GODS Pre-Training - Guiding Object Detection Through Segmentation

1st Yalala Mohit
Khoury college of computer science
Northeastern University
Boston, USA
mohit.y@northeastern.edu
CS 7180 : Advanced Perception

***Abstract*—In the evolving landscape of computer vision, the shift from closed-set to open-vocabulary object detection (OVOD) presents new challenges and opportunities. This paper introduces "GODS Pre-Training - Guiding Object Detection Through Segmentation", a novel approach that enhances the precision of the Owl-Vit2 [6] model through segmentation-guided bounding box objectives and an attention-focused learning mechanism. Our methodology combines the inherent strengths of the Owl-Vit2 [6] model with a novel guidance mechanism obtained from available segmentation data along with an attention mechanism refinement. By directing the model's focus towards segmented, object-relevant features, we anticipate improvements in object detection and classification accuracy.**

Experiments conducted on the LVIS dataset [3] demonstrate the effectiveness of our approach. The integration of segmentation and attention objectives alongside existing classification and detection losses leads to significant improvements in model performance. Our GODS model achieves a notable Average Precision of 46.6, showcasing its potential in OD. While outperforming several benchmarks, it highlights the importance of further research into self-training integration, inspired by the OWL-V2 [6] model, to enhance performance in rare category detection. This study sets a new precedent in the field of OVOD, with implications for various real-world applications.

***Index Terms*—Open-Vocabulary Object Detection, Computer Vision, Segmentation-Guided Bounding Box, Attention Mechanism, Owl-Vit2 Model, Machine Learning, Image Processing.**

I. Introduction

In the realm of computer vision, the transition from closed-set object detection to open-vocabulary object detection (OVOD) marks a significant advancement. Traditional methods, while effective in recognizing pre-defined object categories, fall short in the dynamic, unpredictable environments encountered in real-world scenarios. Addressing this gap, our research introduces GODS Pre-Training, a novel approach that synergizes object detection with segmentation, enhancing the performance of the Owl-Vit2 model in OVOD tasks.

Several recent efforts try to apply the models' linguistic capabilities to object detection. Examples of these techniques include weak super vision with image-level labels [10], distillation against embeddings of image crops, and self-training. Owl-V2 [6] presents an intuitive architecture and end-to-end training protocol that, even on categories not observed during training, performs robust open-vocabulary identification without these techniques.

Our methodology is grounded in the hypothesis that integrating segmentation-guided bounding box objectives with a refined attention mechanism can significantly boost the precision of object detection models. By focusing on segmented, object-relevant features, we aim to improve both the accuracy and efficiency of the Owl-Vit2 [6] model. This is achieved by overlaying bounding box predictions onto segmentation ground truths and refining the attention mechanism to concentrate on these localized features.

The primary contribution of our work lies in its innovative use of segmentation data to guide the object detection process, a departure from conventional techniques that rely solely on bounding box predictions. This integration not only enhances the model's ability to accurately detect objects but also enables it to discern and classify objects that were previously unrecognizable due to the limitations of traditional methods.

Our experiments, conducted on the LVIS dataset [3], underscore the efficacy of GODS Pre-Training. The model demonstrates a remarkable improvement in detection precision, especially in the context of open-vocabulary scenarios, where it must recognize and classify objects beyond its training data.

In summary, GODS Pre-Training represents a leap forward in the field of OVOD. By effectively combining segmentation data with an attention-focused learning paradigm, our approach sets a new standard in object detection, paving the way for more accurate and adaptable computer vision systems in real-world applications.

II. Related Work

- **Simple Open-Vocabulary Object Detection with Vision Transformers :** In 2022, the "Simple Open-Vocabulary Object Detection with Vision Transformers" (SOT) paper [5] paved the way for a new era. It recognized the limitations of closed-set approaches and proposed a simple yet effective OVOD method using Vision Transformers (ViTs). These powerful neural networks, trained on both images and text descriptions, learn to extract rich visual and semantic features, enabling them to recognize novel objects beyond their training data. SOT demonstrated impressive performance on benchmark datasets, showcasing the viability of ViTs for OVOD.

