Open-World Objectness Modeling Unifies Novel Object Detection

Shan Zhang[1,6], Yao Ni[2,6], Jinhao Du[3], Yuan Xue[4], Philip Torr[5], Piotr Koniusz[6,2,*], Anton van den Hengel[1,*]
[1]Australian Institute for Machine Learning, [2]Australian National University, [3]Peking University
[4]Ohio State University, [5]University of Oxford, [6]Data61♥CSIRO
[1]firstname.lastname@adelaide.edu.au, [6]piotr.koniusz@data61.csiro.au

Abstract

*The challenge in open-world object detection, similarly to few- and zero-shot learning, is to generalize beyond the class distribution of the training data. In this paper, we propose a general class-agnostic objectness measure to limit bias toward labeled samples. One issue in open-world detection is that previously unseen objects are often misclassified as known categories or filtered as background by classifiers. To prevent this, we explicitly model the joint distribution of objectness and category labels using variational approximation. However, without sufficient labeled data, minimizing the KL divergence between the estimated posterior and a static normal prior fails to converge. Our theoretical analysis identifies the root cause of this failure and motivates adopting a Gaussian prior with variance dynamically adapted to the estimated posterior as a surrogate. To further reduce misclassification, we introduce an energy-based margin loss that encourages unknown objects to move toward high-density regions of the distribution, thus reducing the uncertainty of unknown detections. Our **O**pen-**W**orld **OBJ**ectness modeling (OWOBJ) boosts novel object detection, especially in low-data regimes. OWOBJ is a flexible plugin that outperforms baselines in Open-World, Few-Shot, and zero-shot Open-Vocabulary Object Detection.*

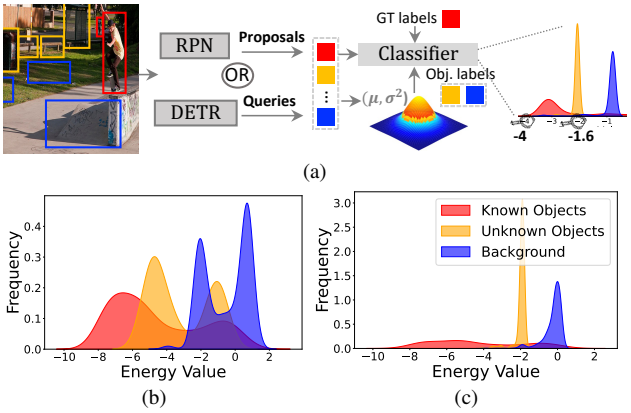


Figure 1. Fig. 1a: The objectness distribution parameters ($\boldsymbol{\mu}$,$\boldsymbol{\sigma}^2$) are estimated across class-agnostic latent embeddings ('Queries' in DETR [69] or 'Proposals' in RPN [39]). Our $Z+1$-dim. classifier includes Z known classes and one additional class for unknown/background detection. The objectness probabilities serve as labels for unlabeled objects ■ and background ■, alongside ground-truth labels for known objects ■, during classifier training. The rightmost plot shows their energies in the classier. Specifically, the energy of known classes is the negative sum of the logits across all Z classes, while the $Z+1$-th logit is used to compute the energy of unknowns/background. Fig. 1b demonstrates the effectiveness of objectness modeling, showing that random soft unknown labels fail to distinguish unknowns from knowns and background. Fig. 1c illustrates that without regularization in the energy space, known energies are spread across low values (high classification confidence), leading to significant misclassifications.

1. Introduction

The ability to recognize general objects as being distinct from the background in images is important for vision systems as they have to deal with the inherently open-set nature of the real world. Deep learning, large datasets and large compute have enabled significant progress towards this goal [4, 7, 8, 16, 50]. However, existing object detectors, *e.g.*, Faster R-CNN [39] and D-DETR [69], remain biased toward training on a fixed set of objects, often misclassifying unknown objects as background. This limitation reduces their value in real-world applications, *e.g.*, autonomous vehicles and robotics [19, 32, 66–68] that must handle unknown objects to ensure good results and safety.

To address this limitation, objectness modeling helps detect objects irrespective of their class labels, facilitating recognition beyond a closed set of categories. Detectors trained on large datasets implicitly learn objectness distributions, enabling generalization in zero-shot scenarios [5, 34, 55, 65]. In situations where large representative datasets are not available, models remain biased toward the known classes [39, 69], limiting their effectiveness for open-world tasks. Older methods [1, 70] relied on hand-

*Corresponding authors. Code: https://github.com/AI4Math-ShanZhang/OWOBJ.

crafted descriptors (*e.g.*, boundaries and edges) or object location [22] to identify potential objects, requiring additional fine-tuning on downstream tasks for class-aware detection.

We capture general objectness given limited samples, enabling the class-aware classifier to perceive this objectness, and prevent unseen objects from being filtered as background or misclassified as known categories. We model the joint distribution of objectness and data (*e.g.*, class distribution in the classifier) via the variational approximation [24] where objectness is treated as a latent variable. The corresponding latent embeddings take the form of 'queries' in DETR-based detectors [17, 69, 71] or 'proposals' in Faster R-CNN detectors [37, 39, 61], highlighting class-agnostic design [6, 22]. The Evidence Lower Bound (ELBO) is then maximized using the Expectation-Maximization (EM) to optimize the distribution [24], equivalent to minimizing the *cross-entropy* between class and objectness variables; and the *KL divergence* between the approximate objectness posterior and the prior normal distribution (§4.1).

For the first item, we employ the output of the probabilistic objectness measure as guidance helping us recognize the potential unknown objects during training, ensuring successful novel object inference. We retain the ground-truth class labels for known objects, as shown in Fig. 1a. The second item, the KL divergence loss, prevents the latent space collapse and overfitting. However, it fails to converge in the absence of sufficient training data, leading to the objectness posterior collapse. Our theoretical analysis (§A.1) reveals the root cause of this issue – the latent variance $\boldsymbol{\sigma}^2$, *e.g.*, the variance of known embeddings approaches zero rapidly during EM-based optimization in the low-data regime. The KL term includes the log of the estimated variance, causing rapid divergence as variance decreases.

Thus, we replace the static prior $\mathcal{N}(\mathbf{0}, \mathbf{1})$ with a dynamic Gaussian prior $\mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 + \boldsymbol{\beta}^2)$, where $\boldsymbol{\beta}$ enables the adaptive behavior. This simple yet effective adjustment transforms the unstable $\log(1/\boldsymbol{\sigma})$ into the stabilized $\log(\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\beta}^2}/\boldsymbol{\sigma})$ (Fig. 4a in §5). Additionally, our theory (§A.1) shows that the dynamic Gaussian prior preserves diversity in latent embeddings, reducing overfitting.

