AirDet: Few-Shot Detection without Fine-tuning for Autonomous Exploration

Bowen Li[1,2], Chen Wang[1], Pranay Reddy[1,3],
Seungchan Kim[1], and Sebastian Scherer[1]

[1] Robotics Institute, Carnegie Mellon University, USA
chenwang@dr.com, {bowenli2,seungch2,basti}@andrew.cmu.edu
[2] School of Mechanical Engineering, Tongji University, China
[3] Electronics and Communication Engineering, IIITDM Jabalpur, India
2018033@iiitdmj.ac.in

Abstract. Few-shot object detection has attracted increasing attention and rapidly progressed in recent years. However, the requirement of an exhaustive offline fine-tuning stage in existing methods is time-consuming and significantly hinders their usage in online applications such as autonomous exploration of low-power robots. We find that their major limitation is that the little but valuable information from a few support images is not fully exploited. To solve this problem, we propose a brand new architecture, AirDet, and surprisingly find that, by learning *class-agnostic relation* with the support images in all modules, including cross-scale object proposal network, shots aggregation module, and localization network, AirDet without fine-tuning achieves comparable or even better results than many fine-tuned methods, reaching up to **30-40%** improvements. We also present solid results of onboard tests on real-world exploration data from the DARPA Subterranean Challenge, which strongly validate the feasibility of AirDet in robotics. To the best of our knowledge, AirDet is the first feasible few-shot detection method for autonomous exploration of low-power robots. The code and pre-trained models are released at https://github.com/Jaraxxus-Me/AirDet.

Keywords: Few-shot object detection, Online, Robot exploration

1 Introduction

Few-shot object detection (FSOD) [36,39–41,45] aims at detecting objects out of base training set with few support examples per class. It has received increasing attention from the robotics community due to its vital role in autonomous exploration, since robots are often expected to detect novel objects in an unknown environment but only a few examples can be provided online. For example, in a mission of rescue shown in Fig. 1 (a), the robots are required to detect some uncommon objects such as drill, rope, helmet, vent.

Despite its recent promising developments, most of existing methods [2, 7, 16, 19, 24, 31, 36, 37, 39, 42] require a careful *offline* fine-tuning stage on novel
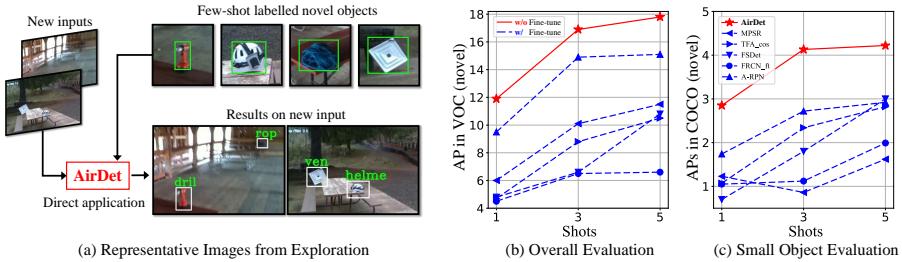
(a) Representative Images from Exploration   (b) Overall Evaluation   (c) Small Object Evaluation

Fig. 1. Representative images from robots' exploration and performance comparison of the state-of-the-arts [6,28,36,39,40] and proposed AirDet. Solid lines denote results with no fine-tuning and dashed lines indicate results fine-tuned on few-shot data. Without further updating, AirDet can outperform prior work. Besides, unlike the fine-tuned models meeting bottleneck in small objects, AirDet has set an outstanding new level.

images before inference. However, the requirement of the fine-tuning process is infeasible for robotic *online* applications, since (**1**) new object categories can be dynamically added during exploration, thus re-fine-tuning the model with limited onboard computational resources for novel classes is extremely inefficient for the time-starved tasks such as search and rescue [8,33–35]. (**2**) to save human effort, only very few samples can be provided online[4], thus the fine-tuning stage [7, 16, 19, 24, 31, 36, 37, 39, 42] needs careful *offline* hyper-parameter tuning to avoid over-fitting, which is infeasible for *online* exploration, and (**3**) fine-tuned models usually perform well for in-domain test [19, 24, 28, 36, 39, 40], while suffer from cross-domain test, which is unfavourable for robotic applications.

Therefore, we often expect a few-shot detector that is able to inference without fine-tuning such as [6]. However, the performance of [6] is still severely hampered for challenging robotics domain due to (**1**) ineffective multi-scale detection; (**2**) ineffective feature aggregation from multi-support images; and (**3**) inaccurate bounding box location prediction. Surprisingly, in this paper, we find that all three problems can be effectively solved by learning *class-agnostic relation* with support images. We name the new architecture AirDet, which can produce promising results even without abominable fine-tuning as shown in Fig. 1, which, to the best of our knowledge, is the first feasible few-shot detection model for autonomous robotic exploration. Specifically, the following three modules are proposed based on *class-agnostic relation*, respectively.