- **Scaling Up-Enter Scaling Open-Vocabulary Object Detection (OWL-ST) :** However, SOT faced a critical challenge: data scarcity. Recognizing this limitation, the "Scaling Open-Vocabulary Object Detection" (OWL-ST) paper [6] emerged in 2023, aiming to unlock the full potential of OVOD through scalable training. OWL-ST introduces a groundbreaking approach: self-training on massive web image-text pairs. This allows the model to learn from a vast and diverse dataset, significantly expanding its vocabulary and ability to handle rare or unseen objects.
- **The Power of Self-Training and Refined Representations:** OWL-ST takes things a step further by introducing OWLv2, an improved ViT architecture specifically designed for handling the noisy and diverse data encountered during self-training. Additionally, the paper presents a comprehensive self-training recipe, addressing challenges like label selection, noise filtering, and training instability. This meticulous approach leads to remarkable performance gains compared to SOT and other state-of-the-art methods, particularly for rare and unseen object categories.
- **A Bridge to the Future-Implications and Beyond :** OWL-ST [6] represents a significant leap in OVOD, opening doors to a future where machines can seamlessly adapt to the ever-changing visual landscape. This has far-reaching implications for real-world applications, from autonomous vehicles navigating dynamic environments to robotic assistants handling unfamiliar objects. While OWL-ST marks a significant achievement, the journey of OVOD is far from over. The field continues to grapple with challenges like optimizing self-training strategies, incorporating temporal information for video object detection, and developing interpretable models. Future research holds immense promise for further refining OVOD, making it even more robust and adaptable to the real world.
- **Multi-Task Learning Through the Prism of Multi-Objective Optimization:** Traditionally, multi-task learning (MTL) trained models on multiple tasks simultaneously, but struggled with conflicting objectives and trade-offs. The 2018 paper "Multi-Task Learning as Multi-Objective Optimization" (MTLM as MO) [7] tackled this by reframing MTL as a multi-objective optimization problem. This allowed researchers to leverage powerful optimization algorithms to balance task priorities and efficiently converge to high-quality solutions. With its custom-designed MGDA-UB algorithm, MTLM as MO demonstrated significant performance gains in diverse domains, solidifying its position as a crucial advancement in the field, inspiring further research and propelling MTL towards greater effectiveness.In essence, MTLM as MO shifted the perspective on MTL, unlocking its potential through optimization algorithms and paving the way for more versatile and effective multi-task learning models.

III. Methodology

A. Hypothesis

Our research is predicated on the hypothesis that the precision and effectiveness of object detection models can be substantially enhanced through a novel integration of segmentation-guided bounding box prediction and a refined attention mechanism. This approach is grounded in the belief that segmentation data, offering a more detailed representation of object boundaries, when coupled with an attention mechanism that focuses on these segmented areas, can lead to significant improvements in object detection and classification accuracy. Essentially, we hypothesize that guiding the model's local focus towards segmented, object-relevant features, rather than the broader image context, will yield a more precise and efficient object detection model.

B. Approach and Implementation

1) Dataset and Ground Truth Preparation: The foundation of our approach begins with the preparation of the dataset. We utilize the LVIS dataset [3], a large-scale benchmark for object detection and segmentation. For each image in the dataset, comprehensive segmentation masks are generated by overlaying individual object masks. These masks form the primary ground truth for our segmentation tasks, offering a detailed and accurate representation of object boundaries.

2) Integration of Bounding Box Prediction and Attention Mechanism:

