Analytical Assessment of Pre-Trained Prompt-Based Multimodal Deep Learning Models for UAV-Based Object Detection Supporting Environmental Crimes Monitoring

geomatics

MDPI

Article

Analytical Assessment of Pre-Trained Prompt-Based Multimodal Deep Learning Models for UAV-Based Object Detection Supporting Environmental Crimes Monitoring

Andrea Demartis [1,*], Fabio Giulio Tonolo [1], Francesco Barchi [2], Samuel Zanella [2] and Andrea Acquaviva [2]

[1] LabG4CH—Laboratory of Geomatics for Cultural Heritage, Department of Architecture and Design (DAD), Politecnico di Torino, Viale Mattioli 39, 10125 Torino, Italy; fabio.giuliotonolo@polito.it
[2] Department of Electrical, Electronic, and Information Engineering, Università di Bologna, Viale del Risorgimento 2, 40136 Bologna, Italy; francesco.barchi@unibo.it (F.B.); samuel.zanella@unibo.it (S.Z.); andrea.acquaviva@unibo.it (A.A.)
* Correspondence: andrea.demartis@polito.it

Abstract

Illegal dumping poses serious risks to ecosystems and human health, requiring effective and timely monitoring strategies. Advances in uncrewed aerial vehicles (UAVs), photogrammetry, and deep learning (DL) have created new opportunities for detecting and characterizing waste objects over large areas. Within the framework of the EMERITUS Project, an EU Horizon Europe initiative supporting the fight against environmental crimes, this study evaluates the performance of pre-trained prompt-based multimodal (PBM) DL models integrated into ArcGIS Pro for object detection and segmentation. To test such models, UAV surveys were specially conducted at a semi-controlled test site in northern Italy, producing very high-resolution orthoimages and video frames populated with simulated waste objects such as tyres, barrels, and sand piles. Three PBM models (CLIPSeg, GroundingDINO, and TextSAM) were tested under varying hyperparameters and input conditions, including orthophotos at multiple resolutions and frames extracted from UAV-acquired videos. Results show that model performance is highly dependent on object type and imagery resolution. In contrast, within the limited ranges tested, hyperparameter tuning rarely produced significant improvements. The evaluation of the models was performed using low IoU to generalize across different types of detection models and to focus on the ability of detecting object. When evaluating the models with orthoimagery, CLIPSeg achieved the highest accuracy with F1 scores up to 0.88 for tyres, whereas barrels and ambiguous classes consistently underperformed. Video-derived (oblique) frames generally outperformed orthophotos, reflecting a closer match to model training perspectives. Despite the current limitations in performances highlighted by the tests, PBM models demonstrate strong potential for democratizing GeoAI (Geospatial Artificial Intelligence). These tools effectively enable non-expert users to employ zero-shot classification in UAV-based monitoring workflows targeting environmental crime.

Keywords: GeoAI; deep learning; UAV; multimodal; prompt-based models; object detection; instance segmentation; environmental crimes monitoring

Academic Editor: Enrico Corrado Borgogno Mondino

Received: 19 November 2025
Revised: 15 January 2026
Accepted: 27 January 2026
Published: 3 February 2026

1. Introduction

Illegal dumping is a serious and growing problem in human settlements, and many studies demonstrate the risks to human health posed by living in areas close to improperly

disposed waste [1]. Waste monitoring plays a central role in avoiding illegal dumping and correctly disposing of municipal solid waste. However, where it fails, the correct intervention in removing illegally dumped waste objects is crucial and has to happen as quickly as possible to avoid risks to the environment and local human communities [2]. The formation of leachate from the waste decomposition, for example, can contaminate surface and underground water, consequently poisoning local fauna or potable water for nearby human settlements [3]. Fighting illegal dumping, however, is not a trivial problem. The areas at risk have to be monitored in order to know if waste is present and where it is, and subsequently the intervention to remove it has a cost to the local responsible administration [1]. To support waste monitoring efforts, uncrewed aerial vehicles (UAVs) have become a widely adopted tool for surveying waste and high-risk areas [4].

In recent years, the acquisition of geospatial data has become increasingly efficient thanks to the development of new platforms including affordable commercial off-the-shelf (COTS) UAV systems equipped with increasingly efficient sensors [5,6]. Additionally, the possibility to plan flights—now integrated into a growing number of photogrammetric flight planning tools—optimizes and consequently accelerates primary data acquisition procedures.

Combining these technologies with advanced computer vision and structure from motion techniques [7] within photogrammetry software accelerated the processing of raw (or primary) data. This allowed for efficient creation of 3D point clouds and derived products, including digital surface models (DSMs), digital terrain models (DTMs), and orthomosaics.

Recently, the advent of deep learning (DL) has opened up new possibilities in this field. The integration of these products with DL models is currently being extensively studied by the scientific community with applications in various tasks such as data classification, object detection (OD), instance segmentation (IS), and others [8]. While DL offers several possibilities, they often require a certain level of coding expertise to be implemented into a workflow, which can be a limitation for many users.

To address this challenge, commercial platforms for the management of geospatial data have started to introduce DL interfaces that allows the use of pre-trained predictive models. Many of these models are available online (examples include MMDetections [9], the YOLO family that has reached YOLOv8 [10], and DINO [11] which is part of this research); however, one of the advantages of their integration in commercial platforms lies in their user-oriented interface, which shifts the expertise from coding to the critical assessment of results. This allows users with relatively low expertise in coding to apply DL to their work to perform tasks that otherwise would be time consuming—or impossible. OD tasks made available by DL, integrated with UAV surveys, can support interventions against illegal dumping, making the monitoring process faster and more cost-effective [4].

DL models, however, must be trained for the specific task for which they are intended to perform; OD models for illegally dumped waste must be trained—or fine-tuned—on a dataset containing the objects that are intended to be found in the model's inference. This is a non-trivial problem; the acquisition of a training dataset for fine-tuning applications or from-scratch training remains a major challenge in the scientific community [12], as it is one of the most time-consuming steps in the training of a DL model.

Recent advancements in deep learning can help address this issue. Multimodal DL models are now available in commercial platforms [13,14]. These models process multiple input types simultaneously, such as images combined with text prompts describing the target object. This capability enables OD and IS tasks with reduced technical barriers compared to traditional approaches. However, some programming knowledge and familiarity with API usage remain necessary for implementation. The key advantage of multimodal models lies in their zero-shot classification performance [15]. These models can classify inputs into classes they were not explicitly trained for, namely, classes that were

not present in their original training data. This feature significantly reduces the need for task-specific model training and labelled datasets, making the technology more accessible for specialized applications.

This paper evaluates the performance of several pre-trained prompt-based multimodal (PBM) DL models within a commercial GIS platform. The goal is to determine if their zero-shot classification capabilities can effectively detect common waste objects in relatively open field environments. This paper investigates the potential of PBM DL models for automatically detecting illegal waste dumping in diverse aerial imagery.

Specifically, we evaluate whether these models exhibit invariance to variations in image resolution and orientation, crucial for real-world application with imagery acquired from varying drone platforms and flight parameters, when performing zero-shot classification.

A key contribution of this work is a newly created, annotated dataset of simulated illegal dumping scenarios captured during aerial surveys conducted by Politecnico di Torino within the EMERITUS Project, an EU Horizon project focused on creating a unified platform for investigating environmental crimes. The dataset comprises both high-resolution orthoimages and frames extracted from drone videos, acquired during dedicated survey campaigns. These surveys were specifically designed and implemented to simulate realistic illegal dumping sites, allowing us to evaluate model performance under controlled yet representative conditions. Section 2 details the test area, data acquisition procedures (for both orthoimagery and video frames), and the rationale behind our selection of PBM models.

Furthermore, Section 2 outlines the inference and validation methodologies employed to test and assess model performance. Section 3 presents the results obtained from both orthoimages and video frames, including an analysis of how hyperparameter tuning impacts model outputs. Finally, Section 4 offers a discussion of these results and explores their potential application within supervised operational workflows for environmental monitoring and enforcement.

2. Materials and Methods

2.1. Description of the Test Area

The chosen test area is the former NATO military base "Calvarina", located in the municipality of Roncà (Verona, Italy). The site is situated in the hills near Verona and Vicenza, in north-eastern Italy, with approximate WGS84 geographical coordinates of 45.508° N, 11.281° E (Figure 1). The site, currently managed by SAFE, a non-profit foundation dedicated to the promotion and implementation of security, defence, and civil protection initiatives, was used as a semi-controlled environment to simulate the presence of an illegal dumping site. To do so, objects attributable to an illegal dumping site were arranged in the area. Specifically, barrels, tyres, and piles of sand on plastic sheets were distributed in the area (Figure 2). The majority of the objects were tyres, scattered across the area in different configurations, including single tyres, pairs, and larger groups, with varying orientations (vertical, horizontal, tilted). The metallic barrels varied in colour and orientation (horizontal, vertical). Additionally, some have lids, while others do not. Lastly, a varying number of sand piles were located in the middle of the asphalt-covered area. Occasionally, other smaller objects were present, but not the main focus of any analysis. The objects were moved in different scenarios to allow the creation of multitemporal products, mainly orthoimagery.

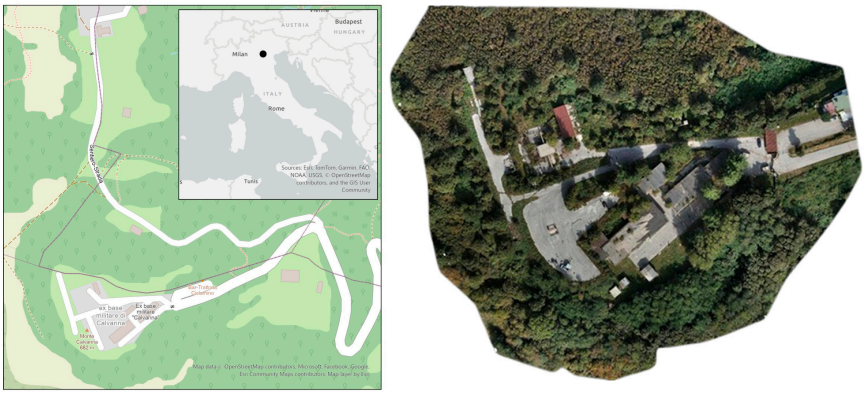

Figure 1. Location of the test area (**left**) and UAV-based orthoimagery (**right**).

Figure 2. Examples of the waste objects in the semi-supervised testing site.

2.2. Flight Plan, Data Acquisition, and Photogrammetric Processing

The imagery used in this research was generated deploying UAVs to carry out photogrammetric flights. During the duration of the project, the objects previously cited were moved before every UAV acquisition in order to allow change detection analysis between the multitemporal products acquired.

The first survey consisted of two photogrammetric UAV flights performed on 5 October 2023 with a DJI Matrice 300 equipped with a DJI ZENMUSE P1 optical sensor (full-frame 35, 9 mm × 24 mm) at a flight height of 80 m above take-off point to have a ground sampling distance (GSD) of about 1 cm. A photogrammetric flight plan was designed to cover the test area with side and cross overlap of about 80% and with orthogonal flight-paths, leading to the acquisition of >500 images for each flight (Figure 3).

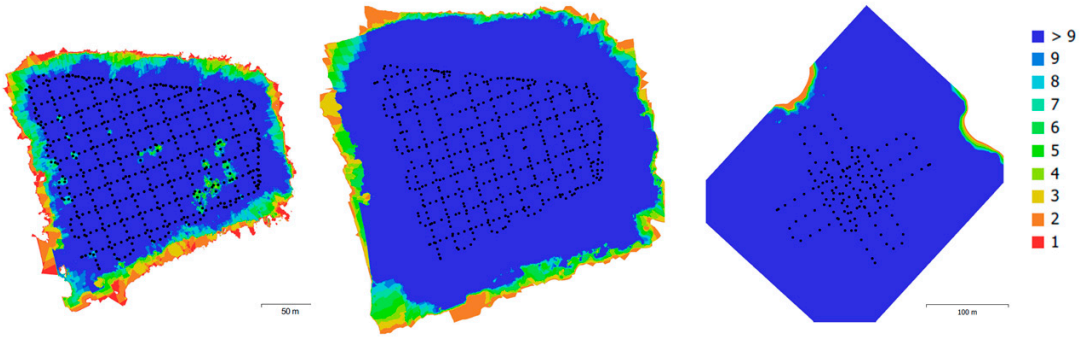

Figure 3. Camera locations (black dots) and images overlap for the UAV flights carried out on 5 October 2023, morning (**left**) and afternoon (**center**), and 27 June 2024 (**right**).