Another issue we investigate is that a significant number of unknown objects are misclassified as known classes. Thus, we visualize object energies in the Z+1-dim. classifier (Z-dim. for known classes and Z+1-th dim. for unknown/background). The energy is aligned with the probability density of inputs [26], *i.e.*, samples with lower energy correspond to data residing in high-probability density regions around the distribution center [31]. Fig. 1c shows energies of first Z-elements for known classes are notably lower and flattened (*e.g.*, ranging from -10 to -2) compared to those of the Z+1-th dim. for unknown/background, leading to overconfidence in known detections.

To address this issue, we apply an energy loss (a margin loss between the energies of known and unknown objects) to regulate the tendency of unknown objects toward high-density regions, and control the margin gap between them. The effectiveness of this approach is illustrated in Fig. 1a (rightmost plot), with a more centered energy distribution for known objects and increased values for unknown objects. Specifically, on Task1 of Open-World Object Detection (OWOD), the number of unknown objects misclassified as known is reduced by over 200, with a 1.2% increase in unknown recall. Moreover, we present the energy distribution from a variant in which unknown labels are not generated by objectness modeling but by randomly sampled values from a uniform distribution (ensuring values range from 0 to 1) to highlight the importance of our objectness modeling. As shown in Fig. 1b, it is evident that the classifier struggles to distinguish unknowns from both the background and known objects, resulting in over 8K misclassification cases and a $\sim$50% decrease in unknown recall.

In summary, our contributions are as follows:

i. We are the first to model the joint distribution of latent objectness and category information probabilistically in the low-data regime, which improves detectors trained on small-scale datasets in open-world tasks.
ii. We theoretically identify the root causes of non-convergence in the variational approximation of objectness. While the KL divergence loss acts as a variance regularizer to ensure an informative latent space, it fails in the low-data regime due to progressively reduced diversity. Based on this analysis, we introduce a dynamic Gaussian prior to stabilize training.
iii. We introduce an energy-based margin loss that guides unknown objects toward high-probability density regions in unknown classifier spaces, reducing unknown detection uncertainty and misclassification rates.

The proposed energy-based **O**pen-**W**orld **OBJ**ectness modeling (OWOBJ) method is evaluated on the COCO dataset for Open-World Object Detection (OWOD) and Few-Shot Object Detection (FSOD). For OWOD, OWOBJ outperforms the state-of-the-art PROB [71] and DETR-style OW-DETR [17] with absolute gains ranging from 5.3% to 19.4% in unknown recall, while also enhancing known object detection by approximately 3% to 6% across the following incremental learning tasks. For FSOD, under the K-shot protocol (K=1, 2, 3, 5, 10, 30), OWOBJ achieves consistent improvements of 2.6%–3.8% over the base model DeFRCN [37] on novel classes. Finally, we demonstrate the effectiveness of OWOBJ (built upon CORA [49]) for zero-shot Open-Vocabulary Object Detection (OVOD) on the OV-LVIS dataset, achieving +3.6% in rare categories.

2. Related Works

Below we discuss related works (see §E for more works).

Objectness Modeling. Objectness differentiates objects from backgrounds without assigning class labels. Early methods used hand-crafted descriptors (*i.e.*, boundaries, edges and colors) to capture generic objects [1, 70]. Region Proposal Network (RPN) [39] learns a two-way classifier to generate candidate proposals but it does not generalize well to novel/unseen objects. OLN [22] leverages localization cues independent of classification for object proposals. The data-driven MAVL [34] builds on multimodal vision transformers, and is trained on large-scale aligned image-text pairs to implicitly recognize objectness from extensive training samples. Noteworthy are also salient object detectors and models leveraging them [21, 42]. None of previous works explore joint modeling of general objectness and class labels as a means of improving performance in the low-data regime. To perform class-aware object detection, they require further fine-tuning on downstream tasks.

Open-World Object Detection (OWOD). More challenging than open set classification [3, 9, 18, 28], OWOD has to tackle instances (of unknown classes) present but unlabeled in the training data. Unlabeled instances are at risk of being learned as background [20]. The first OWOD method, ORE [20], adapted the faster-RCNN by incorporating feature-space contrastive clustering, an RPN-based unknown detector, and an Energy-Based Unknown Identifier (EBUI). Transformer-based methods show significant promise in addressing OWOD. OW-DETR [17] employs a deformable DETR (D-DETR)-serial detector [69] tailored for OWOD tasks. OW-DETR employs a pseudo-labeling scheme that selects high-activation, unmatched proposals (Top-5 by default) as unknown objects. However, this often captures parts of known objects or background regions, leading to unreliable labels and poor performance on unknown objects. PROB [71] employs a probabilistic objectness head alongside D-DETR but it fails to measure the objectness posterior, hindering the classifier's ability to differentiate real objects from non-objects. Without calibrating classification confidence with objectness scores during inference, OWOD degrades significantly.

Few- and Zero-shot Object Detection. Few-Shot Object Detection (FSOD) detects objects using only a few novel samples. Existing methods include transfer learning [43, 47, 48, 54] and meta-learning [10, 13, 45, 46, 52, 59–61]. In practice, transfer learning approaches often outperform meta-learning methods due to their training settings and simpler architectures. MPSR [48] deals with scale invariance by ensuring the detector is trained over multiple scales of positive samples. NP-RepMet [54] introduces negative- and positive-representative learning via triplet losses that bootstrap the classifier. DeFRCN [37] proposes to perform stop-gradient between the RPN and the backbone to deal with the inconsistent optimization goals between them. Zero-shot detection facilitates identifying novel categories absent from the training data. OVR-CNN [58] proposes the Open-Vocabulary Object Detection (OVOD) benchmark to bridge the performance gap between Zero-Shot and supervised learning. Current techniques are divided into four strategies: pseudo-labeling [58, 62, 64, 65], distillation [11, 15], conditional matching [49, 57] and parameter transfer [25, 49]. Although these methods leverage strong zero-shot generalization capabilities of vision-large language models for open-vocabulary detection, they rely on base-category box-level annotations during training, which biases these models toward base classes, underscoring the need for general objectness.

3. Background

Below, we introduce relevant background for our work.

