Improving Open-World Object Detection through Richer Instance Representation using Vision Foundation Models

Minsik Jeon*[1], Sunoh Lee*[1], Junwon Seo[2]

***Abstract*— While humans naturally identify novel objects and understand their relationships, deep learning-based object detectors struggle to detect and relate objects that are not observed during training. To overcome this issue, Open World Object Detection (OWOD) has been introduced to enable models to detect unknown objects in open-world scenarios. However, OWOD methods fail to capture the finer relationships between detected objects, which are crucial for comprehensive scene understanding and applications. In this paper, we propose a method to train an object detector that can both detect novel objects and extract semantically rich features in open-world conditions by leveraging the knowledge of Vision Foundation Models (VFM). We first utilize the semantic masks from the Segment Anything Model to supervise the box regression of unknown objects, ensuring accurate localization. By transferring the instance-wise similarities obtained from the VFM features to the detector's instance embeddings, our method then learns a semantically rich feature space of these embeddings. Extensive experiments show that our method learns a robust and generalizable feature space, additionally increasing the detector's applicability to tasks such as open-world tracking.**

I. INTRODUCTION

Humans can identify novel objects and associate them with similar objects based on their attributes, achieving a comprehensive scene understanding. While deep learning-based methods have improved perception, object detectors still struggle to detect out-of-distribution objects. To perform robustly in real-world scenarios, detectors should be able to detect unexpected objects and grasp the semantic relationships between them. Recently, Open World Object Detection (OWOD) [1] has been introduced to enhance the detectors' generalizability by enabling the detection of unknown objects not labeled in the training set.

While OWOD aims to learn a generalized understanding of *object* across diverse categories, they often overlook the finer relationships between detected objects. Such relationships are critical for a comprehensive understanding of a scene, particularly in open-world applications. For example, tasks such as object tracking [2]–[4] and class discovery [5], [6] rely on feature similarity among proposals for association and obstacle identification. However, existing approaches that leverage object features from detectors fail to capture detailed, meaningful information [6]–[8].

This work was supported by the Agency For Defense Development Grant funded by the Korean Government in 2024.

[1]Agency for Defense Development, Daejeon 34186, Republic of Korea. {mikejeon001123,sunoh0131}@gmail.com

[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. junwonse@andrew.cmu.edu

*These authors contributed equally to this work.

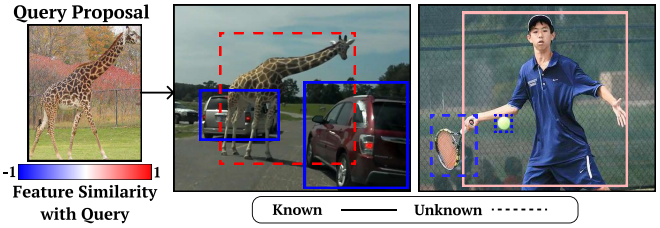

Fig. 1: We propose a method for training an object detector that can accurately detect unknown objects and extract semantically rich features in an open-world setting, thereby effectively capturing the relationships between proposals.

Since existing OWOD methods lack supervision to learn rich features of unknown objects, a few studies have tried to enhance the feature quality by self-supervised learning using unknown proposals [5], [6], [9]. However, the inaccurate proposals from detectors hinder the methods of learning robust features. Moreover, these methods focus on learning representations of unknown objects present in the training set, which we termed *known-unknown*s. In real-world deployment, systems may also encounter *unknown-unknown* objects that were never observed during training [10]. The feature space should be able to embed both of them properly based on their semantics.

We propose a method for training an object detector that can detect unknown objects and extract semantically rich features in an open-world setting by leveraging the knowledge of the Vision Foundation Model (VFM). Specifically, segmentation masks from Segment Anything Model (SAM) [11] are utilized to supervise the regression of bounding boxes for unknown objects. Moreover, the detector's instance features are enhanced by distilling the similarity between instances obtained from VFM's pixel-level features [12] to the detector. This distillation is performed using a relaxed contrastive loss [13], which provides a rich supervisory signal that enables the learning of a generalizable feature space.

Through extensive experiments, we demonstrate that our method significantly improves the unknown detection performance and the feature embedding quality. Evaluation on a novel benchmark shows the generalizability of our method to embed both *known-unknown* and *unknown-unknown*s. Furthermore, we show that learning semantically rich features benefits downstream tasks, including open-world tracking.

