Waste Management Bulletin 3 (2025) 100246

Contents lists available at ScienceDirect

Waste Management Bulletin

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

A Deep Learning Pipeline for Solid Waste Detection in Remote Sensing Images

Federico Gibellini *, Piero Fraternali, Giacomo Boracchi, Luca Morandini, Thomas Martinoli, Andrea Diecidue, Simona Malegori

Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Via Ponzio 34/5, Milan 20133, Italy

ARTICLE INFO

Keywords:
Solid waste detection
Remote sensing
Earth observation
Geospatial artificial intelligence
Computer vision

ABSTRACT

Improper solid waste management represents both a serious threat to ecosystem health and a significant source of revenues for criminal organizations perpetrating environmental crimes. This issue can be mitigated thanks to the increasing availability of Very-High-Resolution Remote Sensing (VHR RS) images. Modern image-analysis tools support automated photo-interpretation and large territory scanning in search of illegal waste disposal sites. This paper illustrates a semi-automatic waste detection pipeline, developed in collaboration with a regional environmental protection agency, for detecting candidate illegal dumping sites in VHR RS images. To optimize the effectiveness of the waste detector at the core of the pipeline, extensive experiments evaluate such design choices as the network architecture, the ground resolution and geographic span of the input images, as well as the pretraining procedures. The best model attains remarkable performance, achieving 92.02 % F1-Score and 94.56 % Accuracy. A generalization study assesses the performance variation when the detector processes images from various territories substantially different from the one used during training, incurring only a moderate performance loss, namely an average 5.1 % decrease in the F1-Score. Finally, an exercise in which expert photo-interpreters compare the effort required to scan large territories with and without support from the waste detector assesses the practical benefit of introducing a computer-aided image analysis tool in a professional environmental protection agency. Results show that a reduction of up to 30 % of the time spent for waste site detection can be attained.

Introduction

Improper waste handling threatens both ecosystems and human health, causing air, soil, and water pollution (Dabrowska et al., 2023; Vaverková et al., 2019). Illegal solid waste disposal may alter the dynamics of ecological systems and boost the diffusion of alien or dangerous species, as demonstrated by the correlation between the presence of landfills and the incidence of mosquito-borne illnesses, such as dengue and chikungunya (Khan et al., 2023). In addition to health hazards, illegal waste management is among the most fruitful activities for environmental crime organizations. In 2020, the annual revenues for hazardous waste trafficking in the EU were estimated between 1.5 and 1.8 EUR billions, whereas profits from non-hazardous waste trafficking ranged between 1.3 and 10.3 EUR billions (Europol, 2022). All these threats highlight the need for innovation in the investigation processes of environmental protection agencies and law enforcement authorities against improper solid waste management.

The recent advances in Computer Vision (CV), Deep Learning (DL) and Artificial Intelligence (AI), along with the increased availability of Very-High-Resolution (VHR) Remote Sensing (RS) images, i.e., images whose Ground Sample Distance (GSD) is lower than 50 cm/px, open new opportunities for contrasting this phenomenon. The application of image analysis methods to RS imagery enables automatic scanning of wide regions in search of illegal waste dumping sites and can reduce the need for expensive monitoring infrastructures as well as on-site inspections.

This paper illustrates a semi-automatic pipeline, developed in collaboration with a regional environmental protection agency, for detecting candidate illegal dumping sites in VHR RS images. The pipeline exploits a waste detector implemented by means of a DL-based binary image classifier, which receives as input a VHR RS image and outputs a binary prediction (i.e., "*waste*"/"*no waste*"). To optimize the

* Corresponding author at: Via Ponzio 34/5, Milan 20133, Italy.
E-mail addresses: federico.gibellini@polimi.it (F. Gibellini), piero.fraternali@polimi.it (P. Fraternali), giacomo.boracchi@polimi.it (G. Boracchi), luca.morandini@polimi.it (L. Morandini), thomas.martinoli@polimi.it (T. Martinoli), andrea.diecidue@polimi.it (A. Diecidue), simona.malegori@polimi.it (S. Malegori).

https://doi.org/10.1016/j.wmb.2025.100246

Available online 9 September 2025

detector accuracy, extensive experiments, yet widely unexplored in existing literature, compare alternative architectures of the underlying neural network, GSD and geographic spans of the input images, as well as the pretraining procedures. A generalization study assesses the performance variation when the detector processes images from regions significantly different from the one used for model training, a topic rarely addressed in previous works. Finally, a novel exercise in which professional photo-interpreters compare the territory scanning effort with and without the support of the waste detector enables a quantification of the practical benefits of introducing CV techniques for waste detection in VHR RS images in the professionals' everyday working routine.

Literature review

The ever-increasing quality and availability of RS products, along with the exceptional performance of DL-powered image analysis techniques, have driven interest in various Earth Observation (EO) applications, such as land use and land cover classification (Jodhani et al., 2025). In the domain of waste management, applications include territory analysis for landfill site planning (Chaturvedi et al., 2025) and automated waste detection for identifying and monitoring solid waste dumps from RS images. Over the years, this latter field has witnessed the evolution from manual photo-interpretation of satellite images to the application of the most advanced DL architectures for CV tasks such as image and scene classification, image segmentation and object detection.

Early works. Early works appeared when automatic image processing on a large scale was still beyond practical feasibility. RS images were manually photo-interpreted (Lyon, 1987), requiring a significant human effort which motivated the shift towards computer-aided solutions. The earliest approaches combined satellite and Geographic Information System (GIS) data (Notarnicola et al., 2004) or applied traditional (i.e., pre-DL) machine learning and CV methods based on the analysis of spectral or textural features of RS images and on the derivation of spectral indices (Lavender, 2022; Parrilli et al., 2021; Vambol et al., 2019).

DL-based approaches. Following the success of DL architectures for natural image analysis, researchers started applying DL models to waste detection in RS imagery, addressing tasks such as image and scene classification, pixel-level classification, image segmentation, and object detection. The most relevant approaches addressing these tasks are described in the following paragraphs and summarized in Table 1.

Image and scene classification methods. (Torres and Fraternali, 2021) implemented a binary scene classification model based on a ResNet backbone (He et al., 2016) and trained on VHR optical images at different resolutions (with GSD ranging between 20 and 50 cm/px) from the Lombardy region in Italy. The binary waste classifier achieves 94.5 % Average Precision (AP) on the 20 cm/px GSD images (Torres and Fraternali, 2021) and 87.99 % on the complete data set (Torres and Fraternali, 2023). (Kruse et al., 2023) developed a semi-supervised approach based on teacher-student DL architecture for plastic waste detection in Sentinel-2 satellite data. Their method exploits a pixel and a patch classifier, where the former creates a heat map of candidate waste locations and the latter processes the scene in each tile to improve classification accuracy. The pixel classifier achieves 90.46 % F1-Score, whereas the patch classifier scores 97.56 % in the same metric.

Pixel-level classification and segmentation methods. (Faizi et al., 2020) localized urban waste disposals using a pixel-based Multi-Spectral (MS) image classification approach, whereas (Ulloa-Torrealba et al., 2023; Đidelija et al., 2022) applied segmentation networks to cluster nearby pixels into objects, which are eventually classified as waste. (Yong et al., 2023) used a segmentation model based on the DeepLabv3+ architecture (Chen et al., 2018) to find construction and demolition waste in High Resolution (HR) optical images of urban areas, attaining an F1-Score of 77.4 %. (Zin et al., 2024) trained a classic U-Net binary image segmentation network and applied it to VHR Pléiades Neo images (30 cm/px GSD) over the central region of the West Coast of Malaysia, attaining an Accuracy of 80.6 % and deriving a risk exposure map from the prediction maps. The recent study (Yu et al., 2025) illustrates a two-step approach to detect waste materials in China. The first stage exploits a modified Yolov8 architecture to segment HR (GSD < 2 m/px) 4-band (RGB + NIR) images with respect to three types of materials: industrial solid waste, tailing ponds and other solid waste. Then, a subsequent classification phase operated by a ResNet model discriminates the subclasses of household garbage, construction waste, and mixed garbage. The detector achieves a mean Average Precision (mAP) of 84.1 % for the instance segmentation of the five waste categories.

Object detection methods. (Zhou et al., 2023) implemented a solid waste detector (SWDet), based on a modified YOLO network (Redmon et al., 2016) to locate industrial and household waste. The authors tested

Table 1
Summary and comparison of recent DL approaches for waste detection in RS images.

