OSR-ViT: A Simple and Modular Framework for Open-Set Object Detection and Discovery

Matthew Inkawhich
Duke University
Durham, NC, USA
matthewinkawhich@gmail.com

Nathan Inkawhich
Air Force Research Laboratory
Rome, NY, USA
ninkawhich315@gmail.com

Hao Yang
Johns Hopkins University
Baltimore, MD, USA
haofrankyang@jhu.edu

Jingyang Zhang
Duke University
Durham, NC, USA
jingyang.zhang@duke.edu

Randolph Linderman
Duke University
Durham, NC, USA
randolph.linderman@duke.edu

Yiran Chen
Duke University
Durham, NC, USA
yiran.chen@duke.edu

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10826036

***Abstract*—An object detector's ability to detect and flag novel objects during open-world deployments is critical for many real-world applications. Unfortunately, much of the work in open object detection today is disjointed and fails to adequately address applications that prioritize unknown object recall in addition to known-class accuracy. To close this gap, we present a new task called Open-Set Object Detection and Discovery (OSODD) and as a solution propose the Open-Set Regions with ViT features (OSR-ViT) detection framework. OSR-ViT combines a class-agnostic proposal network with a powerful ViT-based classifier. Its modular design simplifies optimization and allows users to easily swap proposal solutions and feature extractors to best suit their application. Using our multifaceted evaluation protocol, we show that OSR-ViT obtains performance levels that far exceed state-of-the-art supervised methods. Our method also excels in low-data settings, outperforming supervised baselines using a fraction of the training data.**

***Index Terms*—Computer vision, robustness, open-set.**

I. Introduction

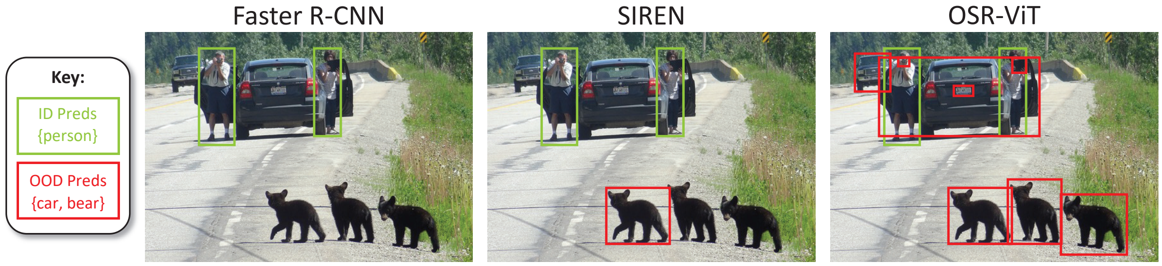
Traditional object detection models are designed, trained, and evaluated under closed-set conditions [1]–[7], where all potential classes of interest are assumed to be exhaustively labeled in the training dataset. If such a model is deployed in an open-set environment [8], [9] where there exists unknown objects from outside the training class distribution, the detector will either misclassify the object as a known class or miss it altogether – leading to serious safety and reliability concerns. This motivates the need for open-set object detection [10], where unknown "out-of-distribution" (OOD) objects are explicitly handled in addition to the known "in-distribution" (ID) objects.

Although there have been many works that attempt to address open-set detection [10]–[15], we posit that the way they choose to handle unknown objects severely limits their practical usefulness. Namely, none of them consider **OOD object recall**. For example, seminal works by Miller et al. [11], [12] and Dhamija et al. [9] define proper "Open-Set Object Detection" (OSOD) behavior as simply avoiding detecting any OOD objects as ID classes. More recent works by Du et al. [13]–[15] tackle "Unknown-Aware Object Detection" (UAOD), where the model is expected to accurately flag OOD objects that happen to be proposed to the detector's classifier head, but does not encourage OOD proposals. While these behaviors may be sufficient for some tasks, many applications require the explicit detection (i.e., discovery) of *all* objects of interest, both ID and OOD. For example, autonomous vehicles are often exposed to unforeseeable obstacles that demand detection for safe operation [16], [17]. Content moderation systems must also accurately identify evolving types of content while navigating the complexities of insufficient filtering [18]. Further, medical image processing models are relied upon to detect abnormalities [19]. In such cases, the consequences of poor OOD recall are severe, necessitating a new open-set task that prioritizes it.

In this work, we introduce **Open-Set Object Detection and Discovery (OSODD): a <u>task</u> that explicitly prioritizes both ID-class accuracy *and* OOD object recall**. OSODD more appropriately models many realistic applications like the ones mentioned above. To measure performance on the OSODD task, we devise a new evaluation protocol that makes no simplifying assumptions about the test data and includes a novel threshold-independent Average Open Set Precision (AOSP) summary metric. We test models on three new benchmarks that are designed to simulate a broad spectrum of feasible settings, including low-data environments and multiple image domains. Not only does our OSODD evaluation protocol enable a more comprehensive analysis of model performance, but it also is the first that allows for a unified comparison of models from several subdivisions of open detection (e.g., OSOD [9], UAOD [13], Open-World Object Detection (OWOD) [10]). Such a comparison highlights how poorly those solutions perform in OSODD (see Fig. 1).

To address the OSODD task, we create a new **highly-modular detection framework, called Open Set Regions with ViT features (OSR-ViT)**. This framework is comprised of a dedicated class-agnostic proposal network combined with a classifier module that leverages powerful off-the-shelf ViT-

979-8-3503-6248-0/24/$31.00 ©2024 IEEE



Faster R-CNN		SIREN		OSR-ViT	
ID mAP ↑	1.0	ID mAP ↑	1.0	ID mAP ↑	1.0
IDvOOD AUROC ↑	1.0	IDvOOD AUROC ↑	1.0	IDvOOD AUROC ↑	1.0
OOD AR@k ↑	**0/5 = 0.0**	**OOD AR@k ↑**	**1/5 = 0.2**	**OOD AR@k ↑**	**5/5 = 1.0**

	Faster R-CNN	SIREN	OSR-ViT
OSOD says:	**Good!** Perfect ID mAP!	**Good!** Perfect ID mAP!	**Good!** Perfect ID mAP!
UAOD says:	**Good!** Perfect ID mAP and IDvOOD AUROC!	**Good!** Perfect ID mAP and IDvOOD AUROC!	**Good!** Perfect ID mAP and IDvOOD AUROC!
OSODD says:	**Bad!** The OOD Recall is zero despite good ID performance!	**Better.** But the OOD Recall could be improved.	**Good!** Both the ID mAP and OOD Recall are perfect!