II. RELATED WORKS

A. Open World Object Detection and Discovery

Open World Object Detection (OWOD) addresses the problem of detecting objects in open-world scenarios [1].

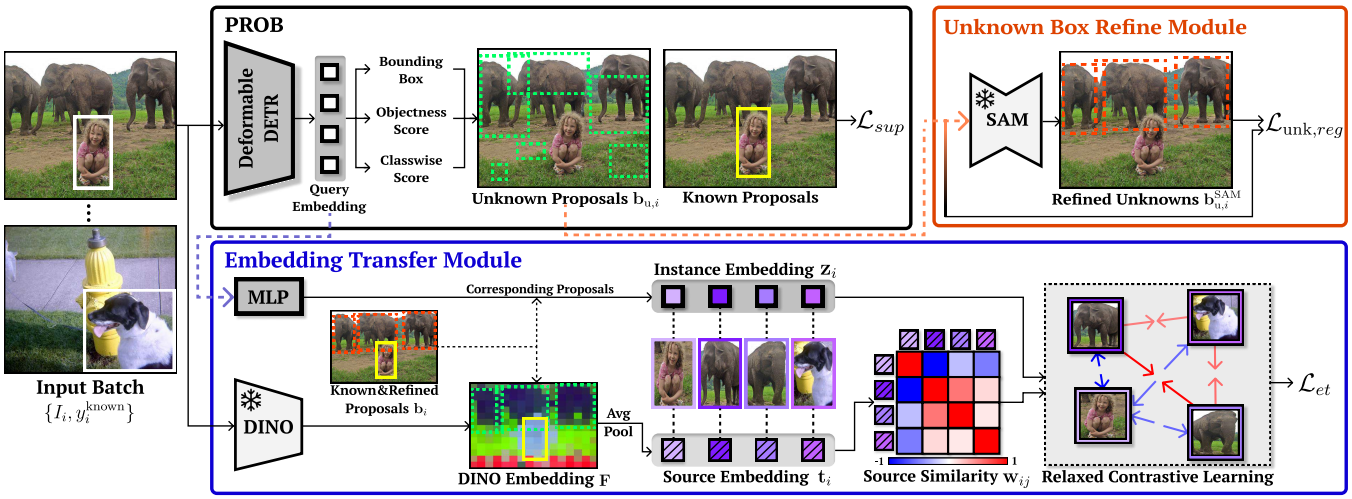

Fig. 2: Overview of our method. *PROB* [14] is adopted for an open-word object detector. The *Unknown Box Refine Module* enhances the regression of unknown proposals using the segmentation masks from SAM. The *Embedding Transfer Module* distills instance-wise relationships obtained from DINOv2's rich feature space to detector's instance embeddings.

These methods use knowledge from labeled objects to acquire an overall understanding of *object* [1], [14]–[19], by directly classifying unknown proposals [1], [15], or learning a continuous objectness score [14]. However, as they focus on learning a generalized concept of objects, distinguishing between different unknown objects remains challenging.

For a comprehensive scene understanding and OWOD applications, detailed relationships between detected objects are essential. Some methods attempt this through clustering the detector's features [7], [8], but the features often lack meaningful information. To enrich the instance-wise features, other methods employed self-supervised learning [13], [20]–[23]. They applied contrastive learning between detected unknown proposals [5], [6], [9], [22], [24], [25]. However, the inaccurate proposals of detectors hinder the learning of robust features [14], [15].

B. Foundation Model for Open World Object Detection

Efforts have been made to employ the foundation models in OWOD due to their strong generalizability to previously unseen scenarios [16], [26]–[28]. As these models are computationally demanding, several approaches focus on distilling knowledge from foundation models to detectors rather than using them directly [29], [30]. For instance, semantic masks obtained from SAM can enhance the box regression of unknown objects [31], or features from large language models are utilized to learn attributes of unknown objects [16]. Despite the advances, current approaches have not yet attempted to distill the instance-level features from foundation models into the detector's embeddings.

