LOWA: Localize Objects in the Wild with Attributes

Xiaoyuan Guo*, **Kezhen Chen***, **Jinmeng Rao, Yawen Zhang, Baochen Sun, Jie Yang**
Mineral
{xiaoyuanguo,kezhenchen,jinmengrao,yawenz,baochens,yangjie}@mineral.ai

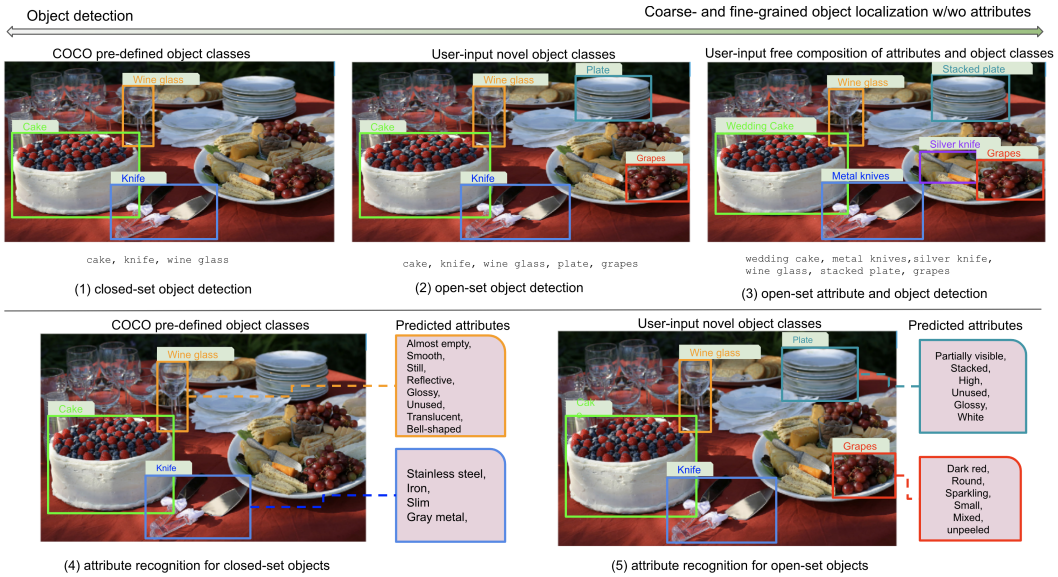

Figure 1: Our model, LOWA, performs open-vocabulary free-text object localization (top row) and attribute inference for both common and novel classes (bottom row). (1) Most current open-vocabulary attribute recognition models limit their object classes into COCO pre-defined classes. (2) However, our model can handle a broader range of object classes in a real open-world setting; (3) We exemplify a real case of flexible composition of text queries for object localization; (4) Not only our model can infer object attributes for common COCO object categories; (5) But also transfer general attribute knowledge to novel classes and infer accurately. Note we only show one detected instance for each class for better visualization.

Abstract

Existing open-vocabulary object detectors can struggle with uncommon or fine-grained classes, as the model and users may have different understandings of object names. Incorporating attributes such as color, shape, and size can help to reduce this inconsistency and make interactive detection more convenient and flexible. Motivated by this, we present LOWA, a new method for localizing objects with attributes effectively in the wild. To train LOWA, we propose a multi-step vision-language training strategy to learn object detection and recognition with class names as well as attribute information, which empowers users to flexibly customize text queries and extend to fine-grained detection with attribute and object information for a wider range of applications. LOWA is built on top of a two-tower vision-language architecture and consists of a standard vision transformer

*Xiaoyuan and Kezhen have equal contributions.

37th R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023(NeurIPS 2023).

as the image encoder and a similar transformer as the text encoder. To learn the alignment between visual and text inputs at the instance level, we train LOWA with three training steps: object-level training, attribute-aware learning, and free-text joint training of objects and attributes. This training strategy first ensures correct object detection, then incorporates instance-level attribute information, and finally balances the object class and attribute sensitivity. We evaluate our model performance of attribute classification and attribute localization on the Open-Vocabulary Attribute Detection (OVAD) benchmark and the Visual Attributes in the Wild (VAW) dataset, and experiments indicate strong zero-shot performance. Ablation studies additionally demonstrate the effectiveness of each training step of our approach.

