Detecting Open World Objects via Partial Attribute Assignment

Muli Yang[1] Gabriel James Goenawan[1] Huaiyuan Qin[1] Kai Han[2] Xi Peng[3] Yanhua Yang[4] Hongyuan Zhu[1†]
[1]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[2]The University of Hong Kong, Hong Kong SAR [3]Sichuan University, China [4]Xidian University, China
{yangml, goenawan, qinhy, zhuh}@i2r.a-star.edu.sg, kaihanx@hku.hk, pengxi@scu.edu.cn, yanhyang@xidian.edu.cn

Abstract

Despite being trained on massive data, today's vision foundation models still fall short in detecting open world objects. Apart from recognizing known objects from training, a successful Open World Object Detection (OWOD) system must also be able to detect unknown objects never seen before, without confusing them with the backgrounds. Unlike prevailing prior works that rely on probability models to learn "objectness", we focus on learning fine-grained, class-agnostic attributes, allowing the detection of both known and unknown objects in an explainable manner. In this paper, we propose Partial Attribute Assignment (PASS), aiming to automatically select and optimize a small, relevant subset of attributes from a large attribute pool. Specifically, we model attribute selection as a Partial Optimal Transport (POT) problem between known visual objects and the attribute pool, in which more relevant attributes signify more transported mass. PASS follows a curriculum schedule that progressively selects and optimizes a targeted subset of attributes during training, promoting stability and accuracy. Our method enjoys end-to-end optimization by minimizing the POT distance and the classification loss on known visual objects, demonstrating high training efficiency and superior OWOD performance among extensive experimental evaluations.[‡]

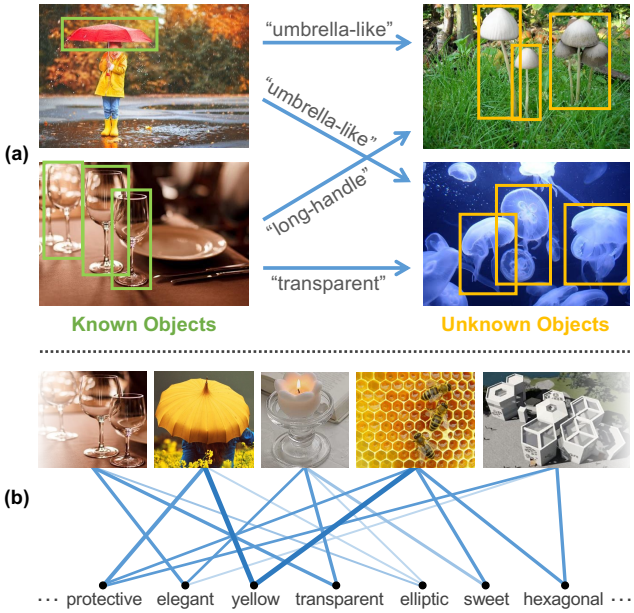

Figure 1. The role of attributes in Open World Object Detection (OWOD). **(a)** Unknown objects can be detected via shared visual/functional attributes with known objects. **(b)** Our method is able to automatically identify and optimize a small subset of relevant attributes from a large pool, effectively enabling the detection of both known and unknown objects.

1. Introduction

Deep learning and foundation models have emerged as critical AI technologies [1, 55], impacting various aspects of modern life. However, their success has been largely based on the closed-world assumption, which presumes a static and well-defined environment [76]. In practice, this assumption is difficult to uphold, as real-world conditions are dynamic and unpredictable, with unforeseen situations and unseen objects regularly appearing. Open World Object Detection (OWOD) [22] has emerged as a remedy to deliver reliable detection in the face of new environments with unknown objects. In addition to recognizing known objects that appeared in training, a successful OWOD system must also detect and continually learn previously unseen objects without mistakenly classifying them as backgrounds.

Recent works mainly use probability distribution estimation to model the "objectness" information, *i.e.*, the probability of a candidate proposal being an object rather than backgrounds [79]. In this way, OWOD can be decomposed into two sub-tasks that respectively focus on objectness estimation and known class recognition [54]. Although this approach has significantly improved the detection of unknown objects, it has certain limitations in explainability and ambiguity. On the one hand, the learned probabilistic

[†]Corresponding author. [‡]Code: https://github.com/muliyangm/PASS

model is unable to explain how the objects—both the known and the unknown—are detected. On the other hand, when background features resemble those of the unknown objects, the model can still misclassify them.

In this paper, we focus on detecting open world objects using fine-grained, class-agnostic attributes that are visually and/or functionally related to the known objects. These attributes can be described using natural languages with rich semantic information, and thus are naturally explainable and less ambiguous. The emphasis on intrinsic object attributes offers stronger resistance to confusion in scenarios where background and unknown object features overlap, making it particularly suitable for dynamic open-world environments. As illustrated in Fig. 1(a), attributes can be useful in detecting both known and unknown object classes, *e.g.*, the learned "`umbrella-like`" attribute from known classes can be used to detect the unknown `fungi`. Despite the promising potential of using attributes in OWOD, effectively curating a set of attributes that are closely related to the specific task from a large pool of attributes remains a significant challenge. To this end, some existing method [78] seeks a set of representative attributes that can effectively classify known objects with multistage processes—such as attribute selection and refinement—which are time-consuming and computationally expensive. More importantly, the separation of the selection and refinement processes may accumulate errors, leading to inaccuracies in the results of the attribute selection and refinement.

We propose to tackle this dilemma using Partial Attribute Assignment (PASS), an approach that progressively selects and optimizes a targeted subset of relevant attributes throughout the training process in an end-to-end manner. In particular, we model attribute selection as a Partial Optimal Transport (POT) problem [6] between known visual objects and the attribute pool, in which more relevant attributes signify more mass transported to the visual objects. Unlike conventional OT [56] that assumes identical mass in both distributions to be completely transported, POT allows the transport of only a targeted fraction of total mass from one distribution to another. As such, the attribute pool can be viewed as a collection of both in-distribution (ID) and out-of-distribution (OOD) attributes *w.r.t.* the visual objects, where we can use the visual object distribution as the baseline distribution to filter out redundant OOD attributes.

To achieve stable, accurate attribute selection and optimization simultaneously, PASS follows a curriculum learning [4] schedule that progressively targets and optimizes a subset of attributes throughout training. PASS ensures that each filtered subset is highly representative and provides broad coverage of the visual objects, by iteratively focusing on and refining a smaller subset of attributes guaranteed by the OT theory. As training progresses, the selected attributes reach a balanced state between representativeness and coverage, allowing the model to better adapt to both known and unknown objects. This process also reduces computational costs by gradually limiting the number of attributes introduced. Our method enjoys end-to-end optimization by minimizing both the POT distance and the classification loss for known visual objects. During inference, the learned attributes play a dual role: they support the recognition of known classes through the mapping relationships established during training, while also enabling the detection of unknown objects by leveraging the correlation between object proposals and the learned attributes. Extensive experiments demonstrate that our method significantly outperforms baseline methods, validating both the feasibility of attribute-based OWOD and the effectiveness of our approach in curating more accurate attributes for OWOD.