III. METHODS

Given a dataset $D = \{I_i, y_i\}_{i=1}^{M}$ consisting of images I_i and their corresponding labels y_i, only objects from K of the total P classes are labeled during the training phase. This known dataset $D_{\mathrm{known}} = \{I_i, y_i^{\mathrm{known}}\}_{i=1}^{M}$, where y_i^{known} denotes the labels excluding any unknown classes, is used to train the baseline open-world detector *PROB* [14]. *PROB* generates query embeddings for each predicted object, which are processed to compute box regression outputs, objectness scores, and classification into one of the K known or *background* classes. Proposals classified as known are matched with the ground truth via Hungarian matching to compute the supervised loss $\mathcal{L}_{sup}$ [14]. For *background* proposals, those with high objectness scores are considered as unknowns. The top-k proposalsare selected as pseudo-unknown objects, with their bounding boxes $\mathbf{B}_u = \{\mathbf{b}_{\mathrm{u},i}\}_{i=1}^{k}$.

A. Unknown Box Refine Module

PROB localizes objects using bounding box labels of known objects, making it difficult to accurately localize unknown objects. Inaccurate bounding boxes also negatively impact instance-wise feature learning, as they rely on predicted proposals during the learning process [5], [9]. Additionally, numerous false-positive proposals hinder meaningful feature extraction. To enhance detection accuracy and feature learning, more precise bounding boxes for unknown objects are essential.

We employ SAM's high-quality semantic masks to supervise the regression of unknown objects and obtain accurate boxes [11]. Given an input image I, each pseudo-unknown box $\mathbf{b}_{\mathrm{u},i}$ is used as a prompt for SAM to generate a refined unknown box $\mathbf{b}^{\mathrm{SAM}}_{\mathrm{u},i}$. Only the refined boxes that have an Intersection over Union (IoU) above a certain threshold κ with pseudo-unknown boxes are used for training, to remove false positive proposals and retain only accurate unknown boxes. The regression head is then trained using these refined boxes as the ground truth, with unknown regression loss calculated as follows:

$$\mathcal{L}_{\mathrm{unk},reg} = \frac{1}{|U|}\sum_{i\in U}\|\mathbf{b}_{\mathrm{u},i} - \mathbf{b}^{\mathrm{SAM}}_{\mathrm{u},i}\|_1 - \mathrm{GIoU}(\mathbf{b}_{\mathrm{u},i}, \mathbf{b}^{\mathrm{SAM}}_{\mathrm{u},i}), \quad (1)$$

where U is the set of indices corresponding to pseudo-unknowns with highly overlapping refined boxes, and GIoU denotes the Generalized Intersection over Union [32].

TABLE I: Quantitative results on M-OWODB and U-OWODB.

Method	M-OWODB					U-OWODB			
	K-mAP	U-Recall	Recall@1	DetRecall@1	DetRecall@2	U-Recall	Recall@1	DetRecall@1	DetRecall@2
PROB [14]	58.88	18.84	26.6	5.01	6.89	37.86	35.33	13.38	19.06
OSODD [9]	58.88	18.84	28.61	5.39	7.14	37.89	34.65	13.12	17.82
RNCDL [6]	57.53	18.94	11.15	2.11	3.08	37.59	16.97	6.38	10.82
Ours	**59.32**	**30.43**	**40.54**	**12.34**	**15.18**	**44.09**	**44.46**	**19.60**	**25.70**

B. Embedding Transfer Module

Previous self-supervised methods for learning semantically rich features often suffer from imprecise supervision, underscoring the need for a robust supervisory signal. Furthermore, the feature space should effectively embed *unknown-unknown*s by accurately capturing instance-wise relationships. To achieve this, we transfer the rich and generalizable features of the VFM to the detector using contrastive loss, weighted by pairwise similarities from the VFM. This ensures that similar proposals in the VFM feature space remain close in the detector's embeddings.

We utilize the pixel-level feature from DINOv2 [12] as our source embedding space. First, the DINOv2 embedding $\mathbf{F}$ for image I is acquired and then resized to the original image dimensions. To obtain proposal-wise features, the bounding boxes $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^{N}$ which include both the known and refined unknown boxes are utilized, where N represents the total number of such boxes. Mean pooling on $\mathbf{F}$ over each bounding box region yields the proposal-wise source embedding $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^{N}$.