1 Introduction

Open-Vocabulary Object Detection (OVD) foundation models (Minderer et al., 2022; Liu et al., 2023; Li et al., 2022) align the text of object names with their visual features, enabling zero/few-shot object detection with free-text inputs. However, these models are limited by their reliance on common object vocabularies, which can make it difficult to use customized text queries, especially in fine-grained domains. There are two main challenges with existing OVD foundation models: (1) Data bias: The training datasets typically contain more instances of common objects than rare objects, which leads to poor performance on text queries with terminologies; (2) Limited expressiveness: Using object names as detection queries is not always optimal. In many real-world scenarios, users prefer to describe objects with their attributes (e.g., shape, color, texture) instead of precise terminologies. For example, using "*a black and white bird with red head feather*" is more accessible in communication than using "*Antioquia Brushfinc*". To address these issues, a promising direction is to use attributes like color, shape, texture, pattern, and action as anchors to extend the generalization ability of existing OVD foundation models. Moreover, the knowledge of seen attributes can transfer to rare/unseen object categories sharing the same properties. With attributes, users can describe objects with more general free-text queries and specify the unusual aspects of a familiar object (*stacked plate*, not just *plate*), see Fig. 1.

However, attribute detection at instance-level is challenging due to the high cost of collecting and annotating datasets. Researchers have proposed various datasets to address this challenge, ranging from patch-labeled (Isola, Lim, and Adelson, 2015) , partially-labeled (Pham et al., 2021), sparsely-labeled (Krishna et al., 2017; Pham et al., 2022) to densely-labeled (Bravo et al., 2022). Popular works on instance-level attribute knowledge learning typically skip the object detection step by freezing the base object detector (Faster-RCNN (Ren et al., 2015)) and then adding a secondary attribute classifier (Bravo et al., 2022; Chen et al., 2023). This strategy has two drawbacks: (1) it is complex to design and implement; (2) it is not end-to-end trainable. Additionally, these works often limit attribute and object name lists to a fixed number of vocabularies, which is not suitable for the real open-world setting, where new objects and attributes may be encountered at any time.

In contrast, we propose a novel method, called *LOWA* (**L**ocalize **O**bjects in the **W**ild with **A**ttributes), that is simple, effective and can handle a broader number of object categories and attributes. LOWA is also the first end-to-end model that can simultaneously perform object bounding box regression, object name classification, and object attribute classification. LOWA overcomes the challenge of incomplete annotations by disentangling the learning of object names and attributes. This allows users to query images using the diverse compositions of object classes and attribute descriptions to get customized detection results. The architecture of LOWA includes a standard vision transformer as the image encoder and another transformer as the text encoder. To enhance the attribute-awareness of the model, we design a three-step strategy for LOWA to align visual features with both object classes and attributes. We summarize our contributions as follows:

- We propose a novel one-stage model for open-vocabulary object detection with attribute classification. To the best of our knowledge, this is the first open-vocabulary fine-grained detection model that is trained in an end-to-end fashion to optimize all the parameters. Compared with other models such as OvarNet (Chen et al., 2023) or GLIP (Li et al., 2022), our model is more flexible and has stronger fine-grained attribute-awareness.

- We design a customized three-step training approach to train LOWA following a coarse-to-detailed learning schema, which enhances the ability to automatically focus on both object classes and attributes in free-style texts and localize the objects accurately.
- We explore the under-represented attribute localization task based on open-vocabulary object detection. We evaluate our model on the attribute classification task and attribute localization task with two benchmarks to show the attribute awareness. On both tasks, our model outperforms all the baselines.
- We conduct ablation studies to show the effectiveness of our proposed multi-step training approach. Results indicate that each step provides significant improvements.

2 Method

Problem Definition Given an image I and a set of free-text queries $Q = \{q_1, q_2, ..., q_n\}$ (n defines the number of input queries) including object names, attributes and composition of both, the goal of LOWA is to localize target objects $O = \{o_1, o_2, ..., o_m\}$ (m is the total number of objects) and output their corresponding bounding boxes $B = \{b_1, b_2, ..., b_m\}$ with a score matrix $S_{m \times n}$. Each object is assigned with n scores in response to each query. A score threshold τ helps determine the final positive predictions. To enable free-text object localization without explicitly separating attributes and object names, we elaborate on our method LOWA in terms of the model architecture, the instance attribute learning process, training optimization and user inference.