Support-guided Cross-Scale Fusion (SCS) for Object Proposal One reason for performance degradation in multi-scale detection is that the region proposals are not effective for small scale objects, even though some existing works adopt multiple scale features from query images [42,45]. We argue that the proposal network should also include cross-scale information from the support

---

[4] Since online annotation is needed during mission execution, only 1-5 samples can be provided in most of the robotic applications, which is the main focus of this paper.

images. To this end, we present a novel SCS module, which explicitly extracts multi-scale features from cross-scale relations between support and query images.

Global-Local Relation (GLR) for Shots Aggregation Most prior work [6, 39, 41] simply average multi-shot support feature to obtain a class prototype for detection head. However, this cannot fully exploit the little but valuable information from every support image. Instead, we construct a shots aggregation module by learning the relationship between the multi-support examples, which achieves significant improvements with more shots.

Prototype Relation Embedding (PRE) for Location Regression Some existing works [6, 42] introduced a relation network [32] into the classification branch; however, the location regression branch is often neglected. To settle this, we introduce cross-correlation between the proposals from SCS and support features from GLR into the regression branch. This results in the PRE module, which explicitly utilize support images for precise object localization.

In summary, AirDet is a fully relation-based few-shot object detector, which can be applied directly to the novel classes without fine-tuning. It surprisingly produces comparable or even better results than exhaustively fine-tuned SOTA methods [6, 28, 36, 39, 40], as shown in Fig. 1 (b). Besides, as shown in Fig. 1 (c), AirDet maintains high robustness in small objects due to the SCS module, which fully takes advantage of the multi-scale support feature. Note that in this paper, fine-tuning is undesired because it cannot satisfy the online responsive requirement for robots, but it can still improve the performance of AirDet.

2 Related Works

2.1 General Object Detection

The task of object detection [9, 10, 13, 23, 25, 28] is to find out all the pre-defined objects in an image, predicting their categories and locations, which is one of the core problems in the field of computer vision. Object detection algorithms are mainly divided into: two-stage approaches [9, 10, 13, 28] and one-stage approaches [23, 25–27]. R-CNN [10], and its variants [9, 10, 13, 28] serve as the foundation of the former branch; among them, Faster R-CNN [28] used region proposal network (RPN) to generate class-agnostic proposals from the dense anchors, which greatly improved the speed of object detection based on R-CNN [9]. On the other hand, YOLO series [25–27] fall into the second branch, which tackles object detection as an end-to-end regression problem. Besides, the known SSD series [20, 23] propose to utilize pre-defined bounding boxes to adjust to various object scales inspired by [28].

One shortcoming of the above methods is that they require abundant labeled data for training. Moreover, the types and number of object categories are fixed after training (80 classes in COCO, for instance), which is not applicable to robot's autonomous exploration, where unseen, novel objects often appear online.

2.2   Few-shot Object Detection

Trained with abundant data for base classes, few-shot object detectors can learn to generalize only using a few labeled novel image shots. Two main branches leading in FSOD are meta-learning-based approaches [6,12,39–41] and transfer-learning-based approaches [24,31,36,38,45].

Transfer-learning approaches seek for the best learning strategy of general object detectors [28] on a few novel images. Wang *et al.* [36] proposed to fine-tune only the last layer with a cosine similarity-based classifier. Using manually defined positive refinement branch, MPSR [39] mitigated the scale scarcity issue. Recent works have introduced semantic relations between novel-base classes [45] and contrastive proposal encoding [31].

Aiming at training meta-models on episodes of individual tasks, meta-learning approaches [6,12,15,40,41,43] generally contain two branches, one for extracting support information and the other for detection on the query image. Among them, Meta R-CNN [41], and FSDet [40] target at support guided query channel attention. With novel attention RPN and multi-relation classifier, A-RPN [6] has set the current SOTA. Very recent works also cover support-query mutual guidance [43], aggregating context information [15], and constructing heterogeneous graph convolutional networks on proposals [12].

2.3   Relation Network for Few-shot Learning

In few-shot image classification, relation network [32], also known as learning to compare, has been introduced to train a classifier by modeling the class-agnostic relation between a query image and the support images. Once trained and provided with a few novel support images, inference on novel query images can be implemented without further updating.

For few-shot object detection, such relation has only been utilized for the classification branch so far in very few works. For example, Fan *et al.* proposed a multi-relation classification network, which consists of global, local, and patch relation branches [6]. Zhang *et al.* leveraged general relation network [32] architecture to build multi-level proposal scoring, and support weighting modules [43]. In this work, we thoroughly explore such relation in few-shot detection and propose a fully relation-based architecture.