The flights were carried out in the morning (T1) and in the afternoon (T2) to allow the scenario to be changed in terms of waste object positions and locations. The change in acquisition time inevitably also had an impact on the brightness conditions of the scene.

Fifteen photogrammetric markers were accurately placed across the area and measured using a GNSS RTK approach with centimetre-level accuracy, enabling 3D accuracies of the final products of about 5 cm. Among these products obtained with rigorous and consolidated photogrammetric processing procedures using the OTS software Agisoft Metashape (v 1.8.5), two orthoimages with different objects distribution and a GSD of 0.01 m covering an area of approximately 200 m × 200 m were used as input imagery for the OD tests.

The second survey was carried out on 27 June 2024 using DJI Mavic 3M (4/3 sensor, 17.3 mm × 13 mm) with a flight height of 65 m above the take-off point. The third photogrammetric flight (T3), based on settings similar to the those used for the previous survey —also in terms of control points—lead to the acquisition of more than 600 images, enabling the generation of an orthoimage with a GSD of 3 cm with a different configuration of waste objects compared to the other two orthoimages.

The three orthoimages are the input imagery for the tested DL models for different tasks, in particular OD, IS, semantic segmentation (SS), and change detection (CD).

While they were not the main focus of the UAV flights, the scenes also contain decaying buildings, while the surrounding area is forested. These elements introduce challenges for detection models, potentially leading to false positives, but also provide a realistic testing environment with varied surfaces. The complexity provided by these elements is given by the introduction of features not belonging to the target objects but with patterns similar to them. In the example of this work, dark shadowed areas in tree crowns can look like black tyres to the models, metallic roofs were mistakenly identified as barrels, and different types of ground (asphalted, cemented, or bare) were consistently mistakenly identified as sand piles. Given the impossibility to unequivocally define a text prompt for sand piles, which leads to different interpretations of sand-related features from the text encoders (TE) of the models, this research focused on tyres and barrels only. We acknowledge that real-world illegal dumping scenarios include a much broader variety of heterogeneous objects, and this limited object diversity constitutes an important limitation of the present models.

Moreover, target objects were placed both on asphalt and on grass, allowing also to test object recognition with different backgrounds (Figure 4). This variety makes the scenarios an ideal testing ground for these models, simulating real-world scenarios where orthophotos have a large coverage and are not necessarily focused only on the target object. The orthoimages used in this research can be found in Figures A1–A3.

Figure 4. Details of the objects in one of the orthoimages.

2.3. Description of the Models

This work was conducted in ArcGIS Pro (versions 3.4.2 to 3.5.3) using a workstation equipped with an Intel Core i9-14900KF CPU @ 3.200 Mhz, 64 GB RAM, an NVIDIA GeForce RTX 4070 GPU with 12 GB VRAM, and Windows 11 Pro.

The model selection process began with an analysis of the living atlas (LA), Esri's platform for imagery, maps, and pre-trained DL models. These models are provided in the ".dlpk" (Deep Learning Package) format, which is required by ArcGIS Pro. Although the LA supports user-uploaded models, this analysis focused specifically on predictive models provided by Esri, indicated by the "esri_analytics" label (https://www.arcgis.com/home/user.html?user=d117460815774d4bbdf6fa2df5e4a7fd, accessed on 29 January 2026). The available models can be subdivided in two main categories: traditional and PBM models. Under traditional models are grouped all the models that are based on a training dataset containing a defined set of target objects and that will be able to recognize the same objects into new imagery. PBM models in this study include all the models that require as input both an image and a text prompt written by the user. The selected models all follow an architecture that resembles the architecture reported in Figure 5, extracted from [16].

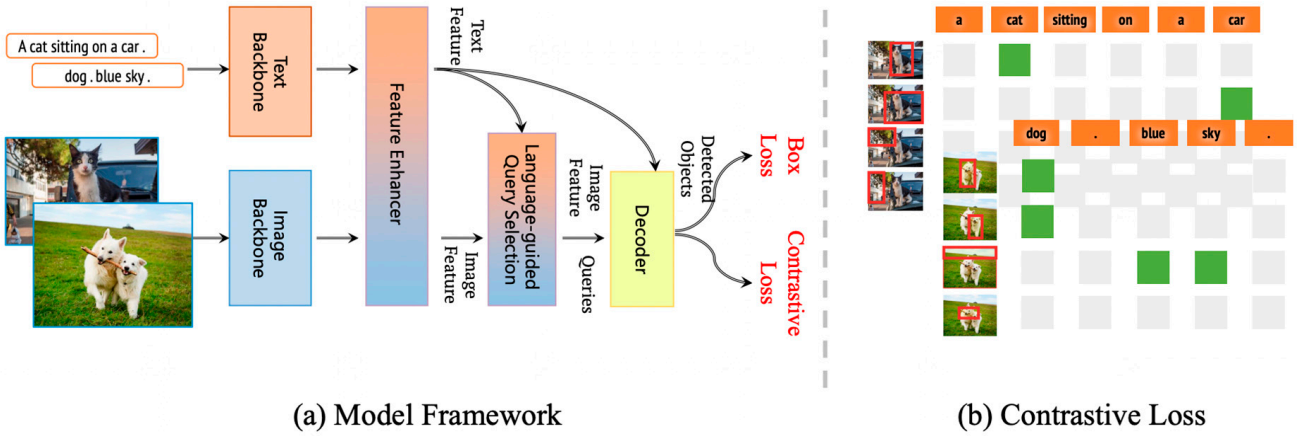

Figure 5. Architecture of GroundingDINO [16], an example of a PBM model.

Prompt-based multimodal models offer several advantages, primarily including their simplified inference process from a user perspective. Unlike traditional object detectors, models like GroundingDINO can perform zero-shot detection when desired classes are absent from training data, enabling object localization through natural language descriptions without explicit annotation during training. This open-set generalization capability stems from the cross-modality fusion architecture, where TEs process free-form textual prompts that are grounded to visual features through attention mechanisms, enabling language-guided OD. This is achieved by exploiting the TE's ability to trace features by similarity with other objects present in the dataset. This flexibility and advantage come at the expenses of performance, which is lower compared to a traditional model tailored for the same objects.

Below is a list of the selected models along with a brief description of their functionalities. Due to fixes to the work and updates in the models uploaded in the LA released during the research, the first part of the research uses more recent versions of the models (Summer 2025), while the work that carried out on frames extracted by UAV videos (that will be explained later in this section) uses a previous version. More recent releases of the models might perform differently. Nevertheless, the models with the used releases include the following:

- CLIPSeg [17] (released in LA: 10 December 2024; version used: 9 July 2025):

CLIPSeg, available in the LA under the name prompt-based segmentation, is based on CLIP [18], a multimodal learning model trained on a dataset of 400 million image-text pairs collected from the internet. Unlike traditional OD or IS models, CLIPSeg is a pixel classification model that performs semantic segmentation and produces a binary segmentation mask based on a given text or visual prompt. Although it does not inherently perform OD or IS, its output can be converted into polygonal features using a 'raster to polygon' tool, effectively enabling IS.

- GroundingDINO [19] (released in LA: 12 July 2024; version used: 9 July 2025):

GroundingDINO is an open-set, open-source OD model that can be prompted using free-form text. It achieves this by integrating a TE with a traditional OD model, enabling it to generalize beyond predefined object categories. The model is trained on multiple datasets including COCO [20]. However, this information is not reported in the model's documentation uploaded in the LA directly; it requires some further research. The outputs are bounding boxes (BBs) that, since the model is integrated in the software, are converted into polygons and transformed into GIS features.

- TextSAM [21] (released on 2 February 2024; version used: 7 August 2025):

TextSAM is an open-source model that enables IS of objects in a given scene by processing an input image and a free-form text prompt describing the object. This model combines two different DL models:

- GroundingDINO: An open-set OD model that locates objects based on a previously described text prompt.
- Segment Anything Model (SAM) [22]: A segmentation model that performs pixel-wise classification into different categories. The workflow involves detecting objects using GroundingDINO and passing the obtained BBs to the SAM model, which refines the segmentation by highlighting relevant features. The final output is a precise polygon representing the segmented object(s).

2.4. Data Preparation for Deep Learning

The photogrammetric processing of the UAV flights produced multitemporal orthophotos of the study area, acquired at different times and with varying resolutions (ranging from 0.01 m to 0.03 m). Three orthophotos were selected for analysis, each representing a different point in time with objects arranged in distinct configurations to simulate three different waste disposal scenarios. To correctly assess the performances of the models, the orthoimages were cropped to retain only the central part of the mosaics and to avoid areas with information gaps or geometrical distortions. As previously explained, two of the selected orthophotos were captured on the same day featuring the same objects in slightly different arrangements. Due to the similarity in acquisition and processing, they were cropped on the same boundary. The third orthophoto, acquired the following year, features the same classes of objects but arranged in a completely different distribution. Due to the difference characteristics of the flights, a different cropping boundary was used. This different cropping boundary resulted in a final orthoimage that is less abundant in vegetation.

2.5. Ground Truth Dataset Annotation and Metric Computation Procedure

The ground truth (GT) datasets were generated using a manual labelling process carried out directly in ArcGIS Pro by creating distinct feature datasets for each object class in each orthoimage by means of visual interpretation by an image analyst. The created GT datasets precisely follow the borders of the objects, emulating IS. Even if some artifacts at

the border of the orthoimage were subject to some distortion, they were still labelled. This choice was justified by the high density of objects near the affected areas and by preliminary tests showing that the model could recognize with some success the distorted or partially visible objects.

The labelled features served as the reference data for evaluating the model's performance in OD and segmentation tasks (Figure 6, on the left side). This assessment was conducted using the 'Compute Accuracy for Object Detection' tool available in the software, which generates an accuracy report and an accuracy table for each model inference based on a given intersection over union (IoU) threshold, namely the ratio between the intersection of the predicted and GT feature over their union. The report includes the key metrics required for the assessment of these tasks, namely precision, recall, F1 score, average precision, true positives, false positives, and false negatives (Figure A4). Additionally, the report is subdivided across multiple IoU thresholds, starting from the value specified as input under the hyperparameter "IoU Threshold" when the tool is launched and extending to IoU $\geq$ 0.5, 0.75, and 0.95. During evaluation, a deliberately low IoU threshold equal to 0.1 was adopted. The PBMs used in this research return different type of masks. This different behaviour would lead to different IoUs for two models that detect the same object but only due to the nature of the generated mask. Under conventional thresholds (e.g., 0.5), many practically useful detections would be misclassified as false negatives, preventing meaningful cross-model comparison. The chosen threshold therefore evaluates whether a predicted mask successfully includes the target object, rather than the precision of its delineation. We emphasize that this approach potentially inflates classical OD metrics when intended as detection and segmentation model and that reported F1 scores should be interpreted strictly as indicators of object-finding capability rather than segmentation accuracy. This choice considers masks containing both correct and incorrect regions as true positives (Figure 6), introducing an optimistic bias. To mitigate artefactual inflation, preprocessing steps such as minimum-area filtering and dissolve operations were applied. The limitations of this evaluation design are explicitly discussed in Section 2.7.

Figure 6. Left: Examples of ground truth labels with tyres. **Right**: Examples of the TextSAM model mask that includes both a true positive and a false positive, considered a true positive.

2.6. DL Models Inference

Inference of the models was performed using the DL tools available in ArcGIS Pro, after previous installations of the required libraries. In particular, the tools utilized for inference of the models are 'Detect Object using Deep Learning' for the models TextSAM and GroundingDINO and 'Classify Pixels using Deep Learning' for CLIPSeg. The tools require three parameters, namely an input raster, an output raster, and a pre-trained model as a ".dlpk". After the paths to the inputs and outputs are indicated, the tools allow the definition of parameters specifically supported by the model. Such parameters are reported in Table A1, where descriptions of the parameter are reported as it is in the ArcGIS Pro documentation. It is important to note that the release 3.5 of ArcGIS Pro has changed the GUI in the tools mentioned before in the subsection, removing the possibility to add hyperparameters easily. Adding hyperparameters is still possible using the python command in a python environment. After preliminary tests, the hyperparameters (namely the settings that the user can define when launching the inference) identified as most impactful were text_prompt, padding, batch_size, and tile_size, which are the focus of the extensive parametrization performed in this research. Before starting with the main experiments, preliminary analyses were performed to understand the ranges in which the hyperparameters could be changed effectively. Given the nature of the chosen hyperparameters, the only one that needed deep testing was tile_size; a threshold could be set around 1500 pixels. These hyperparameters were then modified using a grid approach with fixed width. It is important to note that filters based on the confidence value, returned by models both for rasters and polygons, were not used in this research to improve the results due to their ineffectiveness. Increasing the confidence threshold during inference led to the removal of many true positive predictions while having minimal impact on false positives. Consequently, the confidence threshold was kept at default value, and different strategies were applied to improve results, as described in the next paragraph.

