OW-DETR: Open-world Detection Transformer

Akshita Gupta[*1] Sanath Narayan[*1] K J Joseph[2,4]
Salman Khan[4,3] Fahad Shahbaz Khan[4,5] Mubarak Shah[6]
[1]Inception Institute of Artificial Intelligence [2]IIT Hyderabad [3]Australian National University
[4]Mohamed Bin Zayed University of Artificial Intelligence [5]CVL, Linköping University [6]University of Central Florida

Abstract

Open-world object detection (OWOD) is a challenging computer vision problem, where the task is to detect a known set of object categories while simultaneously identifying unknown objects. Additionally, the model must incrementally learn new classes that become known in the next training episodes. Distinct from standard object detection, the OWOD setting poses significant challenges for generating quality candidate proposals on potentially unknown objects, separating the unknown objects from the background and detecting diverse unknown objects. Here, we introduce a novel end-to-end transformer-based framework, OW-DETR, for open-world object detection. The proposed OW-DETR comprises three dedicated components namely, attention-driven pseudo-labeling, novelty classification and objectness scoring to explicitly address the aforementioned OWOD challenges. Our OW-DETR explicitly encodes multi-scale contextual information, possesses less inductive bias, enables knowledge transfer from known classes to the unknown class and can better discriminate between unknown objects and background. Comprehensive experiments are performed on two benchmarks: MS-COCO and PASCAL VOC. The extensive ablations reveal the merits of our proposed contributions. Further, our model outperforms the recently introduced OWOD approach, ORE, with absolute gains ranging from 1.8% to 3.3% in terms of unknown recall on MS-COCO. In the case of incremental object detection, OW-DETR outperforms the state-of-the-art for all settings on PASCAL VOC. Our code is available at https://github.com/akshitac8/OW-DETR.

1. Introduction

Open-world object detection (OWOD) relaxes the closed-world assumption in popular benchmarks, where only seen classes appear at inference. Within the OWOD paradigm [15], at each training episode, a model learns to detect a given set of *known* objects while simultaneously capable of identifying *unknown* objects. These flagged unknowns can then be forwarded to an oracle (*e.g.*, human annotator), which can label a few classes of interest. Given these *new knowns*, the model would continue updating its knowledge incrementally without retraining from scratch on the previously known classes. This iterative learning process continues in a cycle over the model's life-span.

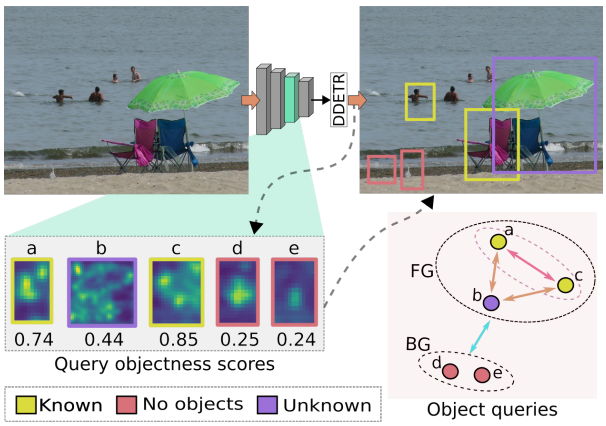

Figure 1. **Visual illustration of the proposed OW-DETR for open-world object detection (OWOD)**. Here, attention maps obtained from the intermediate features are utilized to score the object queries. The objectness scores of queries are then used to identify the pseudo-unknowns. A separation is enforced between these pseudo-unknowns and ground-truth knowns to detect novel classes. In addition, a separation is also learned between the background and foreground (knowns + unknowns) for effective knowledge transfer from known to unknown class w.r.t. characteristics of foreground objects. Our OW-DETR explicitly encodes multi-scale context, has less inductive bias, and assumes no supervision for unknown objects, thus well suited for OWOD problem.

The identification of unknown object classes in OWOD setting poses significant challenges for conventional detectors. *First,* besides an accurate proposal set for seen objects, a detector must also generate quality candidate boxes for potentially unknown objects. *Second,* the model should be able to separate unknown objects from the background utilizing its knowledge about the already seen objects, thereby learning what constitutes a valid object. *Finally*, objects

[*]Equal contribution

of different sizes must be detected while flexibly modeling their rich context and relations with co-occurring objects.

Recently, the work of [15] introduces an open-world object detector, ORE, based on the two-stage Faster R-CNN [31] pipeline. Since unknown object annotations are not available during training in the open-world paradigm, ORE proposes to utilize an auto-labeling step to obtain a set of pseudo-unknowns for training. The auto-labeling is performed on class-agnostic proposals output by a region proposal network (RPN). The proposals not overlapping with the ground-truth (GT) known objects but having high 'objectness' scores are auto-labeled as unknowns and used in training. These auto-labeled unknowns are then utilized along with GT knowns to perform latent space clustering. Such a clustering attempts to separate the multiple known classes and the unknown class in the latent space and aids in learning a prototype for the unknown class. Furthermore, ORE learns an energy-based binary classifier to distinguish the unknown class from the class-agnostic known class.

While being the first to introduce and explore the challenging OWOD problem formulation, ORE suffers from several issues. (i) ORE relies on a held-out validation set with weak supervision for the unknowns to estimate the distribution of novel category in its energy-based classifier. (ii) To perform contrastive clustering, ORE learns the unknown category with a single latent prototype, which is insufficient to model the diverse intra-class variations commonly present in the unknown objects. Consequently, this can lead to a sub-optimal separation between the knowns and unknowns. (iii) ORE does not explicitly encode long-range dependencies due to a convolution-based design, crucial to capture the contextual information in an image comprising diverse objects. Here, we set out to alleviate the above issues for the challenging OWOD problem formulation.

