Analysis of using Zero Shot Open Vocabulary Detection Methods for Plastic Waste Classification

Madhini B
Dept. of Networking and Communications
School of Computing
SRM Institute of Science and Technology
Chengalpattu, India
mb7992@srmist.edu.in

Supraja P
Dept. of Networking and Communications
School of Computing
SRM Institute of Science and Technology
Chengalpattu, India
suprajap@srmist.edu.in

2024 5th International Conference on Communication, Computing & Industry 6.0 (C2I6) | 979-8-3503-7988-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/C2I663243.2024.10894834

***Abstract*—Pollution from plastic garbage has grown to be a major environmental issue, requiring the creation of effective techniques for identifying and categorizing it. Supervised machine learning is a proven technique for automatic waste detection and classification but requires heavy training. Unsupervised learning techniques such as transfer learning and few-shot learning (FSL) have greatly reduced the training time and the need for huge datasets. Vision Language Models (VLMs) with open vocabulary and zero-shot detection techniques have reduced the shortcomings of traditional models. This paper analyzes the scope of using zero-shot combined with open-vocabulary detection techniques such as YOLO World, Grounding DINO, ViLD, and OWL-ViT for plastic waste detection. Zero-shot open vocabulary models eliminate the need for training custom data for long hours, enabling models to classify new classes with only their semantic encoders and contrastive pre-training.**

Keywords—Plastic waste detection, classification, supervised learning, unsupervised learning, convolutional neural networks, zero-shot detection, contrastive pre-training, Grounding DINO, ViLD, OWL-ViT, YOLO World, Visual Language Model (VLM)

I. INTRODUCTION

As per the Central Pollution Control Board (CPCB) 2020-21 annual report, India generates 4,126,997 tonnes per annum (TPA) of plastic waste. The primary disposal method is by landfilling, which is inefficient, expensive, and pollutes the natural environment, affecting the quality of life of the people who stay near the dump site [31]. Sensitizing the people with the waste classification techniques helps better manage these dumps. The imperative need for recycling waste helps minimize the quantity of landfills, leading to a sustainable planet. Effective methods for identifying and classifying plastic garbage are needed due to its growing presence in our ecosystems.

Automated sorting methods have been a phenomenal stage of development and an alternative to manual sorting methods that demand a huge amount of time and human labor. Utilizing deep learning to classify trash proves to have a positive effect on both the environment and the economy by making waste processing plants more efficient [19]. Machines driven with robotic arms and sensors deployed with supervised learning algorithms such as K-Nearest Neighbour (K-NN), Support Vector Machine (SVM), Random Forest, and CNNs are revolutionizing the waste management sector in developed countries. However, these supervised learning machine learning models need hours of training on huge datasets.

Then with the advent of transfer and self-supervised learning, the training hours were significantly reduced and required few known classes for training the model. SSD MobileNetV2, Siamese Networks, and triplet loss perform well as a few-shot learning model in waste identifications. The few-shot or single-shot learning techniques need few or single-seen classes to train the model. However, when unseen classes are encountered, the model struggles to identify and needs to be retrained and also lacks accuracy. The primary challenge is the availability of training data. Scaling up the number of classes present in the pre-trained model dataset demands significant costs and effort.

To overcome the shortcomings of the previous models, the zero-shot learning (ZSL) technique emerges as a new methodology to predict novel (unseen classes) with no prior training. ZSL models learn to recognize objects or concepts that are unseen and eliminate the cost involved in annotations and data imbalance. Accurate prediction of real-world scenario object detection has been facilitated by the usage of semantic information. Open Word Language (OWL), along with ZSL, emerges as a good choice of algorithm in detecting unknown waste categories, as the concept of waste is subjective, and hence redefining the class category is not needed. Real-world applications are far from ideal, where images contain multiple objects and can appear in any part of the image. Zero-shot detection (ZSD) is proposed to simultaneously localize and recognize unseen objects belonging to novel categories. Open-vocabulary object detection, which detects objects described by arbitrary text inputs, resolves this problem.

This paper compares and contrasts the usage of ZSD algorithms such as ViLD, YOLO WORLD, OWL-VIT, and Grounding DiNO for efficient plastic waste detection. These models have low latency and quick response times.

The rest of this study is organized as follows. The related works on plastic waste classification utilizing computer vision are surveyed in Section 2. In section 3, this study analyzes the best-suited zero detection model for plastic waste classification, such as Grounding DINO, YOLO WORLD, etc. In Section 4, the methodology adopted for assessment of the

979-8-3503-7988-4/24/$31.00 ©2024 IEEE

vision language model's performances are discussed. Lastly, the results of this study and future works identified are analyzed and proposed in Section 5.

II. RELATED WORKS

For the past few decades, owing to the tremendous growth seen in economic development, demographic growth, and urbanization, the magnitude of waste generated has increased exponentially [6]. The insurmountable wastes seen at landfills become the foundational cause of waterborne and airborne diseases in their neighborhood. Human beings should be sensitized to follow the 3R principle of Reduce, Reuse, and Recycle. To facilitate effective recycling, source segregation of waste material is needed. Secondary source segregation can be hastened with the help of automated waste sorting machines. Waste identification and classification are machine-driven and are trained with deep learning algorithms [23].

The use of plastics in the packaging industry presents a huge opportunity for the respective stakeholders to shift to a circular economy from a linear model [37].

The field of automatic waste segregation has seen significant advancements over the past few years, with researchers identifying various techniques to improve the accuracy and effectiveness of plastic waste classification. Starting from the traditional methods, such as SVN and Random Forest to Convolutional Neural Networks, the scope of accuracy in detection and speed has seen a steady improvement. However, these methods involve training on large datasets, annotations, fine- tuning of hyperparameters, etc. Therefore, the model becomes bulky and cumbersome as the cost of retraining the model is huge.

