Enhancing Robotic Vision through Vision-Language Models: A Trash Sorting Framework with Physically Grounded Understanding

Joe Lisk
liski003@umn.edu

Prakrit Sinha
sinha239@umn.edu

Sam Komonosky
komon002@umn.edu

Abstract

This work aimed to perform zero-shot trash detection using generated image captions as class labels. We utilized a pipeline consisting of BLIP-2, a vision-language model (VLM) for image captioning, and OWL-ViT, an object detection model built on the vision transformer architecture (ViT). We fine-tuned the BLIP-2 model on the TrashNet dataset and ran inference on the TACO dataset. Due to computational bottlenecks, we weren't able to perform a complete analysis on a test set. Qualitative analysis revealed the architecture lacked robust image captioning and object detection, but the results showed promise that an improved model could perform zero-shot trash detection.

1. Introduction

In recent years, there has been an increased interest in deep-learning models that combine vision and language modalities, known as vision-language models (VLMs) [6]. Previously, most visual recognition models relied heavily on crowd-labelled data, and usually trained a single deep-neural network (DNN) for each single visual recognition task - a laborious and time-consuming process [29]. In contrast, VLMs are capable of learning vision-language correlation from image-text pairs widely available on the Internet, and enabling zero-shot predictions on various visual recognition tasks with a single model.

While VLMs have seen a surge in popularity in recent years, relatively little work has been done on their application to robotics [12]. In addition, most current VLMs are limited in their understanding of the physical concepts of common objects, which restricts their usefulness for robotic manipulation tasks that involve interaction and physical reasoning about such objects [7].

One possible application of VLMs and robotics is the detection and sorting of trash [28]. Relying on manpower to detect and classify domestic waste is highly inefficient and time-consuming, which makes it a good candidate for robotic applications. There are relatively few datasets relating to trash detection, and most have a small amount of data. The definition of trash is ambiguous, and highly dependent on scene context. Objects considered trash can also have a large, diverse number of possible class labels.

This work explores the applications of VLMs for the purpose of detecting trash. In this project, we train a VLM on images of trash to generate captions. Those captions are then parsed and fed as labels into a zero-shot object detection model to attempt to identify the trash objects in the images.

2. Related Work

2.1. Vision-Language Models (VLMs)

The integration of vision and language modalities, commonly known as Vision-Language Models (VLMs), has emerged as a focal point in contemporary research [29]. Traditionally, visual recognition models have heavily relied on labor-intensive processes, where a single deep-neural network (DNN) is trained for each specific visual recognition task using crowd-labeled data. This approach is not only time-consuming but also limits the adaptability of models to diverse visual tasks.

In contrast, VLMs present a paradigm shift by capitalizing on the wealth of image-text pairs available on the Internet. Leveraging these pairs, VLMs can learn intricate correlations between vision and language, unlocking the potential for zero-shot predictions across a spectrum of visual recognition tasks using a unified model. This versatility stands in stark contrast to the conventional approach, offering a more efficient and scalable solution for complex visual understanding.

[22] introduced CLIP, a VLM that predicts image-text pairings using contrastive pretraining and a classifier using text as label data. The authors note that CLIP is able to perform zero-shot classification on a wide range on vision-related tasks. Using text data for classification labels

may be useful in trash classification and detection due to the large number of potential classes.

Despite the surge in popularity and success in various domains, there exists a noteworthy research gap concerning the application of VLMs to robotics [12]. Most existing VLMs grapple with limitations in comprehending the physical concepts associated with everyday objects, thus impeding their efficacy in tasks requiring robotic manipulation and nuanced physical reasoning [7]. The present study aims to address this gap by tailoring and deploying a VLM for a specific and practical application – sorting trash from non-trash in cluttered scenes. By doing so, it endeavors to extend the capabilities of VLMs to the realm of robotics, unlocking their potential for real-world applications and enhancing their understanding of the physical world.

