

# Class-agnostic Object Detection

Ayush Jaiswal, Yue Wu, Pradeep Natarajan, Premkumar Natarajan  
 Amazon Alexa Natural Understanding  
 Manhattan Beach, CA, USA  
 {ayujaisw, wuayue, natarap, premknat}@amazon.com

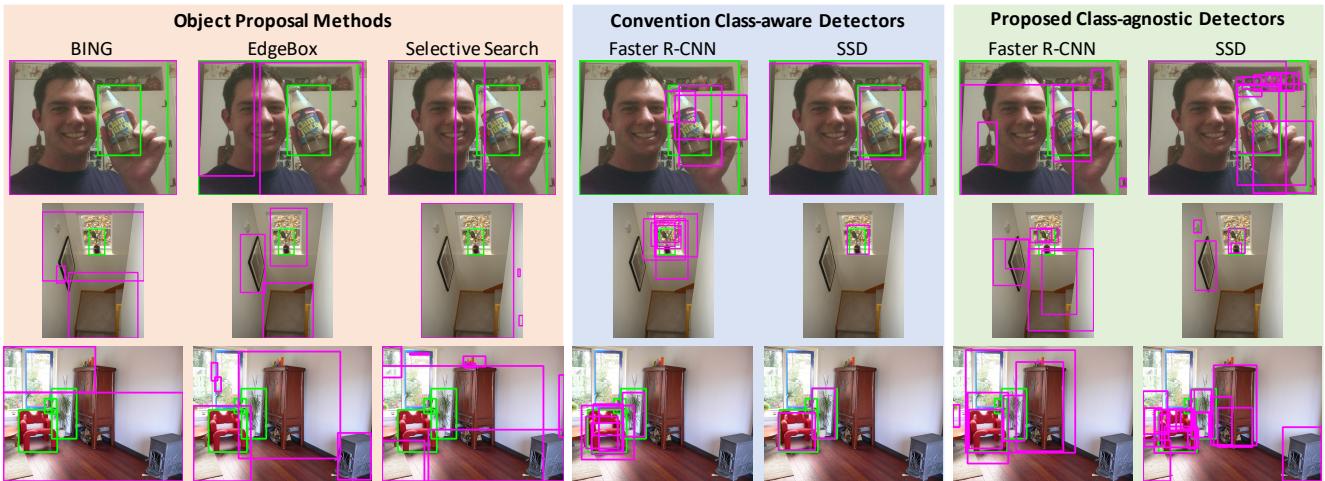


Figure 1: Class-agnostic object detection using object proposal methods (OPMs), conventional class-aware detectors, and the proposed adversarially trained class-agnostic detectors. **Green** and **magenta** boxes indicate **ground-truth** and **detection outputs**, respectively. OPMs generate regions of interest that could contain objects without precisely locating any object. Class-aware detectors detect objects from a set of *known* object-types. The proposed class-agnostic detection aims to localize all objects irrespective of their types including those of unknown classes. Results show that the proposed adversarially trained class-agnostic models detect novel objects for which no annotations are available. Best viewed digitally and zoomed in.

## Abstract

*Object detection models perform well at localizing and classifying objects that they are shown during training. However, due to the difficulty and cost associated with creating and annotating detection datasets, trained models detect a limited number of object types with unknown objects treated as background content. This hinders the adoption of conventional detectors in real-world applications like large-scale object matching, visual grounding, visual relation prediction, obstacle detection (where it is more important to determine the presence and location of objects than to find specific types), etc. We propose class-agnostic object detection as a new problem that focuses on detecting objects irrespective of their object-classes. Specifically, the goal is to predict bounding boxes for all objects in an image but not their object-classes. The predicted boxes can then be consumed by another system to perform application-specific classification, retrieval, etc. We propose training and eval-*

*uation protocols for benchmarking class-agnostic detectors to advance future research in this domain. Finally, we propose (1) baseline methods and (2) a new adversarial learning framework for class-agnostic detection that forces the model to exclude class-specific information from features used for predictions. Experimental results show that adversarial learning improves class-agnostic detection efficacy.*

## 1. Introduction

Human visual scene understanding relies on the ability to detect and recognize objects in one's field of view. This naturally carries over to machine scene understanding through computer vision techniques. Hence, the field of object detection has garnered tremendous research interest [54] and large advances have been made in improving detection systems, especially since the adoption of deep learning.

The object detection task is formulated as the prediction of bounding boxes and classes for objects in a given image.

This requires densely labeled data that contains annotations for all objects in training images. However, creating such datasets is extremely challenging and expensive. Therefore, conventional object detection focuses instead on the reduced task of locating and recognizing “objects of interest” corresponding to limited types of objects that are labeled in the training data, with objects of unknown types treated as background content. Notably, detection of unknown object-types is explicitly penalized in the widely adopted mean average precision metric used for benchmarking. This hinders the adoption of trained detectors in real-world applications due to the added cost of retraining them for application-specific object types. Furthermore, such detectors cannot be used in applications like obstacle detection, where it is more important to determine the location of all objects present in the scene than to find specific kinds of objects.

In order to address the aforementioned limitations of conventional class-aware detection, we propose class-agnostic object detection as a new research task with the goal of predicting bounding boxes for all objects present in an image irrespective of their object-types. Intuitively, this task additionally seeks to detect objects of unknown types, which are not present or annotated in the training data. This challenging yet high-reward goal of generalization to unseen object types would benefit downstream applications (e.g., application-specific object classification, object retrieval from large databases, etc.) that can consume such class-agnostic detections.

Besides the problem formulation, we propose training and evaluation protocols for class-agnostic detection with generalization and downstream utility as primary goals. Generalization is evaluated on PASCAL VOC [9], MS COCO [28], and Open Images [24] in the form of recall of unseen classes. Specifically, models trained on VOC are evaluated on unseen classes of COCO and those trained on COCO are tested on non-overlapping classes of Open Images. Furthermore, the VOC dataset is split into seen and unseen classes in order to measure recall of seen and unseen classes within the same dataset. Utility of trained detectors is evaluated as the accuracy of pretrained ImageNet [40] classifiers on ObjectNet [2] images cropped using the bounding boxes predicted by the detectors.

