

StreetView-Waste: A Multi-Task Dataset for Urban Waste Management

Diogo J. Paulo^{1,2}, João Martins¹, Hugo Proença^{1,2}, João C. Neves^{1,3}

¹University of Beira Interior, Portugal ²IT: Instituto de Telecomunicações ³NOVA LINCS

diogo.paulo@ubi.pt

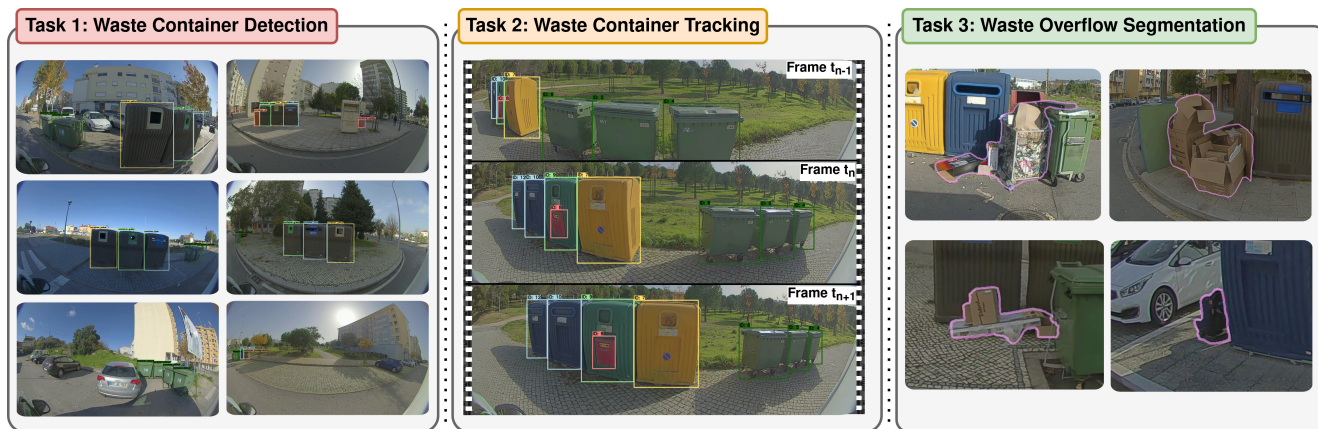


Figure 1. We introduce **StreetView-Waste**, the first fisheye image dataset tailored for urban waste analysis. Captured using two 180° field of view cameras, the dataset mirrors the settings of real urban waste collection, providing high-quality annotations for three core tasks: 2D object detection, object tracking, and instance segmentation. These tasks are critical for logistics, with detection and overflow segmentation enabling status assessment, while tracking is essential for mapping municipal assets and optimizing collection routes. StreetView-Waste serves as a foundation for developing robust, real-world waste analysis models.

Abstract

Urban waste management remains a critical challenge for the development of smart cities. Despite the growing number of litter detection datasets, the problem of monitoring overflowing waste containers — particularly from images captured by garbage trucks — has received little attention. While existing datasets are valuable, they often lack annotations for specific container tracking or are captured in static, decontextualized environments, limiting their utility for real-world logistics. To address this gap, we present *StreetView-Waste*, a comprehensive dataset of urban scenes featuring litter and waste containers. The dataset supports three key evaluation tasks: (1) waste container detection, (2) waste container tracking, and (3) waste overflow segmentation. Alongside the dataset, we provide baselines for each task by benchmarking state-of-the-art models in object detection, tracking, and segmentation. Additionally, we enhance baseline performance by proposing two complementary strategies: a heuristic-based method for

improved waste container tracking and a model-agnostic framework that leverages geometric priors to refine litter segmentation. Our experimental results show that while fine-tuned object detectors achieve reasonable performance in detecting waste containers, baseline tracking methods struggle to accurately estimate their number; however, our proposed heuristics reduce the mean absolute counting error by 79.6%. Similarly, while segmenting amorphous litter is challenging, our geometry-aware strategy improves segmentation mAP@0.5 by 27% on lightweight models, demonstrating the value of multimodal inputs for this task. Ultimately, *StreetView-Waste* provides a challenging benchmark to encourage research into real-world perception systems for urban waste management.

1. Introduction

Effective urban waste management is critical for public health and environmental sustainability. While waste collection has been done manually at fixed-schedule time slots,

there is a growing interest in leveraging computer vision for either inferring statistics of waste distribution or even automating waste collection. This explains the growing number of computer vision datasets for litter detection. However, existing datasets often focus on general litter classification or lack the specific, dynamic context needed for real-world logistics. To bridge this gap, this paper introduces StreetView-Waste, a comprehensive, publicly available dataset comprising 36,478 fisheye images collected in real-world urban settings. The data was captured from a collection vehicle over several weeks to encompass a wide range of real-world urban settings, weather patterns, and lighting conditions, ensuring high environmental diversity. Our dataset is uniquely designed to support three key computer vision tasks: (1) **waste container detection**, annotated with over 71,000 bounding boxes across diverse container type, (2) **waste container tracking**, which includes nearly 400 unique container tracks to evaluate temporal consistency, and (3) **waste overflow segmentation**, featuring over 5,000 detailed instance masks of litter surrounding containers.