Fig. 1. While other settings ignore OOD recall, the proposed OSODD task prioritizes it in addition to established metrics. In this example, a "perfect" model according to the OSOD or UAOD protocol may cause severe safety consequences.

based foundation models. OSR-ViT's bipartite architecture does not require end-to-end training, so users can easily replace either of the components with future or custom models. In this paper's instantiation of the framework we use the state-of-the-art Tunable Hybrid Proposal Network (THPN) [20] and DINOv2 [21] foundation model. We find that our simple, modular, and user-friendly OSR-ViT framework far exceeds the performance of all fully supervised open-set-specific baselines. Our framework has a particular advantage in low-data settings, where even our most lightweight configuration trained on 25% of the PASCAL VOC [22] training data outperforms all other baselines trained on 100% of the data.

Overall, our contributions are as follows:

- We create a new **joint open-set object detection and discovery task** that prioritizes both ID and OOD object detection and is more closely aligned with realistic open-set detection applications. To measure performance we develop a new comprehensive evaluation protocol and AOSP summary metric that allows for a unified comparison of previously uncompared works.
- We propose a novel OSR-ViT framework for tackling the OSODD task. OSR-ViT's modularity allows for the immediate use of the latest foundation models being developed, future-proofing our design.
- We show that OSR-ViT vastly outperforms fully-supervised alternatives with minimal configuration and no finicky end-to-end training. We also demonstrate its effectiveness with sparse-data and in the remote-sensing domain.

II. Limitations of Existing Work

A major limitation of existing "open" object detection literature is the incongruity of task goals. The evaluation protocols used in different works vary significantly, making it challenging to directly compare methods. Here, we detail the existing subdivisions of open detection.

Class-Agnostic Object Proposal. Object discovery models separate objects from background without supervision [23]. Early works in this area identify salient regions with respect to image transformations [23], [24] or noise [25]. More recent works leverage convolutional features instead of the images directly [26], [27]. Class-agnostic object proposal networks seek to maximize ID and OOD object recall (without further classification) [20], [28]–[30]. Kim et al. [29] showed that standard object proposal networks such as Region Proposal Network (RPN) [3] and its variants [31], [32] overfit to the ID categories because of its discriminative learning approach. They instead propose an Object Localization Network (OLN) which replaces the classification heads of a class-agnostic Faster R-CNN [3] with localization-quality prediction heads, yielding a model that more readily generalizes to OOD objects. Konan et al. [28] and Saito et al. [30] use unknown object masking and a background erasing augmentation, respectively, to further reduce ID-bias. While class-agnostic detection is useful, class separation is often necessary for practical tasks.

Open-Set & Unknown-Aware Object Detection. An OSOD detector should ignore OOD objects and not let the presence of OOD or "wilderness" data effect ID accuracy [9], [11], [12], [33], [34]. In other words, the goal is to simply avoid mistaking OOD objects as ID classes. Miller et al. [11] first introduce the notion of open-set object detection and use dropout sampling [35] to improve label uncertainty. Dhamija et al. [9] show that closed-set detectors tend to misclassify OOD objects as ID classes. Recently, Han et al. [34] use a contrastive feature learner to identify OOD objects from their latent representations. The limitation of OSOD is that the recall of OOD objects is irrelevant, which render these methods unfit for many real applications. An UAOD model should maximize ID performance and precisely flag any OOD objects that happen to be proposed to the classifier head [13]–[15]. Du et al. [13] generate near-OOD virtual outliers to learn more compact ID clusters to ease the separation of ID and OOD objects. SIREN [14] maps ID-class representations onto a von Mises-Fisher (vMF) distribution to provide a powerful distance-based OOD algorithm for detectors. Finally, STUD [15] distills unknown objects from video data to improve OOD detection in object detection models. A major limitation of this subdivision is that most works [13]–[15] make several unrealistic and invalid assumptions to evaluate performance. For example, they require mutually exclusive ID and OOD validation sets, and incorrectly assume that all detections with confidence over a threshold are valid ID and OOD predictions, respectively.

Open-World & Open-Vocabulary Object Detection. An OWOD model's goal is to maximize ID performance and incrementally learn new classes by forwarding it's unknown predictions to a human annotator [10], [36]–[42]. Joseph et al.'s ORE model [10] uses a conventional RPN with a contrastive clustering regularization to create a baseline. Gupta et al. [36] introduce a DETR-based [6], [7] OW-DETR model that boosts performance via attention-driven pseudo-labeling. Wu et al. [41] propose a two-branch objectness-centric model that leverages the benefits of OLN's localization-quality prediction head to improve object recall. Finally, Zohar et al.'s PROB model [42] specifically addresses unknown object recall with an additional probabilistic objectness head. While the incremental learning aspect of the OWOD task is interesting, several outside factors (e.g., threshold choice, semantic drift between tasks, training data quality) contribute heavily to perceived performance, making it difficult to judge a model's true usability. Also, while some work does enhance OWOD performance via increased unknown recall [41], [42], their OOD recall performance still remains modest. Open-Vocabulary Object Detection (OVOD) models use natural language models to enable the detector to directly generalize beyond the ID classes using text prompting [43]–[47]. While these approaches are powerful under certain circumstances (i.e., where object classes of interest are well-represented in language datasets), their practical usefulness is limited in many domains (i.e., fine-grained ship detection). For fair comparison, we do not consider OVOD baselines in this work.

III. Open-Set Object Detection and Discovery

In this section we describe in detail the OSODD task. In section III-A we formalize the problem with notation and in section III-B we detail our novel evaluation protocol.

A. Problem Formulation

As with any supervised detection task, we assume access to a training dataset that contains labels for a set of object classes of interest. We refer to this set of classes as the *known* set $\mathcal{K} = \{1, 2, \dots, C\} \subset \mathbb{N}^+$. In OSODD, we also formally acknowledge the existence of instances of *unknown* object classes $\mathcal{U} = \{C+1, \dots\} \subset \mathbb{N}^+$ that coexist with the known instances in both the training and deployment data. The goal is to train a model $\mathcal{M}$ parameterized by θ to detect and localize *all* object instances of interest in a test set (i.e., all instances in the set $\mathcal{K} \cup \mathcal{U}$). For a given test image X, the model's function is $\mathcal{M}(X; \theta) = \{[x, y, w, h, c, s]_{i=1\dots N}\}$, where x, y, w, and h denote the center coordinates, width, and height of the bounding box, respectively. The predicted class $c \in \mathcal{K} \cup \{-1, 0\}$ describes the class category that the ith prediction belongs to. Here, $c = 0$ denotes an *unknown* object of interest and $c = -1$ represents *background* (i.e., no object). Finally, each prediction has a score $s \in [0, 1]$ which represents the model's confidence that box i contains an object of class c.

