Generation of Synthetic Data Based on Real Images for the Detection of Plastic Waste in an Urban Canal

Ana Sofia da Silva Barbosa[1], Antonio Celso Dantas Antonino[1], Manoel Henriques de Sá Campos Filho[2], Tsang Ing Ren[1], Jean Firmino Cardoso[1]

[1] Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)
Recife – PE – Brazil

[2]Escola Politécnica de Pernambuco (POLI)
Universidade de Pernambuco (UPE)
Recife – PE – Brazil

assb@cin.ufpe.br, antonio.antonino@ufpe.br

manoel.sacampos@upe.br, tir@cin.ufpe.br

jean.firmino@ufpe.br

Abstract. *Monitoring plastic waste in urban waterways demands continuous image analysis, yet building annotated datasets for this task is costly, requiring manual inspection of thousands of recorded frames to find usable examples. This paper investigates whether synthetic data can replace or reduce that annotation effort in a Brazilian urban canal under permanent camera surveillance. We contribute a real annotated dataset (S1) and three synthetic variants: images generated by DALL·E conditioned on real canal photographs (S2), copy-paste composites using SAM-extracted waste instances (S3), and a hybrid combining all three sources equally (S4). All datasets are intended to be made publicly available to support future research. Training YOLOv11s on each configuration and evaluating all four models on a fixed held-out set of 1,000 real canal images, the real dataset (S1) achieved the best generalisation (Precision = 0.8723, Recall = 0.9244), while the copy-paste strategy (S3) demonstrated the best balance among synthetic configurations (Precision = 0.3516, Recall = 0.5790). The T2I model (S2) generalised poorly to real conditions despite its strong internal validation score (mAP@0.5 = 0.9866). These results highlight a critical gap between synthetic validation performance and real-world generalisation, and suggest that visual consistency within the training set matters more than data source diversity.*

1. Introduction

The accumulation of solid waste in urban waterways represents a serious environmental and public health challenge, in tropical cities with high population density [UNEP 2021]. Automated detection of such waste through computer vision emerges as a promising alternative to manual monitoring, enabling faster and continuous responses from authorities responsible for cleaning and managing water bodies.

In this context, an urban canal in a tropical Brazilian city is monitored by a surveillance camera that continuously transmits images to a server. Although this system pro-

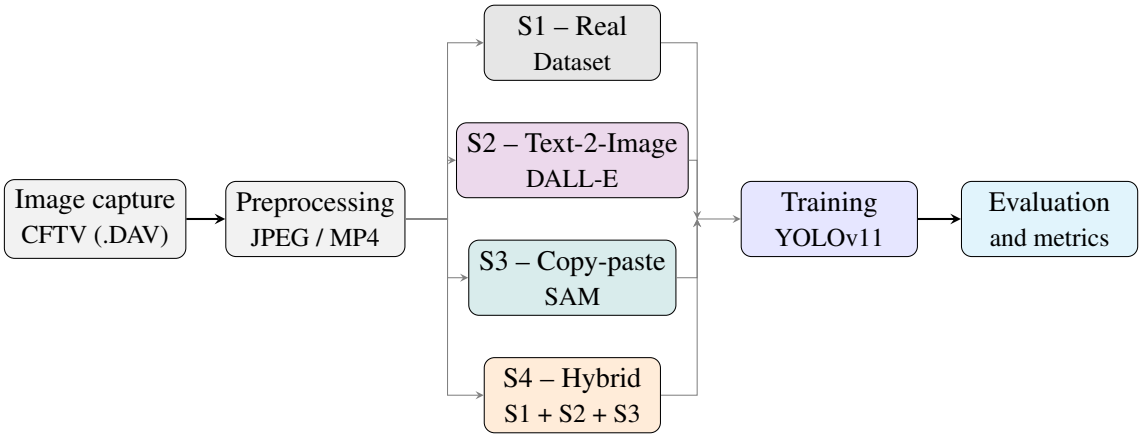


Figure 1. Methodology pipeline showing the four dataset composition strategies (S1–S4)

vides a constant stream of visual data, building an annotated dataset for training detection models is labor-intensive: it requires manually inspecting a large volume of stored images to find frames where plastic waste is sufficiently visible to compose quality training samples. Furthermore, canal conditions vary considerably. Lighting, water turbidity, and the diversity of waste types hinder dataset standardization.