Contributions: Motivated by the aforementioned observations, we introduce a multi-scale context aware detection framework, based on vision transformers [37], with dedicated components to address open-world setting including attention-driven pseudo-labeling, novelty classification and objectness scoring for effectively detecting unknown objects in images (see Fig. 1). Specifically, in comparison to the recent OWOD approach ORE [15], that uses a two-stage CNN pipeline, ours is a single-stage framework based on transformers that require less inductive biases and can encode long-term dependencies at multi-scales to enrich contextual information. Different to ORE, which relies on a held-out validation set for estimating the distribution of novel categories, our setting assumes no supervision given for the unknown and is closer to the true open-world scenario. Overall, our novel design offers more flexibility with broad context modeling and less assumptions to address the open-world detection problem. Our main contributions are:

- We propose a transformer-based open-world detector, OW-DETR, that better models the context with mutli-scale self-attention and deformable receptive fields, in addition to fewer assumptions about the open-world setup along with reduced inductive biases.
- We introduce an attention-driven pseudo-labeling scheme for selecting the object query boxes having high attention scores but not matching any known class box as unknown class. The pseudo-unknowns along with the ground-truth knowns are utilized to learn a novelty classifier to distinguish the unknown objects from the known ones.
- We introduce an objectness branch to effectively learn a separation between foreground objects (knowns, pseudo-unknowns) and the background by enabling knowledge transfer from known classes to the unknown class w.r.t. the characteristics that constitute a foreground object.
- Our extensive experiments on two popular benchmarks demonstrate the effectiveness of the proposed OW-DETR. Specifically, OW-DETR outperforms the recently introduced ORE for both OWOD and incremental object detection tasks. On MS-COCO, OW-DETR achieves absolute gains ranging from 1.8% to 3.3% in terms of unknown recall over ORE.

2. Open-world Detection Transformer

Problem Formulation: Let $\mathcal{K}^t = \{1, 2, \cdots, C\}$ denote the set of known object categories at time t. Let $\mathcal{D}^t = \{\mathcal{I}^t, \mathcal{Y}^t\}$ be a dataset containing N images $\mathcal{I}^t = \{I_1, \cdots, I_N\}$ with corresponding labels $\mathcal{Y}^t = \{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_N\}$. Here, each $\boldsymbol{Y}_i = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_K\}$ denotes the labels of a set of K object instances annotated in the image with $\boldsymbol{y}_k = [l_k, x_k, y_k, w_k, h_k]$, where $l_k \in \mathcal{K}^t$ is the class label for a bounding box represented by x_k, y_k, w_k, h_k. Furthermore, let $\mathcal{U} = \{C+1, \cdots\}$ denote a set of unknown classes that might be encountered at test time.

As discussed in Sec. 1, in the open-world object detection (OWOD) setting, a model $\mathcal{M}^t$ at time t is trained to identify an unseen class instance as belonging to the unknown class (denoted by label 0), in addition to detecting the previously encountered known classes C. A set of unknown instances $\boldsymbol{U}^t \subset \mathcal{U}$ identified by $\mathcal{M}^t$ are then forwarded to an oracle, which labels n novel classes of interest and provides a corresponding set of new training examples. The learner then incrementally adds this set of new classes to the known classes such that $\mathcal{K}^{t+1} = \mathcal{K}^t + \{C+1, \cdots, C+n\}$. For the previous classes $\mathcal{K}^t$, only few examples can be stored in a bounded memory, mimicking privacy concerns, limited compute and memory resources in real-world settings. Then, $\mathcal{M}^t$ is incrementally trained, without retraining from scratch on the whole dataset, to obtain an updated model $\mathcal{M}^{t+1}$ which can detect all object classes in $\mathcal{K}^{t+1}$. This cycle continues over the life-span of the de-