Article	Task	Data set	Image source and resolution	Score
(Torres and Fraternali, 2021)	Binary scene classification	~3000 RGB images	Aerial survey, 0.2 m/px	AP: 94.5 % F1-Score: 88.2 %
(Torres and Fraternali, 2023)	Binary scene classification	10,434 RGB images	Aerial survey, 0.2 m/px WorldView-3, 0.3 m/px Google Earth, 0.5 m/px	AP: 87.99 % F1-Score: 80.70 %
(Kruse et al., 2023)	Binary classification	~4,200,000 pixel spectrograms and 5,407 MS patches	Sentinel-2, 10 m/px	F1-Score: 97.56 %
(Faizi et al., 2020)	Pixel-based classification	MS images (number unavailable)	Sentinel-2, 10 m/px Landsat-8, 15 m/px	Accuracy: 94.82 % Accuracy: 85 %
(Ulloa-Torrealba et al., 2023)	Object-based classification	1,000 RGB-NIR images	Aerial survey, 0.08 m/px	Overall Accuracy: 80.18 %
(Đidelija et al., 2022)	Object-based classification	~175 RGB images	Pléiades 1B, 2 m/px	Overall Accuracy: 95 %
(Yong et al., 2023)	Semantic segmentation	3,322 RGB images	Esri World Image, 1 m/px	F1-Score: 77.40 %
(Zin et al., 2024)	Binary image segmentation	2,775 RGB images	Pléiades Neo, 0.3 m/px	Accuracy: 80.62 %
(Yu et al., 2025)	Instance segmentation	3,466 RGB-NIR images	Gaofen-1, 2 m/px Gaofen-2, 1 m/px Gaofen-6, 2 m/px	mAP: 84.1 %
(Zhou et al., 2023)	Object detection	1,996 RGB images	WorldView-2 and SPOT, 1.8 m/px	mAP: 77.58 %
(Li and Zhang, 2024)	Object detection			mAP: 58.6 %
(Sun et al., 2023)	Object detection	2,219 RGB images	Satellites from Gaofen and SuperView series, 0.3–1 m/px	AP: 70.1 %
(Zhang and Ma, 2024)	Object detection			mAP: 84.6 %

their approach on a public data set of their conception, called SWAD (Zhou et al., 2023), attaining 77.58 % mAP. (Sun et al., 2023) designed an object detection DL method that exploits a Blocked Channel Attention module to increase the discriminative power of the network and emphasize the relevant features in the input channels. The authors processed satellite images acquired in 2021 over 28 cities worldwide and detected over 1,000 dumpsites. They also performed a generalization study and identified the contextual factors that contribute to the formation of illegal dumping sites. (Zhang and Ma, 2024) introduced CascadeDumpNet, a 2-stage DL network for construction and domestic waste detection. In the first stage, an object detection architecture equipped with a novel Contextual Feature Synthesis module optimizes the processing of aerial images. The output of such object detector feeds the second-stage classifier, which receives as input also GIS information and is optimized with automatic hyper-parameter tuning to improve accuracy. The resulting model is applied to MS VHR images (GSD = 0.5 m/px) from Shenzhen, China. The authors compared CascadeDumpNet to other object detectors, demonstrating superior performance by attaining 84.6 % mAP. The authors also conducted a generalization study on other Chinese regions and discussed the key environmental factors contributing to the formation of dumpsites. (Li and Zhang, 2024) built an object detection network exploiting a guidance fusion module, a context awareness module, and a multi-scale interaction module which leverages spatial attention and large kernel convolutions to improve the localization of objects at different scales. The authors tested their detector on two data sets, attaining 58.6 % mAP on the publicly available SWAD data set.

Surveys. Further details on the topic of solid waste detection in RS images can be found in recent surveys. (Papale et al., 2023) review the relevant methods based on satellite data for landfill identification with a focus on case studies. (Fraternali et al., 2024) provide a comprehensive review of the approaches for detecting and monitoring large-scale landfills or urban waste dumps, discussing both methods based on traditional CV techniques and recent DL models. (Wang et al., 2024) survey waste monitoring methods based on both drone and satellite images and summarize benchmark results from recent studies.

Data sets. The AerialWaste data set (Torres and Fraternali, 2023) contains more than 11,700 RS images from 3 different sources and annotated for the tasks of binary scene classification, multi-label image classification and weakly supervised localization. The SWAD data set collects binary-labelled images representing various types of waste materials. It contains 1,996 images at 1.8 m/px resolution acquired in various cities of the Henan Province, China. Finally, the Global Dumpsite Test Data contains the images of four types of dumpsites (domestic, construction, industrial and covered) of the 28 cities analyzed in (Sun et al., 2023).

Literature gaps

Despite the growing interest in waste detection from RS images, several research topics still require further exploration. Among these are the contribution of different image factors and alternative pretraining procedures to detection performance, the ability of the model to generalize on unseen regions, as well as the practical utility of such models in real-world scenarios.

Given the large size of RS images, most methods divide the original imagery into smaller tiles for processing with waste detection algorithms. The chosen size determines the visible content within each patch, henceforth the *context size*, which affects the ability of the detector to isolate waste materials. The image GSD is another critical parameter: higher resolution improves the capacity to discriminate between different waste materials at the cost of increased acquisition expenses for RS products. Moreover, CV models trained on images with a given GSD may experience reduced performance when applied to imagery acquired at different resolutions. Similarly, the initialization of the DL network represents another design choice potentially impacting the model performance. Traditionally, network weights derived from training on natural images were used to initialize the early layers of the network. More recently, large foundation models specialized for EO tasks rely on backbone networks pretrained on extensive RS image data sets, such as Million-AID (Long et al., 2021); SatlasPretrain (Bastani et al., 2023), and the NASA/IBM Harmonized Landsat Sentinel-2 data set, which was released along with the HLS-FM foundation model.[1] These data sets offer a potentially advantageous alternative to pretraining on natural images (Wang et al., 2023).

Another important research question concerns generalization, i.e., the ability of models trained in one region to produce accurate predictions when applied to territories with different morphology, land cover, land use, and waste distribution.

Finally, the current literature lacks public data sets and benchmarks for head-to-head comparison of different methods, as well as studies assessing the practical benefits of automated waste detection within the operational investigation processes of environmental and law enforcement agencies.

This study aims to address these gaps. Our novel contributions include in the first place investigating the impact of image GSD, context size, and different pretraining weights on detection performance. Subsequently, we test the generalization of the best model on various geographical regions. Finally, we assess the practical utility of the detector in the routine activities of professionals.

Methodology

Overview

The proposed waste detection pipeline, illustrated in Fig. 1, receives as input a VHR RS optical image previously acquired by the user or by their agency and covering the entire area of interest. The image is split into geo-referenced squared tiles, which are processed by a binary image classifier that labels them as positive if they contain clues of solid waste and negative otherwise. The classifier returns two outputs: *i)* a numerical score, expressing the model confidence for the input tile to belong to the positive class, and *ii)* a saliency map, computed via Grad-CAM (Selvaraju et al., 2017), highlighting the image pixels where the classifier focused to deliver a positive prediction. Both outputs can be visualized in any GIS software, overlaid to the input satellite images as shown in Fig. 1. Tiles are color-coded on a yellow-to-red scale according to the confidence score output by the classifier. Saliency maps help the user zoom in on the relevant portions of the image to discern the presence of waste materials. This enables the users to filter locations based on the output classification score, thus potentially focusing only on sites detected with high confidence. Using GIS software, they can further integrate additional information about each site, such as population density in the surrounding area or the proximity to protected areas and water bodies. During this inspection phase, environmental agency operators may also assign a risk level to each detected location, combining proprietary information from previous inspections or sensitive legal data that cannot be publicly disclosed. This targeted approach reduces the time otherwise spent on labor-intensive tasks such as visually surveying the entire area of interest, while enabling the operators to shortlist high-priority sites for further action. These shortlisted locations can then be investigated through on-site inspections or drone overflights.

Data set preparation

The RS image data set used for training and testing the binary waste classifier was created following a location-based approach, thanks to the support from professionals of a local environmental agency. These

[1] https://www.earthdata.nasa.gov/news/nasa-ibm-openly-release-geospatial-ai-foundation-model-nasa-earth-observation-data.