Model Architecture LOWA uses a Vision Transformer Encoder as the image encoder E_I and another Transformer Encoder as the text encoder E_T, as depicted in Fig. 2. The text encoder generates a text embedding e_q for each text query q. Similarly as Minderer et al. (2022), we remove the token pooling and final projection layer from the text encoder. The image encoder encodes the image into a sequence of image embeddings, where each image embedding represents a candidate object. A text prediction head H_{cls} applies linear projection on each image embedding to generate an instance-level visual embedding for classification, and a box prediction head H_{box} uses a Multi-Layer Perceptron (MLP) for bounding box regression. Thus, the total number of object proposals is equal to the length of image encoder inputs. We choose the architecture of ViT-L (768 embedding dimensions and 24 layers) with patch size 14 at input size 840 x 840 as the image encoder. Thus, the number of image embeddings is 3,600. Both the image encoder and text encoder are initialized from a pre-trained CLIP model (Radford et al., 2021). As the model architecture only uses Transformer encoders, all the parameters take advantage of image-level pre-training in CLIP.

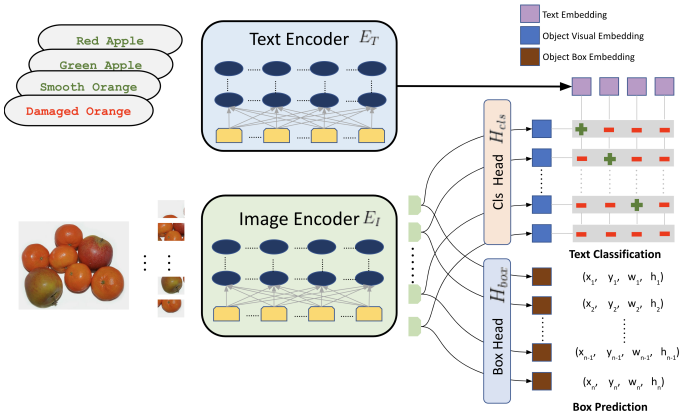

Figure 2: LOWA Architecture Overview: A text encoder E_T takes a set of text queries Q and generates a text embedding for each query. The image encoder E_I takes an image I as input and generates object-level embeddings. The object-level embeddings are passed to a text prediction head and a box prediction head. The text prediction head H_{cls} generates the object visual embeddings for object classification. The box prediction head H_{box} generates the object box embeddings for box coordinate regression.

Instance Attribute Learning Learning instance-level attribute knowledge is challenging for existing OVD foundation models with free-text queries as input, as it requires the models to automatically identify the keywords without additional instruction information. To achieve that, we propose a multi-step training strategy for LOWA with a customized three-step learning process for a coarse-to-detailed alignment: **(1)** object-level training step (Step$_O$ in short), **(2)** attribute-aware training step (Step$_A$ in short) and **(3)** free-text training step (Step$_F$ in short).

(1) Object-level training step trains the model to detect objects by aligning object class names with visual representations. For each image, we provide the object names as the positive text queries. A set of non-overlapped object class names is randomly picked from a candidate set as the negative text queries (We get the object label set by combining the object labels of Objects365 (Shao et al., 2019) and LSA (Pham et al., 2022) datasets.). The objective of this training step is to facilitate the matching of provided object classes with the corresponding object-level visual embeddings. Through ablation studies in Sec. 3.4, we have demonstrated that introducing coarse-grained object detection prior to fine-grained attributes yields a substantial improvement in accuracy for fine-grained object detection.

(2) Attribute-aware training step ensures that the correct objects are localized for Step_A and thus the corresponding attribute details associated with a specific object can be learned. Specifically, as each object could have multiple attributes, we regard each attribute as an additional label to the object. The model is trained to align each visual embedding with multiple text queries. For example, *"a red peeled apple"* corresponds to three positive text queries: *"red"*, *"peeled"*, and *"apple"*. To create negative samples, the object classes and attributes are merged together to construct a label candidate set. For each object, we randomly sample some objects or attributes from the union set as the negative text queries. With these multi-label settings, the image encoder is optimized to disentangle the semantic meaning of attributes and classes.

(3) Free-text training step aims to enhance the vision-language alignment on more free-styled text queries containing attributes and class names. For each object, we randomly pick one attribute from its attribute lists and concatenate it with the object class to construct a positive text query. For negative queries, the positive text queries are broken by replacing either the object class or the attribute. We randomly picked either an object class or an attribute without overlaps. After this, LOWA learns to match text queries with fine-grained descriptions and instance-level visual embeddings.