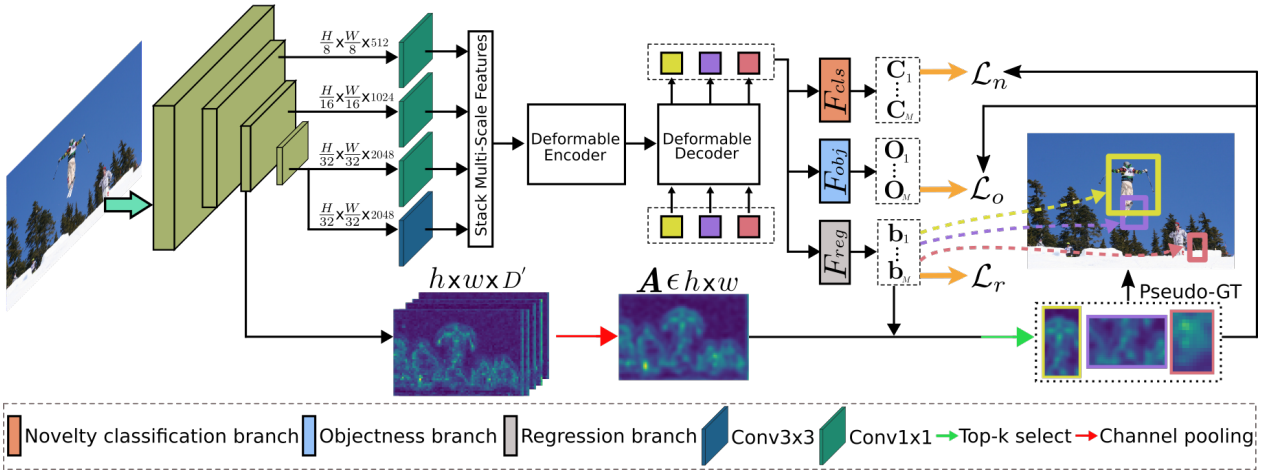

Figure 2. **Proposed OW-DETR framework.** Our approach adapts the standard Deformable DETR for the OWOD problem formulation by introducing (i) an attention driven pseudo-labeling scheme to select the candidate unknown queries, (ii) a novelty classification branch F_{cls} to distinguish the pseudo unknowns from each of the known classes and (iii) an objectness branch F_{obj} that learns to separate foreground objects (known + pseudo unknowns) from the background. In our OW-DETR, D-dimensional multi-scale features for an image I are extracted from the backbone and input to the deformable encoder-decoder along with a set of M learnable object queries $\boldsymbol{q} \in \mathbb{R}^D$ to the decoder. At the decoder output, each object query embedding $\boldsymbol{q}_e \in \mathbb{R}^D$ is input to three different branches: box regression, novelty classification and objectness. The box co-ordinates are output by the regression branch F_{reg}. The objectness branch outputs the confidence of a query being a foreground object, whereas the novelty classification branch classifies the query into one of the known and unknown classes. Our OW-DETR is jointly learned end-to-end with novelty classification loss $\mathcal{L}_n$, objectness loss $\mathcal{L}_o$ and box regression loss $\mathcal{L}_r$.

tector, which updates itself with new knowledge at every episode without forgetting the previously learned classes.

2.1. Overall Architecture

Fig. 2 shows the overall architecture of the proposed open-world detection transformer, OW-DETR. The proposed OW-DETR adapts the standard Deformable DETR (DDETR) [37] for the problem of open-world object detection (OWOD) by introducing (i) an attention-driven pseudo-labeling mechanism (Sec. 2.3) for selecting likely unknown query candidates; (ii) a novelty classification branch (Sec. 2.4) for learning to classify the object queries into one of the many known classes or the unknown class; and (iii) an 'objectness' branch (Sec. 2.5) for learning to separate the foreground objects (ground-truth known and pseudo-labeled unknown instances) from the background. In the proposed OW-DETR, an image I of spatial size $H \times W$ with a set of object instances $\boldsymbol{Y}$ is input to a feature extraction backbone. D-dimensional multi-scale features are obtained at different resolutions and input to a transformer encoder-decoder containing multi-scale deformable attention modules. The decoder transforms a set of M learnable object queries, aided by interleaved cross-attention and self-attention modules, to a set of M object query embeddings $\boldsymbol{q}_e \in \mathbb{R}^D$ that encode potential object instances in the image.

The $\boldsymbol{q}_e$ are then input to three branches: bounding box regression, novelty classification and objectness. While the novelty classification (F_{cls}) and objectness (F_{obj}) branches are single layer feed-forward networks (FFN), the regression branch F_{reg} is a 3-layer FFN. A bipartite matching loss, based on the class and box co-ordinate predictions, is employed to select unique queries that best match the ground-truth (GT) known instances. The remaining object queries are then utilized to select the candidate unknown class instances, which are crucial for learning in the OWOD setting. To this end, an attention map $\boldsymbol{A}$ obtained from the latent feature maps of the backbone is utilized to compute an objectness score s_o for a query $\boldsymbol{q}_e$. The score s_o is based on the activation magnitude inside the query's region-of-interest in $\boldsymbol{A}$. The queries with high scores s_o are selected as candidate instances and pseudo-labeled as 'unknown'. These pseudo-labeled unknown queries along with the collective GT known queries are employed as foreground objects to train the objectness branch. Moreover, while regression branch predicts the bounding box, the novelty classification branch classifies a query into one of the many known classes and an unknown class. The proposed OW-DETR framework is trained end-to-end using dedicated loss terms for novelty classification ($\mathcal{L}_n$), objectness scoring ($\mathcal{L}_o$), in addition to bounding box regression ($\mathcal{L}_r$) in a joint formulation. Next, we present our OW-DETR approach in detail.

2.2. Multi-scale Context Encoding

As discussed earlier in Sec. 1, given the diverse nature of unknown objects that can possibly occur in an image, detecting objects of different sizes while encoding their rich context is one of the major challenges in open-world object detection (OWOD). Encoding such rich context re-

quires capturing long-term dependencies from large receptive fields at multiple scales of the image. Moreover, having lesser inductive biases in the framework that make fewer assumptions about unknown objects, occurring during testing, is likely to be beneficial for improving their detection.

Motivated by the above observations about OWOD task requirements, we adapt the recently introduced single-stage Deformable DETR [37] (DDETR), which is end-to-end trainable and has shown promising performance in standard object detection due to its ability to encode long-term multi-scale context with fewer inductive biases. DDETR introduces multi-scale deformable attention modules in the transformer encoder and decoder layers of DETR [3] for encoding multi-scale context with better convergence and lower complexity. The multi-scale deformable attention module, based on deformable convolution [5, 36], only attends to a small fixed number of key sampling points around a reference point. This sampling is performed across multi-scale feature maps and enables encoding richer context over a larger receptive field. For more details, we refer to [3, 37]. Despite achieving promising performance for the object detection task, the standard DDETR is not suited for detecting unknown class instances in the OWOD setting. To enable detecting novel objects, we introduce an attention-driven pseudo-labeling scheme along with novelty classification and objectness branches, as explained next.

