of object detection models across diverse environmental and situational contexts. **More detailed descriptions are mentioned in the Supplementary material**.

Experiments

Object Detection Qualitative Result In this section, we establish benchmarks for the TSBOW dataset using neural network-based object detection methods. We utilize YOLOv8x, YOLO11x, YOLOv12x, and RT-DETR-x models, pretrained on the COCO dataset (Lin et al. 2014) and fine-tuned on our dataset at an image resolution of 1280 pixels, as lower resolutions impair inference performance, particularly for occluded vehicles. Evaluation metrics include average precision (AP), mean average precision (mAP), intersection over union (IoU), precision, and recall.

Each video is segmented into three subsets: the first 5 minutes for testing, the next 2 minutes for validation, and the final 3 minutes for training, with frame details provided in Tab. 5. To ensure reliable model performance, manually labeled sets are used for training, validation, and testing. Inference parameters include an IoU threshold of 0.6, an image size of 1280 pixels, and a confidence score of 0.5.

Tab. 6 illustrates the precision, recall, mAP50, and mAP50-95 scores of the YOLOv8x, YOLO11x, YOLOv12x, and RT-DETR-x models after training for 100 epochs. In the evaluation, RT-DETR-x achieves the highest recall, prioritizing broad object coverage, but exhibits lower precision and mAP scores, indicating weaker localization performance. Conversely, YOLOv12x outperforms others in precision, mAP50, and mAP50-95, attributed to its reduced false positive rate. Thus, YOLOv12x demonstrates superior robustness for general object detection, and was selected to annotate the remaining frames.

Datasets Comparison To ensure a fair comparison, we created a subset of medium-scale scenes distinct from the TSBOW dataset, featuring unique road structures and vehicle characteristics. While snowy conditions were recorded in Suwon, additional videos capturing normal, haze, and rain conditions were collected in Seoul (Fig. 8). Unlike UA-DETRAC, which includes only fine- and medium-scale videos captured by a color camera, and UAVDT, which focuses on medium- and coarse-scale drone footage, TSBOW encompasses fine, medium, and coarse scales. Therefore,

Figure 8: Selected scenes for comparison with other datasets

Method	Precision	Recall	mAP50	mAP50-95
YOLOv12x trained on **UAVDT**	0.647	0.141	0.383	0.328
YOLOv12x trained on **UA-DETRAC**	**0.820**	0.295	0.558	0.459
YOLOv12x trained on **TSBOW** (ours)	0.743	**0.869**	**0.846**	**0.792**

Table 7: Models performance for *car* across different metrics **on the comparison set**.

the comparison subset comprises medium-scale scenes, included in the UAVDT, UA-DETRAC, and TSBOW datasets.

Tab. 7 details detection performance for the *car* class across various metrics on this comparison set. YOLOv12x models were trained on UAVDT, UA-DETRAC, and TSBOW datasets with identical setups. The UAVDT-trained model yielded the lowest scores, as its high-altitude drone footage is less applicable to ground-based CCTV surveillance. The UA-DETRAC-trained model achieved high precision but low recall, mAP50, and mAP50-95, due to its emphasis on clear vehicle features at specific distances, overlooking distant vehicles. Conversely, TSBOW mitigates these limitations by incorporating vehicles across diverse scales and optimizing region-of-interest (ROI) settings to enhance detection. Consequently, the TSBOW-trained model balances precision and recall, achieving superior performance in recall, mAP50, and mAP50-95.

Ablation Study on Object Classes Tab. 8 presents the detailed detection performance of the fine-tuned YOLOv12x model across various object classes. The class "car," with high occurrence, achieves the highest scores in three of four metrics, while "bus" excels in mAP50-95 due to the distinct features of fixed-shape objects. For smaller objects, such as "pedestrians" and "micromobility", the model demonstrates promising detection performance.