A recent addition to the world of object detection algorithms that has gained attention is the use of zero-shot detection for plastic waste classification. Zero-shot detection refers to the ability to recognize and classify objects, eliminating the requirement of labeled training data. This approach is particularly relevant in the aspects of plastic waste classification, where the diversity of plastic types and the presence of impurities can challenge the accuracy of traditional supervised learning techniques.

In this research paper, we conduct a comparative analysis of the performance of a few popular zero-shot detection methods that can be more suitable for plastic waste classification by drawing insights from the existing literature.

In paper [1], the SSD MobileNet model was used, which needed 24,000 training steps with a completion time of 9 hours and a loss value of approximately 0.2711. However, real-time image classification requires minimal processing power consumption. An ample amount of training data with high-performance hardware becomes the conditional criteria for this process to be effective.

The primary requirement for this process might be that one has to have adequate data for training and fast computational hardware.

In paper [2], the author proposes using one-shot learning to generalize over the complete dataset, where the data requirement for certain classes is insufficient for training CNN. If not, this will lead to an imbalance of data. The similarities between objects of the same classes and the differences between objects of different classes are learned directly by one-shot learning methods.

In paper [3], ResNet18 achieved 87.8% accuracy with pre-processing, resizing of images, and augmentation with fine-tuning. The accuracy was achieved with heavy training.

The same ResNet has been used to design a lightweight neural network trash classification system that achieves an accuracy of 96.10% on TrashNet, but it also struggles to identify multi-label garbage classification [4].

In 2019, [19] developed an innovative waste classification model based on ResNet called DNN-TC that classified waste into organic, inorganic, and medical wastes with an accuracy of 98% and 94%, respectively.

An automatic waste sorting multilayer hybrid deep-learning system was introduced by [38] to sort waste present in urban public areas. The system used a CNN-based algorithm for feature extraction and multilayer perceptron (MLP) for feature consolidation. The Trashnet Dataset released by Yang and Thung [39] was used to evaluate the model. However, when many objects were present in the image, the model failed to identify them and needs improvement.

[20] experimented with images of large and small resolution (225 × 264 and 80 × 45) on a 5-layer bespoke CNN classifier and concluded that the lighter model with small image resolution trained with less training time when compared to the large resolution .

In the paper, [36] a convolutional neural network using a custom model for image classification using MobileNet was built to identify recyclable, compostable, as well as landfill materials. The model, in addition, classifies food waste as compostable but requires a lot of training data and pre-training.

OrgalidWaste, a dataset of 5600 images with four classes--organic waste class and three solid waste classes (glass, metal, and plastic) was used for waste classification and tested upon several CNN architectures such as CNN, VGG16, VGG19, Inception-V3, and ResNet50. VGG16 showed an accuracy of 88.42% [21].

The paper [9] proposed a multiple-layered CNN Inception-v3 model for waste classification. Plastics, glass, paper, card-board, metal, and others are the six different types of classified waste. Online-trained datasets achieved an accuracy of 92.5%.

[14] proposed benchmark approaches for classifying litter with EfficientNet-B0 for demographic-based classification, reaching an accuracy of 74% to 84%.

A. Mitra used a classifier to detect wastes into 6 different types (plastic, trash, glass, metal, paper, and cardboard) present in multiple objects in a single image by using Faster R- CNN algorithm [10].

[11] Accuracy surpassing human prediction is the future hope of the world of object detection using artificial intelligence. Identifying the tiniest difference is yet to be conquered

by these existing detectors, whereas they are efficient in localizing large and medium-sized objects.

A 2015 report states that 97% of the total consumption in the packaging industry sector produces 141 million metric tons of garbage waste. The plastic garbage collected at the waste segregation sites is mostly contaminated and hence lands up in incineration centers or landfills. However, if the plastic garbage collected from the source is mostly cleaner, it can be recycled or repurposed. [12] Recycling waste is useful for the environment as well as for the economy [8].

The rise of smart cities as a global model with the goal of sustainable development has tried to incorporate networking, data management advancement, and computing to implement well-organized waste management systems that raise the quality of life [6].

The wastes recovered from such places will be of mixed varieties, and it is extremely difficult to manually sort and identify the different types of plastics. Hence the need for using automated sorting techniques with artificial intelligent machines or models. After the initial model training with the relevant dataset, the plastic wastes can be sorted into the respective categories with the help of CNN [12]. However, these models need large amounts of training data and are not available at times. By incorporating transfer learning, the training of neural networks can be fastened using a limited number of input images. Transfer learning also improves the classification of multiple types of waste types in a dataset [12].

In [5], the transfer learning technique on seven pre-trained models such as DenseNet201, InceptionResNetV2, Xception, VGG19, ResNet50, InceptionV3, and MobileNet has been utilized for waste classification among seven different classes (cardboard, glass, metal, organic, paper, plastic, and trash). MobileNet with 93.82% validation accuracy reached 93.82%, 89%, and 86% for MobileNet, Xception, and InceptionResNetV2, respectively.

For solving inter-domain learning problems, transfer learning [28] extracts useful information from data in a particular field. Overfitting is solved by applying transfer learning [27]. Object detection is much more efficient and accurate when pre-processing is done using SVM-like YOLOv3 algorithms and at the same time reduces memory utilization [22]. In [24], all layers in MobileNetV1 and MobileNetV2 are unfrozen. After unfreezing, the size reduction is obtained by a Global Average Pooling (GAP). Due to the presence of imbalanced data for training the classes , pre-processing is adapted.

In paper [26], an accuracy of 87.2% is achieved from 2527 images. The categories used were trash, plastic, paper, metal, and cardboard with a MobileNet model. The trash images used were in .jpg extension, but the validation accuracy needs to be improved.

In paper [17]Supervised learning methods have proven their mark in various application fields and shown remarkable progress over the past years. However, their capacity to identify unseen classes goes down rapidly due to the unavailability of suitable labeled data. To identify images seen in real-world applications, the data has to be annotated precisely, which involves a lot of human effort, specifically related to semantic segmentation and analogy learning.