3.1. Problem Formulation

Open World Object Detection. Define T incremental tasks, $\mathcal{T}=\{\mathbf{T}_t\}_{t=1}^{T}$, each introducing known $\mathcal{K}^t=\{C_t^i\}_{i=1}^{K^t}$ and unknown $\mathcal{U}^t=\{C_t^{K^t+1}, C_t^{K^t+2}, \ldots\}$ classes, with $\mathcal{K}^t \cap \mathcal{U}^t = \emptyset$. For each task, the associated training dataset comprises N^t image-label pairs $\mathcal{D}^t = \{(I_i, \mathcal{Y}_i)\}_{i=1}^{N^t}$. Labels, $\mathcal{Y}_i = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots\}$, detail instances within image I_i, with instance label $\boldsymbol{y}_j = [\boldsymbol{l}_j, x_j, y_j, w_j, h_j]$ representing a one-hot class label $\boldsymbol{l}_j \in \{0,1\}^{K^t}$ and the bounding box coordinates $[x_j, y_j, w_j, h_j]$ (the x and y locations, width, and height respectively). OWOD addresses two pivotal challenges: open-set and incremental object detection [20, 71]. In open-set object detection, the object detector for task $\mathbf{T}_t$ trained on a dataset $\mathcal{D}^t$ with annotations for known class only, can categorize objects into known classes $\mathcal{K}^t$ while also identifying any unrecognized objects as 'unknown' during testing. In incremental object detection, the model evolves by integrating new knowledge without retraining from scratch on entire dataset. Specifically, in task $\mathbf{T}_t$, model f^t is developed from dataset $\mathcal{D}^t$, which comprises known classes $\mathcal{K}^t$ and unknown classes $\mathcal{U}^t$. Moving to $\mathbf{T}_{t+1}$, f^t is used to to identify new classes of interest $\bar{\mathcal{U}}^t \in \mathcal{U}^t$, which are then labeled by an oracle. With these newly labeled data, $\bar{\mathcal{D}}^{t+1}$ is formed by expanding $\mathcal{K}^{t+1} = \mathcal{K}^t \cup \bar{\mathcal{U}}^t$ and updating $\mathcal{U}^{t+1} = \mathcal{U}^t \setminus \bar{\mathcal{U}}^t$. The resulted $\bar{\mathcal{D}}^{t+1}$, and a few instances from the datasets of previous tasks $\mathcal{D}^i (i \in \{0, \ldots t\})$, are then employed to refine the model and form f^{t+1}. This process repeats until no unknown classes remain, *i.e.*, $\mathcal{U}^t = \emptyset$.

Few- and Zero-Shot Open-Vocabulary Object Detection. We adopt the standard few-shot object detection approach – a two-stage fine-tuning based model. The detector is first trained on $\mathcal{D}_{base}$ with abundant annotated instances, then quickly fine-tuned on a novel support set $\mathcal{D}_{novel}$ with only a few samples per category. $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$ each consist of N image-label pairs, denoted by $\{(I_i, Y_i)\}_{i=1}^{N}$, where

Y_i contains instance-level annotations. If the support set contains L categories and K examples for each category, the problem is dubbed an L-way K-shot detection. The base classes $\mathcal{C}_{base}$ in $\mathcal{D}_{base}$ and the novel classes $\mathcal{C}_{novel}$ in $\mathcal{D}_{novel}$ are non-overlapping, $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$. Without the second fine-tuning stage, this setting becomes a zero-shot object detection. Similarly to OWOD, we define labeled samples as known objects and unlabeled 'candidate objects' as unknowns.

3.2. Faster-RCNN/DETR & Novel Object Detection

Faster-RCNN [39] includes a convolutional backbone (*i.e.*, ResNet-101) for extracting image features, a Region Proposal Network (RPN) for generating class-agnostic proposals, and a task-specific detection head for class-relevant classification and localization. Specifically, the input image is first processed by the backbone to generate a high-level feature map, which the RPN uses to produce coordinates (top-right and bottom-left) for a sparse set of regions of interest (ROIs) a.k.a., proposals. The global feature maps are cropped by RoIs according to coordinates and aggregated by RoI pooling and passed to the detection head for classification and boundary refinement. In the baseline, DeFRCN [37], the classifier has initially Z+1-dimensions ($Z = |\mathcal{C}_{base}|$ known classes with an additional dimension for non-objects/backgrounds) for base training. Z is gradually extended to $|\mathcal{C}_{base}|+|\mathcal{C}_{novel}|$ in the few-shot fine-tuning stage. To ensure that unlabeled objects are supervised rather than marked as non-objects, we assign objectness probabilities to unlabeled RoIs during training, guiding the class-agnostic classifier in RPN and Z+1-th dimension in class-aware classification. Inference remains same.

DETR-inspired models [7, 69] with simpler designs than two-stage Faster-RCNN models, have been adapted for novel object detection tasks, *e.g.*, OWOD and OVOD tasks. Refer to §4.2 for pipeline details. OWOD employs a one-dimensional classifier for unknown object detection, as in PROB [71], while OVOD further assigns those potential objects to specific novel class names by matching their visual features with text embeddings, as in the baseline CORA [49]. Our approach differs in the labeling process for supervising this one-dimensional classifier. We employ probabilistic objectness scores as soft labels in range $[0, 1]$ for unmatched queries instead of the hard $\{0, 1\}$ labels used in CORA and PROB. No modifications are made to the inference stage.

4. Proposed Approach

4.1. Motivation

We derive our method from a variational approximation perspective [44], aiming to build a more reliable model capable of better capturing the underlying structure (objectness) of the data. One may think of the observed data as being represented by random latent variables, denoted o. These latent variables link the observations $\boldsymbol{x}$ to the model parameters $\boldsymbol{\Theta}$ via the Bayes' law. Mathematically, one can model the observed data and latent variables with a joint distribution $p(\boldsymbol{x}, o; \boldsymbol{\Theta})$. In likelihood-based probabilistic modeling, the goal is to learn a model by maximizing the likelihood $p(\boldsymbol{x}; \boldsymbol{\Theta})$, thereby achieving optimal parameters $\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{x}; \boldsymbol{\Theta})$. This can be accomplished by marginalizing out the latent variables from the joint distribution, as shown by $p(x; \boldsymbol{\Theta}) = \int p(\boldsymbol{x}, o; \boldsymbol{\Theta})\, \mathrm{d}o$. However, directly computing and maximizing the likelihood $p(\boldsymbol{x}; \boldsymbol{\Theta})$ is challenging because it requires integrating over all latent variables. The Evidence Lower Bound (ELBO) serves as a proxy objective for optimizing the latent variable model. Maximizing the ELBO lets us approximate the true latent posterior distribution $p(o|\boldsymbol{x})$ [24]. Formally, the ELBO is defined as: $\mathbb{E}_{q_{\boldsymbol{\phi}}(o|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\Theta}}(\boldsymbol{x},o)}{q_{\boldsymbol{\phi}}(o|\boldsymbol{x})}\right]$, where $q_{\boldsymbol{\phi}}(o|\boldsymbol{x})$ is an approximation of the latent posterior distribution with parameters $\boldsymbol{\phi} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$ that we seek to optimize. The ELBO term can be decomposed further:

$$\begin{aligned}
&\mathbb{E}_{q_{\boldsymbol{\phi}}(o|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\Theta}}(\boldsymbol{x}, o)}{q_{\boldsymbol{\phi}}(o|\boldsymbol{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(o|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\Theta}}(\boldsymbol{x}|o)p(o)}{q_{\boldsymbol{\phi}}(o|\boldsymbol{x})}\right] \\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(o|\boldsymbol{x})}[\log p_{\boldsymbol{\Theta}}(\boldsymbol{x}|o)]}_{\text{negative cross-entropy term}} - \underbrace{D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(o|\boldsymbol{x}) \parallel p(o))}_{\text{prior matching term}} \qquad (1)
\end{aligned}$$

where $D_{\mathrm{KL}}\big(q(\cdot) \parallel p(\cdot)\big)$ represents the Kullback-Leibler divergence between distributions $q(\cdot)$ and $p(\cdot)$.