To distill the inter-sample relations of source embeddings to the detector's embedding, the semantic similarity of pairwise source embeddings is computed from their Euclidean distance, using a Gaussian kernel [13] as follows:

$$\mathbf{w}_{ij} = \exp\left(-\frac{\|\mathbf{t}_i - \mathbf{t}_j\|_2^2}{\sigma}\right), \tag{2}$$

where σ is the kernel bandwidth. Then, the detector's instance embeddings $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N}$ with dimension d are obtained by forwarding the query embeddings corresponding to $\mathbf{B}$ through an MLP head. The instance embeddings are trained using the following relaxed contrastive loss [13] formulated as follows:

$$\mathcal{L}_{et} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\mathbf{w}_{ij}\mathbf{x}_{ij}^2 + (1-\mathbf{w}_{ij})\left[\delta - \mathbf{x}_{ij}\right]_+^2\right], \tag{3}$$

where $\mathbf{x}_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2$ denote the Euclidean distance between instance embeddings and δ is margin.

By minimizing this loss, the network adjusts the instance embeddings based on their semantic similarity as captured by DINOv2. In contrast to binary contrastive loss which neglects the degree of similarity between samples, relaxed contrastive loss leverages detailed inter-sample relations, thereby enhancing its generalizability. Our final loss function is as follows, where α and β are the weight coefficients:

$$\mathcal{L} = \mathcal{L}_{sup} + \alpha\mathcal{L}_{\text{unk},reg} + \beta\mathcal{L}_{et}. \tag{4}$$

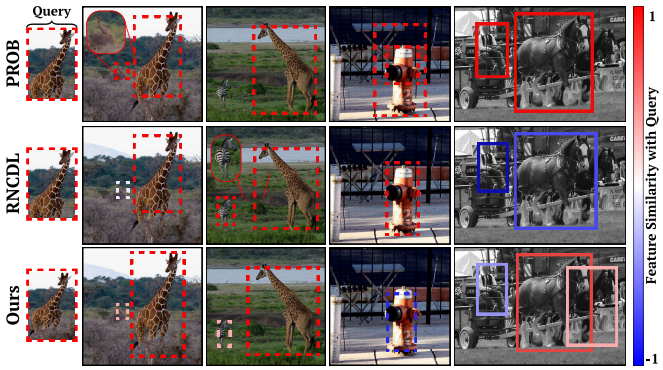

Fig. 3: Qualitative results of inter-proposal relationships. Proposals with high feature similarity to the query proposal are colored red, while highly dissimilar are colored blue.

IV. EXPERIMENTS

A. Experimental Setup

Datasets. We used the *Superclass-Mixed OWOD Benchmark* (M-OWODB) introduced in previous OWOD methods for evaluation [1], [14], [15]. This benchmark combines images from multiple datasets, including MS-COCO [33], PASCAL VOC2007 [34], and PASCAL VOC2012, and divides it into four non-overlapping tasks $\{T_1, \ldots, T_4\}$. We train the detector on images from task T_1 where only 20 classes are labeled. The remaining 60 classes in $\{T_2, T_3, T_4\}$ are left unlabeled and considered as unknowns. To better simulate real-world scenarios with unseen objects, we introduce the *Unknown-Unknown OWOD Benchmark* (U-OWODB), where all images containing unknown classes are excluded from the training set. This prevents the detector from learning any specific features of unknowns, thus evaluating its ability to learn a generalizable feature space for truly unseen objects.

Implementation Details. We utilized *PROB* [14] for the open-world object detector, with query embedding dimension of $d = 256$. The training is conducted for 41 epochs with batch size 32, starting with only the $\mathcal{L}_{sup}$ using labels from known classes. The losses $\mathcal{L}_{\text{unk},reg}$ and $\mathcal{L}_{et}$ are applied after 36 and 15 epochs, respectively. For embedding transfer, ViT-L DINOv2 [12] is utilized with $k = 10$. The ViT-H SAM [11] is utilized, and four regular grid points from each proposal are sampled to use as prompts to SAM. The remaining hyperparameters are set as follows: $\kappa = 0.5$, $\sigma = 1.0$, $\delta = 1.0$, $\alpha = 0.1$, $\beta = 1.0$.

Evaluation Metrics. Detection quality for known objects is evaluated using mean Average Precision (mAP). For unknown objects, false positives become unreliable as not

TABLE II: Ablation studies for each module. Note that **R@1** indicates Recall@1 and **DetR@1** indicates DetRecall@1.