Few-shot learning (FSL) [32] tackles the problem of training the classifier in the case of insufficient labeled data. FSL adapts the famous method of meta-learning that uses a substantial number of jobs similar to the target and applies it to obtain a good initial model value. This approach significantly reduces the training time as the training data is considerably small. When novel labeled data is encountered by the trained classifier, the model uses the techniques of additive [33] and reiterative learning [34] methods to handle the tasks. In some applications, most classes are without labeled data. It is an imperative need for the classifier to identify unseen class objects post-training.

III. METHODOLOGY

A. Zero-Shot Models

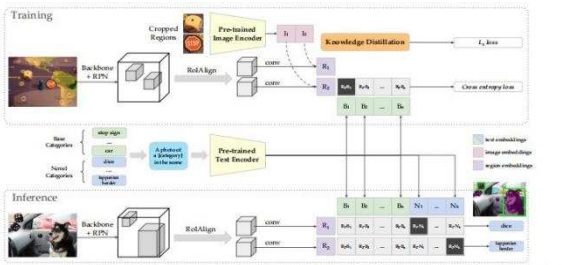
When a model does object detection in images without any previous knowledge of the classes or prior training, the methodology is called zero-shot object detection. During the training phase, the correspondence between text and image is learned by the zero-shot object detection model. The provision of textual descriptions of an object helps the model detect novel images that have never been seen before.

This work will explore popular model architectures of open-set computer vision models and analyze the model's prediction on images containing multiple waste objects. The model's performance will determine which vision language model can be used to automate plastic waste detection effectively.

B. Open-Vocabulary Zero-Shot Models

Open-vocabulary detection models can detect novel class categories using text prompts; the model is named an open-set object detection model from unbounded vocabulary at inference. Possible models that are taken into consideration in this work are 1) ViLD (Xiuye Gu et al., 2021), 2) OWL-ViT (Minderer et al., 2022), 3) GDINO (Shilong Liu et al.), 4) YOLO World [18]

1) Open-vocabulary Object Detection via Vision and Language Knowledge Distillation -- ViLD: ViLD employs a pre-trained model (teacher) and a two-stage detector with Mask R-CNN (student) acting as the backbone. The method aligns the region embeddings associated with the text and the image. Text embeddings play the role of the classifier (teacher) in the detection module. Text embeddings are obtained by inputting category names into the pre-trained text encoder. The next step is to reduce the distance between a region embedding and an image embedding. During the inference stage, the novel category text embeddings are included in the classifier for zero-shot detection. The CLIP model is used as a pre-trained open-vocabulary classification model with an input size



An overview of using ViLD for open-vocabulary object detection. ViLD distills the knowledge from a pretrained open-vocabulary image classification model. First, the category text embeddings and the image embeddings of cropped object proposals are computed, using the text and image encoders in the pretrained classification model. Then, ViLD employs the text embeddings as the region classifier (ViLD-text) and minimizes the distance between the region embedding and the image embedding for each proposal (ViLD-image). During inference, text embeddings of novel categories are used to enable open-vocabulary detection.

Fig. 1. ViLD - Vision Language with Language and Knowledge Distillation - Architecture overview

Source: https://arxiv.org/abs/2104.13921

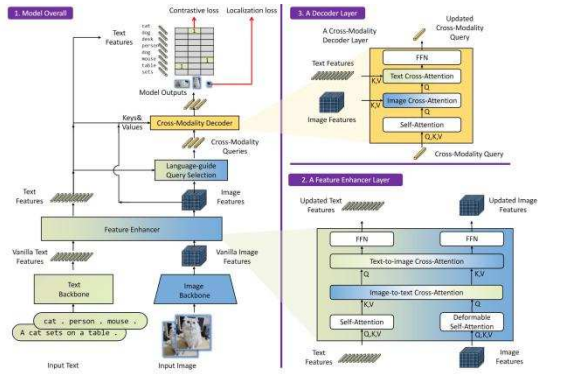


Figure 3. The framework of Grounding DINO. We present the overall framework, a feature enhancer layer, and a decoder layer in block 1, block 2, and block 3, respectively.

Fig. 3. Grounding DINO

Source: https://arxiv.org/abs/2303.05499

of 224x224. ViLD model is trained from scratch for 180,000 iterations shown in "Fig.1" . ResNet with Mask RCNN is used as the backbone, with an input image of 1024x1024, with 0.32 as initial learning rate, and batch size of 256.

2) OWL-ViT: OWL-ViT is a text-conditioned, zero-shot state-of-the-art (SOTA) object detection model that enables querying images with single or multiple texts with no prior training. OWL-ViT uses ViT (a transformer-based) architecture as the backbone as shown in "Fig. 2". A BERT pre-trained text embedding module and a detection head are added to the backbone architecture to condition the object detection task. The detection loss and text embedding loss are combined, and the model is trained with a multi-task loss function to achieve great success in computer vision tasks.

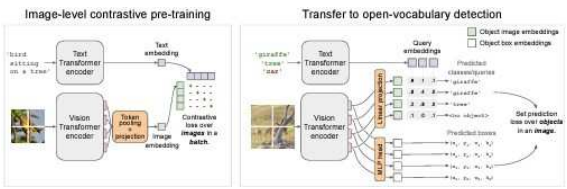


Fig. 2. OWL -ViT

Source: https://arxiv.org/abs/2205.06230

The same set of image-text datasets and loss values are used to pre-train the text encoder and image contrastively. Both the encoders are trained from scratch with a contrastive loss and random initialization. The publicly available pre-trained CLIP model, along with multi-head attention pooling (MAP), is used for image representation. The end-of-sequence (EOS) tokens are used for text representation.

3) Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection: Grounding DINO is a fusion between the transformer-based open-set object detector DINO and grounded pre-training model as shown in "Fig. 3". Grounding DINO takes manual inputs such as expressions, classes, or class names and predicts random objects. Concept generalization is achieved in Grounding DINO by passing language as an input to a closed set detector. The closed-set detector is divided into three phases to achieve strong fusion solution between language and vision modalities. Evaluation of attribute-object pairs is also added in addition to open-set object detection on novel class objects. Grounding DINO performance was evaluated on benchmarks like COCO, LVIS, ODinW, and RefCOCO/+/g with 52.5 AP on COCO and 26.1 AP on ODinW. By incorporating large-scale pre-training, the phrase grounding task is being redefined by the Grounding DINO 1.5 version. The combination of pseudo-labeled grounding data with self-training enhances the model's readiness for open-world settings.

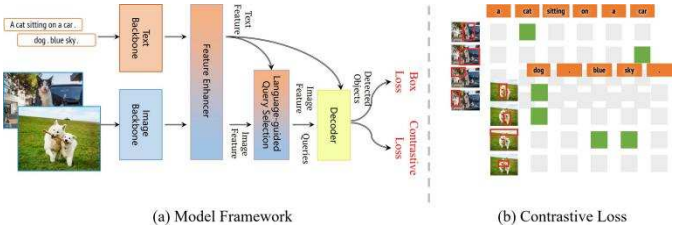


Fig. 4. Grounding DINO 1.5 - Edge/Pro

Source: https://arxiv.org/pdf/2405.10300

Grounding DINO 1.5 version has a larger, robust model architecture with an advanced backbone. The inference speed and capability to handle real-world scenarios are better than its predecessor's. Version 1.5 has two variants, namely. Pro and Edge. Grounding DINO 1.5 Edge model was developed to deploy on the edge devices to facilitate applications such as medical image processing, unmanned vehicle driving, etc. The cost of computation present in an open-set detection model and the restricted resources present on the edge devices is a large gap to be addressed. To combat this, the traditional feature enhancer has been replaced with an EfficientViT-L1, which has multi-scale feature-enhancing capability. The model was deployed on NVIDIA Orin NX platform, with an inference speed more than 10 FPS at an input size of 640 × 640.

4) YOLO World: YOLO-World is the work of [18]. The authors introduced a new version of the YOLO series that improves upon the prior works by adding the possibility of using open vocabulary as textual input and can detect novel categories not included in the closed-set siblings of the YOLO series. Moreover, the zero-shot object detector YOLO-World enables the fusion of text embeddings with vi- sion image embeddings. The novel component that the authors introduced, called the Re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN), can match text input information with regions of interest that are obtained from the input image, enabling a multi-modal approach to object detection.

Fine-tuning increases the performance concerning speed and accuracy of segmentation and object detection of YOLO-World on the LVIS Dataset with a score of 35.4 AP and 52.0 FPS.

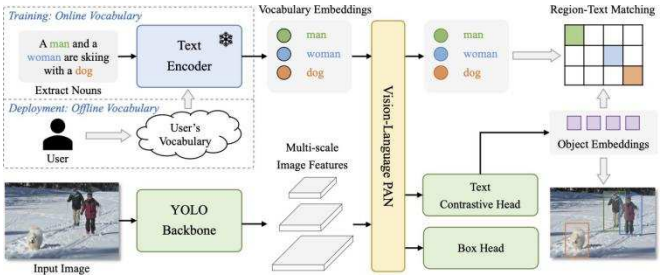


Fig. 5. YOLO World Architecture
Source:https://arxiv.org/abs/2401.17270

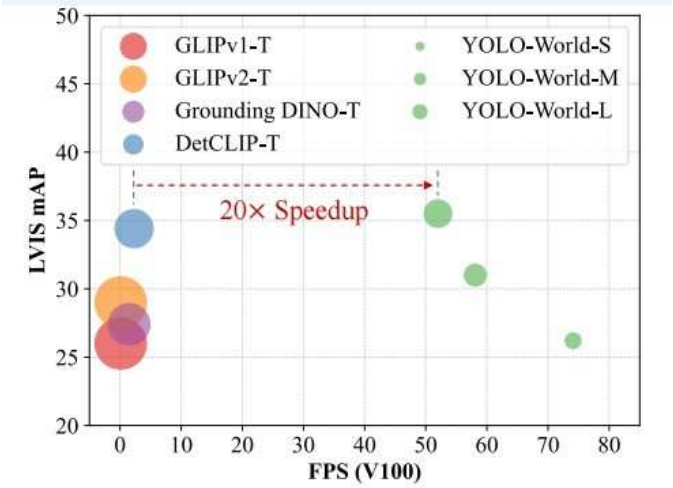


Fig. 6. YOLO world performance comparison with other VLMs
Source: https://arxiv.org/abs/2401.17270

This approach can open up opportunities to use more diverse datasets and expanded datasets to help accelerate object detection development. In "Fig. 5", the architecture of the model can be seen.

Regarding the architecture of the YOLO-World model, it has three key parts: • Multi-scale features of the input image are extracted by the YOLO detector.

- CLIP text encoder that encodes the text into text embeddings.
- Cross-modality that is a multi-level fusion of image and text embeddings performed by the RepVL-PAN custom network

Reasons for YOLO-World's real-time detection speed: • The backbone that this model uses compared to other open-vocabulary object detectors (e.g. GroundingDINO, which uses a computationally heavier DINO backbone, is a faster and lighter Convolutional Neural Network (CNN) based on Dark-net architecture, which highly accounts for its higher inference speed. • At the inference stage, instead of encoding the user's prompt in real-time, CLIP is being used in YOLO World to transform text input (generated while prompting the model) into offline vocabulary embeddings that are cached and re-used, thus circumventing the necessity for real-time text encoding. Overall, YOLO World can make open vocabulary object detection quicker, cost-effective, and commonly available. While maintaining roughly the same precision compared to its predecessors, it is 20 times faster than other architectures in the same category. Last but not least, YOLO World makes it easier to deploy and can be integrated with other architectures for further enhancing its object detection capabilities (e.g., EfficientSAM).