A promising approach to reduce this dependence on manual curation is the use of synthetic data. Copy-paste techniques allow segmented object instances to be inserted into new backgrounds, increasing dataset diversity without requiring new data collection. Text-to-image generative models, such as DALL-E [Ramesh et al. 2022], enable the creation of artificial scenes depicting plastic waste under varying visual conditions. Automatic segmentation models such as the Segment Anything Model (SAM) [Kirillov et al. 2023] further facilitate the extraction of object instances from real images. However, it remains unclear to what extent these synthetic strategies match the use of real images when training object detectors for this specific domain.

This work addresses this gap by comparatively evaluating four dataset composition strategies for plastic waste detection in an urban canal: a manually annotated real dataset (S1), a synthetic dataset generated by a text-to-image model (S2), a copy-paste dataset (S3), and a mixed hybrid combining all three approaches (S4), using YOLOv11 as the detection model. Figure 1 illustrates the overall methodology pipeline.

The main contributions of this work are: (i) a real annotated dataset from the target urban canal; (ii) three synthetic dataset variants: generative, copy-paste, and hybrid; (iii) a comparative evaluation of all four strategies using YOLOv11; and (iv) public release of the prompts used for T2I generation and the annotated datasets, intended to support future research.

2. Related Work

Object detection in aquatic environments has gained increasing attention in recent years, driven by the growing concern over plastic pollution in rivers and urban waterways. Several studies have explored the use of YOLO-based models for detecting floating waste on water surfaces.

Plastic waste detection in rivers using YOLOv5 has been shown to be feasible even in challenging outdoor conditions [Sio et al. 2022]. More recently, [Tomas et al. 2024] evaluated YOLOv8 for real-time detection of floating debris in waterways, reporting competitive results in terms of speed and accuracy. [Peng et al. 2024] proposed an improved YOLOv7 architecture incorporating a global feature pyramid network and long-range attention mechanisms specifically for floating waste detection, achieving gains over the baseline model.

In the context of plastic bottle detection in rivers, [Ramsurrun et al. 2024] combined YOLOv8 with object tracking to count plastic bottles advecting downstream, achieving recall values above 0.947. [Devi et al. 2024] compared YOLOv6 and YOLOv7 for detecting plastic bottles on Indonesian water surfaces using the FloW-Img dataset, finding that both models achieve satisfactory detection rates.

A key limitation across these works is the scarcity of annotated data for specific environments such as urban canals. To address data scarcity, synthetic data generation has emerged as a promising direction. [Ghiasi et al. 2021] demonstrated that simple copy-paste augmentation, inserting segmented object instances into new backgrounds, yields strong improvements for instance segmentation. [Rombach et al. 2022] introduced Latent Diffusion Models, enabling high-quality text-to-image synthesis that can generate training images depicting specific objects and scenes. The Segment Anything Model (SAM) [Kirillov et al. 2023] further facilitates the automatic extraction of object masks from real images, enabling scalable copy-paste pipelines without manual segmentation.

2.1. Visual Challenges in Aquatic Environments

Object detection in aquatic environments faces inherent perceptual difficulties that go beyond model architecture. Water surfaces introduce turbidity, light reflections, and varying illumination that reduce image contrast and distort object appearance. Floating waste items are highly heterogeneous, comprising plastic bottles, bags, food packaging, wrappers, and other everyday consumer products in various states of degradation, and are frequently partially submerged, deformed, or clumped together, causing their color and shape to blend with natural elements such as foam, sediment, and vegetation. This wide variety of object types, combined with high intra-class variability and low contrast against the aquatic background, hinders the formation of stable visual patterns that detection models can reliably learn [Peng et al. 2024, Tomas et al. 2024]. These challenges are further amplified in urban canals such as the one studied here, where environmental conditions, including seasonal lighting changes, water turbidity, and the diversity of discarded materials, differ substantially from those represented in existing public datasets, reinforcing the need for locally collected and annotated data.

2.2. Synthetic Data for Object Detection

The scarcity of annotated data is a well-known bottleneck in supervised object detection, in domain-specific settings where manual labelling is costly. Two complementary strategies have gained traction to address this limitation.