Another open issue with VLMs is fine-tuning them for specific tasks. In order to use VLMs for trash detection, VLM fine-tuning methods will need to be further developed and tested. [27] observed that many VLMs are deployed as black-box models, which limits fine-tuning methods due to restricted access to the VLM's parameters. The authors of [27] proposed a black-box method, Collaborative Fine-Tuning (CraFT), for fine-tuning VLMs, where the prompt was optimized with the CMA-ES algorithm and the model prediction was further optimized with a MLP with AdamW. The authors reported that CraFT was able to outperform baseline methods such as zero-shot CLIP. Fine-tuning is relevant to trash detection due to large number of potential classes for different types of trash.

2.2. Physically Grounded Vision Language Models (PGVLMs)

Recent advancements in physically grounded vision-language models (PGVLM) [7] have addressed some limitations in existing trash detection datasets. This model not only classifies objects in cluttered scenes but also captures physical characteristics, such as deformability. Utilizing images from the Ego objects dataset by Meta [30] and annotating them with their PhysObjects dataset, the PGVLM introduces an innovative approach. However, it currently lacks a mechanism to classify objects as trash or non-trash beyond relying on characteristics like deformability, which may be unreliable for distinguishing between the two categories [7].

2.3. Trash Detection

2.3.1 TACO Dataset

The TACO (Trash Annotations in Context) dataset [21] serves as a representative dataset for trash detection, offering detailed annotations of trash objects. While proficient at detecting multiple parts of an object, such as the plastic lid and paper cup of a coffee cup (Figure 1), the TACO dataset currently focuses on distinguishing only between trash objects and the background. Notably, it does not provide a distinction between trash and non-trash objects within the same scene.

2.3.2 COCO Dataset

The COCO (Common Objects in Context) dataset [16] has been a cornerstone in object detection research, providing a comprehensive benchmark for evaluating object recognition algorithms. Mask R-CNN (Region-based Convolutional Neural Network) [8], an extension of the Faster R-CNN architecture, has shown exceptional performance in instance segmentation tasks. An example image from the COCO dataset is shown in Figure 2. While valuable for general object detection, neither COCO nor Mask R-CNN specifically address the challenges associated with trash detection and physical reasoning in cluttered scenes.

2.3.3 TrashNet

The TrashNet dataset[25] contains around 2527 images of trash objects classified into paper, glass, plastic, metal, carton and garbage subclasses. The images on this dataset consist of photographs of garbage taken on a simple, white background. An example image is shown in Figure 3. TrashNet has been used in other work, notably by Aral et. al.[1] who trained and tested several deep learning models on the dataset for the purposes of waste identification.

2.3.4 YOLO System

The YOLO (You Only Look Once) series, with its real-time object detection capabilities, has been a significant contribution to the field [11]. YOLO v8, the latest iteration, continues to excel in object detection tasks. However, similar to COCO and Mask R-CNN, YOLO v8 may not inherently handle the nuances of trash detection and physical characteristics crucial for sorting tasks.

2.4. Deep Learning Methods

Some research has been done on the application of deep learning to trash detection. Li et. al.[13] proposed a multi-model cascaded convolutional neural network (MCCNN) for domestic waste image detection and classification, and created their own dataset consisting of 30 000 domestic waste multilabeled images with 52 categories. In other work, Fulton et. al.[5] evaluated several deep learning algorithms for the purpose of detecting plastic debris in marine environments.

One issue with trash detection datasets is their relatively small size. To overcome this limitation, research has been done on the use of methods such as generative-adversarial networks (GANs)[4] and two-stage variational autoencodera (VAEs) and binary classifiers[9] to generate synthetic images of trash to supplement existing datasets.

2.5. Vision Transformers

More recently, transformer architectures have been studied in computer vision tasks. [3] introduced the Vision Transformer (ViT) architecture, which feeds image patches with positional embeddings through a transformer architecture. The authors note that when pretrained on a sufficiently large dataset, ViT can perform image recognition with excellent results compared to CNN architectures. Due to the large number of labels in trash classification, a pretrained ViT could be an ideal trash classifier.

