A semi-supervised learning-based framework for quantifying litter fluxes in river systems

Jia, Tianlong; Taormina, Riccardo; de Vries, Rinze; Kapelan, Zoran; van Emmerik, Tim H.M.; Vriend, Paul; Okkerman, Imke

DOI
10.1016/j.watres.2025.124833

Publication date
2025

Document Version
Final published version

Published in
Water Research

Citation (APA)
Jia, T., Taormina, R., de Vries, R., Kapelan, Z., van Emmerik, T. H. M., Vriend, P., & Okkerman, I. (2025). A semi-supervised learning-based framework for quantifying litter fluxes in river systems. *Water Research*, *289*, Article 124833. https://doi.org/10.1016/j.watres.2025.124833

Water Research 289 (2026) 124833

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

A semi-supervised learning-based framework for quantifying litter fluxes in river systems

Tianlong Jia[a,b,*], Riccardo Taormina[a], Rinze de Vries[c], Zoran Kapelan[a], Tim H.M. van Emmerik[d], Paul Vriend[e], Imke Okkerman[e]

[a] *Delft University of Technology, Faculty of Civil Engineering and Geosciences, Department of Water Management, Stevinweg 1, 2628 CN Delft, The Netherlands*

[b] *Karlsruhe Institute of Technology (KIT), Institute of Water and Environment, Karlsruhe, Germany*

[c] *Noria Sustainable Innovators, Schieweg 13, 2627 AN Delft, The Netherlands*

[d] *Wageningen University and Research, Hydrology and Environmental Hydraulics Group, Wageningen, The Netherlands*

[e] *Rijkswaterstaat, Ministry of Infrastructure and Water Management, Griffioenlaan 2, 3526 LA Utrecht, The Netherlands*

ARTICLE INFO

Dataset link: https://github.com/TianlongJia/deep_plastic_Flux_SSL, https://doi.org/10.5281/zenodo.17387612

Keywords:
Artificial intelligence
Object detection
Contrastive learning
SwAV
Macroplastic flux
Environmental monitoring
Pollution
Camera images

ABSTRACT

Supervised deep learning methods have been widely employed to detect floating macroplastic litter (> 5 mm) in (fresh)water bodies. However, few studies used them to quantify floating litter fluxes in rivers with wide cross-sections, that is important for pollution assessment. Additionally, commonly used supervised learning (SL) models rely on extensive labeled data, that is time-consuming and expensive to obtain. Moreover, regardless of the model type, current deep learning models for litter detection usually fail to correctly identify small litter items. To overcome these issues, we propose a semi-supervised learning (SSL)-based framework combined with Slicing Aided Hyper Inference (SAHI) for quantifying cross-sectional floating litter fluxes in rivers. The framework includes four steps: (a) collecting camera images of river surfaces from multiple locations across the river, (b) developing a robust litter detection model using SSL, (c) applying this model with SAHI to detect litter items in images, and (d) post-processing the detection results to quantify fluxes. The SSL method involves: (i) self-supervised pre-training of a ResNet50 on a large amount of unlabeled data, and (ii) supervised fine-tuning of a Faster R-CNN with the ResNet50 backbone on a limited amount of labeled data. We evaluated the in-domain detection performance of SSL models with varying pre-training epochs and pre-training dataset sizes, using images from waterways of The Netherlands, Indonesia and Vietnam, that were used for model pre-training and fine-tuning. Additionally, we assessed the zero-shot out-of-domain detection performance of SSL models and litter flux quantification performance of the proposed framework on a Vietnam case study, that was not used for model development. We benchmarked our results against the SL methods and human visual counting. The results show that SSL models benefit from longer pre-training time and larger pre-training dataset, achieving an in-domain F1-score increase of 0.2 and a zero-shot out-of-domain increase of up to 0.14, over baseline SL benchmarks. Furthermore, the SAHI method correctly identifies 45 additional small litter items (areas < 1,000 cm^2), improving the F1-score by up to 0.19, compared to the results obtained without SAHI. The flux measurement results indicate that the SSL-based framework substantially underestimates fluxes by a factor of 3–4 compared to human measurements, due to missed detections of transparent litter items and items entrapped in water hyacinths. However, it estimates nearly twice the fluxes of the baseline SL-based framework, aligning more closely with human measurements. These findings highlight the potential of SSL-based framework to enhance litter flux measurement. Scaling it with broader datasets could significantly advance global-scale litter monitoring systems.

1. Introduction

Plastic pollution in water bodies is a pressing global issue, threatening aquatic life and human health (Bellou et al., 2021; Lebreton et al., 2018). Rivers are the main pathways of land-based plastic waste to the ocean (Meijer et al., 2021), but they also act as potential temporary and long-term plastic sinks, where significant amounts of

[*] Corresponding author at: Delft University of Technology, Faculty of Civil Engineering and Geosciences, Department of Water Management, Stevinweg 1, 2628 CN Delft, The Netherlands.

E-mail addresses: T.Jia@tudelft.nl, tianlong.jia@kit.edu (T. Jia).

https://doi.org/10.1016/j.watres.2025.124833
Received 5 February 2025; Received in revised form 15 October 2025; Accepted 17 October 2025
Available online 27 October 2025

plastic waste accumulate, and even remain trapped for decades (van Emmerik et al., 2022b). Regardless of waste type, monitoring floating litter fluxes (i.e., the number of litter items transported across the river width per unit of time) is key to assess pollution levels in river systems. Such assessment is essential for developing effective pollution reduction measures, such as source reduction and targeted cleaning campaigns (van Emmerik et al., 2022c, 2019).

Traditional measurement approaches include debris sampling and human visual observation (Hurley et al., 2023). Nevertheless, debris sampling is labor-intensive and requires specialized equipments (e.g., nets and booms), that limits its feasibility and scalability across different locations and over long periods of time (van Lieshout et al., 2020). Human visual observation is not feasible for monitoring rivers with high litter fluxes, where human counters face challenges in accurately tracking debris over time (van Lieshout et al., 2020). Additionally, it may be dangerous during extreme events, such as floods (van Emmerik et al., 2023).

Addressing these limitations highlights the need for more automated and scalable measurement approaches. Recently, using deep learning (DL) approaches based on Convolutional Neural Networks (CNNs) to process camera images of river surfaces have attracted significant research interest as efficient alternatives (Gnann et al., 2022). Some studies have demonstrated their capabilities to enhance automated monitoring through various computer vision tasks, such as object detection. For example, Li et al. (2023) proposed a novel network based on Faster Region-based Convolutional Neural Network (Faster R-CNN) and the attention mechanism to detect floating litter from images collected at Dai Lake, China, achieving an average precision of 80.8%. Kataoka et al. (2024) collected images using cameras fixed on bridges and handheld cameras from seven rivers in Japan, and used a You Only Look Once Version 8 (YOLOv8) model to detect floating plastic debris with an average precision (AP50) of 78%. While current studies have shown promising results, several challenges remain to be addressed as follows.

First, few studies have used DL methods to measure floating litter fluxes across rivers with broader cross-section. Most studies have focused on using DL models to detect floating litter (Jia et al., 2023a). For example, van Lieshout et al. (2020) employed a Faster R-CNN model with a single bridge-mounted camera, to estimate plastic fluxes in a narrow waterway in Jakarta, Indonesia. However, a study reported that the horizontal distribution of floating litter fluxes along some wider rivers is highly uneven, based on observations from 24 locations in rivers across seven countries in Europe and Asia, e.g., the Saigon River (300 m wide), in Vietnam (van Calcar and van Emmerik, 2019). Relying on observation from a single or low number of locations to estimate litter fluxes across a wide river may lead to significant under- or overestimation (van Emmerik et al., 2019).

Second, most research used supervised learning (SL) methods for detecting floating litter, that rely on a large amount of labeled images. Jia et al. (2023a) reported an average dataset size of around 9000 images, across 34 papers on using DL for detecting litter in water bodies. Labeling these images is labor-intensive and time-consuming. While transferring the representations learned from general computer vision datasets can reduce the data requirement, these representations are not sufficiently effective to generalize across different locations and environmental conditions (Jia et al., 2023a; Wu et al., 2024). To overcome this issue, Jia et al. (2024b) proposed a two-stage semi-supervised learning (SSL) method for developing DL models to detect floating litter. First, they pre-trained a model using a self-supervised learning method to learn hidden data representations from a large amount of relevant unlabeled data (e.g., ≈100k images). Then, they fine-tuned the model using a limited amount of labeled data in a supervised manner. This approach obtains comparable or superior performance (AP50 = 53.3%) while requiring only 20% of the labeled data, compared to conventional SL methods (AP50 = 51.1%). However, no studies have evaluated its capabilities for quantifying floating litter fluxes, and whether its performance improves with (1) larger pre-training datasets and (2) longer pre-training time. Literature reports the significant impact of these two factors on SSL performance on ImageNet (Deng et al., 2009) classification tasks (Caron et al., 2020; Goyal et al., 2022).

Finally, detecting small litter or litter located far away from the imaging devices still remains a significant challenge (Jia et al., 2024b, 2023b). These litter items are represented by a limited number of pixels in images, resulting in insufficient details, that hinders their accurate detection with common object detection models (e.g., Faster R-CNN and YOLO). Specifically, the input images are usually resized to a smaller size (e.g., 640 × 640 pixel for YOLO network) by DL models before model training and inference, which causes small items to appear even smaller, further complicating detection (van Emmerik et al., 2025).

To address the above three challenges, we proposed a SSL-based framework combined with the Slicing Aided Hyper Inference (SAHI) method for measuring cross-sectional floating litter fluxes in river systems. The SSL-based framework can effectively quantify fluxes with the limited availability of labeled data for model development. Additionally, the SAHI method enhances the detection of small litter by slicing input images into small tiles and resizing them to a larger dimension. We developed and validated this framework using images collected from canals and waterways in the Netherlands, Indonesia, and Vietnam.

2. Material and methods

2.1. Methodology

2.1.1. Overview of the semi-supervised learning-based framework for quantifying litter fluxes

Fig. 1 shows the proposed SSL-based framework for quantifying cross-sectional floating litter fluxes in flowing rivers. This framework includes four steps: (a) collecting data from locations of target rivers with digital cameras; (b) developing a DL model for litter detection using SSL methods; (c) applying the DL model to detect and count litter items in each collected image; and (d) post-processing the detection results to quantify litter fluxes. This framework can be flexibly integrated into real-world monitoring campaigns. This framework reduces the reliance on human visual observation, and enables scalable, high-frequency, long-term monitoring of floating litter transport in river networks. In step (b), we can develop models using existing openly available plastic datasets (e.g., see Section 2.2), parts of data from target rivers, or a combination of both. We described the details on data collection in Section 2.1.2. The development and application of the DL model for litter detection is presented in Section 2.1.3. Section 2.1.4 gives details on the litter flux estimation.