The first is copy-paste augmentation, in which segmented object instances are composited onto background images from the target domain. [Ghiasi et al. 2021] showed that this simple technique yields consistent gains for instance segmentation even without

sophisticated blending. Automatic mask extraction, as provided by the Segment Anything Model [Kirillov et al. 2023], further reduces the annotation effort required to build object libraries for such pipelines. Despite its practicality and low computational cost, copy-paste augmentation has a known limitation: composited images may lack realism, as pasted instances do not always blend naturally with the background in terms of lighting, scale, and context. In this work, plastic waste instances were extracted from a public floating waste dataset [ttylu 2024] and pasted onto real images captured at the target canal.

The second strategy uses text-to-image generative models to synthesise training images from natural language descriptions. [Ramesh et al. 2022] demonstrated that large-scale diffusion models conditioned on text can produce photorealistic and semantically controllable images, making them attractive for generating domain-specific training data without real-world collection. However, this approach introduces its own challenges: generation quality is highly sensitive to prompt design, requiring careful crafting to accurately convey the target scene, object appearance, and environmental conditions. Furthermore, querying large text-to-image models at scale can incur significant computational or financial cost. In this work, synthetic images were generated using DALL-E prompted with descriptions grounded on real photographs from the target canal, aiming to preserve the visual characteristics of the target environment. The full list of prompts used is publicly available [Barbosa 2025].

Both strategies were evaluated against a real-data baseline, allowing an empirical comparison of their respective contributions to floating waste detection performance.

3. Methodology

This section presents the methodology employed to assess the extent to which synthetic data can mitigate the reliance on manually curated real-world images for plastic waste detection in urban waterways. Four dataset composition strategies were systematically designed and compared under controlled experimental conditions. This section also describes the image acquisition pipeline developed for the target canal, the annotation procedure adopted for each configuration, and the object detection architecture selected for the evaluation.

3.1. Image Capture Pipeline

Images of the target urban canal are captured by a fixed camera (Intelbras VIP 3225 SD) installed on a utility pole at the canal monitoring site. The camera angle and optical zoom were adjusted to provide sufficient sharpness and detail for the subsequent detection and modelling stages. Images are recorded by an Intelbras NVD 1404 network video recorder, which transfers footage to a local workstation via an FTP server, where files are stored on an SSD. The workstation is equipped with an Intel Core i7-13700 processor (13th Gen, 24 threads), 32 GB of RAM, a 500 GB SSD and a 6 TB external hard drive for storage, and an NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM for model training. Footage is periodically transferred to the workstation in encrypted .DAV files, which were manually converted to .JPEG or .MP4 format for inspection. Frames were then individually reviewed and selectively extracted to compose the dataset, retaining only images with sufficient visual quality and representative canal conditions.

3.2. Real Dataset (S1)

A total of 1,500 real images from the target canal footage were selected, retaining only frames in which plastic waste was clearly visible. These images were divided into training (70%), validation (15%), and test (15%) subsets, yielding 1,050 training images, 225 validation images, and 225 test images. Images were annotated in Roboflow using bounding boxes with a single class: `plastic`, comprising visible plastic waste such as bottles, bags, and food packaging. These images compose configuration S1.

3.3. Text-to-Image Dataset (S2)

Synthetic images were generated using the DALL·E text-to-image model, accessed via the ChatGPT interface, by submitting textual prompts describing urban canal scenes with floating plastic waste under varying lighting conditions and waste density. Prompts were grounded on real photographs of the canal without visible waste, so as to preserve the visual characteristics of the target environment. The full list of prompts used will be made publicly available upon publication of this work. A total of 500 images were initially generated; the final dataset (S2) was then expanded to 1,000 images through data augmentation techniques, including random rotations, horizontal flipping, brightness and contrast adjustments, and cropping variations, to increase sample diversity. All images were annotated in Roboflow with a single class: `plastic` and divided into training (70%), validation (15%), and test (15%) subsets, yielding 700 training images, 150 validation images, and 150 test images.