Furthermore, to demonstrate the value of our dataset for exploring novel research directions, we focus on the most challenging tasks: waste container tracking and overflowing waste segmentation. We propose and evaluate a model-agnostic approach that enhances segmentation models by fusing RGB data with geometric information and a heuristic-based method for improving waste container tracking. While traditional approaches [4, 14, 28, 35] rely solely on RGB input information or from IoT sensors, we enhance the power of segmentation models by introducing geometric information and developing a model-agnostic solution. We conduct extensive experiments on a variety of state-of-the-art models, including detection, tracking, and segmentation methods. Our results demonstrate significant performance gains in segmentation accuracy when combined with geometric information. Nonetheless, the accuracy of our system is inherently connected to the quality of depth and normal predictions, which are susceptible to noise; hence, we also discuss the limitations exposed by our diagnostic tools. Our findings confirm that tracking waste containers and segmenting overflowing litter in street-level scenes remains an open problem. To this end, we present StreetView-Waste as a challenging new benchmark, paving the way to accelerate research in this domain. While our proposed strategies for segmentation and tracking achieve significant gains over baselines — reducing container counting error by 79.6% and boosting overflow segmentation mAP@0.5 by 27% on lightweight models — the results show that there is still room for improvement. We believe our dataset and initial findings will serve as a valuable resource for the community to develop and validate the next generation of solutions for automated urban

waste management. Our main contributions are as follows:

- **A Novel, Large-Scale Dataset for Urban Waste Management.** We introduce and make publicly available StreetView-Waste, a new, large-scale, multi-task dataset of over 36,000 fisheye images captured from the street-level perspective of waste collection vehicles.
- **A Comprehensive Benchmark on Litter and Waste Container Analysis.** We provide an extensive benchmark of state-of-the-art models on the proposed dataset for three key tasks: waste container detection, tracking and counting, and waste overflow segmentation. This work establishes a thorough performance benchmark, providing a reference for future research in this domain.
- **Innovative Strategies for Segmentation and Tracking.** We propose and validate two distinct strategies that enhance general-purpose models for critical waste management tasks. The first is a model-agnostic method that improves overflow segmentation by fusing RGB data with geometric information. The second is a novel heuristic-based approach that increases the accuracy and robustness of waste container tracking and counting.

The full dataset and code will be made publicly available.

2. Related Work

Litter Detection Datasets. Several works have focused attention on the detection of litter. Initial studies focused on waste classification, aiming to sort materials for recycling [3, 8, 9, 16, 18, 25, 33, 36, 43–47, 53]. While these are vital for sorting applications, these datasets typically feature decontextualized images of isolated objects with simple backgrounds, limiting their utility for in-the-wild scene understanding. To address real-world scenarios, research has progressed towards litter detection and segmentation in complex environments. Among these works, [5, 6, 20, 27, 38, 50] provide instance segmentation masks for different litter categories across various settings. Many other datasets provide bounding box annotations for litter detection, including [8, 9, 15, 29]. For underwater environments, [18, 21] offer benchmarks for detecting marine debris. For aerial surveillance, UAVWaste [27] provides data captured from drones. Others address unique surface conditions, such as LOTS [5] for segmenting litter on sandy beaches, or specific object types, like the BePLi Dataset v1 [20] for plastic litter. Although these domain-specific datasets are invaluable, they do not address the distinct perceptual challenges of ground-level urban waste management; a critical gap remains from the operational perspective of municipal service vehicles. Furthermore, large-scale datasets like WoodScape [54] use fisheye cameras, commonly found on such vehicles, but are annotated for autonomous driving tasks (e.g., vehicles, pedestrians), and entirely lack the labels required for waste management.

The task of monitoring waste containers for efficient collection logistics remains underexplored. While some datasets include waste containers as a class [1, 34, 49], no existing resource provides a comprehensive benchmark for the joint tasks of waste container detection, multi-object tracking from a moving vehicle, and waste overflow segmentation. Our work is explicitly designed to fill this void. To the best of our knowledge, StreetView-Waste is the first large-scale dataset to combine a street-view fisheye perspective with rich, multi-task annotations for container detection, tracking, and overflow segmentation, thereby fostering the development of smart waste management systems. For a direct comparison, we summarize the key statistics of these datasets, including image and class counts, as well as the type of task, in a comparative table in the supplementary material (Table 1).

Automated Litter Analysis. The datasets previously mentioned have enabled the benchmarking of various models for automated litter analysis, focusing on object detection and segmentation. For waste and container detection, several models have been employed as benchmarks [12, 17, 30, 31, 39, 40, 48, 56]. Two-stage models like Faster R-CNN [40] provide high accuracy and are often used for baseline evaluations on datasets like TrashICRA19 [15] and TrashCan [21]. For applications requiring real-time performance, such as on-vehicle processing, one-stage detectors like the YOLO family [39] and SSD [31] are predominantly used due to their computational efficiency. Regarding object segmentation tasks, instance segmentation models have been adopted to produce fine-grained, pixel-level masks. The typical approach for high-quality segmentation is Mask R-CNN [19], which has been benchmarked on datasets like TACO [38] and TrashCan [21]. For real-time scenarios, faster models such as YOLACT [10], SOLO [51], and its successor SOLOv2 [52] have been developed. More recently, vision transformers have set a new standard, with models like Mask2Former [13] unifying panoptic, instance, and semantic segmentation into a single powerful framework.

3. Dataset

StreetView-Waste comprises images of waste containers captured under various real-world conditions, as well as instances of overflowing waste and litter in the surrounding areas. Our dataset’s primary value lies in its challenging scenarios, such as severe occlusions by passing vehicles, illumination changes, and the wide range of field of view, which often cause state-of-the-art models to fail. To situate our contribution within the existing landscape, we provide a detailed comparison with other relevant public datasets in the supplementary material (Table 1) and a figure illustrating several images of StreetView-Waste (Figure 1).

3.1. Dataset Acquisition

To ensure a high degree of environmental and temporal diversity, our dataset was collected across multiple, distinct sessions on different days. These recording sessions were intentionally varied, taking place in both the morning and afternoon under different weather conditions, ranging from bright, direct sunlight creating harsh shadows to overcast skies with diffuse, low-contrast lighting. This strategic collection ensured that StreetView-Waste captures a wide spectrum of real-world operational scenarios, avoiding biases that might arise from a single, short-term collection period. Our dataset was captured from a vehicle equipped with two fisheye cameras, which were mounted on the vehicle’s flanks, each providing a 180° field of view to ensure wide environmental coverage. Raw video sequences were recorded at 1920×1080 resolution and 30 fps. To minimize temporal correlation between frames and ensure a diverse training and evaluation set, we applied a uniform subsampling procedure, resulting in an effective frame rate of 10 fps for annotation. To maximize reproducibility, the full set of camera intrinsics and distortion coefficients is provided in the supplementary material (Table 2), estimated using the generic fisheye model of Kannala and Brandt [24]. No rectification or cropping was applied, as the raw fisheye projection preserves critical context from both road and sidewalk regions.