2.1.2. Data collection from target rivers

In this framework, we used digital cameras to capture images at multiple sampling points on an infrastructure (e.g., a bridge) of target rivers over the river surface (see Fig. 1(a)). Due to their affordable costs and user-friendliness, digital cameras are commonly used to collect data from any water body polluted by macroplastic litter, compared to other devices, e.g., drones (Jia et al., 2023a). Thus, we selected such device to enhance the practical applicability of the proposed framework. The cameras can either be: (1) fixed on the bridge at each sampling point for continuous monitoring, or (2) handheld for a pre-defined period to survey multiple sampling points, both with a time-lapse recording. Among these, fixed cameras are more suitable for long-term structured monitoring, as they can be deployed to automatically capture images at pre-defined time periods and frequencies over extended periods of time, while requiring little equipment maintenance (e.g., camera power supply). Their installation and removal are relatively easy and time-saving. The time-lapse interval (seconds per frame) is determined based on the actual river plastic flow rate, ensuring that all floating litter

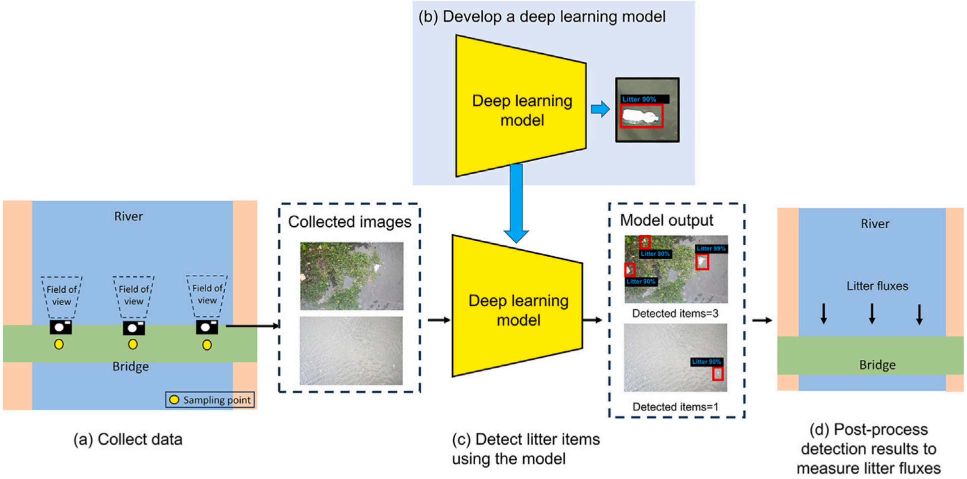


Fig. 1. The schematic illustration of deep leaning-based framework for quantifying cross-sectional floating litter fluxes. First, we used digital cameras to collect images at multiple sampling points on a bridge over the river surface (a). These images capture all floating litter items in camera's field of view. Second, we developed a deep learning model for litter detection using a semi-supervised learning method (b). Third, we used the developed model to detect litter from the collected images, providing the number of items detected in each image (c). Lastly, we post-processed the detection results to measure cross-sectional floating litter fluxes (d).

items within the observation area are captured in images. The width of the observation area, that is smaller than the full river width, depends on the camera's field of view and the height of the bridge above the water surface. Each sampling point can be measured multiple times during a pre-defined period $\Delta t_{i,m}$ [h] on one measurement day (van Emmerik et al., 2022a). For example, van Emmerik et al. (2025) mounted a single camera at 5 sampling points on three bridges along the Saigon River. They captured 31 images at 10-s intervals, up to 8 times for each sampling point.

2.1.3. Model development and application for litter detection

Given that the promising detection performance of SSL methods, we adopted the same SSL approach from Jia et al. (2024b) to develop a robust model for litter detection. Then, we applied a SAHI method (Akyon et al., 2022) to enhance the model's generalization to small litter in target rivers. In this study, we did not develop DL models capable of automatically identifying and counting the same litter item appearing in multiple consecutive images as a single instance. Thus, we manually reviewed the detected litter items and corrected the counts before estimating floating litter fluxes.

Semi-supervised learning approaches for litter detection. Fig. 2 presents the schematic diagram of the SSL approach based on Swapping Assignments between multiple Views of the same image (SwAV) (Caron et al., 2020). The method involves two phases: self-supervised pre-training and supervised fine-tuning stage. In the first phase, we employed SwAV to pre-train a Residual Network with 50 layers (ResNet50) (He et al., 2016), and a projection head, using a large set of unlabeled data. SwAV is a self-supervised learning method that enables models to learn data representations using a "swapped" prediction mechanism (see Fig. 2(a)). The detailed information of SwAV is shown in Appendix A. To develop the final model for object detection, we constructed the Faster R-CNN architecture (Ren et al., 2015) by integrating additional DL networks to the pre-trained ResNet50. Finally, we fine-tuned the Faster R-CNN using a limited amount of labeled data in a supervised fashion, to perform the specific downstream task for detecting floating litter. The labeled images contain manually annotated bounding boxes, indicating the class and location of each litter item. In this task, the Faster R-CNN identifies the locations of litter items using bounding boxes along with confidence scores, representing the probability assigned by Faster R-CNN to each bounding box. The detailed information of Faster R-CNN is shown in Appendix A.

Slicing aided hyper inference for small litter detection. Fig. 3 shows the schematic illustration of the SAHI method (Akyon et al., 2022) for detecting floating litter. First, the SAHI method slices the original input image into smaller overlapping tiles with a width of W_s and height of H_s (e.g., 400 × 400 pixels) with an overlap ratio. For simplicity, Fig. 3(a) shows slicing process with the overlap ratio of 0. Then, each sliced tile is resized into a larger dimension with a weight of W_r and height of H_r. Each resized tile is fed into the Faster R-CNN. Finally, the predictions in tiles (i.e., the yellow bounding boxes in Fig. 3(c)) are mapped back to the original input image dimensions. The SAHI method employs Non-Maximum Suppression (NMS) to refine duplicate predictions for the same object in overlapping regions of adjacent tiles (Hosang et al., 2017). The NMS measures the overlap between the predicted bounding boxes in overlapping regions using Intersection over Union (IoU), and filters out redundant boxes with higher IoU overlap than a predefined IoU NMS threshold, retaining the boxes with confidence score higher than a certain confidence threshold (Akyon et al., 2022). The detailed description of IoU is shown in Appendix A.

2.1.4. Litter flux estimation

We post-processed the detection results from the DL model to quantify cross-sectional floating litter fluxes. First, we calculated the mean litter fluxes f_i [items/h] for sampling point i, using the following equation (Schreyers et al., 2023):

$$f_i = \frac{1}{M_i} \sum_{m=1}^{M_i} \frac{N_{i,m}}{\Delta t_{i,m}} \tag{1}$$

where $N_{i,m}$ [items] is the total number of litter items detected by the model in the images collected at sampling point i during the mth measurement within the time period $\Delta t_{i,m}$ [h]. M_i denotes the total number of sampling events at sampling point i.

Then, we calculated the total cross-sectional floating litter fluxes F [items/h] using the following equation, as derived from (van Emmerik et al., 2022a):

$$F = \frac{1}{S} \sum_{i=1}^{S} \frac{f_i}{w_i} \cdot W \tag{2}$$

where S is the total number of sampling points in a bridge. w_i [m] is the width of the observation area at sampling point i. W [m] is the total river width.

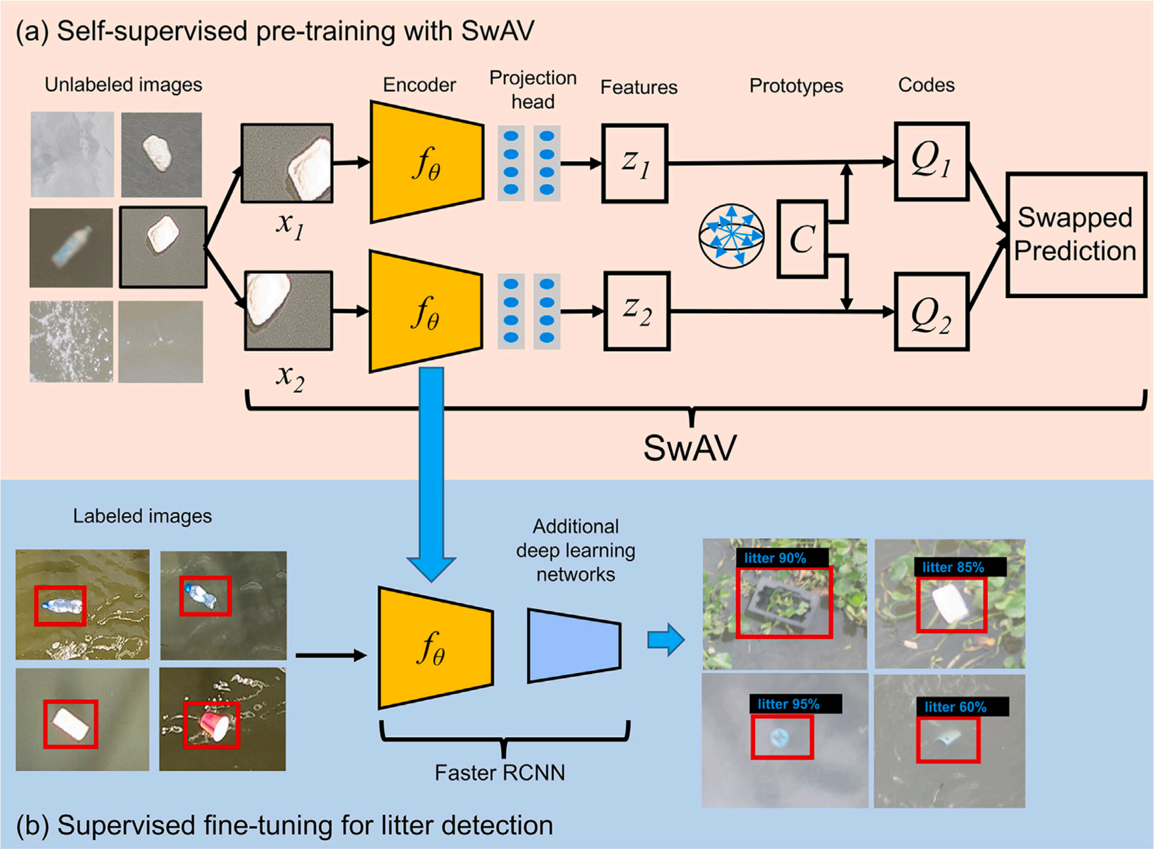


Fig. 2. The schematic diagram of the two-stage semi-supervised learning method.

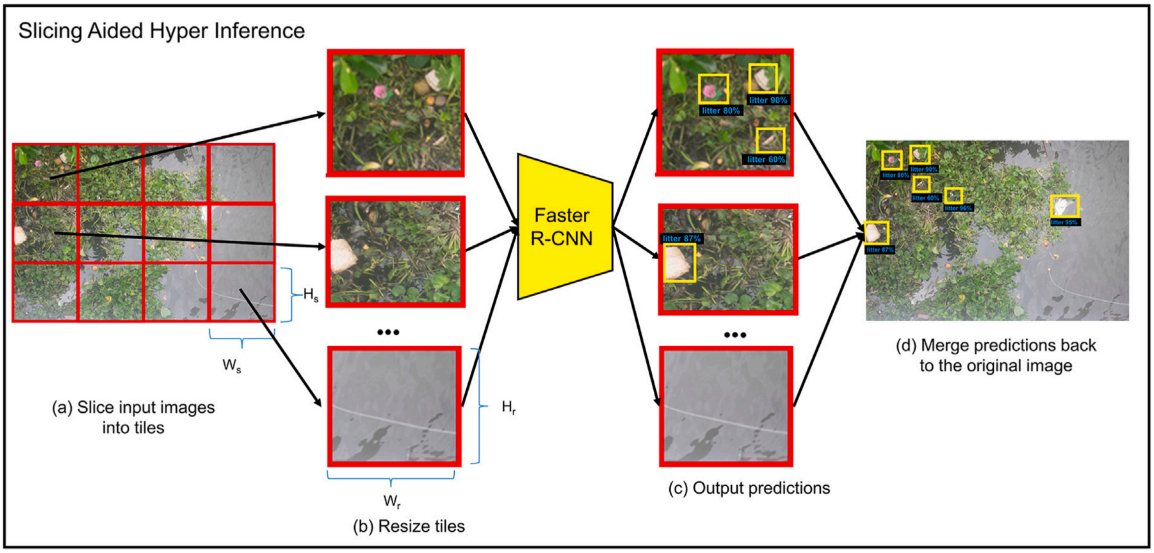


Fig. 3. The schematic illustration of Slicing Aided Hyper Inference (SAHI) for detecting floating litter. First, the SAHI method divides the input images into smaller (overlapping) tiles (a), and resizes them into a larger scale (b). Then, we used the Faster R-CNN to detect litter in each resized tile (c). Finally, these detections (yellow bounding boxes) are merged back to the original input image (d). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Datasets and case study

We developed litter detection models using data from six locations: (1) The TU Delft - Green Village (TUD-GV), the Netherlands (Jia et al., 2023b), (2) Oostpoort, the Netherlands (Jia et al., 2024b), (3) Amsterdam, the Netherlands (Jia et al., 2024b), (4) Groningen, the Netherlands (Jia et al., 2024b), (5) Jakarta, Indonesia (van Lieshout et al., 2020), and (6) Wageningen UR - Ho Chi Minh City (WUR-HCMC) (van

Table 1
Details on datasets for model development.