We present a few baseline methods for class-agnostic object detection – region proposals of two-stage detectors, pretrained class-aware models used *as-is* or finetuned for binary object versus not classification, and detection models trained from scratch for the said binary classification. Finally, we propose a new adversarial training framework that forces the detection model to exclude class-specific information from the features used for making predictions. Experimental results show that our adversarial framework improves class-agnostic detection efficacy for both Faster R-CNN [39] (two-stage) and SSD [29] (one-stage).

The major contributions of this paper are:

- a novel class-agnostic object detection problem formulation as a new research direction,
- training and evaluation protocols for benchmarking and advancing research,
- a new adversarial learning framework for class-agnostic detection that penalizes the model if object-type is encoded in embeddings used for predictions.

## 2. Related Work

Zou *et al.* [54] provide a comprehensive survey of advances in object detection in the past 20 years. Some conventional detection works [20, 25, 36, 46, 50] describe the task as “generic” object detection but they use it to signify their focus on common objects of *known* types, which is distinct from our work on detecting objects of *all* types.

Two-stage object detectors like Faster R-CNN [39] employ a region proposal module that identifies a set of regions in a given image that could contain objects, which are then used by a downstream detector module to localize and identify objects. A number of works [5, 7, 12, 23, 33, 35, 43, 52] have been proposed recently that aim to improve the quality of such *object proposals* and reduce their number in order to speed up the final detection task. A majority of proposal methods are trained end-to-end with the detector module, which biases them to objects of known types. Methods also exist for generating [6, 32, 44, 53] or filtering [30, 31, 38] object proposals based on edge-related features and objectness metrics. However, as defined, these proposals do not correspond directly to final detections and must be fed to a detector module to infer final detections. In contrast, the proposed task of class-agnostic detection is aimed at predicting final bounding boxes for all objects in an image. Two-stage methods for class-agnostic detection could employ these works to generate intermediate object proposals.

Some works [14, 49] focus on generating pixel-level objectness scores in order to segment object regions out of background content. These works produce objectness heatmaps which could sometimes be used to detect individual objects if they do not overlap and are strikingly distinct.

A few works [8, 13] have shown the efficacy of convolutional neural networks like AlexNet [21] for localizing objects irrespective of their classes. In this work, we take this idea further and formally define the task of class-agnostic object detection along with training and evaluation protocols for benchmarking current and future research.

Kuen *et al.* [22] show how weights from object recognition networks can be transferred to detection models to scale-up the number of object-types for detection. This line of research is also related to few-shot [10, 19, 34, 45, 47, 48, 51] and zero-shot [1, 26, 37] object detection, which are targeted towards detecting objects of novel types given a few reference samples or descriptions, respectively.

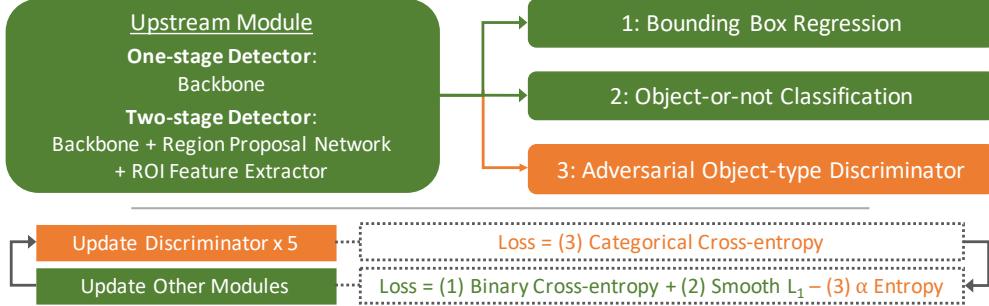


Figure 2: Adversarial framework for class-agnostic object detection for both one- and two-stage models. The upstream module with the backbone (e.g., VGG16, ResNet-50, etc.), Region Proposal Network (two-stage) and ROI feature extractor (two-stage) remain unchanged, along with the box regressor. The conventional object-type classification is replaced with a binary object-or-not classifier, and a new adversarial object-type discriminator is attached during training. The bottom part of the figure shows the training procedure – iterating between five discriminator updates for each update to the other modules.

### 3. Class-agnostic Object Detection

In this section, we describe the proposed task of class-agnostic object detection and contrast it with conventional class-aware object detection. We then discuss a few intuitive baseline models for the proposed task.

#### 3.1. From Class-aware to Class-agnostic Detection

The conventional formulation of object detection can be described as the task of locating and classifying objects in a given image. The goal of models developed to solve this task is to predict bounding box coordinates and class labels for objects present in the image. Given this formulation, a plethora of models have been proposed that train on images with object annotations. However, due to the difficulty and cost associated with collecting and labeling large datasets, annotations are typically collected for a limited number of object categories. This has inadvertently reduced the original formulation to the task of detecting “objects of interest”, corresponding to object types that are annotated in training datasets. Although reduced in form, the conventional formulation remains a challenging problem, with direct application in use-cases with a fixed set of known object types.

In this work, we propose a sibling task of class-agnostic object detection that aims to detect all objects in a given image irrespective of their object-types. More specifically, the goal of this task is to predict bounding boxes for all objects present in an image but not their category. Furthermore, given that most available training datasets do not contain dense annotations for all kinds of objects, an additional implicit goal for models developed for this task is to generalize to objects of unknown types, i.e., those for which annotations are not available in the training dataset. This is in direct contrast to conventional class-aware object detection, which treats unknown objects as background content. As compared to conventional detectors, class-agnostic models can be more easily adopted in complex real-life

applications like object retrieval from a large and diverse database, recognition of application-specific object-types through training a downstream object recognition model (instead of retraining the full detection network), etc.