3.2. Tasks and Metrics

Our StreetView-Waste dataset is designed to benchmark three critical tasks for automated waste management. We establish baselines for each using state-of-the-art models and standard evaluation protocols.

Waste Container Detection. The first task evaluates the ability of models to perform multi-class detection of seven distinct container types from individual frames. Accurate detection is the foundational step for any subsequent analysis, such as tracking or status assessment. To establish robust baselines, we evaluate two distinct state-of-the-art paradigms. We use a high-performance, frame-based detector, YOLOv11 [23], representing the efficiency required for real-time, on-vehicle applications, and a video object detector, DiffusionVID [42], which uses temporal information across frames to potentially improve detection robustness and consistency. Performance is measured using the standard mean average precision (mAP) metric.

Multi-Object Tracking and Counting. This task addresses the challenge of associating container detections across consecutive frames to maintain unique identities, which is an important task for route optimization and mapping municipal assets. We benchmark two strong tracking-by-detection baselines: ByteTrack [55] and BoT-SORT [2]. Their ability to use low-confidence detections to handle occlusions is particularly relevant for cluttered street-view

scenes where containers are frequently occluded by vehicles. Tracking performance is assessed by analyzing two aspects: the tracking accuracy and object counting estimation. In the former, we adopt a) the standard multiple object tracking accuracy (MOTA) proposed in [7]; b) the identity f1 score (IDF1), introduced in [41], which is a metric that is more sensitive to long-term identity preservation compared to MOTA; and c) the higher order tracking accuracy (HOTA) [32]. For counting accuracy, we use the mean absolute error (MAE), root mean square error (RMSE) between the predicted count and the ground truth, mean absolute percentage error (MAPE) for providing a normalized estimation of counting deviation, and the sum of absolute count differences per sequence (SAD).

Overflowing Waste Segmentation. This task aims to produce pixel-accurate instance masks for unstructured waste spilling out of containers. Although operational alarms could be binary, pixel masks are required to estimate overflow volume and prioritize routes, which is central to the logistics goal; hence, we keep overflowing waste segmentation as the primary benchmark. Precise segmentation is a prerequisite for quantifying overflow volume and triggering collection alerts. To achieve this, we benchmark a diverse set of architectures, offering a comprehensive overview of how various architectural biases address this task. We evaluate the segmentation quality using the mean average precision (mAP) over masks and the boundary intersection over union (B-IoU).

3.3. Statistics per Task

To facilitate rigorous and reproducible research, we provide standardized training, validation, and test splits for each task. These splits are performed at the video level to prevent data leakage, ensuring that frames from the same recording do not appear in different sets. The full StreetView-Waste dataset comprises a total of 36,478 images, captured to maximize environmental diversity. Of these, 14,219 images contain labeled objects, forming the core of our benchmarks. The statistics for each specific task are detailed below. Our dataset covers seven municipal container types reflecting common European curbside collection. Default containers are general-purpose bins deployed on most streets. Green containers are for glass. Blue containers collect paper and cardboard. Yellow containers handle lightweight packaging (plastics and metals). Biodegradable containers are for organic waste. Oil containers are dedicated units for household cooking oil due to environmental risk. Battery containers are compact drop-boxes for small batteries, often narrow and frequently occluded by street furniture. This variety mirrors real segregation rules and yields strong class imbalance for minority streams, which we capture in StreetView-Waste. An additional per-class analysis of the size and location of the con-

tainers in the image is provided in the supplementary material (Figure 2). For waste container detection, the dataset

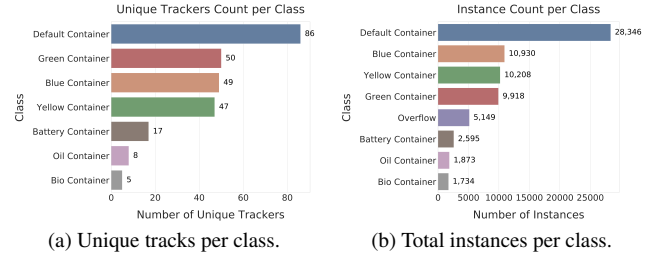


Figure 2. **Class distribution statistics for our StreetView-Waste dataset.** (a) Distribution of the 376 unique container tracks, highlighting the long-tail nature of the tracking task. (b) Distribution of the 71,170 total annotated instances for the detection task.

is substantial, featuring 71,170 annotated instances across seven classes, as shown in Figure 2b. This scale provides a robust foundation for developing and evaluating object detectors in complex urban environments, capturing a wide diversity of container appearances and scenarios. The multi-object tracking and counting benchmark is built upon dense annotations identifying 376 unique container identities. As illustrated in Figure 2a, these tracks exhibit a realistic long-tail class distribution. While common types, such as the default container, are abundant, minority classes, such as the biodegradable container and oil container, present a challenging real-world scenario for evaluating long-term tracking and re-identification. Finally, the overflowing waste segmentation task is supported by a dedicated set of annotations focusing on unstructured waste. It includes 5,149 fine-grained, pixel-level instance masks for litter and overflow. These annotations are distributed across 4,197 positive images (with the presence of waste). The complete segmentation benchmark contains 7,230 images, including negative samples, to ensure a robust evaluation of a model’s ability to identify and segment overflowing waste precisely.