Module		M-OWODB			U-OWODB		
ET	BR	U-Recall	R@1	DetR@1	U-Recall	R@1	DetR@1
✗	✗	18.84	26.60	5.01	37.86	35.33	13.38
✓	✗	18.60	**41.73**	7.76	36.99	43.92	16.25
✗	✓	31.33	23.69	7.42	**46.12**	34.79	16.05
✓	✓	**31.61**	40.54	**12.34**	44.09	**44.46**	**19.60**

every object in the dataset is labeled, so we use Unknown-Recall (U-Recall) as a metric [2], [14], [15].

To assess the quality of instance embeddings, we compute Recall@K for the prediction with the highest overlap for each ground truth (GT) object. However, Recall@K alone is unreliable due to the absence of consideration for undetected object's embeddings. To resolve this issue, we introduce a new metric dubbed *Detection Recall* (DetRecall), which assumes that undetected GT objects have no corresponding samples among their K-nearest neighbors. By adjusting the denominator of Recall@K from the number of predictions to the number of GT objects, DetRecall is computed as follows:

$$\text{DetRecall@}K = \text{Recall@}K \times \text{U-Recall}. \tag{5}$$

Comparison Methods. The baseline method *PROB* [8], [14] uses the query embeddings from the detector. We also evaluated several methods that apply self-supervision to enhance feature quality. *OSODD* [9] employs augmentation-based self-supervision by generating multiple views of cropped proposal images to train a separate embedding network [22]. On the other hand, *RNCDL* directly optimizes the detector's features by clustering all the features based on their similarity and treating all embeddings within a cluster as positive pairs. All methods are implemented based on the *PROB* framework.

B. Experimental Results

Comparison under M-OWODB. The quantitative results are summarized in Table. I. By transferring rich and robust supervisory signals from VFM, our method outperforms other approaches in both feature embedding quality and unknown object detection. The substantial drop in *RNCDL*'s performance indicates that numerous false positives lead to misassigned clusters during self-supervised learning, resulting in ineffective features. Despite the presence of numerous false positives, our method successfully learns meaningful feature space and inter-proposal relations, as shown in Fig. 3.

Comparison under U-OWODB. Table I shows that our method successfully embeds the objects of *unknown-unknown* classes. Methods relying on self-supervision show a larger performance drop compared to M-OWODB, suggesting that rigid pair assignments limit feature space generalizability by focusing on *known-unknown* classes. In contrast, by distilling detailed instance-wise relationships from VFM, our method learns a more generalizable feature space.

Ablation Studies. We conduct ablation studies to validate the effectiveness of each component, as summarized in Table II. The *Embedding Transfer Module* (ET) improves the feature embedding quality, demonstrating that distilling VFM's instance-wise relationships enhances the detector's feature space. Additionally, the adoption of an *Unknown Box Refine Module* (BR) improves the accurate localization of boxes. Overall, combining both components achieves the highest performance in terms of DetRecall by learning the feature embedding from the accurate unknown boxes.

TABLE III: Quantitative results for the open-world tracking. **A-Accuracy** represent Association Accuracy [2].

Method	U-Recall	A-Accuracy	OWTA
PROB [14]	48.67	8.49	19.72
Ours	**54.38**	**9.41**	**22.18**

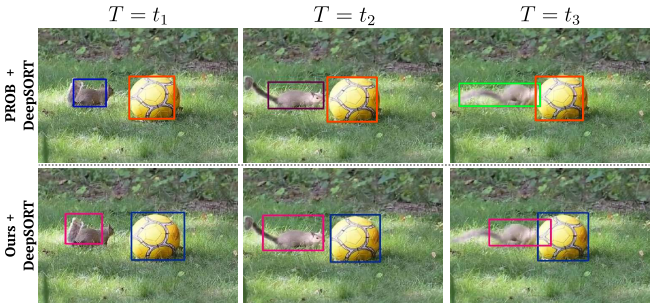

Fig. 4: Qualitative results in an open-world tracking scenario. Bounding box colors represent track IDs.

Application to Open-World Tracking. To validate the applicability of our method, we conduct open-world tracking [2] using the proposals and features from the detector. DeepSORT [4] is used as the tracker, which matches proposals across adjacent frames by combining the IoU with the similarity of the instance embeddings. The experiment is conducted on a subset of the TAO-OW dataset [2], which contains objects exhibiting dynamic movement, and evaluated on unknown objects using OWTA metric [2]. Table III shows our method improves both unknown object detection and frame-to-frame association accuracy. Fig. 4 illustrates an example tracking scenario. Using feature embeddings from *PROB* fails to properly associate the rapidly moving *squirrel*, even when it is successfully detected and meets the IoU threshold. In contrast, our method learns semantically rich instance embeddings and successfully matches squirrel detections.