#### 3.2. Baseline Models

We identify three intuitive and straightforward baseline models for solving the class-agnostic detection task, selected due to their ease of implementation and natural curiosity about their performance. The baseline models are:

- region proposal network of a two-stage detector,
- class-aware detector trained for object-type classification and bounding box regression,
- pretrained class-aware model finetuned end-to-end for object-or-not binary classification instead of object-type, along with bounding box regression,
- detection model trained from scratch for object-or-not binary classification and bounding box regression.

## 4. Adversarial Learning Framework for Class-agnostic Object Detection

### 4.1. General Framework

Conventional class-aware detection focuses on detecting “objects of interest”, which inherently requires models to be able to distinguish between types of objects in a closed known set. Intuitively, models achieve this by encoding features that are discriminative of object-types. However, for class-agnostic detection and for models to be able to detect previously unseen types of objects, detectors should encode features that more effectively distinguish objects from background content and individual objects from other objects in the image, without discriminating between object-types.

Naïvely training conventional object detectors for the binary classification task of object-or-not along with bounding box regression is not sufficient to ensure that models fo-

cus on class-agnostic features and more importantly, ignore type-distinguishing features so that they can better generalize to unseen object-types. In order to overcome this problem, we propose to train class-agnostic object detectors in an adversarial fashion such that models are penalized for encoding features that contain object-type information.

We begin with observing a popular two-part pattern in the model design of both one-stage and two-stage conventional object detectors. The first *upstream* part of a detection model learns a set of convolutional features from entire images (one-stage) or regions of interest (two-stage). The second *downstream* part consumes these features and passes them through classification and regression branches for object-type and bounding box prediction, respectively. This two-part setup allows for external control on the information output by the first part and consumed by the second.

We propose to augment class-agnostic detectors with adversarial discriminator branches that attempt to classify object-types (annotated in the training data) from the features output by the upstream part of detection networks, and penalize the model training if they are successful [15–18]. The models are trained in an alternating manner such that the discriminators are frozen when the rest of the model is updated and vice versa. While updating the discriminators, we use the standard categorical cross-entropy loss with object-types used as prediction targets. On the other hand, while training the rest of the model, we minimize (a) the cross-entropy loss for object-or-not classification, (b) smooth L<sub>1</sub> loss for bounding box regression, and (c) the negative entropy of discriminator predictions. This entropy maximization forces the upstream part of detection models to exclude object-type information from the features it outputs. The discriminator is updated five times for every update to the rest of the model and the negative entropy is weighted with a multiplier  $\alpha$  (tuned on {0.1, 1}) in the overall objective. Figure 2 summarizes the complete framework.

During test-time inference, the discriminators are detached from the model, giving back the original network with the standard layers and parameter-count. Thus, our framework does not cause performance delays.

## 4.2. Model Instantiation

We demonstrate the applicability of the proposed adversarial framework to both Faster R-CNN (FRCNN), a two-stage detector, and SSD, a one-stage detector. We use the publicly available MMDetection [4] framework for implementing the models and running experiments. We train two versions of the adversarial models – one is trained from scratch and the other is finetuned from a pretrained baseline class-aware model. Further details are as follows.

**Faster R-CNN.** The FRCNN model first generates regions of interest in the input image, which is followed by extracting features from and making object-type and bounding

box predictions for each region. We create a class-agnostic adversarial version of FRCNN by (1) replacing the multi-class object-type classification layer with a binary object-or-not layer, and (2) attaching an adversarial discriminator on top of the feature extraction layer that provides inputs to the classification and regression heads. Thus, during training, this feature layer serves three prediction heads instead of the standard two. In our experiments, we use the standard FRCNN model available in MMDetection, which includes a ResNet-50 backbone and a Feature Pyramid Network [27].

**SSD.** An SSD model utilizes features from several layers of its backbone network to detect objects of different scales, corresponding to the depth-levels of the backbone layers. Specifically, SSD models contain classification and regression layers for making predictions at each depth-level. In order to create a class-agnostic adversarial version of the SSD model, we (1) replace each object-type classification layer with a binary object-or-not layer, and (2) attach an adversarial discriminator at each depth-level where predictions are made. Thus, during training, each prediction level in the resulting model has three prediction heads instead of the conventional two. We use the standard SSD-300 with VGG-16 pretrained on ImageNet as the backbone.

## 5. Training and Evaluation Protocols

We propose two kinds of experiments for evaluating class-agnostic detection, geared towards measuring (1) generalization to unseen object types, and (2) the downstream utility of trained models. While there are several ways to design experiments for (1) and (2), we propose first steps with potential for refinement in future work.

### 5.1. Generalization to Unseen Object-types.

Class-agnostic detectors should, by definition, not be limited to object types seen during training. Hence, it is important to evaluate their efficacy at identifying unseen types of objects. We measure this performance as average recall (AR; also known as detection rate [53]) at various number ( $k \in \{3, 5, 10, 20, 30, 100, 300, 1000\}$ ) of allowed bounding box predictions, i.e., AR@ $k$  (IoU  $\geq 0.5$ ). The goal here is to achieve high AR@ $k$  at all  $k$  levels. We design two sets of generalization experiments based on this setup.

**Experiment I.** We split the VOC dataset into 17 seen classes and three unseen classes. Seen classes in the VOC 07+12 training set are used for learning the models while both seen and unseen classes in the VOC 07 validation set are used for evaluation as AR@ $k$  of seen and unseen classes. The harmonic mean of seen and unseen AR@ $k$ , for each  $k$ , is reported as a measure of overall performance.