3. Baseline Method

A representative dataset for trash detection is the TACO dataset [21]. One of the strengths of the TACO dataset is its annotation of trash objects. In some instances, it can detect multiple parts of an object, i.e. the plastic lid and paper cup of a coffee cup. However, a current limitation of the TACO dataset is that it only differentiates trash objects from the background. TACO doesn't differentiate between trash and non-trash objects in the same scene.

Initially, we intended to use the physical characteristics of the detected objects from the PG-Instruct BLIP model to assess whether or not they were trash. Due to issues with fine-tuning the PG-Instruct BLIP model on a custom dataset (see Discussion), we opted instead to perform zero-shot trash detection. While this changed the structure of our research problem somewhat, it is still an important problem to solve in trash detection. This is mainly due to the multitude of different kinds of trash that would make it difficult to train a detector on all possible labels.

To test the zero-shot capabilities of trash detectors, we decided to use images from the TACO dataset while using the labels from the TrashNet dataset. For the baseline method, we chose the OWL-ViT model for zero-shot object detection. We used the 6 labels of the TrashNet dataset (paper, plastic, metal, glass, cardboard, trash) and images from the TACO dataset as input to the OWL-ViT model without any training or fine-tuning. This allowed us to establish a baseline for the effectiveness of zero-shot trash detection.

4. Proposed Method

The Vision-Language Model selected for this project was BLIP-2. BLIP-2 leverages frozen pre-trained image encoders and large language models by training a lightweight, 12-layer Transformer encoder in between them, achieving state-of-the-art performance on various vision-language tasks, despite having significantly fewer trainable parameters than existing methods[14].

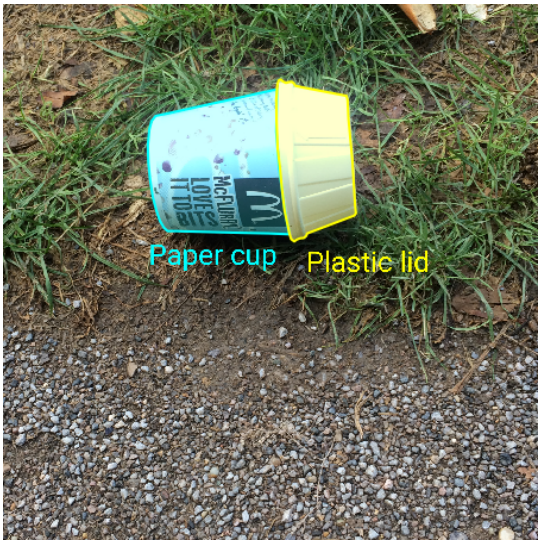

Figure 1. Sample image from the TACO dataset

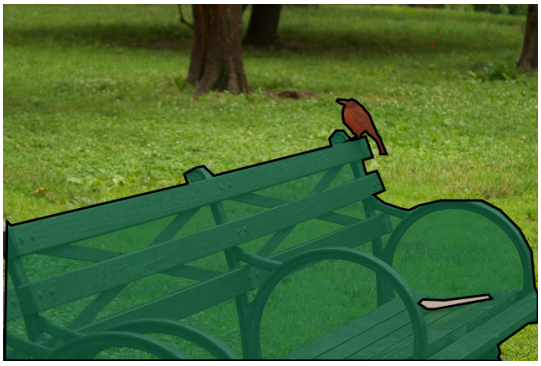

Figure 2. Sample image from the COCO dataset

Figure 3. Sample image from the TrashNet dataset

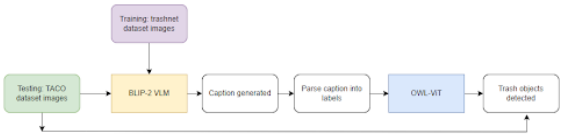

Figure 4. Block diagram of proposed architecture

The TrashNet dataset[25] was used for training the BLIP-2 model. TrashNet was selected due to the images being mostly "clean" and showing only trash objects. The numeric ids of the dataset images were mapped to text categories (0:cardboard, 1:glass, 2:metal, 3:paper, 4:plastic, 5:trash) in order to generate captions for each image using BLIP-2. The BLIP-2 model was then trained using images from the TACO dataset.