2.7. Preliminary Elaborations

The results from each model inference underwent a cleaning process to filter and improve the results. This included non-maximum suppression (NMS, a technique that keeps only the highest-confidence detection among overlapping predictions) with a maximum overlap ratio of 0.9 to reduce redundant features over the same detected object, followed by a feature merging step using the 'Dissolve' tool to decrease the number of features without further reducing their detection potential (applied only to TextSAM and GroundingDINO inferences since CLIPSeg returned first a raster and only after the conversion to polygons features were clearly distinct). While confidence was not a reliable parameter for filtering predictions, NMS utilizes it to eliminate overlapping features, minimizing the loss of potentially correct masks. Therefore, a high overlap ratio threshold was chosen to activate NMS only in cases of highly overlapping detections. Additionally, a noise filter was applied based on feature area, retaining only features with an area greater than 0.2 m^2, as the target objects are never expected to be smaller under any condition. No upper area threshold was applied, as model outputs could represent either individual objects or groups of objects covering larger areas. For the CLIPSeg model specifically, a feature extraction process was necessary to evaluate the inferences. This was performed using a dedicated Python script integrated into ArcGIS Pro that automated the process exporting raster returned by the inference and the launch of the 'Raster to polygon' tool to convert them into polygons. After this additional preprocessing step, the polygons went under the same cleaning process as outputs from the other models.

A summary of the preliminary elaborations is reported in Figure 7.

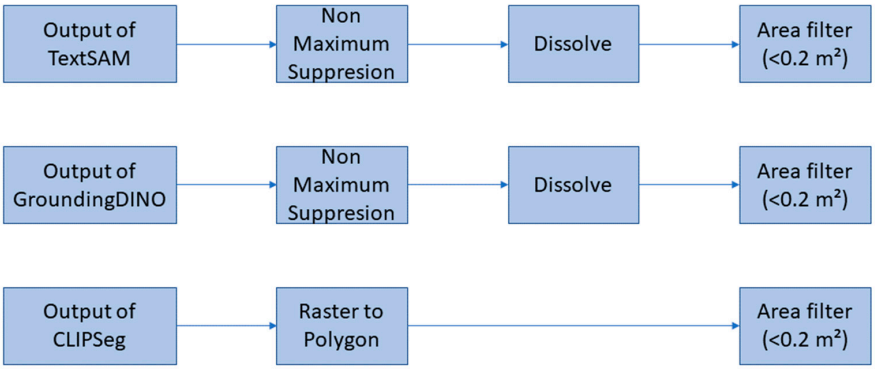

Figure 7. Summary of preliminary elaborations according to the used model.

2.8. Frame Extraction from UAV-Acquired Videos

During the photogrammetric survey carried out at T3, video footage of the test area was also recorded. Among this video footage, five were selected to test the models also on non-ortho-corrected and mosaicked images and/or non-nadiral images, and a total of 14 frames were extracted. Looking at Figure 8, the top left frame is referred to as "Single Frame", and the top right frame is referred to as "Horizontal Frame". The second, third, and fourth rows are referred to as "T3_Video", "Approaching_Video", and "Circular_Video", respectively.

Figure 8. Frames extracted from the videos. Single frames (upper row, **left**: single frame, **right**: horizontal frame) and four frames from the longer video referred as the nadiral frames.

2.9. Generative AI

Generative AI provided a contribution with code generation. Specifically, coding was used to automate the process of converting geospatial data types, to manage big data related to the results produced by the evaluation of the inferences launched with the models on the imagery, and to create graphs from said results. Regardless of the level of AI involvement, all AI-generated content was critically reviewed and edited to align with academic rigor. The author takes full responsibility for the accuracy, originality, and

integrity of the final work. Generative AI tools were also used for editorial support, such as grammar correction, translations, paraphrasing suggestions, or improving readability.

3. Results

As detailed in the previous paragraphs, the experiments were conducted both on orthoimages and on video footage frames, both based on UAV acquisitions. All models were tested varying the hyperparameters with fixed intervals for a total of 79 inferences per orthoimage, for a total 237 inferences. Inferences and inferences on frames extracted from videos were launched with default parameters for a total of 36 inferences.

3.1. Model Inference on the Original Orthoimages

The first tests conducted were on orthoimages using a comprehensive hyperparameter sensitivity analysis to assess the accuracy metrics and their dependence from hyperparameters, namely inferencing each model multiple times by varying the target hyperparameters. An example of the output from the model in orthoimage at time T3 can be found in Figure 9a, while some examples of true positive and false positive in orthoimages T1 and T2 can be found in Figure 9b,c. Blurred areas (Figure 9c) correspond to typical artifacts in orthoimagery, especially in border areas.

The analysis was repeated for each orthoimage representing the three multitemporal scenarios and for each class of object. By consulting the code attached in the LA page of the models—not available at the start of the research, but available now—it is possible to confirm that the models accept complex text prompts as input. The output of the models has a "Class" label. In the case of a simple prompt (one word), the models will recall the text prompt exactly. However, in the case of a complex text prompt (more than one word), the models will test all the words in the prompt, first singularly and then combined. This means that the output masks can be associated to the correct class (the complete prompt) or just part of it (a combination of the other words). For the purpose of this research, however, the implemented prompting design was as simple as possible, specifically only one word. The motivation behind this choice is to establish how the models perform under straightforward inputs, avoiding potential biases or noises introduced by complex text prompting. Future experiments will evaluate how complex text prompting can influence model performance.

In Figure 10, for each object, orthoimage, and model, the inferences were too numerous; thus, a boxplot was used to visualize the results. The graphs report results regarding barrels in the top row and tyres in the bottom row. The height of each box represents results from Q1 (the 25th percentile) to Q3 (the 75th percentile). The lower whisker limit is $Q1 - 1.5 \times IQR$, and the upper whisker limit is $Q3 + 1.5 \times IQR$ ($IQR = Q3 - Q1$). Points outside the whiskers (e.g., results on the right in Figure 10a) are considered outliers. The results in Figure 10 show the evaluation of each inference used in this research, which are reported in Tables A2–A4, subdivided per orthoimage, object, and model used. Default inferences use batch_size = 4, padding = 256, and tile_size = 1024 for TextSAM and GroundingDINO, while they use batch_size = 4, padding = 88, and fixed tile_size = 352 for CLIPSeg. The results of imagery at times T1 and T2 (Figure 10, first two columns) highlighted how the hyperparametrization, in the ranges tested in this study, does not impact sensibly the output of the models, with an F1 score from default conditions that can vary by $\pm 10\%$. An exception is made for the model CLIPSeg that always shows better results with default conditions. These results apply only to the hyperparameter ranges that were be feasibly tested in this study. Because these ranges were constrained by model behaviour and computational limitations, no conclusions can be drawn for broader or more principled hyperparameter optimization strategies, which may identify different or more effective configurations.

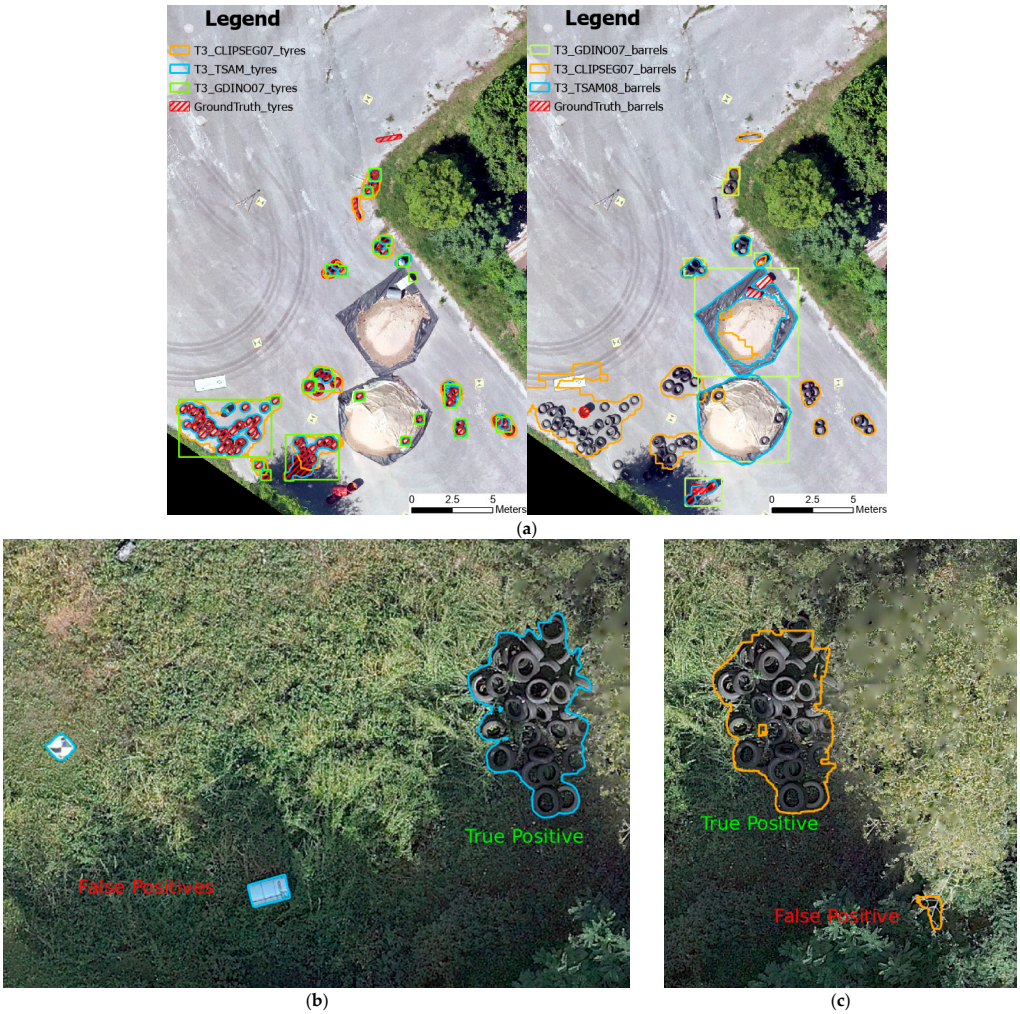

Figure 9. (**a**) Example of the outputs from the three models in orthoimages acquired at time T3; (**b**) snapshot of true and false positives with the TextSAM model; (**c**) snapshots of true positives and false positives with the CLIPSeg model.

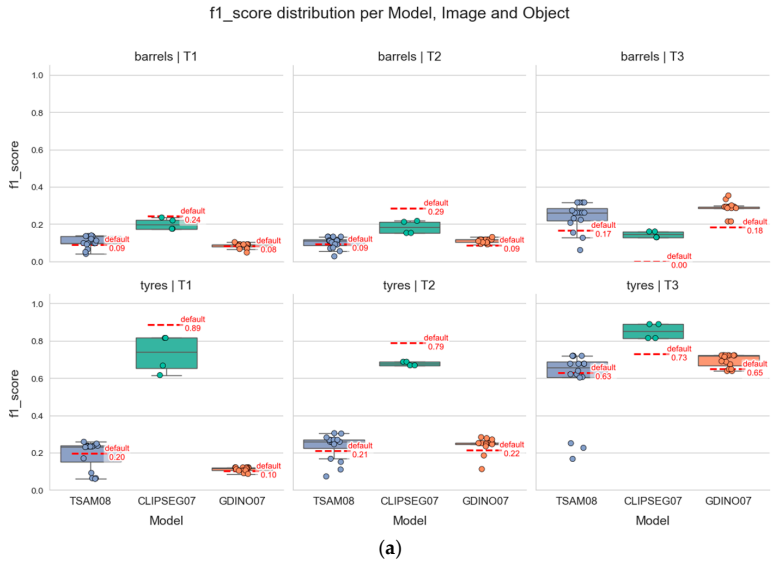

Figure 10. *Cont*.