3.4. Copy-Paste Dataset (S3)

Plastic waste instances were segmented from a public floating waste dataset [ttylu 2024] using the Segment Anything Model (SAM) [Kirillov et al. 2023] and composited onto real canal background images containing no visible waste, following the copy-paste augmentation strategy [Ghiasi et al. 2021]. This procedure initially yielded 675 images, which were then expanded to 1,000 images through data augmentation techniques, including random rotations, horizontal flipping, brightness and contrast adjustments, and cropping variations. All images were divided into training (70%), validation (15%), and test (15%) subsets, yielding 700 training images, 150 validation images, and 150 test images, and annotated in Roboflow with a single class: `plastic`, composing configuration S3.



(a) Real (S1)



(b) Synthetic/T2I (S2)



(c) Synthetic/Copy-Paste (S3)

Figure 2. Representative samples from three of the four dataset composition strategies (S4 comprises images from S1, S2, and S3).

3.5. Hybrid Dataset (S4)

The hybrid dataset combines samples from all three strategies, comprising 500 real annotated images, 500 T2I synthetic images, and 500 copy-paste images, totalling 1,500 samples (S4). This configuration aims to leverage the complementary strengths of each approach. All images share a unified single-class annotation scheme (`plastic`) and were divided into training (70%), validation (15%), and test (15%) subsets, yielding 1,050 training images, 225 validation images, and 225 test images.

3.6. Detection Model and Training Setup

The YOLOv11s architecture [Jocher and Qiu 2024] was selected for all experiments. Among the YOLO11 variants, YOLOv11s offers a suitable balance between computational efficiency and detection accuracy: it achieves 47.0 mAP@0.5:0.95 on COCO with only 9.4M parameters, outperforming the lighter YOLOv11n (39.5 mAP@0.5:0.95) [Jocher and Qiu 20
Additionally, its compact size makes it compatible with edge deployment scenarios, which may be relevant for future on-site inference at the canal monitoring station. All models were trained for 100 epochs with early stopping (patience = 20), a batch size of 16, an initial learning rate of 0.01, and input resolution of 640×640 pixels, using pre-trained COCO weights for initialisation. A total of four training runs were conducted, one per

dataset configuration (S1–S4). To enable a fair cross-configuration comparison on real-world data, a fixed held-out test set of 1,000 real canal images, previously unseen during any training or validation stage, was reserved exclusively for final evaluation. This set was constructed to assess real-world generalisation across all configurations under identical domain conditions, with no overlap with any training or validation split. Validation metrics reported during training reflect each configuration's internal validation subset.

3.7. Evaluation Metrics

Models were evaluated using four standard object detection metrics: Precision, Recall, mean Average Precision at IoU threshold 0.5 (mAP@0.5), and mAP at IoU thresholds from 0.5 to 0.95 (mAP@0.5:0.95). Precision measures the fraction of correct detections among all detections made; Recall measures the fraction of ground-truth instances correctly detected; mAP@0.5 summarises detection performance across all confidence thresholds at a fixed IoU of 0.5; and mAP@0.5:0.95 provides a stricter evaluation across multiple IoU thresholds, penalising imprecise localisation.

4. Results and Discussion

Table 1 presents the internal validation metrics obtained during training for each dataset composition strategy, reflecting performance on each configuration's own validation subset. Table 2 presents the results on the fixed held-out test set of 1,000 real canal images, which provides a domain-consistent evaluation of real-world generalisation. Unlike the internal validation sets, this held-out set was not used in any training or validation stage of any configuration, ensuring a fair and unbiased comparison across all four strategies (S1–S4) under identical real-world conditions. All configurations achieved mAP@0.5 above 0.90 on their internal validation sets.

The T2I dataset (S2) achieved the highest mAP@0.5 (0.9866), surpassing the real dataset baseline (S1, mAP@0.5 = 0.9269) by 6.0 percentage points. This result is notable, as it demonstrates that images generated by a text-to-image model conditioned on real canal photographs without visible waste can produce training data that generalises effectively within the synthetic validation domain. S3 also achieved the highest Recall (0.9336), indicating fewer missed detections, a relevant property in environmental monitoring where false negatives carry practical consequences such as unreported pollution events.

The copy-paste strategy (S3) yielded the highest Precision (0.9801) and mAP@0.5:0.95 (0.8180), suggesting that compositing segmented waste instances onto real canal backgrounds produces well-localised detections even under stricter IoU thresholds. The real dataset (S1), despite being the most labour-intensive to produce, requiring manual inspection of thousands of recorded frames, ranked second in mAP@0.5 (0.9269), reinforcing the potential of synthetic strategies to reduce annotation effort without sacrificing detection quality.