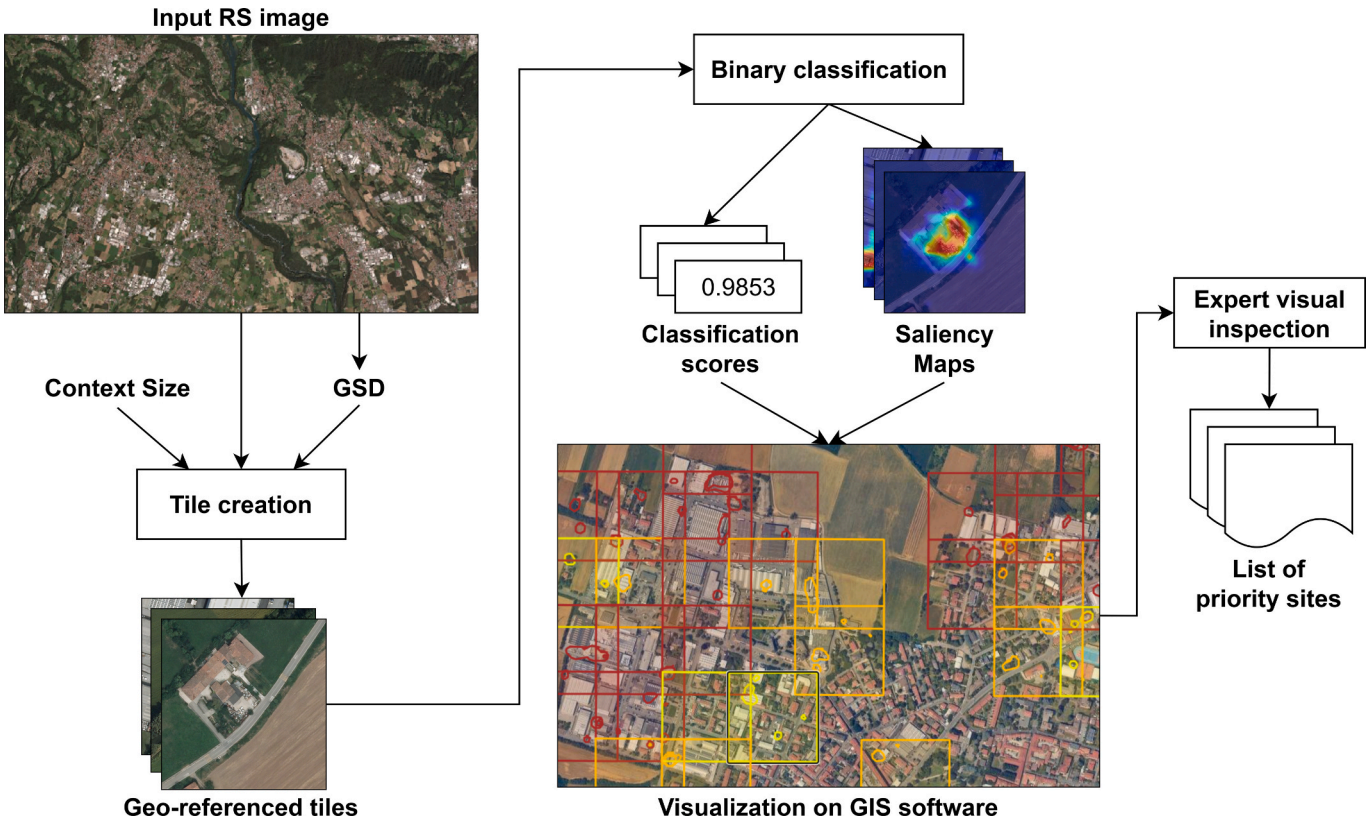

Fig. 1. The RS photo-interpretation process supported by the proposed pipeline.

experts provided the coordinates of waste dumping localities in Lombardy (Italy) known from previous investigations, which determined the positive samples. The locations of negative samples were defined by randomly picking places sufficiently close to the positive sites to guarantee the same geographic context, but also far enough to avoid significant overlaps. During the tile generation step, images were extracted centered in the selected positive and negative locations.

To foster diversity in the RS image data set, tiles were collected from various sources: **Google Earth** (≈21 cm/px GSD in the target area), **WorldView-3** (≈30 cm/px GSD) and **AGEA** (≈20 cm/px GSD). The latter images derive from flights conducted between 2021 and 2023 by the Italian National Agricultural Agency. Professional photo-interpreters visually inspected all the tiles to verify the presence or absence of waste, thus assuring correct labeling. During this revision process, the interpreters ensured that waste appeared in a 100 × 100 m area in the center of the positive samples. This guarantees that cropping an image to reduce its geographic context, as in the experiments described in Section Architecture, image and training, does not invalidate the image label. Following these validation phases, the data set consists of ≈11,700 tiles, of which approximately 80 % belong to the training set, whereas the remaining belong to the test set. In both sets, the number of negative images is twice the number of positive ones, to account for the real class imbalance between locations with and without waste. This approach is consistent with previous versions of AerialWaste (Torres and Fraternali, 2023), for which this data set was published as Version 3. Fig. 2 shows some examples of positive and negative tiles. The presence of waste is highlighted by the saliency maps computed with the best classifier described in the following sections.

Experimental setup

The design of the waste detector must account for various factors that affect its performance. These characteristics describe specific features of the input images, such as their GSD and the context size (i.e., the size of the represented geographic area), as well as the detector architecture and its training process.

Image Factors: GSD, context Size, and image size

Two factors primarily influence the tile content: *i)* the *GSD*, which denotes the size of one pixel on the ground, expressed in (centi)meters per pixel, and *ii)* the geographic *context size*, which describes the dimensions of the squared geographic area enclosed in the tile, measured in meters. The combination of GSD and context size determines the size in pixels of the resulting tiles. As an example, given a RS image with 30 cm/px GSD, a tile covering an area of 150 × 150 m yields a size of 500 × 500 pixels. In addition, the combination of these two factors affects the detection performance. As illustrated in Fig. 3.a, increasing the context size implies including a wider region in a single image, thus providing the network with more information. This may benefit the classification task, because the evaluation of the surrounding context should influence the detection of a specific waste object: for instance, a car in a parking lot should not be considered waste, whereas a car abandoned in a forest should. However, if the GSD is not adjusted accordingly, increasing the context size implies enlarging the image, thus potentially hindering the model classification capabilities. On the other hand, increasing the GSD while keeping the context size fixed (Fig. 3.b) produces smaller images with a coarser level of details. This results in losing the distinctive

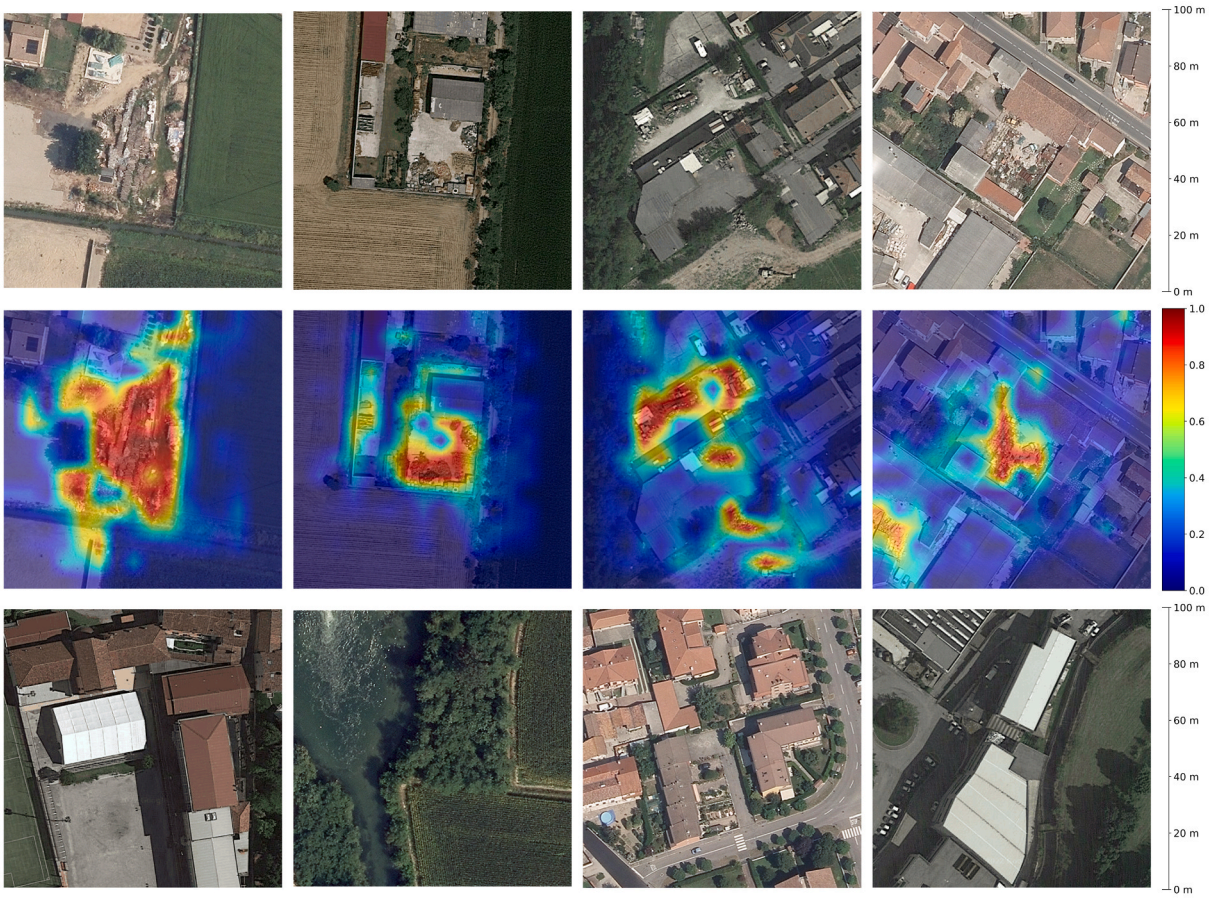

Fig. 2. Examples of positive (first row) and negative (third row) samples from the training data set. The second row highlights the presence of waste (red areas) in positive images by overlaying the saliency maps obtained with the waste detector to the original samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