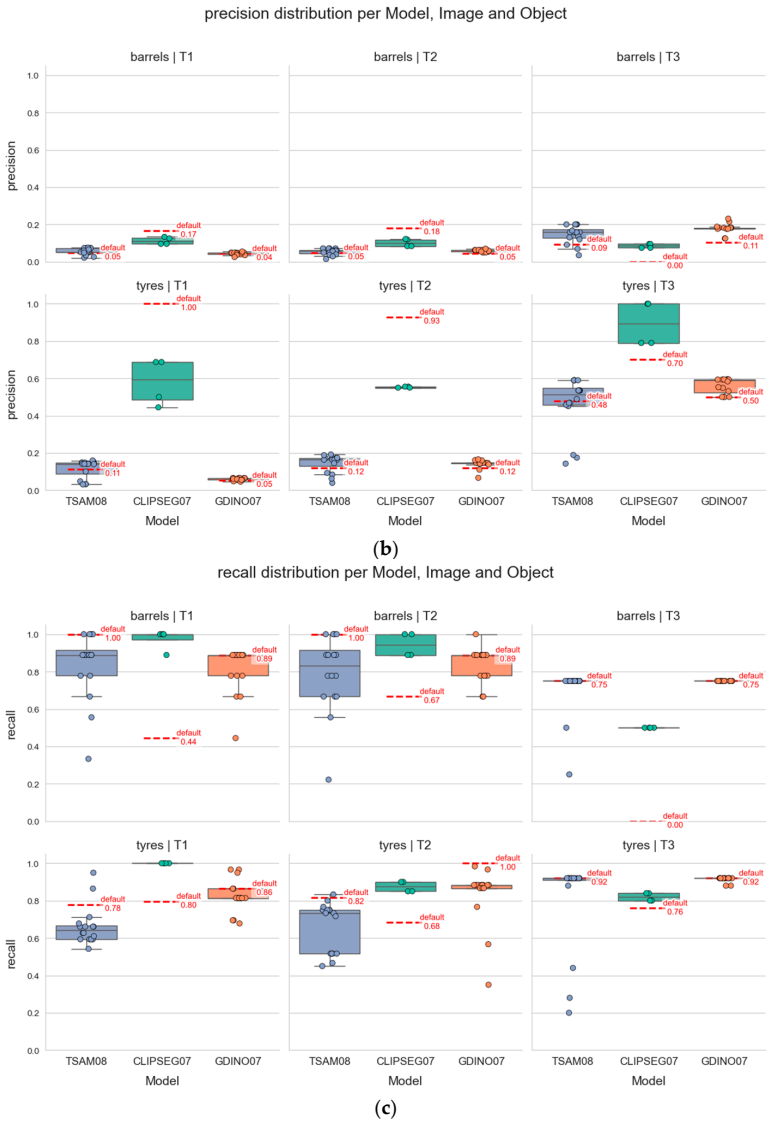

Figure 10. Evaluation of the performances of the models subdivided per model, object, and orthoimage according to (**a**) F1-score; (**b**) precision; (**c**) recall.

Additionally, considering only this last model, the hyperparametrization settings that produced good results for the orthoimage at time T1 did not return the same range of results for the orthoimage at time T2. Specifically, the results with this second orthoimage were less variable but also with a ~−10% F1 score. This occurred even though the input imagery has the same technical specifications and only the acquisition time and the object distribution varied. The results related to orthoimage acquired at time T3 show more comparable results across the models compared to the previous orthoimages. Moreover, results at T3 shows how default conditions on average perform worse than any other hyperparametrization.

According to F1 score, the best results are obtained with "tyres" and the model CLIPSeg that returned an F1 score = 0.88 in both orthoimages T1 and T3, making it also the most robust model across resolution changes.

Lastly, evaluating only F1 score, it is interesting to highlight that hyperparameters different from default do not always improve the results, which was an expected output given the variability of the type of input data supported by the models. This means that the models are influenced by the resolution of the input data more than the hyperparametrization, which is an important result. The most impacting change between default inferences

and the other inferences is the high padding, but this does not necessarily demonstrate that high padding is a good solution since the default configuration of hyperparameters could—by coincidence—create padded areas that remove areas that are rich in false positives. While it is an outlier, the worst output was obtained with default conditions for orthoimage T3 and default conditions for CLIPSeg, which is also the model that returned the best results in other conditions, as previously shown in this section.

While all models analysed in this work show potential, it is clear that while correctly performing zero-shot classification, the user that intends to use these models must consider a quality control step in their workflow. The level of consistency of the models showed that they cannot be entrusted with unsupervised OD with specific objects if the input is a single word text prompt and an orthophoto with sources of noise. The results of the experiments, paired with the code released in the LA documentation, allowed the formation of a hypothesis for a general role for each parameter analysed in the work. Table A1 presents the descriptions of the hyperparameters as found in the ArcGIS Pro documentation.

3.1.1. The Role of "Padding"

The description of padding in ArcGIS Pro documentation states that it defines the number of pixels in the border of a tile that are also included in the adjacent tile, making it responsible for the blending of predictions in adjacent ones (Table A1). This description, however, is in contrast with what can be found when analysing the code, where masks whose centroid falls inside this strip of pixels are filtered out. This means that objects in the orthoimage that fall inside the padded areas will never be detected. The role of padding is therefore to filter masks that are close to tile borders that have less context around them, while keeping only those with in the central part of the tile that have higher context. This small difference between documentation and actual functioning, however, can highly impact the inferences. The definition written in the documentation suggests that high padding will improve the results (since the objects will be detected twice and therefore only high-quality masks will be returned as output), but the actual results of high padding is to filter out high strips of pixels, which might remove a high number of correct masks. Padding can go up to half of the tile size (a higher one would be equal to a central area of the tile equal to 0). The experiments were conducted using a padding equal to 0 and equal to 1/16 of the tile size.

3.1.2. The Role of "Tile_Size"

When a large image is provided as input to the models, as in the case of this research, the model automatically subdivides it into tiles. Tile_size, in pixels, corresponds to the dimension of the side of a single tile, which is then analysed—entirely or paired with other tiles according to the batch_size—by the model. This parameter is important for different reasons. A sub-optimal tile size might not give enough context (or give too much context) to the model, which will have worse performance. Context is obtained both with other objects and features in the image, as well as from relative dimensions in the tile. In fact, by consulting the code attached in the documentation, it is possible to see that these types of pre-trained models do not use any information about the GSD of images. The fact that GSD is not used by the models, although it could improve the outputs of the models, allows the usage of the models with other input images. This was confirmed in other tests, which showed it was possible to appreciate their performances on non-nadiral and non-correctly scaled images. It is important to highlight that for the model "CLIPSeg", the model does not allow changes in the tile_size, which is fixed at 352 pixels. Obviously, the tile_size must not be smaller than the object under analysis. Adopting a value as large as the whole

orthophoto would be an error since it does not help the model in the OD task and greatly increases the inference time.

3.1.3. The Role of "Batch_Size"

When the input image is large and is subdivided in tiles, and when the GPU allows it, it is possible for the models to make inferences on more than one tile at a time. This parameter is called batch_size and corresponds to the number of tiles processed at the same time by the models. Batch_size can be defined as any value greater than zero, but it will assume values that make sure that the number of tiles processed together can be as close as possible to a square. The batch_size does not affect in any way the results of the inferences, but its interaction with padding, which will be described in more detail later, makes it important to evaluate it carefully. Other than this interaction, the value of this parameter can be as high as the available hardware allows, since it greatly reduces the computational time required by the model while decreasing of some percentage point the F1 score of the output.

3.1.4. Interaction Among Padding, Batch_Size, and Tile_Size

While they can be modified singularly in TextSAM and GroundingDINO, it is important to understand how interconnected these parameters are. If padding is greater than 0 and batch_size is equal to 1, strips of pixels at the border of each tile will be filtered out. However, if padding is positive and batch_size creates a shape larger than a 1×1 grid, padding will be applied to the borders of the whole grid. This means that masks that might be discarded with a specific batch_size because the border of the grid might be considered if the batch_size changes. Moreover, it means that a higher tile_size results in a lower padded area if padding is constant. However, it must be considered that high tile_sizes might penalize the model, worsening the outputs and increasing computational time.

3.2. Model Inference on Frames Extracted from UAV-Acquired Videos

The research then continued assessing the performances of the models on non-orthoprojected images, namely the frames extracted from the videos described in 2.9. An example of the inference on the frames can be found in Figure 11 ([Text Prompt = Tyres]).

The overall results of the tests are reported in Figure 12. Results are represented with swarm plots, where each dot represents a frame. The labels with frame numbers represent the best and the worst results. With non-orthoprojected imagery, considering F1 score for evaluation, the results obtained analysing tyres are sensibly better than those obtained for barrels. The same disparity between objects observed with orthoimages can be appreciated with this different type of imagery. Looking at Figure 12, it is possible to see that F1 scores with barrels hardly reach an F1 score of 0.7. Results obtained with tyres instead are consistently higher, with some very good cases reaching an F1 score of 1.

Moreover, considering only tyres, it is possible to appreciate how precision is higher compared to orthoimages. This is symptom of fewer false positives, which are wrong detections that all models tended to produce.

The apparent increase in performance observed in these few frames could be a reflection of the similarity between the non-orthoprojected imagery and the training datasets of the models; however, given the extremely limited size and heterogeneity of the video-frame dataset, this observation remains purely exploratory and does not support any generalizable conclusion.

Figure 11. Examples of the masks obtained with the models with frame extracted from UAV-acquired videos.

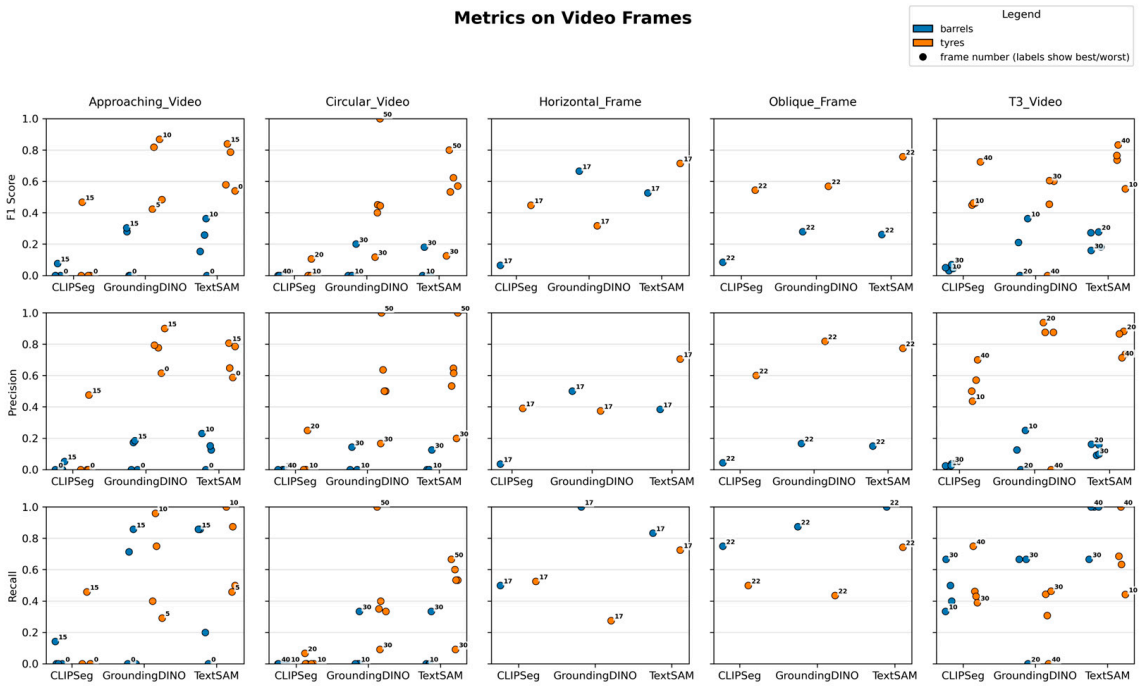

Figure 12. Average metrics of the models in the video frames, subdivided per metrics, model, object, and video.

4. Discussion

All the results in this study are obtained using experiments on datasets acquired in a single semi-controlled test area. The present analysis, therefore, does not provide any evidence regarding the models' ability to generalize to different geographic environments. Generalization to other contexts remains an open question and represents a limitation of this study. PBM DL models applied to orthoimages highlighted how their performances are not yet reliable for non-supervised monitoring applications of the tested objects in broad areas. The chosen models still present many limitations, especially while working with orthoimages. The most impacting limitation is due to the number of false positives that all models create (which leads to low values of precision). This makes it time-consuming for the end user to check them and then filter them. The results for the different orthoimages highlighted that inferences obtained using orthoimage T3 (which was generated with a focus on the non-vegetated area) as input returned less false positives and therefore better results on average.