3.4. Privacy Considerations and Data Integrity

Our dataset, captured on public streets, contains faces and license plates, which are subject to the GDPR. While we considered anonymization methods like GAN-based replacement [26], we concluded that any alteration poses a significant risk to the data’s scientific utility, as artifacts could be misidentified as litter or occlude key features. Therefore, to avoid introducing harmful domain shifts, we decide to preserve the original data, a position supported by the creators of other large-scale datasets [11, 54]. Consequently, access will be managed through a formal data license agreement that restricts use to academic research and requires users to comply with GDPR. This approach balances privacy obligations with the need for high-quality, unaltered data. Moreover, the data collection was con-

ducted exclusively in public spaces where no additional ethics board approval was required under national regulations. Nevertheless, the study protocol, including privacy handling, was reviewed internally to ensure GDPR compliance. For illustrative purposes in this paper, all exemplar figures have been redacted to prevent identification.

4. Proposed Strategies

Our initial benchmarking experiments, detailed in Section 5, reveal varying levels of difficulty across the three evaluation tasks. We observe that fine-tuned object detectors achieve reasonable performance on waste container detection. However, the more complex tasks of tracking and segmentation expose significant limitations in current state-of-the-art models when applied to our challenging street-view domain. For tracking, standard methods struggle with frequent occlusions and the visual similarity of containers, resulting in high rates of identity switches and track fragmentation. For segmentation, delineating amorphous, overflowing waste from its container and the cluttered urban background proves exceptionally challenging for models that rely solely on RGB data. To address these specific challenges, we introduce two complementary, model-agnostic strategies designed to enhance the performance of off-the-shelf models. These strategies target the distinct failure modes observed in tracking and segmentation, respectively.

4.1. Heuristic-based Tracking Refinement

While tracking-by-detection frameworks like ByteTrack [55] or BoT-SORT [2] provide a strong foundation, their performance degrades in scenarios with prolonged occlusions or when containers briefly exit and re-enter the camera’s field of view. To mitigate these errors, we introduce a set of heuristics not only as a practical method to improve tracking and counting accuracy, but also as a diagnostic tool that benchmarks the specific failure modes of existing trackers in this challenging vehicular context. These rules are designed to enforce domain-specific constraints based on the physical behavior of static containers, and the thresholds for these rules were empirically determined by optimizing counting performance on our validation set.

Our approach introduces three key rules:

- **H₁: Minimum Track Duration Filter** - This heuristic dictates that tracks shorter than 15 frames are removed. These short-lived tracks often result from unreliable detections when the container is distant from the camera, as indicated by small bounding box areas and unstable ID assignments. Thus, removing them reduces false positives and temporal noise.
- **H₂: Temporal Track Merging Based on Gaps** - This heuristic addresses track fragmentation by assuming such gaps may still belong to the same object. Two tracks of

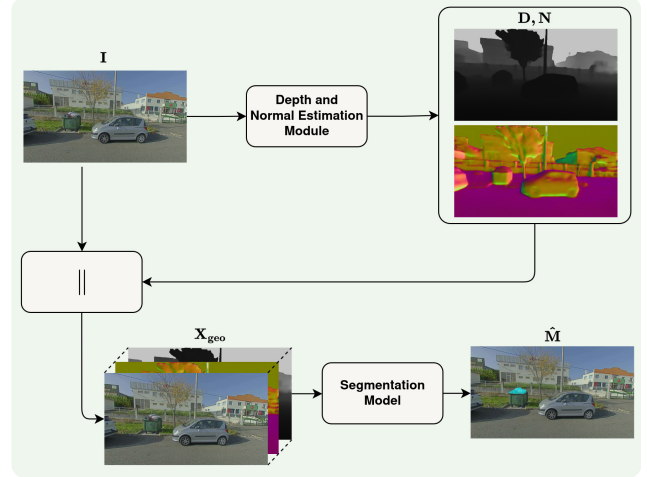


Figure 3. **Overview of the geometry-aware method for the segmentation task.** The input RGB image I is processed using a geometry estimation module, which produces both a depth map D and a surface normal map N . These are then concatenated with the original image to form an enriched input tensor $X_{\text{geo}} \in \mathbb{R}^{H \times W \times 7}$. This new representation is then fed to adapted segmentation models capable of handling multi-channel input, which output the predicted mask for overflowing waste.

the same class are merged if the temporal gap between them is less than or equal to 20 frames.

- **H₃: Spatial Proximity Constraint for Merging** - This heuristic extends H_2 by enforcing a spatial proximity condition: the distance between the last detection of the first track and the first detection of the second track must be within a maximum center distance of 0.10 normalized image units.

4.2. Geometry-Aware Overflow Segmentation

To compensate for the information lost during the camera’s projective transform, we introduce a geometry-aware strategy to address the ambiguity of overflowing waste segmentation by leveraging 3D spatial data to supplement the 2D RGB input. As illustrated in Figure 3, our approach aims to reduce both false positives (e.g., background clutter mistaken for waste) and false negatives (missed overflow) by resolving the inherent scale ambiguity, allowing the model to differentiate between near and distant objects.

Our primary hypothesis is that depth and surface normal information can provide additional and useful cues. We use a zero-shot, single-image geometry estimation model, Metric3Dv2 [22], to infer a depth map (D) and a surface normal map (N) from each input image (I). Depth maps are clipped at 50 m and normalized by this value to map distances into $[0, 1]$, while surface normals are stored as images and rescaled into $[0, 1]$. These geometric maps are then concatenated with the original RGB channels to form an en-

riched 7-channel input tensor $X_{\text{geo}} = I \| N \| D \in \mathbb{R}^{H \times W \times 7}$.

To process this channel-extended input, the kernels of the first convolutional layer in existing segmentation architectures are adapted to accept seven input channels instead of three. The rest of the network architecture and its pre-trained weights remain unchanged, allowing us to use pre-existing models with minimal modification. This fusion provides advantages as depth information helps the model distinguish foreground waste from similarly textured background elements (e.g., walls, distant vehicles) and surface normals describe the local orientation of surfaces, enabling the model to better understand the shape and form of the overflowing waste.

5. Benchmarks and Results

To demonstrate the utility of StreetView-Waste and establish performance baselines, we conduct a comprehensive empirical evaluation across the three proposed tasks. This section details our experimental setup and presents a thorough analysis of the results, highlighting the unique challenges presented by our dataset. All results are reported as mean $\pm$ standard deviation computed over 10 bootstrap re-samples of the test set.