2.3. Attention-driven Pseudo-labeling

For learning to detect unknown objects without any corresponding annotations in the train-set, an OWOD framework must rely on selecting potential unknown instances occurring in the training images and utilizing them as pseudo-unknowns during training. The OWOD approach of ORE [15] selects proposals having high objectness scores and not overlapping with the ground-truth (GT) known instances as pseudo-unknowns. These proposals obtained from a two-stage detector RPN are likely to be biased to the known classes since it is trained with strong supervision from known classes. Distinct from such a strategy, we introduce a bottom-up attention-driven pseudo-labeling scheme that is better generalizable and applicable in a single-stage object detector. Let $\boldsymbol{f}$ denote intermediate $D^{'}$-dimensional feature maps extracted from the backbone, with a spatial size $h \times w$. The magnitude of the feature activations gives an indication of presence of an object in that spatial position, and thereby can be used to compute the confidence of objectness within a window. Let $\boldsymbol{b} = [x_b, y_b, w_b, h_b]$ denote a box proposal with center (x_b, y_b), width w_b and height h_b. The objectness score $s_o(\boldsymbol{b})$ is then computed as,

$$s_o(\boldsymbol{b}) = \frac{1}{h_b \cdot w_b} \sum_{x_b - \frac{w_b}{2}}^{x_b + \frac{w_b}{2}} \sum_{y_b - \frac{h_b}{2}}^{y_b + \frac{h_b}{2}} \boldsymbol{A}, \qquad (1)$$

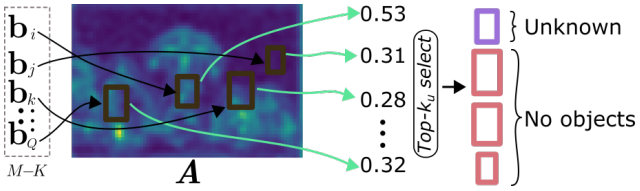

Figure 3. **An example illustration showing our attention-driven pseudo-labeling**. An objectness score for each of the $M-K$ object queries $\boldsymbol{q}_e$ is computed as the mean confidence score in a region-of-interest, corresponding to its box proposal $\boldsymbol{b}_i$, in the attention feature map $\boldsymbol{A}$. A *top-*k_u selection is performed on these $M-K$ scores for obtaining k_u pseudo-unknowns.

where $\boldsymbol{A} \in \mathbb{R}^{h \times w}$ is the feature map $\boldsymbol{f}$ averaged over the channels $D^{'}$. The object proposals in our framework are obtained as the bounding boxes $\boldsymbol{b}$ predicted by the regression branch for the M object query embeddings $\boldsymbol{q}_e$ output by the deformable transformer decoder. For an image with K known object instances, the objectness score s_o is computed for the $M-K$ object queries not selected by the bipartite matching loss[1] of DDETR as best query matches to the GT known instances. The *top-*k_u queries among $M-K$ with the high objectness scores s_o are then pseudo-labeled as unknown objects with bounding boxes given by their corresponding regression branch predictions (see Fig. 3).

2.4. Novelty Classification

The ORE [15] approach introduces an energy-based unknown identifier for classifying a proposal between known and unknown classes. However, it relies on a held-out validation set with weak unknown supervision to learn the energy distributions for the known and unknown classes. In contrast, our OW-DETR does not require any unknown object supervision and relies entirely on the pseudo-unknowns selected using attention-driven pseudo-labeling described in Sec. 2.3. Furthermore, the classification branch F_{cls} in the standard DDETR classifies an object query embedding $\boldsymbol{q}_e$ into one of the known classes or background, *i.e.*, $F_{cls} : \mathbb{R}^D \rightarrow \mathbb{R}^C$. However, when an unknown object is encountered, it fails to classify it into a novel class. To overcome these issues and enable our OW-DETR framework to be trained with only the selected pseudo-unknown objects, we introduce a class label for novel objects in the classification branch. Query embeddings $\boldsymbol{q}_e$ selected as pseudo-unknowns are then trained with the pseudo-label (set to 0 for ease) associated with the novel class in the novelty classification branch $F_{cls} : \mathbb{R}^D \rightarrow \mathbb{R}^{C+1}$. Such an introduction of the novelty class label in classification branch enables $\boldsymbol{q}_e$ to be classified as unknown objects in OW-DETR, which otherwise would have been learned as background, as in the standard object detection task. This helps our model to discriminate potential unknown objects from the background.

[1] Bipartite matching selects one unique object query per GT instance.

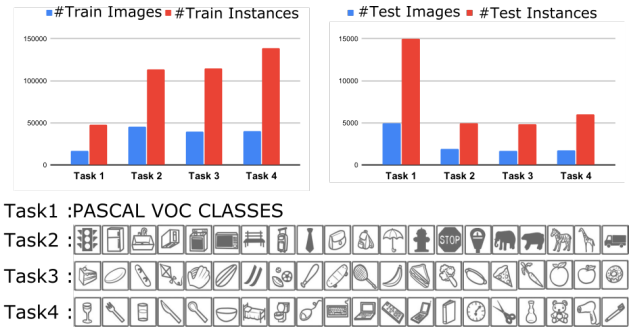

Figure 4. **Task composition in the OWOD evaluation protocol**. The MS-COCO classes in each task along with the number of images and instances (objects) across splits are shown.

