This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Annealing-based Label-Transfer Learning for Open World Object Detection

Yuqing Ma[1], Hainan Li[4], Zhange Zhang[1], Jinyang Guo[1],
Shanghang Zhang[3], Ruihao Gong[1], Xianglong Liu[1,2,4*]
[1] SKLSDE Lab, Beihang University, [2] Zhongguancun Laboratory,
[3] National Key Laboratory for Multimedia Information Processing, Peking University,
[4] Institute of Data Space, Hefei Comprehensive National Science Center

Abstract

Open world object detection (OWOD) has attracted extensive attention due to its practicability in the real world. Previous OWOD works manually designed unknown-discover strategies to select unknown proposals from the background, suffering from uncertainties without appropriate priors. In this paper, we claim the learning of object detection could be seen as an object-level feature-entanglement process, where unknown traits are propagated to the known proposals through convolutional operations and could be distilled to benefit unknown recognition without manual selection. Therefore, we propose a simple yet effective Annealing-based Label-Transfer framework, which sufficiently explores the known proposals to alleviate the uncertainties. Specifically, a Label-Transfer Learning paradigm is introduced to decouple the known and unknown features, while a Sawtooth Annealing Scheduling strategy is further employed to rebuild the decision boundaries of the known and unknown classes, thus promoting both known and unknown recognition. Moreover, previous OWOD works neglected the trade-off of known and unknown performance, and we thus introduce a metric called Equilibrium Index to comprehensively evaluate the effectiveness of the OWOD models. To the best of our knowledge, this is the first OWOD work without manual unknown selection. Extensive experiments conducted on the common-used benchmark validate that our model achieves superior detection performance (200% unknown mAP improvement with the even higher known detection performance) compared to other state-of-the-art methods. Our code is available at https://github.com/DIG-Beihang/ALLOW.git.

1. Introduction

Traditional object detection models [7, 19, 34, 35] are under an ideal closed world assumption, meaning the detected classes must be labeled and given during training. However, an object detection system is highly likely to encounter unknown objects that do not appear in the training phase. To address the above problem, Open World Object Detection (OWOD) has been pioneeringly proposed in [17], where the detection models should identify both known (annotated) and unknown (unannotated) categories and incrementally learn the unknown classes once their annotations are given.

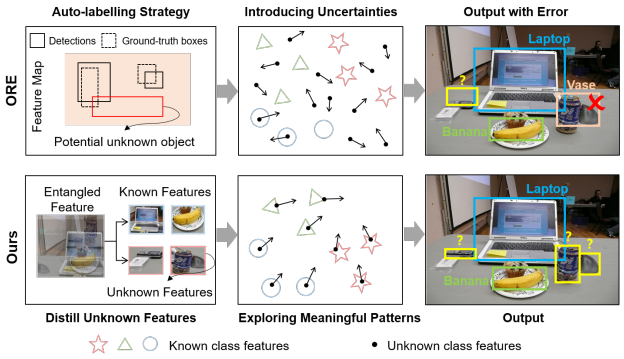


Figure 1. Comparison of the ORE [17] and Ours. ORE designed an auto-labelling strategy to select potential unknown objects from the background region, introducing uncertainties. In contrast, we view object detection as an object-level feature entanglement and explore unknown information from all the known proposals.

As shown in Figure 1, to accomplish unknown identification, previous OWOD works manually designed complex unknown-discover strategies to select specific regions from the background as unknown proposals, which will introduce uncertainties in the open world learning. For example, ORE [17] adopted an auto-labelling strategy in RPN, which selected top-k background region proposals sorted by their objectness scores as unknown objects. OW-DETR [10] further designed an attention-driven pseudo-labeling for selecting query boxes with high attention scores but not matched with known class boxes. Although these works [10, 17, 39, 42, 45] achieved promising performance and improved the unknown recall, these unknown-discover strategies are much likely to choose parts of the known pro-

*Corresponding author: Xianglong Liu, xlliu@buaa.edu.cn

posals or the authentic background regions, introducing too many uncertainties. These low-quality unknown proposals do not necessarily contain meaningful semantic patterns to accomplish the unknown recognition, and may also degrade the detection performance of known categories.

Different from previous works, we claim that the learning of object detection could be seen as an object-level feature-entanglement process through the local-connected convolution computation. That is to say, the features of each region inside the image could be propagated and closely entangled with each other, and thus a known proposal encompasses not only its own class-specific traits but also the propagated context information with potential unknown features. Empirical evidence also supports this point that even a closed world detection model is capable of extracting unknown features without unknown-discovery strategies, which will be further illustrated in Section 3.2.

Therefore, in this paper, to alleviate the uncertainties of manually selecting unknown proposals, we proposed a simple yet effective Annealing-based Label-Transfer framework for the open world object detection task. Our model sufficiently utilizes all the known proposals to distill unknown features of meaningful semantic patterns and promotes the collaborative learning of known and unknown classes. It follows a Label-Transfer Learning with a Sawtooth Annealing Scheduling.

Specifically, the Label-Transfer Learning disentangles the meaningful unknown features from all known proposals to accomplish unknown recognition, effectively reducing the uncertainties. Since it is non-trivial to directly decouple the features within a known proposal, the known label is transferred to the unknown class and thus decoupled into two classes, to guide the model decouple the learning of unknown and known traits. This strategy is intuitive and easy to implement, without complex and excessive computations in the forward process. Since most images simultaneously contain known and unknown objects, it could effectively distill the unknown traits of meaningful semantic patterns from the known proposals. Hence, it subtly avoids selecting unknown proposals and alleviates uncertainties.