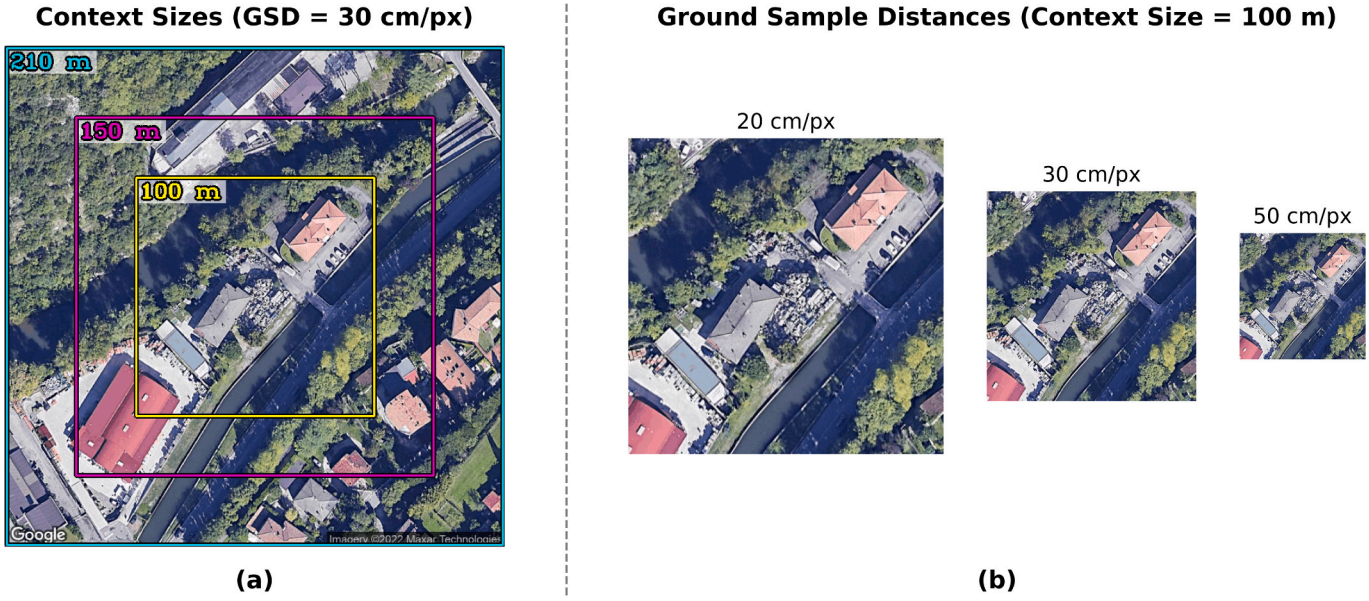

Fig. 3. Effects on the image size of: (a) varying the context size with a fixed GSD; (b) varying the GSD with a fixed context size (100 m). The positive tiles are manually validated to ensure that the smallest context size retains the presence of the waste within the image.

pattern of certain waste objects, such as asbestos plates, whose characteristic corrugated surface may be unnoticeable at a small scale. The chosen GSD values are 20, 30 and 50 cm/px. The 20 cm/px GSD value is the most common one in the training set, whereas the others are offered by commercial VHR satellite products. The selected context size values are 100, 150 and 210 m, with the latter being the default size of the images in the training data set. Smaller values have been defined to reduce the area portrayed by the positive tiles without altering the image label. The resulting image sizes used in the experiments were rounded to the closest multiple of 4 to provide non-padded images as input to the early convolutional layers of the chosen architectures.

Training factors: network architectures and training procedure

The experiments compare two network architectures, ResNet-50 (He et al., 2016) and Swin-T (Liu et al., 2021), which are representatives of two popular classes of image classification models, i.e., Convolutional Neural Networks and Vision Transformers. The chosen networks have a comparable number of parameters, 23M for ResNet-50 and 27M for Swin-T, and share a similar four-staged architecture. For the binary classification task, the original network head was replaced by a fully connected layer with a single output neuron, followed by a sigmoid activation to produce the classification score for the positive class.

For the selected architectures, two initializations with different pretraining weights were tested: traditional ImageNet (Deng et al., 2009) Pretraining (INP) and pretraining on Million-AID, a large data set of RS images for aerial scene recognition (Wang et al., 2023), henceforth Remote Sensing Pretraining (RSP). Training is conducted following a two-step procedure. In the *Transfer Learning* (TL) phase, the weights of the pretrained model are loaded for the entire network except for the classification head, the backbone is frozen, and the head is trained. In the subsequent *Fine-Tuning* (FT) step, the fourth and final stage in the backbone of the model obtained from TL is unfrozen and trained along with the network head to better adapt high-level features to the waste detection task.

All the combinations of GSD, context size and pretraining weights were used to train both network architectures. All training sessions were executed with a batch size of 120 and with a phase-dependent learning rate: 0.001 during TL and 0.0001 during FT.

Results and Discussion

Section Architecture, image and training illustrates the results of the tests with a focus on the impact of the factors presented in Section Experimental setup. Section Generalization study presents a generalization study to assess the performance variation when the waste detector processes images from areas different from the region used for training. Section Utility evaluation reports the findings of a practical evaluation exercise comparing the photo-interpretation effort of professionals with and without support from the waste detector.

Architecture, image and training

The experiments compare the 36 combinations of network architecture (2), GSD (3), context size (3) and pretraining weights (2), to assess the impact of such factors on classification performance. Table 2 and Fig. 4 summarize the results of these experiments. The computation of the Precision, Recall and F1-Score metrics assigns an input tile to the positive class if the output confidence score exceeds 0.5.

Architectures

Both Swin-T and ResNet-50 networks achieve excellent performance on the waste detection task, consistently averaging around 90 % F1-Score. However, for each combination of GSD and context size, Swin-T yields better values than ResNet-50, as shown in Fig. 4, where the solid lines representing Swin-T configurations are higher than the dashed lines for ResNet-50 with the same GSD and context size. This behavior may be attributed to the improved ability of transformers to capture global features and relationships, enabling a more comprehensive contextual understanding of the image and, consequently, a more effective evaluation.

Image GSD and context size

The performance of both networks seems not to be influenced by the GSD of the input images, as demonstrated by the horizontal lines in Fig. 4.a-b. One possible explanation for this phenomenon is that the determinant features of waste, such as colors and textures, which should guide the model in detection, do not vary significantly across the considered GSD values. These features remain instead recognizable and similar at all tested resolutions, thus resulting in stable performance. Moreover, only a moderate dependency on the context size appears for ResNet-50 models (Fig. 4.c-d), with smaller context areas resulting preferable. This behavior might be due to the different nature of the chosen architectures: the convolutional nature of ResNet-50 is accountable for a more significant local dependency, overcome by Swin-T thanks to its attention module, which increases the transformer

Table 2
Summary of experiment results.

Network	GSD [cm/px]	Context Size [m]	Image Size [px]	Metrics (RSP)				Metrics (INP)	
				F1-Score	Precision	Recall	Accuracy	F1-Score	Accuracy
ResNet-50	20	100	500	90.17 %	91.48 %	88.91 %	93.54 %	86.89 %	91.60 %
		150	748	89.54 %	91.62 %	87.57 %	93.18 %	85.11 %	90.68 %
		210	1048	86.95 %	**91.96 %**	82.51 %	91.76 %	84.25 %	90.02 %
	30	100	332	**90.98 %**	91.37 %	90.66 %	**94.02 %**	87.53 %	**91.88 %**
		150	500	90.81 %	91.16 %	90.49 %	93.90 %	85.69 %	91.06 %
		210	700	87.43 %	91.50 %	83.75 %	91.98 %	85.84 %	90.91 %
	50	100	200	89.12 %	87.56 %	**90.75 %**	92.61 %	**87.68 %**	91.86 %
		150	300	88.97 %	89.95 %	88.03 %	92.73 %	85.62 %	90.77 %
		210	420	87.53 %	89.95 %	85.29 %	91.91 %	85.47 %	90.47 %
Swin-T	20	100	500	**92.02 %**	90.02 %	**94.13 %**	**94.56 %**	90.40 %	93.63 %
		150	748	91.67 %	90.51 %	92.89 %	94.38 %	89.50 %	93.15 %
		210	1048	91.47 %	**92.17 %**	90.79 %	94.35 %	89.69 %	93.27 %
	30	100	332	91.63 %	90.08 %	93.28 %	94.32 %	90.40 %	93.59 %
		150	500	91.83 %	91.03 %	92.66 %	94.51 %	90.58 %	93.72 %
		210	700	91.52 %	91.06 %	92.04 %	94.32 %	**90.74 %**	**93.78 %**
	50	100	200	90.98 %	89.10 %	92.98 %	93.86 %	89.57 %	93.01 %
		150	300	90.61 %	89.87 %	91.44 %	93.69 %	89.86 %	93.17 %
		210	420	91.48 %	90.57 %	92.41 %	94.26 %	89.83 %	93.20 %