Name	Collection location	Collection device[a]	Image resolution (pixel × pixel)	Device height (m)	No. images[a]
TUD-GV (Jia et al., 2023b)	Delft, the Netherlands	GoPro Hero 4, GoPro MAX 360	1920 × 1080	2.7, 4	3777
Oostpoort (Jia et al., 2024b)	Delft, the Netherlands	GoCam3, GoPro MAX 360	3840 × 2160, 1920 × 1440	5	562
Amsterdam (Jia et al., 2024b)	Amsterdam, the Netherlands	GoPro Hero 10	5568 × 4176	1–2	92
Groningen (Jia et al., 2024b)	Groningen, the Netherlands	Obscape HQ	2592 × 1944	4	63
Jakarta (van Lieshout et al., 2020)	Jakarta, Indonesia	Dahua Easy4ip	2560 × 1440, 1920 × 1080	4.5	526
WUR-HCMC (van Emmerik et al., 2025)	Ho Chi Minh City, Vietnam	DJI Phantom 4 Pro	5464 × 3070	11–14	935

[a] In these columns, we only reported the number of images we used for model development, and the corresponding collection devices (see Section 2.3.1).

Emmerik et al., 2025). Table 1 provides an overview of these six openly available datasets. Additional information about the actual data used for model development is shown in Section 2.3.1. Readers are referred to the cited publications for more details. Additionally, we evaluated the zero-shot out-of-domain detection performance of the models, as well as the zero-shot out-of-domain flux quantification capability of the proposed framework, using the TU Delft - Ho Chi Minh City (TUD-HCMC) case study.

2.2.1. The TU Delft - Green Village dataset

The TUD-GV dataset was generated by Jia et al. (2023b). They collected images using action cameras and a phone from semi-controlled experiments conducted during 10 days from February to April 2021, in a drainage canal in the TU Delft Campus, the Netherlands.

2.2.2. The Oostpoort dataset

The Oostpoort dataset was created from experiments performed during 26 days between February and March 2022, in a canal at Oostpoort, Delft, the Netherlands (Jia et al., 2024b). Jia et al. (2024b) recorded video sequences using action cameras with a time-lapse recording, and extracted images from these recorded videos.

2.2.3. The Amsterdam dataset

The Amsterdam dataset was generated by Jia et al. (2024b) from a data sampling activity on the 1st March 2023. They sampled images with an action camera in canals and ponds at Amsterdam, the Netherlands.

2.2.4. The Groningen dataset

The Groningen Dataset was generated from experiments conducted between January and September 2023, in a canal in Groningen, the Netherlands (Jia et al., 2024b). Jia et al. (2024b) installed security cameras on a bridge and recorded images with a time-lapse recording.

2.2.5. The Jakarta dataset

The Jakarta dataset is created from experiments conducted from 30 April to 12 May 2018, at five waterways in Jakarta, Indonesia (van Lieshout et al., 2020). The images were recorded using a security camera installed on five bridges.

2.2.6. The Wageningen UR - Ho Chi Minh City dataset

The WUR-HCMC dataset was created by van Emmerik et al. (2025) from WUR. They captured images using drones at the Thanh Ho and Quy Kien locations, as well as action cameras with a time-lapse recording at the Phu Long, Binh Loi and Thu Thiem bridges, along the Saigon River in Ho Chi Minh City, Vietnam, from 6 February to 1 April 2023.

2.2.7. Case study: TU Delft - Ho Chi Minh City

We conducted measurements at the Binh Loi and Thu Thiem bridges across the Saigon River in Ho Chi Minh City, Vietnam, over two days during the wet season in September 2023. Fig. 4 shows the location of these bridges in the Saigon River, and our sampling points on each bridge. The Binh Loi bridge is located in the central part of the city, while the Thu Thiem bridge is situated at the downstream end. Additional information about the Saigon River is provided in Appendix B.

Table 2 shows the details of the measurements. We divided each bridge into five transects, and monitored floating litter at the center of each transect. The length of these transects was carefully selected to ensure that the bridge piers were not visible within the camera's field of view during sampling. All measurements were performed in the southernmost side of bridges during the ebb tide. On each measurement day, we conducted 4 or 6 rounds of measurements. During each round, we captured images (6016 × 4000 pixels) sequentially from the sampling point 1 to 5, using a handheld camera (Pentax K-series) over a period $\Delta t_{i,m}$ of 130 s. The camera was oriented nearly vertically with respect to the water surface, with a time-lapse recording (1 image/10 s). The observation area width for each sampling point is 7 m, and the ground sampling distance (GSD) of each image is 0.12 cm/pixel. Due to instability in the operation of the handheld camera, we only selected the images without heavy blur for measuring litter fluxes, as reported in Table 2. Finally, we built the TUD-HCMC dataset, including the $\mathrm{Test}_{\mathrm{Thu\ Thiem}}$ and $\mathrm{Test}_{\mathrm{Binh\ Loi}}$ subsets with 199 and 309 images collected from the Thu Thiem and Binh Loi bridge, respectively. We annotated litter items in images using bounding boxes to indicate their locations. Since some items appear in multiple consecutive images, the number of annotated litter items exceeds the actual number of litter items in the rivers (see Table 2). Examples of images are shown in Appendix B.

2.3. Experimental procedure

We conducted multiple experiments to investigate the effectiveness of SSL methods for litter detection, and the capability of the proposed SSL-based framework for floating litter flux quantification. In Experiment 1 and 2, we evaluated models' in-domain and zero-shot out-of-domain litter detection performance, respectively. In-domain generalization performance indicates the model performance on new, unseen data collected from the same geographic locations as the training data. In contrast, out-of-domain generalization performance indicates the performance on unseen data sourced from different geographic locations. Zero-shot out-of-domain generalization refers to the capability of DL models to detect previously unseen objects from different geographic locations, without requiring training data of these unseen objects. This capability is especially crucial for large-scale structured monitoring, enabling the monitoring of multiple geographic locations with varying environmental conditions in extensive river system, without well-labeled and location-specific data for further refinement of DL models (Jia et al., 2023a). Moreover, we compared the litter detection results against those obtained from a SL benchmark. In Experiment 2,

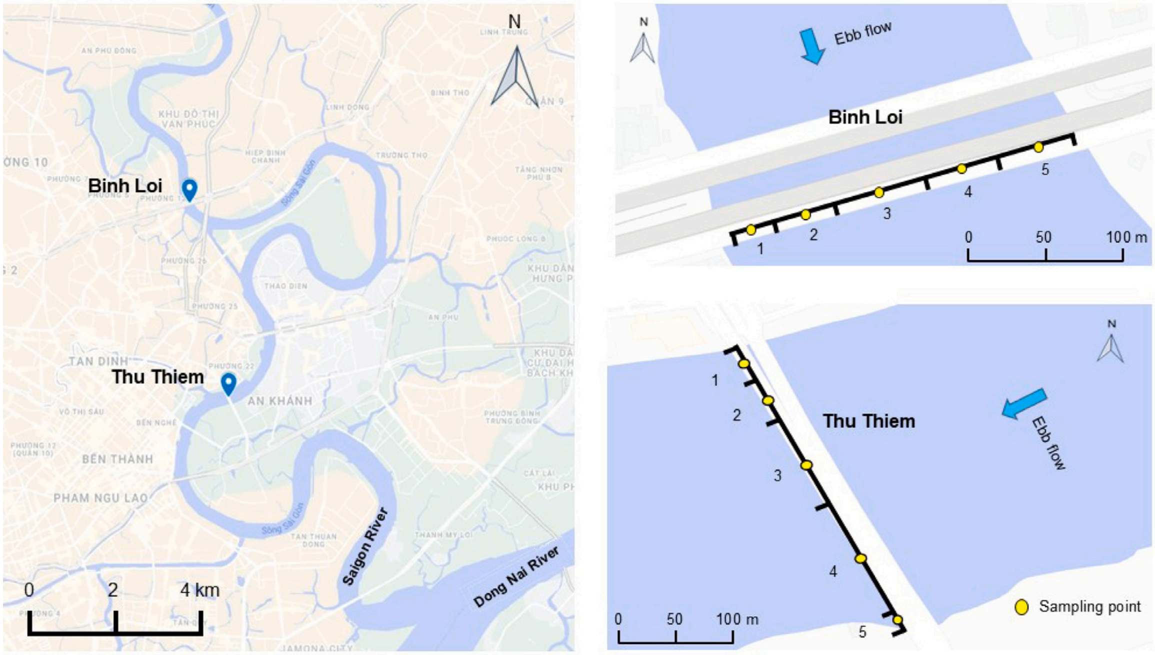


Fig. 4. The location of Binh Loi and Thu Thiem bridges in the Saigon River (left) and sampling points for each bridge (right).

Table 2
Details of the measurements at Thu Thiem and Binh Loi bridges on the Saigon River.

Bridge	River width (m)	Date	Sampling point	Transect width (m)	Observation area width (m)	No. measurement rounds	Sampling duration per point (s)	Time-lapse interval (s)	No. images	No. litter items	No. annotated litter items
Thu Thiem	285	09/09	1 2 3 4 5	35 58 70 60 62	7	4	130	10	199	51	64
Binh Loi	228	12/09	1 2 3 4 5	28 33 69 85 12	7	6	130	10	309	108	114

we evaluated the benefits of the SAHI method in small litter detection. Finally, in Experiment 3, we compared the flux quantification results of the SSL-based framework with those of a SL-based framework and a conventional human visual counting method.

To evaluate model performance for litter detection, we employed four widely used metrics: (1) AP50, which is the Average Precision (AP) at an IoU threshold of 50%, (2) F1-score, (3) precision, and (4) recall (Jia et al., 2023a). Precision, recall, and F1-score are calculated based on true positives (TP), false positives (FP) and false negatives (FN), with the same IoU threshold. TPs and FPs are the number of litter items correctly and incorrectly identified, respectively. FNs are the number of undetected litter items (Jia et al., 2024a). The detailed description of these metrics is shown in Appendix A.

2.3.1. Data selection

For developing models, we randomly selected images from the TUD-GV, Oostpoort, Amsterdam, Groningen, Jakarta and WUR-HCMC dataset, as detailed in the “Total images” column of Table 3. We aimed to evaluate models’ out-of-domain generalization performance to a new TUD-HCMC case study in Experiment 2 and 3. Thus, we only selected 935 images collected at the Quy Kien and Thanh Ho location from the WUR-HCMC dataset for model development. This selection ensures that images from the Binh Loi and Thu Thiem location in the TUD-HCMC dataset remain unseen during model pre-training and fine-tuning. We sliced the selected images into tiles and achieved a total of 501,983 image tiles with a standard size of 224 × 224 pixels, matching the input size required for ResNet50. Example image tiles are shown in Appendix B.