Moreover, the Sawtooth Annealing Scheduling strategy meticulously adjusts the disentanglement degree to encourage the collaborative learning of both known and unknown classes. In our Label-Transfer Learning, since we cannot access the unknown supervision, it is hard to tell its entanglement extent and thus nontrivial to determine disentanglement level. If the disentanglement level is too small, we cannot well sufficiently mine the unknown traits. In contrast, if we over-disentangle the features, the informative known semantic will be dramatically destroyed and may adversely affect unknown identification either. Thus, it is critical to determine the suitable disentanglement level, to appropriately guide the unknown identification while maintaining the known detection performance. The concept "annealing" is borrowed from material science, where atoms within a solid material will redistribute under certain temperature control and reach an equilibrium state. Through our Sawtooth Annealing Scheduling strategy, the decision boundaries will be rebuilt with the consideration of both known and unknown classes, thus harmoniously promoting both the known and unknown learning to reach equilibrium.

The proposed Annealing-based Label-Transfer OWOD model is the first OWOD work without manually selecting unknown proposals, and achieves state-of-the-art performance. In addition, unlike previous models only reporting the unknown recall performance, we further present the unknown mAP performance and introduce a novel evaluation metric Equilibrium Index (EI) to comprehensively measure the unknown and known detection performance. In summary, our contributions are:

- We view object detection as an object-level feature entanglement and propose an Annealing-based Label-Transfer Learning, which is the first OWOD work without manually selecting unknown proposals.
- We introduce Label-Transfer Learning to disentangle the meaningful unknown traits from all known proposals, alleviating the uncertainties.
- We design a Sawtooth Annealing Scheduling strategy to adjust the disentanglement degree, ensuring the collaborative learning of known and unknown classes.
- A new OWOD evaluation metric Equilibrium Index is proposed to comprehensively evaluate known and unknown detection performance.
- Extensive experiments conducted on the commonly-used dataset prove the effectiveness of the proposed method. Specifically, we report a substantial increase in unknown mAP performance (200% gains compared to previous state-of-the-arts with an even better known detection performance).

2. Related work

The development of deep learning [3, 5, 9, 13, 15, 21, 22, 24, 47] has promoted the research of object detection where multiple objects should be recognized and localized inside an image. Traditional object detection models are under an ideal closed world assumption, which means the classes to be detected must be labeled and given in the training phase. However, it is highly possible that an object detection system will encounter the unknown objects that is not appeared in the training phase. To handle this problem, previous approaches have explored open set and open world settings.

Open set classification and detection. In the open set setting, the knowledge obtained through the training set is

incomplete, and thus the classifier during inference may encounter categories that do not appear in the training set. To meet this challenge, several works [8, 14, 18, 29, 33, 36] explored this task under a number of assumptions. The open set classification problem was first defined in [31] as a constrained minimization problem and extended to multi-class classifier by follow up works [16,32]. Bendale and Boult [2] proposed a method to identify the unknowns in the feature space of the model and use the OpenMax classifier to estimate the ensemble risk. Liu et al. [23] developed a metric learning framework that identifies invisible classes as unknown classes through a long-tail recognition setting for category coexistence. PROSER [46] encouraged the discrimination between known and unknown classes, neglecting the dynamic balance between the known and unknown instances. In addition, self-supervised learning [28] and unsupervised learning with reconstruction [43] method has been used in the recognition problem of open set.

Dhamija et al. [4] studied the open set object detection task and proposed the open set object detection protocol. Subsequent works [11, 26, 27] improved the detection performance of by measuring the uncertainties. OpenDet [12] also learns from known proposals from a feature-density perspective, but it manually designed an unknown-discover strategy that selects a few high-uncertainty known proposals to help improve the unknown identification.

Open world classification and detection. Unlike open set tasks that only focus on the identification of unknown classes, open world tasks also learn incrementally based on newly obtained category data. Bendale et al. [1] proposed the first open world image recognition model and presented a protocol for the evaluation of open world recognition systems. Xu et al. proposed a meta-learning method [41] to match a new sample with a dynamic set of known classes and identifies the new sample as an unknown class when it shows low similarities to all known classes. Some recent works [25, 25, 38] attempted to address the open world classification with long-tail distribution [44], few-shot learning [37] and zero-shot learning [40], respectively, to explore more complex scenes.

For open world detection, Joseph et al. [17] proposed the ORE method, in which an unknown object aware RPN is designed to endow the model with the capacity of detecting unknown objects. The work SA [42] utilizes semantic topology by defining a semantic centroid in feature space for each category, and pushing object instances close to their belonging centroids during the learning. OW-DETR [10] proposed an end-to-end framework comprising pseudo-labeling, novelty classification and object scoring. Wu et al. [39] defined the Unknown-Classified OWOD problem and designed a two-stage detector based on similarity and clustering to distinguish multiple different unknown classes. Zhao et al. [45] further proposed an auxiliary proposal advisor and a class-specific expelling classifier to improve the performance of the unknown detection.

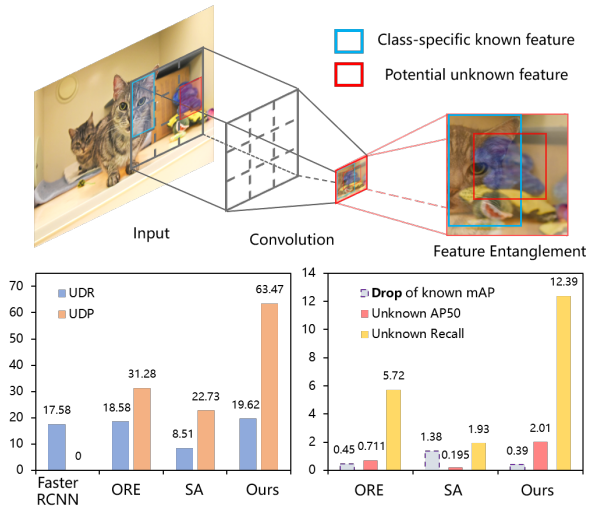


Figure 2. Object-level Feature Entanglement. Through local-connected convolutional operations of the multi-layer neural networks, potential unknown features could be propagated to known proposals and entangled with the known class-specific features. Thus, even closed world detector Faster RCNN could extract unknown features but mistakenly recognize them as known classes.