In a nutshell, our main contributions are

- We model the attribute curation issue in OWOD as a Partial Optimal Transport (POT) problem, enabling end-to-end attribute selection and optimization for effectively detecting both known and unknown objects;
- We propose a curriculum selection strategy by progressively optimizing a targeted subset of attributes with strong representativeness and broad coverage, benefiting both training stability and selection effectiveness;
- Extensive experiments demonstrate the absolute superiority of the proposed method in terms of both training efficiency and detection performance.

2. Related Work

Open World Object Detection. Open World Object Detection (OWOD) [22] enhances traditional object detectors by shifting away from closed-set assumptions, enabling dynamic adaptation to novel classes as encountered in real-world settings [20, 29, 30, 33, 34, 67, 74]. Existing works [27, 51, 61, 65, 69, 70] reduce the overlap between known, unknown classes and the background distributions using different ways, *e.g.*, pseudo-labeling [18, 38], objectness estimation [54, 79], and hierarchical modeling [13, 37]. OWOD is also related to similar tasks such as Class-Agnostic Object Detection [14, 21, 23, 36, 44, 45, 50, 59, 60] and Open-Vocabulary Object Detection (OVOD) [17, 25, 41, 42, 58, 71, 75]. Recent advances, such as FOMO [78] and a concurrent work UMB [62], extend the standard OVOD method to OWOD by utilizing vision and language foundation models. By identifying objects based on semantic attributes [26, 49, 53], they learn class-agnostic attributes to detect task-relevant unknown objects, yet require multistage attribute selection and refinement, which is time-consuming and computationally inefficient. In this paper, we propose an end-to-end attribute curation method that not only benefits the training efficiency but also largely improves the detection performance in both known and unknown classes.

Optimal Transport. Optimal Transport (OT), originally introduced by Monge [43] to minimize the cost of transporting items, has become a prominent tool in machine learning and computer vision for matching distributions across domains [46, 56]. OT has been widely applied in distribution-based tasks [8, 57, 73] such as generative models [2], clustering [19, 66], domain adaptation [11], object detection [16] and structural tasks such as sequence [9], graph [63], and image matching [32, 72]. To address OT's high computational demands, the Sinkhorn OT [12] offers a solution by approximating OT through an efficient iterative algorithm. Partial OT (POT), a variant of OT that transports only a portion of the mass, addresses cases where distributions only partially overlap, making it especially relevant in scenarios such as open-set domain adaptation [6, 7, 15, 35, 48, 64, 68]. The Sinkhorn algorithm has been extended to POT to further reduce computational costs [3]. In our work, we utilize POT to dynamically match attribute embeddings with visual embeddings of known object classes, selecting and optimizing the most representative attributes for effectively detecting both known and unknown objects.

3. Background

OWOD with Foundation Models. Open World Object Detection (OWOD) requires a model to detect and incrementally learn unknown objects. In particular, there are multiple stages/tasks for training and evaluation, indicated by t. In task t, the model can be trained on K^t known object classes from $\mathcal{K}^t = \{O_1^t, O_2^t, \dots, O_{K^t}^t\}$, and to be evaluated on a test set that contains both known and unknown classes. To achieve that, we use an additional O_0^t to denote the unknown classes, alongside the K^t known classes. In fact, O_0^t represents U^t classes of interest that are unknown to the model in the task t, *i.e.*, $\mathcal{U}^t = \{O_{K^t+1}^t, O_{K^t+2}^t, \dots, O_{K^t+U^t}^t\}$. After discovering the unknown object classes, the model can be updated with the knowledge of these new classes using annotations from an oracle (*e.g.*, a human annotator) in task $t+1$. As a result, the model is ready to detect $K^{t+1} = K^t + U^t$ known object classes in the test set, as well as additional unknown classes for the subsequent task cycle.

To push the boundary of OWOD in real-world applications, we follow a recent work [78] to make use of existing pretrained foundation models in open-vocabulary object detection [41, 71], *e.g.*, vision and language encoders, and adapt them to OWOD. The training follows a few-shot paradigm: for the z-shot, only z images per known class are given, with only one bounding box label and one object class label for each image. For each dataset, we use the attribute texts proposed by a Large Language Model (LLM), *i.e.*, GPT-3.5, by prompting with the known class names using the template described in recent works [40, 77, 78]. This will result in an attribute pool with numerous class-agnostic attributes visually and/or functionally related to the known objects, *e.g.*, `shape is straight`.

Optimal Transport Preliminary. Optimal Transport (OT) is a powerful tool for measuring the distance between two distributions. Here we review the discrete OT scenario that is more related to our framework and refer the readers to [12] for more details.

Assuming we have two sets of data points (features) $\mathbf{X} = \{\boldsymbol{x}_m\}_{m=1}^M$ and $\mathbf{Y} = \{\boldsymbol{y}_n\}_{n=1}^N$, their discrete distributions can be formulated as

$$\mathcal{X} = \sum_{m=1}^{M} \frac{1}{M} \delta_{\boldsymbol{x}_m}, \qquad \mathcal{Y} = \sum_{n=1}^{N} \frac{1}{N} \delta_{\boldsymbol{y}_n}, \tag{1}$$

where δ is the Dirac function and we use the uniform distribution to equally view each data point. The OT distance is defined as the smallest cost of transporting $\mathcal{X}$ to $\mathcal{Y}$:

$$d_{\text{OT}}(\mathcal{X}, \mathcal{Y}; \mathbf{C}) \triangleq \min_{\mathbf{T} \in \Pi(\mathcal{X}, \mathcal{Y})} \langle \mathbf{T}, \mathbf{C} \rangle_F, \tag{2}$$

in which $\mathbf{C} \in \mathbb{R}_+^{M \times N}$ is the cost matrix between $\mathbf{X}$ and $\mathbf{Y}$ where $C_{m,n} = c(\boldsymbol{x}_m, \boldsymbol{y}_n)$ is the transport cost from $\boldsymbol{x}_m$ to $\boldsymbol{y}_n$, with $c(\cdot)$ being the cost function (*e.g.*, the cosine distance), $\mathbf{T}$ is the transport plan to be solved, and $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. The optimization of OT distance satisfies two marginal constraints:

$$\Pi(\mathcal{X}, \mathcal{Y}) \triangleq \{\mathbf{T} \in \mathbb{R}_+^{M \times N} | \mathbf{T}\mathbf{1}_N = \mathcal{X}, \mathbf{T}^\top \mathbf{1}_M = \mathcal{Y}\}, \tag{3}$$

where $\mathbf{1}_N$ is an N-dimensional vector of all ones, and we slightly abuse $\mathcal{X}$, $\mathcal{Y}$ to denote $\mathcal{X} = \frac{1}{M}\mathbf{1}_M$, $\mathcal{Y} = \frac{1}{N}\mathbf{1}_N$ to reflect the discrete probability in Eq. (1).

4. Method

We propose Partial Attribute Assignment (PASS) to address the challenging Open World Object Detection (OWOD) task. PASS is composed of three modules shown in Fig. 2, which are introduced in separate subsections below.