Ablation Study on Data Characteristics The fine-tuned YOLOv12x model is evaluated across diverse data characteristics, including weather, scenario, scale, road type, and traffic. Tab. 9 provides detailed performance metrics, including precision, recall, mAP50, and mAP50-95. In the "disaster" scenario, heavy snow significantly obscures object features, markedly impairing detection. Under weather conditions, "normal" yield lower scores than "rain" due to frequent vehicle overlap. Similarly, the "coarse" scale, characterized by numerous small and heavily occluded objects, poses significant detection challenges. For road type, "boulevards," with high vehicle density and occlusion, present substantial obstacles to detection accuracy. Object detectors often struggle with heavily occluded objects, frequently misidentifying two to three occluded vehicles as a single object, leading to numerous missed detections.

Class	Images	Instances	Precision	Recall	mAP50	mAP50–95
All	**29,621**	**689,839**	**0.806**	**0.662**	**0.744**	**0.615**
Unidentified	2,177	4,855	0.475	0.223	0.317	0.221
Others	14,248	90,276	0.769	0.628	0.696	0.619
Pedestrian	11,807	32,779	0.833	0.605	0.715	0.447
Micromobility	9,124	18,490	0.793	0.574	0.726	0.519
Car	29,244	479,560	0.959	0.932	0.959	0.849
Bus	12,378	21,037	0.917	0.929	0.951	0.876
Small truck	16,176	36,152	0.878	0.830	0.870	0.750
Truck	5,390	6,690	0.824	0.575	0.720	0.643

Table 8: YOLOv12x performance across different classes.

Group	Category	Precision	Recall	mAP50	mAP50–95
Scenario	Road	0.756	0.608	0.687	0.575
	Intersection	**0.827**	0.677	**0.759**	**0.633**
	Special cases	0.800	**0.681**	0.747	0.615
	Disaster	0.765	0.559	0.656	0.510
Weather	Normal	0.791	0.622	0.726	0.605
	Haze	0.799	**0.732**	0.785	**0.684**
	Rain	**0.851**	0.721	**0.789**	0.664
	Snow	0.783	0.641	0.723	0.597
Scale	Fine	0.733	0.619	0.686	0.559
	Medium	**0.810**	0.666	**0.751**	**0.630**
	Coarse	0.791	**0.676**	0.733	0.581
Road Type	Urban	0.821	**0.675**	**0.757**	**0.632**
	Standard	**0.828**	0.661	0.754	0.622
	Boulevard	0.745	0.664	0.702	0.579
Traffic	Light	0.792	0.658	0.726	0.617
	Moderate	0.795	0.647	0.734	0.606
	Heavy	**0.841**	**0.678**	**0.769**	**0.640**

Table 9: Influence of dataset characteristics on object detection performance

Conclusion and Future Works

This study introduces the Traffic Surveillance Benchmark for Occluded vehicles under various Weather conditions (TSBOW), a comprehensive, semi-automatically annotated traffic surveillance dataset designed to improve monitoring system training, particularly under extreme weather conditions such as heavy haze and snow. Collected across all seasons and diverse road scenarios, TSBOW comprises 32 hours of footage from 198 videos, encompassing a variety of road types and scales, and providing multiple viewing angles for vehicles and pedestrians. The dataset includes over 3.2 million frames, each annotated with weather conditions and scenarios, alongside detailed object annotations derived from extracted images. Capturing complex, high-density scenes of vehicles and pedestrians in crowded urban settings, TSBOW features approximately 71.1 million bounding boxes across eight distinct traffic participant classes. As a robust resource for traffic surveillance research, TSBOW offers substantial potential to deepen insights into traffic dynamics and support advancements in intelligent transportation systems. The initial version focuses on daytime traffic flow under varying weather conditions. Future versions will include ground truth annotations for nighttime scenarios and additional computer vision tasks, such as multi-object tracking, semantic segmentation, vehicle counting, and speed estimation, to further enhance its utility.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-01364, An intelligent system for 24/7 real-time traffic surveillance on edge devices).