Eq. (1) outlines two principles for effectively capturing relationship between the latent objectness variable and the data: 1) *minimizing the cross-entropy* between the observed and latent variables $H(\boldsymbol{x}, o)$ and 2) *minimizing the KL divergence* between the learned variational distribution and the prior distribution. In object detection, the continuous data distribution is discretized into a category distribution via the softmax activation function. To satisfy the first requirement, we use probabilistic objectness as pseudo-labels to provide auxiliary guidance for the (Z+1)-dimensional classifier, with Z-dimensions for known object detection and one dimension for unknown/non-object prediction. Specifically, the (Z+1)-th classifier is supervised by pseudo-labels, while the Z-dimensional classifier is guided by ground-truth labels for Z known classes. We minimize the cross-entropy loss between predictions and these labels.

As noted above, the KL divergence term in Eq. (1) causes the optimization to fail to converge, particularly in the low-data regime such as Task1 of OWOD, where only 20 classes are labeled as known objects while the remaining 60 classes are treated as unknown objects. Our theoretical analysis identifies reduced variance $\boldsymbol{\sigma}^2$ as the cause, resulting in logarithmic instability in the prior matching term, as shown be-

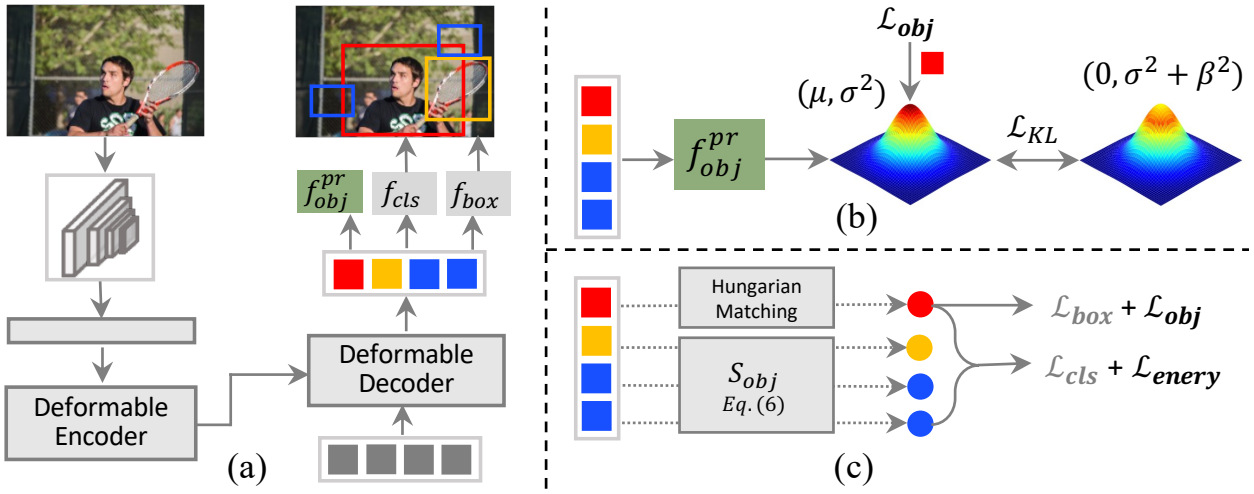


Figure 2. **Pipeline of OWOBJ instantiated in D-DETR.** (a) Query embeddings generated by D-DETR are used by the classifier, the bounding box head, and probabilistic objectness modeling. We introduce the objectness modeling module ($f^{\text{pr}}_{\text{obj}}$). Unaltered modules are shown in gray. (b) The probability distribution $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ is estimated using the exponential moving average across all query embeddings. The likelihood for labeled known queries ■ is maximized by penalizing the Mahalanobis distance, $\mathcal{L}_{obj}$, along with the KL divergence $\mathcal{L}_{KL}$ to mitigate bias toward known objects ■ and to avoid the latent posterior collapse. (c) Queries matched with labeled known objects are assigned ground-truth labels ● through Hungarian matching, while other unmatched queries (*e.g.*, unknown objects ■ and non-objects ■) are assigned soft pseudo-labels ●/●, indicating the probability of objectness.

low:

$$D_{\text{KL}}\big(q_{\boldsymbol{\phi}}(o|\boldsymbol{x}) \parallel p(o)\big) = \log \frac{1}{\boldsymbol{\sigma}} + \frac{\boldsymbol{\sigma}^2 + \boldsymbol{\mu}^2}{2} - \frac{1}{2}, \quad (2)$$

where $q_{\boldsymbol{\phi}}(o|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ represents an estimated posterior and $p(o) = \mathcal{N}(\mathbf{0}, \mathbf{1})$ denotes the prior normal distribution. This KL divergence term encourages the model to learn a distribution rather than collapse to a Dirac delta function. It serves as a variance regularizer that preserves diversity in latent embeddings. Reduced variance underscores the need to effectively optimize the KL divergence loss and mitigate bias toward labeled samples. As a solution, we employ a dynamic Gaussian prior $g_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 + \boldsymbol{\beta}^2)$, where $\boldsymbol{\beta}$ modulates the adaptive behavior, which can mitigate logarithmic instability by effectively regularizing the KL term and stabilize training in the low-data regime:

$$D_{\text{KL}}\big(q_{\boldsymbol{\phi}}(o|\boldsymbol{x}) \,\|\, g_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\epsilon})\big) = \log \frac{\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\beta}^2}}{\boldsymbol{\sigma}} + \frac{\boldsymbol{\sigma}^2 + \boldsymbol{\mu}^2}{2(\boldsymbol{\sigma}^2 + \boldsymbol{\beta}^2)} - \frac{1}{2}, \quad (3)$$

where $g_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\epsilon}) = \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}^{(\mathbf{1})} \oplus \boldsymbol{\epsilon}^{(\boldsymbol{\beta})}$ with $\boldsymbol{\epsilon}^{(\mathbf{1})} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ and $\boldsymbol{\epsilon}^{(\boldsymbol{\beta})} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\beta}^2)$. Operators $\odot$ and $\oplus$ denote element-wise multiplication and addition, respectively.

4.2. Overall Architecture

OWOBJ is adaptable to both Faster R-CNN-based and DETR-based detectors as a plug-in module. The core component of OWOBJ is the objectness modeling which operates on class-agnostic candidate objects, whether in the form of ROIs in Faster R-CNN or queries in DETR. Below, we demonstrate the integration of OWOBJ within the D-DETR [69] serial pipeline. We build our pipeline upon PROB [71] for OWOD and CORA [49] for OVOD with a focus on improving identification of unlabeled objects. Both tasks require an additional one-dimensional classifier to learn objectness. Figure 2 shows our integration of the probabilistic objectness model ($f^{\text{pr}}_{\text{obj}}$), outlined in green, into the training stage, with the inference process unchanged. This integration forms a robust framework for generating pseudo-labels to supervise objectness prediction in the classifier (the right panel of Fig. 2).