1) **Bounding Box Prediction:** The Owl-Vit2 [6] model is employed to predict bounding boxes on original input images, which are then overlaid onto the segmentation ground truth masks. To ensure training efficiency, a model free solution is leveraged to generate the predicted segmentation masks with incorporated bounding box information. For this, after the bounding boxes are overlaid onto the ground truth segmentation masks, an operation is performed to only include the regions inside the predicted bounding boxes, while eliminating any regions that lie outside these predictions. Essentially, a 100 accurate segmentation decoder, would have achieved the same results in terms of segmentation masks. Our approach is more scalable and efficient in this context.
2) **Attention Mechanism Refinement:** As observed in DINO [2], the emergent properties of the model, allow its' attention masks to be focused on the foreground of the image, thereby, leading to great downstream classification and detection capabilities. However, such a trend is non-observant when the attention masks of a supervised model are visualized. Essentially, the attentions of a patch with itself, focuses on the local features, while the attentions of a patch with other patches focuses on the global features. Hence, we come up with the attention hypothesis, which aims to obtain better results by guiding the attentions of a patch with itself, to be focused on the foreground objects. The hypothesis suggests that locally, the attention masks

should be focused on learning the object boundaries or masks, which would then allow the detection decoder to extract non-linear information about the boundaries of the object along with the other features that the model learned to extract. In essence, we aim to partially guide the attention masks in learning the object boundaries to make the model more robust for downstream tasks that could be beneficial from this information. To obtain this the attention masks from all attention heads of the model are obtained, and the diagonal elements that represent local features of the image, are used to reconstruct the attention mask. This results in an 'attention mask predictions' that reflects the model's focus areas, and needs to be guided.

3) Extraction of Segmentation and Attention Regions: The next phase involves the extraction of three key types of regions:

- **Original Segmentation Masks:** These are the compounded masks obtained from the ground truth segmentation.
- **Predicted Segmentation Regions:** Masks derived from regions within the predicted bounding boxes, layered over the ground truth segmentation masks.
- **Attention Image Regions:** Reconstructed attention masks, focusing on the local features of each image, essentially the reconstructed attention mask obtained by reconstructing the attentions of a patch with itself.

4) Multi-Objective Loss Computation and Model Fine-Tuning: Our approach utilizes a multi-objective loss function that incorporates various aspects of the model's performance:

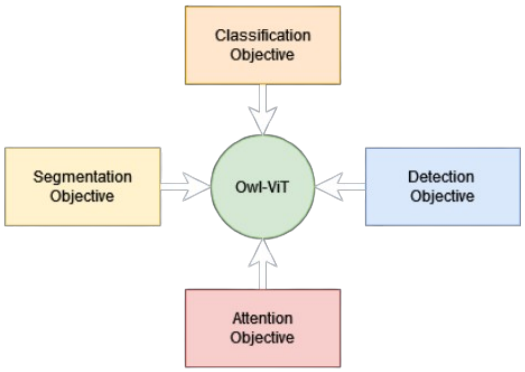

Fig. 1: Our proposed Objective function

- **Classification Loss:** The Focal Loss is employed to focus on hard-to-classify examples in imbalanced datasets. It is defined as:

$$\text{Focal Loss} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

where p_t is the model's estimated probability for the class label t, α_t is a weighting factor for class t, and γ is the focusing parameter.

- **Detection Loss:** The Smooth L1 Loss, also known as Huber Loss, is used for balancing sensitivity and robustness. It is given by:

$$\text{Smooth L1 Loss} = \begin{cases} 0.5 \cdot (x - y)^2 & \text{if } |x - y| < 1 \\ |x - y| - 0.5 & \text{otherwise} \end{cases} \tag{2}$$

where x is the predicted value, and y is the true value.

- **IoU and GIoU Losses:** These losses measure the overlap between predicted and ground truth masks. The IoU Loss is defined as:

$$\text{IoU Loss} = 1 - \frac{\text{Intersection}}{\text{Union}} \tag{3}$$

The GIoU Loss extends IoU by considering the minimum enclosing box:

$$\text{GIoU Loss} = 1 - \text{IoU} + \frac{\text{Enclosing Area} - \text{Union}}{\text{Convex Area}} \tag{4}$$

- **Segmentation & Attention Loss:** Assessed using Dice Loss, which compares attention mask predictions with ground truth segmentation:

$$\text{Dice Loss} = 1 - \frac{2 \times \text{Intersection}}{\text{Union}} \tag{5}$$

The integration of these loss metrics—segmentation, attention, along with the existing classification and detection loss is hypothesized to not only enhance the accuracy of bounding box predictions but also to refine the model's attention mechanism. By prioritizing local features within the areas of interest, the model is expected to exhibit increased precision in object detection and classification.