Contrary to the initial expectation that combining all three sources would yield the best performance, the hybrid configuration (S4) achieved the lowest results across all internal validation metrics (mAP@0.5 = 0.9094, Precision = 0.9011, Recall = 0.8278). A possible explanation is that mixing images with substantially different visual characteristics, namely real canal frames, AI-generated scenes, and composited backgrounds,

introduced distributional heterogeneity that hindered the formation of stable visual patterns during training.

Table 1. Validation metrics per dataset composition strategy (each evaluated on its own internal validation subset).

Dataset	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Real (S1)	0.9648	0.9103	0.9269	0.7542
T2I (S2)	0.9112	0.9091	0.9866	0.8093
Copy-Paste (S3)	0.9801	0.9336	0.9149	0.8180
Hybrid (S4)	0.9011	0.8278	0.9094	0.8077

Table 2. Model performance on the fixed held-out real test set (1,000 Real Canal images, confidence threshold = 0.70).

Model	Precision	Recall
Real (S1)	0.8723	**0.9244**
T2I (S2)	0.2568	0.2262
Copy-Paste (S3)	0.3516	**0.5790**
Hybrid (S4)	0.1713	**0.8603**

The real-world evaluation reveals a behaviour markedly different from that observed in the internal validation metrics. The real dataset (S1) achieved the best overall performance on the held-out test set, with Precision of 0.8723 and Recall of 0.9244, confirming that models trained on real domain-specific data generalise most effectively to real deployment conditions.

Among the synthetic configurations, S3 (copy-paste) demonstrated the best balance between Precision (0.3516) and Recall (0.5790), making it the most stable synthetic strategy for real-world deployment. The S4 (hybrid) model achieved the highest Recall among synthetic configurations (0.8603), indicating a tendency to detect most waste instances present, but at the cost of low Precision (0.1713), suggesting a high rate of false detections in regions without waste. The S2 (T2I) model yielded the weakest performance among synthetic strategies (Precision = 0.2568, Recall = 0.2262), evidencing that the visual domain generated by DALL·E did not generalise adequately to real canal conditions despite its strong internal validation score. These results suggest that validation metrics obtained on synthetic sets may overestimate real-world model performance, and that evaluation on real images from the target domain is essential for a fair comparison between dataset composition strategies.

Figure 3 illustrates representative inference errors produced by each configuration on real canal images. The S2 (T2I) model misclassified vegetation as plastic waste, generating a false positive detection with no true positives in the scene. The S3 (copy-paste) model failed to detect a plastic bottle clearly visible on the water surface, evidencing a missed detection. The S4 (hybrid) model produced false positive detections in regions without waste, corroborating its low Precision observed in Table 2. These examples highlight the domain gap between synthetic training conditions and real-world canal imagery.



(a) S2 – T2I (b) S3 – Copy-Paste (c) S4 – Hybrid

Figure 3. Representative inference results on real canal images for the three synthetic configurations at confidence threshold 0.70.

Figure 4 presents representative successful detections produced by each configuration on images from their respective internal validation subsets, illustrating that all strategies are capable of correctly localising plastic waste instances under their own training domain conditions. These examples contrast with the failure cases observed in Figure 3, reinforcing that the performance gap is primarily attributable to domain shift rather than to a fundamental inability of the models to detect plastic waste.



(a) S2 – T2I (b) S3 – Copy-Paste (c) S4 – Hybrid

Figure 4. Representative successful detections produced by each synthetic configuration on their respective internal validation subsets.

5. Conclusion

This work investigated four dataset composition strategies for plastic waste detection in an urban canal using the YOLOv11s architecture. The strategies compared were: a manually annotated real dataset (S1), a text-to-image synthetic dataset generated with DALL·E conditioned on real canal backgrounds (S2), a copy-paste dataset produced by compositing segmented waste instances onto real canal backgrounds using SAM (S3), and a hybrid dataset combining equal proportions of all three sources (S4).