In Experiment 1, we trained and validated models, and evaluated their in-domain generalization capability using these 501,983 image tiles. We randomly sampled tiles to create the non-overlapping subsets, including (1) 499,477 tiles (99.5%) for SwAV pre-training, (2) 1128 tiles for supervised fine-tuning, (3) 125 tiles for model validation, and (4) 1253 tiles for model testing, as outlined in Table 4. We used a maximum of 499,477 tiles without annotations for SwAV pre-training (Train_{500k}). To further investigate model performance regarding to the availability of unlabeled data for SwAV pre-training, we generated five additional smaller pre-training subsets by gradually reducing the number of tiles down to 25k (Train_{300k} to Train_{25k}). We used up to 1128 image tiles for fine-tuning SSL and baseline SL models in a supervised manner ($\text{Train}_{100\%}$). These tiles contain 1349 litter items annotated by bounding boxes indicating their locations, without further classification. To evaluate model performance with respect to the availability of labeled data, we generated two smaller fine-tuning subsets by decreasing the number of labeled tiles down to 20% ($\text{Train}_{60\%}$ and $\text{Train}_{20\%}$). We used a maximum of 125 tiles containing 158 annotations to validate models ($\text{Validation}_{100\%}$), with a 9:1 ratio relative to the tiles for fine-tuning ($\text{Train}_{100\%}$). For consistency, we generated two smaller

Table 3
Details on the images for model development.

Image source	Total images	Total image tiles	No. image tiles without labels	No. image tiles with litter annotated	No. annotated litter items
TUD-GV	3777	91,565	90,112	1453	1719
Oostpoort	562	78,043	77,710	333	342
Amsterdam	92	36,864	36,712	152	204
Groningen	63	5350	5193	157	167
Jakarta	526	16,789	16,433	356	501
WUR-HCMC	935	273,372	273,317	55	63
Total	5955	501,983	499,477	2506	2996

Table 4
The subsets for model development in Experiment 1.

Learning method	Training dataset			Validation dataset			Test dataset		
	Name	No. annotated litter items	No. tiles	Name	No. annotated litter items	No. tiles	Name	No. annotated litter items	No. tiles
Self-supervised	Train_{500k}	0	499,477						
	Train_{300k}	0	299,679						
	Train_{200k}	0	203,454						
	Train_{100k}	0	99,887						
	Train_{50k}	0	49,941						
	Train_{25k}	0	24,966						
Semi-supervised and supervised	$\text{Train}_{100\%}$	1349	1128	$\text{Validation}_{100\%}$	158	125	Test	1489	1253
	$\text{Train}_{60\%}$	800	677	$\text{Validation}_{60\%}$	91	75			
	$\text{Train}_{20\%}$	276	226	$\text{Validation}_{20\%}$	33	25			

validation subsets ($\text{Validation}_{60\%}$ and $\text{Validation}_{20\%}$). We generated the Test subset with 1253 tiles and 1489 annotations for testing models' in-domain generalization performance.

2.3.2. Experiment 1: In-domain detection performance

With the first experiment, we assessed the benefits of (1) varying pre-training epochs, and (2) varying pre-training dataset sizes on the in-domain generalization performance of SSL models. This examination is essential for assessing the effectiveness of representations learned from different scales of pre-training dataset for generalization. It also offers insights into the effectiveness of SSL methods in scenarios with limited labeled samples, but with abundant unlabeled images and sufficient computational resources for extensive hyperparameter tuning.

For developing SSL models, we first initialized the ResNet50 backbone with ImageNet weights, and then using SwAV to pre-train all the layers of the ResNet50 network with a projection head of 2-layer multilayer perceptron on all six pre-training subsets, (i.e., Train_{500k} to Train_{25k} subset) in the self-supervised learning stage. Due to the limited computational resources, we performed SwAV pre-training for 100, 200, and 300 epochs (Caron et al., 2020; Chen et al., 2020). In the supervised fine-tuning phase, we fine-tuned the Faster R-CNN architecture derived from the SSL backbone on $\text{Train}_{100\%}$ to $\text{Train}_{20\%}$ subset. During fine-tuning, we froze the first four convolutional blocks of the ResNet50 backbone network (see Appendix A). It allows the Faster R-CNN to retain relevant low-level features (e.g., edges and texture) in the first two blocks, as well as high-level features (e.g., object shapes) in the last two blocks, learned from SwAV pre-training. Most important, these high-level features significantly improve the model's in-domain and out-of-domain generalization performance in data scarce conditions (Jia et al., 2024b). Model validation was conducted on the respective Validation subsets, i.e., $\text{Validation}_{100\%}$ to $\text{Validation}_{20\%}$ subset. We selected the SSL model that achieved the highest validation accuracy across the three different pre-training epoch settings. Then, we evaluated its in-domain performance on the Test subset.

We compared the effectiveness of SSL models with baseline SL models, developed with the supervised fine-tuning phase (see Fig. 2(b)), but without the SwAV pre-training phase (see Fig. 2(a)). These SL models are Faster R-CNNs supervised fine-tuned on images with annotated litter, with ResNet50 backbones initialized using ImageNet weights. During fine-tuning, the first four convolutional blocks of the ResNet50 backbone network were frozen. We chose ImageNet weights because transferring data representations obtained from the ImageNet classification task to other domains is a common approach for litter detection (Jia et al., 2023a). They were fine-tuned, validated, and tested on the same subsets used for SSL model development. The training setup and procedure for SSL and SL models are shown in Appendix A.

2.3.3. Experiment 2: Zero-shot out-of-domain detection performance

To evaluate the zero-shot out-of-domain generalization performance for litter detection, we tested the best-performing SSL and SL model developed in Experiment 1, on the $\text{Test}_{\text{Thu Thiem}}$ and $\text{Test}_{\text{Binh Loi}}$ subsets, as outlined in Table 2. We did not re-train these models on any data from the Thu Thiem and Binh Loi location.

Evaluation of the SAHI methods. We compared performance of the SSL model using the SAHI method and that without SAHI during model inference on $\text{Test}_{\text{Thu Thiem}}$ and $\text{Test}_{\text{Binh Loi}}$ subsets. Inspired by Akyon et al. (2022) and Gia et al. (2024), we tested four configurations of width W_s and height H_s for the selected SSL model: (1) 400 × 400, (2) 640 × 640, (3) 1280 × 1280, and (4) 1920 × 1920 pixels. The configuration yielding the best detection performance for each subset was selected for subsequent steps of this experiment.

When applied to detect litter in the TUD-HCMC case study, models may produce a high number of misdetections, due to the limited data available for SwAV pre-training and supervised fine-tuning (Jia et al., 2024b). To reduce these misdetections, we refined the output bounding boxes by setting a high confidence threshold value before making the final predictions and computing performance metrics. This threshold defines the minimum confidence level required for a detected object to be considered as a valid detection. Increasing this threshold excludes low-confidence predictions, but may also result in missing some true positives with confidence scores below the threshold. Thus, we compared the SSL model's performance using three confidence threshold values (0.5, 0.7 and 0.9) with the best W_s and H_s settings. The confidence threshold value yielding the best performance for each subset was chosen for following steps of this experiment.

It is noted that selecting optimal hyperparameters based on test performance is not a standard practice in machine learning. However, the aim of this experiment was to evaluate the benefit of the SAHI method, while utilizing as much data from TUD-HCMC case study for testing as possible.

Evaluation of the SSL and SL methods. To minimize the influence of randomization, we repeated the fine-tuning process for a total of 10 times for both SSL and SL models. Then, we evaluated the detection performance of all models, using the SAHI method with W_s, H_s and confidence threshold settings that yielded the best performance in the previous evaluation, ensuring that the pre-processed input images by SAHI were the same before being fed to both SSL and SL models.

2.3.4. Experiment 3: Litter flux measurement

To evaluate the zero-shot out-of-domain flux quantification capability of the proposed SSL-based framework, we used the best-performing SSL models for the Thu Thiem and Binh Loi locations from the 10 runs conducted in Experiment 2. We estimated floating litter fluxes, using the approach introduced in Section 2.1.4. Additionally, we evaluated flux quantification capability of the SL-based framework similarly, but replacing the SSL model with the best-performing SL model from 10 runs, as illustrated in Fig. 1(a) and (c). Furthermore, we compared these results against those obtained using the conventional human counting method, where litter items were manually observed and counted directly from the images. We used the Pearson correlation coefficient (r) (Benesty et al., 2009) to assess the linear correlation between fluxes measured by DL-based frameworks (i.e., the SSL- and SL-based framework) and human counting methods across 10 sampling points in the case study. This coefficient ranges from −1 to 1. A higher positive value indicates a stronger positive correlation between two variables. The reader is referred to the work of Benesty et al. (2009) for more details on this coefficient. Litter items appearing in multiple consecutive images were counted only once across all methods and frameworks.

3. Results and discussion

3.1. Experiment 1: In-domain detection performance

3.1.1. SSL model performance for varying pre-training epochs

Table 5 presents the AP50 detection performance of the SSL methods on the Validation$_{100\%}$ subset, evaluated with varying pre-training epochs and pre-training dataset sizes. Results for the Validation$_{60\%}$ and Validation$_{20\%}$ subsets are shown in Appendix C. We observed that increasing the pre-training epochs from 100 to 200 usually leads to an improvement in model performance, as indicated by AP50 improvements ranging from 0.2% to 4.2%, while an additional 100 pre-training epochs requires substantial computational resources (e.g., 278 h per 100 epochs on the Train$_{500k}$ subset). This finding is similar to that reported by Caron et al. (2020). The authors pre-trained the ResNet50 using SwAV for 100, 200, 400, and 800 epochs on 1.28 million unlabeled images from the ImageNet dataset. Their results demonstrate a 3.2% improvement in top-1 accuracy on the ImageNet classification task as pre-training epochs increase from 100 to 800. Furthermore, we found that this improvement is more noticeable, when a large amount of data is available for pre-training. For example, the SSL models pre-trained on Train$_{200k}$ and Train$_{500k}$ achieve an AP50 improvement ranging from 3.4% to 4.2% by increasing epochs, while the SSL models pre-trained on Train$_{50k}$ and Train$_{100k}$ only obtain a AP50 improvement ranging from 0.2% to 0.4%. We attribute this superior performance to the more robust feature representations learned from SwAV pre-training from a larger amount of data for longer training time, which enhance the performance of Faster R-CNN for the downstream litter detection task.

Table 5 also demonstrates a decline in AP50 ranging from 0.3% to 3.1%, when epochs increase from 200 to 300. It could be attributed to the limited size of pre-training dataset (500k images). Caron et al. (2020) reported improved performance with longer pre-training time, but used a significantly larger dataset (1.28 million). Another reason is the single pre-training run conducted, due to computational limitations. The inherent stochasticity of neural network training leads to variations in results across multiple runs, potentially affecting the observed performance (Punjani and Fleet, 2021).

3.1.2. Performance for varying pre-training dataset sizes

The benefit of larger pre-training dataset on model performance is more noticeable from the results shown in Fig. 5, that shows in-domain generalization performance of the SSL and baseline SL methods on the Test subset, with varying proportion of labeled data for fine-tuning. It reveals a general upward trend in AP50 and F1-score for SSL models, as the pre-training dataset size increases, irrespective of the amount of labeled data available for fine-tuning. The performance improvement is particularly noticeable when scaling the pre-training dataset from a small size (<100k) to a larger size, with AP50 increasing by 5.6% to 14.7% and F1-score improving by 0.06 to 0.25. For instance, when models are fine-tuned on the Train$_{100\%}$ subset, the AP50 improves from 76.3% to 82.3%, and the F1-score increases from 0.69 to 0.75, as the pre-training dataset size increases from 25k to 500k. These findings underscore the advantages of large-scale datasets, enabling models to learn more effective low-level and high-level representations. This improvement is especially significant in scenarios with limited labeled data for fine-tuning (i.e., Train$_{20\%}$), where AP50 increases by 14.7% and F1-score improves by 0.25.