5) Anticipated Outcomes: We anticipate that this integrated approach, combining segmentation-guided bounding box predictions with an attention-focused learning paradigm, will set a new benchmark in object detection accuracy. The model's increased precision in detecting and classifying objects is expected to demonstrate a significant advancement in computer vision, particularly in applications requiring high precision and efficiency.

IV. Experiments and Results

A. Experimental Setup

Our experiments were meticulously designed to evaluate the effectiveness of the GODS Pre-Training model. The training was conducted on the LVIS dataset, a comprehensive benchmark for object detection and segmentation. Firstly, the model was trained with a blend of classification, detection, and segmentation objectives for 15 epochs using the ADAM optimizer with a learning rate of 0.0001. Subsequently, the attention objective was integrated, and the model underwent a fresh 20 epochs of training. Notably, a significant improvement in model performance was observed after the 18^{th} epoch of this second experiment.

(a) Detections with False positives - streetsign, backpack

(b) Tiny object detection capabilities

(c) Accurate detections in a cluttered environment

(d) Accurate detections in a cluttered environment

Fig. 2: GODS model tested on a wide range of objects exhibit good detection capabilities, on instances it hasn't seen during train.

TABLE I: Open-Vocabulary Object Detection Performance on LVIS (AP (%))

Method	Backbone	Self-training Data	Self-training Vocabulary	Human Box Annotations	LVIS AP $^{val}_{all}$	LVIS AP $^{val}_{rare}$
RegionCLIP [9]	R50x4	CC3M	6K concepts	$LVIS_{base}$	32.3	22.0
OWL [5]	CLIP B/16	-	-	O365+VG	27.2	20.6
OWL [5]	CLIP L/14	-	-	O365+VG	34.6	31.2
F-VLM [4]	R50x4	-	-	$LVIS_{base}$	28.5	26.3
F-VLM [4]	R50x64	-	-	$LVIS_{base}$	34.9	32.8
3Ways [1]	NFNet-FO	TODO	captions	$LVIS_{base}$	35.7	25.6
3Ways [1]	NFNet-F6	TODO	captions	$LVIS_{base}$	44.6	30.1
OWL V2-ST [6]	CLIP B/16	WebLI	N-grams	O+VG	27.0	29.6
OWL V2-ST [6]	CLIP L/14	WebLI	N-grams	O+VG	33.5	34.9
OWL V2-ST [6]	SigLIP G/14	WebLI	N-grams	O+VG	33.7	37.5
OWL V2-ST+FT [6]	CLIP B/16	WebLI	N-grams	O+VG, $LVIS_{base}$	41.8	36.2
OWL V2-ST+FT [6]	CLIP L/14	WebLI	N-grams	O+VG, $LVIS_{base}$	49.4	44.6
OWL V2-ST+FT [6]	SigLIP G/14	WebLI	N-grams	O+VG, $LVIS_{base}$	47.0	47.2
Detic [10]	R50	IN-21k	LVIS classes	$LVIS_{base}$	32.4	24.6
DetCLIPv2 [8]	Swin-T	CC15M	Nouns+curated	O365+GoldG	32.8	31.0
DetCLIPv2 [8]	Swin-L	CC15M	Nouns+curated	O365+GoldG	36.6	33.3
OWL-ST+FT [6]	CLIP B/16	WebLI	N-gram+curated	O+VG, $LVIS_{base}$	45.6	40.5
OWL-ST+FT [6]	CLIP L/14	WebLI	N-gram+curated	O+VG, $LVIS_{base}$	50.4	45.9
GODS (Ours)	CLIP B/16	-	-	$LVIS_{base}$	46.6	26.3

B. Performance Evaluation

For the evaluation, output logits were ranked by their objectness score, and only the top 50% of the 3600 output logits per image were used for multi-objective loss calculation. The models from both sets of experiments were evaluated on the LVIS dataset, comprising 100,000 images. Importantly, exposure to rare category objects was deliberately omitted during training to test the model's open-vocabulary performance.