B. Evaluation Protocol

A key contribution of our work is the novel evaluation procedure we develop for the OSODD task. Our evaluation uses four types of metrics to comprehensively evaluate models with minimal assumptions and thresholds:

- **Closed-Set ID mean Average Precision (ID-mAP)**: Measures the maximum potential ID-mAP by assuming all detections are knowns.
- **Class-Agnostic Average Recall (CA-AR)**: Measures performance of the proposal network by computing AR@100 assuming a single foreground (FG) class.
- **Area Under the Receiver Operating Characteristic (AUROC)**: Measures the classifier's separability across all possible thresholds. Recent UAOD works [13]–[15] only measure ID vs. OOD AUROC, since they assume that an input is always either ID or OOD (binary). However, such an assumption is inadequate for the OSODD task where we face a ternary decision: A proposal can either be an ID object, an OOD object, or background (BG). Thus, we also compute AUROC for the following separation axes: ID vs. Non-ID, OOD vs. BG, and FG vs. BG.
- **Average Open-Set Precision (AOSP)**: Our new AOSP metric provides a threshold-independent summary of a model's tradeoff between ID-mAP and OOD Recall. This metric is described in detail below.

Computing AUROC. Unlike existing works [13], [14] that use AUROC for open-set detection, we do *NOT* require that ID and OOD data are in mutually exclusive sets, and we do *NOT* assume that all high-confidence predictions are valid object regions. Instead, we take a more scrupulous approach and partition all proposed regions in the mixed test set (i.e., the images contain both ID and OOD objects) into their corresponding ID/OOD/BG bin based on their IoU overlap with the ground-truth annotations. Note that during evaluation, we always pretend that some subset of classes are OOD, so we have ground truth matches for OOD objects too. Once the predictions are partitioned, we compute our AUROC scores. ID vs. OOD and ID vs. Non-ID AUROC are computed using the proposal's ID score, which should be high for ID objects and low for OOD objects (e.g., energy [48], Mahalanobis distance [49], etc.). BG vs. OOD and FG vs. BG AUROC

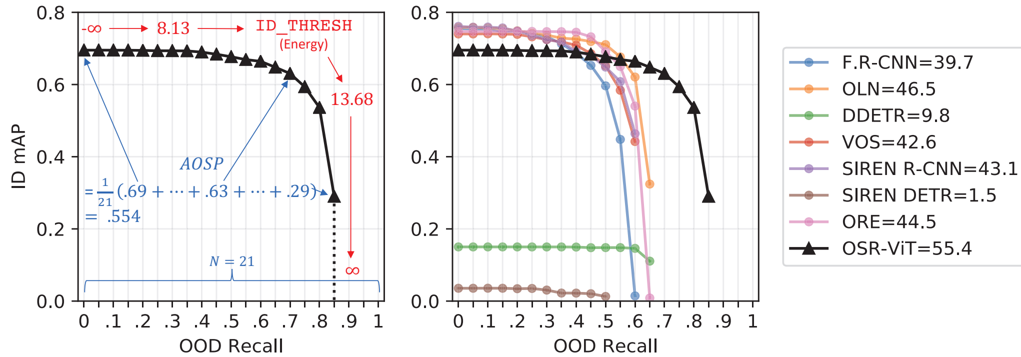


Fig. 2. Our threshold-agnostic Average Open Set Precision (AOSP) performance metric provides a holistic view of the ID-OOD performance trade-off.

are computed using the objectness score, which represents the likelihood that a region contains a foreground object (either ID or OOD).

The AOSP metric. While it is tempting to try to use mAP to measure OOD performance, this is invalid because computing *precision* requires that *all* OOD objects are labeled. Due to limitations of current datasets, we do not have exhaustive annotations for every single object. Thus, the accepted standard for measuring OOD performance is *recall* given a fixed number k of detections per image. However, we argue that the true performance of an OSODD model cannot be fully understood from a single recall measure, as it only captures performance at a one operating point. This point is determined by a model's `ID_THRESH`, the threshold which determines the minimum ID score for a prediction to be deemed an ID object. We argue that the best way to evaluate a model is to use a threshold-independent metric that summarizes the tradeoff between ID and OOD performance, as different applications require different thresholds.

To this end, we propose **Average Open-Set Precision (AOSP)**. AOSP summarizes the tradeoff between ID-mAP (@IoU=0.5) and OOD recall (@k=100 detections per image), and provides us with a single scalar metric to compare methods on the OSODD task. Fig. 2 shows a visualization of the AOSP computation. We specifically find the minimum `ID_THRESH` to achieve 21 discrete target OOD recall points in $\{0{:}.05{:}1\}$. At each of these, we set c=0 (*unknown*) for all detections with ID score $<$ `ID_THRESH` and compute ID-mAP on the updated set. AOSP is the average of ID-mAP over these OOD recall points:

$$\text{AOSP} := \frac{1}{21} \sum_{r \in \{0:.05:1\}} \text{ID-mAP}_{@\text{OOD Recall} = r} \qquad (1)$$

Note that at `ID_THRESH`$=-\infty$ every detection is deemed ID (max ID-mAP), and at `ID_THRESH`$=\infty$ every detection is deemed OOD (max OOD recall). At OOD recall points beyond the detector's maximum capability (e.g., r=$\{0.9,0.95,1\}$ in Fig. 2), we consider ID-mAP=0.

IV. OSR-ViT Modular Detection Framework

An effective OSODD model must excel at two key subtasks: (1) localizing all objects in an image, and (2) accurate discernment between ID and OOD classes. Thus, our proposed solution is a modular bipartite framework that combines an arbitrary strong proposal network with a classifier module that leverages an arbitrary Vision Transformer (ViT) [50] foundation model (see Fig. 3). **An important reason for this design choice is that in today's fast-paced ML climate, modularity is critical for future-proofing**. New state-of-the-art models are being released almost daily, necessitating frameworks that allow for seamless transitioning between solutions. The task-agnostic nature of these foundation models is also critical to being adaptable to dynamic environments and tasks. This is opposed to developing highly task-specific solutions that require extra hyperparameters, regularization terms, and underlying assumptions. We call our solution Open-Set Regions with ViT features (OSR-ViT), taking inspiration from the seminal "Regions with CNN features" (R-CNN) model family [1]–[3], [51]. The remainder of this section details the Proposal Network (section IV-A), the Foundational Classifier (section IV-B), and model training (section IV-C).