Previous methods [10, 17, 39, 42, 45] usually adopted complex unknown-discover strategies to deal with the unknown detection, and cannot always accurately select unknown proposals and thus introduce too many uncertainties, harming the learning of unknown objects and influencing the known classification as well. In contrast, our method only explores the unknown information from the known proposals through a reasonable disentanglement process, which improves the detection performance of the unknown objects while maintaining that of the known objects.

3. Method

Our Annealing-based Label Transferring model effectively improves the open world object detection performance through disentangling the known traits and unknown information from the known proposals, without specifically selecting unknown proposals. We will elaborate on the details of the proposed model in this section. First, we will introduce the OWOD formulation. And then, we will analyze the object-level feature entanglement in OWOD, and put forward our motivation. Finally, we will present the overall Annealing-based Label-Transfer OWOD framework, and further demonstrate our Label-Transfer Learning and Sawtooth Annealing Scheduling strategy.

3.1. Problem Formulation

The open world object detection contains T incremental tasks. In the t-th task, where $t \in \{1, ..., T\}$, the known class

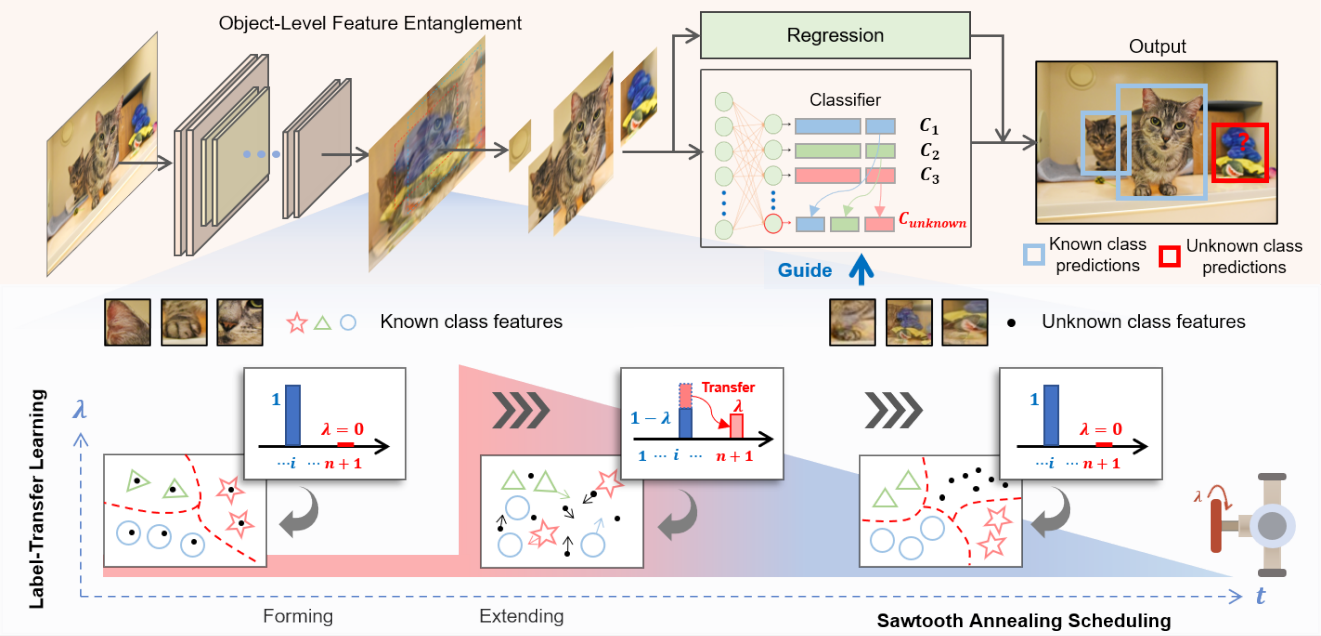


Figure 3. Overview of our Annealing-based Label-Transfer Framework. By viewing the learning of object detection as a process of object-level feature-entanglement process, the proposed model follows a Label-Transfer learning paradigm with a Sawtooth Annealing Scheduling strategy to accomplish the collaborative learning of known and unknown classes.

set and unknown class set are denoted as $\mathcal{K}^t$ and $\mathcal{U}^t$, where $\mathcal{K}^t \cap \mathcal{U}^t = \varnothing$. During the training phase, ONLY instances of known classes are assigned class labels and bounding box annotations. For example, an instance of the i-th known classes, denoted as $\mathbf{x}^{(\mathcal{K}_i^t)}$, and its corresponding class labels $\mathbf{y}^{(\mathcal{K}_i^t)}$ and bounding box coordinates $b = [x, y, w, h]$ would be given to guide the learning. However, both known class set $\mathcal{K}$ and unknown class set $\mathcal{U}$ exist during inference and should be identified by the OWOD models. The known instances should be accurately classified into their own categories, while the unknown instances are ought to be recognized as "unknown" class.