2.5. Foreground Objectness

As discussed above, the novelty classification branch F_{cls} is class-specific and classifies a query embedding $\boldsymbol{q}_e$ into one of the $C+1$ classes: C known classes or 1 unknown class or background. While this enables the learning of class-specific separability between known and unknown classes, it does not permit a transfer of knowledge from the known to the unknown objects, which is crucial in understanding as to what constitutes an unknown object in the OWOD setting. Furthermore, the attention-driven pseudo-labeling is likely to be less accurate due to absence of unknown class supervision resulting in most of the query embeddings to be predicted on the background. To alleviate these issues, we introduce a foreground objectness branch $F_{obj} : \mathbb{R}^D \rightarrow [0,1]$ that scores the 'objectness' [18, 30] of the query embeddings $\boldsymbol{q}_e$ in order to better separate the foreground objects (known *and* unknown) from the background. Learning to score the queries corresponding to foreground objects higher than the background enables improved detection of unknown objects which otherwise would have been detected as background. Such a class-agnostic scoring also aids the model to transfer knowledge from the known classes to the unknowns w.r.t. the characteristics that constitute a foreground object.

2.6. Training and Inference

Training: Our OW-DETR framework is trained end-to-end using the following joint loss formulation,

$$\mathcal{L} = \mathcal{L}_n + \mathcal{L}_r + \alpha \mathcal{L}_o, \tag{2}$$

where $\mathcal{L}_n$, $\mathcal{L}_r$ and $\mathcal{L}_o$ denote the loss terms for novelty classification, bounding box regression and objectness scoring, respectively. While the standard focal loss [19] is employed for formulating $\mathcal{L}_n$ and $\mathcal{L}_o$, the term $\mathcal{L}_r$ is the standard ℓ_1 regression loss. Here, α denotes the weight factor for the objectness scoring. When a set of new categories are introduced for the incremental learning stage at each episode in OWOD, motivated by the findings in [15, 28, 34], we employ an exemplar replay based finetuning to alleviate catastrophic forgetting of previously learned classes. Specifically, the model is finetuned after the incremental step in each episode using a balanced set of exemplars stored for each known class.

Inference: M object query embeddings $\boldsymbol{q}_e$ are computed for a test image I and their corresponding bounding box and class predictions are obtained, as in [37]. Let C^t be the number of known classes at time t in addition to the unknown class, *i.e.*, $C^t = |\mathcal{K}^t| + 1$. A *top-k* selection is employed on $M \cdot C^t$ class scores and these selected detections with high scores are used during the OWOD evaluation.

3. Experiments

Datasets: We evaluate our OW-DETR on MS-COCO [20] for OWOD problem. Classes are grouped into set of non-overlapping tasks $\{T_1, \cdots, T_t, \cdots\}$ s.t. classes in a task T_λ are not introduced till $t = \lambda$ is reached. While learning for task T_t, all the classes encountered in $\{T_\lambda : \lambda \leq t\}$ are considered as *known*. Similarly, classes in $\{T_\lambda : \lambda > t\}$ are considered as *unknown*. As in [15], the 80 classes of MS-COCO are split into 4 tasks (see Fig. 4). The training set for each task is selected from the MS-COCO and Pascal VOC [9] train-set images, while Pascal VOC test split and MS-COCO val-set are used for evaluation.

Evaluation Metrics: For known classes, the standard mean average precision (mAP) is used. Furthermore, we use recall as the main metric for unknown object detection instead of the commonly used mAP. This is because all possible unknown object instances in the dataset are not annotated. Recall has been used in [1, 21] under similar conditions.

Implementation Details: The transformer architecture is similar to DDETR in [37]. Multi-scale feature maps are extracted from a ResNet-50 [14], pretrained on ImageNet [6] in a self-supervised manner [4]. Such a pretraining mitigates a possible open-world setting violation, which could occur in fully-supervised pretraining (with class labels) due to possible overlap with the novel classes. The number of queries $M = 100$, while $D = 256$. The k_u for selecting pseudo-labels is set to 5. Moreover, *top*-50 high scoring detections per image are used for evaluation during inference. The OW-DETR framework is trained using ADAM optimizer [17] for 50 epochs, as in [37]. The weight α is set to 0.1. Additional details are provided in the supplementary.

3.1. State-of-the-art Comparison

Tab. 1 shows a comparison of our OW-DETR with the recently introduced ORE [15] on MS-COCO for the OWOD problem. We also report the performance of Faster R-CNN [31] and the standard Deformable DETR (DDETR) [37] frameworks. The comparison is shown in terms of the known class mAP and unknown class recall (U-Recall). U-Recall quantifies a model's ability to retrieve unknown object instances in the OWOD setting. Note that

Table 1. **State-of-the-art comparison for OWOD on MS-COCO.** The comparison is shown in terms of known class mAP and unknown class recall (U-Recall). The unknown recall (U-Recall) metric quantifies a model's ability to retrieve the unknown object instances. The standard object detectors (Faster R-CNN and DDETR) in the top part of table achieve promising mAP for known classes but *are inherently not suited for the OWOD setting since they cannot detect any unknown object*. For a fair comparison in the OWOD setting, we compare with the recently introduced ORE [15] not employing EBUI. Our OW-DETR achieves improved U-Recall over ORE across tasks, indicating our model's ability to better detect the unknown instances. Furthermore, our OW-DETR also achieves significant gains in mAP for the known classes across the four tasks. Note that since all 80 classes are known in Task 4, U-Recall is not computed. See Sec. 3.1 for more details.

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
Faster-RCNN [31]	-	56.4	-	3.7	26.7	15.2	-	2.5	15.2	6.7	0.8	14.5	4.2
Faster-RCNN + Finetuning	Not applicable in Task 1		-	51.0	25.0	38.0	-	38.2	13.6	30.0	29.7	13.0	25.6
DDETR [37]	-	60.3	-	4.5	31.3	17.9	-	3.3	22.5	8.5	2.5	16.4	6.0
DDETR + Finetuning	Not applicable in Task 1		-	54.5	34.4	44.8	-	40.0	17.8	33.3	32.5	20.0	29.4
ORE − EBUI [15]	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
Ours: OW-DETR	**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**

Table 2. **State-of-the-art comparison for incremental object detection (iOD) on PASCAL VOC.** We experiment on 3 different settings. The comparison is shown in terms of per-class AP and overall mAP. The 10, 5 and 1 class(es) in gray background are introduced to a detector trained on the remaining 10, 15 and 19 classes, respectively. Our OW-DETR achieves favorable performance in comparison to existing approaches on all the three settings. See Sec. 3.2 for additional details.