In order to select the three unseen classes, we compute the confusion matrix<sup>1</sup> of a 300-layer SSD model trained on

<sup>1</sup>[https://github.com/kaanakan/object\\_detection\\_confusion\\_matrix](https://github.com/kaanakan/object_detection_confusion_matrix)

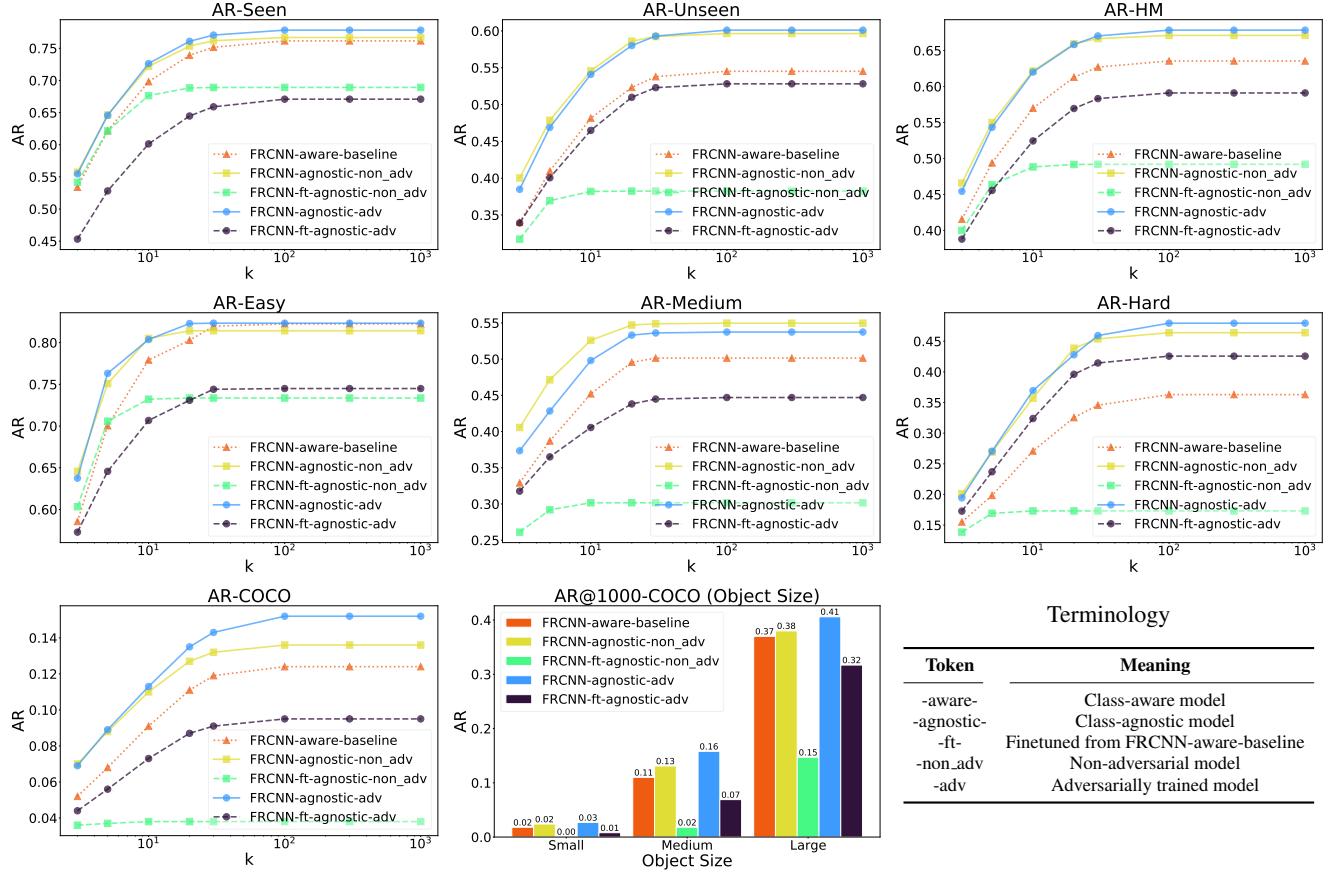


Figure 3: Generalization results for FRCNN models trained on the seen VOC dataset. The top row shows macro-level AR@ $k$  for seen and unseen classes in VOC and their harmonic mean (AR-HM). FRCNN-agnostic-adv performs the best overall. The second row shows micro-level results for the easy, medium, and hard unseen classes. FRCNN-agnostic-adv performs the best on the hard and easy classes with recall drop for the medium class. The last row provides results of evaluation on the COCO data of 60 unseen classes. FRCNN-agnostic-adv achieves the best AR@ $k$  for objects of all sizes.

the standard 20-class VOC 07+12 dataset, and use  $F_1$  scores to determine easy, medium, and hard classes. Specifically, the class with the highest  $F_1$  is the hardest for generalization as it is the most distinguishable from the other classes, making it the most difficult to generalize to if it were not seen during training. Similarly, the class with the lowest  $F_1$  is picked as the easy class. The medium class is the one with the median  $F_1$ . Thus, the unseen set comprises *cow* (easy), *boat* (medium) and *tvmonitor* (hard) classes, while the rest are considered seen. In addition to the macro AR@ $k$  scores, micro AR@ $k$  are reported for each unseen class. Besides the evaluation on VOC, the (VOC) models are tested for generalization on the 60 non-VOC classes in the COCO 2017 validation set using the same AR@ $k$  metric. AR@1000 is further reported for small, medium, and large sized objects belonging to the 60 classes.

**Experiment II.** In this experiment, models are trained on the COCO 2017 training set and evaluated on non-overlapping classes in the Open Images V6 test set. We

first normalize the class names in both COCO and Open Images to find exact matches between classes in the two datasets. We identify a few classes that have slightly different names in the two datasets and normalize them manually. We then build a semantic tree using the publicly available Open Images class hierarchy and exclude all classes that either match exactly with a COCO class or have a COCO class as a predecessor or successor. The remaining classes are used for testing. AR@ $k$  is used as the evaluation metric.