We observed a performance plateau when increasing the pre-training dataset size from 100k to 500k. It could be attributed not only to the limited size of the pre-training dataset (500k) and the limited pre-training epochs, but also to the constraints of conducting only a single training run, imposed by computational resource limitations. Literature demonstrates notable performance improvements for SSL models with larger SwAV pre-training datasets, scaling from 1.28 million to over 1 billion images (Goyal et al., 2022). Thus, we believe that better performance could be achieved by scaling the unlabeled dataset size to over 1 million and conducting a large number of training runs.

Fig. 5 also demonstrate that the most SSL models significantly outperform the baseline SL benchmarks, irrespective of the amount of data used for SwAV pre-training or fine-tuning. The SSL method performs best in most cases, obtaining AP50 values ranging from 59.3% to 82.3%, and F1-scores from 0.48 to 0.77. In comparison, the baseline SL method achieves AP50 values varying from 61.8% to 72.5%, and F1-scores from 0.53 to 0.73. These values are particularly low when fine-tuning data is limited (i.e., Train$_{20\%}$ subsets), where SSL models yield improvements of up to 12% in AP50 and 0.20 in F1-score. The SSL model requires only 20% of the labeled images (226 images with 276 annotated litter items) combined with 500k unlabeled images to achieve comparable or superior performance (AP50 = 74.0%, and F1-score = 0.72) to that of the baseline SL method, which relies on 100% of labeled images (1128 images with 1349 annotated litter items, AP50 = 72.5%, and F1-score = 0.73). These findings highlight the benefits of transferring low-level and high-level representations learned by SwAV from unlabeled yet domain-relevant data, leading to notable improvements compared to simple transfer from ImageNet. Jia et al. (2024b, 2025) also reported a similar finding. While the features extracted from ImageNet are general, they are not sufficiently relevant to the specific litter detection task.

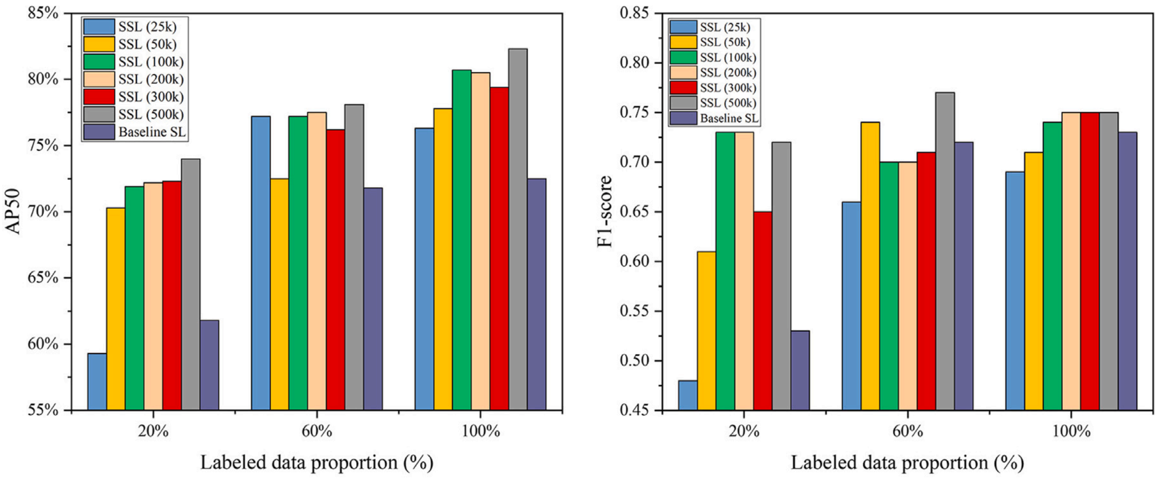
3.2. Experiment 2: Zero-shot out-of-domain detection performance

3.2.1. Performance for SAHI methods

Tables 6 and 7 present the performance of SSL models with or without SAHI methods on the Test$_{\text{Thu Thiem}}$ and Test$_{\text{Binh Loi}}$ subset, respectively. The SSL model achieving the highest AP50 on the Test subset in Experiment 1, was selected for these evaluations (i.e., model pre-trained on the Train$_{500k}$ subset and fine-tuned on the Train$_{100\%}$ subset). The results demonstrate that SSL models using SAHI methods significantly outperform those without SAHI in all metrics for the Thu Thiem location, and in recall and F1-score for the Binh Loi location under the same confidence threshold settings (0.5). Especially, the SAHI method achieves an improvement in F1-score of up to 0.19, compared to models without SAHI across two locations.

Table 5
Pre-training time and validation accuracy (AP50) on the $\text{Validation}_{100\%}$ subset of all SSL models for Experiment 1. The bold entities are the best results for models pre-trained on each pre-training dataset.

Pre-training dataset	No. pre-training epochs	Pre-training time (h/100 epochs)	AP50
Train_{25k}	100	17	**80.2%**
	200		78.8%
	300		77.1%
Train_{50k}	100	33	80.2%
	200		**80.4%**
	300		79.4%
Train_{100k}	100	56	81.8%
	200		**82.2%**
	300		81.9%
Train_{200k}	100	117	78.0%
	200		**82.2%**
	300		80.7%
Train_{300k}	100	168	82.4%
	200		**82.9%**
	300		81.0%
Train_{500k}	100	278	80.2%
	200		**83.6%**
	300		80.5%



(a) AP50 detection performance

(b) F1-score detection performance

Fig. 5. AP50 (a) and F1-score (b) detection performance of the SSL and baseline SL methods on the Test subset with different proportion of labeled data for fine-tuning. The six SSL models were pre-trained on Train_{25k}, Train_{50k}, Train_{100k}, Train_{200k}, Train_{300k}, and Train_{500k} subset, respectively.

Table 6
Confusion matrix, Precision, Recall and F1-score on the $\text{Test}_{\text{Thu Thiem}}$ subset for SSL models, evaluated with varying inference hyperparameters (i.e., W_s, H_s and confidence threshold score). The model was fine-tuned on the $\text{Train}_{100\%}$ subset. The bold entity is the best F1-score.

$W_s \times H_s$ (pixel × pixel)	Confidence threshold	TP	FP	FN	Precision	Recall	F1-score
No SAHI	0.5	0	6	64	0.00	0.00	0.00
400 × 400	0.5	21	838	43	0.02	0.33	0.05
640 × 640	0.5	27	539	37	0.05	0.42	0.09
1280 × 1280	0.5	22	95	42	0.19	0.34	0.24
	0.7	21	74	43	0.22	0.33	0.26
	0.9	19	40	45	0.32	0.30	**0.31**
1920 × 1920	0.5	9	32	55	0.22	0.14	0.17

Tables 6 and 7 also present the performance of SSL models with SAHI, along with the best W_s and H_s settings under varying confidence thresholds. Increasing confidence threshold from 0.5 to 0.9 usually yields a slight decline in TP and a significant reduction in FP, since a large number of low-confidence FPs are filtered out. This adjustment leads to a slight decrease in recall, but a notable improvement in precision and F1-score. For example, the model with SAHI (W_s, H_s = 1280 pixel) achieves a substantial increase in precision of 0.13 and F1-score of 0.07, with a minor decrease in recall of 0.04 for the Thu Thiem location, when the confidence threshold is raised from 0.5 to 0.9.

For the Thu Thiem location, the model without SAHI fails to detect any litter items correctly (TP = 0) and produces 6 FPs, resulting in precision, recall, and F1-score values of 0. In contrast, the SAHI method

Table 7

Confusion matrix, Precision, Recall and F1-score on the Test$_{\text{Binh Loi}}$ subset for SSL models, evaluated with varying inference hyperparameters (i.e., W_s, H_s and confidence threshold score). The model was fine-tuned on the Train$_{100\%}$ subset. The bold entity is the best F1-score.

$W_s \times H_s$ (pixel × pixel)	Confidence threshold	TP	FP	FN	Precision	Recall	F1-score
No SAHI	0.5	7	22	107	0.24	0.06	0.10
400 × 400	0.5	49	3438	65	0.01	0.43	0.03
640 × 640	0.5	68	2749	46	0.02	0.60	0.05
1280 × 1280	0.5	69	625	45	0.10	0.61	0.17
1920 × 1920	0.5	39	229	75	0.15	0.34	0.20
	0.7	36	153	78	0.19	0.32	0.24
	0.9	30	70	84	0.30	0.26	**0.28**

under the same confidence threshold settings (0.5) correctly detects 9~27 litter items (TP) depending on the W_s and H_s settings, achieving higher precision (0.02~0.22), recall (0.14~0.42), and F1-score (0.05~0.24), while it generates a significant number of false positives (FP = 32~838). For the Binh Loi location, the model without SAHI correctly detects only a few litter items (TP = 7) and generates few FPs (22), resulting in a precision of 0.24, but with very low recall (0.06) and F1-score (0.10). In contrast, the SAHI method detects a significantly higher number of litter items (TP = 39~69), but also generates a large number of FPs (229~3438). This leads to lower precision (0.01~0.15), but higher recall (0.34~0.61) and F1-score (0.03~0.20), compared to those obtained by the model without SAHI.

Fig. 6 shows the area of all litter items correctly detected by SSL models with or without SAHI method in the Test$_{\text{Thu Thiem}}$ (W_s, H_s = 1280 pixel) and Test$_{\text{Binh Loi}}$ (W_s, H_s = 1920 pixel) subset. The area of each litter item is approximately calculated by multiplying its ground-truth bounding box area (pixel^2) by the square of the GSD of images (cm^2/pixel^2). The results show that the model without the SAHI method only correctly detect 7 "big" litter items with area above 1000 cm^2 (see Fig. 7(e)), while fails to detect all "small" litter items with area below 1000 cm^2. In contrast, the model with the SAHI method not only correctly identifies these 7 "big" litter items, but also detects 9 additional "big" litter items, and 45 additional "small" litter items. Visual inspection of the predicted bounding boxes highlights the effectiveness of the SAHI method in handling diverse object sizes, as shown for examples in Fig. 7. The accurate detection of "small" litter by SAHI methods can be primarily attributed to its slicing and resizing process (see Section 2.1.3), which enlarges these objects, thereby providing sufficient details for the model to recognize them effectively.

These findings provide clear evidence that the SAHI method can correctly detect a large number of litter items with area smaller than a specific threshold, that models without the SAHI method fail to identify. To the best of our knowledge, no study has reported the precise value of specific threshold, as it depends on the GSD, that is determined by sensor elevation and properties (Andriolo et al., 2023). For example, a CNN model without the SAHI method may correctly detect a plastic bottle in images captured by sensors at a low elevation, where the GSD is low and the bottle appears relatively large (i.e., represented by many pixels). However, this model may fail to detect the same bottle in images taken by the same sensors at a higher elevation, where the GSD is higher, making the bottle appears relatively small (i.e., represented by fewer pixels). This issue arises from the lack of scale invariance in CNNs, that refers to a model's ability to maintain consistent outputs regardless of object scale (Singh and Davis, 2018).