10 + 10 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ILOD [32]	69.9	70.4	69.4	54.3	48	68.7	78.9	68.4	45.5	58.1	59.7	72.7	73.5	73.2	66.3	29.5	63.4	61.6	69.3	62.2	63.2
Faster ILOD [27]	72.8	75.7	71.2	60.5	61.7	70.4	83.3	76.6	53.1	72.3	36.7	70.9	66.8	67.6	66.1	24.7	63.1	48.1	57.1	43.6	62.1
ORE − (CC + EBUI) [15]	53.3	69.2	62.4	51.8	52.9	73.6	83.7	71.7	42.8	66.8	46.8	59.9	65.5	66.1	68.6	29.8	55.1	51.6	65.3	51.5	59.4
ORE − EBUI [15]	63.5	70.9	58.9	42.9	34.1	76.2	80.7	76.3	34.1	66.1	56.1	70.4	80.2	72.3	81.8	42.7	71.6	68.1	77	67.7	64.5
Ours: OW-DETR	61.8	69.1	67.8	45.8	47.3	78.3	78.4	78.6	36.2	71.5	57.5	75.3	76.2	77.4	79.5	40.1	66.8	66.3	75.6	64.1	**65.7**
15 + 5 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ILOD [32]	70.5	79.2	68.8	59.1	53.2	75.4	79.4	78.8	46.6	59.4	59	75.8	71.8	78.6	69.6	33.7	61.5	63.1	71.7	62.2	65.8
Faster ILOD [27]	66.5	78.1	71.8	54.6	61.4	68.4	82.6	82.7	52.1	74.3	63.1	78.6	80.5	78.4	80.4	36.7	61.7	59.3	67.9	59.1	67.9
ORE − (CC + EBUI) [15]	65.1	74.6	57.9	39.5	36.7	75.1	80	73.3	37.1	69.8	48.8	69	77.5	72.8	76.5	34.4	62.6	56.5	80.3	65.7	62.6
ORE − EBUI [15]	75.4	81	67.1	51.9	55.7	77.2	85.6	81.7	46.1	76.2	55.4	76.7	86.2	78.5	82.1	32.8	63.6	54.7	77.7	64.6	68.5
Ours: OW-DETR	77.1	76.5	69.2	51.3	61.3	79.8	84.2	81.0	49.7	79.6	58.1	79.0	83.1	67.8	85.4	33.2	65.1	62.0	73.9	65.0	**69.4**
19 + 1 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ILOD [32]	69.4	79.3	69.5	57.4	45.4	78.4	79.1	80.5	45.7	76.3	64.8	77.2	80.8	77.5	70.1	42.3	67.5	64.4	76.7	62.7	68.2
Faster ILOD [27]	64.2	74.7	73.2	55.5	53.7	70.8	82.9	82.6	51.6	79.7	58.7	78.8	81.8	75.3	77.4	43.1	73.8	61.7	69.8	61.1	68.5
ORE − (CC + EBUI) [15]	60.7	78.6	61.8	45	43.2	75.1	82.5	75.5	42.4	75.1	56.7	72.9	80.8	75.4	77.7	37.8	72.3	64.5	70.7	49.9	64.9
ORE − EBUI [15]	67.3	76.8	60	48.4	58.8	81.1	86.5	75.8	41.5	79.6	54.6	72.8	85.9	81.7	82.4	44.8	75.8	68.2	75.7	60.1	68.8
Ours: OW-DETR	70.5	77.2	73.8	54.0	55.6	79.0	80.8	80.6	43.2	80.4	53.5	77.5	89.5	82.0	74.7	43.3	71.9	66.6	79.4	62.0	**70.2**

all 80 classes are known in Task 4 and thereby U-Recall cannot be computed due to the absence of unknown test annotations. Since both Faster R-CNN and DDETR can only classify objects into known classes but not the unknown, they are *not suited for OWOD setting* and U-Recall cannot be computed for them. For a fair comparison in the OWOD setting, we report ORE without its energy-based unknown identifier (EBUI) that relies on held-out validation data with weak unknown object supervision. The resulting ORE−EBUI framework achieves U-Recall of 4.9, 2.9 and 3.9 on Task 1, 2 and 3, respectively. Our OW-DETR improves the retrieval of unknown objects, leading to improved performance with significant gains for U-Recall, achieving 7.5, 6.2 and 5.7 on the same tasks 1, 2 and 3, respectively. Furthermore, OW-DETR outperforms the best existing OWOD approach of ORE in terms of the known class mAP on all the four tasks, achieving significant absolute gains up to 3.6%. While we use the same split as [15] here for fairness, our OW-DETR also achieves identical gains on a stricter data split (included in supplementary) obtained by removing any possible information leakage. The consistent improvement of OW-DETR over ORE, vanilla Faster R-CNN and DDETR emphasizes the importance of proposed contributions towards a more accurate OWOD.