C. Comparative Analysis

Our GODS model achieved an Average Precision (AP) of 46.6 on the LVIS validation set, which is a remarkable achievement, particularly when compared to other state-of-the-art methods as detailed in Table 1. For instance, RegionCLIP [9] achieved an AP of 32.3, while the more advanced OWL-ViT [5] models ranged from 27.2 to 34.6 AP, depending on the backbone used. Notably, our model even outperformed the F-VLM [4] with an R50x64 backbone, which scored 34.9 in AP.

However, it is crucial to note that our model showed a significant drop in AP for LVIS rare categories, achieving only 26.3 AP. This is a notable area for future improvement, especially when compared to the OWL-ST+FT model using the CLIP L/14 backbone [6], which achieved an impressive 44.6 AP in rare categories. The superior performance of OWL-ST+FT, especially in rare category detection, can be attributed to its self-training on web image-text pairs and fine-tuning on the LVIS dataset. This drop reaffirms the findings in the Owl-V2 [6] paper which suggested a fusion of self-training and fine-tuning to address this challenge. Acknowledging the absence of self-training in GODS, future work could explore its integration, inspired by insights from Owl-V2, to enhance performance and mitigate the observed drop in LVIS rare category AP.

D. Attentions Demystified

One of the key aspects of our study is the visualization and comparison of attention mechanisms between the OWL V2 model and our GODS model, as illustrated in Figure 3. This comparison is vital to understanding the improvements in model performance.

The OWL V2 model, as observed in the top two images of Figure 3, has shown an inherent tendency to focus on the foreground object boundaries in certain scenarios. However, this is not consistent across all cases. As depicted in the bottom two images of the same figure, there are instances where the