C. Pre-processing

The Vision Language pre-training phase aims to pre-train by learning from image-text correlation and predicting visual recognition tasks. ViLD uses a process called distillation to extract the knowledge from its image classification model. Image and text embeddings are extracted and computed using the encoders present in the pre-trained classification model. Vision Transformer architecture (ViT) uses a technique called contrastive pre-training and applies it to a large image-text dataset. The final token present in the pooling layer has been removed, and a classification light-weight box head is added to each output token of the transformer model. Gounding20M which is pre-trained on large-scale publicly available 20M images. YOLO World encodes the input texts using the pre-trained CLIP encoder and connects the text and image features for better semantic-visual representation, incorporating the Re- parameterizable Vision-Language Path Aggregation Network (RepVL-PAN).

D. Datasets

The datasets to be considered are the waste that is discarded by humans and industries as part of their livelihood. The waste detection models have utilized traditional machine learning models such as SVN, Random Forests, etc. CNN-based mod-els such as ResNet50 [8], DenseNet, MobileNetV2, Incep-tion V3 model[9], SSD Mobilenet [1], Faster R-CNN [10], EfficentDetD2[11], EfficientNetB0-B3 and Region Proposal Network. The waste classification done by the models caters to organic, recyclables (cardboard, metal, paper, plastic, glass), non-recyclables, and others. Plastic waste is also classified based on resin code and categorized into seven types, namely PETE, PE-HD, PVC, PE-LD, PP, PS, and others

TABLE I. COMPILED VERSION OF WASTE DATASETS

Name	Size	Classes	Year
OpenLitterMap	≥100K	≥100	2018
Waste Pictures	~ 24K	34 classes	2019
TrashNet	≥ 2100	6 classes	2016
Extended TACO	1500	60	2020
Wade AI	~ 400	1	2016
UAV Waste	772	1 rubbish	2016
Trash ICRA	7668	7	2021
TrashCan(UnderWater)	7212	16	2020
Drinking Waste	4800	4	2020
MJU Waste	2475	1	2020
Places365	10 million	434	2018
TACO	1500	28	2020

TABLE II PERFORMANCE METRICS OF ZERO-SHOT OPEN VOCABULARY MODELS ON VARIOUS DATASETS

Model	DataSet	AP
ViLD	PASCAL VOC	72.2
ViLD	COCO	36.6
ViLD	Object365	11.8
OWL-ViT	LVIS	23.3
OWL-ViT	Object365	30.6
GDINO 1.5 Pro	COCO	54.3
GDINO 1.5 Edge + TensorRT	LVIS minval	36.2
YOLO world	LVIS	35.4

[12]. But using open-vocabulary object detection, the waste classification vocabulary paradigm can be changed according to the secondary segregation nomenclature. Waste datasets are available in abundance, ranging from private to public Table I. Benchmarking on waste datasets has been done by a few researchers [40] but we need to have a machine learning model that can adapt to changing waste classification types. Table II Table III refers to the performance of zero-shot open-vocabulary model on various datasets

An object is considered a waste based on its need and place of use. A fixed classification type will not be useful for segregating waste objects. Hence, with the help of open-vocabulary object detection, flexibility in class categorization can be added. E.g., plastic boxes that are black get segregated separately when compared to their other color counterparts because the black-colored plastic box is pre-recycled and has a lower market value. Using open-vocabulary object detection, the class category can be changed by using the text prompt as required.

Fig. 7. Predicted Image - Black Plastic cup PT1 (Plastic Type 2)

Fig. 8. Predicted Image - White Plastic cup PT1 (Plastic Type 1

In this example, black-colored plastics are classified as Plastic Type 1 or PT1, and white-colored plastics are classified as Plastic Type 2 or PT2 (name categorization taken for this particular example). Refer to "Fig. 7" and "Fig. 8" where the ViLD model classifies the plastic boxes based on their color. This is achieved by adding the required class category using the textual prompt of the model.

The performance of the model's outcome is determined by four parameters, namely True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

The model's capacity to rightly categorize the occurrences of classes among the positive samples is determined by the term **Precision** and the formula used as shown in Eq. (1).

$$\frac{TP}{TP + FP} \tag{1}$$

The **Recall** value helps to measure the machine learning model's capacity to detect the number of positive samples present in the test or train environment and the formula used as shown in Eq (2).

$$\frac{TP}{TP + FN} \tag{2}$$

When the precision and recall value of any model's performance are plotted, a curve is obtained, and the area under the curve gives the **Average Precision (AP)**, and it displays the trade-off between precision and recall.

Accuracy of OwL-ViT, GDINO, ViLD, and YOLO World are derived by the formula as shown in Eq. (3).

$$\frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3}$$

E. Normalization

To achieve an equal data distribution, data normalization is the fundamental step to achieve. Normalization helps in faster convergence. When the standard deviation is used to divide the difference of each pixel's mean, we get normalization. When normalization distribution data is plotted, a Gaussian curve is achieved with the center as zero.

F. Fine Tuning

Feature extraction is followed by another technique called fine-tuning that involves freezing the last layer and building a custom layer which is used for classification. ViLD does not add fine-tuning. From end to end, OWL-ViT uses bipartite matching loss on standard detection datasets as the process

TABLE III SUMMARY OF THE DATASETS TRAINED FOR ZERO-SHOT VLMS

Models	Year	Dataset trained	Availability
ViLD	2021	LVIS	Private
OWL Vit	2022	LVIS	Private
GDINO - Swin-T	2023	Object365,GoldG,Cap4M	Private
GDINO - Swin-B	2023	COCO,GoldG,O365,Cap4M,ODinW-35	Private
YOLO World	2024	Objects365 GQA Flickr30K CC3M	Public

of fine-tuning. In YOLO World, 80 epochs are used for fine-tuning with the AdamW optimizer. On the COCO dataset, the RepVL-PAN is removed to increase acceleration. After fine-tuning, the performance of YOLO-World has shown good improvement on the standard LVIS dataset.