Implementation Details. For the three tasks, all experiments were conducted on NVIDIA GeForce RTX 5070 with 12 GB of VRAM, and the framework utilized was PyTorch [37]. Resource usage (#Params, GFLOPs, latency/FPS at 640×640 , and peak VRAM) is summarized in the supplementary material (Table 6). To ensure consistency across experiments, all models were trained using a uniform image size of 640×640 . The batch size was set to 8 for all models, except for Mask2Former [13], for which we used a reduced batch size of 4. To ensure a fair comparison, the key hyperparameters for each model (e.g., learning rate, weight decay) were independently optimized on our validation set. Regarding the dataset proportion of positive and negative samples, we employed a sampling strategy for images containing target objects versus background-only images in order to ensure a ratio of approximately 1:2 for the detection and tracking tasks and a balanced 1:1 for the overflow segmentation task. As described in 4.2, depth maps were clipped at 50 m and normalized to [0,1], while surface normals were stored as images and rescaled to [0,1] before concatenation with the RGB channels. Additionally, standard data augmentation techniques, including horizontal flipping, were applied during training to mitigate overfitting.

5.1. Waste Container Detection

As shown in Table 1, the single-frame model consistently outperforms the video-based one across most categories, achieving an overall mAP@[0.5:0.95] of 0.77 versus 0.70

Table 1. **Benchmark of StreetView-Waste for per-class detection accuracy with the standard deviation ($\pm$ std).** The YOLOv11 [23] model consistently outperforms DiffusionVID [42] across most classes, except for Battery container.

Class	DiffusionVID [42]			YOLOv11 [23]		
	AP@0.5	AP@[.5:.95]	AR@[.5:.95]	AP@0.5	AP@[.5:.95]	AR@[.5:.95]
Default	0.96 $\pm$ 0.02	0.75 $\pm$ 0.03	0.78 $\pm$ 0.03	0.98$\pm$0.01	0.82$\pm$0.02	0.82$\pm$0.02
Green	0.92 $\pm$ 0.03	0.76 $\pm$ 0.03	0.82 $\pm$ 0.03	0.97$\pm$0.02	0.82$\pm$0.03	0.84$\pm$0.02
Biodegradable	0.92 $\pm$ 0.02	0.75 $\pm$ 0.02	0.77 $\pm$ 0.03	0.96$\pm$0.01	0.85$\pm$0.02	0.86$\pm$0.02
Blue	0.95 $\pm$ 0.02	0.73 $\pm$ 0.03	0.79 $\pm$ 0.03	0.96$\pm$0.02	0.80$\pm$0.02	0.81$\pm$0.02
Yellow	0.90 $\pm$ 0.03	0.73 $\pm$ 0.03	0.80 $\pm$ 0.03	0.96$\pm$0.01	0.82$\pm$0.02	0.84$\pm$0.02
Oil	0.92 $\pm$ 0.02	0.62 $\pm$ 0.04	0.69 $\pm$ 0.04	0.93$\pm$0.01	0.70$\pm$0.03	0.70$\pm$0.03
Battery	0.82$\pm$0.05	0.57 $\pm$ 0.05	0.61$\pm$0.05	0.80 $\pm$ 0.06	0.61$\pm$0.05	0.55 $\pm$ 0.05
All	0.91 $\pm$ 0.03	0.70 $\pm$ 0.03	0.76 $\pm$ 0.03	0.94$\pm$0.02	0.77$\pm$0.03	0.77$\pm$0.03

Table 2. **Benchmark of StreetView-Waste for the tracking task using our post-processing heuristics.** We report the overall metrics for the baseline tracker and with the progressive application of each heuristic. The arrows ($\uparrow/\downarrow$) indicate whether a higher or lower value is better.

Model	Experiment	MOTA $\uparrow$	IDF1 $\uparrow$	HOTA $\uparrow$	DetA $\uparrow$	AssA $\uparrow$
ByteTrack [55]	Baseline	76.80%	81.40%	69.76%	69.96%	69.83%
	H_1 (Duration)	77.10%	82.20%	69.73%	69.30%	70.40%
	$H_1 + H_2$ (Temporal)	77.00%	55.10%	50.98%	45.62%	57.03%
	$H_1 + H_2 + H_3$ (Spatial)	77.20%	66.80%	59.60%	56.35%	63.15%
BoT-SORT [2]	Baseline	82.50%	79.60%	71.50%	75.27%	68.17%
	H_1 (Duration)	82.40%	82.10%	72.09%	73.80%	70.61%
	$H_1 + H_2$ (Temporal)	82.50%	75.80%	67.34%	65.82%	69.06%
	$H_1 + H_2 + H_3$ (Spatial)	82.50%	79.80%	70.30%	70.69%	70.06%

and an mAR@[0.5:0.95] of 0.77 versus 0.76. These results suggest that for this task, the high-quality features learned by a state-of-the-art static detector are more impactful than the temporal information exploited by DiffusionVID [42]. However, a key exception provides critical insight. DiffusionVID [42] surpasses YOLOv11 [23] on the *Battery Container* class. Our analysis suggests this is due to the unique physical properties of this class. Battery containers are typically smaller and narrower than other types, making them more susceptible to being fully occluded for several consecutive frames by street objects like poles, signs, or cars. In these scenarios, DiffusionVID’s [42] temporal propagation mechanism can effectively fill in detections during brief occlusions where a single-frame detector would fail, demonstrating the specific conditions under which video-based models offer a distinct advantage.

5.2. Waste Container Tracking and Counting

This task focuses on tracking and counting waste container instances within video sequences in urban scenes, which is essential for applications such as inventory management and mapping. The results, shown in Tables 2 and 3, reveal significant challenges in temporal reasoning that current methods struggle with.

The comparison between ByteTrack [55] and BoT-SORT [2] highlights the dataset’s difficulty. BoT-SORT achieves a stronger baseline with 82.5% of MOTA and

Table 3. **Benchmark of StreetView-Waste for object counting accuracy.** The results evidence strong performance when using our proposed heuristics.