Training Optimization and Inference All the training steps use the same bipartite matching loss $\mathcal{L}_m$ introduced by DETR (Carion et al., 2020) for object classification and bounding box prediction. We use the L1 loss $\mathcal{L}_1$ and the GIOU loss $\mathcal{L}_g$ for bounding box prediction. We also use the focal loss $\mathcal{L}_c$ as the classification loss for text labels, as each single object could correspond to multiple text labels. Instead of predicting the logits in a fixed global label space as closed-vocabulary detectors, the classification projection head outputs logits over the per-image label space defined by the queries. The matching loss $\mathcal{L}_m$ is formally defined as: $\mathcal{L}_m = \mathcal{L}_1 + \mathcal{L}_g + \mathcal{L}_c$ During inference, LOWA predicts the bounding boxes of objects in an image based on the provided text queries. Each text query could be a fine-grained description of an object with certain attributes.

3 Experiments

3.1 Experimental Settings

Baselines Since our model is designed to localize objects in the open-vocabulary environment, we choose two SOTA open-vocabulary detection models as the baselines: OWL-ViT (Minderer et al., 2022) and GroundingDINO (Liu et al., 2023). Both models have impressive open-vocabulary object detection ability and have shown attribute-awareness.

Evaluation Tasks Our model can perform multiple tasks, but we focus on its most significant feature: object localization with attributes. Therefore, we evaluate our model primarily on tasks that involve this capability: **attribute classification** and **attribute localization**. Attribute classification is the task of identifying the attributes of an object given its class name. The current attribute evaluation benchmarks are generally for this task. In contrast, attribute localization is the task of identifying the localization of an object given its attributes. Additionally, we report object detection performance on OVAD dataset following the previous works.

Evaluation Datasets and Protocol We use two attribute-annotated datasets, VAW (Pham et al., 2021) and OVAD (Bravo et al., 2022) for evaluation. VAW contains 620 attributes covering object color, material, shape, size, texture and action. OVAD is a densely annotated objects and attributes benchmark with 117 attribute classes for over 14,300 object instances (no training set is provided). We tested the two baselines under the **box-free** setting. Without target bounding boxes provided, we obtain predictions of bounding boxes and probabilities for a given image with a list of candidate attributes and object names. Under the box-free setting, we first find the predicted box that has the largest Intersection over Union (IoU) with the ground truth box among all the predicted bounding boxes, and then use the corresponding object feature for attribute classification. We follow the same

workflow for evaluating GroundingDINO and OWL-ViT. We calculate the mean average precision (mAP) over all the attributes as the metric to compare all the methods.

Evaluation Metrics For attribute recognition, we follow VAW and OVAD to measure attribute prediction from a different perspective: mAP, mean average precision over all classes; mR@K, mean recall over all classes at top K predictions in each image; and F1@K, F1 scores over all classes at top K predictions in each image. K is selected as 8 for OVAD and 10 for VAW following the previous works' default settings. For attribute localization, we report mAR@10 results, which is the mean average recall at top 10 predictions.

3.2 Attribute Classification Results and Analysis

Comparison with baseline models *(1) OVAD Benchmark:* We report the overall attribute recognition performance in Tab. 1 using the metrics mAP, mR@15 and F1@15. As the occurrence of attributes is long-tailed, we use the provided standard splits (Head, Medium and Tail) of the attributes and present the model mAP performance for each of them. This helps us to inspect model's generalization ability in handling common attributes and rare attributes. Additionally, we evaluate mAP at 0.5 IoU for the open-vocabulary object detection task on the 80 class object set, called OVD-80. As we can see, our model LOWA exhibits the best performance overall and is able to recognize attributes with different frequencies, especially the rare attributes. In contrast, GroundingDINO and OWLViT are not good at attribute classification, despite their decent performance on object detection tasks, as shown in the OVD-80 results. For a fair comparison, we also continued training OWL-ViT for 100K steps using the same training data. We report this model's performance with the model as OWL-ViT (ViTL/14, Cont.). However, the attribute classification ability drops compared to the original OWL-ViT model.

(2) VAW Benchmark: Similar to the OVAD dataset, we evaluate models using the zero-shot attribute classification task on the VAW validation set under a box-free setting. Table 2 shows the performance comparison between all the baseline models and our model. Based on the results, our model outperforms all the baseline models on the attribute classification task and thus shows enhanced attribute-awareness.

Table 1: Performance comparison with baseline models on the OVAD benchmark for attribute classification and object detection. (Bold indicates the optimal performance.)