The process is initiated by inputting an image of size $H \times W$ into a backbone network (ResNet-101), which extracts D-dimensional multi-scale features at varying resolutions. These features are subsequently processed by a transformer encoder containing multi-scale deformable attention modules. Next, the decoder leverages N_{query} learned query vectors to cross-correlate them with the encoded image features, generating corresponding query embeddings $\boldsymbol{Q} \in \mathbb{R}^{N_{\text{query}} \times D}$. Each query embedding $\boldsymbol{q} \in \mathbb{R}^D$ is then passed into three branches: the bounding box regression branch (f_{box}, a 3-layer feed-forward network (FFN) to produce box coordinate), the classification head (f_{cls}), and the probabilistic objectness model ($f^{\text{pr}}_{\text{obj}}$). To train the classification head and bounding box regression, the Hungarian matching algorithm is used to assigns queries with the best-matched ground-truth instances according to their class and box coordinate predictions [69]. This process produces a set of matched queries $(\boldsymbol{q}_j, j \in \mathcal{Q})$, each linked to a label

from the Z known classes, and a set of unmatched queries ($\boldsymbol{q}_i, i \in \bar{\mathcal{Q}}$), totaling $|\mathcal{Q}| + |\bar{\mathcal{Q}}| = N_{\text{query}}$. The classifier is then trained to classify these matched queries as one of the Z known classes, whereas the regression branch provides bounding box locations. Unmatched queries, representing either unknown objects or background, are classified into the Z+1-th category for objectness prediction. These are then incrementally learned for specific class names in OWOD, or directly matched with text embeddings of novel class names in zero-shot OVOD to assign class labels.

Without labels indicating whether unmatched queries are objects or background, identifying unknown objects via the Z+1-th category during testing becomes challenging. We thus generate soft pseudo-labels indicating the probability of objectness, guiding the Z+1-th classifier for unmatched queries and reducing the cross-entropy between them. To further enhance detection of unknown objects, we propose $\mathcal{L}_{energy}$ to increase the likelihood of unknown objects, reducing their misclassification as known classes. Below we provide details on the energy-based margin loss $\mathcal{L}_{energy}$, the objectness loss $\mathcal{L}_{obj}$ for optimizing the objectness posterior and the objectness scores S_{obj} used as labels for unknowns.

4.3. Objectness Modeling

Objectness loss. Following [27, 71], we model the objectness probability as a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean $\boldsymbol{\mu}$ and diagonal covariance $\boldsymbol{\Sigma}$=$\boldsymbol{\sigma}^2\mathbf{I}$, both updated via exponential moving average (EMA) across all query embeddings. The empirical mean and covariance of query embeddings are computed channel-wise, *i.e.*, $\boldsymbol{\mu} \in \mathbb{R}^{D\times 1}, \boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$. Typically, the energy function, defined as the distance between the sample and the mean of objectness distribution, is chosen to optimize the probability distribution. We use the Mahalanobis distance, effective in few-shot learning [2]. The probilistic model assigns higher probabilistic density to objects and lower to backgrounds. For matched queries, which are known to be objects, we maximize their likelihood by minimizing the Mahalanobis distance $d_M(\cdot)$ as follows:

$$d_M(\boldsymbol{q}) = (\boldsymbol{q} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{q} - \boldsymbol{\mu}), \tag{4}$$

$$\mathcal{L}_{obj} = \textstyle\sum_{i\in\mathcal{Q}} d_M(\boldsymbol{q}_i). \tag{5}$$

Objectness score. After estimating the objectness distribution, we generate pseudo-labels for unmatched queries to enhance classifier training. Denote $\hat{\boldsymbol{l}} \in \mathbb{R}^{Z+1}$ as the classifier prediction. We modify the original one-hot label $\boldsymbol{l} \in \{0,1\}^{Z+1}$ by assigning objectness scores S^i_{obj} to the $(Z+1)$-th element for unmatched queries $\boldsymbol{q}_i, i \in \bar{\mathcal{Q}}$:

$$\boldsymbol{l}_i[Z+1] = S^i_{\text{obj}}, \;\; \text{with} \;\; S^i_{\text{obj}} = \exp\big(-d_M(\boldsymbol{q}_i)\big). \tag{6}$$

Nothing is changed for matched queries, where one of the first Z elements set to 1 and the others to 0. We adopt the focal loss [30] as the classification loss $\mathcal{L}_{cls} = \text{FL}(\hat{l}, l)$, as in baseline models.

Energy-based margin loss. The energy of a classifier with a softmax activation is defined as the negative sum of the logits across all classes [31], a.k.a. the denominator of the softmax function: $E(\boldsymbol{q}; f_{\text{cls}}) = -\log \sum_i e^{f^i_{\text{cls}}(\boldsymbol{q})}$. For the known class classifier (with Z dimensions), the energy is computed across matched queries $\boldsymbol{q}_k, k \in \mathcal{Q}$ as follows:

$$E_k = -\frac{1}{|\mathcal{Q}|} \sum_{k}^{|\mathcal{Q}|} \log \sum_{i}^{Z} e^{f^i_{\text{cls}}(\boldsymbol{q}_k)}. \tag{7}$$

For the classifier's $(Z+1)$-th element, the energy is computed among unmatched queries $\boldsymbol{q}_u, u \in \bar{\mathcal{Q}}$ as:

$$E_u = -\frac{1}{|\bar{\mathcal{Q}}|} \sum_{u}^{|\bar{\mathcal{Q}}|} \log Z \cdot e^{f^{Z+1}_{\text{cls}}(\boldsymbol{q}_u)}. \tag{8}$$

Our energy-based margin loss is defined as:

$$\mathcal{L}_{\text{energy}} = \left(E_u - E_k + \delta\right)_+, \tag{9}$$

where δ ($\delta = 0.2$ as default) is a positive margin parameter that controls the separation between the classifier energies for known and unknown objects. The energy is negatively correlated with probability density or classification confidence, so encouraging $\boldsymbol{q}_u$ to have higher energy values helps reduce uncertainty for unknown objects.

The final losses include: a classification loss denoted as $\mathcal{L}_{cls}$; a regression loss $\mathcal{L}_{reg}$, which is a combination of the ℓ_1 loss and the GIoU loss; an energy-based margin loss $\mathcal{L}_{energy}$; and the maximization of the likelihood $\mathcal{L}_{obj}$ which is applied to matched queries, along with the KL divergence $\mathcal{L}_{KL}$ (Eq. (3)). We follow baselines for other implementation details (learning rate, weight decay, and batch size).