2.4   Multi-Scale Feature Extraction

Multi-scale features have been exhaustively exploited for multi-scale objects in general object detection [17,20,21,23,26,30]. For example, FSSD [20] proposed to fuse multi-scale feature and implement detection on the fused feature map. Lin *et al.* constructed the feature pyramid network (FPN) [21], which builds a top-down architecture and employs multi-scale feature map for detection. For few-shot detection, standard FPN [21] has been widely adopted in prior transfer-learning-based methods [31,36,39,45]. In meta-learning, existing meta-learner [43] employs all scales from FPN and implements detection on each scale in parallel, which is computationally inefficient.
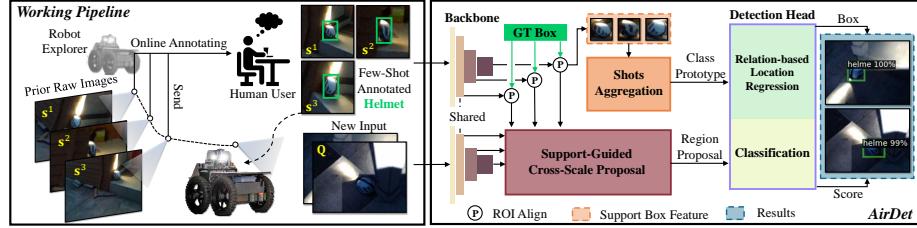
Fig. 2. The pipeline of the autonomous exploration task and the framework of AirDet. During exploration, a few prior raw images that potentially contain novel objects (helmet) are sent to a human user first. Provided with online annotated few-shot data, the robot explorer is able to detect those objects by observing its surrounding environment. AirDet includes 4 modules, *i.e.*, the shared backbone, support-guided cross-scale (SCS) feature fusion module for region proposal, global-local relation (GLR) module for shots aggregation, and relation-based detection head, which are visualized by different colors.

3   Preliminary

In few-shot object detection [4,39–41], the classes are divided into B base classes $\mathcal{C}_\mathrm{b}$ and N novel ones $\mathcal{C}_\mathrm{n}$, satisfying that $\mathcal{C}_\mathrm{b} \cap \mathcal{C}_\mathrm{n} = \varnothing$. The objective is to train a model that can detect novel classes in $\mathcal{C}_\mathrm{n}$ by only providing k-shot labeled samples for $\mathcal{C}_\mathrm{n}$ and abundant images from base classes $\mathcal{C}_\mathrm{b}$.

During training, we adopt the episodic paradigm [41]. Basically, images from the base classes $\mathcal{C}_\mathrm{b}$ are split into query images $\mathbf{Q}_\mathrm{b}$ and support images $\mathbf{S}_\mathrm{b}$. Given all support images $\mathbf{S}_\mathrm{b}$, the model learns to detect objects in query images $\mathbf{Q}_\mathrm{b}$. During test, the model is to detect objects in novel query images $\mathbf{Q}_\mathrm{n}$ by only providing few (1-5) labeled novel support images $\mathbf{S}_\mathrm{n}$.

Remark 1: Most existing methods [2,24,31,36,38–41,45] have to be fine-tuned on $\mathbf{S}_\mathrm{n}$ due to the class-specific model design, while AirDet can be applied directly to $\mathbf{Q}_\mathrm{n}$ by providing $\mathbf{S}_\mathrm{n}$ without fine-tuning.

4   Methodology

Since only a few shots are given during model test, information from the support images is little but valuable. We believe that the major limitation of the existing algorithms is that such information from support images is not fully exploited. Therefore, we propose to learn *class-agnostic relation* with the support images in all the modules of AirDet. As exhibited in Fig. 2, the structure of AirDet is simple: except for the shared backbones, it only consists of three modules, *i.e.*, a support-guided cross-scale fusion (SCS) module for regional proposal, a global-local relation (GLR) module for shots aggregation, and a relation-based detection head, containing prototype relation embedding (PRE) module for location regression and a multi-relation classifier [6]. We next introduce two kinds of *class-agnostic relation*, which will be used by the three modules.