A caption was generated for each image, which was then parsed to match the categories mentioned above. As an example, a generated caption for an image of a styrofoam/paper cup is "*cardboard cup with acardboard cup*", while the parsed caption is simply "*cardboard*". It was decided that the ideal use for the parsed captions would be to use them to detect trash objects in the images. For this, OWL-ViT, a zero-shot text-conditioned object detection model, was chosen. OWL-ViT is an open-vocabulary object detection network trained on a variety of (image, text) pairs. It can be used to query an image with one or multiple text queries to search for and detect target objects described in text, and can attain very strong performance on zero-shot text-conditioned and one-shot image-conditioned object detection[19].

Figure 4 shows a block diagram of the proposed method. The parsed captions generated by the images from the TACO dataset were fed as labels into OWL-ViT, along with the respective image. Bounding boxes, labels, and confidence scores were then created by OWL-ViT and drawn on the output images, as shown in Figure 6 and Figure 7.

For the evaluation metric, IoU for bounding boxes generated by the baseline method and the proposed method are compared against the ground truth bounding boxes present in the TACO dataset. Additionally, a mapping of TACO class labels to TrashNet labels was performed to test the classification accuracy of the detected trash.

5. Results

Due to computing resource and time constraints (see Discussion), we were only able to perform an analysis on a small subset of images. For example, in Figure 8 and Figure 9, the baseline method was able to detect all instances

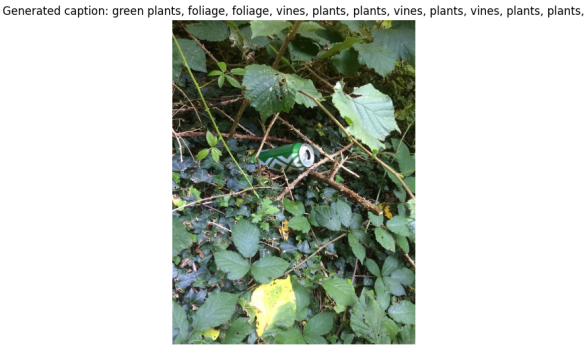

Figure 5. Sample image from BLIP-2 captioning (failure)

of trash, while the proposed method missed a small piece of trash. The labels of the proposed method were all "cardboard", while the baseline method labels were varied. The bounding boxes generated from the proposed method were less noisy than the baseline method. In another image, Figure 5 ,the trash object was neither detected by the baseline or the proposed method. As previously mentioned, the captions were sometimes unusable even after cleaning. Figure 10 only produced a caption about a red table, with none of the bottles accounted for.

5.1. Discussion

During the research phase of the project, the fine-tuning of the VLM PG-Instruct BLIP on custom datasets proved to be intractable due to the scarcity in documentation and tutorials available. This was overcome by using the VLM BLIP2, the precursor of PG-Instruct BLIP, which is a used ubiquitously in the domain of vision-language tasks.

Subsequently, we also realised that there was a lot of ambiguity in our previously chosen COCO dataset (for fine-tuning the model), as a clear classification if the object in the images were trash or not trash was not present. This was handled by choosing the TrashNet dataset, for the primary reasons being that all the objects in the images in this dataset were trash objects, and there were well defined and documented categories to which the trash objects belonged to.