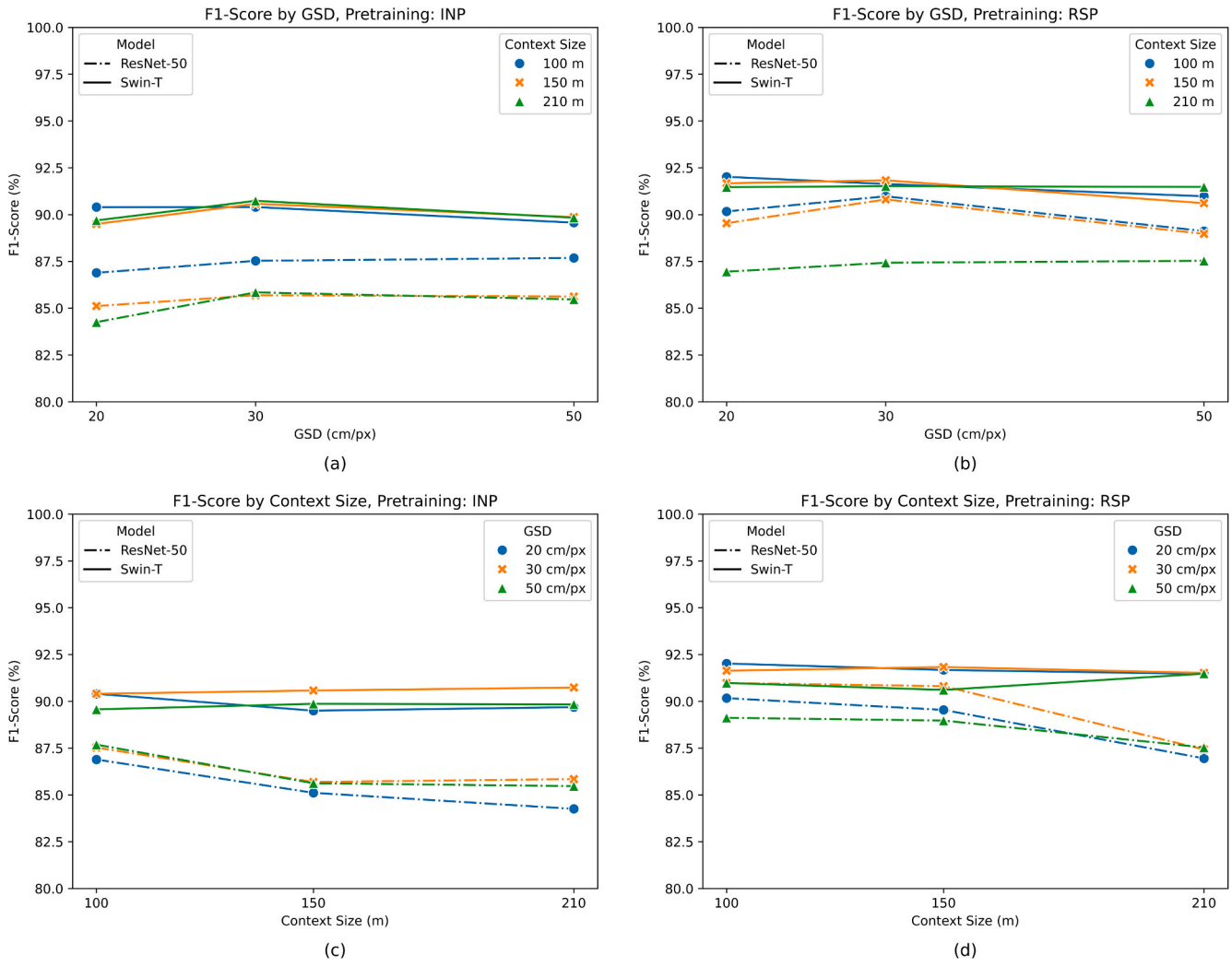

Fig. 4. Distribution of F1-Score across GSD (top) and context size (bottom) values.

robustness to variations in image and context size.

Training weights

In general, RSP guarantees a minor performance increase to both networks, especially to ResNet-50 models, whose metrics approach those of Swin-T models (Fig. 4.b-d). However, such an increase does not suggest a significant improvement in adopting RS images for pretraining rather than natural images (INP). This behavior might be motivated by the different GSD of the images used for backbone pretraining, which do not match those of the waste detection task. Indeed, RSP weights are obtained by training the networks on a very large data set of RS images, whose GSD values vary in the wide range between 0.5 and 153 m/px (Long et al., 2021). Therefore, RSP results in learning features at GSD values higher than 50 cm/px, thus being potentially beneficial when fine-tuning on images with larger GSD values, closer to those of the pretraining images rather than those considered here.

Best configuration

In summary, results show that the Swin-T architecture consistently outperforms ResNet-50 across all experiments, thanks to the increased contextual understanding of transformer models. The impact of GSD variations appears negligible, likely due to the absence of significant changes in the characteristic features of waste instances across the considered GSD values. Context size has only a minor influence, which is more evident for ResNet-50 models and tends to favor smaller patches. This behavior can be attributed to the more limited robustness to context increases for networks without attention modules. Finally, pretraining with RSP yields a modest but consistent performance improvement, albeit constrained by the different GSDs of images in the pretraining and fine-tuning data sets.

The best model overall is Swin-T pretrained on RSP and fine-tuned on images with a GSD of 20 cm/px and a context size of 100 m. This model, as shown Table 2, achieves the highest performance in terms of both F1-Score (92.02 %) and Accuracy (94.56 %).

Generalization study

To assess the classifier generalization capabilities, the model was tested for inference on images from 3 European countries[2]: Greece, Sweden, and Romania. These territories portray landscapes from diverse climate regions as well as differences in visual appearance compared to Lombardy, which provided the images for fine-tuning the waste detector. The Swedish region is the most similar to Lombardy, differing mainly for the significant presence of woods and for the architectural peculiarities of Northern Europe buildings, such as grey roofs. The Romanian region primarily covers an urban landscape where buildings

[2] The regions were suggested by partner agencies of the PERIVALLON project (https://perivallon-he.eu).

Table 3
Distribution of samples across the generalization test sets and metrics scored upon inference with the waste detector.

Region	Positives	Negatives	Samples	F1-Score	Precision	Recall	Accuracy
Greece	116	137	253	85.45 %	90.38 %	81.03 %	87.35 %
Sweden	187	193	380	83.82 %	91.19 %	77.54 %	85.26 %
Romania	185	133	318	91.48 %	96.41 %	87.03 %	90.57 %

are significantly denser and more cluttered than in Lombardy. Finally, the Greek region covers a hilly territory characterized by dry lands differing from the typical greenery of Lombardy.

The test sets to assess the model generalization performance were prepared following a location-based approach. Initially, positive and negative sites were manually identified on Google Maps, with care taken to select points that maximize the land-cover heterogeneity, thus portraying urban, rural and industrial areas. Then, at the selected locations, 100 × 100 m squared tiles were extracted from different sources depending on the target country: a WorldView-3 acquisition performed on July 2nd 2024 for the Greek region, Google Earth for the Swedish and Romanian areas. The extracted tiles were visually double-checked to verify the label correctness. This process led to 3 different test sets, whose sample distribution is reported in Table 3. Exemplary samples are presented in Figs. 5-7.

The best model identified in Section Best configuration, i.e., Swin-T pretrained with RSP weights and trained on images with context size of 100 m and GSD of 20 cm/px, was run on such generalization test sets, achieving the results presented in Table 3. Averaging across the three regions, the detector scores 87.73 % Accuracy and 86.92 % F1-Score, thus losing respectively 6.83 % and 5.10 % compared to the performance on the Lombardy test set. This demonstrates that the classifier generalizes well to regions which are visually and morphologically different from the training area.

The saliency maps in Figs. 5-7 demonstrate that the classifier can correctly focus on regions covered by waste, confirming its generalization capabilities. For true positives, the network successfully attends to the correct areas, precisely identifying the locations where waste is present. This effect is particularly evident in the Romanian data set, where the saliency maps appear especially accurate. However, this may be partially due to the prevalence of large landfills in the region, which are easier to detect. In the case of false positives, the network often focuses on genuinely suspicious regions that were annotated as negatives, probably because of their reduced dimensions or for being difficult to classify even for a human interpreter. False positives in Sweden highlight these issues since they mainly portray industrial areas where small amounts of residual materials were left outside of factories. Examples of false negatives show that the network can still focus on potentially relevant regions, even though the classification score is lower than the threshold. The Swedish false negatives typically correspond to small landfills in residential areas whose limited size reduces the overall confidence of the classifier. In Greece waste appears in larger quantities of scattered materials, which visually resemble the rocky surroundings. Finally, regarding true negatives, the model correctly rejects clean areas across a variety of contexts, including both rural and urban environments. In rural contexts, the prevalence of fields with regular and uniform textures enhances the model prediction confidence, whereas the abundant cluttering of urban scenes might hinders it. The tested model correctly evaluates both situations. Nevertheless, further improvements could be achieved by fine-tuning the classifier on a data set including images from different regions with diverse contexts.

Utility evaluation

The practical utility of the proposed pipeline was evaluated through a study conducted in collaboration with the same environmental monitoring professionals who provided the database of waste locations to initially build the data set. The goal of such study was to quantify the time saved when scanning a large territory with support from the designed pipeline as opposed to without. The study addressed 12 municipalities in Lombardy, which were processed following two procedures: traditional manual photo-interpretation and an AI-supported approach. The first procedure consisted in manually inspecting the satellite imagery covering the region of interest via a GIS tool, without further assistance. The second method leveraged the proposed pipeline to analyze the entire territory by dividing it into squared tiles and processing them with the waste detector. Then, human interpreters visualized on a GIS tool the scores and saliency maps output by the classifier for all and only the tiles whose classification score was greater than 0.2. Therefore, only the area covered by such tiles was inspected.