5. Experiments

We assess the generalization ability of OWOBJ across three challenging low-shot scenarios, achieving superior performance over all methods and strong baselines. Specifically, we evaluate: 1) Open-World Object Detection (OWOD) on COCO, measuring mean average precision (mAP) for known classes and average recall (U-recall) for unknown objects (refer to §B for the Absolute Open-Set Error (A-OSE) quantifying unknown objects misclassified as knowns); 2) Few-Shot Object Detection (FSOD) on COCO, using COCO-style mAP where 60 categories serve as base classes and 20 as novel classes under $K = 1, 2, 3, 5, 10, 30$ shots for few-shot fine-tuning, with evaluation on 5k validation images; and 3) Zero-shot Open-Vocabulary Object Detection (OVOD) on OV-LVIS [16], where the model, trained on 461 common and 405 frequent classes, is evaluated on rare classes using mAP (AP_r). We use the same training and testing pipelines as baselines to ensure a fair compar-

Table 1. Comparison for OWOD on M-OWODB (top) and S-OWODB (bottom), measured by unknown recall (U-Recall) and known class mean Average Precision at IoU threshold 0.5 (known mAP@0.5) for previously and currently learned known objects.

Task IDs (→)	**Task 1**		**Task 2**				**Task 3**				**Task 4**		
	U-Recall	mAP (↑)	U-Recall	mAP (↑)			U-Recall	mAP (↑)			mAP (↑)		
	(↑)	Current known	(↑)	Previously known	Current known	Both	(↑)	Previously known	Current known	Both	Previously known	Current known	Both
ORE* [20]	4.9	56.0	2.9	52.7	26.0	39.4	3.9	38.2	12.7	29.7	29.6	12.4	25.3
UC-OWOD [56]	2.4	50.7	3.4	33.1	30.5	31.8	8.7	28.8	16.3	24.6	25.6	15.9	23.2
OCPL [34]	8.26	56.6	7.65	50.6	27.5	39.1	11.9	38.7	14.7	30.7	30.7	14.4	26.7
2B-OCD [51]	12.1	56.4	9.4	51.6	25.3	38.5	11.6	37.2	13.2	29.2	30.0	13.3	25.8
OW-DETR [17]	7.5	59.2	6.2	53.6	33.5	42.9	5.7	38.3	15.8	30.8	31.4	17.1	27.8
PROB [71]	19.4	59.5	17.4	55.7	32.2	44.0	19.6	43.0	22.2	36.0	35.7	18.9	31.5
PROB+OWOBJ	**23.6**	**61.4**	**23.8**	**58.4**	**34.4**	**45.7**	**25.1**	**44.8**	**27.8**	**38.8**	**36.4**	**20.7**	**32.0**
ORE* [51]	1.5	61.4	3.9	56.5	26.1	40.6	3.6	38.7	23.7	33.7	33.6	26.3	31.8
OW-DETR [17]	5.7	71.5	6.2	62.8	27.5	43.8	6.9	45.2	24.9	38.5	38.2	28.1	33.1
PROB [71]	17.6	73.4	22.3	66.3	36.0	50.4	24.8	47.8	30.4	42.0	42.6	31.7	39.9
CAT [33]	23.7	60.0	19.1	55.5	32.7	44.1	24.4	42.8	18.7	34.8	34.4	16.6	29.9
MEPU-FS [14]	37.9	74.3	35.8	68.0	41.9	54.3	35.7	50.2	38.3	46.2	43.7	33.7	41.2
PROB+OWOBJ	22.3	76.2	28.7	69.8	41.0	54.8	30.9	50.6	35.7	46.8	46.7	36.9	43.2
MEPU-FS+OWOBJ	**39.7**	**77.4**	**37.8**	**71.5**	**43.1**	**57.2**	**38.1**	**53.1**	**39.2**	**49.0**	**49.4**	**38.8**	**43.9**

Table 2. Impact of progressively integrating our contributions into the baseline. The comparison is conducted on M-OWODB for OWOD task. We also include the performance of D-DETR and an upper limit (D-DETR trained with ground-truth annotations for unknown classes), reported by OW-DETR.

Task IDs (→)	**Task 1**		**Task 2**			
	U-Recall	mAP (↑)	U-Recall	mAP (↑)		
	(↑)	Current known	(↑)	Previously known	Current known	Both
Upper Bound	32.9	63.8	40.5	60.8	39.2	49.3
D-DETR [69]	-	60.3	-	54.5	34.4	44.7
OWOBJ−S_{obj}	10.2	56.1	13.7	53.4	30.9	41.2
OWOBJ−$\mathcal{L}_{obj}$	13.4	58.5	15.6	57.3	33.2	44.3
OWOBJ−$\mathcal{L}_{KL}$	19.2	**62.3**	20.2	**60.1**	**35.2**	**46.1**
OWOBJ−$\mathcal{L}_{energy}$	21.1	60.1	21.6	57.5	33.1	44.6
OWOBJ	**23.6**	61.4	**23.8**	58.4	34.4	45.7

Table 3. Results on the OV-LVIS benchmark.

Method	**Extra Data**	**Pre-train Model**	**AP**$_r$
ViLD [15]	-	CLIP (ViT-B/32)	16.3
OV-DETR [57]	-	CLIP (ViT-B/32)	17.4
RegionCLIP [64]	CC3M	CLIP (RN50x4)	22.0
MEDet [6]	CC3M	CLIP (ViT-B/32)	22.4
Detic [65]	IN-21k	CLIP (text encoder)	26.2
OWL-ViT [35]	-	CLIP (ViT-L/14)	25.6
RO-ViT [23]	ALIGN	CLIP (ViT-B/16)	28.4
CORA [49]	IN-21k	CLIP (RN50x4)	28.1
CORA+OWOBJ	IN-21k	CLIP (RN50x4)	**31.7**

ison. We also conduct analyses of each component within our framework (refer to §D for visualization results).

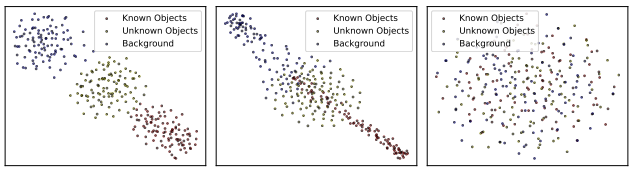
5.1. Open World Object Detection

Setup. In what follows, we employ the widely utilized OWOD benchmark known as the 'superclass-mixed OWOD benchmark' (M-OWODB) [20] and the 'superclass-separated OWOD benchmark' (S-OWODB) [17]. M-OWODB organizes images from COCO [29], PASCAL VOC2007 [12], and PASCAL VOC2012 into $T = 4$ incremental tasks, adding 20 labeled classes per task. During testing, all classes from current and previous tasks are detected. S-OWODB, which focuses on COCO dataset, adds classes by super-categories per increment to preserve superclass integrity.

Table 4. Results on Few-Shot Object Detection across varying K.