In the $(t+1)$-th task, unknown classes of interests $\bar{\mathcal{U}}^t \in \mathcal{U}^t$, are labeled and then integrated into known class set $\mathcal{K}^{t+1} = \mathcal{K}^t \cup \bar{\mathcal{U}}^t$. The current unknown class set $\mathcal{U}^{t+1} = \mathcal{U}^t \backslash \bar{\mathcal{U}}^t$. This process continues until $\mathcal{U}^T = \varnothing$. Furthermore, the OWOD models are required to incrementally identify previous known classes and current known classes, while recognizing the remained unknown classes as "unknown" class as well. It is worth noting that only a few instances of previous known classes in each task are stored for incremental learning, in order to save the computation and storage resources.

3.2. Motivation: Object-Level Feature Entanglement in Object Detection

Since multiple objects, including the unknown ones, are contained in the same image, their object-level features will be simultaneously perceived and entangled together through local-connected convolutional operations. We term this phenomenon as object-level feature entanglement. As shown in Figure 2, output by the multiple-layer network, the "cat" feature is entangled with the context (such as the "schoolbag" and "the toy"). Therefore, a proposal matched with a known groundtruth box contains both the known class-specific characteristics and the potential unknown information benefiting the unknown recognition. Simply classifying such an entangled known proposal as a known class will induce mis-classification of unknown instances.

Figure 2 also provides the empirical evidence to prove that. We adopt the evaluation metrics UDP and UDR used in [45]. UDR could illustrate the localization rate of unknown objects (even misclassified as the known ones), while the UDP is the rate of correct classification of the localized unknown objects. From the UDR performance, we can conclude that both close world and open world object detection models are capable of capturing the unknown characteristics due to the object-level feature entanglement. But with the guidance of classic one-hot supervisions, they tend to mistakenly identify the unknown instances as the known ones from the UDP results. Moreover, previous OWOD models merely reported the unknown recall results, veiling the fact that the unknown mAP is extremely low (less than 1%) due to the mis-classification. What is worse, the learning between the unknown and known classes is nearly a trade-off, where the unknown detection is improved with known mAP performance drop of the similar extent.

Therefore, we could simply distill the unknown informa-

tion with appropriate guidance from all the known proposals. Since most images simultaneously contain known and unknown instances, and we can extract unknown traits with meaningful semantic patterns while not conflicting with the known discriminative features. In this way, it could improve the unknown detection while maintaining the known detection performance, accomplishing the OWOD task.

3.3. Annealing-based Label-Transfer Framework

The overall architecture is shown in Figure 3. Our model could be based on any closed world detection architecture. The model follows an Label-Transfer Learning under the Sawtooth Annealing Scheduling strategy. The Label-Transfer Learning could guide the model to distinguish the known and potential unknown information to advance the unknown learning, while the Sawtooth Annealing Scheduling strategy could adjust the learning process through modifying the disentanglement degree, to reach an equilibrium between the unknown and known learning.

Following the Sawtooth Annealing Scheduling, the whole Label-Transfer Learning pipeline could be seen as two successive phases, namely the forming phase and the extending phase, respectively with different disentanglement degrees. In the forming phase, the model inclines to form entangled known proposals. Then, as the disentanglement degree is increased to learn unknown information, the known detection performance is adversely affected. Therefore, in the extending phase, we adjust the disentanglement degree changing to a sawtooth shape, to rebuild the known decision boundaries with the consideration of unknown objects. Finally, the decision boundaries of both unknown and known classes are formed, and the OWOD model is empowered with the unknown recognition ability and preserves the known identification accuracy. Besides, following the incremental learning approaches adopted in previous OWOD models [10, 17, 45], we store a balanced set of examples of the previous tasks and further finetune the current model to alleviate the forgetting problem in the incremental process.

3.3.1 Label-Transfer Learning

Since the object detection is an object-level feature-entanglement process, a known proposal may simultaneously contain discriminative class-specific traits and potential unknown information, which could be represented as:

$$\mathbf{x}^{(\mathcal{K}_i^t)} = \lambda \ddot{\mathbf{x}}^{(\mathcal{U}^t)} + (1-\lambda)\ddot{\mathbf{x}}^{(\mathcal{K}_i^t)} \tag{1}$$

where $\ddot{\mathbf{x}}$ indicates the uncoupled information of either i-th known class $\mathcal{K}_i^t$ or the unknown class $\mathcal{U}^t$ rather than a real instance. λ indicates the disentanglement degree.

It is hard to directly decouple the object-level feature, especially without explicit unknown labels. However, we can project these feature respectively to their corresponding

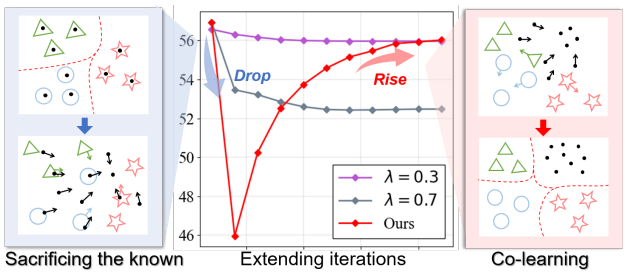


Figure 4. Comparison of Label-Transfer Learning with a fixed λ value and our Sawtooth Annealing Scheduling strategy on the validation set. With a fixed λ value, the known mAP declines until convergence, transferring the known semantic to learn unknown instances. In contrast, following the initial dramatic decline, the known mAP performance of the model with Sawtooth Annealing Scheduling subsequently rises to collaboratively learn the known and unknown information.

labels, to encourage the classifier to better recognize them. Despite assigning a feature-entangled proposal with a one-hot label that could strengthen its class-specific discrimination, it will suppress its generalized information. Thus, we could guide the classifier to recognize different features and accomplish the unknown and known classification.