The extensive hyperparametrization required to obtain a satisfactory result is not generalizable across different orthophotos even when all the conditions (same UAV flight parameters in the same day) are as similar as possible. An exception is made for the model CLIPSeg, which returned results with an F1 score of around 0.7, as well as for all orthoimages when working with tyres. This low generalizability makes it impossible for an end user to know which hyperparametrization configuration performs best since knowing this would require the user to have identified all the objects first and then assess the results. However, since the results obtained with default conditions are generally better than those obtained with a tailored hyperparametrization, it is not essential to vary the hyperparameters from the default conditions. Moreover, the inferences that return unsatisfactory metrics (F1 score lower than 0.5) can still be useful under certain circumstances. If time is taken out of the equation and the segmented features can be supervised, false positives have less impact on the performance, and the models allow localization of all the objects. On the contrary, if time is an important factor in the monitoring workflow, a hyperparametrization configuration that always returns high recall can be used to remove some of the objects with precise interventions. Moreover, this study demonstrates that, in the limited tested ranges, the interaction between the hyperparameters can highly vary the performances of the results but never effectively enough to be relevant. This result is valuable. Due to the low generalization offered by the model, it would not be possible to know which is the best hyperparametrization configuration because a ground truth is needed to evaluate the best hyperparametrization configuration. However, if the default hyperparameter configuration is close to optimal, this low generalization is less impactful. It is important to emphasize that these conclusions apply only to the hyperparametrization configuration that could be feasibly explored, and any extrapolation of these conclusions to similar situations with different model releases, inputs, or hyperparametrization configurations cannot be safely made. Different optimization approaches or strategies may return different results, including more effective configurations.

Regarding the research conducted on the PBM DL models applied to frames extracted by UAV videos, and analysing their higher performances on this input data, some hypotheses can made. DL models are usually trained on imagery that is neither from the nadiral perspective nor orthoprojected; this means that the datasets on which they are trained probably contain a majority of images that do not match the input data used in this research. The same hypothesis can be made for the TE, which is trained on large text datasets that probably do not describe common objects, such as tyres or barrels, from a nadiral perspective. The results obtained with frames extracted by video acquired with UAV tend to favour these hypotheses, since the performances of the models are generally higher in terms of

F1 scores. Moreover, distance played a factor in the results with frames, with models struggling to return accurate results with frames extracted from videos recorded at higher distances. Specifically, in the videos "Approaching_Video" and "Circular_Video", frames acquired from further distance generally perform poorly compared to the others. This does not automatically means that videos acquired from a closer distance always returned good performances. A clear example is highlighted by frame 40 in "T3_Video", where the model GroundingDINO could not find any tyres, although numerous were present. Compared with the results obtained with orthoimagery, it is possible to appreciate a growing trend in precision across the models that highlights behaviour that discourages false positive creation with this type of imagery—or acquisition geometry.

The last consideration regards the limitations of TEs. It is important for such architectures to univocally associate features to the text prompt. This is the main reason why working with the prompt "sand" was less consistent since this single word can represent a range of different scenarios. This same limitation is probably behind the worse results achieved with the text prompt "barrels" compared to "tyres". The first word can represent a range of shapes and materials, for example wooden barrels, or different colours, while the second one refers to objects that are almost always identical to one another.

Nevertheless, the limited variety of objects that could be tested in this study is not representative of the heterogeneous and mixed-material waste typically found at real illegal dumping sites. Since the evaluated models were not exposed to more complex object types such as plastics, metals, fabrics, irregular heaps, and composite waste, it is not possible to draw conclusions about their ability to generalize to the full spectrum of materials encountered in operational scenarios. The present results therefore cannot be extrapolated beyond the specific objects evaluated here, and expanding the dataset to include more diverse waste categories will be essential for generalized assessment of the models. The limited variety of objects tested in this study also makes it premature to assess whether the models can support environmental crimes from an ecological point of view. Despite the promising results, it is premature to extrapolate these considerations to real case scenarios and to different waste objects.

Lastly, given the inconsistent results between the objects under analysis and to successfully use these models, the authors advise to test and assess if the model is able to successfully detect the desired object before use.

5. Conclusions

The research had the objective to test and validate the DL models available in the LA to allow users without expertise in coding to exploit the advantages of these tools, with a focus on PBM models that allow the use of the same model for different objects. The testing fields of these models were very high-resolution orthoimages. Tests were also performed with frames extracted by videos acquired from multiple multitemporal UAV surveys. This research highlights how the current state of the art models chosen in this study do not perform well enough to be reliable in any type of non-supervised monitoring application, especially if no preliminary tests are performed in advance, since the performances vastly differ between different target objects. Specifically, in the case of this research, the models showed good metrics with tyres (F1 scores consistently above 0.7, with the best results obtained by the model CLIPSeg (F1 score = 0.88)), but improvable metrics with barrels (F1 scores lower than 0.4 even for the best results). Moreover, it was possible to assess that PBM models struggle with objects that do not have a non-ambiguous definition. The tests on the frames showed that the models can work in other UAV monitoring contexts, with performances on specific frames that yielded F1 scores of up to 0.82 with tyres and better for barrels (best F1 score = 0.65), both using GroundingDINO. One limitation seems

to be the tendency of the models to create false positives with buildings and in vegetated areas. Due to the black box nature of the models, is impossible to assess which part of the model is the bottleneck. However, given that the best performances were achieved with CLIPSeg, it is possible to suspect a different TE for GroundingDINO and TextSAM. Despite not performing well at an absolute level, the models can still be relevant in applications where only a high recall is needed, given the fact that it is the metric that all the model consistently return with considerably high values.

Future directions for this research include the continuous testing of the new releases to assess the improvements of the models, descriptive complex text prompting to help the models in the detection of the desired object, and testing with satellite imagery to assess the performances of the models with different platforms and geographic regions. Moreover, the possibility of fine-tuning will be explored to assess if it is more efficient to fine-tune a PBM model or to train a new task specific model with a fresh training dataset based on time and performance. In conclusion, the models show high potential in the democratisation of DL and in its integration in powerful non-supervised monitoring workflows in the next future for monitoring applications. All the results and metrics obtained in this research are specific to the tested target objects (tyres and barrels), which are untypical for these applications. Other research assessing the performances of these models with standard datasets containing common objects [19], or with fine-tuning techniques [23], highlight how the models already perform well enough to be implemented in operative workflows in those contexts.

Author Contributions: Conceptualization, A.D. and F.G.T.; methodology, A.D.; validation, A.D.; formal analysis, A.D.; investigation, A.D., F.G.T., F.B., S.Z. and A.A.; resources, F.G.T. and A.A.; data curation, A.D.; writing—original draft preparation, A.D.; writing—review and editing, A.D., F.G.T., F.B., S.Z. and A.A.; visualization, A.D.; supervision, F.G.T.; project administration, F.G.T.; funding acquisition, F.G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by EMERITUS project, which was funded by the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101073874.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy reasons.

Acknowledgments: We would like to thank the DAD—Laboratory of Geomatics for Cultural Heritage and DIATI—Laboratory of GIS and Photogrammetry from Politecnico di Torino for the field data acquisition and the data processing resources. We would like to thank SAFE for their support in the planning of the acquisition campaigns, as well as for their assistance during the data collection activities and the setup of the scenarios. During the preparation of this manuscript/study, the author(s) used ChatGPT, GPT 5, and GPT 4o mini for the purposes of code production and editorial support. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The orthoimages are reported in Figures A1–A3. The images highlight how the original orthophoto was cropped to maintain only the most complete part of it, but also how it still contains a lot of non-target objects, introducing a lot of noise in the image and therefore representing a real case application of the models.

Figure A1. Orthoimage used in the study, acquired at time (and referred to as) "T1".

Figure A2. Orthoimage used in the study, acquired at time (and referred to as) "T2".

https://doi.org/10.3390/geomatics6010014

Figure A3. Orthoimage used in the study, acquired at time (and referred to as) "T3".

Accuracy	Predictions/ Classifications	$\frac{\text{Correct}}{\text{Correct + Incorrect}}$
Precision	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Positive}}$
Recall	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Negative}}$
F1	Predictions/ Classifications	$\frac{\text{2 * True Positive}}{\text{True Positive + 0.5 (False Positive + False Negative)}}$
IoU	Object Detections/ Segmentations	$\frac{\text{Pixel Overlap}}{\text{Pixel Union}}$

Figure A4. Metric explanations [24].

Table A1. Summary of the hyperparameters the user can modify via the GUI or the equivalent python script.

Model	Parameters	Description
CLIPSeg	prompt	Text that describes the feature to be segmented. For visual prompting, an image URL or a local path can be provided.
	threshold	The pixel classifications with probability values higher than this threshold are included in the result. The allowed values range from 0 to 1.0. This threshold value will be effective only when return_probability_raster is set to false.
	batch_size	Number of image tiles processed in each step of the model inference. This depends on the memory of your graphic card.
	return_probability_raster	If true, the output classified raster will be a continuous magnitude raster indicating the probability value at each pixel.
	padding	Number of pixels at the border of image tiles from which predictions are blended for adjacent tiles. Increase its value to smooth the output while reducing edge artifacts. The maximum value of the padding can be half of the tile size value.
	test_time_augmentation	Performs test time augmentation while predicting. If true, predictions of flipped and rotated variants of the input image will be merged into the final output.
	merge policy	Policy for merging predictions (mean, min, or max). Applicable when test_time_augmentation is true.
TextSAM	text_prompt	Text that describes the objects to be detected. The input can be multiple text prompts, separated by commas, allowing the detection of multiple classes.
	padding	Number of pixels at the border of image tiles from which predictions are blended for adjacent tiles. Increase its value to smooth the output while reducing edge artifacts. The maximum value of the padding can be half of the tile size value.
	batch_size	Number of image tiles processed in each step of the model inference. This depends on the memory of your graphics card.
	box_threshold	The confidence score used for selecting the detections to be included in the results. The allowed values range from 0 to 1.0.
	text_threshold	The confidence score used for associating the detected objects with the provided text prompt. A higher value ensures strong association but potentially fewer matches. The allowed values range from 0 to 1.0.
	tta_scales	Performs test time augmentation while predicting by changing the scale of the image. Values in the range of 0.5 to 1.5 are recommended. Multiple scale values separated by commas can also be provided, for example, 0.9, 1, 1.1.
	box_nms_thresh	The box IoU cut-off used by non-maximal suppression to filter duplicate masks.
GroundingDINO	text_prompt	Text that describes the objects to be detected. The input can be multiple text prompts, separated by commas, allowing the detection of multiple classes.
	padding	Number of pixels at the border of image tiles from which predictions are blended for adjacent tiles. Increase its value to smooth the output while reducing edge artifacts. The maximum value of the padding can be half of the tile size value.

Table A1. *Cont.*

Model	Parameters	Description
GroundingDINO	batch_size	Number of image tiles processed in each step of the model inference. This depends on the memory of your graphics card.
	box_threshold	The confidence score used for selecting the detections to be included in the results. The allowed values range from 0 to 1.0.
	text_threshold	The confidence score used for associating the detected objects with the provided text prompt. A higher value ensures strong association but potentially fewer matches. The allowed values range from 0 to 1.0.
	tta_scales	Performs test time augmentation while predicting by changing the scale of the image. Values in the range of 0.5 to 1.5 are recommended. Multiple scale values separated by commas can also be provided, for example, 0.9, 1, 1.1.
	nms_overlap	The maximum overlap ratio for two overlapping features, which is defined as the ratio of intersection area over union area. The default is 0.1
	exclude_pad_detections	If true, filters potentially truncated detections near the edges that are in the padded region of image chips.

Table A2. Evaluation of inferences in orthophoto "T1" (red = tyres, blue = barrels). Source name is composed of the orthoimage, model used (TSAM stands for TextSAM, GDINO stands for GroundingDINO), object analysed, padding (p), batch size (bs), and tile size (ts).