Method / Shots	**Shot Number** K					
	1	**2**	**3**	**5**	**10**	**30**
Meta R-CNN [52]	-	-	-	-	8.7	12.4
TFA [47]	4.4	5.4	6.0	7.7	10.0	13.7
MPSR [48]	5.1	6.7	7.4	8.7	9.8	14.1
FSCE [43]	-	-	-	-	11.9	16.4
KFSOD [61]	-	-	-	-	18.5	22.8
TENET [60]	-	-	-	-	19.1	23.7
DeFRCN [37]	9.3	12.9	14.8	16.1	18.5	22.6
DeFRCN+OWOBJ	**11.9**	**15.6**	**17.9**	**19.4**	**23.8**	**26.4**

Figure 3. t-SNE visualization of query embeddings derived from three variants. Best viewed in color with zoom.



(a) OWOBJ (b) OWOBJ−$\mathcal{L}_{KL}$ (c) OWOBJ−$\mathcal{L}_{obj}$

Comparisons. Table 1 presents a comparison of our OWOBJ (build upon PROB and MEPU-FS) with other OWOD methods. S-OWODB has been recently introduced

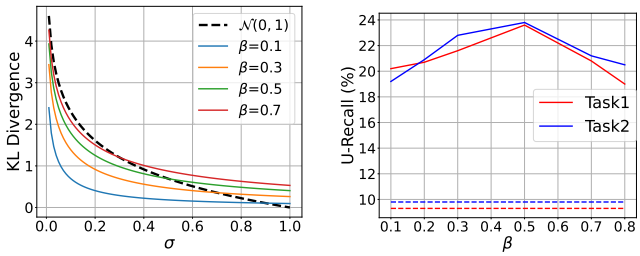
to OW-DETR [17] and only ORE [20], OW-DETR [17], PROB [71] and MEPU-FS [14] provide performance metrics on such a new split dataset (bottom). This dataset involves a full super-category separation across tasks, making it more challenging for OWOD settings. Our MEPU-FS+OWOBJ demonstrates notable improvements in unknown recall (U-Recall), along with additional enhancements in known object mAP, compared to other methods.

In comparisons across four tasks and two benchmarks, PROB+OWOBJ demonstrates a significant improvement in detecting **unknown objects**, with a 5% increase in U-Recall over the previous state-of-the-art PROB. This highlights the importance of jointly modeling probabilistic objectness and categorical labels for effective OWOD detection. A closer examination of the results in §B reveals that OWOBJ achieves higher accuracy in unknown object detection, as reflected by significantly lower A-OSE values (a ∼3K decrease across three tasks). In terms of **known object detection**, our method consistently outperforms all previous state-of-the-art OWOD techniques in terms of known object mAP. Specifically, compared to the strong baseline PROB, our approach achieves ∼+3% mAP in both M-OWODB (Table 1, top) and S-OWODB (Table 1, bottom). This improvement underscores the effectiveness of capturing the underlying objectness distribution, which, in turn, enhances the detection of known objects in OWOD. Further evidence is shown in §C.1.

Ablations. Due to the limited training samples in the initial two tasks, we conduct our ablation study on the challenging Task1 and Task2 of M-OWODB, as detailed in Table 2. We observe that removing objectness scores as unknown pseudo-labels and substituting randomly generated soft labels (OWOBJ−S_{obj}) significantly decreases performance for both known and unknown detection. As shown in Fig. 1b, random labels lead to greater overlap across object energies, indicating reduced discriminative power. In the variant OWOBJ−$\mathcal{L}_{obj}$, which excludes the objectness loss in the training stage, the model fails to optimize the latent distribution and to effectively capture what an object should look like, leading to unreliable guidance and impaired performance (-10.2%). As expected, omitting the KL divergence (OWOBJ−$\mathcal{L}_{KL}$) leads to overfitting towards known classes, which adversely impacts unknown recall (-4.4%). Finally, without the energy-based margin loss (OWOBJ−$\mathcal{L}_{energy}$), potential unlabeled object scores are become significantly lowered, resulting in increased misclassification negatively impacting both unknown recall and known mAP. See Fig. 3a, 3b and 3c. Without $\mathcal{L}_{KL}$, known embedding variance decreases, overfitting the model to known objects and misclassifying unknown objects as known ones (yellow *vs.* red points). Without $\mathcal{L}_{obj}$, the detector fails to perceive objectness, clustering all embeddings together.

Fig. 4a shows how a dynamic Gaussian prior stabilizes KL divergence, preventing rapid divergence as $\boldsymbol{\sigma}$ decreases. Fig. 4b displays U-Recall for dynamic Gaussian prior adjusted by β values, with $\beta = 0.5$ performing best.

Figure 4. Fig. 4a: Dynamic prior w.r.t. β stabilizes the KL divergence *vs.* normal prior $\mathcal{N}(0, 1)$; Fig. 4b: U-Recall of dynamic prior across various β with dashed lines representing the static prior.



(a) (b)

5.2. Few- and Zero-Shot Object Detection

Results. Table 4 shows COCO evaluation results for FSOD with 1, 2, 3, 5, 10, 30-shot setting, reported as the standard COCO-style mAP. Our approach consistently surpasses the strong baseline across all settings, achieving a notable 4.2% improvement in the 30-shot setting and even a 2.6% gain in the challenging 1-shot scenario. Zero-Shot OVOD evaluates generalization to novel classes, unseen during training. Again, without altering the inference step, Table 3 shows that CORA+OWOBJ outperforms CORA by +3.6%.

6. Conclusions

Detecting objects in the open-world is challenging due to limited training data and the need for generalization to unseen classes. A crucial aspect is effectively modeling the objectness conditioned on observed samples. Previous methods rely on heuristic designs or require fine-tuning on downstream tasks for class-aware detection. We are the first to model the joint distribution of latent objectness and category labels, achieved through the variational approximation. Our theory also identifies the root cause of objectness posterior collapse, which we address with a dynamic Gaussian prior. Additionally, the proposed energy-based margin loss regularizes the energy space, guiding unknown objects toward high-probability density regions and reducing uncertainty in detection. Experimental results show that our approach achieves superior performance in detecting novel/unseen objects in the low-data regime in Open-World, Few-Shot, and Zero-Shot object detection tasks. We believe OWOBJ represents a step forward in open-world novel object understanding, offering new insights for future research.

Acknowledgment

Shan Zhang and part of this work are supported by the Centre for Augmented Reasoning, an initiative by the Department of Education, Australian Government. PK and part of this work are funded by the CSIRO Science Digital (SD) and Advanced Engineering Biology Future Science Platforms (AEB-FSP).