To promote the unknown classification and guide the model to distinguish the unknown from the known ones, we correspondingly transfer the original one-hot label $\mathbf{y}^{(\mathcal{K}_i^t)}$ of the object to the unknown class. The label-transfer modification could be denoted as $\widetilde{\mathbf{y}}'$:

$$\widetilde{\mathbf{y}}' = \underbrace{\lambda(\mathbf{y}^{(\mathcal{U}^t)}|\mathbf{y}^{(\mathcal{K}_i^t)})}_{\text{Label-Transfer term}} + (1-\lambda)\mathbf{y}^{(\mathcal{K}_i^t)}, \tag{2}$$

where the first term represents that $\mathbf{y}^{(\mathcal{K}_i^t)}$ transfers to the unknown label $\mathbf{y}^{(\mathcal{U}^t)}$ with a tangling degree λ. The classification regularization with our Label-Transfer modifications can be written as follows:

$$\begin{aligned}\widetilde{\ell}_{RCNN}^{cls} &= -\sum_i \widetilde{\mathbf{y}}' \log \mathbf{x}^{(\mathcal{K}_i^t)} \\ &= -\sum_i (1-\lambda)\mathbf{y}^{(\mathcal{K}_i^t)} \log \mathbf{x}^{(\mathcal{K}_i^t)} - \sum_i \lambda \mathbf{y}^{(\mathcal{U}^t)} \log \mathbf{x}^{(\mathcal{K}_i^t)}.\end{aligned} \tag{3}$$

The Label-Transfer Learning could effectively and efficiently encourage the model to explore the unknown characteristics of the known proposals and thus generalize to unknown categories without ground-truth supervision.

3.3.2 Sawtooth Annealing Scheduling

Although the Label-Transfer Learning could improve the unknown detection, the disentanglement degree λ is hard to determine without specific unknown supervision. A large disentanglement degree will lead to a dramatic known per-

formance loss, while a small disentanglement degree cannot sufficiently guide the learning of unknown information.

Therefore, we further present a Sawtooth Annealing Scheduling strategy, ensuring the mutual advancement of unknown and known learning. The known identification performance is first sacrificed to improve the unknown detection, and then is slowly enhanced with consideration of unknown information, learning to form the decision boundaries of both known and unknown classes.

Specifically, the whole training follows an end-to-end pipeline, which could be seen as two successive phases, namely the forming phase and the extending phase. We denote p as the iteration number, and P_f indicates the total iteration of forming phase. The Sawtooth Annealing Scheduling could be formalized as:

$$\lambda = \begin{cases} 0, & p < P_f \\ 1 - \tau(p - P_f), & p \geq P_f \end{cases} \tag{4}$$

Where τ is a constant demonstrating the annealing speed, regulating the decoupling as the training iteration increases.

As we can see, in the forming phase, the model is actually trained in a closed world setting where parameters of unknown classifiers $\mathbf{W}_u$ are removed, in order to form informative but mixed semantics under the one-hot supervision of known categories:

$$\arg\min_{[\Theta, \mathbf{W}/\mathbf{W}_u]} \ell_f = \ell^{cls}_{RPN} + \ell^{reg}_{RPN} + \widetilde{\ell}^{cls}_{RCNN} + \ell^{reg}_{RCNN}, \tag{5}$$

Where $\mathbf{W}$ is the full classifier parameters, and $\mathbf{W}/\mathbf{W}_u$ means the full classifier parameters without the unknown classifier's. In addition, Θ represents the other parameters in the whole network except the classifiers.

In the extending phase, the disentanglement degree is abruptly changed to the peak to promote the unknown learning, and then gradually decreased to encourage the co-learning of both unknown and known categories. In this stage, we only fine-tune the classifier to avoid of destroying the informative features:

$$\arg\min_{\mathbf{W}} \ell_e = \widetilde{\ell}^{cls}_{RCNN}, \tag{6}$$

Figure 4 demonstrates the known mAP performance on the validation datasets and also proves our point. First of all, with a fixed disentanglement degree λ, the known mAP decreases all the time. The larger the λ is, the more obvious the mAP decrease is. It manifests that the model is sacrificing known performance to improve the unknown ones. However, with our annealing strategy, after the initial known mAP drop, it increases as the λ diminishes. Along with the known mAP boosts, the model rebuilds the decision boundaries with consideration of both known and unknown classes, to reach a final equilibrium.

4. Experiments

In this section, we perform comprehensive experiments and detailed analysis to demonstrate the effectiveness of the proposed method for open world object detection.

Datasets. According to the typical setup of OWOD, all classes from the training set are grouped in T incremental tasks. Following [17], we set T as 4, and adopt the Pascal-VOC [6] and MS-COCO [20] dataset. When learning the t_1 task, we treat classes and data from Pascal-VOC as training set, and the remaining 60 classes of MS-COCO are considered unknown. For subsequent tasks, the class division strategy is exactly the same as that in ORE. For evaluation, we use Pascal-VOC test set and MS-COCO validation set.

Evaluation metrics. Except for the common-used OWOD evaluation criteria, such as the mean average precision (mAP), recall, Wilderness Impact (WI) [4], Absolute Open-Set Error (A-OSE) [27], UDP and UDR, we also propose a new metric, namely the Equilibrium Index (EI). Previous work usually respectively report known and unknown detection performance, while our EI is designed to comprehensively evaluate the OWOD performance of both known and unknown classes.