## 5.2. Downstream Utility

A major motivation for research in the domain of class-agnostic object detection is the potential for widespread downstream utilization of class agnostic bounding boxes for extracting objects in images and using them for various visual reasoning tasks. In this work, we propose evaluation of downstream utility in terms of object recognition on images in the ObjectNet [2] dataset. Detection models are used to predict  $M$  bounding boxes for each image, which are then

### Terminology

Token	Meaning
-aware	Class-aware model
-agnostic	Class-agnostic model
-ft	Finetuned from FRCNN-aware-baseline
-non_adv	Non-adversarial model
-adv	Adversarially trained model

Table 1: Generalization results for experiments I and II. AR@1000-Unseen shows results for unseen VOC classes, AR@1000-COCO for unseen COCO classes, and AR@1000-OI for unseen Open Images classes. “Ovr”, “Med”, “Sml”, and “Lrg” stand for overall, medium, small, and large, respectively. “-aw-” and “-ag-” in the model name indicate whether the model is class-aware or -agnostic, “-ft-” tells whether the model was finetuned from the class-aware baseline, and “ad” represents models trained adversarially. FRCNN-aw-prop refers to the evaluation of the FRCNN-aw proposals from its first stage Region Proposal Network. Class-agnostic models generalize better than class-aware models with those trained adversarially from scratch performing the best overall.

Model	I: Training on seen VOC								II: Training on COCO			
	AR@1000-Unseen				AR@1000-COCO				AR@1000-OI			
	Ovr	Easy	Med	Hard	Ovr	Sml	Med	Lrg	Ovr	Sml	Med	Lrg
FRCNN-aw-prop	0.374	0.388	0.383	0.355	0.107	0.006	0.071	0.375	0.153	0.005	0.037	0.329
FRCNN-aw	0.545	0.822	0.501	0.363	0.124	0.018	0.110	0.370	0.170	0.016	0.083	0.330
FRCNN-ft-ag	0.382	0.733	0.302	0.173	0.038	0.001	0.018	0.147	0.184	<b>0.028</b>	<b>0.138</b>	0.313
FRCNN-ag	0.596	0.814	<b>0.550</b>	0.464	0.136	0.024	0.131	0.380	0.182	<b>0.028</b>	0.125	0.318
FRCNN-ft-ag-ad	0.528	0.745	0.447	0.426	0.095	0.008	0.069	0.317	0.181	0.026	0.112	0.328
FRCNN-ag-ad	<b>0.601</b>	<b>0.823</b>	0.537	<b>0.479</b>	<b>0.152</b>	<b>0.027</b>	<b>0.158</b>	<b>0.406</b>	<b>0.192</b>	<b>0.028</b>	0.131	<b>0.337</b>
SSD-aw	0.663	0.831	0.594	0.588	0.156	0.019	0.194	0.391	0.199	0.013	0.145	<b>0.350</b>
SSD-ft-ag	0.671	0.819	0.668	0.555	0.177	<b>0.040</b>	0.238	0.366	0.188	0.020	0.168	0.301
SSD-ag	0.702	0.810	0.669	0.645	0.181	0.031	0.239	0.404	0.206	0.023	0.206	0.314
SSD-ft-ag-ad	0.683	0.809	0.651	0.612	0.155	0.028	0.197	0.354	0.194	0.018	0.181	0.308
SSD-ag-ad	<b>0.722</b>	<b>0.833</b>	<b>0.689</b>	<b>0.660</b>	<b>0.196</b>	0.036	<b>0.252</b>	<b>0.443</b>	<b>0.210</b>	<b>0.024</b>	<b>0.211</b>	0.319

use to crop the image into  $M$  versions. Pretrained<sup>2</sup> ImageNet classifiers are then used to predict the object class from the cropped images. Here we use ResNet-152 [11], MobileNet-v2 [41], and Inception-v3 [42] models.

Two metrics are used for evaluation – Accuracy@ $M$  for  $M \in \{1, 5, 10\}$ , and Best-overlap (BO) accuracy. For Accuracy@ $M$ , the classifier’s prediction on at least one of the  $M$  crops needs to be correct for the image to be considered as successfully classified. BO-accuracy is calculated using bounding boxes that have the highest intersection over union (IoU) with the ground-truth [3] bounding boxes.

## 6. Results

### 6.1. Generalization to Unseen Object-types

**Experiment I.** We train the baseline models and the adversarial models on the VOC seen set and evaluate them on both the VOC set and the COCO unseen classes. Results of this experiment with FRCNN as the base model are presented in Figure 3. Class-agnostic models outperform standard class-aware models in overall performance, especially for unseen VOC classes. Breakdown of recall for unseen classes reveals that the margin in recall becomes larger as the difficulty of generalizability is increased, with the largest performance gains on the hard unseen class. Results on the COCO unseen classes show that the adversarial model performs the best overall and for objects of all sizes.

<sup>2</sup><https://pytorch.org/docs/stable/torchvision/models.html>

Figure 4 summarizes the results with SSD as the base model. The results show that the class-agnostic SSD models largely outperform the conventional class-aware SSD model. Furthermore, the proposed adversarial model performs the best overall with some reduction in average recall for the VOC seen set. Breakdown of recall for the easy, medium, and hard unseen classes shows that the adversarial model performs the best across the board. Finally, results on COCO show consistent improvements from adversarial learning, especially for medium and large sized objects.

Hence, class-agnostic models perform better than class-aware models in generalizing to objects of unseen types and the proposed adversarial model performs the best overall.

**Experiment II.** In this experiment, the baseline and the adversarial models are trained on the COCO dataset and evaluated on the non-overlapping classes of the Open Images, as outlined in Section 5.1. ?? summarizes the results for both FRCNN and SSD models. In both cases, the class-agnostic models generalize better than the class-aware models overall, with the proposed class-agnostic models trained from scratch achieving the best AR scores.