3.2.2. Performance for SSL and SL methods

Fig. 8 illustrates the zero-shot generalization performance of SSL and baseline SL models with SAHI on the unseen Thu Thiem (W_s, H_s = 1280 pixel) and Binh Loi (W_s, H_s = 1920 pixel) location. The confidence threshold is 0.9. These models were fine-tuned on the Train$_{100\%}$ subset across 10 runs. The results demonstrate that SSL methods significantly outperform the baseline SL methods across all metrics for both locations. For the Thu Thiem location, the SSL method achieves substantial improvements of 0.25 in median precision, 0.11 in median recall and 0.14 in median F1-score, compared to the baseline SL method. Similarly, for Binh Loi location, the SSL method show enhancement of 0.09 in median precision, 0.09 in median recall, and 0.07 in median F1-score. Additionally, we conducted a one-way analysis of variance (ANOVA) (Sawyer, 2009) to compare the statistical significance of performance differences between the two methods. The results show that the p-values for all comparisons are lower than 0.05 (see Fig. 8). These findings indicate that the observed performance differences are statistically significant, and unlikely to be due to random chance.

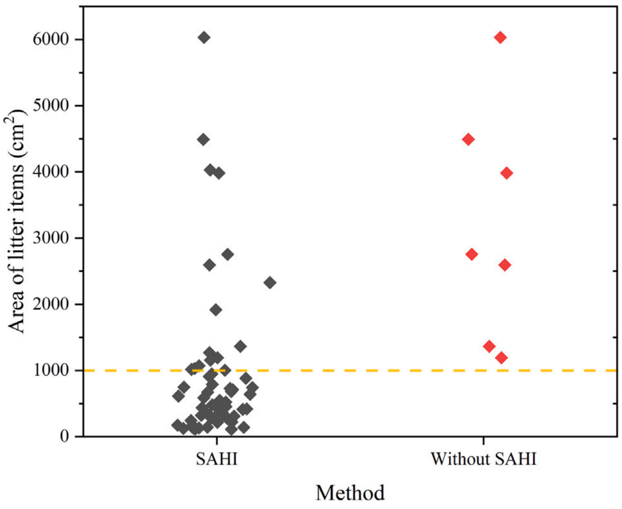


Fig. 6. The areas of litter items correctly detected by SSL models with or without SAHI method in the Test$_{\text{Thu Thiem}}$ (W_s, H_s = 1280 pixel) and Test$_{\text{Binh Loi}}$ (W_s, H_s = 1920 pixel) subset. The confidence threshold is 0.5.

The superior performance of SSL methods is further reflected by visual inspection of the predicted bounding boxes, as depicted in Appendix D. The baseline SL method yields few correct detections and a high misdetection probability, particularly with respect to water hyacinth, reflective elements on the river surface, and other disturbances. These findings indicate that SSL models learn a broader set of features from both unlabeled and labeled domain-relevant data along with ImageNet, compared to baseline SL models that only learn features from labeled data and ImageNet. Due to the richer features, SSL models demonstrate superior out-of-domain generalization to new environments, compared to SL models (Jia et al., 2024b).

While the SSL methods achieve much better performance than the baseline SL methods, all metric values remain low for practical application purposes in the TUD-HCMC case study. Specifically, the SSL method obtains a median F1-score of 0.33 and 0.16 for the Thu Thiem and Binh Loi location, respectively. After conducting a quantitative analysis of litter items in TUD-HCMC images, we found that

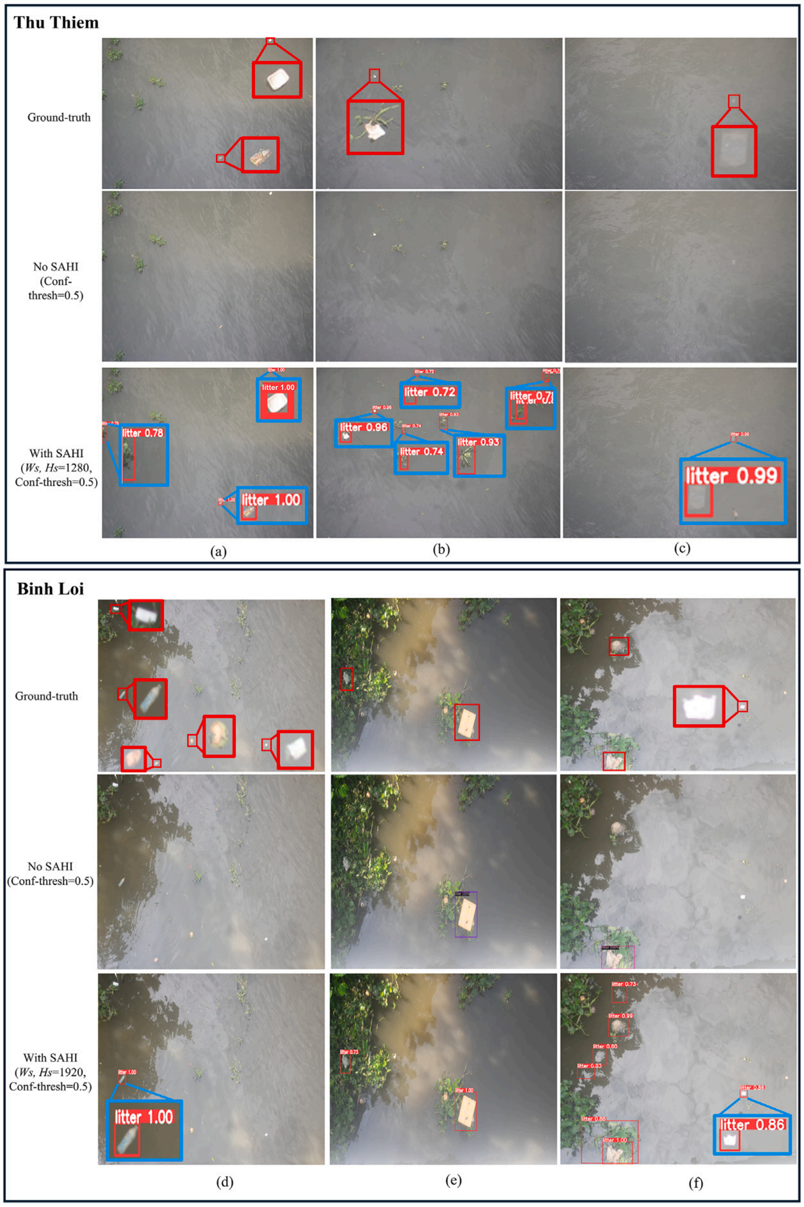
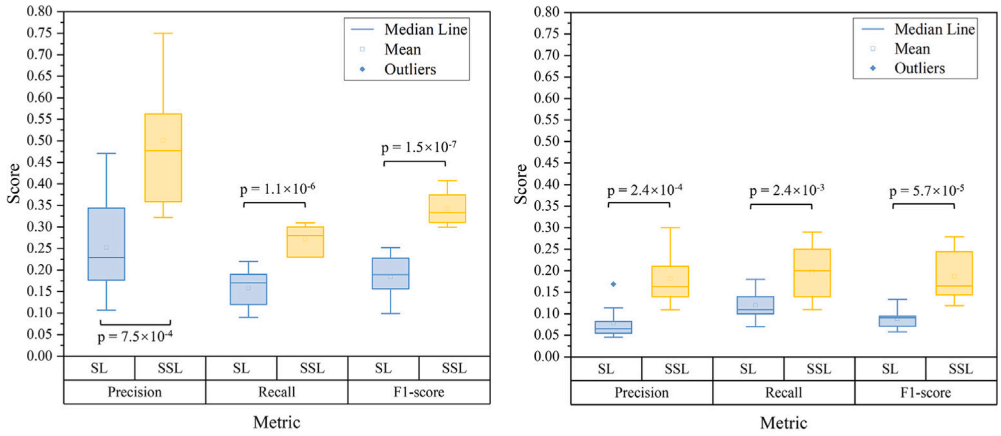


Fig. 7. Examples of predicted bounding boxes for the Faster R-CNN model with and without the SAHI method on the Test$_{\text{Thu Thiem}}$ ((a)–(c)) and Test$_{\text{Binh Loi}}$ ((d)–(f)) subsets. We used the SSL method to develop the Faster R-CNN model, that was fine-tuned on the Train$_{100\%}$ subset. During inference, we set W_s and H_s to 1280 pixel for Thu Thiem case and 1920 for Binh Loi case, with a confidence threshold score of 0.5. Without the SAHI method, the model fails to detect all "small" litter items with area below 1000 cm^2 in (a)–(f), while correctly detects two "big" items with area above 1000 cm^2 in (e) and (f). With SAHI, the model correctly detects some "small" items in (a)–(f) as well as two "big" items in (e) and (f). Acronyms used: Confidence threshold (Conf-thresh).

approximately 40% of all litter items are transparent or are entrapped in water hyacinths, as shown in Appendix D. The developed model usually fails to detect transparent litter, due to the insufficient differentiation between the features of transparent litter and water surface. This challenge is further exacerbated under poor lighting conditions. Additionally, the water hyacinths cover large areas of the litter, resulting in insufficient visible details of litter for accurate detection. Such occlusions also alter shape and texture of litter, further complicating recognition. For example, in the Thu Thiem, the SSL model fails to detect the majority of transparent litter items (14 out of 17 cases) and all entrapped items (5 cases). Similarly, in Binh Loi, all transparent litter items (8 cases) and most entrapped items (26 out of 33 cases) remain undetected. Thus, we explained the low recall and in turn the low F1-score, mainly by the failure to detect these two types of litter. An additional contributing factor could be dataset imbalance. While the pre-training dataset maintains a relatively balanced distribution of samples between rivers in the Netherlands (42%) and those in Vietnam (55%), the fine-tuning dataset is highly imbalanced (81% vs. 2%, see Table 3), that hinders the model's generalization capability to rivers in Vietnam.

3.3. Experiment 3: Litter flux measurement

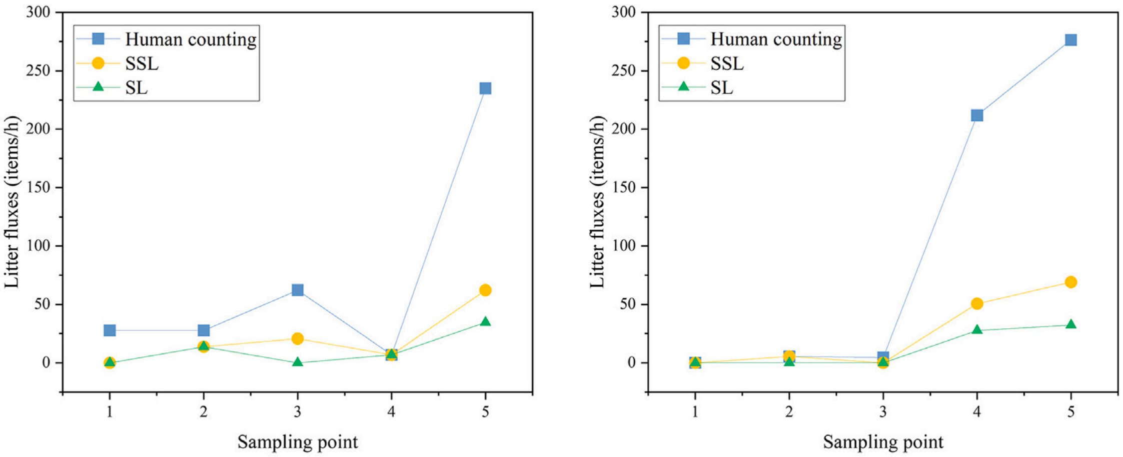
Fig. 9 shows the horizontal distribution of cross-sectional floating litter fluxes, measured by multiple frameworks and methods. We measured fluxes by including the correctly detected litter items by models (i.e., TPs). For this measurement evaluation, we selected the best-performing SSL and baseline SL models (achieving the highest F1-score) across the 10 runs from Experiment 2. The results indicate that the SSL-based and baseline SL-based frameworks yield identical flux measurements for most low-fluxes region (human-measured fluxes < 50 items/h), such as 14 items/h for sampling point 2 at Thu Thiem and



(a) Detection performance at Thu Thiem

(b) Detection performance at Binh Loi

Fig. 8. Zero-shot generalization performance on precision, recall, and F1-score metrics of SSL and baseline SL methods for the two unseen locations: Thu Thiem and Binh Loi bridge. The models were fine-tuned on the $\text{Train}_{100\%}$ subset. P-values for one-way analysis of variance (ANOVA) are presented above (or below) the corresponding boxplots.