Method	OVAD						Generalized OVD-80 (AP_{50})		
	mAP(all)	mR@8	F1@8	Head	Medium	Tail	$Novel_{(32)}$	$Base_{(48)}$	$All_{(80)}$
GroundingDINO (Swin-B)	8.9	8.0	9.0	36.0	7.6	0.7	**70.5**	64.6	67.0
GroundingDINO (Swin-T)	8.9	7.8	8.1	34.4	8.1	0.8	56.1	54.8	55.3
OWL-ViT (ViTL/14)	11.1	13.9	15.8	42.8	10.0	1.3	58.6	56.4	57.3
OWL-ViT (ViTL/14, Cont.)	10.2	12.3	14.0	40.3	9.0	0.9	69.3	**67.4**	**68.2**
LOWA (Ours)	**18.7**	**35.3**	**39.4**	**58.0**	**20.4**	**2.6**	68.2	67.0	67.5

Table 2: Performance comparison on attribute classification with baseline models on VAW validation set. (Bold indicates the best performance.)

Model	mAP(all)	mR@15	F1@15	Head	Medium	Tail
GroundingDINO (Swin-B)	30.2	2.0	3.7	33.8	28.7	20.8
GroundingDINO (Swin-T)	30.0	2.7	5.0	33.9	28.4	20.2
OWL-ViT (ViTL/14)	37.4	12.3	20.4	41.8	35.9	25.3
OWL-ViT (ViTL/14, Cont.)	36.1	9.4	16.1	40.3	34.4	25.6
LOWA (Ours)	**50.0**	**36.9**	**45.5**	**54.4**	**47.8**	**39.4**

Table 3: Additional performance comparison of attribute classification models with public OVAD benchmark.

Method	Box	OVAD						Generalized OVD-80 (AP_{50})		
		mAP(all)	mR@8	F1@8	Head	Medium	Tail	$Novel_{(32)}$	$Base_{(48)}$	$All_{(80)}$
OV-Faster-RCNN (Bravo et al., 2022)	given	11.7	–	–	34.4	13.1	1.9	0.3	53.3	32.1
VL-PLM (Zhao et al., 2022)	given	13.2	–	–	32.6	16.3	2.6	19.7	58.8	43.2
Detic (Zhou et al., 2022)	given	13.3	–	–	44.4	13.4	2.3	20.0	49.2	37.5
LocOv (Bravo, Mittal, and Brox, 2022)	given	14.9	–	–	42.8	17.2	2.2	22.5	52.5	40.5
OVR (Zareian et al., 2021)	given	15.1	–	–	46.3	16.7	2.1	17.9	51.8	38.2
OVAD (Bravo et al., 2022)	given	18.8	–	–	47.7	22.0	4.6	24.7	49.1	39.3
OvarNet(ViT-B16) (Chen et al., 2023)	free	**27.2**	–	–	56.8	**33.6**	**8.9**	35.2	60.4	54.2
LOWA (Ours)	free	18.7	**35.3**	**39.4**	**58.0**	20.4	2.6	**68.2**	**67.0**	**67.5**

Additional comparison with other relevant models We also put the comparison results between our model and recent traditional attribute classification models in Tab. 3. Most of the models in this

Table 4: Performance (AR@10) of Attribute Localization on OVAD dataset

Model	Cleanliness	Pattern	State	Type	Material	Size	Texture	Length	Tone
OWL-ViT	20.1	24.4	19.9	12.8	29.7	23.8	12.9	10.8	8.8
OWL-ViT (ViTL/14. Cont.)	**25.0**	37.3	21.5	15.4	28.6	22.3	16.4	20.2	10.7
LOWA (Ours)	23.3	**37.9**	**21.8**	**29.6**	**32.2**	**26.6**	**22.5**	**27.4**	**15.3**

Table 5: Ablation studies on OVAD benchmark.

Method	Training Steps	OVAD			Generalized OVD-80 (AP_{50})		
		mAP(all)	mR@8	F1@8	$Novel_{(32)}$	$Base_{(48)}$	$All_{(80)}$
$Step_O$+$_A$ (Ours)	250K	17.0	27.7	31.7	**75.4**	**71.8**	**73.3**
$Step_A$+$_F$ (Ours)	250K	16.3	27.3	30.5	45.9	40.1	42.4
$Step_O$+$_F$ (Ours)	250K	18.1	32.4	36.3	66.4	66.1	72.1
$Step_O$+$_A$+$_F$ (Ours)	350K	18.5	34.4	38.5	70.2	68.7	69.3
$Step_O$+$_A$+$_F$ (Ours)	250K	**18.7**	**35.3**	**39.4**	68.2	67.0	67.5

table lack the object detection components, and thus they are evaluated under a simpler **box-given** setting. The ground truth locations of objects are provided and the models only perform attribute classification on each given object. Based on the results, our model achieves competitive results even compared with these traditional attribute classification models. The OvarNet is also a box-free model designed for open-vocabulary attribute classification and achieves impressive performance. However, OvarNet is trained via a teacher-student process to learn attribute detection. Our model uses a one-stage detection process, which is more efficient and general to adapt to multiple visual tasks. Also, our model significantly outperforms OvarNet on object detection results (improved by 24.5% on all objects). Results from Tab. 1 and Tab. 3 provide promising evidence to show that our model has better attribute-awareness compared to other OVD baselines while keeping strong object detection ability.