4.1. Partial Attribute Assignment

As discussed above, now we have z-shot images of K known object classes (the superscript t is omitted for brevity) and N attribute texts in the attribute pool. Our goal is to select and optimize a small set of relevant attributes from the attribute pool, with the target attribute quantity denoted as $N' \in (0, N]$. The resulting N' attributes are to be used for effectively detecting both known and unknown objects.

Formally, we collect the visual object embeddings as $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_M] \in \mathbb{R}^{D \times M}$ and the attribute embeddings as $\mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_N] \in \mathbb{R}^{D \times N}$ in the D-dimensional embedding space using pretrained vision and text encoders [41, 47], where $M = z \cdot K$ is the number of known image patches obtained using bounding box labels,

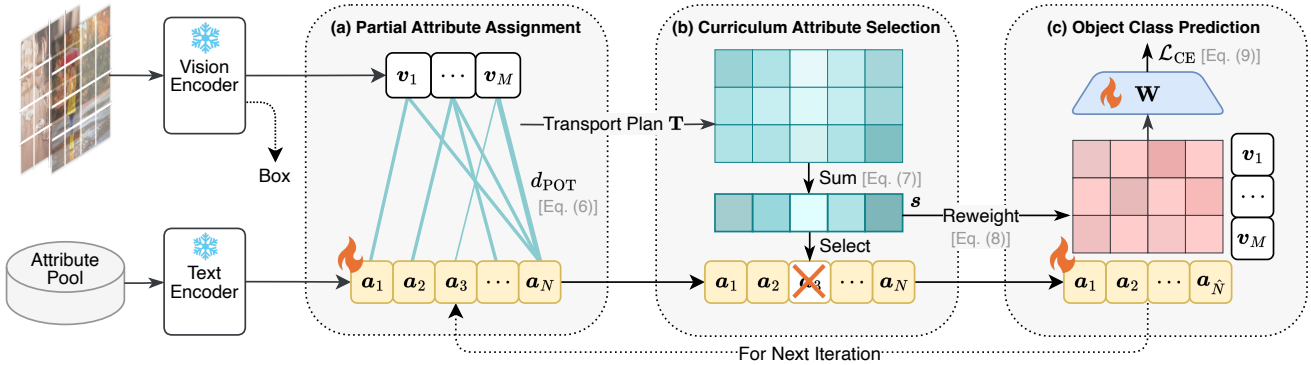

Figure 2. Overview of the proposed Partial Attribute Assignment (PASS) for OWOD. PASS is constructed upon pretrained vision and text encoders [41, 47] and is designed to identify and refine a set of relevant attributes from an attribute pool, which are essential for detecting both known and unknown object classes. **(a)** We employ Partial Optimal Transport (POT) to model the assignment problem between attributes and known visual objects (Sec. 4.1). **(b)** The resulting transport plan $\mathbf{T}$ is used to generate attribute in-distribution (ID) scores $\boldsymbol{s}$, which help in selecting the most relevant attributes for the specific task (Sec. 4.2). **(c)** These attribute ID scores are further used to reweight the selected attributes, which are subsequently applied to compute object class predictions via a learnable mapping matrix (Sec. 4.3).

and all embeddings are ℓ_2 normalized. Likewise, we can denote the discrete distributions of $\mathbf{V}$ and $\mathbf{A}$ as $\mathcal{V}$ and $\mathcal{A}$ following Eq. (1). The conventional OT assumes that the two distributions have the identical total probability mass, *i.e.*, $\|\mathcal{V}\|_1 = \|\mathcal{A}\|_1$, and that all the mass of $\mathcal{A}$ must be transported. Although this assumption works effectively for tasks that involve identical distributions, it faces challenges in our scenario, where only a subset of samples in $\mathcal{A}$ aligns with $\mathcal{V}$.

A workaround is to transport only the mass of in-distribution (ID) samples between $\mathcal{V}$ and $\mathcal{A}$ while retaining the out-of-distribution (OOD) mass within $\mathcal{A}$. This leads to the concept of Partial OT (POT) [6], which relaxes the assumption of equal total mass between $\mathcal{V}$ and $\mathcal{A}$. Instead, it allows for transporting a portion of the mass within the range $[0, \min(\|\mathcal{V}\|_1, \|\mathcal{A}\|_1)]$. To this end, we use the visual distribution $\mathcal{V}$ as a baseline distribution to filter out redundant outliers in $\mathcal{A}$. Accordingly, the visual distribution $\mathcal{V}$ and attribute distribution $\mathcal{A}$ are redefined as

$$\mathcal{V} = \sum_{m=1}^{M} \frac{1}{M} \delta_{\boldsymbol{v}_m}, \qquad \mathcal{A} = \sum_{n=1}^{N} \frac{\alpha}{N} \delta_{\boldsymbol{a}_n}, \tag{4}$$

where $\alpha = \frac{N}{N'} \geqslant 1$ is the attribute redundant rate reflecting the redundant mass in $\mathcal{A}$, which enables transporting only a subset of mass in $\mathcal{A}$. In this manner, we formulate the resulting problem as a POT problem with the following marginal constraints:

$$\Pi(\mathcal{V}, \mathcal{A}) \triangleq \{\mathbf{T} \in \mathbb{R}_{+}^{M \times N} | \mathbf{T}\mathbf{1}_N = \mathcal{V}, \mathbf{T}^\top \mathbf{1}_M \leqslant \mathcal{A}\}. \tag{5}$$

It is important to highlight that, unlike the traditional POT problem, which imposes inequality constraints on both distributions, Eq. (5) introduces equality constraints within $\mathcal{V}$ [48]. This modification explicitly aligns the attribute distribution with the visual distribution by redistributing weights among the attributes. Moreover, the concise constraints accelerate the convergence of POT computations. The POT distance is then denoted as

$$d_{\text{POT},\epsilon}(\mathcal{V}, \mathcal{A}; \mathbf{C}) \triangleq \min_{\mathbf{T} \in \Pi(\mathcal{V}, \mathcal{A})} \langle \mathbf{T}, \mathbf{C} \rangle_F - \epsilon h(\mathbf{T}), \tag{6}$$

where $\epsilon > 0$ is the regularization coefficient, and $h(\mathbf{T}) = -\sum_{m,n} T_{m,n} \ln(T_{m,n})$ is the entropic constraint to accelerate the optimization [12], so that the optimal $\mathbf{T}$ can be estimated in a few iterations. We defer the detailed optimization procedure of Eq. (6) to Supplementary Material.

4.2. Curriculum Attribute Selection

Upon the optimal transport plan $\mathbf{T}$ derived in Eq. (6), we can define an in-distribution (ID) score for each attribute as $\boldsymbol{s} = [s_1, s_2, \ldots, s_N]^\top$, by aggregating the mass transferred from one attribute to all visual objects:

$$s_n = \frac{N}{\alpha} \sum_{m=1}^{M} T_{m,n}, \tag{7}$$

where $s_n \in [0, 1]$ is the scaled mass transported that can be interpreted as the probability of the specific attribute being in-distribution to the visual objects in $\mathbf{V}$. The underlying rationale is that the transport cost among ID samples is relatively low. To minimize the total transport cost, ID attributes in $\mathcal{A}$ are allocated a greater proportion of transportation mass compared to the redundant OOD attributes, in accordance with the principles of POT.