The real dataset (S1) demonstrated solid and consistent performance across both internal validation and the held-out real test set (Precision = 0.8723, Recall = 0.9244), establishing itself as a reliable baseline for this domain. Its results confirm that carefully

curated real annotated data remains a strong foundation for plastic waste detection in urban canal surveillance, even when the volume of annotated samples is limited. Motivated by the practical difficulty of scaling manual annotation in continuous monitoring scenarios, this work explored whether synthetic strategies could serve as viable complements or alternatives to real data collection.

On internal validation metrics, S2 reached the highest mAP@0.5 (0.9866) and S3 yielded the best localisation accuracy (mAP@0.5:0.95 = 0.8180). These results initially suggested that synthetic strategies could match or exceed real annotated data. However, evaluation on a fixed held-out set of 1,000 real canal images revealed a markedly different picture: S2 generalised poorly to real conditions (Precision = 0.2568, Recall = 0.2262), while S3 achieved the best balance among synthetic strategies (Precision = 0.3516, Recall = 0.5790). Taken together, these results indicate that synthetic datasets, in their current form, are not yet capable of replacing real annotated data for this domain. Rather, they represent a promising avenue for data augmentation, particularly when domain-specific backgrounds are preserved, as in the copy-paste approach.

The underperformance of the hybrid configuration (S4) on both internal validation (mAP@0.5 = 0.9094) and the real test set (Precision = 0.1713, Recall = 0.8603) suggests that combining visually heterogeneous sources does not guarantee improved performance and may introduce distributional inconsistencies that hinder robust pattern learning. Visual consistency within the training set appears to play a more critical role than data source diversity.

The absolute precision and recall values on the held-out test set remain low across synthetic configurations, which can be partially attributed to the domain gap between training and deployment conditions and to the fixed confidence threshold adopted for evaluation. These results should therefore be interpreted as relative comparisons between strategies rather than as indicators of deployment-ready performance.

As future work, we intend to continue investigating the effect of dataset size and composition ratio when combining real and synthetic sources, as these factors were not fully isolated in the present study. We also plan to explore improved blending techniques for copy-paste augmentation, investigate prompt engineering strategies that better preserve the visual characteristics of the target environment, and evaluate the contribution of a negative class to reducing false positives. Ultimately, we aim to validate the most promising configurations through in-field deployment at the canal monitoring station, consolidating the path from synthetic data research toward practical environmental monitoring applications. The datasets and prompts used in this work are intended to be made publicly available to support future research on plastic waste detection in urban waterways.

References

Barbosa, A. S. d. S. (2025). Prompts for DALL·E synthetic image generation — Cavouco Canal plastic waste detection.

Devi, B. S., Gupta, D., and Singh, R. P. (2024). A deep learning approach to plastic bottle waste detection on the water surface using YOLOv6 and YOLOv7. *Engineering, Technology & Applied Science Research*.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2918–2928.

Jocher, G. and Qiu, J. (2024). Ultralytics YOLO11. `https://github.com/ultralytics/ultralytics`. Version 11.0.0.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026.

Peng, C., He, B., Xi, W., and Lin, G. (2024). Improved YOLOv7 algorithm for floating waste detection based on GFPN and long-range attention mechanism. *Wuhan University Journal of Natural Sciences*, 29(4):338–348.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Ramsurrun, N., Suddul, G., Armoogum, S., and Foogooa, R. (2024). Plastic water bottle detection model using computer vision in aquatic environments. *Scientific Reports*, 14.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Sio, G. A., Guantero, D., and Villaverde, J. (2022). Plastic waste detection on rivers using YOLOv5 algorithm. In *Proceedings of the International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE Press.

Tomas, J. P. Q., Tupas, J. E. E., Soniel, M. T., Caruz, C. H. M. E., and Babar, D. B. (2024). Real-time detection of floating debris in waterways using YOLOv8. In *Proceedings of the 14th International Workshop on Computer Science and Engineering (WCSE)*, pages 19–21, Phuket Island, Thailand.

ttylu (2024). Floating plastic waste river dataset. `https://universe.roboflow.com/ttylu/floating-plastic-waste-river`. Available at Roboflow Universe. Accessed: May 2026. Contains 319 images.

UNEP (2021). From pollution to solution: A global assessment of marine litter and plastic pollution. Technical report, United Nations Environment Programme.