Model	Experiment	MAE↓	SAD↓	RMSE↓	MAPE↓
ByteTrack [55]	Baseline	3.48	73	7.80	82.96%
	H_1 (Duration)	1.05	22	1.93	24.96%
	$H_1 + H_2$ (Temporal)	<u>0.76</u>	<u>16</u>	1.75	<u>17.77%</u>
	$H_1 + H_2 + H_3$ (Spatial)	0.71	15	<u>1.70</u>	16.03%
BoT-SORT [2]	Baseline	6.43	135	9.60	187.61%
	H_1 (Duration)	1.19	25	1.91	27.19%
	$H_1 + H_2$ (Temporal)	0.81	17	1.36	16.03%
	$H_1 + H_2 + H_3$ (Spatial)	0.90	19	1.79	17.96%

71.5% of HOTA, which reflects its improved detection association. However, both trackers suffer from frequent identity switches and degraded performance once heuristics are introduced. For example, while our spatio-temporal heuristics ($H_1 + H_2 + H_3$) reduce ByteTrack’s counting error drastically (MAE from 3.48 to 0.71, with SAD dropping from 73 to 15), they simultaneously reduce identity preservation (IDF1 81.4% $\rightarrow$ 66.8%, HOTA 69.8% $\rightarrow$ 59.6%). A similar pattern occurs with BoT-SORT, where heuristics improve counting accuracy (MAE 6.43 $\rightarrow$ 0.90) but disrupt balanced association quality. A detailed breakdown of these counting improvements for each class is available in the supplementary material (Tables 4 and 5). As illustrated in Figure 4, this happens frequently when containers of the same type appear sequentially along a collection route. Therefore, we interpret these heuristics not as improvements but as diagnostic tools; they show that even simple, rule-based temporal or spatial constraints expose the fundamental difficulty of maintaining consistent identities in cluttered street-level scenes. This confirms that container tracking under realistic operational conditions remains an open challenge, and StreetView-Waste provides a benchmark for studying these limitations in depth.

5.3. Waste Overflow Segmentation

For this task, we evaluated five state-of-the-art instance segmentation models to establish baselines for the challenging task of segmenting amorphous, overflowing waste. Table 4 compares the performance of each model on the standard RGB fisheye projection [24] against our proposed geometry-aware strategy, which incorporates geometric priors (depth and surface normals).

The results reveal an architectural division in the ability to use multi-modal information for this task. On the one hand, lightweight models show significant performance gains with the added geometric data, with YOLACT [10] achieving the highest overall mAP@[0.5:0.95] and B-IoU of 0.31 and 0.90, respectively. On the other hand, more complex models exhibit performance degradation. Our analysis suggests that this divergence oc-

Table 4. **Benchmark of StreetView-Waste for the waste overflow segmentation task using our geometry-aware strategy.** The addition of geometric priors improves performance for the majority of the models.

	Model	Experiment	mAP@0.5	mAP@[0.5:0.95]	B-IoU
Complex Architectures	SOLOv2 [52]	Baseline (RGB)	0.20±0.03	0.10±0.01	0.30±0.02
		Ours (Geometric Cues)	0.07±0.01	0.03±0.00	0.20±0.01
	Mask2Former [13]	Baseline (RGB Data)	0.18±0.02	0.11±0.01	0.38±0.02
		Ours (Geometric Cues)	0.29±0.02	0.13±0.01	0.32±0.02
	Mask R-CNN [19]	Baseline (RGB)	0.41±0.02	0.26±0.01	0.54±0.02
		Ours (Geometric Cues)	0.12±0.01	0.06±0.01	0.38±0.01
Lightweight Models	YOLACT [10]	Baseline (RGB)	0.41±0.02	0.22±0.01	0.87±0.02
		Ours (Geometric Cues)	0.52±0.02	0.31±0.02	0.90±0.01
	YOLOv11 [23]	Baseline (RGB)	0.50±0.01	0.30±0.01	0.77±0.01
		Ours (Geometric Cues)	0.52±0.01	0.31±0.01	0.77±0.01

curs due to fundamental architectural limitations when faced with unstructured targets like litter. The models that successfully adapt, such as YOLACT [10] and Mask2Former [13], possess more flexible instance representation mechanisms, such as YOLACT’s [10] prototype-based design and Mask2Former’s [13] transformer-based cross-attention mechanism, which allows them to effectively learn relationships between modalities across the entire scene. In contrast, models like Mask R-CNN [19] are built on rigid structural assumptions that are violated by our dataset’s challenges. Two-stage models rely on a Region Proposal Network (RPN) that fails when there are no well-defined, “object-like” structures to propose from the amorphous litter. Similarly, SOLOv2’s [52] grid-based design, which assumes one compact instance per grid cell, cannot handle waste that expands across multiple grid cells. For these architectures, the introduction of new geometric channels does not provide a helpful signal but instead exposes and amplifies these core weaknesses, leading to a significant degradation in performance. Furthermore, Figure 5 depicts original images processed using the geometry-aware method, and to complement, **ablation studies** are provided in the supplementary material (Table 3), analyzing the individual contributions of depth and surface normal cues.

5.4. Discussion and Limitations

Our empirical evaluation establishes baselines for StreetView-Waste and highlights key challenges for future research. While geometric priors improve segmentation, the approach remains sensitive to errors in the predicted depth and surface normals, which can degrade performance in cluttered or low-contrast scenes. Moreover, complex architectures such as Mask R-CNN [19] and SOLOv2 [52] did not benefit from the added modalities, suggesting that larger training sets or more specialized multimodal fusion blocks (e.g., cross-attention mechanisms) may be required. These observations point to a need for further exploration of how geometric cues interact with different architectural designs.



Figure 4. **Qualitative results for the multi-object tracking.** This scenario, common in our dataset, shows the difficulty of keeping track and explains the results of the lower IDF1 score when introducing temporal heuristics. This improves track continuity but corrupts identity (lowering IDF1).