(a) Litter fluxes for each sampling point at Thu Thiem

(b) Litter fluxes for each sampling point at Binh Loi

Fig. 9. Horizontal distribution of cross-sectional floating litter fluxes, measured by the SSL-based and baseline SL-based framework, and human counting method. We measured the mean litter fluxes by including the correctly detected litter items by models (i.e., true positives). The SSL and baseline SL models are best-performing models in 10 runs from Experiment 2.

0 items/h for sampling point 1 at Binh Loi (see Fig. 4). However, for the high-flux region (human-measured fluxes > 50 items/h), the SSL-based framework significantly outperforms the SL-based framework by consistently measuring higher fluxes, aligning more closely with those measured by humans. For example, at sampling point 5, fluxes measured by the SSL-based framework is 27 items/h and 37 items/h higher than that by the SL-based framework for the Thu Thiem and Binh Loi bridges, respectively. This is mainly attributed to the higher recall achieved by the SSL models compared to the baseline SL models, as described in Section 3.2.

Fig. 9 also demonstrates that the concentration of litter items is highest near the eastern riverbanks (i.e., sampling point 5), accounting for approximately 70% at Thu Thiem and 60% at Binh Loi. This spatial distribution can be mainly explained by the flow direction and river morphology. Floating litter fluxes are likely highest in the outer curves of the river (van Emmerik et al., 2018), as observed at sampling point 5 for the Binh Loi and Thu Thiem bridges during ebb tides. van Emmerik et al. (2018) also reported a similar spatial distribution of litter fluxes based on measurement taken at 12 sampling points on the Thu Thiem bridge.

Fig. 10 presents the linear fit of fluxes measured by the SSL-based and SL-based framework against those measured via human counting. The results indicate a strong positive correlation between fluxes measured by human and that by DL-based frameworks, regardless of whether the SSL or baseline models are used. However, the SSL-based framework demonstrates a stronger correlation with human counting (the Pearson correlation coefficient $r = 0.99$), compared to the SL-based framework ($r = 0.93$).

Nevertheless, both DL-based frameworks significantly underestimates fluxes in high-flux regions, compared to human measurements. Interestingly, van Lieshout et al. (2020) reported contrasting findings, where DL models estimate relatively higher fluxes for video clips with high litter fluxes, compared to human measurements. The high fluxes, reaching up to 35 items/(min m) in some video clips, poses a significant challenge for human counters to accurately identify and count each transported litter item. Thus, this discrepancy can be attributed to the limitation on how many objects per minute human observers can realistically count. However, human observers in this study did not face such challenge, since we counted litter items directly from images rather than videos, ensuring reliable human measurements. The lower

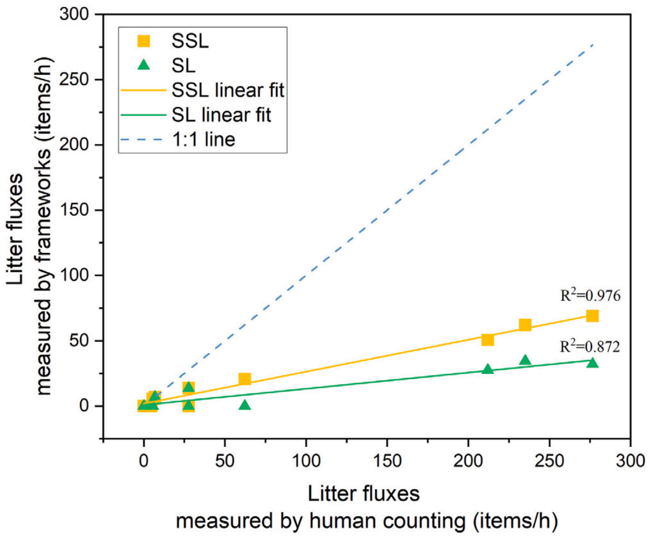


Fig. 10. Comparison of the mean litter fluxes of 10 sampling points with linear fit analysis: SSL-based framework, baseline SL-based framework, and human counting method.

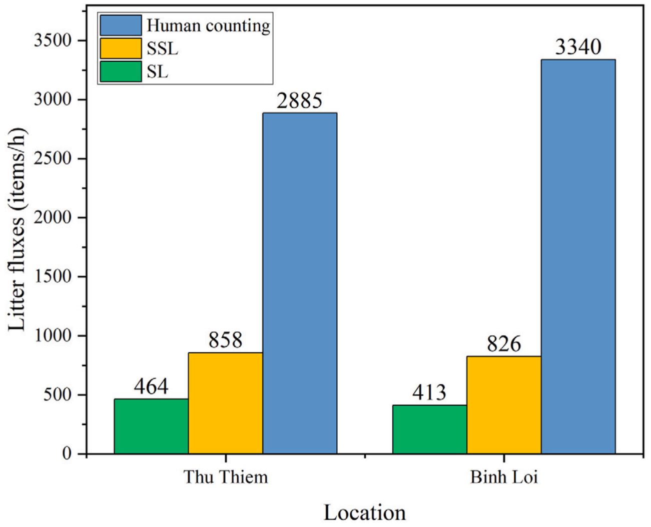


Fig. 11. The cross-sectional floating litter fluxes at the Thu Thiem and Binh Loi bridge, measured by the SSL-based and baseline SL-based framework, and human counting method.

fluxes measured by DL-based frameworks in this study is explained by the low detection accuracy of DL models, as discussed in Section 3.2.2.

Fig. 11 shows the total cross-sectional floating litter fluxes at the Thu Thiem and Binh Loi bridges. Both DL-based frameworks substantially underestimate the fluxes, compared to human counting. Specifically, the fluxes measured by the SSL-based framework is approximately 3 times lower than human measurements at the Thu Thiem bridge, and 4 times lower at the Binh Loi bridge. Despite this underestimation, the fluxes measured by the SSL-based framework (858 items/h at Thu Thiem and 826 items/h at Binh Loi) are nearly double those of the baseline SL-based framework (464 items/h and 413 items/h, respectively). This improvement also highlights the superior capability of the SSL-based framework for flux measurement, compare to the SL-based framework.

Our human-measured cross-sectional fluxes at the Thu Thiem bridge align with the findings of van Emmerik et al. (2019). They reported fluxes measured by human visual counting methods at 12 sampling points on this bridge in 2018, including data from 5 days in September. Using the approach presented in Section 2.1.4, we estimated cross-sectional fluxes, and found that fluxes during ebb tides for these 5 days in September ranged from 1409 to 31,195 items/h. Our measured fluxes (2885 items/h) in this study falls within this range.

3.4. Limitations and future works

The aim of this study is not to deploy and optimize an automated system for floating litter flux measurement, but rather to demonstrate that self-supervision can serve as an effective approach toward developing such a system. The performance improvements achieved by SSL methods over SL methods further strengthen the findings of our previous study (Jia et al., 2024b). While their zero-shot detection accuracy remains lower than the expensive and time-consuming human counting method in the TUD-HCMC case study, their performance holds significant potential for improvement through cost-effective approaches.

We proposed several recommendations to further enhance model performance. First, increasing the amount of unlabeled data and extending the training time are beneficial, as shown in Section 3.1. Second, collecting a limited number of labeled images from target rivers for fine-tuning SSL models could further enhance model accuracy (van Lieshout et al., 2020) and solve the data imbalance problem discussed in Section 3.2.2. This is particularly effective when using images containing specific litter types from target rivers, e.g., transparent litter and litter entrapped in water hyacinths in the TUD-HCMC case study. Due to the highly limited number of images we collected in this case study, we did not perform standard hyperparameter optimization in this study. We believed that optimizing hyperparameters in SAHI could further enhance model accuracy (such as W_s and H_s). Third, we suggest applying data augmentation method to increase the number of labeled instances of specific litter types (Jia et al., 2023a). For example, water hyacinths of varying sizes can be extracted from source images, and pasted within the bounding boxes of litter items in target images, provided that the hyacinth area is smaller than the corresponding bounding box. This operation simulates scenarios where litter is partially occluded by vegetation, thereby enhancing the model's robustness to such challenging scenarios. Lastly, the framework could also benefit from replacing the ResNet50 backbone with state-of-the-art architectures, such as transformers that have demonstrated their effectiveness in foundational models like GPT, DINOv2, and Prithvi (Dosovitskiy et al., 2020).

A major shortcoming of developed models in this study is their inability to automatically identify and count the same litter item appearing in multiple consecutive images as a single item, resulting in overestimated fluxes. While we manually corrected the number of litter items detected by models to avoid flux overestimation, more automated methods are needed to address this issue, such as employing DeepSORT for tracking the detected objects across consecutive images (Wojke et al., 2017).

3.5. Towards foundational models for quantifying litter fluxes in river system

The establishment of large-scale monitoring networks for automated litter flux measurement in rivers requires models with robust generalization capabilities (Jia et al., 2023a). These models should perform accurately under zero-shot or few-shot scenarios, enabling measurement with minimal prior training data specific to the target locations. However, traditional SL models often fail to generalize effectively across varied locations, environmental conditions, and device setups (van Lieshout et al., 2020; Jia et al., 2023b). This limitation highlights the critical need for developing a more robust model, particularly a foundation model specifically designed for floating litter detection.

Artificial intelligence (AI) is undergoing a paradigm shift with the rise of foundation models, such as BERT, DALL-E, and GPT-3 (Kolides et al., 2023). These models learn general data representations from broad data — typically using self-supervised learning methods at scale, that enable them to be adapted to a wide range of downstream tasks when fine-tuned for the specific downstream tasks. The powerful OpenAI GPT series exemplifies this paradigm shift, driving the current AI revolution by enabling the development of ChatGPT (Achiam et al., 2023). Especially, the latest version, GPT-4, demonstrates a remarkable ability to generalize beyond its original training, performing tasks such as generating LaTeX code for drawing graphics, despite not being explicitly trained for this task (Bubeck et al., 2023). Another notable example is AlphaFold from Google DeepMind, which has had a transformative impact on understanding protein structure and dynamics (Jumper et al., 2021; Varadi et al., 2022). It represents a major breakthrough in computational biology, facilitating significant advances in drug discovery and disease research.

We believe foundational models tailored for floating litter detection could significantly improve the ability of the proposed SSL-based framework to monitor litter fluxes. In this study, we developed models using approximately 500k images sourced from a limited number of locations through self-supervised learning. We argue that expanding the dataset to include millions of images from diverse geographical locations is essential to mitigate pollution in river systems on a global scale.