With the ID scores $\boldsymbol{s}$ at hand, one can simply select N' attributes with the highest ID scores at a certain stage of training, *e.g.*, at the beginning or the end. However, this may suffer from either the inaccurate transport cost due to the modality gap, or the inaccurate mapping from attributes to known classes. In light of this, we devise an iterative attribute

selection strategy that follows the spirit of Curriculum Learning [4]. Instead of single-time selection, we divide it into η steps alongside the training, such that only a slightly smaller set of attributes are to be selected in each step, making the POT problem easier to solve, and thus benefit the training stability and attribute selection effectiveness.

Formally, α in Eq. (4) is now rewritten as $\alpha = (\frac{N}{N'})^{\frac{1}{\eta}}$. In each of the η selection steps, we select the $\frac{1}{\alpha}$ amount of the remaining attributes according to the indices of the top $\frac{1}{\alpha}$ ID scores $\boldsymbol{s}$. The detailed selection method is summarized in Lines 5 to 8 in Algorithm 1. We denote the resulting selected attribute embeddings as $\hat{\mathbf{A}} \in \mathbb{R}^{D\times N'}$, with the ID scores being $\boldsymbol{s} = [s_1, s_2, \dots, s_{N'}]^\top \in \mathbb{R}^{N'}$.

4.3. Object Class Prediction

After gaining the selected attributes in each step, we can learn a mapping matrix $\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_K] \in \mathbb{R}^{N'\times K}$ to assign N' attributes to K known object classes, with the prediction probability computed as

$$p(O_k|\boldsymbol{v}) = \frac{\exp(\boldsymbol{w}_k^\top \mathbf{A}'^\top \boldsymbol{v})}{\sum_{k'=1}^{K} \exp(\boldsymbol{w}_{k'}^\top \mathbf{A}'^\top \boldsymbol{v})}\,, \tag{8}$$

in which $\mathbf{A}' = [\boldsymbol{a}'_1, \boldsymbol{a}'_2, \dots, \boldsymbol{a}'_{N'}]$ is the element-wise multiplication between $\boldsymbol{s}$ and $\hat{\mathbf{A}}$ where $\boldsymbol{a}'_n = s_n \hat{\boldsymbol{a}}_n$. The inclusion of ID scores $\boldsymbol{s}$ constitutes a complementary "soft" filtering strategy by reweighting selected attributes, whose effectiveness is clearly shown in Sec. 5.2.

After acquiring the class predictions, we can use a standard cross-entropy loss to supervise the optimization of both $\mathbf{W}$ and $\hat{\mathbf{A}}$:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} l_{m,k} \log p(O_k|\boldsymbol{v}_m)\,, \tag{9}$$

where $\boldsymbol{l}_m$ is the one-hot label vector for $\boldsymbol{v}_m$.

4.4. Overall Objective

Training. Our method can be optimized using the below training objective in an end-to-end fashion:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda d_{\text{POT},\epsilon}\,, \tag{10}$$

in which $\lambda > 0$ is a trade-off coefficient to balance the effect of the POT loss in Eq. (6). Algorithm 1 summarizes the whole training procedure.

Inference. We follow the prior work [78] to detect if a test visual embedding $\boldsymbol{v}$ contains unknown objects by combining two probabilities: (a) *task relevance*—the probability of $\boldsymbol{v}$ being related to the optimized attribute embeddings (p_{ID}), and (b) *unknownness*—the probability of $\boldsymbol{v}$ not being from the known classes (p_{OOD}), *i.e.*, $p_{\text{unk}} = p_{\text{ID}} \cdot p_{\text{OOD}}$:

$$p_{\text{ID}} = \max_{n\in\{1,\dots,N'\}} \sigma(\boldsymbol{a}_n'^\top \boldsymbol{v})\,, \tag{11}$$

$$p_{\text{OOD}} = 1 - \max_{k\in\{1,\dots,K\}} p(O_k|\boldsymbol{v})\,, \tag{12}$$

Algorithm 1: Training Procedure of PASS for OWOD

Input: Visual object embeddings $\mathbf{V} \in \mathbb{R}^{D\times M}$ with the one-hot labels $\boldsymbol{l}$, initial attribute embeddings $\mathbf{A} \in \mathbb{R}^{D\times N}$, target attribute quantity N', selection steps η, total training epochs E

Output: Optimal selected attribute embeddings $\hat{\mathbf{A}}$, mapping matrix $\mathbf{W}$, attribute ID scores $\boldsymbol{s}$

1 **Initialize:** Selected attribute embeddings $\hat{\mathbf{A}} = \mathbf{A}$, selected attribute quantity $\hat{N} = N$, Randomly initialized $\mathbf{W} \in \mathbb{R}^{N\times K}$, attribute redundancy rate $\alpha = (\frac{N}{N'})^{\frac{1}{\eta}}$;
2 **for** e *in* $(1, 2, \dots, E)$ **do**
/* Partial Optimal Transport */
3 Get transport plan $\mathbf{T}$ and distance d_{POT} by solving the POT problem between $\mathbf{V}$ and $\hat{\mathbf{A}}$ using Eq. (6) and α;
4 Compute attribute ID scores as $\boldsymbol{s} = \frac{\hat{N}}{\alpha}\mathbf{T}^\top \mathbf{1}_M$;
/* Curriculum Attribute Selection */
5 **if** $e \bmod \lfloor \frac{E}{\eta+1} \rfloor = 0$ **then**
6 Update selected attribute quantity $\hat{N} \leftarrow \lfloor \frac{\hat{N}}{\alpha} \rfloor$;
7 Update $\hat{\mathbf{A}}$ and $\mathbf{W}$ by retaining only $\hat{N}$ columns in $\hat{\mathbf{A}}$ and $\hat{N}$ rows in $\mathbf{W}$ with top $\hat{N}$ indices in $\boldsymbol{s}$;
8 **end**
/* Object Class Prediction */
9 Calculate object class prediction $p(O_k|\boldsymbol{v})$ for $\mathbf{V}$ in Eq. (8) using $\mathbf{W}$, $\hat{\mathbf{A}}$, and $\boldsymbol{s}$;
10 Calculate loss $\mathcal{L}_{\text{CE}}$ in Eq. (9) using $p(O_k|\boldsymbol{v})$ and $\boldsymbol{l}$;
11 Get final loss $\mathcal{L}$ by combining $\mathcal{L}_{\text{CE}}$ and d_{POT} in Eq. (10);
12 Update $\mathbf{W}$ and $\hat{\mathbf{A}}$ using $\nabla\mathcal{L}$;
13 **end**

where $\sigma(\cdot)$ is the sigmoid function.

5. Experiment

In this section, we present the experimental evaluations of our proposed PASS. We start with the experimental setup, followed by the evaluation results and analysis.