References

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 73–80. IEEE, 2010. 1, 3

[2] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14493–14502, 2020. 6

[3] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 7

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[5] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training, 2024. 1

[6] Peixian Chen, Kekai Sheng, Mengdan Zhang, Mingbao Lin, Yunhang Shen, Shaohui Lin, Bo Ren, and Ke Li. Open vocabulary object detection with proposal mining and prediction equalization, 2022. 2, 7

[7] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085*, 2022. 1, 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1

[9] Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *ArXiv*, abs/1812.02765, 2018. 3, 7

[10] Jinhao Du, Shan Zhang, Qiang Chen, Haifeng Le, Yanpeng Sun, Yao Ni, Jian Wang, Bin He, and Jingdong Wang. σ-adaptive decoupled prototype for few-shot object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18904–18914. IEEE, 2023. 3

[11] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guo Chun Li. Learning to prompt for open-vocabulary object detection with vision-language model. *CVPR*, 2022. 3, 8

[12] 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 (IJCV)*, 88(2):303–338, 2010. 7

[13] Qi Fan, Wei Zhuo, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. *CoRR*, abs/1908.01998, 2019. 3, 8

[14] Ruohuan Fang, Guansong Pang, Lei Zhou, Xiao Bai, and Jin Zheng. Unsupervised recognition of unknown objects for open-world object detection. *arXiv preprint arXiv:2308.16527*, 2023. 7, 8, 4, 5

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2021. 3, 7, 8

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1, 6

[17] Akshita Gupta, Sanath Narayan, KJ 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)*, 2022. 2, 3, 7, 8, 4, 5

[18] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv*, abs/1610.02136, 2016. 3, 7

[19] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision (FTCGV)*, 12(1–3):1–308, 2020. 1

[20] 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)*, 2021. 3, 7, 8, 4, 5

[21] Dahyun Kang, Piotr Koniusz, Minsu Cho, and Naila Murray. Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19627–19638, 2023. 3

[22] 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*, 7(2):5453–5460, 2022. 2, 3

[23] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers, 2023. 7

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 4

[25] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *ICLR*, 2023. 3, 8

[26] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 2

[27] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 6

[28] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *ArXiv*, abs/1706.02690, 2017. 3, 7

[29] 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)*, 2014. 7

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

[31] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems (NIPS)*, 2020. 2, 6

[32] Zhe Liu, Xin Zhao, Tengteng Huang, Ruolan Hu, Yu Zhou, and Xiang Bai. Tanet: Robust 3d object detection from point clouds with triple attention. In *AAAI Conference on Artificial Intelligence(AAAI)*, 2020. 1

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19681–19690, 2023. 7, 4, 5

[34] 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*, pages 512–531. Springer, 2022. 1, 3, 7, 4

[35] 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. In *ECCV*, pages 728–755. Springer, 2022. 7

[36] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *Pattern recognition letters*, 140:109–115, 2020. 5

[37] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection, 2021. 2, 3, 4, 7, 8

[38] 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 *ICML*, pages 8748–8763. PMLR, 2021. 8

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 1, 2, 3, 4

[40] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European signal processing conference*, pages 606–610. IEEE, 2007. 3

[41] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*, pages 3400–3409, 2017. 5

[42] Peipei Song, Jing Zhang, Piotr Koniusz, and Nick Barnes. Multi-modal transformer for rgb-d salient object detection. *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2466–2470, 2022. 3

[43] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. FSCE: few-shot object detection via contrastive proposal encoding. *CoRR*, abs/2103.05950, 2021. 3, 7, 8

[44] Dimitris G Tzikas, Aristidis C Likas, and Nikolaos P Galatsanos. The variational approximation for bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, 2008. 4

[45] Lei Wang and Piotr Koniusz. Temporal-viewpoint transportation plan for skeletal few-shot action recognition. In *Computer Vision – ACCV 2022: 16th Asian Conference on Computer Vision, Macao, China, December 4–8, 2022, Proceedings, Part IV*, page 307–326, Berlin, Heidelberg, 2022. Springer-Verlag. 3

[46] Lei Wang, Jun Liu, Liang Zheng, Tom Gedeon, and Piotr Koniusz. Meet JEANIE: A similarity measure for 3d skeleton sequences via temporal-viewpoint alignment. *Int. J. Comput. Vis.*, 132(9):4091–4122, 2024. 3

[47] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection, 2020. 3, 7, 8

[48] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, pages 456–472. Springer, 2020. 3, 7, 8

[49] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching, 2023. 2, 3, 4, 5, 7, 8

[50] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[51] 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. 7, 4

[52] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta R-CNN: towards general solver for instance-level low-shot learning. In *ICCV 2019*, pages 9576–9585. IEEE, 2019. 3, 7, 8

[53] Shuo Yang, Peize Sun, Yi Jiang, Xiaobo Xia, Ruiheng Zhang, Zehuan Yuan, Changhu Wang, Ping Luo, and Min Xu. Objects in semantic topology. In *International Conference on Learning Representations (ICLR)*, 2022. 7

[54] Yukuan Yang, Fangyun Wei, Miaojing Shi, and Guoqi Li. Restoring negative information in few-shot object detection. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3, 8

[55] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection, 2022. 1

[56] Jinan Yu, Liyan Ma, Zhenglin Li, Yan Peng, and Shaorong Xie. Open-world object detection via discriminative class

prototype learning. *arXiv preprint arXiv:2302.11757*, 2023. 7

[57] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. 2022. 3, 7, 8

[58] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021. 3, 8

[59] Shan Zhang, Dawei Luo, Lei Wang, and Piotr Koniusz. Few-shot object detection by second-order pooling. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3, 8

[60] Shan Zhang, Murray Naila, Lei Wang, and Piotr Koniusz. Time-reversed diffusion tensor transformer: A new tenet of few-shot object detection. In *ECCV*, 2022. 7, 8

[61] Shan Zhang, Lei Wang, Naila Murray, and Piotr Koniusz. Kernelized few-shot object detection with efficient integral aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19207–19216, 2022. 2, 3, 7, 8

[62] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, pages 159–175. Springer, 2022. 3, 8

[63] Xiaowei Zhao, Xianglong Liu, Yifan Shen, Yixuan Qiao, Yuqing Ma, and Duorui Wang. Revisiting open world object detection. *ArXiv*, abs/2201.00471, 2022. 7

[64] 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 *CVPR*, pages 16793–16803, 2022. 3, 7, 8

[65] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pages 350–368. Springer, 2022. 1, 3, 7, 8

[66] Yu Zhou, Xiang Bai, Wenyu Liu, and Longin Latecki. Fusion with diffusion for robust visual tracking. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. 1

[67] Yu Zhou., Yinfei Yang., Yi Meng., Xiang Bai, Wenyu Liu, and Longin Jan Latecki. Online multiple person detection and tracking from mobile robot in cluttered indoor environments with depth camera. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 28(1): 1455001.1–1455001.28, 2014.

[68] Yu Zhou, Xiang Bai, Wenyu Liu, and Longin Jan Latecki. Similarity fusion for visual tracking. *International Journal of Computer Vision (IJCV)*, 118(3):337–363, 2016. 1

[69] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 3, 4, 5, 7

[70] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014. 1, 3

[71] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2023. 2, 3, 4, 5, 6, 7, 8