References

Alif, M. A. R.; and Hussain, M. 2025. YOLOv12: A Breakdown of the Key Architectural Features. arXiv:2502.14740.

Alkhammash, E. H. 2025. Multi-Classification Using YOLOv11 and Hybrid YOLO11n-MobileNet Models: A Fire Classes Case Study. *Fire*, 8(1): 17.

Bahnsen, C.; and Moeslund, T. 2019. Rain Removal in Traffic Surveillance: Does it Matter? *I E E E Transactions on Intelligent Transportation Systems*, 20(8): 2802–2819.

Canon. 2010. Canon EOS 500D. https://www.canon.co.uk/for_home/product_finder/cameras/digital_slr/eos_550d/. Accessed: December 25, 2024.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, Z.; Cheng, Y.; Liu, L.; Wang, S.; Ke, R.; Schönlieb, C.-B.; and Aviles-Rivero, A. I. 2024. TrafficCAM: A versatile dataset for traffic flow segmentation. *IEEE Transactions on Intelligent Transportation Systems*.

Diaz-Ruiz, C. A.; Xia, Y.; You, Y.; Nino, J.; Chen, J.; Monica, J.; Chen, X.; Luo, K.; Wang, Y.; Emond, M.; Chao, W.-L.; Hariharan, B.; Weinberger, K. Q.; and Campbell, M. 2022. Ithaca365: Dataset and Driving Perception Under Repeated and Challenging Weather Conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21383–21392.

Doersch, C.; Gupta, A.; Markeeva, L.; Recasens, A.; Smaira, L.; Aytar, Y.; Carreira, J.; Zisserman, A.; and Yang, Y. 2022. TAP-Vid: A Benchmark for Tracking Any Point in a Video. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 13610–13626. Curran Associates, Inc.

Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6569–6578.

Fernández Llorca, D.; Hernández Martínez, A.; and García Daza, I. 2021. Vision-based vehicle speed estimation: A survey. *IET Intelligent Transport Systems*, 15(8): 987–1005.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gupta, A.; Dollár, P.; and Girshick, R. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5351–5359.

Hafiz, A. M.; and Bhat, G. M. 2020. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3): 171–189.

Han, J.; Liang, X.; Xu, H.; Chen, K.; HONG, L.; Ye, C.; Zhang, W.; Li, Z.; Liang, X.; and Xu, C. 2021. SODA10M: Towards Large-Scale Object Detection Benchmark for Autonomous Driving.

He, J.; Chen, H.; Liu, B.; Luo, S.; and Liu, J. 2024. Enhancing YOLO for occluded vehicle detection with grouped orthogonal attention and dense object repulsion. *Scientific Reports*, 14(1): 19650.

He, Q.; Xu, A.; Ye, Z.; Zhou, W.; and Cai, T. 2023. Object detection based on lightweight YOLOX for autonomous driving. *Sensors*, 23(17): 7596.

Hidayatullah, P.; Syakrani, N.; Sholahuddin, M. R.; Gelar, T.; and Tubagus, R. 2025. YOLOv8 to YOLO11: A Comprehensive Architecture In-depth Comparative Review. *arXiv preprint arXiv:2501.13400*.

Hussain, M. 2023. YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection. *Machines*, 11(7).

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s): 1–41.

Li, B.; Huang, H.; Zhang, A.; Liu, P.; and Liu, C. 2021. Approaches on crowd counting and density estimation: a review. *Pattern Analysis and Applications*, 24: 853–874.

Li, Y.; Wang, Y.; Wang, W.; Lin, D.; Li, B.; and Yap, K.-H. 2024. Open world object detection: a survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Liang, H.; and Han, R. 2024. OVT-B: A New Large-Scale Benchmark for Open-Vocabulary Multi-Object Tracking. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 14849–14863. Curran Associates, Inc.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, L.; Cheng, Y.; Deng, Z.; Wang, S.; Chen, D.; Hu, X.; Liò, P.; Schönlieb, C.-B.; and Aviles-Rivero, A. 2024. TrafficMOT: A Challenging Dataset for Multi-Object Tracking in Complex Traffic Scenarios. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 1265–1273. Association for Computing Machinery. ISBN 9798400706868.