Inferences with Orthophoto "T1"														
Text Prompt: Tyres								Text Prompt: Barrels						
Source.Name	Precision	Recall	F1_ Score	AP	True_ Positive	False_ Positive	False_ Negative	Source.Name	Precision	Recall	F1_Score	AP	True_ Positive	False_ Positive	False_ Negative
T1_CLIPSEG07_ tyres_default_.csv	1.00	0.80	0.89	1.26	8	0	12	T1_CLIPSEG07_ barrels_default_.csv	0.17	0.44	0.24	0.38	4	20	5
T1_CLIPSEG07_ tyres_p0_bs1_ts352.csv	0.69	1.00	0.81	0.69	11	5	0	T1_CLIPSEG07_barrels_ p22_bs1_ts352.csv	0.13	1.00	0.24	0.13	8	52	0
T1_CLIPSEG07_ tyres_p0_bs4_ts352.csv	0.69	1.00	0.81	0.69	11	5	0	T1_CLIPSEG07_barrels_ p22_bs4_ts352.csv	0.13	0.89	0.22	0.14	7	49	1
T1_CLIPSEG07_ tyres_p22_bs4_ts352.csv	0.50	1.00	0.67	0.50	8	8	0	T1_CLIPSEG07_barrels_ p0_bs1_ts352.csv	0.10	1.00	0.17	0.10	9	85	0
T1_CLIPSEG07_ tyres_p22_bs1_ts352.csv	0.44	1.00	0.62	0.44	8	10	0	T1_CLIPSEG07_barrels_ p0_bs4_ts352.csv	0.10	1.00	0.17	0.10	9	85	0
T1_TSAM08_tyres_p37_bs1_ ts600_NMS_dis.csv	0.16	0.68	0.26	0.24	26	137	19	T1_TSAM08_barrels_ p11_bs1_ts180_NMS_dis.csv	0.08	0.89	0.14	0.08	8	98	1
T1_TSAM08_tyres_p11_bs1_ ts180_NMS_dis.csv	0.15	0.63	0.25	0.25	24	131	22	T1_TSAM08_barrels_ p0_bs1_ts1020_NMS_dis.csv	0.07	0.89	0.14	0.08	8	101	1
T1_TSAM08_tyres_p0_bs1_ ts1020_NMS_dis.csv	0.15	0.59	0.24	0.25	20	115	24	T1_TSAM08_barrels_ p0_bs1_ts1480_NMS_dis.csv	0.07	0.89	0.14	0.08	8	101	1
T1_TSAM08_tyres_p0_bs1_ ts1480_NMS_dis.csv	0.15	0.59	0.24	0.25	20	115	24	T1_TSAM08_barrels_ p0_bs1_ts180_NMS_dis.csv	0.07	0.89	0.14	0.08	8	101	1
T1_TSAM08_tyres_p0_bs1_ ts180_NMS_dis.csv	0.15	0.59	0.24	0.25	20	115	24	T1_TSAM08_barrels_p0_bs1_ ts600_NMS_dis.csv	0.07	0.89	0.14	0.08	8	101	1
T1_TSAM08_tyres_p0_bs1_ ts600_NMS_dis.csv	0.15	0.59	0.24	0.25	20	115	24	T1_TSAM08_barrels_p37_bs1_ ts600_NMS_dis.csv	0.07	0.78	0.12	0.08	7	100	2
T1_TSAM08_tyres_p0_bs4_ ts1020_NMS_dis.csv	0.14	0.66	0.23	0.21	25	152	20	T1_TSAM08_barrels_p63_bs1_ ts1020_NMS_dis.csv	0.06	0.89	0.11	0.07	8	127	1
T1_TSAM08_tyres_p0_bs4_ ts1480_NMS_dis.csv	0.14	0.66	0.23	0.21	25	152	20	T1_GDINO07_barrels_p92_bs4_ ts1480_NMS_dis.csv	0.05	0.89	0.10	0.06	8	138	1
T1_TSAM08_tyres_p0_bs4_ ts180_NMS_dis.csv	0.14	0.66	0.23	0.21	25	152	20	T1_TSAM08_barrels_p11_bs4_ ts180_NMS_dis.csv	0.05	0.89	0.10	0.06	7	128	1
T1_TSAM08_tyres_p0_bs4_ ts600_NMS_dis.csv	0.14	0.66	0.23	0.21	25	152	20	T1_TSAM08_barrels_p0_bs4_ ts1020_NMS_dis.csv	0.05	1.00	0.10	0.05	8	148	0
T1_TSAM08_tyres_p63_bs1_ ts1020_NMS_dis.csv	0.14	0.63	0.23	0.22	25	153	22	T1_TSAM08_barrels_p0_bs4_ ts1480_NMS_dis.csv	0.05	1.00	0.10	0.05	8	148	0
T1_TSAM08_tyres_ default_NMS_dis.csv	0.11	0.78	0.20	0.14	29	230	13	T1_TSAM08_barrels_p0_bs4_ ts180_NMS_dis.csv	0.05	1.00	0.10	0.05	8	148	0
T1_TSAM08_tyres_p92_bs1_ ts1480_NMS_dis.csv	0.10	0.54	0.17	0.19	21	188	27	T1_TSAM08_barrels_p0_bs4_ ts600_NMS_dis.csv	0.05	1.00	0.10	0.05	8	148	0
T1_GDINO07_tyres_p0_bs1_ ts1020_NMS_dis.csv	0.07	0.81	0.12	0.08	17	244	11	T1_GDINO07_barrels_p0_bs4_ ts1020_NMS_dis.csv	0.05	1.00	0.10	0.05	8	148	0
T1_GDINO07_tyres_p0_bs1_ ts1480_NMS_dis.csv	0.07	0.81	0.12	0.08	17	244	11	T1_TSAM08_barrels_p92_bs1_ ts1480_NMS_dis.csv	0.05	0.78	0.09	0.06	7	137	2
T1_GDINO07_tyres_p0_bs1_ ts180_NMS_dis.csv	0.07	0.81	0.12	0.08	17	244	11	T1_GDINO07_barrels_p0_bs1_ ts1020_NMS_dis.csv	0.05	0.89	0.09	0.05	8	159	1

Table A2. *Cont.*

Inferences with Orthophoto "T1"															
Text Prompt: Tyres								**Text Prompt: Barrels**							
T1_GDINO07_tyres_p0_bs1_ts600_NMS_dis.csv	0.07	0.81	0.12	0.08	17	244	11	T1_GDINO07_barrels_p0_bs1_ts1480_NMS_dis.csv	0.05	0.89	0.09	0.05	8	159	1
T1_GDINO07_tyres_p0_bs4_ts1020_NMS_dis.csv	0.07	0.81	0.12	0.08	17	244	11	T1_GDINO07_barrels_p0_bs1_ts180_NMS_dis.csv	0.05	0.89	0.09	0.05	8	159	1
T1_GDINO07_tyres_p0_bs4_ts1480_NMS_dis.csv	0.07	0.81	0.12	0.08	17	244	11	T1_GDINO07_barrels_p0_bs1_ts600_NMS_dis.csv	0.05	0.89	0.09	0.05	8	159	1
T1_GDINO07_tyres_p0_bs4_ts180_NMS_dis.csv	0.07	0.81	0.12	0.08	17	244	11	T1_GDINO07_barrels_p0_bs4_ts1480_NMS_dis.csv	0.05	0.89	0.09	0.05	8	159	1
T1_GDINO07_tyres_p0_bs4_ts600_NMS_dis.csv	0.07	0.81	0.12	0.08	17	244	11	T1_GDINO07_barrels_p0_bs4_ts180_NMS_dis.csv	0.05	0.89	0.09	0.05	8	159	1
T1_GDINO07_tyres_p37_bs4_ts600_NMS_dis.csv	0.06	0.95	0.11	0.06	16	246	3	T1_GDINO07_barrels_p0_bs4_ts600_NMS_dis.csv	0.05	0.89	0.09	0.05	8	159	1
T1_GDINO07_tyres_p63_bs4_ts1020_NMS_dis.csv	0.06	0.86	0.11	0.07	15	238	8	T1_GDINO07_barrels_p92_bs1_ts1480_NMS_dis.csv	0.05	0.89	0.09	0.05	8	159	1
T1_GDINO07_tyres_p63_bs1_ts1020_NMS_dis.csv	0.06	0.86	0.11	0.06	15	253	8	T1_TSAM08_barrels_default_NMS_dis.csv	0.05	1.00	0.09	0.05	9	181	0
T1_GDINO07_tyres_p11_bs1_ts180_NMS_dis.csv	0.06	0.97	0.10	0.06	16	273	2	T1_GDINO07_barrels_p11_bs1_ts180_NMS_dis.csv	0.04	0.78	0.08	0.06	7	150	2
T1_GDINO07_tyres_p11_bs4_ts180_NMS_dis.csv	0.06	0.97	0.10	0.06	16	274	2	T1_GDINO07_barrels_default_NMS_dis.csv	0.04	0.89	0.08	0.05	8	175	1
T1_GDINO07_tyres_p92_bs4_ts1480_NMS_dis.csv	0.06	0.69	0.10	0.08	14	237	18	T1_GDINO07_barrels_p11_bs4_ts180_NMS_dis.csv	0.04	0.78	0.08	0.06	7	152	2
T1_GDINO07_tyres_default_NMS_dis.csv	0.05	0.86	0.10	0.06	12	212	8	T1_GDINO07_barrels_p37_bs1_ts600_NMS_dis.csv	0.04	0.78	0.08	0.05	7	170	2
T1_TSAM08_tyres_p11_bs4_ts180_NMS_dis.csv	0.05	0.95	0.09	0.05	8	159	3	T1_GDINO07_barrels_p37_bs4_ts600_NMS_dis.csv	0.04	0.67	0.07	0.05	6	160	3
T1_GDINO07_tyres_p37_bs1_ts600_NMS_dis.csv	0.05	0.69	0.09	0.07	13	258	18	T1_GDINO07_barrels_p63_bs1_ts1020_NMS_dis.csv	0.03	0.67	0.07	0.05	6	166	3
T1_GDINO07_tyres_p92_bs1_ts1480_NMS_dis.csv	0.05	0.68	0.09	0.07	13	269	19	T1_TSAM08_barrels_p63_bs4_ts1020_NMS_dis.csv	0.03	0.67	0.06	0.05	6	171	3
T1_TSAM08_tyres_p63_bs4_ts1020_NMS_dis.csv	0.03	0.71	0.06	0.05	7	204	17	T1_TSAM08_barrels_p92_bs4_ts1480_NMS_dis.csv	0.03	0.56	0.05	0.05	5	186	4
T1_TSAM08_tyres_p37_bs4_ts600_NMS_dis.csv	0.03	0.86	0.06	0.04	6	177	8	T1_GDINO07_barrels_p63_bs4_ts1020_NMS_dis.csv	0.03	0.44	0.05	0.06	4	155	5
T1_TSAM08_tyres_p92_bs4_ts1480_NMS_dis.csv	0.03	0.61	0.06	0.05	7	216	23	T1_TSAM08_barrels_p37_bs4_ts600_NMS_dis.csv	0.02	0.33	0.04	0.06	3	137	6

Table A3. Evaluation of inferences in orthophoto "T2" (red = tyres, blue = barrels). Source name is composed of the orthoimage, model used (TSAM stands for TextSAM, GDINO stands for GroundingDINO), object analysed, padding (p), batch size (bs), and tile size (ts).