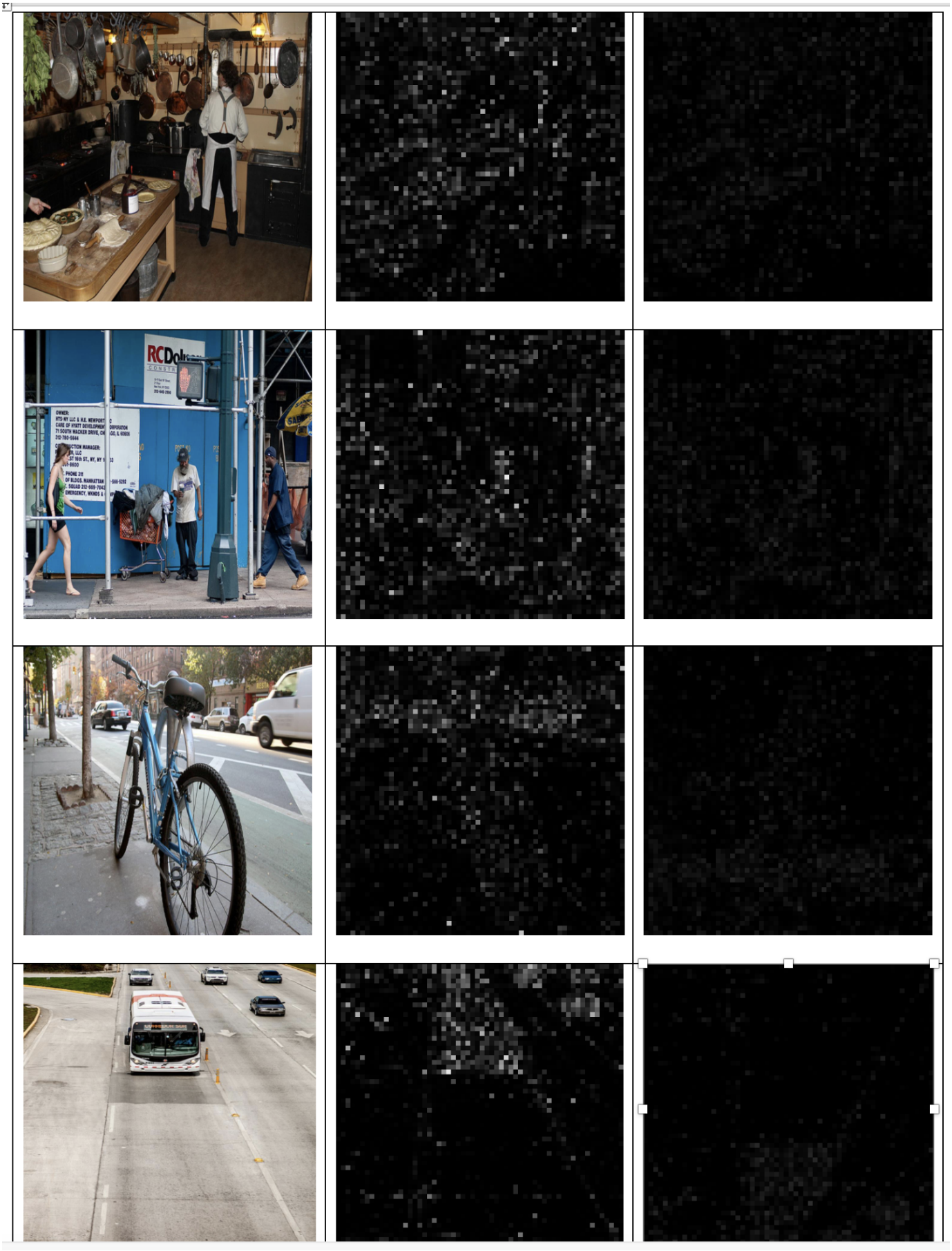

Fig. 3: Comparative visualization of the attention masks. To the left is the source image, in the middle are the attention masks generated by the model we trained, to the left is the attention masks obtained from the OWL V2 model

OWL V2 model's attention deviates towards the background, potentially compromising the model's effectiveness in object detection tasks.

Conversely, our GODS model exhibits a more consistent behavior. The attention mechanism in our model consistently focuses on the object boundaries and extends its scope to cover the entire region of the foreground objects. This consistency and expanded focus area, as demonstrated in our model's attention visualizations, suggest a significant enhancement in the model's ability to accurately detect and classify objects. We hypothesize that this shift in attention focus, from partial to more holistic coverage of the object areas, is a fundamental factor contributing to the observed increase in model performance.

V. Discussion and Future Work

A. Why Pre-training and not Fine-tuning

A pivotal aspect of our research approach is the decision to focus on pre-training rather than fine-tuning. This choice stems from the nature of using segmentation masks in object detection models. Traditionally, fine-tuning processes often rely on human-labeled data. However, the reliance on such data is not only resource-intensive but also limits the scalability of the training process.

The advent of advanced models like the 'Segment Anything' approach has opened new avenues for generating pseudo labels. These pseudo labels, derived from unsupervised or semi-supervised learning techniques, provide a viable alternative to manually annotated data. By utilizing these pseudo labels, we can pre-train our model on a massive scale, harnessing a broader range of data without the constraints of human labeling.

Furthermore, we propose an innovative methodology where the ideology of generating guidance images from predicted bounding boxes could be applied to the RGB input images. This approach involves incorporating bounding box information into the input images and then calculating a channel-wise Dice loss. The integration of this strategy into the pre-training phase allows for a more comprehensive and data-driven model development. This process, fundamentally different from the typical fine-tuning approach, justifies our emphasis on pre-training.

B. Conclusion

The results demonstrate that while the GODS model presents a novel and effective approach to open-vocabulary object detection, there is still room for improvement, particularly in rare category detection. The comparative analysis highlights the potential benefits of integrating self-training methodologies, as seen in the performance of OWL-ST+FT models. Future work will explore the integration of self-training inspired by the OWL-V2 [6] insights, to enhance our model's performance in detecting rare category objects.

References

[1] Relja Arandjelović, Alex Andonian, Arthur Mensch, Olivier J. Hénaff, Jean-Baptiste Alayrac, and Andrew Zisserman. Three ways to improve feature alignment for open vocabulary detection, 2023.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021.

[3] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019.

[4] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models, 2023.

[5] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*.

[6] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection, 2023.

[7] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[8] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment, 2023.

[9] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining, 2021.

[10] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.