4. Conclusions

Deep learning methods provide automated and efficient solutions for detecting floating litter in (fresh)water bodies. However, few studies have used them to quantify cross-sectional floating litter fluxes. Conventional supervised learning (SL) models require extensive labeled data, which is costly and labor-intensive to obtain. Moreover, current deep learning models for litter detection struggle with small litter detection. These limitations hinder the development of robust and large-scale monitoring networks necessary for addressing global-scale pollution problems. To overcome these limitations, we proposed a semi-supervised learning (SSL)-based framework combined with the Slicing Aided Hyper Inference (SAHI) method, to measure cross-sectional floating litter fluxes in river systems. We validated its effectiveness through experiments on images from waterways of the Netherlands, Indonesia, and Vietnam. Our main findings are as follows:

(1) The SSL models benefit from longer pre-training time and larger pre-training dataset size. Especially, when a large amount of data (200k images) is available, increasing pre-training epochs from 100 to 200 achieves an improvement in average precision (AP50) of 4.2%. Moreover, scaling the pre-training dataset size from 20k to 500k yields an improvement in AP50 of 14.7% and F1-score of 0.24, with a limited amount of labeled data for fine-tuning (226 images with 276 annotated litter items).

(2) The SSL methods significantly outperform the baseline SL methods in in-domain and out-of-domain detection performance. Compared to baseline SL benchmarks, SSL methods achieve an in-domain AP50 increase of 12% and F1-score increase of 0.2, and a zero-shot out-of-domain median F1-score increase of up to 0.14. It can be primarily attributed to the extraction of more informative and robust feature representations through self-supervised pre-training on relevant unlabeled images.

(3) The SAHI method enables the SSL models' ability to accurately detect 45 additional “small” litter items (area < 1000 cm^2) in the Vietnam case study, compared to the results obtained from the same SSL models without SAHI. This improvement also lead to an increase in F1-score by 0.34 and 0.19 for two locations in the case study, respectively.

(4) The cross-sectional floating litter fluxes measured by the SSL-based framework are nearly double those of the baseline SL-based framework, demonstrating closer alignment with human-measured fluxes in the Vietnam case study. While the SSL-based framework exhibits a strong positive correlation with human-measured fluxes across 10 sampling points (the Pearson correlation coefficient $r = 0.99$), it significantly underestimates fluxes by a factor of 3 to 4, compared to human measurements. One of the main reasons is the challenge of correctly detecting transparent litter and litter entrapped in water hyacinth, which together account for around 40% of the litter items in the Vietnam case study.

While we tested this new framework with cameras only for river surfaces, it can also be used with drones. It can be even extended to measure litter fluxes under river surface, provided that images are captured using similar devices (e.g., underwater cameras) or sonar technologies. Additionally, combining SSL methods with images collected from manned aircraft can enhance the detection of macroplastic hotspots on a larger scale, e.g., marine surfaces (Jia et al., 2023a; Garcia-Garin et al., 2021). This study aims to establish a pathway for developing a robust framework for floating litter flux measurement by incorporating a foundation model specifically designed for floating litter detection. Accurate measurements are expected to facilitate the assessment of pollution levels, thereby supporting the development of effective pollution reduction strategies.

CRediT authorship contribution statement

Tianlong Jia: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Riccardo Taormina:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Rinze de Vries:** Writing – review & editing, Supervision, Data curation. **Zoran Kapelan:** Writing – review & editing, Supervision, Funding acquisition. **Tim H.M. van Emmerik:** Writing – review & editing. **Paul Vriend:** Writing – review & editing. **Imke Okkerman:** Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Author Rinze de Vries is employed by Noria Sustainable Innovators. Authors Paul Vriend and Imke Okkerman are employed by Rijkswaterstaat. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The work is supported by China Scholarship Council (No. 202006160032) and the Directorate-General for Public Works and Water Management of The Netherlands (Rijkswaterstaat). We are grateful to Edoardo Antonio Forte, Francesca Anita Lena, Agatha Zamuner and Khiet Bui for the TUD-HCMC dataset. We acknowledge the use of computational resources of DelftBlue supercomputer (https://www.tudelft.nl/dhpc).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.watres.2025.124833.

Data availability

The code for this study is available on https://github.com/TianlongJia/deep_plastic_Flux_SSL. The TUD-HCMC dataset used in this study including images and bounding box annotations is available for download at: https://doi.org/10.5281/zenodo.17387612.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Akyon, F.C., Altinuc, S.O., Temizel, A., 2022. Slicing aided hyper inference and fine-tuning for small object detection. In: 2022 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 966–970.

Andriolo, U., Topouzelis, K., van Emmerik, T.H., Papakonstantinou, A., Monteiro, J.G., Isobe, A., Hidaka, M., Kako, S., Kataoka, T., Gonçalves, G., 2023. Drones for litter monitoring on coasts and rivers: suitable flight altitude and image resolution. Marine Poll. Bull. 195, 115521.

Bellou, N., Gambardella, C., Karantzalos, K., Monteiro, J.G., Canning-Clode, J., Kemna, S., Arrieta-Giron, C.A., Lemmen, C., 2021. Global assessment of innovative solutions to tackle marine litter. Nat. Sustain. 4 (6), 516–524.

Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–4. http://dx.doi.org/10.1007/978-3-642-00296-0_5.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al., 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inf. Process. Syst. 33, 9912–9924.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Garcia-Garin, O., Monleón-Getino, T., López-Brosa, P., Borrell, A., Aguilar, A., Borja-Robalino, R., Cardona, L., Vighi, M., 2021. Automatic detection and quantification of floating marine macro-litter in aerial images: Introducing a novel deep learning approach connected to a web application in R. Environ. Pollut. 273, 116490.

Gia, B.T., Khanh, T.B.C., Trong, H.H., Doan, T.T., Do, T., Le, D.-D., Ngo, T.D., 2024. Enhancing road object detection in fisheye cameras: An effective framework integrating SAHI and hybrid inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7227–7235.

Gnann, N., Baschek, B., Ternes, T.A., 2022. Close-range remote sensing-based detection and identification of macroplastics on water assisted by artificial intelligence: a review. Water Res. 222, 118902.

Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., Bojanowski, P., 2022. Vision models are more robust and fair when pretrained on uncurated images without supervision. arXiv preprint arXiv:2202.08360.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hosang, J., Benenson, R., Schiele, B., 2017. Learning non-maximum suppression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4507–4515.

Hurley, R., Braaten, H.F.V., Nizzetto, L., Steindal, E.H., Lin, Y., Clayer, F., van Emmerik, T., Buenaventura, N.T., Eidsvoll, D.P., Økelsrud, A., et al., 2023. Measuring riverine macroplastic: Methods, harmonisation, and quality control. Water Res. 235, 119902.

Jia, T., de Vries, R., Kapelan, Z., van Emmerik, T.H., Taormina, R., 2024b. Detecting floating litter in freshwater bodies with semi-supervised deep learning. Water Res. 266, 122405.

Jia, T., Kapelan, Z., De Vries, R., Vriend, P., Peereboom, E.C., Okkerman, I., Taormina, R., 2023a. Deep learning for detecting macroplastic litter in water bodies: A review. Water Res. 231, 119632.

Jia, T., Peng, Z., Yu, J., Piaggio, A.L., Zhang, S., de Kreuk, M.K., 2024a. Detecting the interaction between microparticles and biomass in biological wastewater treatment process with deep learning method. Science of the Total Environment 951, 175813.

Jia, T., Vallendar, A.J., de Vries, R., Kapelan, Z., Taormina, R., 2023b. Advancing deep learning-based detection of floating litter using a novel open dataset. Front. Water 5, 1298465.

Jia, T., Yu, J., Sun, A., Wu, Y., Zhang, S., Peng, Z., 2025. Semi-supervised learning-based identification of the attachment between sludge and microparticles in wastewater treatment. Journal of Environmental Management 375, 124268.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589.

Kataoka, T., Yoshida, T., Yamamoto, N., 2024. Instance segmentation models for detecting floating macroplastic debris from river surface images. Front. Earth Sci. 12, 1427132.

Kolides, A., Nawaz, A., Rathor, A., Beeman, D., Hashmi, M., Fatima, S., Berdik, D., Al-Ayyoub, M., Jararweh, Y., 2023. Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts. Simul. Model. Pr. Theory 126, 102754.

Lebreton, L., Slat, B., Ferrari, F., Sainte-Rose, B., Aitken, J., Marthouse, R., Hajbane, S., Cunsolo, S., Schwarz, A., Levivier, A., et al., 2018. Evidence that the great Pacific garbage patch is rapidly accumulating plastic. Sci. Rep. 8 (1), 1–15.

Li, Q., Wang, Z., Li, G., Zhou, C., Chen, P., Yang, C., 2023. An accurate and adaptable deep learning-based solution to floating litter cleaning up and its effectiveness on environmental recovery. J. Clean. Prod. 388, 135816.

Meijer, L.J., van Emmerik, T., van der Ent, R., Schmidt, C., Lebreton, L., 2021. More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean. Sci. Adv. 7 (18), eaaz5803.

Punjani, A., Fleet, D.J., 2021. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. J. Struct. Biol. 213 (2), 107702.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28, 91–99.

Sawyer, S.F., 2009. Analysis of variance: the fundamental concepts. J. Man. Manip. Ther. 17 (2), 27E–38E.

Schreyers, L.J., Bui, K., van Emmerik, T., Biermann, L., Uijlenhoet, R., Nguyen, H.Q., van der Ploeg, M.J., 2023. Discontinuity in fluvial plastic transport increased by floating vegetation. Authorea Prepr..

Singh, B., Davis, L.S., 2018. An analysis of scale invariance in object detection snip. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3578–3587.

van Calcar, C.v., van Emmerik, T.v., 2019. Abundance of plastic debris across European and Asian rivers. Environ. Res. Lett. 14 (12), 124051.

van Emmerik, T., de Lange, S., Frings, R., Schreyers, L., Aalderink, H., Leusink, J., Begemann, F., Hamers, E., Hauk, R., Janssens, N., et al., 2022a. Hydrology as a driver of floating river plastic transport. Earth's Futur. 10 (8), e2022EF002811.

van Emmerik, T.H., Frings, R.M., Schreyers, L.J., Hauk, R., de Lange, S.I., Mellink, Y.A., 2023. River plastic transport and deposition amplified by extreme flood. Nat. Water 1 (6), 514–522.

van Emmerik, T., Janssen, T., Jia, T., Bui, T.-K.L., Taormina, R., Quan, N.H., Schreyers, L., 2025. Plastic pollution and water hyacinths consistently co-occur in the lower Saigon river. Environ. Res.: Water.

van Emmerik, T., Kieu-Le, T.-C., Loozen, M., van Oeveren, K., Strady, E., Bui, X.-T., Egger, M., Gasperi, J., Lebreton, L., Nguyen, P.-D., et al., 2018. A methodology to characterize riverine macroplastic emission into the ocean. Front. Mar. Sci. 5, 372.

van Emmerik, T., Mellink, Y., Hauk, R., Waldschläger, K., Schreyers, L., 2022b. Rivers as plastic reservoirs. Front. Water 3, 212.

van Emmerik, T., Strady, E., Kieu-Le, T.-C., Nguyen, L., Gratiot, N., 2019. Seasonality of riverine macroplastic transport. Sci. Rep. 9 (1), 13549.

van Emmerik, T., Vriend, P., Copius Peereboom, E., 2022c. Roadmap for long-term macroplastic monitoring in rivers. Front. Environ. Sci. 9, 802245.

van Lieshout, C., van Oeveren, K., van Emmerik, T., Postma, E., 2020. Automated river plastic monitoring using deep learning and cameras. Earth Space Sci. 7 (8), e2019EA000960.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al., 2022. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50 (D1), D439–D444.

Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 3645–3649.

Wu, Y., Ma, X., Guo, G., Jia, T., Huang, Y., Liu, S., Fan, J., Wu, X., 2024. Advancing deep learning-based acoustic leak detection methods towards application for water distribution systems from a data-centric perspective. Water Research 121999.