Inferences with Orthophoto "T2"														
Text Prompt: Tyres								Text Prompt: Barrels						
Source.Name	Precision	Recall	F1_ Score	AP	True_ Positive	False_ Positive	False_ Negative	Source.Name	Precision	Recall	F1_Score	AP	True_ Positive	False_ Positive	False_ Negative
T2_CLIPSEG07_ tyres_default_.csv	0.93	0.68	0.79	1.36	13	1	19	T2_CLIPSEG07_barrels_ default_.csv	0.18	0.67	0.29	0.27	6	27	3
T2_CLIPSEG07_ tyres_p0_bs1_ts352.csv	0.56	0.90	0.69	0.62	15	12	6	T2_CLIPSEG07_barrels_ p22_bs4_ts352.csv	0.12	1.00	0.22	0.12	9	65	0
T2_CLIPSEG07_ tyres_p0_bs4_ts352.csv	0.56	0.90	0.69	0.62	15	12	6	T2_CLIPSEG07_barrels_ p22_bs1_ts352.csv	0.12	1.00	0.21	0.12	9	67	0
T2_CLIPSEG07_tyres_ p22_bs1_ts352.csv	0.55	0.85	0.67	0.65	11	9	9	T2_CLIPSEG07_barrels_ p0_bs1_ts352.csv	0.08	0.89	0.15	0.09	8	88	1
T2_CLIPSEG07_tyres_ p22_bs4_ts352.csv	0.55	0.85	0.67	0.65	11	9	9	T2_CLIPSEG07_barrels_ p0_bs4_ts352.csv	0.08	0.89	0.15	0.09	8	88	1
T2_TSAM08_tyres_ p11_bs1_ts180_NMS_Dis.csv	0.19	0.73	0.30	0.26	29	122	16	T2_TSAM08_barrels_ p0_bs1_ts1020_NMS_Dis.csv	0.07	0.89	0.13	0.08	8	104	1
T2_TSAM08_tyres_ p63_bs1_ts1020_NMS_Dis.csv	0.19	0.77	0.30	0.25	29	125	14	T2_TSAM08_barrels_p0_bs1_ ts1480_NMS_Dis.csv	0.07	0.89	0.13	0.08	8	104	1
T2_GDINO07_tyres_ p63_bs4_ts1020_NMS_Dis.csv	0.17	0.97	0.28	0.17	28	141	2	T2_TSAM08_barrels_p0_bs1_ ts180_NMS_Dis.csv	0.07	0.89	0.13	0.08	8	104	1
T2_TSAM08_tyres_ p37_bs1_ts600_NMS_Dis.csv	0.17	0.73	0.28	0.24	31	147	16	T2_TSAM08_barrels_p0_bs1_ ts600_NMS_Dis.csv	0.07	0.89	0.13	0.08	8	104	1
T2_GDINO07_tyres_p63_bs1 _ts1020_NMS_Dis.csv	0.16	0.98	0.28	0.16	27	140	1	T2_GDINO07_barrels_p92_bs4_ ts1480_NMS_Dis.csv	0.07	1.00	0.13	0.07	9	120	0
T2_GDINO07_tyres_p37_bs4_ ts600_NMS_Dis.csv	0.16	0.87	0.27	0.18	29	152	8	T2_GDINO07_barrels_p0_bs1_ ts1020_NMS_Dis.csv	0.06	0.89	0.12	0.07	8	120	1
T2_TSAM08_tyres_p0_bs4_ ts1020_NMS_Dis.csv	0.16	0.75	0.27	0.22	28	143	15	T2_GDINO07_barrels_p0_bs1_ ts1480_NMS_Dis.csv	0.06	0.89	0.12	0.07	8	120	1
T2_TSAM08_tyres_p0_bs4_ ts1480_NMS_Dis.csv	0.16	0.75	0.27	0.22	28	143	15	T2_GDINO07_barrels_p0_bs1_ ts180_NMS_Dis.csv	0.06	0.89	0.12	0.07	8	120	1
T2_TSAM08_tyres_p0_bs4_ ts180_NMS_Dis.csv	0.16	0.75	0.27	0.22	28	143	15	T2_GDINO07_barrels_p0_bs1_ ts600_NMS_Dis.csv	0.06	0.89	0.12	0.07	8	120	1
T2_TSAM08_tyres_p0_bs4_ ts600_NMS_Dis.csv	0.16	0.75	0.27	0.22	28	143	15	T2_GDINO07_barrels_p0_bs4_ ts1480_NMS_Dis.csv	0.06	0.89	0.12	0.07	8	120	1
T2_TSAM08_tyres_p0_bs1_ ts1020_NMS_Dis.csv	0.17	0.52	0.26	0.33	21	101	29	T2_GDINO07_barrels_p0_bs4_ ts180_NMS_Dis.csv	0.06	0.89	0.12	0.07	8	120	1
T2_TSAM08_tyres_p0_bs1_ ts1480_NMS_Dis.csv	0.17	0.52	0.26	0.33	21	101	29	T2_GDINO07_barrels_p0_bs4_ ts600_NMS_Dis.csv	0.06	0.89	0.12	0.07	8	120	1
T2_TSAM08_tyres_p0_bs1_ ts180_NMS_Dis.csv	0.17	0.52	0.26	0.33	21	101	29	T2_GDINO07_barrels_p37_bs1_ ts600_NMS_Dis.csv	0.06	0.78	0.12	0.08	7	104	2
T2_TSAM08_tyres_p0_bs1_ ts600_NMS_Dis.csv	0.17	0.52	0.26	0.33	21	101	29	T2_TSAM08_barrels_p0_bs4_ ts1020_NMS_Dis.csv	0.06	1.00	0.11	0.06	9	143	0
T2_GDINO07_tyres_p0_bs1_ ts1020_NMS_Dis.csv	0.15	0.88	0.25	0.17	24	140	7	T2_TSAM08_barrels_p0_bs4_ ts1480_NMS_Dis.csv	0.06	1.00	0.11	0.06	9	143	0
T2_GDINO07_tyres_p0_bs1_ ts1480_NMS_Dis.csv	0.15	0.88	0.25	0.17	24	140	7	T2_TSAM08_barrels_p0_bs4_ ts180_NMS_Dis.csv	0.06	1.00	0.11	0.06	9	143	0
T2_GDINO07_tyres_p0_bs1_ ts180_NMS_Dis.csv	0.15	0.88	0.25	0.17	24	140	7	T2_TSAM08_barrels_p0_bs4_ ts600_NMS_Dis.csv	0.06	1.00	0.11	0.06	9	143	0

Table A3. *Cont.*

Inferences with Orthophoto "T2"															
Text Prompt: Tyres								**Text Prompt: Barrels**							
T2_GDINO07_tyres_p0_bs1_ts600_NMS_Dis.csv	0.15	0.88	0.25	0.17	24	140	7	T2_GDINO07_barrels_p0_bs4_ts1020_NMS_Dis.csv	0.06	1.00	0.11	0.06	9	143	0
T2_GDINO07_tyres_p0_bs4_ts1020_NMS_Dis.csv	0.15	0.88	0.25	0.17	24	140	7	T2_TSAM08_barrels_p11_bs1_ts180_NMS_Dis.csv	0.06	0.78	0.11	0.08	7	112	2
T2_GDINO07_tyres_p0_bs4_ts1480_NMS_Dis.csv	0.15	0.88	0.25	0.17	24	140	7	T2_GDINO07_barrels_p92_bs1_ts1480_NMS_Dis.csv	0.06	0.78	0.11	0.08	7	113	2
T2_GDINO07_tyres_p0_bs4_ts180_NMS_Dis.csv	0.15	0.88	0.25	0.17	24	140	7	T2_GDINO07_barrels_p63_bs4_ts1020_NMS_Dis.csv	0.05	0.89	0.10	0.06	8	139	1
T2_GDINO07_tyres_p0_bs4_ts600_NMS_Dis.csv	0.15	0.88	0.25	0.17	24	140	7	T2_TSAM08_barrels_p37_bs1_ts600_NMS_Dis.csv	0.06	0.67	0.10	0.08	6	102	3
T2_TSAM08_tyres_p92_bs1_ts1480_NMS_Dis.csv	0.14	0.80	0.25	0.18	30	177	12	T2_GDINO07_barrels_p37_bs4_ts600_NMS_Dis.csv	0.06	0.67	0.10	0.08	6	103	3
T2_GDINO07_tyres_p11_bs4_ts180_NMS_Dis.csv	0.14	0.87	0.25	0.16	28	168	8	T2_GDINO07_barrels_p11_bs4_ts180_NMS_Dis.csv	0.05	0.78	0.10	0.07	7	125	2
T2_GDINO07_tyres_p11_bs1_ts180_NMS_Dis.csv	0.14	0.87	0.24	0.16	28	169	8	T2_GDINO07_barrels_p63_bs1_ts1020_NMS_Dis.csv	0.05	0.78	0.09	0.06	7	133	2
T2_GDINO07_tyres_p92_bs4_ts1480_NMS_Dis.csv	0.14	0.77	0.23	0.18	25	156	14	T2_TSAM08_barrels_default_NMS_Dis.csv	0.05	1.00	0.09	0.05	9	178	0
T2_GDINO07_tyres_default_NMS_Dis.csv	0.12	1.00	0.22	0.12	31	226	0	T2_TSAM08_barrels_p63_bs1_ts1020_NMS_Dis.csv	0.05	0.78	0.09	0.06	7	137	2
T2_TSAM08_tyres_default_NMS_Dis.csv	0.12	0.82	0.21	0.15	31	227	11	T2_GDINO07_barrels_p11_bs1_ts180_NMS_Dis.csv	0.05	0.67	0.09	0.07	6	117	3
T2_GDINO07_tyres_p37_bs1_ts600_NMS_Dis.csv	0.11	0.57	0.18	0.19	20	161	26	T2_TSAM08_barrels_p11_bs4_ts180_NMS_Dis.csv	0.05	0.78	0.09	0.06	7	140	2
T2_TSAM08_tyres_p11_bs4_ts180_NMS_Dis.csv	0.09	0.83	0.17	0.11	16	156	10	T2_GDINO07_barrels_default_NMS_Dis.csv	0.05	0.89	0.09	0.05	8	166	1
T2_TSAM08_tyres_p37_bs4_ts600_NMS_Dis.csv	0.08	0.72	0.15	0.12	16	174	17	T2_TSAM08_barrels_p92_bs1_ts1480_NMS_Dis.csv	0.04	0.67	0.07	0.06	6	148	3
T2_GDINO07_tyres_p92_bs1_ts1480_NMS_Dis.csv	0.07	0.35	0.11	0.19	12	168	39	T2_TSAM08_barrels_p63_bs4_ts1020_NMS_Dis.csv	0.04	0.67	0.07	0.06	6	157	3
T2_TSAM08_tyres_p92_bs4_ts1480_NMS_Dis.csv	0.06	0.47	0.11	0.13	13	196	32	T2_TSAM08_barrels_p92_bs4_ts1480_NMS_Dis.csv	0.03	0.56	0.05	0.05	5	168	4
T2_TSAM08_tyres_p63_bs4_ts1020_NMS_Dis.csv	0.04	0.45	0.07	0.09	7	170	33	T2_TSAM08_barrels_p37_bs4_ts600_NMS_Dis.csv	0.01	0.22	0.03	0.07	2	136	7

Table A4. Evaluation of inferences in orthophoto "T3" (red = tyres, blue = barrels). Source name is composed of the orthoimage, model used (TSAM stands for TextSAM, GDINO stands for GroundingDINO), object analysed, padding (p), batch size (bs), and tile size (ts).

Inferences with Orthophoto "T3"															
Text Prompt: Tyres								Text Prompt: Barrels							
Source.Name	**Precision**	**Recall**	**F1_Score**	**AP**	**True_ Positive**	**False_ Positive**	**False_ Negative**	**Source.Name**	**Precision**	**Recall**	**F1_ Score**	**AP**	**True_ Positive**	**False_ Positive**	**False_ Negative**
T3_CLIPSEG07_tyres_ p22_bs1_ts352.csv	1.00	0.80	0.89	1.25	15	0	5	T3_GDINO07_barrels_p21_bs1_ ts340_NMS_Dis.csv	0.23	0.75	0.35	0.31	3	10	1
T3_CLIPSEG07_tyres_ p22_bs4_ts352.csv	1.00	0.80	0.89	1.25	15	0	5	T3_GDINO07_barrels_p12_bs1_ ts200_NMS_Dis.csv	0.21	0.75	0.33	0.29	3	11	1
T3_CLIPSEG07_tyres_ p0_bs1_ts352.csv	0.79	0.84	0.81	0.94	15	4	4	T3_TSAM08_barrels_p0_bs1_ ts200_NMS_Dis.csv	0.20	0.75	0.32	0.27	3	12	1
T3_CLIPSEG07_tyres_ p0_bs4_ts352.csv	0.79	0.84	0.81	0.94	15	4	4	T3_TSAM08_barrels_p0_bs1_ ts340_NMS_Dis.csv	0.20	0.75	0.32	0.27	3	12	1
T3_CLIPSEG07_tyres_ default_.csv	0.70	0.76	0.73	0.92	14	6	6	T3_TSAM08_barrels_p0_bs1_ ts495_NMS_Dis.csv	0.20	0.75	0.32	0.27	3	12	1
T3_GDINO07_tyres_ p0_bs1_ts200_NMS_Dis.csv	0.59	0.92	0.72	0.65	19	13	2	T3_TSAM08_barrels_p0_ bs1_ts60_NMS_Dis.csv	0.20	0.75	0.32	0.27	3	12	1
T3_GDINO07_tyres_ p0_bs1_ts340_NMS_Dis.csv	0.59	0.92	0.72	0.65	19	13	2	T3_GDINO07_barrels_p21_ bs4_ts340_NMS_Dis.csv	0.19	0.75	0.30	0.25	3	13	1
T3_GDINO07_tyres_ p0_bs1_ts495_NMS_Dis.csv	0.59	0.92	0.72	0.65	19	13	2	T3_GDINO07_barrels_p0_bs1_ ts200_NMS_Dis.csv	0.18	0.75	0.29	0.24	2	9	1
T3_GDINO07_tyres_p0_bs1_ ts60_NMS_Dis.csv	0.59	0.92	0.72	0.65	19	13	2	T3_GDINO07_barrels_p0_bs1_ ts340_NMS_Dis.csv	0.18	0.75	0.29	0.24	2	9	1
T3_GDINO07_tyres_p0_bs4_ ts200_NMS_Dis.csv	0.59	0.92	0.72	0.65	19	13	2	T3_GDINO07_barrels_p0_bs1_ ts495_NMS_Dis.csv	0.18	0.75	0.29	0.24	2	9	1
T3_GDINO07_tyres_p0_bs4_ ts340_NMS_Dis.csv	0.59	0.92	0.72	0.65	19	13	2	T3_GDINO07_barrels_p0_bs1_ ts60_NMS_Dis.csv	0.18	0.75	0.29	0.24	2	9	1
T3_GDINO07_tyres_p0_bs4_ ts495_NMS_Dis.csv	0.59	0.92	0.72	0.65	19	13	2	T3_GDINO07_barrels_p0_bs4_ ts200_NMS_Dis.csv	0.18	0.75	0.29	0.24	2	9	1
T3_GDINO07_tyres_p0_bs4_ ts60_NMS_Dis.csv	0.59	0.92	0.72	0.65	19	13	2	T3_GDINO07_barrels_p0_bs4_ ts495_NMS_Dis.csv	0.18	0.75	0.29	0.24	2	9	1
T3_TSAM08_tyres_p0_bs1_ ts200_NMS_Dis.csv	0.59	0.92	0.72	0.64	23	16	2	T3_GDINO07_barrels_p0_bs4_ ts60_NMS_Dis.csv	0.18	0.75	0.29	0.24	2	9	1
T3_TSAM08_tyres_p0_bs1_ ts340_NMS_Dis.csv	0.59	0.92	0.72	0.64	23	16	2	T3_GDINO07_barrels_p12_bs4_ ts200_NMS_Dis.csv	0.18	0.75	0.29	0.24	3	14	1
T3_TSAM08_tyres_p0_bs1_ ts495_NMS_Dis.csv	0.59	0.92	0.72	0.64	23	16	2	T3_GDINO07_barrels_p3_bs1_ ts60_NMS_Dis.csv	0.18	0.75	0.29	0.24	3	14	1
T3_TSAM08_tyres_p0_bs1_ ts60_NMS_Dis.csv	0.59	0.92	0.72	0.64	23	16	2	T3_GDINO07_barrels_p3_bs4_ ts60_NMS_Dis.csv	0.18	0.75	0.29	0.24	3	14	1
T3_GDINO07_tyres_p21_bs1_ ts340_NMS_Dis.csv	0.58	0.92	0.71	0.63	21	15	2	T3_TSAM08_barrels_p3_bs1_ ts60_NMS_Dis.csv	0.17	0.75	0.27	0.22	3	15	1
T3_GDINO07_tyres_p21_bs4_ ts340_NMS_Dis.csv	0.55	0.92	0.69	0.60	21	17	2	T3_TSAM08_barrels_p0_bs4_ ts200_NMS_Dis.csv	0.16	0.75	0.26	0.21	3	16	1