IV. IMPLEMENTATION

For this analysis, GPU T4 and Google Colab were used for the initial experimental setup. Real-time and internet-sourced mixed plastic waste images were used for testing various VLMs by altering the text prompts. For example, a yellow plastic cup, blue clip, black spoon, etc. The GitHub demo models of ViLD, OWL-ViT, GDiNO, and YOLOWorld were tested for their prediction capabilities. The performance accuracy (3) of the VLM models under study was observed: ViLD model predicted accurately, but the model's performance was slower than the other models. The OWL-ViT prediction was better than the rest but less than the ViLD model. Grounding DINO and YOLO World prediction performance were on similar lines when compared. The advantage of using GDINO and YOLO World for waste classification is their lightweight model architecture with edge deployment adaptability.

V. RESULTS AND DISCUSSION

The results from open vocabulary and re-parameterized zero-shot object detection models were analyzed in detail. Single images with different types of plastic waste images were used for testing, and the text prompt was added with textual categories relevant to the plastic waste. Due to the leverage of adding text prompts per the input requirement, the novel waste categories can be identified correctly based on the classification categories needed at the secondary segregation facilities. A spoon can be textually prompted as a **plastic** spoon, a cup as a **white** plastic cup, or a cardboard box as a **brown** cardboard box. Refer "Fig. 9 – 13"

VI. CHALLENGES AND FUTURE DIRECTION

Zero-shot open vocabulary detection is a boon for the object detection world since it greatly decreases the requirement of huge amounts of labeled data and seen classes. The challenge with the zero-shot open vocabulary detection methodology is the model size, which can be solved by using suitable optimization techniques. Future research can be in the field of deploying these models on edge devices.

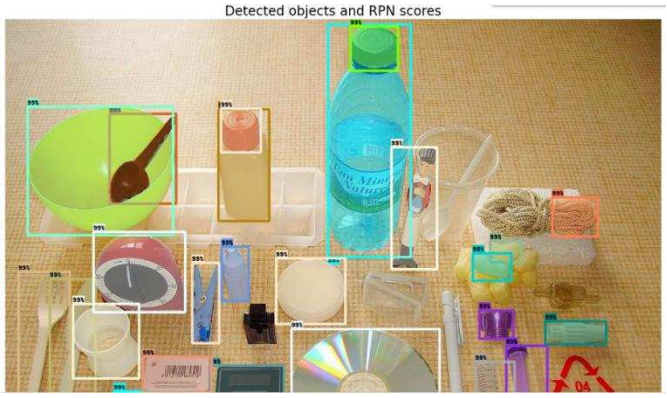


Fig. 9. ViLD - Waste Object Prediction - Summary

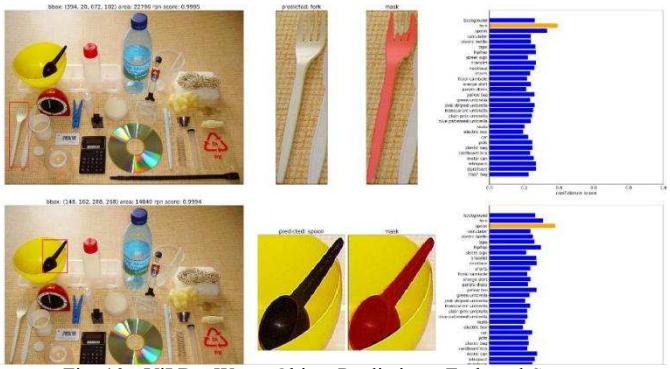


Fig. 10. ViLD - Waste Object Prediction - Fork and Spoon

VII. CONCLUSION

This research paper analyzes the scope of integrating a zero-shot object detection model for the classification of plastic waste. ViLD model's accuracy outperforms other models in prediction, but it is comparatively slower than GDINO and YOLO world because of the ResNet backbone. The accuracy of YOLO World and GDINO is greatly increased by fine-tuning and adjusting the confidence level. Refer TableIV Because of their lightweight nature, GDINO, and YOLO World prove to be a good choice for edge deployment of zero-shot open vocabulary detection models.