5.1. Experimental Setup

Datasets. As the performance on traditional OWOD benchmarks [22]—which are based on common everyday object datasets like COCO [31]—is reported to be highly saturated [78], we use the Real-World Object Detection Benchmark established in [78] to evaluate the proposed method. This benchmark is designed for the few-shot or low-data setting, recognizing that most real-world applications cannot gather datasets on the scale of traditional benchmarks. Specifically, it includes five challenging real-world datasets as outlined below. (a) Aquatic [10] contains 637 underwater images of 7 different sea animals, providing useful data for marine-related applications. (b) Aerial [28] contains 10,000 aerial photos of 20 different structures such as stadiums and storage containers, supporting satellite imaging and intelligence analysis. (c) Game [10] contains 1575 game screen-

shots with 59 different avatars, serving as a benchmark for testing on synthetic data. (d) Medical [10] consists of 182 hand X-rays to identify 12 different bones, offering insights for medical applications such as detecting arthritis, fractures, and structural issues in hands. (e) Surgery [5] is composed of 1829 images of 13 different surgical instruments, captured under neurosurgical microscopes.

Evaluation Protocol. For each dataset, classes were evenly divided into two subsets: one containing the most common classes and the other containing the least common classes. The evaluation follows OWOD protocols and is conducted in two tasks. In Task 1, the model is provided with the most common classes as *known* (K), while the least common classes remain *unknown* (U), and the model is expected to detect both known and unknown object classes. In Task 2, in addition to the *previously known* (PK) classes, the least common classes are introduced as *currently known* (CK) classes. The model's performance is then evaluated on both PK and CK classes.

The mean average precision (mAP) is used to evaluate the performance in known classes (K-mAP). To gain deeper insights into the quality of continual learning in OWOD, mAP is further divided to respectively evaluate PK and CK classes (PK-mAP and CK-mAP) as discussed above. Additionally, we also report the mean average precision for unknown objects (U-mAP).

Implementation Details. Our approach is based on a frozen CLIP-pretrained OWL-ViT model (B/16 and L/14) [41], which has been fine-tuned for detection on a federated dataset combining Objects365 [52] and Visual Genome [24]. We leverage this frozen model, focusing on the construction of optimal attribute embeddings $\hat{\mathbf{A}}$ and a mapping matrix $\mathbf{W}$ to enhance the detection performance for both known and unknown object classes. For training, we use $\epsilon = 0.01$ in Eq. (6). For fair comparison with [78], N' is determined so that the average number of selected attributes per known class is 25. η is default to be 4 for all datasets. Other hyperparameters, including training epochs and learning rate, are searched based on the validation mAP on known classes following [78]. The experiments were carried out on an NVIDIA RTX A6000 GPU. More implementation details are provided in Supplementary Material.

5.2. Ablation Study

We ablate our proposed PASS to evaluate the effectiveness of each proposed module. As shown in Tab. 1, the ablation study primarily focuses on the POT distance loss (d_{POT}) in Eq. (6), the attribute selection method in Sec. 4.2, and the attribute ID score ($\boldsymbol{s}$) reweighting strategy in Eq. (8).

Effect of POT loss. The minimization of POT distance (d_{POT}) constitutes an important ingredient of our proposed PASS. It explicitly optimizes the attribute embeddings to align closely with the corresponding visual object embeddings through optimal assignment, effectively reducing the modality gap and improving the generalizability of attributes. This enables the model to detect both known and unknown object classes. As shown in Tab. 1, performance drops significantly without minimizing d_{POT} (setting $\lambda = 0$ in Eq. (10)), particularly in the detection of unknown objects. While the cross-entropy loss in Eq. (9) helps maintain performance on known objects, relying solely on the classification loss leads to poor generalization of the learned attributes for the unknown objects.

Dataset ($\rightarrow$)	**Aquatic**		**Surgery**	
	U	K	U	K
B/16 Backbone:				
PASS Full (Ours)	**5.2**	**43.4**	**14.3**	**15.6**
PASS w/o Minimizing d_{POT}	3.2	22.8	3.0	5.5
PASS w/o Attribute Selection	4.8	34.4	8.1	12.6
PASS w/o $\boldsymbol{s}$ Reweighting	4.5	41.3	10.6	11.9
L/14 Backbone:				
PASS Full (Ours)	**21.7**	**53.9**	**16.6**	**45.6**
PASS w/o Minimizing d_{POT}	10.1	18.2	8.0	14.1
PASS w/o Attribute Selection	15.5	34.1	15.2	43.6
PASS w/o $\boldsymbol{s}$ Reweighting	18.3	42.2	13.4	40.0

Table 1. Ablation study of PASS. We report U-mAP (U) and K-mAP (K) on two representative datasets: Aquatic and Surgery, with Task 1 evaluation in the 100-shot regime. **Best** results are highlighted in each column. More ablation study results can be found in Supplementary Material.

Effect of Attribute Selection. Effective attribute selection helps retain a streamlined set of in-distribution attributes that are most relevant to the specific task. While attribute selection has a less pronounced impact on performance compared to d_{POT}, since POT already performs an initial filtering by assigning lower weights to out-of-distribution attributes, it still provides benefits. Specifically, it helps reduce the noise in the attribute pool, which can enhance detection performance. Additionally, keeping a smaller attribute set is more computationally efficient. We provide further analysis on the advantage of the proposed curriculum attribute selection in Supplementary Material.

Effect of Attribute ID Score Reweighting. In addition to the "hard" attribute selection, reweighting object class predictions based on the attribute ID score—obtained from the POT transport plan—provides a complementary "soft" filtering strategy. This approach allows the final selected attributes to contribute differently, based on their relevance to the specific task, rather than being treated equally. As demonstrated in Tab. 1, this fine-grained use of selected attributes further enhances the detection performance.