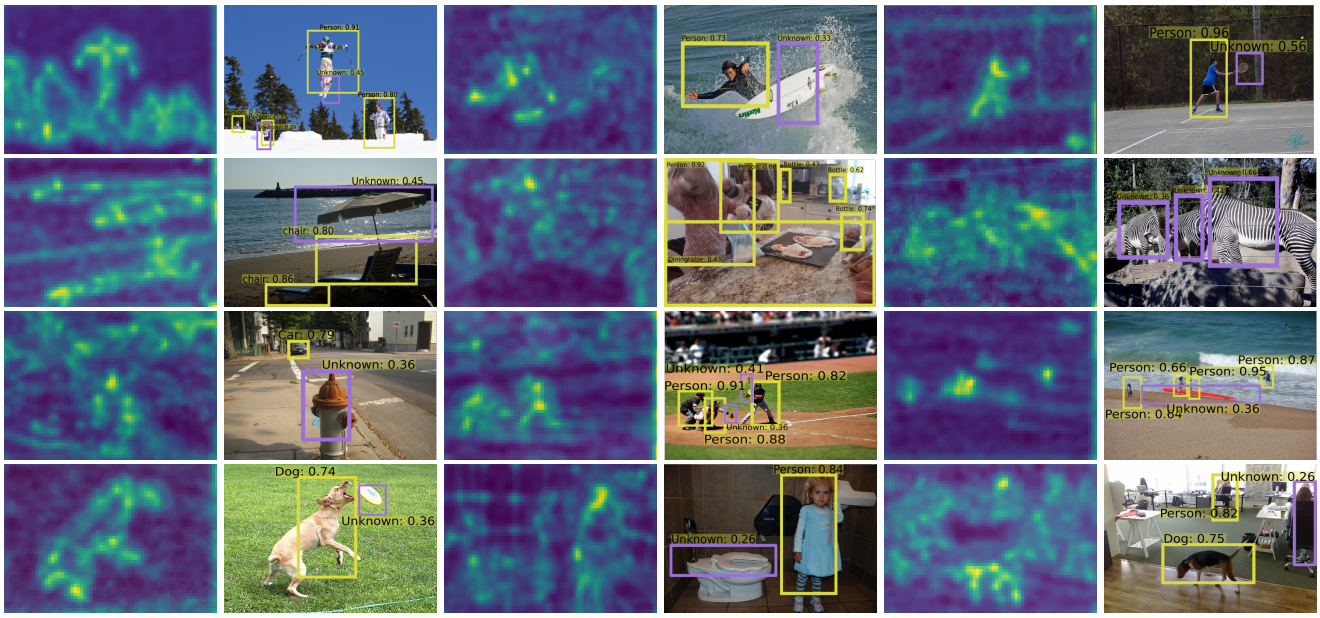

Figure 5. **Qualitative results on example images from MS-COCO test set.** For each example image, its corresponding attention map $\boldsymbol{A}$ computed from the intermediate feature maps is shown on its left. The detections obtained from our OW-DETR are overlayed on the known (yellow) and unknown (purple) class objects. We observe that the attention map activations tend to be higher for regions with foreground objects, illustrating the benefits of attention-driven pseudo-labeling for the unknown objects. The unknown objects like *racket* (row 1, right), *umbrella* (row 2, left), *fire hydrant* (row 3, left) are detected reasonably well. Due to the challenging open-world setting, a few unknown objects are missed, *e.g.*, *sink* (row 2, middle), *table* (row 3, right). Nevertheless, these results indicate the promising performance achieved by our OW-DETR framework in the challenging OWOD setting.

3.2. Incremental Object Detection

As an intuitive consequence of detecting unknown instances, our OW-DETR performs favorably on the incremental object detection (iOD) task. This is due to the decrease in confusion of an unknown object being classified as known class, which enables the detector to incrementally learn the various newer class instances as true foreground objects. Tab. 2 shows a comparison of OW-DETR with existing approaches on PASCAL VOC 2007. As in [27, 32], evaluation is performed on three standard settings, where a group of classes (10, 5 and last class) are introduced incrementally to a detector trained on the remaining classes (10, 15 and 19). Our OW-DETR performs favorably against existing approaches on all three settings, illustrating the benefits of modeling the unknown object class.

3.3. Ablation Study

Tab. 3 shows the impact of progressively integrating our contributions into the baseline framework for the OWOD problem. The comparison is shown in terms of mAP for the known (current and previous) classes and recall for the unknown class, denoted as U-Recall. All the variants shown (except `Baseline`†) include a finetuning step to alleviate the catastrophic forgetting during incremental learning stage. Here, our baseline is the standard Deformable DETR. We also show the upper bound performance of an oracle, *i.e.*, the baseline trained with ground-truth annotations of the unknown class. The `Baseline` achieves higher performance on the known classes but cannot detect *any* unknown object, since it is trained with only known classes and is thereby not suited for OWOD. Integrating the novelty classification branch (denoted by `Baseline+NC`) and employing the pseudo-unknowns selected by our attention-driven pseudo-labeling mechanism for training the novelty classifier enables the detection of unknown instances. Consequently, such an integration achieves unknown recall rates of 6.0, 4.6 and 4.6 for tasks 1, 2 and 3. Our final framework, OW-DETR, obtained by additionally integrating the objectness branch further improves the retrieval of unknown objects in the OWOD setting, achieving U-Recall of 7.5, 6.2 and 5.7 for the same tasks 1, 2 and 3. These results show the effectiveness of our proposed contributions in the OWOD setting for learning a separation between knowns and unknowns through the novelty classification branch and learning to transfer knowledge from known classes to the unknown through the objectness branch.