Lu, D.; and Weng, Q. 2007. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5): 823–870.

Ma, C.; Qiao, L.; Zhu, C.; Liu, K.; Kong, Z.; Li, Q.; Zhou, X.; Kan, Y.; and Wu, W. 2024. HoloVIC: Large-scale Dataset and Benchmark for Multi-Sensor Holographic Intersection and Vehicle-Infrastructure Cooperative. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22129–22138.

Olorunshola, O. E.; Irhebhude, M. E.; and Evwiekpaefe, A. E. 2023. A comparative study of YOLOv5 and YOLOv7 object detection algorithms. *Journal of Computing and Social Informatics*, 2(1): 1–12.

Sohan, M.; Sai Ram, T.; Reddy, R.; and Venkata, C. 2024. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics*, 529–545. Springer.

Soleimanitaleb, Z.; Keyvanrad, M. A.; and Jafari, A. 2019. Object tracking methods: A review. In *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*, 282–288. IEEE.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Sun, Z.; Wei, G.; Fu, W.; Ye, M.; Jiang, K.; Liang, C.; Zhu, T.; He, T.; and Mukherjee, M. 2024. Multiple Pedestrian Tracking Under Occlusion: A Survey and Outlook. *IEEE Transactions on Circuits and Systems for Video Technology*.

Verma, A.; Chakravarthi, B.; Vaghela, A.; Wei, H.; and Yang, Y. 2024. ETraM: Event-Based Traffic Monitoring Dataset. In *Proceedings - 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 22637–22646. IEEE Computer Society.

Wang, W. 2023. Advanced Auto Labeling Solution with Added Features. https://github.com/CVHub520/X-AnyLabeling.

Weber, M.; Xie, J.; Collins, M. D.; Zhu, Y.; Voigtlaender, P.; Adam, H.; Green, B.; Geiger, A.; Leibe, B.; Cremers, D.; Osep, A.; Leal-Taixé, L.; and Chen, L. 2021. STEP: Segmenting and Tracking Every Pixel. *CoRR*, abs/2102.11859.

Wen, L.; Du, D.; Cai, Z.; Lei, Z.; Chang, M.-C.; Qi, H.; Lim, J.; Yang, M.-H.; and Lyu, S. 2020. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193: 102907.

Yilmaz, A.; Javed, O.; and Shah, M. 2006. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4): 13–es.

Yu, H.; Li, G.; Zhang, W.; Huang, Q.; Du, D.; Tian, Q.; and Sebe, N. 2020. The Unmanned Aerial Vehicle Benchmark: Object Detection, Tracking and Baseline. *International Journal of Computer Vision*, 128: 1141–1159.

Zhao, L.; and Li, S. 2020. Object detection algorithm based on improved YOLOv3. *Electronics*, 9(3): 537.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16965–16974.

Zheng, Y.; Harley, A. W.; Shen, B.; Wetzstein, G.; and Guibas, L. J. 2023. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19855–19865.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing Through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, D.; Wang, K.; Gu, J.; Peng, X.; Lian, D.; Zhang, Y.; You, Y.; and Feng, J. 2023. Dataset quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17205–17216.

Zhou, X.; Larintzakis, K.; Guo, H.; Zimmer, W.; Liu, M.; Cao, H.; Zhang, J.; Lakshminarasimhan, V.; Strand, L.; and Knoll, A. 2025. TUMTraf VideoQA: Dataset and Benchmark for Unified Spatio-Temporal Video Understanding in Traffic Scenes. In *Forty-second International Conference on Machine Learning*.

Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; and Ye, J. 2023. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3): 257–276.