Dataset (→)	Aquatic				Aerial				Game				Medical				Surgery				Overall			
Task ID (→)	Task 1		Task 2		Task 1		Task 2		Task 1		Task 2		Task 1		Task 2		Task 1		Task 2		Task 1		Task 2	
	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK
B/16 Backbone:																								
BASE-ZS+GT†	29.8	45.0	45.0	36.7	1.3	5.7	5.7	1.4	15.0	0.4	0.4	0.1	0.5	0.0	0.0	0.1	5.6	1.5	1.4	0.3	**10.4**	10.5	10.5	7.7
BASE-ZS	6.2	45.0	45.0	36.7	0.9	5.7	5.7	1.4	15.7	0.4	0.4	0.1	0.2	0.0	0.0	0.1	1.4	1.5	1.4	0.3	4.9	10.5	10.5	7.7
BASE-ZS+IN	26.5	45.1	45.1	36.7	1.9	5.7	5.7	1.4	2.4	0.3	0.3	0.0	0.6	0.0	0.0	0.1	1.7	1.4	1.0	0.3	6.6	10.5	10.4	7.7
BASE-ZS+LLM	24.7	45.1	45.1	36.5	1.4	5.7	5.7	1.4	15.1	0.4	0.4	0.1	0.6	0.0	0.0	0.1	8.9	1.5	1.3	0.3	10.2	10.5	10.5	7.7
BASE-FS	7.1	41.1	41.1	31.9	1.2	10.4	10.1	4.0	16.0	4.6	4.8	3.9	0.6	6.1	6.1	3.3	1.3	11.9	11.3	10.9	5.2	14.8	14.7	10.8
FOMO [78]	3.5	43.8	44.1	40.8	0.9	12.0	12.6	5.4	13.3	3.8	4.4	4.1	2.1	6.4	5.5	11.5	6.1	12.7	12.9	11.0	5.2	15.7	15.9	14.6
PASS (Ours)	5.2	43.4	43.2	46.6	1.9	14.0	16.0	7.0	21.5	10.0	7.7	9.0	4.9	8.4	6.8	12.1	14.3	15.6	13.1	14.7	9.6	**18.3**	**17.4**	**17.9**
Δ	**+1.7**	**-0.4**	**-0.9**	**+5.8**	**+1.0**	**+2.0**	**+3.4**	**+1.6**	**+8.2**	**+6.2**	**+3.3**	**+4.9**	**+2.8**	**+2.0**	**+1.3**	**+0.6**	**+8.2**	**+2.9**	**+0.2**	**+3.7**	**+4.4**	**+2.5**	**+1.5**	**+2.3**
L/14 Backbone:																								
BASE-ZS+GT†	34.8	36.0	36.0	42.3	1.0	7.9	7.2	0.8	12.4	0.9	0.8	0.3	2.4	0.2	0.2	0.3	2.4	0.2	2.6	1.3	10.6	9.0	9.4	9.0
BASE-ZS	0.7	35.9	36.0	42.3	9.1	8.2	7.2	0.8	6.8	0.9	0.8	0.3	0.0	0.2	0.2	0.3	3.6	2.9	2.6	1.3	4.1	9.6	9.4	9.0
BASE-ZS+IN	19.6	35.8	35.8	41.8	2.3	7.2	6.9	0.9	15.8	0.9	0.8	0.3	0.9	0.1	0.1	0.2	3.1	2.1	1.9	1.1	8.3	9.2	9.1	8.8
BASE-ZS+LLM	24.7	35.8	35.8	42.2	0.6	7.6	7.2	0.8	12.5	0.9	0.8	0.2	1.6	0.1	0.1	0.2	12.6	2.6	2.5	1.3	10.4	9.4	9.3	9.0
BASE-FS	2.4	43.6	42.9	42.8	9.7	23.7	21.9	13.0	8.2	10.4	10.2	13.4	1.1	23.2	21.7	24.2	3.6	26.0	25.0	7.4	5.0	25.4	24.3	20.2
FOMO [78]	18.2	50.1	48.1	47.1	6.0	25.3	23.7	16.0	30.4	10.7	9.9	11.2	9.4	21.8	19.9	34.6	12.0	29.0	28.9	8.5	15.2	27.4	26.1	23.5
PASS (Ours)	21.7	53.9	56.6	58.3	8.4	34.2	36.1	20.2	36.0	24.3	23.7	26.3	13.1	34.3	30.0	32.0	16.6	45.6	47.9	43.3	**19.1**	**38.5**	**38.9**	**36.0**
Δ	**+3.5**	**+3.8**	**+8.5**	**+11.2**	**+2.4**	**+8.9**	**+12.4**	**+4.2**	**+5.6**	**+13.6**	**+13.8**	**+15.1**	**+3.7**	**+12.5**	**+10.1**	**-2.6**	**+4.6**	**+16.6**	**+19.0**	**+34.8**	**+3.9**	**+11.1**	**+12.8**	**+12.5**

Table 2. OWOD results on the five real-world object detection datasets. The evaluation on each dataset is divided into two tasks, and we report U-, K-mAP for Task 1, and PK-, CK-mAP for Task 2, which are introduced in Sec. 5.1. BASE-FS, FOMO [78], and our proposed PASS are evaluated in the 100-shot regime, whereas the results of different few-shot regimes are provided in Supplementary Material. †GT baselines leverage ground-truth class names to detect unknown objects, functioning within the open-vocabulary object detection framework and providing an upper bound for text-conditioned (zero-shot) baselines. **Best** overall results are highlighted in each column.

5.3. Main Results

We perform extensive evaluations of our proposed PASS method, comparing it with existing state-of-the-art approaches (SOTAs). Additionally, we present qualitative results that showcase the detection outcomes along with the most relevant attributes used in the process. More experimental results can also be found in Supplementary Material.

Comparison with State of the Arts. We compare our proposed PASS with FOMO [78] and the strong baselines established therein, by adapting the Open-Vocabulary Object Detection (OVOD) methods [41] that also use foundation models to the OWOD setting. As listed in Tab. 2, BASE-ZS uses a generic "`object`" prompt for unknown objects [39], while BASE-ZS+IN utilizes all ImageNet class names (with the known classes removed) as the unknown classes proposals, and BASE-ZS+LLM employs class names predicted by an LLM (based on the known classes) for detecting potential unknown objects. On the other hand, BASE-ZS+GT uses the ground-truth class names for unknown objects, serving as the upper bound of the OVOD methods by assuming access to all unknown class names. BASE-FS utilizes image exemplars to produce vision-derived object embeddings, which are averaged per class to form class embeddings. Then a general prompt, such as "`a photo of an object`", is applied to help identify unknown objects.

As can be seen in Tab. 2, our proposed PASS consistently outperforms FOMO by a significant margin and achieves performance that is very close to the OVOD upper bound, even when using the B/16 backbone. In particular, our improvement is especially notable in known classes, as reflected in the K-, PK-, and CK-mAP metrics. This can be attributed to the robust attribute selection and optimization capabilities of our proposed PASS framework, which effectively curates a set of accurate and effective attributes to enhance object detection. On the other hand, PASS also demonstrates a significant performance boost in unknown classes as measured by the U-mAP metric. This suggests that the attributes learned by PASS are not only effective for known classes during training but also highly generalizable to unknown objects during evaluation.

Visualizations. For further comprehension of the effectiveness of our proposed PASS in detecting open world objects, we present several qualitative results for each dataset in Fig. 3. When comparing the detection results between FOMO (second row) and our proposed PASS (third row), PASS consistently demonstrates improved detection performance, with results that are more closely aligned with the ground truth (first row). In particular, FOMO sometimes incorrectly classifies known objects as unknown, or vice versa, suggesting that the learned attributes might not be sufficiently accurate to reliably characterize known objects. Furthermore, FOMO occasionally yields significantly fewer detection results in certain domains, such as medical datasets. This suggests that the learned attributes might not adequately capture the full range of potentially useful features necessary for detecting unknown objects. In contrast, PASS delivers