3.3 Attribute Localization Results and Analysis

We further evaluate the effectiveness of our model on attribute localization. To this end, we reuse the OVAD dataset and only use attributes as queries to localize the target objects such as *"red"* or *"smooth"* object. Specifically, we evaluate the attribute awareness for each attribute category. Table 4 reports nine representative categories. From the table, LOWA significantly outperforms all the baselines across various attribute perspectives, which further demonstrates its attribute-awareness.

3.4 Ablation Study

To investigate the effectiveness of our training strategy, we conduct ablation studies by removing each training step. We increase the training steps in each ablation study to keep the same number of total training steps for a fair comparison. We increase the $Step_O$ or $Step_F$ to 200K steps for setting $Step_O$+$_F$ and $Step_A$+$_F$. By removing $Step_F$, we increase 25K steps for each previous step. Furthermore, as we mentioned in Section 2, the goal of $Step_F$ is to enhance vision and language matching instead of improving attribute and object detection. We also investigate whether $Step_F$ follows our assumption by increasing the training steps in $Step_F$ to 150K steps. Table 5 shows the performance comparison of all the studies. Results show that the model trained with three steps outperforms all the other designs that miss one training step. These studies convince us that our training framework is effective in enhancing the fine-grained object detection of OVD models. Training the model to learn open-vocabulary object detection from coarse-grained to fine-grained features is necessary to disentangle classes and attributes. Also, we prove our assumption that longer $Step_F$ does not help attribute and object detection.

4 Conclusion

In this work, we introduce LOWA, a new open-vocabulary object detection model with fine-grained attribute-awareness. We present a novel training framework to incorporate attribute information, which is fully compatible with existing OVD models. Our model trained on large-scale data can perform open-vocabulary object detection with free-text inputs in a one-stage manner. We also compare our model with several SOTA open-vocabulary detection models and observe a significant improvement on two popular benchmarks. Results show that LOWA has strong zero-shot detection performance and can disentangle object classes and attributes. Ablation studies indicate that each step of our training framework is effective and important. We believe that this work could improve the bottleneck of existing OVD models on fine-grained domain adaption. The novel ideas and observations could provide valuable insights to researchers in this field.

References

Bravo, M.; Mittal, S.; and Brox, T. 2022. Localized vision-language matching for open-vocabulary object detection. *GCPR 2022*.

Bravo, M. A.; Mittal, S.; Ging, S.; and Brox, T. 2022. Open-vocabulary Attribute Detection. *arXiv preprint arXiv:2211.12914*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

Chen, K.; Jiang, X.; Hu, Y.; Tang, X.; Gao, Y.; Chen, J.; and Xie, W. 2023. OvarNet: Towards Open-vocabulary Object Attribute Recognition. *arXiv preprint arXiv:2301.09506*.

Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1383–1391.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, L.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; Chang, K.-W.; and Gao, J. 2022. Grounded Language-Image Pre-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

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

Pham, K.; Kafle, K.; Lin, Z.; Ding, Z.; Cohen, S.; Tran, Q.; and Shrivastava, A. 2021. Learning To Predict Visual Attributes in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13018–13028.

Pham, K.; Kafle, K.; Lin, Z.; Ding, Z.; Cohen, S.; Tran, Q.; and Shrivastava, A. 2022. Improving Closed and Open-Vocabulary Attribute Prediction Using Transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, 201–219. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Zareian, A.; Rosa, K.; Hu, H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. *CVPR 2021*.

Zhao, S.; Zhang, Z.; Schulter, S.; Zhao, L.; Kumar, V.; Stathopoulos, A.; Chandraker, M.; and Metaxas, D. 2022. Exploiting Unlabeled Data with Vision and Language Models for Object Detection. *ECCV 2022*.

Zhou, X.; Girdhar, R.; Joulin, A.; Krahenbuhl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. *ECCV 2022*.