A. Proposal Network

The upper bound of overall OSODD performance is directly predicated on the model's ability to discern foreground objects vs. background, as even a detector with a perfect classifier is useless if true positive regions are never proposed in the first place. One major pitfall of open object proposal is overfitting to ID classes. Basic supervised proposal networks like RPN [3] inherently overfit due to their discriminative objective [29]. Several recent works have tried to combat this issue [28], [29], however it has been shown that incorporating such dedicated proposal networks directly into end-to-end open-set/world detectors yields worse overall performance [41], [42]. The other major pitfall is a lack of adaptability. A practically useful OSODD proposal network should be able to be adapted to different application requirements [20]. For example, a security system should prioritize the detection of a couple of key ID classes (e.g., person, car) while ignoring unrelated OOD objects. However, a household robot should

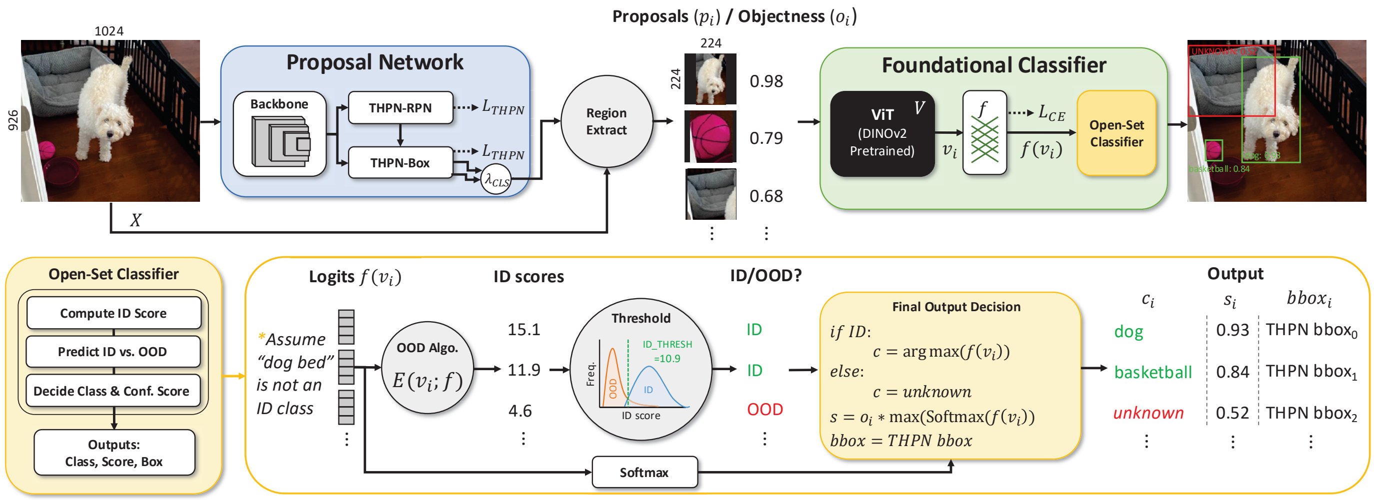


Fig. 3. Our OSR-ViT framework consists of two independently-trained models working in conjunction: (1) a class-agnostic Proposal Network, and (2) a ViT-powered Foundational Classifier. This allows for seamless integration of new or future models.

be much more generalizable to rare and unexpected object classes.

In our instantiation of OSR-ViT, we use a Tunable Hybrid Proposal Network (THPN) [20]. THPN is a state-of-the-art CNN-based proposal network that learns a hybrid objectness representation via dual prediction heads. Critically, THPN provides a single hyperparameter $\lambda_{CLS} \in [0, 1]$ which balances both the loss contribution and final confidence score from each prediction head. The larger λ_{CLS} is set, the more ID-biased the resulting model is, meaning the more propensity the model has for detecting ID objects at the cost of some OOD objects. THPN also leverages a self-training optimization procedure [52] that significantly enhances data efficiency, allowing for impressive performance in low-data or semi-supervised settings. We emphasize that OSR-ViT users can seamlessly plug-and-play with any proposal model of their choosing. For example, if an organization has developed an exquisite proposal network for a specific remote-sensing task, that network can be leveraged here.

B. Foundational Classifier

The recent emergence of large-scale foundation models has begun to revolutionize the pipeline of training and deploying vision AI. Open-source models such as CLIP [53] and DINOv2 [21] are trained on hundreds of millions of images for tens of thousands of GPU hours. They provides users with off-the-shelf task-agnostic models that can be made task-specific with a minimal fine-tuning stage, and outperform supervised specialist models. This strong performance is due to the highly expressive representations that are encoded by the Vision Transformer (ViT) architecture. However, we argue that the true power of these foundation models extends far beyond closed-set recognition. **Our hypothesis is that the highly descriptive ViT representations of the object proposals will enable effective ID *and* OOD separation**. In this work, we use a DINOv2 [21] model as the feature extractor of the foundational classifier in the OSR-ViT. DINOv2 is trained on the extensive LVD-142M dataset [21], meaning it is fully capable of well-representing a wide variety of image domains and object types. Again, we encourage users to plug-and-play beyond DINOv2 with whatever new or custom foundation model they see fit.

As shown in Fig. 3, the input image is first processed by the proposal network $\text{PropNet}(X) : \mathbb{R}^D \rightarrow \{(p_i^{bbox}, o_i)\}_{i=1}^{100}$, which maps a D-dimensional input image X to N=100 pairs of object proposal boxes p_i^{bbox} and their corresponding predicted objectness o_i. The pixel region of each proposal is then cropped from the image and resized to the 224x224 resolution that the DINOv2 model can ingest. We call these resulting resized proposal "images" p_i. The proposal images are then forwarded through the ViT feature extractor $V(p) : \mathbb{R}^{224 \cdot 224} \rightarrow \mathbb{R}^d$, where d is the dimensionality of the ViT's feature space. We refer to the ViT representation of proposal p_i as v_i. We use a simple 2-layer fully connected (non-linear) module $f(v) : \mathbb{R}^d \rightarrow \mathbb{R}^C$ on top of the ViT feature extractor to enable C-way classification. The output logits of each proposal $f(v_i)$ are then forwarded to the Open-Set Classifier which makes the final output decision.

Reaching a final detection involves two sequential predictions. First, we must predict if a proposal is ID or OOD. We use a post-hoc Energy-based OOD detection algorithm [48] that uses a proposal's free energy as its ID score:

$$E(v_i; f) = -T \cdot \log \sum_{j}^{C} e^{f_j(v_i)/T} \quad (2)$$

where T is a temperature parameter. Note that for a given proposal, the larger this energy score is, the more likely it is to be an ID class object. If $-E(v_i; f) > \texttt{ID\_THRESH}$ we call the ith proposal an ID object, else we call it *unknown*. For deployment, one would use a validation set to choose a reasonable `ID_THRESH`.