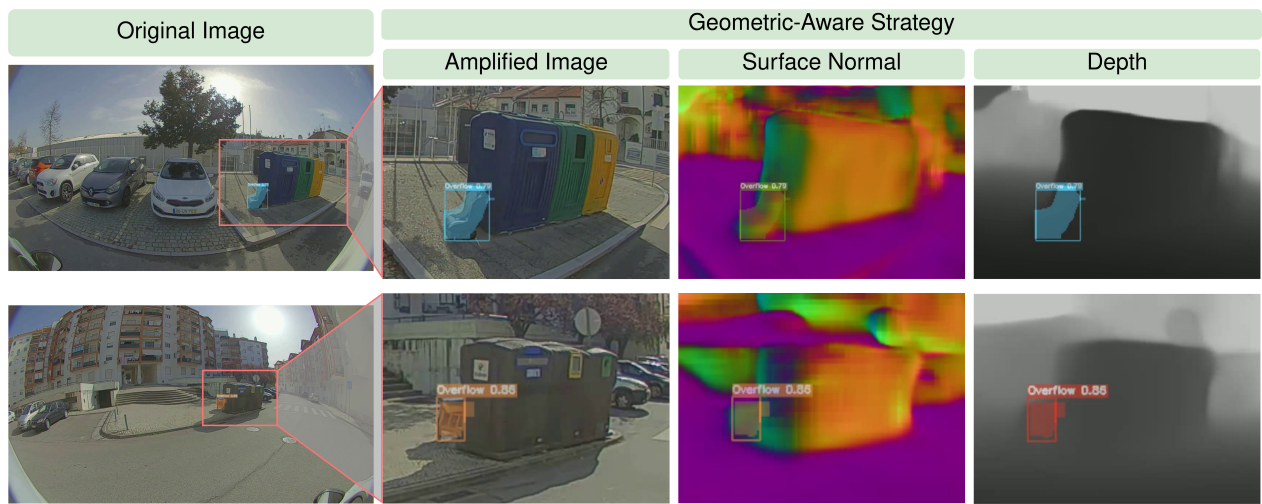


Figure 5. **Qualitative results for the waste overflow segmentation task, comparing the original images with our proposed geometry-aware strategy.** The columns show, from left to right: the original input image, the same image with the segmentation result from our method, the estimated surface normal map, and the estimated depth map.

6. Conclusion

We introduced StreetView-Waste, a large-scale dataset for benchmarking detection, tracking, and overflow segmentation in urban waste management. We applied two simple strategies, not as novel solutions but as diagnostic tools, to probe the limitations of state-of-the-art models. Our experiments yield two key insights: (1) simple heuristics markedly reduce counting errors, revealing that current trackers lack fundamental temporal reasoning; and (2) our geometric fusion strategy lowers the performance of complex architectures, showing their fragility to multi-modal inputs and amorphous objects. These results confirm StreetView-Waste as a critical diagnostic benchmark. Future work will extend the dataset with GPS to enable logistics-oriented applications.

7. Acknowledgments

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., and, when eligible, co-funded by EU funds under project/support UID/50008/2025 – Instituto de Telecomunicações, with DOI identifier <https://doi.org/10.54499/UID/50008/2025>. This work is also financed by the project WATERMARK¹ and supported by UID/04516/NOVA Laboratory for Computer Science and Informatics (NOVA LINC) with the financial support of FCT.IP.

¹WATERMARK project (Watermark-Based Algorithms for Trustworthy Media Authentication and Robust Certification in Public Administration), Project No. 2024.07356.IACDC, supported by “RE-C05-i08.M04 – Support the launch of a program of R&D projects aimed at the development and implementation of advanced systems in cybersecurity, artificial intelligence, and data science in public administration, as well as a scientific training program,” under the Recovery and Resilience Plan (PRR), as part of the funding agreement signed between the Recovery Portugal Task Force (EMRP) and the Foundation for Science and Technology (FCT).

References

- [1] Cathaoir Agnew, Dishant Mewada, Eoin M Grua, Ciarán Eising, Patrick Denny, Mark Heffernan, Ken Tierney, Pepijn Van de Ven, and Anthony Scanlan. Detecting the over-filled status of domestic and commercial bins using computer vision. *Intelligent Systems with Applications*, 18:200229, 2023. 3
- [2] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 3, 5, 6, 7
- [3] arthurcen. Waste images from sushi restaurant. <https://www.kaggle.com/datasets/arthurcen/waste-images-from-sushi-restaurant>, 2020. Kaggle dataset. 2
- [4] Archana Balmik, Subhasish Barik, Mrityunjay Jha, and Anup Nandy. A vision-based litter detection and classification using ssd mobilenetv2. In *International conference on signal processing and integrated networks (SPIN)*, pages 180–185. IEEE, 2023. 2
- [5] Paola Barra, Alessia Auriemma Citarella, Giosué Orefice, Modesto Castrillón-Santana, and Angelo Ciaramella. Lots: Litter on the sand dataset for litter segmentation. In *International Conference on Machine Vision and Applications (MVA)*, pages 1–4. IEEE, 2023. 2
- [6] Dina Bashkurova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21147–21157, 2022. 2
- [7] Keni Bernardin and Rainer Stiefelhausen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008(1): 246309, 2008. 4
- [8] Janusz Bobulski and Mariusz Kubanek. Deep learning for plastic waste classification system. *Applied Computational Intelligence and Soft Computing*, 2021(1):6626948, 2021. 2
- [9] Janusz Bobulski and Jacek Piatkowski. Pet waste classification method and plastic waste database-wadaba. In *International conference on image processing and communications*, pages 57–64. Springer, 2017. 2
- [10] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 3, 7
- [11] Markus Braun, Sebastian Krebs, Fabian Flohr, and Darius M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1844–1861, 2019. 4
- [12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3
- [13] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3, 6, 7
- [14] Berardina De Carolis, Francesco Ladogana, and Nicola Macchiarulo. Yolo trashnet: Garbage detection in video streams. In *IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pages 1–7. IEEE, 2020. 2
- [15] Michael Fulton, Jungseok Hong, Md Jahidul Islam, and Junaed Sattar. Robotic detection of marine litter using deep visual detection models. In *International Conference on Robotics and Automation (ICRA)*, pages 5752–5758. IEEE, 2019. 2, 3
- [16] gale31. Tidy: Trash image dataset “yucky”. <https://github.com/gale31/TIDY>, 2023. Accessed on July 14, 2025. Dataset of classified trash images. 2
- [17] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3
- [18] Henry Haeffliger. Deepseawaste. <https://www.kaggle.com/datasets/henryhaefliger/deepseawaste>, 2020. Kaggle dataset of approximately 3,055 underwater waste images with CSV annotations; accessed July 14, 2025. 2
- [19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 7
- [20] Mitsuko Hidaka, Koshiro Murakami, Kenta Koshidawa, Shintaro Kawahara, Daisuke Sugiyama, Shin’ichiro Kako, and Daisuke Matsuoka. Bepli dataset v1: Beach plastic litter dataset version 1 for instance segmentation of beach plastic litter. *Data in Brief*, 48:109176, 2023. 2
- [21] Jungseok Hong, Michael Fulton, and Junaed Sattar. Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv preprint arXiv:2007.08097*, 2020. 2, 3
- [22] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 5
- [23] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 3, 6, 7
- [24] Juho Kannala and Sami S Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1335–1340, 2006. 3, 7
- [25] Adam Kelly. Cigarette butt dataset. <https://www.immersivelimit.com/datasets/cigarette-butts>, 2018. 2,200 synthetic COCO-formatted images of cigarette butts on ground; accessed July 14, 2025. 2
- [26] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 4
- [27] Marek Kraft, Mateusz Piechocki, Bartosz Ptak, and Krzysztof Walas. Autonomous, onboard vision-based trash and litter detection in low altitude aerial images collected