TABLE IV. PREDICTION ACCURACY OF VLM MODELS

Models	Accuracy
ViLD	0.92
OWL ViT	0.76
GDINO	0.96
YOLO World	0.86

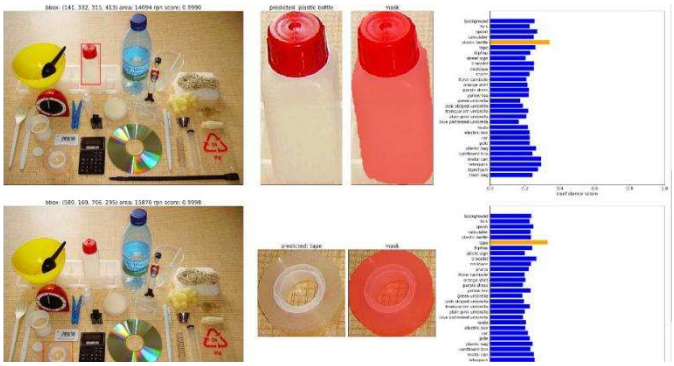


Fig. 11. ViLD - Waste Object Prediction - Plastic cup

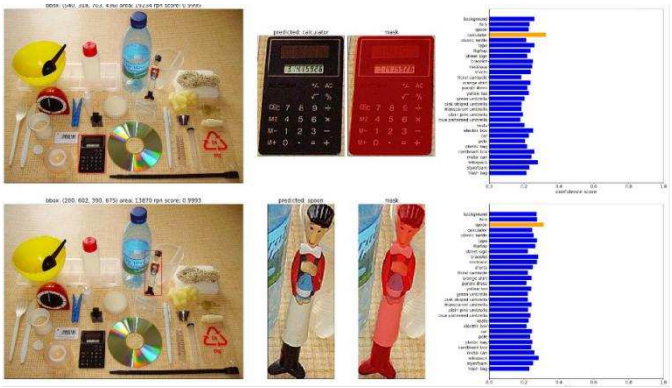


Fig. 12. ViLD - Waste Object Prediction - Calculator and Spoon

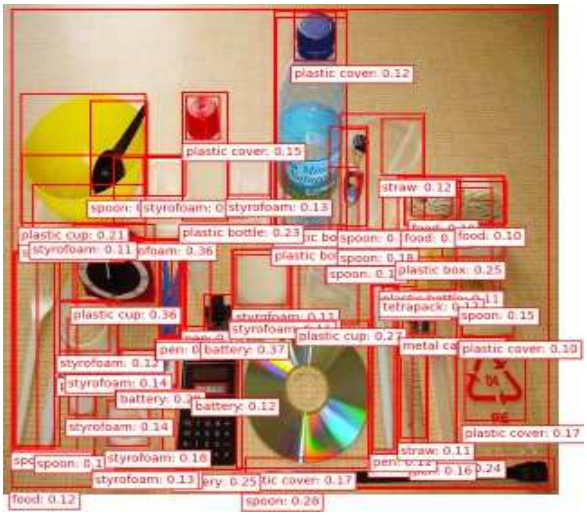


Fig. 13. OWL Vit - Predicted Image

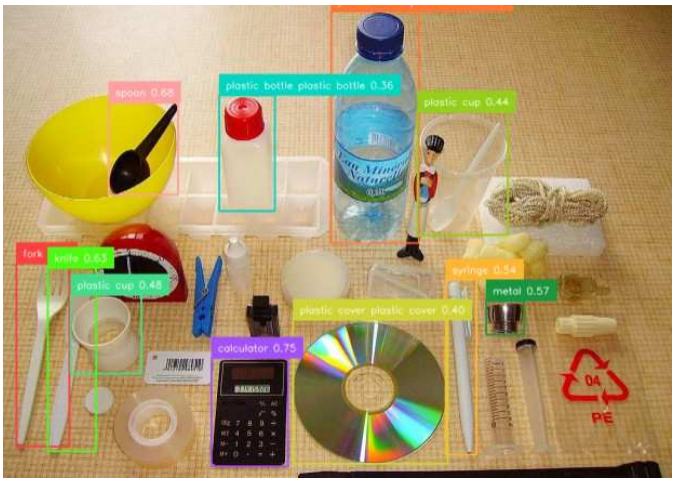


Fig. 14. Grounding DINO - Predicted Image

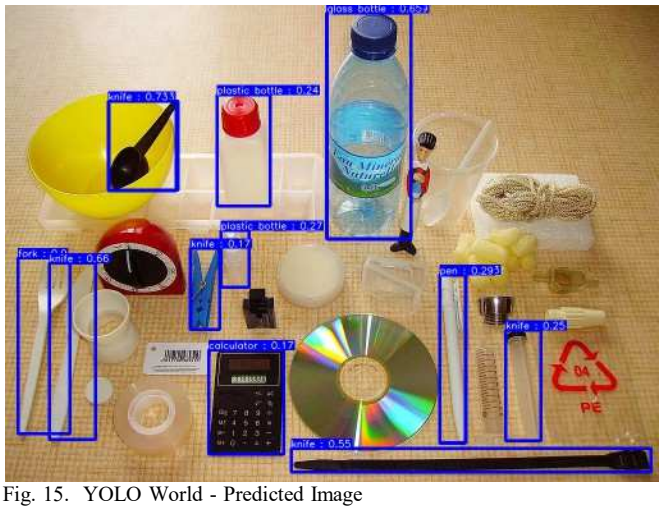


Fig. 15. YOLO World - Predicted Image

REFERENCES

[1] Thokrairak, K. Thibuy, and P. Jitngernmadan, " Valuable Waste Classification Modeling based on SSD-MobileNet," Oct. 2020,doi:10.1109/incit50588.2020.9310928.

[2] S. Agarwal, R. Gudi, and P. Saxena, "One-Shot learning based classification for segregation of plastic waste," Nov. 2020, doi: 10.1109/dicta51227.2020.9363374.

[3] D. Gyawali, A. Regmi, A. Shakya, A. Gautam, S. Shrestha, "Comparative Analysis of Multiple Deep CNN Models for Waste Classification," Apr. 2020, doi.org/10.48550/arXiv.2004.02168

[4] Z. Yang and D. Li, "WasNet: A Neural Network-Based Garbage Collection Management System," IEEE Access, vol. 8, pp. 103984–103993, Jan. 2020, doi: 10.1109/access.2020.2999678.

[5] S. Poudel and P. Poudyal, "Classification of Waste Materials using CNN Based on Transfer Learning," Dec. 2022, doi:10.1145/3574318.3574345

[6] Fang, B., Yu, J., Chen, Z. et al. Artificial intelligence for waste management in smart cities: a review. Environ Chem Lett 21, 1959–1989 (2023).https://doi.org/10.1007/s10311-023-01604-3

[7] N. J. Sinthiya, T. A. Chowdhury, and A. K. M. B. Haque, "Artificial Intelligence Based Smart Waste Management—A Systematic Review," in Green energy and technology, 2022, pp. 67–92. doi: 10.1007/978-3-030-96429-0

[8] A. Sevinc¸ and F. Ozyurt, "Classification of recyclable waste using deep learning architectures," Firat University Journal of Experimental and Computational Engineering, vol. 1, no. 3, pp. 122–128, Jan. 2022, doi: 10.5505/fujece.2022.83997.