Next, after the model was fine-tuned, we faced several caption generation issues. Firstly, sometimes the generated caption words not always usable as a set of class labels for the zero-shot model (Ex. *"can can of soda can with green and white background"*). Then, the generated captions were inconsistent for a single image over multiple runs. To rectify these problems, we matched the words in the generated caption with classes in the unseen TACO dataset to use as labels for OWL-ViT. However, a more robust way could have been to use further natural language processing

methods for the refinement of the generated captions. [17][18]

One of the challenges faced with the dual BLIP-2/OWL-ViT architecture was the time complexity. [24] notes that one of the bottlenecks of transformers is their quadratic time complexity. This was reflected in a 4 minute inference time for 6 classes in the baseline method. This limited the amount of images our method could feasibly test on. Additionally, the bounding boxes obtained from OWL-ViT were incorrect dimensions. As an example, the results obtained in Figure 8 experienced a leftward shift in all detected bounding boxes.

The object detection performed somewhat better than the BLIP-2 captioning method. When OWL-ViT was provided with a sufficiently accurate caption, it was able to perform successful object detection. A caption that failed to capture any of the trash classes led to incorrect trash detection and in some cases, no object detection. This is possibly due to BLIP-2 being trained on the TrashNet dataset while being evaluated on the TACO dataset. The images in Trash Net contain only one piece of trash per image with little clutter and in non-noisy environments. Whereas the TACO dataset images can contain multiple instances of trash per image with noisy environments. It is possible that training BLIP-2 on the TACO dataset may yield better captioning results. However, both datasets have valid use cases: TrashNet could be useful to train a classifier for robot arms sorting trash on a conveyor-belt, while TACO could be used to detect multiple instances of trash in noisy, outdoor environments.

5.2. Future Work

Increasing the diversity and context of training data is generally helpful for improving model performance. Having more varied trash datasets with richer scene context and detailed labels could allow models to better understand objects in their real-world environments. So, the first future work recommendation we can suggest is for to make a larger database with detailed annotations. Having a scene context will be helpful as well, like [10].

Secondly, using a single unified architecture for multiple related tasks like caption generation, label prediction, and object detection can be efficient by sharing representations. However, there are often trade-offs between multi-task performance versus computational cost/complexity. Careful architecture design is needed. A good start can be [26].

Next, including scene context in the labels, rather than just object names, provides useful information to the model about how objects relate to their surroundings. This scene-level reasoning could improve detection accuracy [20].

Also, in order for this method to be viable for trash detection at a larger scale, inference times on VLMs need to improve. [2] introduces MobileVLM V2, a family of significantly improved vision language models that achieve better or on-par performance on standard VLM benchmarks compared with much larger VLMs. Additionally, [15] proposes a novel training strategy for Large Vision-Language models (LVLMs), which can construct a sparse model with a very large number of parameters at a constant computational cost.

Lastly, utilizing advanced natural language processing techniques to better parse and understand the semantics of image captions could indeed extract more useful signal to guide the vision models during training [23].

6. Conclusion

This paper aimed to combine image captioning and zero-shot object detection for trash detection. Despite issues in caption-generation, object detection failures, and long inference times, this method demonstrates promise for trash detection tasks utilizing vision-language models.

6.1. Links to Datasets

TACO Dataset:
http://tacodataset.org/
TrashNet Dataset:
https://github.com/garythung/trashnet

References

[1] Rahmi Arda Aral, Şeref Recep Keskin, Mahmut Kaya, and Murat Hacıömeroğlu. Classification of trashnet dataset based on deep learning models. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2058–2062, 2018. 2

[2] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. Mobilevlm v2: Faster and stronger baseline for vision language model, 2024. 5

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[4] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. Enhancing underwater imagery using generative adversarial networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7159–7165, 2018. 3