The second decision that must be made by the Open-Set Classifier is the final output class c_i and confidence score s_i. Fig. 3 shows this decision in the "Final Output Decision" box. If p_i is deemed OOD, the assigned class label is c_i = *unknown*, but if it is deemed ID then $c_i = \arg\max(f(v_i))$. Regardless of class label, the confidence score is the product of the predicted objectness from the proposal network o_i and the max Softmax confidence over the ID classes:

$$s_i = o_i \cdot \max(\text{Softmax}(f(v_i))). \tag{3}$$

Note that many existing works [13]–[15] simply use the maximum Softmax score for OOD predictions. Although this may be valid for the binary open-set *classification* task (ID vs. non-ID), it is not appropriate for the ternary open-set *detection* task (ID vs. OOD vs. BG). In other words, **just because a proposal does not significantly excite any one ID output node does NOT necessarily mean that it does not have strong general object features**. For this reason, our score measure s_i directly incorporates the objectness score from the class agnostic proposal network, meaning the resulting scores for both ID and OOD predictions will be more appropriately calibrated. Finally, we reuse the boxes output by the proposal network as the final box predictions.

C. Training

Much of OSR-ViT's user-friendliness is due to its disentangled training of the proposal network and the foundational classifier. This allows users to easily incorporate new custom or off-the-shelf models for either role. In this work, we optimize the THPN following the procedure outlined in the paper [20]. We adapt the foundational classifier separately, and in two stages. In the first stage, we freeze the DINOv2-pretrained ViT and update the fully connected classifier module f using cross-entropy loss for 50 epochs. To improve model flexibility while maintaining the expressiveness of the ViT's pre-trained representations, we then perform a short 5-epoch fine-tuning stage in which we train the ViT and the classifier module together with a much smaller learning rate.

V. Experiments

To evaluate models on the OSODD task, we create three separate benchmarks which offer far more diversity than contemporary literature [9], [11], [13], [14], [33], [34]. Section V-A contains our Natural Imagery Benchmark, Section V-B contains our Limited Data Benchmark, and section V-C covers model performance on the Ships Benchmark. Finally, in section V-D we perform additional analysis on our OSR-ViT method.

A. Natural Imagery Benchmark

This benchmark considers two cross-dataset transfer tasks between common natural imagery datasets. The first is to train on the 20-class PASCAL VOC [22] training dataset and test on the 80-class COCO [54] validation set. In this case, the OOD classes are the non-VOC classes of COCO. The second is to train on the COCO training set and test on 40,000 images from the 365-class Objects365 [55] dataset. Here, the OOD classes are the non-COCO classes of Objects365. Since the Objects365 label space is more granular we consider all synsets or hyponyms of the COCO classes as ID. Table I contains the results for this benchmark. Note that the "-S", "-B", and "-L" specifiers on the DINOv2 models indicate the size of the ViT. Our OSR-ViT method outperforms all baselines on all OOD-related metrics on both tasks. In general, OSR-ViT's margin of improvement over the baselines is greater on VOC→COCO compared to COCO→Objects365. This is because the stronger supervised baselines (e.g., DETR-based models) can learn better representations of the ID classes in tasks with more data.

OSR-ViT significantly outperforms all baselines in terms of CA-AR, showcasing the utility of a non-ID-biased proposal network like THPN. The relatively mediocre AOSP and CA-AR scores from the major OWOD methods (ORE [10], OW-DETR [36], and PROB [42]) shows that the incremental learning aspect of the OWOD task does indeed distract from the relatively poor OOD recall, justifying the need for our OSODD task. Finally, OSR-ViT excels in terms of classifier separability (i.e., AUROC metrics). The strong ID score-based separation (ID vs. OOD, ID vs. Non-ID) demonstrates that ViT's strong nuanced representations allow superior OOD detectability, even compared to strong regularized UAOD baselines such as VOS [13] and SIREN [14] that are specifically designed for this capability. The objectness-based separation (OOD vs. BG, FG vs. BG) is also much better than the baselines, with the FG vs. BG AUROC being 16.02% higher than the best baseline (OLN).

As expected, the size of the DINOv2 ViT does positively correlate with performance, but even DINOv2-S can provide state-of-the-art performance on both tasks in terms of AOSP. On the moderately-scaled VOC→COCO task, the smallest DINOv2-S is still sufficient to outperform the UAOD methods in terms of classifier separability, but on the larger COCO→Objects365 task the larger DINOv2-L is required to beat SIREN-DETR [14]. One limitation of our particular OSR-ViT configuration is that it trades off far superior OOD recall for slightly worse closed-set ID-mAP. Our analysis shows that this is not due to Foundational Classifier error, but rather to the ID/OOD tradeoff made by the THPN proposal network. Here, we configure the THPN in these experiments with λ_{CLS}=.10, yielding a more OOD-biased model. In additional experiments we explore the impact of λ_{CLS} and find that this ID-mAP discrepancy can be minimized.

B. Limited Data Benchmark

While performance on large-scale benchmarks is important, in many scenarios and applications we do not have training datasets with hundreds of thousands of annotations at our disposal. For this reason, we devise a Limited Data Benchmark in which models are trained on a random (class-balanced) set of 25%, 50%, and 75% of the VOC training annotations and tested on the COCO validation set.

TABLE I
RESULTS ON THE NATURAL IMAGERY BENCHMARK TASKS.