Open-set Detection Comparison: A detector's ability to handle unknown instances in open-set data can be measured by the degree of decrease in its mAP value, compared to its mAP on closed set data. We follow the same evaluation protocol of [23] and report the performance in Tab. 4. By ef-

Table 3. **Impact of progressively integrating our contributions into the baseline.** The comparison is shown in terms of known class average precision (mAP) and unknown class recall (U-Recall) on MS-COCO for OWOD setting. Apart from the standard baseline (denoted with †), all other models shown include a finetuning step to mitigate catastrophic forgetting. We also show the performance of the oracle (baseline trained with ground-truth unknown class annotations). *Although* `Baseline` *achieves higher mAP for known classes, it is inherently not suited for the OWOD setting since it cannot detect any unknown object*. Integrating the proposed pseudo-labeling based novelty classification (`NC`) with `Baseline` enables unknown class detection. Additionally integrating our objectness branch into the framework further improves the retrieval of unknown objects. Note that since all 80 classes are known in Task 4, U-Recall is not computed.

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
Oracle	31.6	62.5	40.5	55.8	38.1	46.9	42.6	42.4	29.3	33.9	35.6	23.1	32.5
`Baseline`†	-	60.3	-	4.5	31.3	17.8	-	3.3	22.5	8.5	2.5	16.4	6.0
`Baseline`	Not applicable in Task 1		-	54.5	34.4	44.7	-	40.0	17.7	33.3	32.5	20.0	29.4
`Baseline` + `NC`	5.9	58.1	4.6	52.5	32.7	42.6	4.6	36.4	13.4	28.9	30.8	16.3	27.2
Final: **OW-DETR**	**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**

Table 4. **Performance comparison on open-set object detection task**. Our OW-DETR generalizes better by effectively modeling the unknowns and decreasing their confusion with known classes.

Evaluated on →	Pascal VOC 2007	Open-Set (WR1)
Standard Faster R-CNN	81.8	77.1
Standard RetinaNet	79.2	73.8
Dropout Sampling [23]	78.1	71.1
ORE [15]	81.3	78.2
Ours: OW-DETR	**82.1**	**78.6**

fectively modeling the unknowns, our OW-DETR achieves promising performance in comparison to existing methods.

Qualitative Analysis: Fig. 5 shows qualitative results on example images from the MS-COCO test set, along with their corresponding attention maps $\boldsymbol{A}$. The detections for a known class (in yellow) and unknown class (in purple) obtained from our OW-DETR are also overlayed. We observe that unknown objects are detected reasonably well, *e.g.*, *skiis* in top-left image, *tennis racket* in top-right image, *frisbee* in bottom-left image. Although few novel objects are missed (*table* in bottom-right image), these results show that our OW-DETR achieves promising performance in detecting unknown objects in the challenging OWOD setting. Additional results are provided in the supplementary.

4. Relation to Prior Art

Several works have investigated the problem of standard object detection [2, 11, 13, 19, 25, 26, 29, 31]. These approaches work under a strong assumption that the label space of object categories to be encountered during a model's life-cycle is the same as during its training. The advent of transformers for natural language processing [33, 35] has inspired studies to investigate related ideas for vision tasks [8, 10, 16, 24], including standard object detection [3, 37]. Different to standard object detection, incremental object detection approaches [27, 32] model newer object classes that are introduced in training incrementally and tackle the issue of catastrophic forgetting. On the other hand, the works of [7, 12, 22, 23] focus on open-set detection, where new unknown objects encountered during test are to be rejected. In contrast, the recent work of [15] tackles the challenging open-world object detection (OWOD) problem for detecting both known and unknown objects in addition to incrementally learning new object classes. Here, we propose an OWOD approach, OW-DETR, in a transformer-based framework [37], comprising the following novel components: attention-driven pseudo-labeling, novelty classification and objectness scoring. Our OW-DETR explicitly encodes multi-scale contextual information with fewer inductive biases while simultaneously enabling transfer of objectness knowledge from known classes to the novel class for improved unknown detection.

5. Conclusions

We proposed a novel transformer-based approach, OW-DETR, for the problem of open-world object detection. The proposed OW-DETR comprises dedicated components to address open-world settings, including attention-driven pseudo-labeling, novelty classification and objectness scoring in order to accurately detect unknown objects in images. We conduct extensive experiments on two popular benchmarks: PASCAL VOC and MS COCO. Our OW-DETR consistently outperforms the recently introduced ORE for all task settings on the MS COCO dataset. Furthermore, OW-DETR achieves state-of-the-art performance in case of incremental object detection on PASCAL VOC dataset.

Acknowledgements

This work was partially supported by VR starting grant (2016-05543).

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018. 5

[2] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *ECCV*, 2020. 8

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 4, 8

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 4

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[7] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, 2020. 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 8

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 5

[10] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019. 8

[11] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 8

[12] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *WACV*, 2020. 8

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 8

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[15] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021. 1, 2, 4, 5, 6, 8

[16] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021. 8

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Ming Lu, and Yurong Chen. Ron: Reverse connection with objectness prior networks for object detection. In *CVPR*, 2017. 5

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5, 8

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[21] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 5

[22] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *ICRA*, 2019. 8

[23] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, 2018. 7, 8

[24] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *NeurIPS*, 2021. 8

[25] Jing Nie, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Enriched feature guided refinement network for object detection. In *ICCV*, 2019. 8

[26] Yanwei Pang, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Efficient featurized image pyramid network for single shot detector. In *CVPR*, 2019. 8

[27] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *PRL*, 2020. 6, 7, 8

[28] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *ECCV*, 2020. 5

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 8

[30] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 5

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 5, 6, 8

[32] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *ICCV*, 2017. 6, 7, 8

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 8

[34] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 5

[35] Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *ICLR*, 2019. 8

[36] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 4

[37] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2, 3, 4, 5, 6, 8