[9] F. A. Azis, H. Suhaimi, and E. Abas, "Waste Classification using Convo-

lutional Neural Network," Aug. 2020, doi: 10.1145/3417473.3417474.
[10] A. Mitra,"Detection of Waste Materials Using Deep Learning and Image Processing," California State University San Marcos 2020
[11] S. Majchrowska et al., "Deep learning-based waste detection in natural and urban environments," Waste Management, vol. 138, pp. 274–284, Feb. 2022, doi: 10.1016/j.wasman.2021.12.001.
[12] A. A. P. Chazhoor, E. S. L. Ho, B. Gao, and W. L. Woo, "Deep transfer learning benchmark for plastic waste classification," Intelligence & Robotics, Jan. 2022, doi: 10.20517/ir.2021.15.
[13] J. Bobulski and M. Kubanek, "Project of Sorting System for Plastic Garbage in Sorting Plant Based on Artificial Intelligence," Jul. 2020, doi: 10.5121/csit.2020.100903.
[14] M. Malik et al., "Waste Classification for Sustainable Development Using Image Recognition with Deep Learning Neural Network Models," Sustainability, vol. 14, no. 12, p. 7222, Jun. 2022, doi: 10.3390/su14127222.
[15] D. O. Melinte, A.-M. Travediu, and D. N. Dumitriu, "Deep Convolutional Neural Networks Object Detector for Real-Time Waste Identification," Applied Sciences, vol. 10, no. 20, p. 7301, Oct. 2020, doi: 10.3390/app10207301.
[16] H. Abu-Qdais, N. Shatnawi, and E. Al-Alamie, "Intelligent system for solid waste classification using combination of image processing and machine learning models," Journal of Experimental and Theoretical Artificial Intelligence, pp. 1–12, Feb. 2024, doi: 10.1080/0952813x.2024.2323043.
[17] C. Tan, X. Xu, and F. Shen, "A Survey Of zero shot detection: Methods and applications," Cognitive Robotics, vol. 1, pp. 159–167, Jan. 2021, doi: 10.1016/j.cogr.2021.08.001.
[18] T. Cheng , L. Song1, Y. Ge1, W. Liu3 , X.Wang, Y. Shan ,"YOLO-World: Real-Time Open-Vocabulary Object Detection," Feb. 2024. [Online]. Available: https://arXiv:2401.17270v3
[19] A. H. Vo, L. H. Son, M. T. Vo, and T. Le, "A Novel Framework for Trash Classification Using Deep Transfer Learning," IEEE Access, vol. 7, pp. 178631–178639, Jan. 2019, doi: 10.1109/access.2019.2959033.
[20] N. Nnamoko, J. Barrowclough, and J. Procter, "Solid Waste Image Classification Using Deep Convolutional Neural Network," Infrastructures, vol. 7, no. 4, p. 47, Mar. 2022, doi: 10.3390/infrastructures7040047.
[21] G. Rishma and R. Aarthi, "Classification of Waste Objects Using Deep Convolutional Neural Networks," in Lecture notes in electrical engineering, 2021, pp. 533–542.
[22] S. K. Behera, A. Barathwaj SR, V. L, S. G, and H. N. C, "AI Based Waste classifier with Thermo-Rapid Composting," 2020.
[23] O. Adedeji and Z. Wang, "Intelligent Waste Classification System Using Deep Learning Convolutional Neural Network," Procedia Manufacturing, vol. 35, pp. 607–612, Jan. 2019, doi: 10.1016/j.promfg.2019.05.086.
[24] I. F. Nurahmadan, R. M. Arjuna, H. D. Prasetyo, P. A. Hogantara, I. N. Isnainiyah, and R. Wirawan, "A Mobile Based Waste Classification Using MobileNets-V1 Architecture," Oct. 2021, doi: 10.1109/icimcis53775.2021.9699161.
[25] M. Yang and G. Thung, "Classification of trash for recyclability status," Mach. Learn., Stanford, CA, USA, Project Rep. CS229, 2016.
[26] E. J. Rabano, Stephenn L.; Cabatuan, Melvin K.; Sybingco, Edwin; Dadios, Elmer P.; Calilung, "Common Garbage Classification Using MobileNet," 2018 IEEE 10th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Control. Environ. Manag., pp. 1–4, 2018.
[27] R. Faria, F. Ahmed, A. Das, and A. Dey, "Classification of Organic and Solid Waste Using Deep Convolutional Neural Networks," Sep. 2021, doi: 10.1109/r10-htc53172.2021.9641560.
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision.
[29] Serezhkin, "Drinking waste classification,https://www.kaggle.com/arkadiyhacks/drinking-waste-classification, accessed July 5, 2021.
[30] S. Sekar, "Waste classification data," https://www.kaggle.com/techsash/ waste-classification-data, accessed July 5, 2021.
[31] N. Ramsurrun, G. Suddul, S. Armoogum, and R. Foogooa, "Recyclable Waste Classification Using Computer Vision And Deep Learning," May 2021, doi: 10.1109/zinc52049.2021.9499291.
[32] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, ICLR, 2017.
[33] G. Fei, S. Wang, B. Liu, Learning cumulatively to become more knowl- edgeable, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[34] Z. Chen, B. Liu, Lifelong machine learning, Synthesis Lectures on Artificial Intelligence and Machine Learning, 2016.
[35] P. Sane and R. Agrawal, "Pixel normalization from numeric data as input to neural networks: For machine learning and image processing," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2017, pp. 2221-2225.
[36] S. Frost, B. Tor, R. Agrawal, and A. G. Forbes, "CompostNet: An Image Classifier for Meal Waste," IEEE Xplore, Oct. 2019, doi: 10.1109/ghtc46095.2019.9033130. Available: https://doi.org/10.1109/ghtc46095.2019.9033130
[37] J. Backstrom and N. Kumar, "Advancing the Circular Economy of Plastics through eCommerce," DSpace, Jun. 2021, Available: https://dspace.mit.edu/handle/1721.1/130968
[38] Y. Chu, C. Huang, X. Xie, B. Tan, S. Kamal, and X. Xiong, "Multilayer Hybrid Deep-Learning Method for Waste Classification and Recycling," Computational Intelligence and Neuroscience, vol. 2018, pp. 1–9, Nov. 2018, doi: 10.1155/2018/5060857. Available: https://doi.org/10.1155/2018/5060857
[39] M. Yang and G. Thung, "Classification of trash for recyclability status", 2016.
[40] AS Charisis,"Solid waste detection in context: Merging Datasets and using Open-Set Vision Models", 2024, https://fse.studenttheses.ub.rug.nl/id/eprint/33537.