Data	Training	Model	OOD Algo.	AOSP	ID-mAP	CA-AR	IDvOOD AUROC	IDvNONID AUROC	OODvBG AUROC	FGvBG AUROC
VOC → COCO	Plain Supervised	Faster R-CNN	Energy	17.8	31.1	37.2	73.41	64.00	59.77	65.64
			Mahalanobis	18.0	31.1	37.2	56.27	68.32	59.77	65.64
		OLN	Energy	18.8	30.0	38.5	72.42	64.94	59.44	66.29
			Mahalanobis	18.4	30.0	38.5	51.66	65.81	59.44	66.29
		Deformable DETR	Energy	10.1	**34.6**	33.3	58.77	69.05	58.58	57.62
			Mahalanobis	9.8	**34.6**	33.3	55.25	63.35	58.58	57.62
	VOS	Faster R-CNN	Energy	18.6	31.5	36.3	78.68	73.55	61.44	73.68
	SIREN	Faster R-CNN	SIREN-KNN	17.3	31.3	36.7	82.74	77.15	58.91	64.23
		Deformable DETR	SIREN-KNN	12.0	33.6	33.1	75.87	82.70	57.98	57.74
	ORE	Faster R-CNN	Energy	18.3	28.0	35.4	75.13	74.90	53.91	63.01
	OW-DETR	Deformable DETR	Direct Pred.	10.7	30.2	30.9	-	-	-	-
	PROB	Deformable DETR	Direct Pred.	12.6	32.5	31.7	-	-	-	-
	OSR-ViT	THPN+DINOv2-S	Energy	23.6	30.2	**43.2**	84.79	85.08	63.26	80.69
		THPN+DINOv2-B	Energy	25.0	31.4	**43.2**	86.49	**86.28**	63.42	81.86
		THPN+DINOv2-L	Energy	**25.1**	31.5	**43.2**	**87.57**	85.52	**64.87**	**82.31**
COCO → Obj365	Plain Supervised	Faster R-CNN	Energy	17.6	24.5	44.1	61.84	65.10	63.62	66.99
			Mahalanobis	14.6	24.5	44.1	53.71	56.97	63.62	66.99
		OLN	Energy	17.5	23.0	44.9	62.66	65.06	63.25	66.32
			Mahalanobis	13.6	23.0	44.9	52.39	56.60	63.25	66.32
		Deformable DETR	Energy	17.3	**29.0**	43.9	55.57	60.50	58.04	61.34
			Mahalanobis	13.2	**29.0**	43.9	48.11	46.98	58.04	61.34
	VOS	Faster R-CNN	Energy	17.8	24.4	43.6	65.20	68.16	63.25	67.34
	SIREN	Faster R-CNN	SIREN-KNN	17.0	24.4	43.4	68.34	68.68	62.91	66.99
		Deformable DETR	SIREN-KNN	8.2	28.8	43.4	71.45	73.75	58.43	60.75
	ORE	Faster R-CNN	Energy	16.9	22.7	42.4	62.35	66.17	60.09	64.07
	OSR-ViT	THPN+DINOv2-S	Energy	18.7	23.9	**49.7**	67.01	73.57	68.55	73.89
		THPN+DINOv2-B	Energy	19.7	25.1	**49.7**	70.72	75.81	67.16	73.70
		THPN+DINOv2-L	Energy	**20.2**	25.7	**49.7**	**71.60**	**76.67**	**67.33**	**74.04**

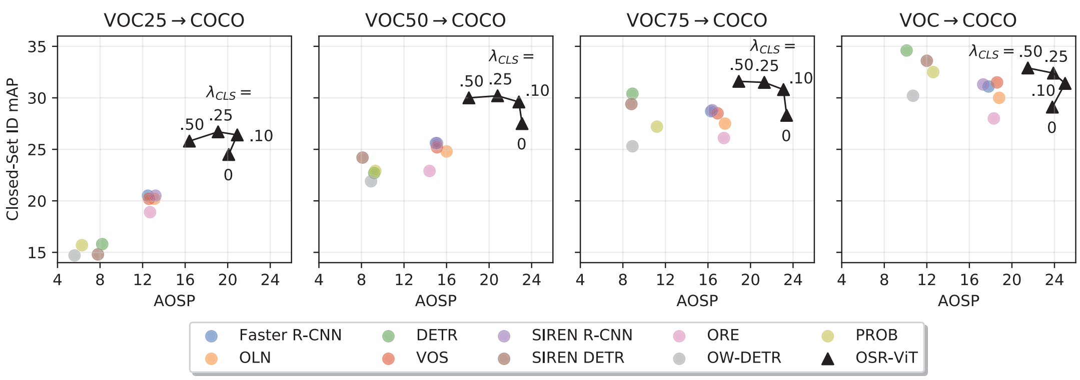


Fig. 4. While supervised baselines struggle in the data-constrained settings of our Limited Data Benchmark, our OSR-ViT model maintains good performance.

Fig. 4 visualizes the results from this benchmark as AOSP vs. closed-set ID mAP. In this experiment we vary the THPN λ_{CLS} parameter in our OSR-ViT(-B) model. The key take-away from this result is that OSR-ViT maintains ID-mAP and AOSP much better than fully supervised models when training data gets scarce. In fact, the most lightweight OSR-ViT model (THPN(λ_{CLS}=.10)+DINOv2-S) trained on 25% of the VOC data achieves 20.6% AOSP, which is higher than *any* baseline method trained on 100% of the VOC data! It should also be noted that the CA-AR of the OSR-ViT models trained on the 25% split is 38.4%, which essentially matches the highest performing baseline (i.e., OLN) trained on the 100% split. This performance can be mainly attributed to the ViT's ability to generalize well with very limited task-specific data.

As discussed in section V-A above, some DETR-based baselines outperform our OSR-ViT configuration in terms of closed-set ID-mAP. However, this challenging benchmark reveals that these methods require significant training data to reach this level of performance. Notice that decreasing the labeled training annotations even to 75% of the original number drastically reduces the performance of these models. In a scenario like VOC25→COCO, where we have less than 12,000 training annotations, these methods are essentially useless. Finally, these results showcase the effect of THPN's λ_{CLS} parameter. In general, the higher we set λ_{CLS}, the higher the ID mAP. Using an adaptable proposal network like THPN in the OSR-ViT model greatly increases its flexibility, as we can more effectively configure the model for a given set of requirements.

C. Ships Benchmark

Our final benchmark evaluates performance in the remote-sensing image domain. We consider the ShipRSImageNet

TABLE II
RESULTS ON THE SHIPS BENCHMARK.

Training	Model	OOD Algo.	AOSP	ID-mAP	CA-AR	IDvOOD AUROC	IDvNONID AUROC	OODvBG AUROC	FGvBG AUROC
Plain Supervised	Faster R-CNN	Energy	39.7	60.8	58.6	70.55	64.65	78.00	77.38
		Mahalanobis	40.9	60.8	58.6	45.35	55.96	78.00	77.38
	OLN	Energy	46.5	**61.3**	59.7	73.17	65.99	78.59	76.14
		Mahalanobis	46.9	**61.3**	59.7	41.66	52.07	78.59	76.14
	Deformable DETR	Energy	9.8	8.5	32.0	49.33	81.21	65.56	64.63
		Mahalanobis	9.5	8.5	32.0	50.70	78.45	65.56	64.63
VOS	Faster R-CNN	Energy	42.6	59.5	59.0	71.49	68.45	75.05	72.70
SIREN	Faster R-CNN	SIREN-KNN	43.1	60.7	58.6	77.11	75.03	72.47	70.44
	Deformable DETR	SIREN-KNN	1.5	1.3	19.6	49.72	79.09	64.62	64.25
ORE	Faster R-CNN	Energy	44.5	58.7	54.1	74.61	62.26	68.30	63.05
OW-DETR	Deformable DETR	Direct Pred.	8.3	11.4	31.7	-	-	-	-
PROB	Deformable DETR	Direct Pred.	14.2	12.5	38.5	-	-	-	-
OSR-ViT	THPN+DINOv2-S	Energy	53.4	57.2	**64.3**	75.22	**87.78**	94.07	95.49
	THPN+DINOv2-B	Energy	**55.4**	58.9	**64.3**	**77.16**	85.72	**94.16**	**95.81**

TABLE III
MODEL DESIGN ANALYSIS ON THE VOC→COCO TASK.