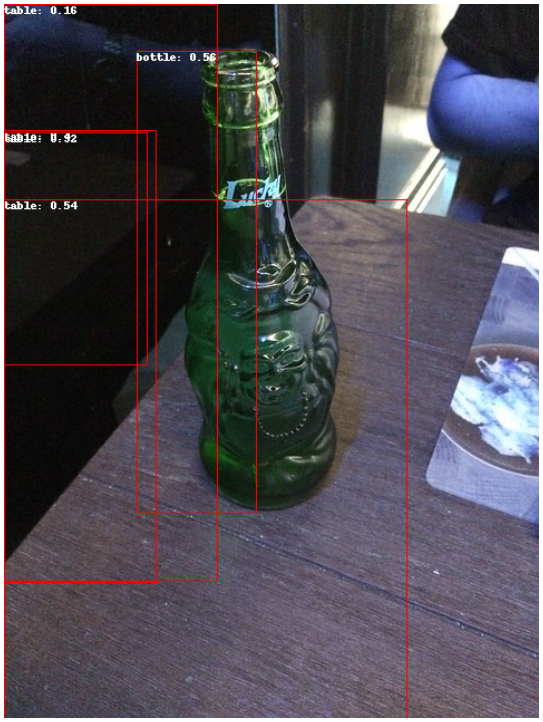

Figure 6. Sample image from the OWL-ViT zero-shot learning

Figure 7. Sample image from the OWL-ViT zero-shot learning

[5] Michael Fulton, Jungseok Hong, Md Jahidul Islam, and Junaed Sattar. Robotic detection of marine litter using deep visual detection models. In *2019 International Conference on*

Figure 8. Sample image from the OWL-ViT zero-shot learning with baseline method

Figure 9. Sample image from the OWL-ViT zero-shot learning with proposed method

Figure 10. Detection failure: Glasses on the table were not detected

Robotics and Automation (ICRA), pages 5752–5758, 2019. 2

[6] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022. 1

[7] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *arXiv preprint arXiv:2309.02561*, 2023. 1, 2

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 2

[9] Jungseok Hong, Michael Fulton, and Junaed Sattar. A generative approach towards improved robotic detection of marine litter. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10525–10531, 2020. 3

[10] Jungseok Hong, Michael Fulton, and Junaed Sattar. Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv preprint arXiv:2007.08097*, 2020. 5

[11] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 2

[12] Kento Kawaharazuka, Yoshiki Obinata, Naoaki Kanazawa, Kei Okada, and Masayuki Inaba. Robotic applications of pre-trained vision-language models to various recognition behaviors. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8, 2023. 1, 2

[13] Jiajia Li, Jie Chen, Bin Sheng, Ping Li, Po Yang, David Dagan Feng, and Jun Qi. Automatic detection and classification system of domestic waste via multimodel cascaded convolutional neural network. *IEEE Transactions on Industrial Informatics*, 18(1):163–173, 2022. 2

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 3

[15] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models, 2024. 5

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2

[17] Koki Maeda, Shuhei Kurita, Taiki Miyanishi, and Naoaki Okazaki. Vision language model-based caption evaluation method leveraging visual context extraction. *arXiv preprint arXiv:2402.17969*, 2024. 5

[18] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 259–274. Springer, 2020. 5

[19] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022. 4

[20] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 5

[21] Pedro F Proença and Pedro Simões. Taco: Trash annotations in context for litter detection, 2020. 2, 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[23] Smriti Sehgal, Jyoti Sharma, and Natasha Chaudhary. Generating image captions based on deep learning and natural language processing. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pages 165–169. IEEE, 2020. 5

[24] Weixuan Sun, Zhen Qin, Hui Deng, Jianyuan Wang, Yi Zhang, Kaihao Zhang, Nick Barnes, Stan Birchfield, Lingpeng Kong, and Yiran Zhong. Vicinity vision transformer, 2023. 5

[25] Gary Thung. Trashnet, 2021. 2, 4

[26] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[27] Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. Connecting the dots: Collaborative fine-tuning for black-box vision-language models, 2024. 2

[28] Qiang Liu Falong Xiao Changyun Li Yuezhong Wu, Xuehao Shen. A garbage detection and classification method based on visual scene understanding in the home environment. In *Complexity, vol. 2021, Article ID 1055604, 14 pages*, 2021. 1

[29] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey, 2023. 1

[30] Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Chang, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2