Table A4. *Cont.*

Inferences with Orthophoto "T3"															
Text Prompt: Tyres								**Text Prompt: Barrels**							
T3_GDINO07_tyres_p12_bs4_ts200_NMS_Dis.csv	0.55	0.92	0.69	0.60	17	14	2	T3_TSAM08_barrels_p0_bs4_ts340_NMS_Dis.csv	0.16	0.75	0.26	0.21	3	16	1
T3_TSAM08_tyres_p0_bs4_ts200_NMS_Dis.csv	0.53	0.92	0.68	0.58	23	20	2	T3_TSAM08_barrels_p0_bs4_ts495_NMS_Dis.csv	0.16	0.75	0.26	0.21	3	16	1
T3_TSAM08_tyres_p0_bs4_ts340_NMS_Dis.csv	0.53	0.92	0.68	0.58	23	20	2	T3_TSAM08_barrels_p0_bs4_ts60_NMS_Dis.csv	0.16	0.75	0.26	0.21	3	16	1
T3_TSAM08_tyres_p0_bs4_ts495_NMS_Dis.csv	0.53	0.92	0.68	0.58	23	20	2	T3_TSAM08_barrels_p30_bs1_ts495_NMS_Dis.csv	0.16	0.75	0.26	0.21	3	16	1
T3_TSAM08_tyres_p0_bs4_ts60_NMS_Dis.csv	0.53	0.92	0.68	0.58	23	20	2	T3_GDINO07_barrels_p0_bs4_ts340_NMS_Dis.csv	0.16	0.75	0.26	0.21	3	16	1
T3_GDINO07_tyres_p12_bs1_ts200_NMS_Dis.csv	0.53	0.92	0.67	0.58	17	15	2	T3_TSAM08_barrels_p12_bs1_ts200_NMS_Dis.csv	0.14	0.75	0.23	0.18	3	19	1
T3_GDINO07_tyres_default_NMS_Dis.csv	0.50	0.92	0.65	0.54	13	13	2	T3_TSAM08_barrels_p3_bs4_ts60_NMS_Dis.csv	0.13	0.75	0.22	0.17	3	20	1
T3_GDINO07_tyres_p3_bs1_ts60_NMS_Dis.csv	0.50	0.92	0.65	0.54	19	19	2	T3_GDINO07_barrels_p30_bs1_ts495_NMS_Dis.csv	0.13	0.75	0.21	0.17	2	14	1
T3_GDINO07_tyres_p3_bs4_ts60_NMS_Dis.csv	0.50	0.92	0.65	0.54	19	19	2	T3_GDINO07_barrels_p30_bs4_ts495_NMS_Dis.csv	0.13	0.75	0.21	0.17	2	14	1
T3_TSAM08_tyres_p3_bs4_ts60_NMS_Dis.csv	0.49	0.92	0.64	0.53	22	23	2	T3_TSAM08_barrels_p21_bs1_ts340_NMS_Dis.csv	0.12	0.75	0.21	0.16	3	22	1
T3_GDINO07_tyres_p30_bs1_ts495_NMS_Dis.csv	0.50	0.88	0.64	0.57	19	19	3	T3_GDINO07_barrels_default_NMS_Dis.csv	0.11	0.75	0.18	0.14	2	17	1
T3_GDINO07_tyres_p30_bs4_ts495_NMS_Dis.csv	0.50	0.88	0.64	0.57	19	19	3	T3_TSAM08_barrels_default_NMS_Dis.csv	0.09	0.75	0.17	0.13	3	29	1
T3_TSAM08_tyres_default_NMS_Dis.csv	0.48	0.92	0.63	0.52	21	23	2	T3_CLIPSEG07_barrels_p0_bs1_ts352.csv	0.10	0.50	0.16	0.19	2	19	2
T3_TSAM08_tyres_p12_bs1_ts200_NMS_Dis.csv	0.47	0.92	0.62	0.51	22	25	2	T3_CLIPSEG07_barrels_p0_bs4_ts352.csv	0.10	0.50	0.16	0.19	2	19	2
T3_TSAM08_tyres_p3_bs1_ts60_NMS_Dis.csv	0.47	0.92	0.62	0.51	22	25	2	T3_TSAM08_barrels_p21_bs4_ts340_NMS_Dis.csv	0.09	0.50	0.15	0.18	2	20	2
T3_TSAM08_tyres_p21_bs1_ts340_NMS_Dis.csv	0.45	0.92	0.61	0.49	23	28	2	T3_CLIPSEG07_barrels_p22_bs1_ts352.csv	0.07	0.50	0.13	0.15	2	25	2
T3_TSAM08_tyres_p30_bs1_ts495_NMS_Dis.csv	0.46	0.88	0.60	0.52	22	26	3	T3_CLIPSEG07_barrels_p22_bs4_ts352.csv	0.07	0.50	0.13	0.15	2	25	2
T3_TSAM08_tyres_p12_bs4_ts200_NMS_Dis.csv	0.18	0.44	0.25	0.40	7	33	14	T3_TSAM08_barrels_p30_bs4_ts495_NMS_Dis.csv	0.07	0.75	0.13	0.09	2	27	1
T3_TSAM08_tyres_p21_bs4_ts340_NMS_Dis.csv	0.19	0.28	0.23	0.68	7	30	18	T3_TSAM08_barrels_p12_bs4_ts200_NMS_Dis.csv	0.03	0.25	0.06	0.14	1	28	3
T3_TSAM08_tyres_p30_bs4_ts495_NMS_Dis.csv	0.14	0.20	0.17	0.71	5	30	20	T3_CLIPSEG07_barrels_default_.csv	0.00	0.00	0.00	0.00	0	22	4

References

1. Ichinose, D.; Yamamoto, M. On the Relationship between the Provision of Waste Management Service and Illegal Dumping. *Resour. Energy Econ.* **2011**, *33*, 79–93. [CrossRef]
2. Porta, D.; Milani, S.; Lazzarino, A.I.; Perucci, C.A.; Forastiere, F. Systematic Review of Epidemiological Studies on Health Effects Associated with Management of Solid Waste. *Environ. Health* **2009**, *8*, 60. [CrossRef] [PubMed]
3. Kjeldsen, P.; Barlaz, M.A.; Rooker, A.P.; Baun, A.; Ledin, A.; Christensen, T.H. Present and Long-Term Composition of MSW Landfill Leachate: A Review. *Crit. Rev. Environ. Sci. Technol.* **2002**, *32*, 297–336. [CrossRef]
4. Bansal, K.; Tripathi, A.K. WasteNet: A Novel Multi-Scale Attention-Based U-Net Architecture for Waste Detection in UAV Images. *Remote Sens. Appl. Soc. Environ.* **2024**, *35*, 101220. [CrossRef]
5. Berra, E.F.; Peppa, M.V. Advances and Challenges of UAV SFM MVS Photogrammetry and Remote Sensing: Short Review. In Proceedings of the 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS), Santiago, Chile, 22–26 March 2020; pp. 533–538.
6. Diara, F.; Roggero, M. Quality Assessment of DJI Zenmuse L1 and P1 LiDAR and Photogrammetric Systems: Metric and Statistics Analysis with the Integration of Trimble SX10 Data. *Geomatics* **2022**, *2*, 254–281. [CrossRef]
7. Colomina, I.; Molina, P. Unmanned Aerial Systems for Photogrammetry and Remote Sensing: A Review. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 79–97. [CrossRef]
8. Mittal, P.; Singh, R.; Sharma, A. Deep Learning-Based Object Detection in Low-Altitude UAV Datasets: A Survey. *Image Vis. Comput.* **2020**, *104*, 104046. [CrossRef]
9. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155. [CrossRef]
10. Yaseen, M. What Is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector. *arXiv* **2024**, arXiv:2409.07813.
11. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
12. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
13. Eitel, A.; Springenberg, J.T.; Spinello, L.; Riedmiller, M.; Burgard, W. Multimodal Deep Learning for Robust RGB-D Object Recognition. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 681–687.
14. Chen, K.; Jiang, X.; Wang, H.; Yan, C.; Gao, Y.; Tang, X.; Hu, Y.; Xie, W. OV-DAR: Open-Vocabulary Object Detection and Attributes Recognition. *Int. J. Comput. Vis.* **2024**, *132*, 5387–5409. [CrossRef]
15. Li, S.; Cao, J.; Ye, P.; Ding, Y.; Tu, C.; Chen, T. ClipSAM: CLIP and SAM Collaboration for Zero-Shot Anomaly Segmentation. *Neurocomputing* **2025**, *618*, 129122. [CrossRef]
16. Ren, T.; Jiang, Q.; Liu, S.; Zeng, Z.; Liu, W.; Gao, H.; Huang, H.; Ma, Z.; Jiang, X.; Chen, Y.; et al. Grounding DINO 1.5: Advance the "Edge" of Open-Set Object Detection. *arXiv* **2024**, arXiv:2405.10300. [CrossRef]
17. Lüddecke, T.; Ecker, A.S. Image Segmentation Using Text and Image Prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
18. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual Event, 8–24 July 2021.
19. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In Proceedings of the Computer Vision—ECCV 2024, Milan, Italy, 29 September–4 October 2024.
20. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014.
21. Mumuni, F.; Mumuni, A. Segment Anything Model for Automated Image Data Annotation: Empirical Studies Using Text Prompts from Grounding DINO. *arXiv* **2024**, arXiv:2406.19057. [CrossRef]
22. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4015–4026.

23. Spiegler, P.; Koleilat, T.; Harirpoush, A.; Miller, C.S.; Rivaz, H.; Kersten-Oertel, M.; Xiao, Y. TextSAM-EUS: Text Prompt Learning for SAM to Accurately Segment Pancreatic Tumor in Endoscopic Ultrasound. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Honolulu, HI, USA, 19–23 October 2025.
24. Goodwin, M.; Halvorsen, K.T.; Jiao, L.; Knausgård, K.M.; Martin, A.H.; Moyano, M.; Oomen, R.A.; Rasmussen, J.H.; Sørdalen, T.K.; Thorbjørnsen, S.H. Unlocking the Potential of Deep Learning for Marine Ecology: Overview, Applications, and Outlook. *ICES J. Mar. Sci.* **2022**, *79*, 319–336. [CrossRef]