Model	OOD Algo.	AOSP	ID-mAP	CA-AR	IDvOOD AUROC	IDvNONID AUROC	OODvBG AUROC	FGvBG AUROC
THPN+DINOv2-B	MSP	24.8	31.4	**43.2**	83.97	83.33	63.42	81.86
	MaxLogit	**25.0**	31.4	**43.2**	86.48	86.20	63.42	81.86
	ODIN	**25.0**	31.4	**43.2**	86.00	85.41	63.42	81.86
	Energy	**25.0**	31.4	**43.2**	**86.49**	**86.28**	63.42	81.86
FT→No FT	Energy	24.8	31.1	**43.2**	85.46	85.41	63.87	**81.88**
THPN→Faster R-CNN	Energy	20.0	**32.4**	37.2	84.38	82.51	61.59	71.81
DINOv2-B→CLIP-B	Energy	22.3	29.0	**43.2**	78.74	83.53	**64.70**	80.21

dataset [56], which contains overhead imagery of coastal regions with 50 fine-grained ship classes. Here, we manually create the ID/OOD class split by deeming all "other" ship categories as OOD. An implicit challenge of this dataset is that there are relatively few annotations to train on compared to the natural imagery benchmarks (i.e., 2k ship instances compared to 47k VOC instances). Table II contains the results. Even in this different domain, OSR-ViT beats all fully-supervised baselines in terms of AOSP and CA-AR. Our method lags OLN slightly in ID-mAP, but achieves a substantial 8.5% higher AOSP than OLN's best post-hoc OOD algorithm (Mahalanobis [49]). OSR-ViT's classifier separability is also superior, specifically in terms of objectness-based separability. Our method outperforms the closest baseline (OLN) in OOD vs. BG AUROC and FG vs. BG AUROC by 15.57% and 19.67%, respectively! We note that DETR-based methods were unable to converge to a reasonable solution on this smaller-scale task, highlighting their limitations in many settings.

D. OSR-ViT Performance Analysis

OSR-ViT's modular design allows for arbitrary proposal networks and feature extractors to be incorporated. In Table III we investigate several different variants of our base configuration using THPN and DINOv2-B on the VOC→COCO task. The exact choice of post-hoc OOD algorithm does not have a massive effect on performance, although Energy is the best overall. The FT→No FT row represents our base configuration but without the 5-epoch end-to-end fine-tuning step described in section IV-B. While this fine-tuning is not necessary, it does boost overall performance. When we swap THPN (λ_{CLS}=.10) for a class-agnostic Faster R-CNN [3] proposal network, we get noticeably worse AOSP and CA-AR, but better ID-mAP due to Faster R-CNN's inherent ID bias. But again, it should be noted that a THPN with λ_{CLS}=.50 can outperform Faster R-CNN with an ID mAP of 32.9. Finally, we compare the impact of swapping the DINOv2 foundation model for a CLIP [53] model of the same size. We find that OSR-ViT with CLIP achieves substandard results across the board.

Fig. 5 depicts 2D t-SNE visualizations [57] of the penultimate object features of four different models on the VOC→COCO task. Note that the colored circle, star, and triangle markers represent detections that positively match ID ground-truth objects, the chartreuse squares represent detections matched to OOD ground truth objects, and the black squares represent detections matched to background. Ultimately, the performance of a model is directly related to how separable these features are, with more compact ID and OOD clusters being indicative of better models. The key takeaway from this analysis is that the DINOv2 feature extractor does a far better job of separating the OOD objects from ID objects and BG compared to the baselines. These findings support our hypothesis from section IV-B: The DINOv2 representations are indeed nuanced enough to not only distinguish ID and OOD objects, but also different OOD objects from each other. This quality of representation is generally not feasible with task-specific supervised training alone. Finally, this OOD separability would make our method a powerful starting starting

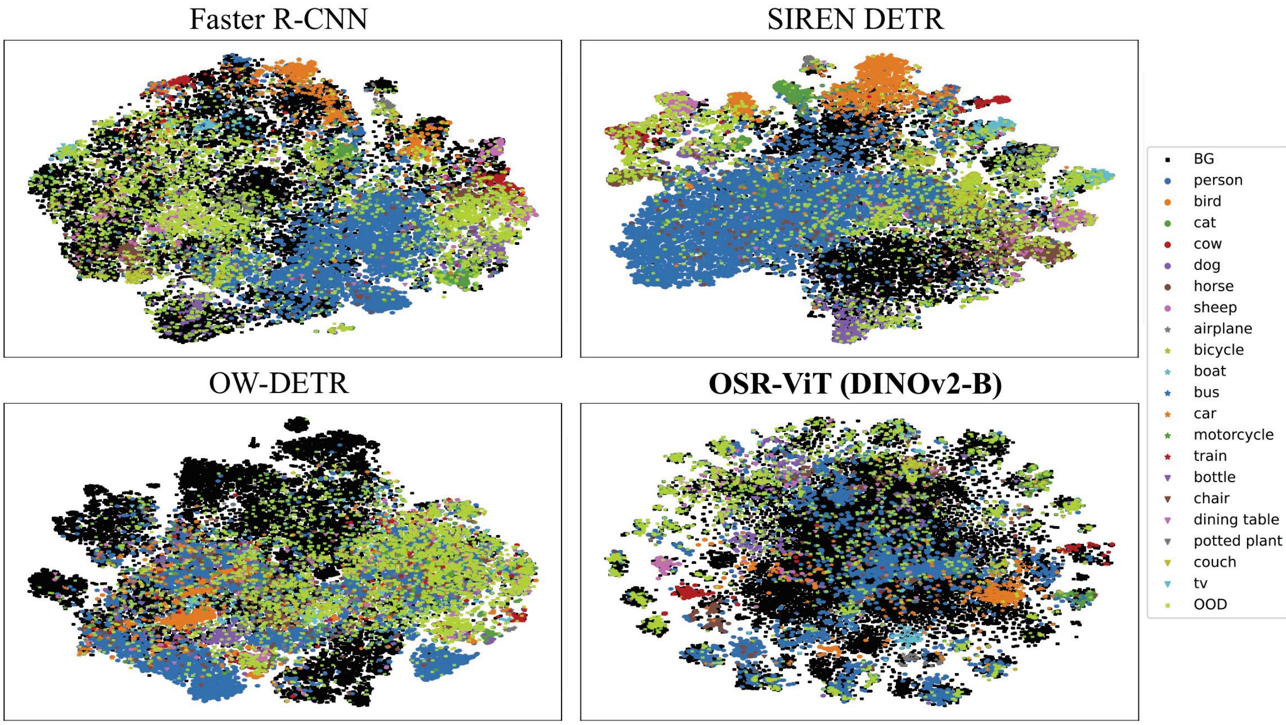


Fig. 5. 2D t-SNE visualization of penultimate features on the VOC→COCO task. OSR-ViT models generate the most compact ID-class clusters, aiding in ID vs. OOD separation. Also, OSR-ViT's ability to segregate OOD instances into different sub-clusters is infeasible with only task-specific supervision.

point for the OWOD task which incrementally learns new classes, but we leave this for future work.