Table 1 presents AR@1000 generalization scores for experiments I and II, showing that the adversarially trained (from scratch) models perform the best in both settings.

We find that the recall of models that are finetuned from the pretrained class-aware baselines are worse than those that are trained from scratch. We attribute this to the dif-

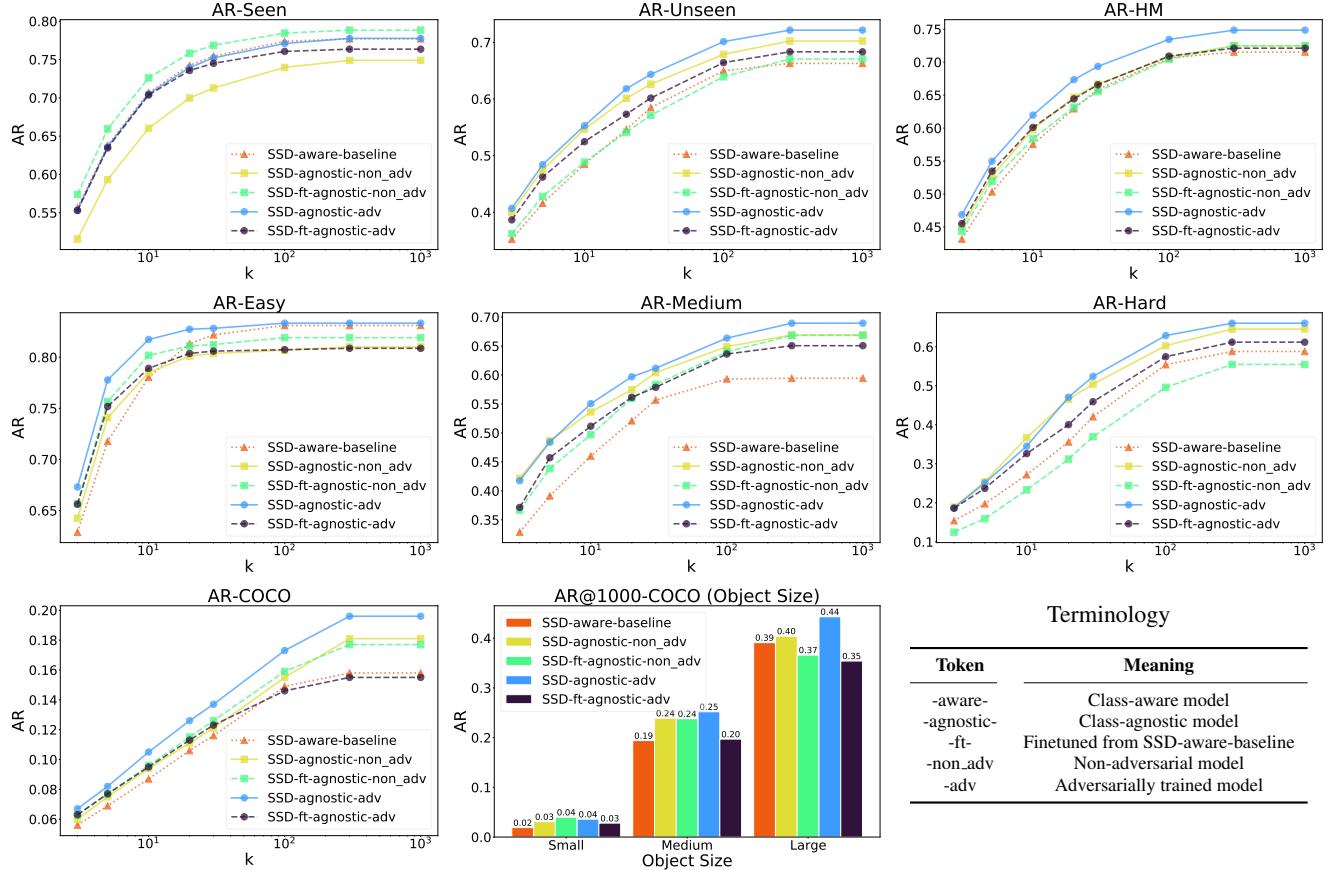


Figure 4: Generalization results for SSD models trained on the seen VOC dataset. The top row shows macro-level AR@ $k$  for seen and unseen classes in VOC as well as their harmonic mean (AR-HM). SSD-agnostic-adv performs the best on AR-Unseen and AR-HM, with a drop in AR-Seen, but the models that outperform SSD-agnostic-adv on AR-Seen do significantly worse on AR-Unseen and AR-HM. The second row shows micro-level results for the easy, medium, and hard unseen classes. SSD-agnostic-adv performs the best in all categories. The last row provides results of evaluation on the COCO data of 60 unseen classes. SSD-agnostic-adv achieves the best AR@ $k$  with a slight reduction for small-sized objects.

ficulty of unlearning discriminative features for object-type classification and realigning to learn type-agnostic features, which prevents the finetuned models from achieving the same performance as those that are trained from scratch.

## 6.2. Downstream Object Recognition

We evaluate the object recognition performance for images cropped using bounding boxes predicted by the class-aware and class-agnostic models trained for the generalization experiments in Section 5.1. Table 2 summarizes the results of these experiments. Results show that the adversarially trained class-agnostic models perform better than the baselines in general on both Accuracy@ $M$  and Best-overlap accuracy (as described in Section 5.2), irrespective of the ImageNet classifier used. These results correlate with the generalization results presented in Section 6.1, indicating that detection models that generalize better to unseen object-types have higher downstream utility.

## 7. Conclusion and Future Work

Conventional class-aware objection detection models perform well at detecting “objects of interest” that belong to known objects types seen during training. However, these models do not generalize well to unseen object-types, limiting their incorporation in real-world applications. In this work, we have formulated a novel task of class-agnostic object detection, where the goal is to detect objects of known *and* unknown types irrespective of their category. Furthermore, we have presented training and evaluation protocols for benchmarking models and advancing future research in this direction. These include two sets of experiments – (1) one for generalization to new object-types and (2) another for downstream utility in terms of object recognition. Finally, we have presented a few intuitive baselines and proposed a new adversarially trained model that penalizes the objective if the learned representations encode object-type.