- by an unmanned aerial vehicle. *Remote Sensing*, 13(5):965, 2021. 2
- [28] Hrushikesh N Kulkarni and Nandini Kannamangalam Sundara Raman. Waste object detection and classification. *CS230 Stanford*, 2019. 2
- [29] Nikhil Venkat Kumsetty, Amith Bhat Nekkare, Sowmya Kamath, et al. Trashbox: trash detection and classification using quantum transfer learning. In *Conference of Open Innovations Association (FRUCT)*, pages 125–130. IEEE, 2022. 2
- [30] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6054–6063, 2019. 3
- [31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3
- [32] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2): 548–578, 2021. 4
- [33] Gaurav Mittal, Kaushal B Yagnik, Mohit Garg, and Narayanan C Krishnan. Spotgarbage: smartphone app to detect garbage using deep learning. In *Proceedings of the ACM international joint conference on pervasive and ubiquitous computing*, pages 940–945, 2016. 2
- [34] Paula Moral, Álvaro García-Martín, Marcos Escudero-Viñolo, José M Martínez, Jesús Bescós, Jesús Peñuela, Juan Carlos Martínez, and Gonzalo Alvis. Towards automatic waste containers management in cities via computer vision: containers localization and geo-positioning in city maps. *Waste Management*, 152:59–68, 2022. 3
- [35] Abdulhalik Oğuz and Ömer Faruk Ertuğrul. Determining the fullness of garbage containers by deep learning. *Expert Systems with Applications*, 217:119544, 2023. 2
- [36] OpenLitterMap. Openlittermap dataset: Open citizen science data on litter & plastic pollution, 2018. Peer-reviewed article in Open Geospatial Data 3:6. 2
- [37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [38] Pedro F Proença and Pedro Simões. Taco: Trash annotations in context for litter detection. *arXiv preprint arXiv:2003.06975*, 2020. 2, 3
- [39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3
- [40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3
- [41] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 4
- [42] Si-Dong Roh and Ki-Seok Chung. Diffusionvid: Denoising object boxes with spatio-temporal conditioning for video object detection. *IEEE Access*, 11:121434–121444, 2023. 3, 6
- [43] sapal6. Waste classification data v2. <https://www.kaggle.com/datasets/sapal6/waste-classification-data-v2>, 2020. Kaggle dataset. Extended version with non-recyclable (“N”) class added. 2
- [44] Suresh Singh, Minwei Luo, and Yu Li. Multi-class anomaly detection. In *International Conference on Neural Information Processing*, pages 359–371. Springer, 2022.
- [45] Sam Single, Saeid Iranmanesh, and Raad Raad. Realwaste: a novel real-life data set for landfill waste classification using deep learning. *Information*, 14(12):633, 2023.
- [46] Joan Sosa-García and Francesca Odone. “hands on” visual recognition for visually impaired users. *ACM Transactions on Accessible Computing (TACCESS)*, 10(3):1–30, 2017.
- [47] sumn2u. Garbage classification v2. <https://www.kaggle.com/datasets/sumn2u/garbage-classification-v2>, 2024. Kaggle dataset for waste classification tasks. Accessed on July 14, 2025. 2
- [48] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 3
- [49] Miguel Valente, Hélio Silva, João Caldeira, Vasco Soares, and Pedro D Gaspar. Computer vision approaches to waste containers detection. In *Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–4. IEEE, 2019. 3
- [50] Tao Wang, Yuanzheng Cai, Lingyu Liang, and Dongyi Ye. A multi-level approach to waste object segmentation. *Sensors*, 20(14):3816, 2020. 2
- [51] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *European conference on computer vision*, pages 649–665. Springer, 2020. 3
- [52] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020. 3, 7
- [53] Mindy Yang and Gary Thung. Classification of trash for recyclability status. *CS229 project report*, 2016(1):3, 2016. 2
- [54] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O’Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the International Conference on Computer Vision*, pages 9308–9318, 2019. 2, 4
- [55] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 3, 5, 6, 7
- [56] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of*

the IEEE/CVF international conference on computer vision,
pages 6748–6758, 2023. [3](#)