The study involved 4 professionals, each of whom analyzed all 12 municipalities, 6 with the traditional method and 6 with support from the classifier. As reported in Table 4, exploiting the classifier predictions allowed the personnel of the environmental agency to detect a larger number of critical sites, 155 against the 95 identified without support from the image-analysis tool, while examining in detail a 60.2 % smaller area. Time spent by the professionals to detect and verify each site was recorded to compare the human effort required by each procedure. As expected, the AI-supported procedure yielded a higher overall inspection time due to the larger number of detected sites, but reduced the average time to inspect a site by 12.2 %.

Since the experiment aimed at quantifying the human effort in the experiment, time for running the classifier was excluded from computations, as such process does not require any human supervision. Moreover, the time required for model inference on a large area (100 km^2) on a single consumer-level NVIDIA GeForce RTX 2080 Ti with 12 GB RAM demonstrated to be negligible with respect to the time needed to visually inspect images.

The experiment was then repeated in a more conservative scenario where the classifier threshold was set to 0.7 following the suggestions of the experts involved in the experiment, who reckoned such value to be the best option for detecting all high-risk sites while minimizing the inspection effort. By raising the threshold to 0.7, only tiles with a high confidence of containing waste sites were reported for inspection, thus further and significantly reducing the area to be examined by 64 % with respect to the case using a 0.2 filtering threshold. Under these conditions, the number of detected sites also decreased by 32 %, as several less confident model detections are filtered out due to the higher threshold. Consequently, inspection with AI support resulted in an approximate 30 % time reduction compared with the fully manual photo-interpretation approach. On top of this, the limited number of municipalities addressed in the study did not allow to assess the impact of fatigue on the operators, a relevant phenomenon occurring when analyzing a significant amount of images. Therefore, the measured time savings must be considered a lower bound to the potential gain achieved when analyzing hundreds of municipalities.

Conclusions

Given the threats posed by improper waste management from private citizens and criminal organizations, the application of CV techniques to the RS image analysis can help environmental and law enforcement agencies to increase the frequency and scale of their territory monitoring activities, thus improving the effectiveness and efficiency of their investigation processes. This paper aims to contribute to a better understanding of the decisions involved in the design of a waste

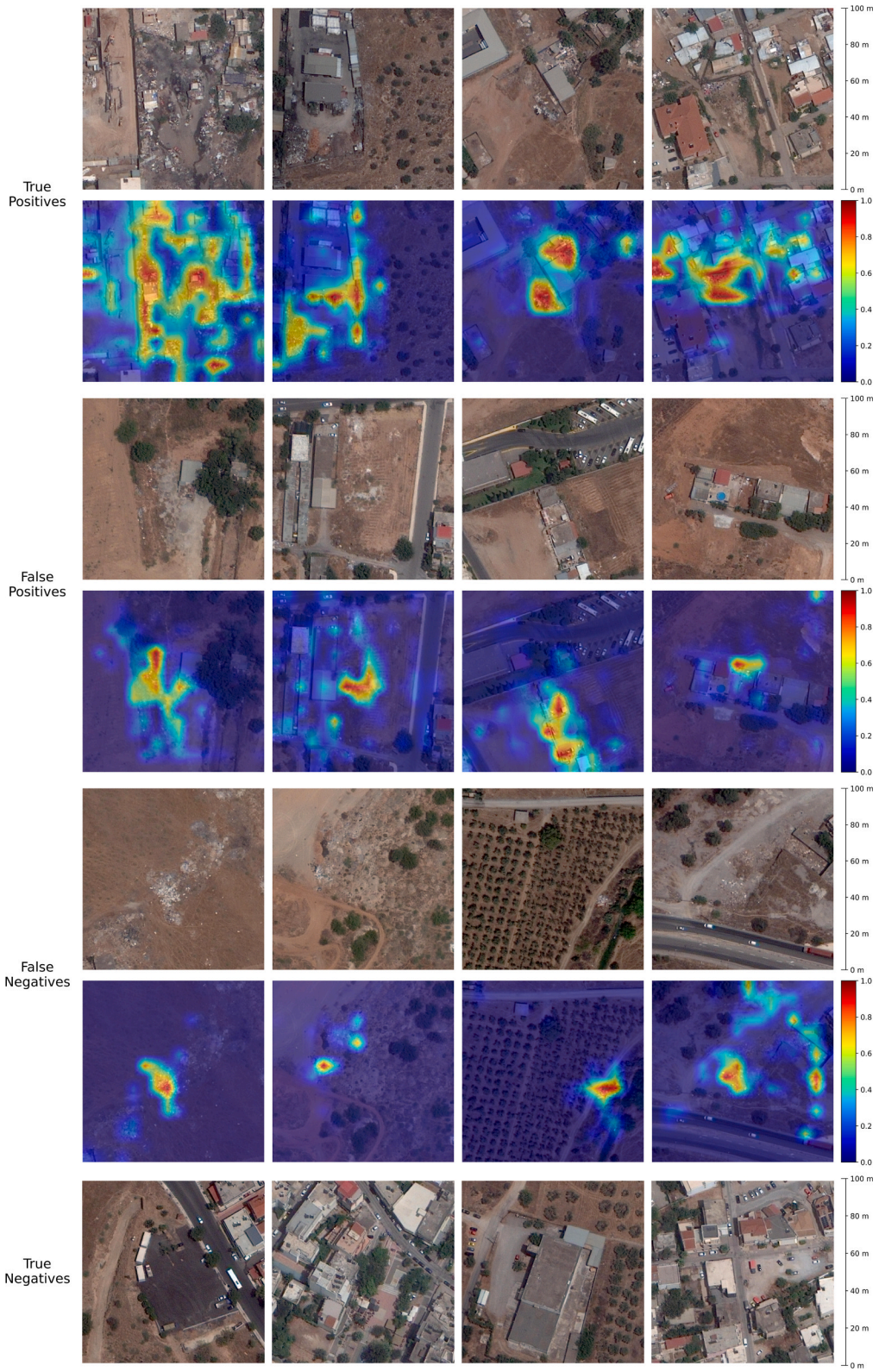

Fig. 5. Examples of true positive, false positive, false negative and true negative samples from the Greek generalization test set [© DigitalGlobe, Inc. (2024), provided by European Space Imaging]. For the first 3 sets of samples, the second row highlights the presence of waste (red areas) by overlaying the saliency maps obtained with the waste detector to the original sample. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

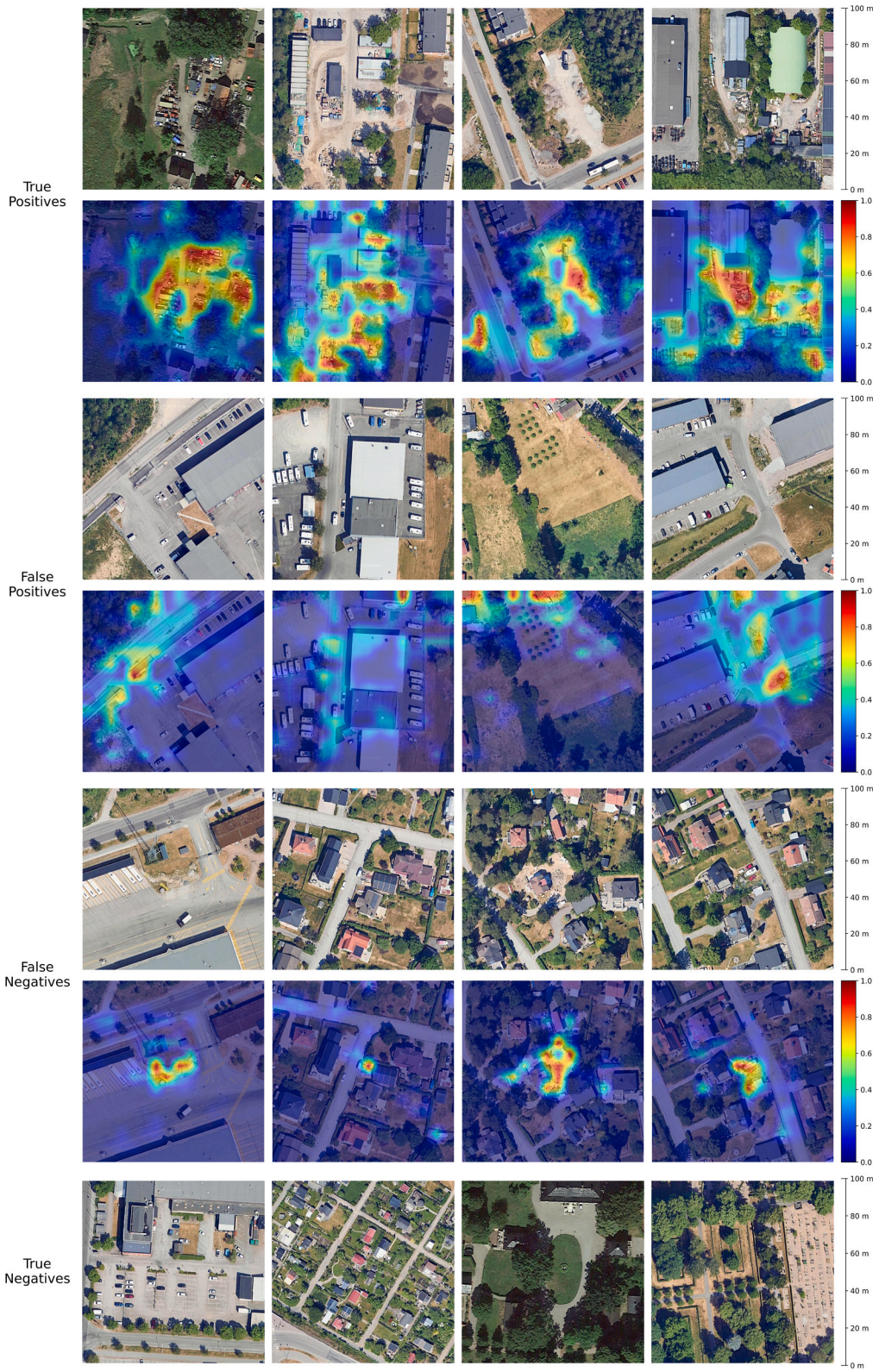

Fig. 6. Examples of true positive, false positive, false negative and true negative samples from the Swedish generalization test set [© Google, 2025]. For the first 3 sets of samples, the second row highlights the presence of waste (red areas) by overlaying the saliency maps obtained with the waste detector to the original sample. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

detection pipeline, of the generalization capacity of CV models and of the practical utility of applying AI-based tools in the RS image interpretation process. Towards such a goal, the paper provided the following contributions.