V. CONCLUSION

This paper presents a method to learn object detectors to successfully detect unknown objects and extract semantically rich features in open-world scenarios. By leveraging the knowledge of Vision Foundation Models, the proposed method accurately localizes unknown objects and learns the nuanced relationships between instance features. We believe that learning such semantically rich and generalizable feature space can expand the applicability of open-world object detectors, as demonstrated by extensive experiments.

REFERENCES

[1] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5830–5840.

[2] Y. Liu, I. E. Zulfikar, J. Luiten, A. Dave, D. Ramanan, B. Leibe, A. Ošep, and L. Leal-Taixé, "Opening up open world tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 19 045–19 055.

[3] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu, "Quasi-dense similarity learning for multiple object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 164–173.

[4] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3645–3649.

[5] Z. Wu, Y. Lu, X. Chen, Z. Wu, L. Kang, and J. Yu, "Uc-owod: Unknown-classified open world object detection," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 193–210.

[6] V. Fomenko, I. Elezi, D. Ramanan, L. Leal-Taixé, and A. Osep, "Learning to discover and detect objects," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 8746–8759, 2022.

[7] S. S. Rambhatla, R. Chellappa, and A. Shrivastava, "The pursuit of knowledge: Discovering and localizing novel categories using dual memory," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9153–9163.

[8] T. L. Hayes, C. R. de Souza, N. Kim, J. Kim, R. Volpi, and D. Larlus, "Pandas: Prototype-based novel class discovery and detection," *arXiv preprint arXiv:2402.17420*, 2024.

[9] J. Zheng, W. Li, J. Hong, L. Petersson, and N. Barnes, "Towards open-set object detection and discovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3961–3970.

[10] A. Dhamija, M. Gunther, J. Ventura, and T. Boult, "The overlooked elephant of object detection: Open set," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1021–1030.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[12] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[13] S. Kim, D. Kim, M. Cho, and S. Kwak, "Embedding transfer with label relaxation for improved metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3967–3976.

[14] O. Zohar, K.-C. Wang, and S. Yeung, "Prob: Probabilistic objectness for open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 11 444–11 453.

[15] A. Gupta, S. Narayan, K. Joseph, S. Khan, F. S. Khan, and M. Shah, "Ow-detr: Open-world detection transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9235–9244.

[16] O. Zohar, A. Lozano, S. Goel, S. Yeung, and K.-C. Wang, "Open world object detection in the era of foundation models," *arXiv preprint arXiv:2312.05745*, 2023.

[17] D. Kim, T.-Y. Lin, A. Angelova, I. S. Kweon, and W. Kuo, "Learning open-world object proposals without learning to classify," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5453–5460, 2022.

[18] Z. Wang, Y. Li, X. Chen, S.-N. Lim, A. Torralba, H. Zhao, and S. Wang, "Detecting everything in the open world: Towards universal object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 11 433–11 443.

[19] Y. Wang, Z. Yue, X.-S. Hua, and H. Zhang, "Random boxes are open-world object detectors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6233–6243.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[21] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9729–9738.

[23] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 268–284.

[24] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," *arXiv preprint arXiv:1911.05371*, 2019.

[25] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 26, 2013.

[26] 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," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 728–755.

[27] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 901–16 911.

[28] Y. Cao, Z. Yihan, H. Xu, and D. Xu, "Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.

[29] N. G. Faber, S. S. M. Ziabari, and F. K. Najadasl, "Leveraging foundation models via knowledge distillation in multi-object tracking: Distilling dinov2 features to fairmot," *arXiv preprint arXiv:2407.18288*, 2024.

[30] S. Li, L. Ke, M. Danelljan, L. Piccinelli, M. Segu, L. Van Gool, and F. Yu, "Matching anything by segmenting anything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 18 963–18 973.

[31] Y. He, W. Chen, Y. Tan, and S. Wang, "Usd: Unknown sensitive detector empowered by decoupled objectness and segment anything model," *arXiv preprint arXiv:2306.02275*, 2023.

[32] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 658–666.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[34] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.