VI. Conclusion

As ML becomes more and more ubiquitous in our real-world systems, it is important to keep safety at the forefront of model design. In this work, we identify a serious vulnerability of state-of-the-art "open-set object detection" models: the detection of unknown objects is not explicitly prioritized. We use this finding to motivate a new OSODD task, and create an evaluation protocol that allows different related works to be directly compared to each other for the first time. We also introduce a modular new OSR-ViT framework that leverages self-contained proposal networks and and off-the-shelf ViT models in a plug-and-play fashion to achieve far superior performance to all previous supervised baselines.

Acknowledgment

This work is supported by NSF CNS-2112562.

References

[1] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
[2] R. B. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.
[3] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
[4] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2016.
[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., 2020.
[7] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: deformable transformers for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2021.
[8] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, 2013.
[9] A. R. Dhamija, M. Günther, J. Ventura, and T. E. Boult, "The overlooked elephant of object detection: Open set," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.
[10] K. J. Joseph, S. H. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[11] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.
[12] D. Miller, F. Dayoub, M. Milford, and N. Sünderhauf, "Evaluating merging strategies for sampling-based uncertainty techniques in object detection," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
[13] X. Du, Z. Wang, M. Cai, and Y. Li, "VOS: learning what you don't know by virtual outlier synthesis," *CoRR*, vol. abs/2202.01197, 2022.
[14] X. Du, G. Gozum, Y. Ming, and Y. Li, "SIREN: shaping representations for detecting out-of-distribution objects," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
[15] X. Du, X. Wang, G. Gozum, and Y. Li, "Unknown-aware object detection: Learning what you don't know from videos in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[16] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2702–2719, 2019.

[17] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[18] W. Moon, J. Park, H. S. Seong, C.-H. Cho, and J.-P. Heo, "Difficulty-aware simulator for open set recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 365–381.

[19] H. Deng, Z. Zhang, J. Bao, and X. Li, "Medical open set recognition via intra-class clustering," in *Medical Imaging with Deep Learning*, 2024.

[20] M. Inkawhich, N. Inkawhich, H. Li, and Y. Chen, "Tunable hybrid proposal networks for the open world," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[21] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Huang, S. Li, I. Misra, M. G. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *CoRR*, vol. abs/2304.07193, 2023.

[22] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, 2010.

[23] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[24] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[25] J. Ekekrantz, N. Bore, R. Ambrus, J. Folkesson, and P. Jensfelt, "Unsupervised object discovery and segmentation of rgbd-images," *CoRR*, vol. abs/1710.06929, 2017.

[26] A. Arun, C. V. Jawahar, and M. P. Kumar, "Dissimilarity coefficient based weakly supervised object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[28] S. Konan, K. J. Liang, and L. Yin, "Extending one-stage detection with open-world proposals," *CoRR*, vol. abs/2201.02302, 2022.

[29] D. Kim, T. Lin, A. Angelova, I. S. Kweon, and W. Kuo, "Learning open-world object proposals without learning to classify," *IEEE Robotics Autom. Lett.*, vol. 7, 2022.

[30] K. Saito, P. Hu, T. Darrell, and K. Saenko, "Learning to detect every thing in an open world," *CoRR*, vol. abs/2112.01698, 2021.

[31] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade RPN: delving into high-quality region proposal network with adaptive convolution," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[32] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] D. Miller, N. Sünderhauf, M. Milford, and F. Dayoub, "Uncertainty for identifying open-set errors in visual object detection," *IEEE Robotics Autom. Lett.*, vol. 7, 2022.

[34] J. Han, Y. Ren, J. Ding, X. Pan, K. Yan, and G. Xia, "Expanding low-density latent regions for open-set object detection," *CoRR*, vol. abs/2203.14911, 2022.

[35] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML)*, M. Balcan and K. Q. Weinberger, Eds., 2016.

[36] A. Gupta, S. Narayan, K. J. Joseph, S. Khan, F. S. Khan, and M. Shah, "OW-DETR: open-world detection transformer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[37] X. Zhao, X. Liu, Y. Shen, Y. Qiao, Y. Ma, and D. Wang, "Revisiting open world object detection," *CoRR*, vol. abs/2201.00471, 2022.

[38] Z. Wu, Y. Lu, X. Chen, Z. Wu, L. Kang, and J. Yu, "UC-OWOD: unknown-classified open world object detection," in *European Conference on Computer Vision (ECCV)*, 2022.

[39] S. Yang, P. Sun, Y. Jiang, X. Xia, R. Zhang, Z. Yuan, C. Wang, P. Luo, and M. Xu, "Objects in semantic topology," in *International Conference on Learning Representations (ICLR)*, 2022.

[40] J. Yu, L. Ma, Z. Li, Y. Peng, and S. Xie, "Open-world object detection via discriminative class prototype learning," in *IEEE International Conference on Image Processing (ICIP)*, 2022.

[41] Y. Wu, X. Zhao, Y. Ma, D. Wang, and X. Liu, "Two-branch objectness-centric open world detection," in *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis (HCMA@MM)*, 2022.

[42] O. Zohar, K. Wang, and S. Yeung, "PROB: probabilistic objectness for open world object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[43] A. Zareian, K. D. Rosa, D. H. Hu, and S. Chang, "Open-vocabulary object detection using captions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[44] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Open-vocabulary DETR with conditional matching," in *European Conference on Computer Vision*, 2022.

[45] H. A. Rasheed, M. Maaz, M. U. Khattak, S. H. Khan, and F. S. Khan, "Bridging the gap between object and image-level representations for open-vocabulary detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[46] X. Wu, F. Zhu, R. Zhao, and H. Li, "CORA: adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[47] M. Minderer, A. A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," *CoRR*, vol. abs/2306.09683, 2023.

[48] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[49] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[51] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[52] D. hyun Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," 2013.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[54] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014.

[55] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[56] Z. Zhang, L. Zhang, Y. Wang, P. Feng, and R. He, "Shiprsimagenet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 2021.

[57] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, 2008.