A CV pipeline is developed to support environmental and law enforcement agencies in their investigation processes. The pipeline receives as input RS imagery over a region of interest, divides it into tiles on a regular grid pattern, processes such tiles with a DL binary waste

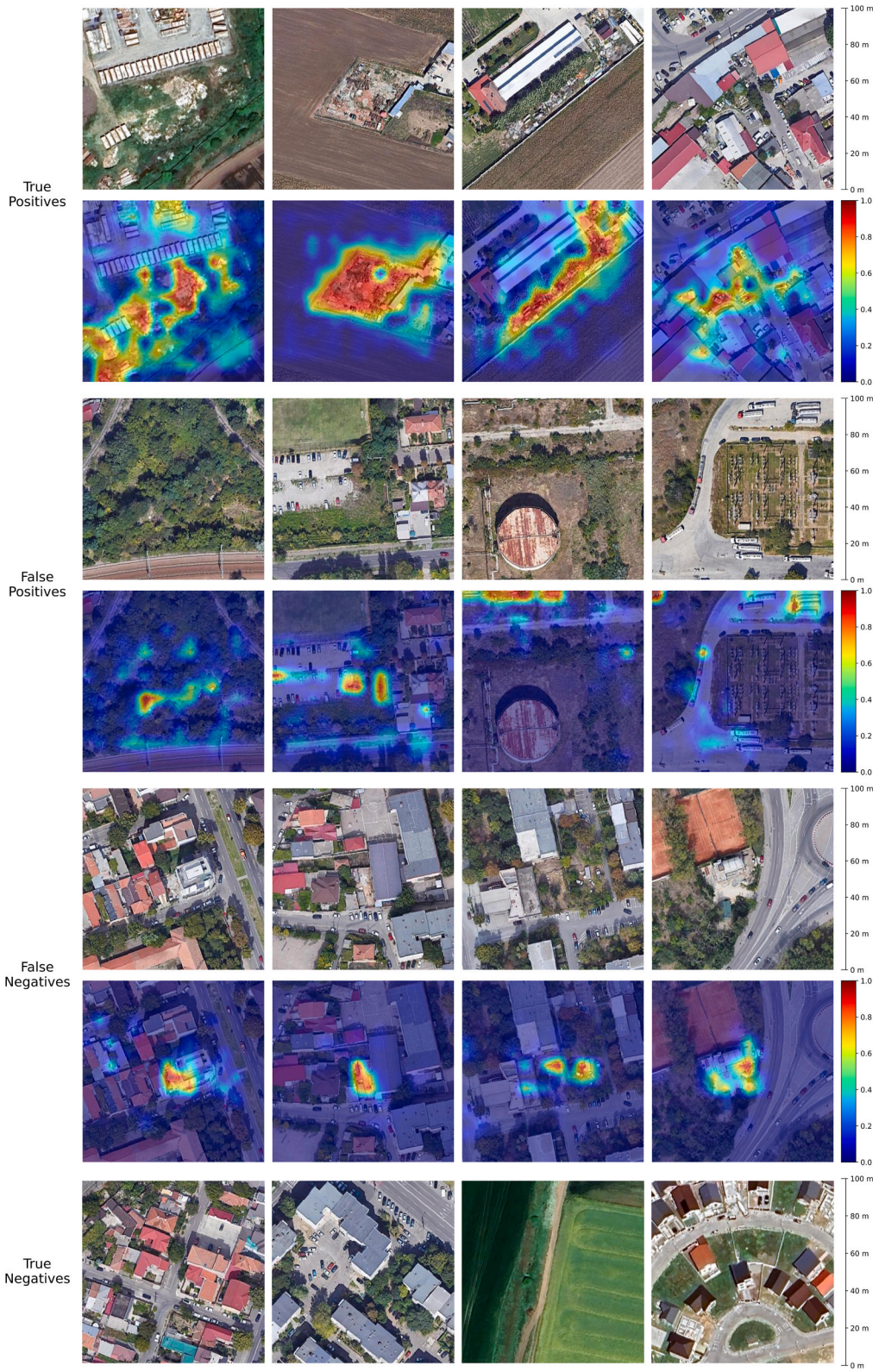

Fig. 7. Examples of true positive, false positive, false negative and true negative samples from the Romanian generalization test set [© Google, 2025]. For the first 3 sets of samples, the second row highlights the presence of waste (red areas) by overlaying the saliency maps obtained with the waste detector to the original sample. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

classifier and outputs for each processed tile a confidence score and a saliency map. The former denotes the trust of the model in the presence of waste materials in the tile, the latter highlights the regions of the image where waste materials have been detected. The pipeline can be easily integrated by environmental monitoring professionals in their routine workflows. The only potential barrier towards a more systematic adoption is represented by the high costs for acquiring VHR RS imagery, which are however expected to decrease in the near future.

Table 4
Results from the experiment involving professionals from a local environmental agency to evaluate time savings implied by the introduction of the proposed pipeline.

Photo-interpretation approach	Manually inspected area [km^2]	Detected sites	Total time [min]	Average time per site [min]
Without AI support	125.24	95	1486	15.6
With AI support	49.84	155	2133	13.7
Variation	−60.2 %	**+63.2 %**	+43.5 %	**−12.2 %**

An empirical study assesses how the performance of the binary waste classifier is affected by various factors: *i)* the network architecture, *ii)* the image GSD, *iii)* the size of the area portrayed in the picture, and *iv)* the weights adopted for pretraining the network. The experiments highlight that the best configuration uses a Swin-T model initialized with RSP weights and trained with images at 20 cm/px GSD portraying a squared area with a side of 100 m. Such model achieves 94.56 % Accuracy and 92.02 % F1-Score. All experiments employ a data set developed in collaboration with expert photo-interpreters from an environmental agency and made publicly available as Version 3 of AerialWaste (Torres and Fraternali, 2021; Torres and Fraternali, 2023).

A generalization study shows that the designed approach retains good performance when applied to RS images of territories with different visual appearance from the training region, averaging 87.73 % Accuracy and 86.92 % F1-Score. In addition, an evaluation exercise quantifies the impact of the proposed pipeline in the working routine of professionals from an environmental agency. In comparison to manual photo-interpretation, the use of the binary waste detector during the image search saves ≈12 % time in the analysis of the region of interest and enables the detection of ≈63 % more candidate waste dumping sites.

This study adds to the very few that perform a generalization test on territories with very distinct characteristics and, to the best of our knowledge, it is the first documented exercise evaluating the practical utility of computer support to photo-interpretation for waste detection. The developed binary waste classifier is being used to locate waste dumping sites in various European countries and its code is publicly available at https://github.com/gblfrc/waste-detection-dl-pipeline.

Future work will focus on extending the training data set with images from various regions to improve its generalization abilities and on developing methods to support prioritization of on-site inspection campaigns. These methods include automatically identifying the specific materials contained in a waste dump and highlighting the presence of particularly hazardous substances such as asbestos or flammable items. Material recognition requires a more complex detector architecture and possibly the use of multi- or hyper-spectral data, to discriminate materials with a similar visual appearance but different spectrographic response, e.g., asbestos plates against corrugated metal sheets. Risk assessment could exploit multi-modal information as input, such as cadaster information, land cover/land use maps and other GIS data. Finally, multi-temporal analysis based on time series of RS images of the same region could be applied to prioritize the investigation of dumping sites with a growing evolution profile.

CRediT authorship contribution statement

Federico Gibellini: Writing – review & editing, Writing – original draft, Visualization, Software, Data curation, Conceptualization. **Piero Fraternali:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Giacomo Boracchi:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Luca Morandini:** Writing – review & editing, Data curation, Conceptualization. **Thomas Martinoli:** Writing – review & editing, Data curation, Conceptualization. **Andrea Diecidue:** Writing – review & editing, Data curation. **Simona Malegori:** Writing – review & editing, Software, Data curation.