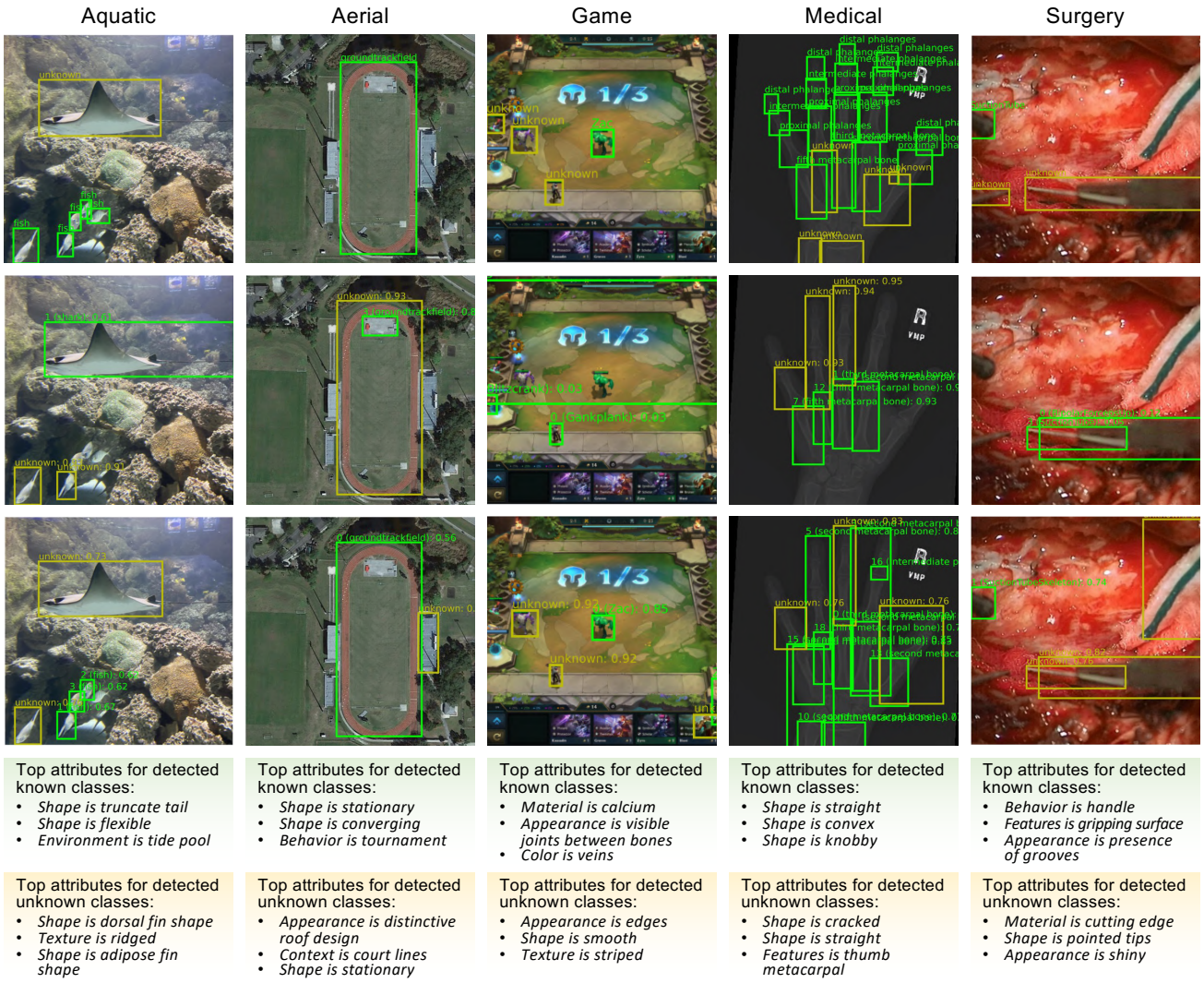

Figure 3. Qualitative results on the five real-world datasets. **First Row:** images in each dataset with ground truth bounding boxes and class names. **Second Row:** Detection results using FOMO [78]. **Third Row:** Detection results using our proposed PASS. We use green and yellow boxes to indicate known and unknown objects, respectively. **Bottom Rows:** The top three attributes that are activated for detecting known and unknown classes in each image using our proposed PASS.

superior detection results due to the more accurate attributes learned during training. These attributes not only provide a strong characterization of known object classes but also demonstrate high generalizability to unknown classes. Furthermore, as illustrated in the bottom two rows of Fig. 3, PASS effectively leverages meaningful attributes to identify both known and unknown objects. For example, it uses the attribute "`shape is dorsal fin shape`" to detect an unknown fish species recognized for its large fin.

6. Conclusion

In this paper, we propose an efficient end-to-end approach to select a compact set of representative attributes tailored to detect open-world objects, including both known and unknown classes. Specifically, we formulate the attribute selection and optimization process as a Partial Optimal Transport (POT) problem, where the most relevant attributes emerge with the highest transported mass. Additionally, we introduce a curriculum-based attribute selection strategy that gradually refines the attribute subset during training, significantly enhancing training stability and selection effectiveness. We evaluate our approach on challenging real-world datasets, where it achieves superior performance that surpasses current state-of-the-art methods by a large margin.

In future work, we aim to expand our method to incorporate hierarchical information within attribute taxonomies. This enhancement will allow for a more fine-grained understanding of the detection process by leveraging attributes at varying levels of granularity.

Acknowledgments. This research is supported by the EDB Space Technology Development Programme under Project S22-19016-STDP, the National Natural Science Foundation of China under Grant 62306251, 62171343, 62176171, U21B2040, the Hong Kong Research Grant Council – Early Career Scheme under Grant 27208022, the HKU Seed Fund for Basic Research, the Fundamental Research Funds for the Central Universities under Grant CJ202303, as well as the Sichuan Science and Technology Planning Project under Grant 24NSFTD0130. The authors sincerely thank the anonymous Reviewers and Area Chairs for their valuable feedback and constructive comments.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017. 3

[3] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015. 3

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, pages 41–48, 2009. 2, 5

[5] David Bouget, Rodrigo Benenson, Mohamed Omran, Laurent Riffaud, Bernt Schiele, and Pierre Jannin. Detecting surgical tools by modelling local appearance and global shape. *IEEE Transactions on Medical Imaging*, 34(12):2603–2617, 2015. 6

[6] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, pages 673–730, 2010. 2, 3, 4

[7] Wanxing Chang, Ye Shi, Hoang Tuan, and Jingya Wang. Unified optimal transport framework for universal domain adaptation. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 29512–29524, 2022. 3

[8] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[9] Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. Improving sequence-to-sequence learning via optimal transport. *arXiv preprint arXiv:1901.06283*, 2019. 3

[10] Floriana Ciaglia, Francesco Saverio Zuppichini, Paul Guerrie, Mark McQuade, and Jacob Solawetz. Roboflow 100: A rich, multi-domain object detection benchmark. *arXiv preprint arXiv:2211.13523*, 2022. 5, 6

[11] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Conference on Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2013. 3, 4

[13] Thang Doan, Xin Li, Sima Behpour, Wenbin He, Liang Gou, and Liu Ren. Hyp-ow: Exploiting hierarchical structure learning with hyperbolic distance enhances open world object detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1555–1563, 2024. 2

[14] Ruohuan Fang, Guansong Pang, and Xiao Bai. Simple image-level classification improves open-vocabulary object detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1716–1725, 2024. 2

[15] Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010. 3

[16] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 303–312, 2021. 3

[17] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[18] Akshita Gupta, Sanath Narayan, K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9235–9244, 2022. 2

[19] Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via wasserstein means. In *International Conference on Machine Learning (ICML)*, pages 1501–1509, 2017. 3

[20] Haiwen Huang, Andreas Geiger, and Dan Zhang. Good: Exploring geometric cues for detecting objects in an open world. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[21] Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Premkumar Natarajan. Class-agnostic object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 919–928, 2021. 2

[22] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5830–5840, 2021. 1, 2, 5