We formally define the known and unknown mAP of the closed world baseline models as K_mAP_c and U_mAP_c (U_mAP_c=0). Correspondingly, the known and unknown mAP of the evaluated open world model is K_mAP_o and U_mAP_o. And the Equilibrium Index is defined as:

$$\begin{aligned} EI = \mathcal{I}(U\_mAP_o > 0)((K\_mAP_o - K\_mAP_c) \\ + \delta(U\_mAP_o - U\_mAP_c)), \end{aligned} \tag{7}$$

Where δ indicates the importance of unknown detection performance, and can be set as 1 to indicate the equal importance of known and unknown classes.

The first term of EI demonstrates the OWOD model should detect the unknown instance, while the second term demonstrates the weighted sum of mAP performance variations of both known and unknown classes. Ideally, both known and unknown mAP variations should be positive and improved by a large margin. However, normally known classes and unknown classes may be negatively correlated. Therefore, the goal of OWOD model is pursuing that the unknown mAP gains should be larger than the known mAP loss. Therefore, the larger the EI is, the OWOD performance is better. The worst case is that the model sacrifices too much known accuracy but improves limited unknown detection performance, where the EI is a negative value.

Implementation details. We implement our method based on two closed world detection models: (1) Faster RCNN [30] with ResNet-50 [13] backbone and the cosine classifier, and (2) DETR [3] following OW-DETR [10]. We simply take the best model in the forming phase as the baseline method for comparison. During training, the SGD optimizer was used and the batch size is set to 8. The peak of

Table 1. State-of-the-art comparison for OWOD according to traditional detection metrics. "K-" indicates the known classes, and "U-" represents the unknown classes. We present models using two closed world detection baselines, Faster RCNN and DETR, with their performance at the top as a reference. Our model achieves superior performance in terms of traditional evaluation metrics in most cases.

Task IDs (→)	Task t_1					Task t_2					Task t_3					Task t_4
	WI-0.8 (↓)	A-OSE (↓)	K-mAP (↑)	U-mAP (↑)	U-Recall (↑)	WI-0.8 (↓)	A-OSE (↓)	K-mAP (↑)	U-mAP (↑)	U-Recall (↑)	WI-0.8 (↓)	A-OSE (↓)	K-mAP (↑)	U-mAP (↑)	U-Recall (↑)	K-mAP (↑)
Faster RCNN [30]	0.0645	10502	56.94	0	0	0.0273	8653	41.56	0	0	0.0164	7345	32.41	0	0	27.03
ORE [17]	**0.0528**	11998	56.49	0.71	5.72	0.0315	9744	39.64	0.14	2.66	0.0209	7769	30.17	0.12	3.34	25.95
SA [42]	0.0563	23320	55.56	0.20	1.93	**0.0181**	16768	39.02	0.03	0.79	**0.0136**	**1428**	31.54	0.003	0.12	26.42
Ours-RCNN	0.0604	**8332**	**56.67**	**2.12**	**12.76**	0.0269	**9454**	**40.55**	**0.41**	**5.02**	0.0157	6635	**32.07**	**0.44**	**9.81**	**27.03**
DETR [47]	0.0600	57430	59.75	0	0	0.0245	27795	46.08	0	0	0.0187	17822	38.28	0	0	30.60
OW-DETR [10]	0.0599	**42331**	58.78	0.07	7.65	0.0319	25857	44.11	0.04	5.83	0.0220	18056	35.96	0.03	5.97	27.94
Ours-DETR	**0.0564**	46589	**59.34**	**4.86**	**13.56**	**0.0274**	**24709**	**45.58**	**0.65**	**10.04**	**0.0194**	**14952**	**37.97**	**0.39**	**14.30**	**30.60**

Table 2. State-of-the-art comparison for OWOD according to the newly proposed detection metrics, including UDR, UDP, and our Equilibrium Index (EI). Our model achieves superior performance on the newly-proposed evaluation metrics in most cases.

Task IDs (→)	Task t_1					Task t_2					Task t_3				
	UDR (↑)	UDP (↑)	EI($\delta=1$) (↑)	EI($\delta=2$) (↑)	EI($\delta=5$) (↑)	UDR (↑)	UDP (↑)	EI($\delta=1$) (↑)	EI($\delta=2$) (↑)	EI($\delta=5$) (↑)	UDR (↑)	UDP (↑)	EI($\delta=1$) (↑)	EI($\delta=2$) (↑)	EI($\delta=5$) (↑)
Faster RCNN [30]	17.58	0	0	0	0	16.32	0	0	0	0	24.69	0	0	0	0
ORE [17]	**18.58**	31.28	0.26	0.97	3.11	17.30	15.37	-1.79	-1.65	-1.23	23.67	14.95	-2.12	-2.00	-1.64
SA [42]	8.51	22.73	-1.18	-0.98	-0.41	5.74	13.83	-2.51	-2.48	-2.39	9.12	1.30	-0.87	-0.86	-0.86
Ours-RCNN	17.95	**71.08**	**1.85**	**3.97**	**10.33**	**17.62**	**28.49**	**-0.61**	**-0.20**	**1.04**	**23.78**	**41.25**	**0.10**	**0.54**	**1.86**
DETR [47]	20.74	0	0	0	0	14.41	0	0	0	0	34.48	0	0	0	0
OW-DETR [10]	18.31	41.77	-0.90	-0.83	-0.63	**16.24**	35.88	-1.94	-1.90	-1.78	**21.53**	27.72	-2.29	-2.25	-2.15
Ours-DETR	**18.47**	**73.42**	**4.45**	**9.31**	**23.89**	13.92	**72.15**	**0.15**	**0.80**	**2.75**	18.53	**77.19**	**0.08**	**0.47**	**1.64**

λ is set to 1, and the annealing speed τ is set to 5e-5. In the extending phase, the initial learning rate is set to 1e-4. The decay steps are set to (12000, 16000), and the learning rate is divided by 10 at each decay step.