### Terminology

Token	Meaning
-aware-	Class-aware model
-agnostic-	Class-agnostic model
-ft-	Finetuned from SSD-aware-baseline
-non_adv	Non-adversarial model
-adv	Adversarially trained model

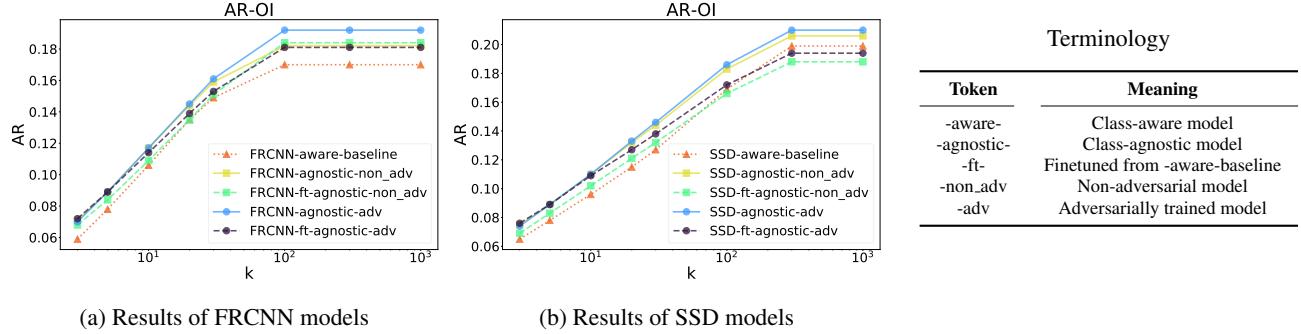


Figure 5: Generalization results of (a) FRCNN and (b) SSD models trained on the COCO dataset and evaluated on the non-overlapping classes in the Open Images dataset. Results show that class-agnostic models generalize better than the class-aware models, with the SSD-agnostic-adv and FRCNN-agnostic-adv models achieving the best recall.

Table 2: Downstream ObjectNet recognition results. Images are cropped using boxes predicted by the detectors and pretrained ImageNet models are used for classification from the cropped images. Averages of accuracies across classifiers are also reported. BO-acc is the accuracy when the predicted box with the highest IoU with the ground-truth box is used. Accuracies are also presented for top  $M$  proposals as Acc@ $M$ . Results for uncropped images and those cropped using ground-truth boxes are provided for reference. “-aw-” and “-ag-” in the model name indicate whether the model is class-aware or -agnostic; “-ad” represents models trained adversarially. Results of other baseline detectors are provided in the supplementary material. Adversarially trained class-agnostic models achieve the best results.

Classifier	Detector	Detectors trained on VOC				Detectors trained on COCO			
		BO-acc	Acc@1	Acc@5	Acc@10	BO-acc	Acc@1	Acc@5	Acc@10
ResNet-152	FRCNN-aw	0.235	0.108	0.255	0.318	0.111	0.011	0.014	0.014
	FRCNN-ag-ad	<b>0.293</b>	<b>0.132</b>	<b>0.297</b>	<b>0.366</b>	<b>0.304</b>	<b>0.131</b>	<b>0.288</b>	<b>0.352</b>
MobileNet-v2	FRCNN-aw	0.153	0.057	<b>0.177</b>	0.226	0.063	0.009	0.011	0.011
	FRCNN-ag-ad	<b>0.198</b>	<b>0.070</b>	<b>0.177</b>	<b>0.248</b>	<b>0.202</b>	<b>0.085</b>	<b>0.198</b>	<b>0.253</b>
Inception-v3	FRCNN-aw	0.194	0.072	0.223	0.288	0.094	0.011	0.013	0.014
	FRCNN-ag-ad	<b>0.249</b>	<b>0.086</b>	<b>0.224</b>	<b>0.310</b>	<b>0.258</b>	<b>0.106</b>	<b>0.255</b>	<b>0.320</b>
Avg. for	FRCNN-aw	0.194	0.079	0.218	0.277	0.089	0.010	0.013	0.013
	FRCNN-ag-ad	<b>0.247</b>	<b>0.096</b>	<b>0.233</b>	<b>0.308</b>	<b>0.255</b>	<b>0.107</b>	<b>0.247</b>	<b>0.308</b>
ResNet-152	SSD-aw	0.291	0.188	0.304	0.362	0.314	0.181	0.333	0.388
	SSD-ag-ad	<b>0.312</b>	<b>0.199</b>	<b>0.339</b>	<b>0.406</b>	<b>0.317</b>	<b>0.218</b>	<b>0.370</b>	<b>0.422</b>
MobileNet-v2	SSD-aw	0.190	0.114	0.202	0.254	0.210	0.116	0.232	0.280
	SSD-ag-ad	<b>0.211</b>	<b>0.121</b>	<b>0.233</b>	<b>0.292</b>	<b>0.213</b>	<b>0.143</b>	<b>0.258</b>	<b>0.307</b>
Inception-v3	SSD-aw	0.247	0.156	0.270	0.332	<b>0.270</b>	0.148	0.301	0.361
	SSD-ag-ad	<b>0.267</b>	<b>0.165</b>	<b>0.308</b>	<b>0.383</b>	<b>0.270</b>	<b>0.182</b>	<b>0.331</b>	<b>0.390</b>
Avg. for	SSD-aw	0.243	0.153	0.259	0.316	0.265	0.148	0.289	0.343
	SSD-ag-ad	<b>0.263</b>	<b>0.162</b>	<b>0.293</b>	<b>0.360</b>	<b>0.267</b>	<b>0.181</b>	<b>0.320</b>	<b>0.373</b>
Accuracy on full uncropped images									0.167
Accuracy on images cropped with ground-truth boxes (empirical upper bound of BO-Acc and Acc@1)									0.378

Results show that the proposed adversarial class-agnostic model outperforms the baselines on both generalization and downstream-utility experiments, for both one-stage (SSD) and two-stage (Faster R-CNN) detectors.