Funding

This research was partially funded by European Union's Horizon Europe project PERIVALLON – Protecting the EuROpean terrItory from organised enVironmentAl crime through inteLLigent threat detectiON tools, under grant agreement no. 101073952.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Special thanks to the professional photo-interpreters from ARPA Lombardia, who participated in the data set creation process and in the utility validation experiment. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

Data availability

The AerialWaste data set employed in the study is openly available on Zenodo at https://doi.org/10.5281/zenodo.12607190.

References

Dabrowska, D., Rykala, W., Nourani, V., 2023. Causes, Types and Consequences of Municipal Waste Landfill Fires—Literature Review. Sustainability 15, 5713. https://doi.org/10.3390/su15075713.

Vaverková, M.D., Maxianová, A., Winkler, J., Adamcová, D., Podlasek, A., 2019. Environmental consequences and the role of illegal waste dumps and their impact on land degradation. Land Use Policy 89, 104234. https://doi.org/10.1016/j.landusepol.2019.104234.

Khan, A., Bisanzio, D., Mutuku, F., Ndenga, B., Grossi-Soyster, E.N., Jembe, Z., Maina, P. W., Chebii, P.K., Ronga, C.O., Okuta, V., LaBeaud, A.D., 2023. Spatiotemporal overlapping of dengue, chikungunya, and malaria infections in children in Kenya. BMC Infect. Dis. 23, 183. https://doi.org/10.1186/s12879-023-08157-4.

Europol, Environmental crime in the age of climate change – Threat assessment 2022, 2022 10.2813/54422.

Jodhani, K.H., Patel, D., Madhavan, N., Gupta, N., Singh, S.K., Pandey, M., 2025. ML-Based Land Use and Land Cover Classification: Assessing Performance and Predicting Future Changes. J. Hydrol. Eng. 30. https://doi.org/10.1061/JHYEFF.HEENG-6416.

Chaturvedi, S., Bhatt, N., Shah, V., Jodhani, K.H., Patel, D., Singh, S.K., 2025. Landfill site selection in hilly terrains: An integrated RS-GIS approach with AHP and VIKOR. Waste Management Bulletin 3 332–348. https://doi.org/10.1016/j.wmb.2025.01.010.

Lyon, J.G., 1987. Use of maps, aerial photographs, and other remote sensor data for practical evaluations of hazardous waste sites, Photogramm Eng. Remote Sens. (Basel) 53, 515–519.

Notarnicola, C., Angiulli, M., Giasi, C.I., 2004. Southern Italy illegal dumps detection based on spectral analysis of remotely sensed data and land-cover maps. In: Ehlers, M., Kaufmann, H.J., Michel, U. (Eds.), Remote Sensing for Environmental Monitoring, GIS Applications, and Geology III. SPIE, pp. 483–493. https://doi.org/10.1117/12.511668.

Lavender, S., 2022. Detection of Waste Plastics in the Environment: Application of Copernicus Earth Observation Data. Remote Sens. (Basel) 14, 4772. https://doi.org/10.3390/rs14194772.

S. Parrilli, L. Cicala, C. VincenzoAngelino, D. Amitrano, Illegal Micro-Dumps Monitoring: Pollution Sources and Targets Detection in Satellite Images with the Scattering Transform. In: IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE 2021, pp. 4892-4895. https://doi.org/10.1109/IGARSS47720.2021.9555072.

Vambol, S., Vambol, V., Sundararajan, M., Ansari, I., 2019. The nature and detection of unauthorized waste dump sites using remote sensing. Ecol. Quest. 30, 1. https://doi.org/10.12775/EQ.2019.018.

Torres, R.N., Fraternali, P., 2021. Learning to Identify Illegal Landfills through Scene Classification in Aerial Images. Remote Sens. (Basel) 13, 4520. https://doi.org/10.3390/rs13224520.

K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, USA, 2016: pp. 770–778. https://doi.org/10.1109/cvpr.2016.90.

Torres, R.N., Fraternali, P., 2023. AerialWaste dataset for landfill discovery in aerial and satellite images. Sci. Data 10, 63. https://doi.org/10.1038/s41597-023-01976-9.
Kruse, C., Boyda, E., Chen, S., Karra, K., Bou-Nahra, T., Hammer, D., Mathis, J., Maddalene, T., Jambeck, J., Laurier, F., 2023. Satellite monitoring of terrestrial plastic waste. PLoS One 18, e0278997. https://doi.org/10.1371/journal.pone.0278997.
F. Faizi, K. Mahmood, M.H. Chaudhry, A.D. Rana, Satellite remote sensing and image processing techniques for monitoring MSW dumps, in: Proceedings of 5th EurAsia Waste Management Symposium, Istanbul, Turkey, 2020: pp. 26–28.
Ulloa-Torrealba, Y.Z., Schmitt, A., Wurm, M., Taubenböck, H., 2023. Litter on the streets - solid waste detection using VHR images. Eur J Remote Sens 56. https://doi.org/10.1080/22797254.2023.2176006.
Đidelija, M., Kulo, N., Mulahusić, A., Tuno, N., Topoljak, J., 2022. Segmentation scale parameter influence on the accuracy of detecting illegal landfills on satellite imagery. A case study for Novo Sarajevo, Ecol Inform 70, 101755. https://doi.org/10.1016/j.ecoinf.2022.101755.
Yong, Q., Wu, H., Wang, J., Chen, R., Yu, B., Zuo, J., Du, L., 2023. Automatic identification of illegal construction and demolition waste landfills: A computer vision approach. Waste Manag. 172, 267–277. https://doi.org/10.1016/j.wasman.2023.10.023.
L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in: Computer Vision – ECCV 2018, Springer International Publishing, 2018: pp. 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.
S.M.M. Zin, N.M. Yusoff, N.I. Zainal, W.A. Numpang, H. Halim, Deep Learning Model for Automated Detection of Solid Waste Dumping Sites using Satellite Imagery, in: Asian Conference on Remote Sensing (ACRS 2024), ACRS-AARS, Sri Jayewardenepura, 2024: pp. 1–22.
Yu, J., Mao, P., Wu, W., Wang, Q., Shao, X., Teng, J., Wang, Y., 2025. TSNET: A solid waste instance segmentation model in China based on a Two-Step detection strategy and satellite remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 136, 104366. https://doi.org/10.1016/j.jag.2025.104366.
Zhou, L., Rao, X., Li, Y., Zuo, X., Liu, Y., Lin, Y., Yang, Y., 2023. SWDet: Anchor-Based Object Detector for Solid Waste Detection in Aerial Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 306–320. https://doi.org/10.1109/JSTARS.2022.3218958.
Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. In: in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.
Sun, X., Yin, D., Qin, F., Yu, H., Lu, W., Yao, F., He, Q., Huang, X., Yan, Z., Wang, P., Deng, C., Liu, N., Yang, Y., Liang, W., Wang, R., Wang, C., Yokoya, N., Hänsch, R., Fu, K., 2023. Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery. Nat. Commun. 14, 1444. https://doi.org/10.1038/s41467-023-37136-1.
Zhang, S., Ma, J., 2024. CascadeDumpNet: Enhancing open dumpsite detection through deep learning and AutoML integrated dual-stage approach using high-resolution satellite imagery. Remote Sens. Environ. 313, 114349. https://doi.org/10.1016/j.rse.2024.114349.
Li, Y., Zhang, X., 2024. Multi-Scale Context Fusion Network for Urban Solid Waste Detection in Remote Sensing Images. Remote Sens. (Basel) 16, 3595. https://doi.org/10.3390/rs16193595.
Papale, L.G., Guerrisi, G., De Santis, D., Schiavon, G., Del Frate, F., 2023. Satellite Data Potentialities in Solid Waste Landfill Monitoring: Review and Case Studies. Sensors 23, 3917. https://doi.org/10.3390/s23083917.
Fraternali, P., Morandini, L., Herrera González, S.L., 2024. Solid waste detection, monitoring and mapping in remote sensing images: A survey. Waste Management 189, 88–102.
Wang, B., Xing, Y., Wang, N., Chen, C.L.P., 2024. Monitoring Waste From Uncrewed Aerial Vehicles and Satellite Imagery Using Deep Learning Techniques: A Review. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17, 20064–20079. https://doi.org/10.1109/JSTARS.2024.3488056.
Long, Y., Xia, G.-S., Li, S., Yang, W., Yang, M.Y., Zhu, X.X., Zhang, L., Li, D., 2021. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews. Guidances, and Million-AID, IEEE J Sel Top Appl Earth Obs Remote Sens 14, 4205–4230. https://doi.org/10.1109/JSTARS.2021.3070368.
F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, A. Kembhavi, SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Paris, France, 2023: pp. 16726–16736. https://doi.org/10.1109/iccv51070.2023.01538.
Wang, D., Zhang, J., Du, B., Xia, G.-S., Tao, D., 2023. An Empirical Study of Remote Sensing Pretraining. IEEE Trans. Geosci. Remote Sens. 61, 1–20. https://doi.org/10.1109/TGRS.2022.3176603.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. https://doi.org/10.1109/ICCV.2017.74.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002. https://doi.org/10.1109/iccv48922.2021.00986.
J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: pp. 248–255. https://doi.org/10.1109/cvprw.2009.5206848.