4.1. State-of-the-art Comparison

Table 1 shows the comparison of our method respectively with the state-of-the-art OWOD methods according to traditional evaluation metrics, such as WI, A-OSE, mAP, and Recall. From Table 1, we have the following observations: our method outperforms other state-of-the-art OWOD methods in most cases. Specifically, based on the Faster RCNN backbone, the unknown mAP of our method is almost three times of that for the suboptimal model ORE with the better known mAP performance, while the unknown class recall of our method reaches 12.76 and is higher than ORE by 7.04, which is the best method in the existing works. Moreover, our method achieves even higher unknown performance based on DETR architecture. These phenomena clearly demonstrate the effectiveness of our Annealing-based Label-Transfer Learning. As novel known classes are added in the subsequent tasks, the recognition ability of unknown classes decreases, which is consistent with the performance of ORE. However, our method still outperforms the comparison methods by a large margin in U-mAP and unknown Recall metrics. And our known mAP performance on the newly-annotated and previous known classes is maintained. That is to say, our model is capable of promoting the collaborative learning of both known and unknown classes. It is worth noting that our method performs the same as the baseline in task t_4 because there are no unknown categories in t_4 in the OWOD setting, so we do not perform extending training.

Table 3. Ablation study. Our full model yields the superior performance, and each module contributes to the proposed model.

	EI($\delta=1$)	K-mAP	U-mAP	UDR	UDP
w/o LT	0	56.94	0	17.58	0
w/o SAS	-0.10	54.57	2.27	21.39	95.45
full FT	-20.36	34.95	1.63	23.01	99.78
with AL	1.61	56.54	2.01	19.72	65.42
Ours(SAS+LT)	1.85	56.67	2.12	17.95	71.08

Table 2 further illustrates the comparison with state-of-the-arts in terms of the recently-proposed UDR, UDP, and our new EI metric with δ respectively set to 1, 2 and 5. We can see a similar trend as shown in Table 1, where our model achieves superior performance. In certain cases, such as ORE in t_1 task or OW-DETR in t_2 task achieves slightly better UDR performance, illustrating their unknown-discovery strategy indeed provides some unknown proposals, but may introduce more background regions in terms of UDP and thus adversely affect the performance. Moreover, according to our new metric, we can find that our model obtains maximum U-mAP gains with minimum K-mAP loss, and thus achieves the best EI performance. However, the U-mAP of SA and DETR in the t_1 task are less than the K-mAP drop compared to their

closed world baseline, therefore their EI performance is below zero. In the incremental tasks, the EI performance of almost all other methods are less than 0, demonstrating that incrementally learning may bring obstacles for unknown learning. In sum, the above observations fully demonstrate that the newly proposed metric can reflect the OWOD performance and comprehensively measure both known and unknown detection performance.

4.2. Ablation Study

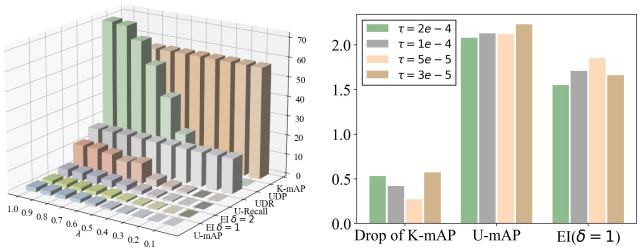
In Table 3, we investigate different components in our proposed framework. It respectively lists the performance of our model without Label-Transfer Learning (w/o LT) which degrades to the Faster RCNN model, our model without Sawtooth Annealing Scheduling (w/o SAS), our model with full Label-Transfer with fixed $\lambda = 1$ (full LT) where all known proposals will be projected into unknown classes, our model with auto-labelling strategies proposed by ORE [17] (with AL), and the full model (SAS+LT). From the table, we can observe that without the Label-Tansfer, the model is not capable of accomplishing unknown detection. Without SAS, we report the fixed $\lambda = 0.5$ which shares similar trends with that of other fixed λ values, while more experiments could be seen in the appendix. Consistent with our intuition, full Label-Transfer will greatly damage the known detection performance, which will in turn do harm to the unknown identification. Employing the unknown-discovery strategy AL cannot bring performance gains for our model. In conclusion, our full model achieves the best performance, and each module contributes to the proposed model.

4.3. Hyperparameter Analysis

Figure 5 respectively shows the hyperparameter analysis of the proposed method. We first investigate the peak of λ in our Sawtooth Annealing scheduling strategy. Then we will analysis the annealing speed τ.

Analysis on the peak of λ. As we can see from the Figure 5 (a), different peaks will not impact the K-mAP performance, but will influence the unknown one. Consistent with our intuition, the higher λ is, the higher unknown detection performance is. We can see the U-mAP, EI, U-Recall, and UDP shows an obvious drop as the peak of λ declines. However, the localization of unknown objects is not affected in terms of UDR. It proves our point that we should first increase the disentanglement degree λ to the maximum, and then decrease it to reach the equilibrium.

Analysis on the annealing speed. We conduct experiments which respectively sets τ to 2e-4, 1e-4, 5e-5 and 3e-5, and keep the peak value of λ unchanged. From Figure 5 (b), it is clearly that our model with different τ achieves similar detection performance with either slightly higher K-mAP and fractionally lower U-mAP or the opposite. Therefore, our model is insensitive to the annealing speed, proving the robustness and efficiency of the proposed model.