This work serves to establish a new direction of research

in object detection. As such, the possibilities for future research are endless, including but not limited to — (1) further refinement of the problem formulation, (2) improved as well as additional training and evaluation protocols, and (3) novel methods for class-agnostic object detection.

Terminology	
Token	Meaning
-aware-	Class-aware model
-agnostic-	Class-agnostic model
-ft-	Finetuned from -aware-baseline
-non_adv	Non-adversarial model
-adv	Adversarially trained model

## References

- [1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.
- [2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pages 9453–9463, 2019.
- [3] Ali Borji. Objectnet dataset: Reanalysis and correction. *arXiv preprint arXiv:2004.02042*, 2020.
- [4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tian-heng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [5] Shuhan Chen, Jindong Li, Xuelong Hu, and Ping Zhou. Saliency detection for improving object proposals. In *2016 IEEE International Conference on Digital Signal Processing (DSP)*, pages 462–466. IEEE, 2016.
- [6] Ming-Ming Cheng, Yun Liu, Wen-Yan Lin, Ziming Zhang, Paul L Rosin, and Philip HS Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. *Computational Visual Media*, 5(1):3–20, 2019.
- [7] Ian Endres and Derek Hoiem. Category-independent object proposals with diverse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):222–234, 2013.
- [8] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.
- [9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
- [10] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [12] Shengfeng He and Rynson WH Lau. Oriented object proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 280–288, 2015.
- [13] Brody Huval, Adam Coates, and Andrew Ng. Deep learning for class-generic object detection. *arXiv preprint arXiv:1312.6885*, 2013.
- [14] K J Joseph, Rajiv Chunilal Patel, Amit Srivastava, Uma Gupta, and Vineeth N Balasubramanian. Mason: A model agnostic objectness framework. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.
- [15] Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, Wael AbdAlmageed, and Premkumar Natarajan. Invariant representations through adversarial forgetting. In *AAAI*, pages 4272–4279, 2020.
- [16] Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised adversarial invariance. In *Advances in Neural Information Processing Systems*, pages 5092–5102, 2018.
- [17] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Unified adversarial invariance. *arXiv preprint arXiv:1905.03629*, 2019.
- [18] Ayush Jaiswal, Shuai Xia, Iacopo Masi, and Wael AbdAlmageed. Ropad: Robust presentation attack detection through unsupervised adversarial invariance. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.
- [19] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8420–8429, 2019.
- [20] Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Ming Lu, and Yurong Chen. Ron: Reverse connection with objectness prior networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5936–5944, 2017.
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [22] Jason Kuen, Federico Perazzi, Zhe Lin, Jianming Zhang, and Yap-Peng Tan. Scaling object detection by transferring classification weights. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6044–6053, 2019.
- [23] Weicheng Kuo, Bharath Hariharan, and Jitendra Malik. Deepbox: Learning objectness with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2479–2487, 2015.
- [24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallochi, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.
- [25] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE, 2004.
- [26] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. Zero-shot object detection with textual descriptions. In *Proceedings of the AAAI Con-*

- ference on Artificial Intelligence}, volume 33, pages 8690–8697, 2019.
- [27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.
- [30] Yun Liu, Shi-Jie Li, and Ming-Ming Cheng. Refinedbox: Refining for fewer and high-quality object proposals. *Neurocomputing*, 2020.
- [31] Cewu Lu, Shu Liu, Jiaya Jia, and Chi-Keung Tang. Contour box: Rejecting object proposals without explicit closed contours. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2021–2029, 2015.
- [32] Yao Lu and Linda G Shapiro. Closing the loop for edge detection and object proposals. In *AAAI*, pages 4204–4210, 2017.
- [33] Jianxiang Ma, Anlong Ming, Zilong Huang, Xinggang Wang, and Yu Zhou. Object-level proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4921–4929, 2017.
- [34] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13846–13855, 2020.
- [35] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):128–140, 2016.
- [36] Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, and Jian Sun. Thundernet: Towards real-time generic object detection on mobile devices. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6718–6727, 2019.
- [37] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Asian Conference on Computer Vision*, pages 547–563. Springer, 2018.
- [38] Esa Rahtu, Juho Kannala, and Matthew Blaschko. Learning a category independent object detection cascade. In *2011 international conference on Computer Vision*, pages 1052–1059. IEEE, 2011.
- [39] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.
- [40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [41] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [43] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.
- [44] H. Wang, Z. Liu, L. Wang, L. Weng, and C. Pan. Objectness estimation using edges. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4136–4140, 2015.
- [45] 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.
- [46] Zhiqian Wang and Jezekiel Ben-Ari. Generic object detection using model based segmentation. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 428–433. IEEE, 1999.
- [47] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. *arXiv preprint arXiv:2007.09384*, 2020.
- [48] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. *arXiv preprint arXiv:2007.12107*, 2020.
- [49] Bo Xiong, Suyog Dutt Jain, and Kristen Grauman. Pixel objectness: learning to segment generic objects automatically in images and videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2677–2692, 2018.
- [50] Hongyu Xu, Xutao Lv, Xiaoyu Wang, Zhou Ren, Navaneeth Bodla, and Rama Chellappa. Deep regionlets: Blended representation and deep learning for generic object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [51] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9577–9586, 2019.
- [52] Ziming Zhang, Yun Liu, Xi Chen, Yanjun Zhu, Ming-Ming Cheng, Venkatesh Saligrama, and Philip HS Torr. Sequential optimization for efficient high-quality object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1209–1223, 2017.
- [53] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.

- [54] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye.  
Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.