[23] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters (RA-L)*, 2022. 2

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017. 6

[25] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 2

[26] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 2

[27] Sunoh Lee, Minsik Jeon, Jihong Min, and Junwon Seo. Open-world object detection with instance representation learning. *arXiv preprint arXiv:2409.16073*, 2024. 2

[28] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 5

[29] Yiming Li, Yi Wang, Wenqian Wang, Dan Lin, Bingbing Li, and Kim-Hui Yap. Open world object detection: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[30] Wenteng Liang, Feng Xue, Yihao Liu, Guofeng Zhong, and Anlong Ming. Unknown sniffer for object detection: Don't turn a blind eye to unknown objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3230–3239, 2023. 2

[31] 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 (ECCV)*, pages 740–755, 2014. 5

[32] Benlin Liu, Yongming Rao, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Multi-proxy wasserstein classifier for image classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 8618–8626, 2021. 3

[33] Xuanyi Liu, Zhongqi Yue, and Xian-Sheng Hua. Proposal-level unsupervised domain adaptation for open world unbiased detector. *arXiv preprint arXiv:2311.02342*, 2023. 2

[34] Yanxin Long, Youpeng Wen, Jianhua Han, Hang Xu, Pengzhen Ren, Wei Zhang, Shen Zhao, and Xiaodan Liang. Capdet: Unifying dense captioning and open-world detection pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15233–15243, 2023. 2

[35] You-Wei Luo and Chuan-Xian Ren. Mot: Masked optimal transport for partial domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3531–3540, 2023. 3

[36] Shuailei Ma, Yuefeng Wang, Ying Wei, Peihao Chen, Zhixiang Ye, Jiaqi Fan, Enming Zhang, and Thomas H Li. Detecting the open-world objects with the help of the brain. *arXiv preprint arXiv:2303.11623*, 2023. 2

[37] Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Thomas H Li, Hongli Liu, and Fanbing Lv. Cat: Localization and identification cascade detection transformer for open-world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19681–19690, 2023. 2

[38] Yuqing Ma, Hainan Li, Zhange Zhang, Jinyang Guo, Shanghang Zhang, Ruihao Gong, and Xianglong Liu. Annealing-based label-transfer learning for open world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11454–11463, 2023. 2

[39] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 7

[40] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[41] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision (ECCV)*, pages 728–755, 2022. 2, 3, 4, 6, 7

[42] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 2

[43] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781. 3

[44] Sahal Shaji Mullappilly, Abhishek Singh Gehlot, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Hisham Cholakkal. Semi-supervised open-world object detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 4305–4314, 2024. 2

[45] David Pershouse, Feras Dayoub, Dimity Miller, and Niko Sünderhauf. Addressing the challenges of open-world object detection. *arXiv preprint arXiv:2303.14930*, 2023. 2

[46] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 3

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 3, 4

[48] Yilong Ren, Chuanwen Feng, Xike Xie, and S Kevin Zhou. Partial optimal transport based out-of-distribution detection for open-set semi-supervised learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. 3, 4

[49] Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605, 1975. 2

[50] Kuniaki Saito, Ping Hu, Trevor Darrell, and Kate Saenko. Learning to detect every thing in an open world. In *European Conference on Computer Vision (ECCV)*, pages 268–284, Cham, 2022. Springer Nature Switzerland. 2

[51] Hiran Sarkar, Vishal Chudasama, Naoyuki Onoe, Pankaj Wasnik, and Vineeth N Balasubramanian. Open-set object detection by aligning known class representations. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 219–228, 2024. 2

[52] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A

large-scale, high-quality dataset for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8430–8439, 2019. 6

[53] Yuyin Sun, Liefeng Bo, and Dieter Fox. Attribute based object identification. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2096–2103, 2013. 2

[54] Zhicheng Sun, Jinghan Li, and Yadong Mu. Exploring orthogonality in open world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17302–17312, 2024. 1, 2

[55] A Vaswani. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 1

[56] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2009. 2, 3

[57] Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3

[58] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11186–11196, 2023. 2

[59] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11433–11443, 2023. 2

[60] Yan Wu, Xiaowei Zhao, Yuqing Ma, Duorui Wang, and Xianglong Liu. Two-branch objectness-centric open world detection. In *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis*, pages 35–40, 2022. 2

[61] Xing Xi, Yangyang Huang, Jinhao Lin, and Ronghua Luo. Ktcn: Enhancing open-world object detection with knowledge transfer and class-awareness neutralization. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. 2

[62] Xing Xi, Yangyang Huang, Zhijie Zhong, and Ronghua Luo. Umb: Understanding model behavior for open-world object detection. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 74233–74261, 2024. 2

[63] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *International Conference on Machine Learning (ICML)*, pages 6932–6941, 2019. 3

[64] Renjun Xu, Pelen Liu, Yin Zhang, Fang Cai, Jindong Wang, Shuoying Liang, Heting Ying, and Jianwei Yin. Joint partial optimal transport for open set domain adaptation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2540–2546, 2020. 3

[65] Haosen Yang, Chuofan Ma, Bin Wen, Yi Jiang, Zehuan Yuan, and Xiatian Zhu. Recognize any regions. *arXiv preprint arXiv:2311.01373*, 2023. 2

[66] Muli Yang, Liancheng Wang, Cheng Deng, and Hanwang Zhang. Bootstrap your own prior: Towards distribution-agnostic novel class discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3459–3468, 2023. 3

[67] Muli Yang, Jie Yin, Yanan Gu, Cheng Deng, Hanwang Zhang, and Hongyuan Zhu. Consistent prompt tuning for generalized category discovery. *International Journal of Computer Vision*, pages 1–28, 2025. 2

[68] Yucheng Yang, Xiang Gu, and Jian Sun. Prototypical partial optimal transport for universal domain adaptation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 10852–10860, 2023. 3

[69] Mısra Yavuz and Fatma Güney. O1o: Grouping of known classes to identify unknown objects as odd-one-out. *arXiv preprint arXiv:2410.07514*, 2024. 2

[70] Jinan Yu, Liyan Ma, Zhenglin Li, Yan Peng, and Shaorong Xie. Open-world object detection via discriminative class prototype learning. In *IEEE International Conference on Image Processing (ICIP)*, pages 626–630, 2022. 2

[71] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14393–14402, 2021. 2, 3

[72] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12203–12213, 2020. 3

[73] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Differentiable earth mover's distance for few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5632–5648, 2022. 3

[74] Xiaowei Zhao, Yuqing Ma, Duorui Wang, Yifan Shen, Yixuan Qiao, and Xianglong Liu. Revisiting open world object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[75] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, 2022. 2

[76] Fei Zhu, Shijie Ma, Zhen Cheng, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. Open-world machine learning: A review and new outlooks. *arXiv preprint arXiv:2403.01759*, 2024. 1

[77] Orr Zohar, Shih-Cheng Huang, Kuan-Chieh Wang, and Serena Yeung. Lovm: Language-only vision model selection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3

[78] Orr Zohar, Alejandro Lozano, Shelly Goel, Serena Yeung, and Kuan-Chieh Wang. Open world object detection in the era of foundation models. *arXiv preprint arXiv:2312.05745*, 2023. 2, 3, 5, 6, 7, 8

[79] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11444–11453, 2023. 1, 2