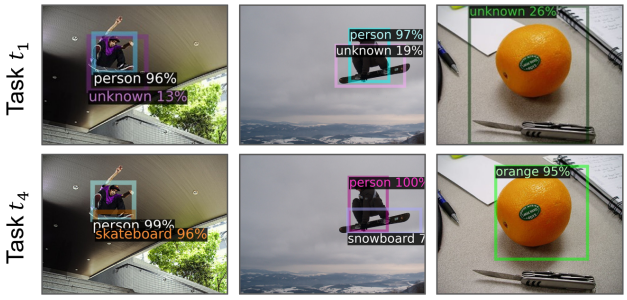


(a) Impact of different peaks of λ (b) Impact of annealing speed

Figure 5. Hyperparameter analysis. (a) The larger the λ is, the better OWOD detection performance the model achieves; (b) Our model is insensitive to τ and performs well with different choices.

4.4. Visualization

Figure 6 presents the visualization of the proposed method respectively in t_1 and t_4 tasks. In t_1 task, the model could detect the known classes and identify the unknown instances. In t_4 task, our model could incrementally learn the semantic class of all instances.



(a) (b) (c)

Figure 6. Visualization of the proposed model.

5. Conclusion

In this work, we propose a simple yet effective Annealing-based Label-Transfer framework for open world object detection. By viewing the learning of object detection as a process of object-level feature-entanglement process, the proposed model follows a Label-Transfer learning paradigm with a Sawtooth Annealing Scheduling strategy to accomplish the collaborative learning of known and unknown classes. The proposed framework is the first OWOD framework without requiring manual unknown-discover strategies. Extensive experiments on the common-used benchmark verify that our model shows state-of-the-art open world detection performance.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62206010 and 62022009.

References

[1] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015. 3

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 6

[4] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1021–1030, 2020. 3, 6

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[7] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021. 1

[8] Geli Fei and Bing Liu. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 506–514, 2016. 3

[9] Jinyang Guo, Dong Xu, and Guo Lu. Cbanet: Towards complexity and bitrate adaptive deep image compression using a single network. *arXiv preprint arXiv:2105.12386*, 2021. 2

[10] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9235–9244, 2022. 1, 3, 5, 6, 7

[11] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1031–1040, 2020. 3

[12] Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9591–9600, 2022. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 6

[14] Brian Heflin, Walter Scheirer, and Terrance E Boult. Detecting and classifying scars, marks, and tattoos found in the wild. In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 31–38. IEEE, 2012. 3

[15] J. Guo, W. Ouyang, and D. Xu. Channel pruning guided by classification loss and feature importance. In *AAAI*, 2020. 2

[16] Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *European Conference on Computer Vision*, pages 393–409. Springer, 2014. 3

[17] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021. 1, 3, 5, 6, 7, 8

[18] Fayin Li and Harry Wechsler. Open set face recognition using transduction. *IEEE transactions on pattern analysis and machine intelligence*, 27(11):1686–1697, 2005. 3

[19] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cbnet: A composite backbone network architecture for object detection. *IEEE Transactions on Image Processing*, 2022. 1

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[21] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *33rd AAAI Conference on Artificial Intelligence*, 2019. 2

[22] Aishan Liu, Xianglong Liu, Hang Yu, Chongzhi Zhang, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation. *IEEE Transactions on Image Processing*, 2021. 2

[23] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 3

[24] Yuqing Ma, Wei Liu, Shihao Bai, Qingyu Zhang, Aishan Liu, Weimin Chen, and Xianglong Liu. Few-shot visual learning with contextual memory and fine-grained calibration. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 811–817, 2021. 2

[25] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5222–5230, 2021. 3

[26] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2348–2354. IEEE, 2019. 3

[27] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3243–3249. IEEE, 2018. 3, 6

[28] Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11814–11823, 2020. 3

[29] Dimitrios A Pritsos and Efstathios Stamatatos. Open-set classification for automated genre identification. In *European Conference on Information Retrieval*, pages 207–217. Springer, 2013. 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6, 7

[31] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012. 3

[32] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014. 3

[33] Matthew D Scherreik and Brian D Rigling. Open set recognition for automatic target classification with rejection. *IEEE Transactions on Aerospace and Electronic Systems*, 52(2):632–642, 2016. 3

[34] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 1

[35] Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei, Yifu Ding, Bowei Jin, Hongping Zhi, Xianglong Liu, and Aishan Liu. Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21157–21167. IEEE, 2022. 1

[36] Rafael Vareto, Samira Silva, Filipe Costa, and William Robson Schwartz. Towards open-set face recognition using hashing functions. In *2017 IEEE international joint conference on biometrics (IJCB)*, pages 634–641. IEEE, 2017. 3

[37] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020. 3

[38] John Willes, James Harrison, Ali Harakeh, Chelsea Finn, Marco Pavone, and Steven Waslander. Bayesian embeddings for few-shot open world recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[39] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. Uc-owod: Unknown-classified open world object detection. *arXiv preprint arXiv:2207.11455*, 2022. 1, 3

[40] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591, 2017. 3

[41] Hu Xu, Bing Liu, Lei Shu, and P Yu. Open-world learning and application to product classification. In *The World Wide Web Conference*, pages 3413–3419, 2019. 3

[42] Shuo Yang, Peize Sun, Yi Jiang, Xiaobo Xia, Ruiheng Zhang, Zehuan Yuan, Changhu Wang, Ping Luo, and Min Xu. Objects in semantic topology. *arXiv preprint arXiv:2110.02687*, 2021. 1, 3, 7

[43] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4025, 2019. 3

[44] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 3

[45] Xiaowei Zhao, Xianglong Liu, Yifan Shen, Yuqing Ma, Yixuan Qiao, and Duorui Wang. Revisiting open world object detection. *arXiv preprint arXiv:2201.00471*, 2022. 1, 3, 4, 5

[46] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2021. 3

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 7