4.1   Class-Agnostic Relation

To exploit the relation between two features from different aspects, we define two relation modules, *i.e.*, spatial relation $\mathcal{R}_{\mathrm{s}}(\cdot, \cdot)$ and channel relation $\mathcal{R}_{\mathrm{c}}(\cdot, \cdot)$.
1. Spatial Relation: Object features from the same category are often correlated along the spatial dimension, thus we define the spatial relation features $\mathcal{R}_{\mathrm{s}}$ in (1) leveraging on the regular and depth-wise convolution.

$$\mathcal{R}_{\mathrm{s}}(\mathbf{A}, \mathbf{B}) = \mathbf{A} \odot \mathrm{MLP}\Big(\mathrm{Flatten}\big(\mathrm{Conv}(\mathbf{B})\big)\Big), \tag{1}$$

where inputs $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{C \times W \times H}$ denote 2 general tensors. Flatten means flatten the features in spatial (image) domain and MLP denotes multilayer perceptron (MLP), so that $\mathrm{MLP}\Big(\mathrm{Flatten}\big(\mathrm{Conv}(\mathbf{B})\big)\Big) \in \mathbb{R}^{C \times 1 \times 1}$. $\odot$ indicates depth-wise convolution [6]. Note that we use convolution to calculate correlation since both operators are composed of inner products.
2. Channel Relation: Inspired by the phenomenon that features of different classes are often stored in different channels [18], we propose a simple but effective channel relation $\mathcal{R}_{\mathrm{c}}(\cdot, \cdot)$ in (2) to extract the cross-class relation features.

$$\mathcal{R}_{\mathrm{c}}(\mathbf{A}, \mathbf{B}) = \mathrm{Conv}\big(\mathrm{Cat}(\mathbf{A}, \mathbf{B})\big) + \mathrm{Cat}\big(\mathrm{Conv}(\mathbf{A}), \mathrm{Conv}(\mathbf{B})\big), \tag{2}$$

where $\mathrm{Cat}(\cdot, \cdot)$ is to concatenate features along the channel dimension.
Remark 2: The two simple but effective *class-agnostic relation* learners are fundamental building blocks of AirDet, which, to the best of our knowledge, is the first attempt towards a fully relation-based structure in few-shot detection.

4.2   Support-guided Cross-Scale Fusion (SCS) for Object Proposal

As mentioned earlier, existing works generate object proposals only using single scale information from query images [16, 37, 39, 40], while such strategy may not be effective for small scale novel objects.

Differently, we propose support-guided cross-scale fusion (SCS) in AirDet to introduce multi-scale features and take the relation between query and support images for region proposal. As shown in Fig. 3 (a), SCS takes support and query features from different backbone blocks (ResNet2, 3, and 4 block) as input. We first apply *spatial relation*, where the query and support features from the same backbone block are $\mathbf{A}, \mathbf{B}$ in (1), respectively. Then we use *channel relation* to fuse the ResNet2 and ResNet3 block features, which are $\mathbf{A}, \mathbf{B}$ in (2), respectively. The fused channel relation feature is later merged with the spatial relation feature from ResNet4 block. The final merged feature is sent to the region proposal network (RPN) [28] to generate region proposals.

4.3   Global-Local Relation (GLR) for Shots Aggregation

In prior attempts [16, 39–41], the support object features from multiple shots are usually averaged to represent the class prototype, which is then used for
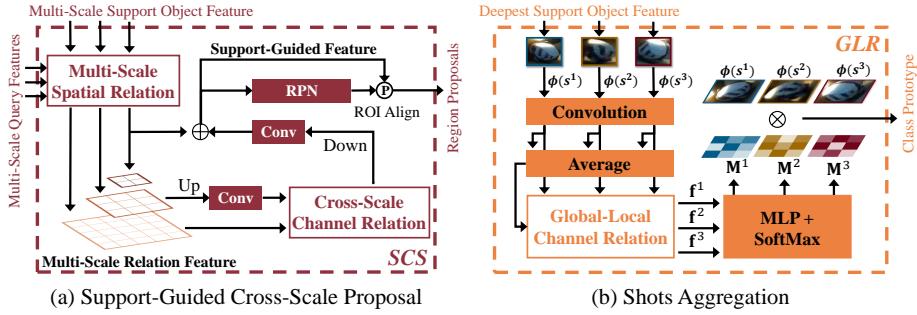
(a) Support-Guided Cross-Scale Proposal   (b) Shots Aggregation

Fig. 3. Network architecture of SCS for region proposal and GLR for shots aggregation.

regression and classification. Although it can be effective with fine-tuning, we argue that a simple average cannot fully utilize the information from few-shot data. To this end, we built global-local relation (GLR) module in Fig. 3 (b), which leverages the features from every shot to construct the final prototype.

Suppose the k-shot deepest support features are $\phi(\mathbf{s}^i)$, $i = 1, \cdots, k$, our final class prototype $\mathbf{e}$ can be expressed as a weighted average of the features:

$$\mathbf{e} = \sum_{i=1}^{k} (\phi(\mathbf{s}^i) \otimes \mathbf{M}^i), \tag{3}$$

where $\otimes$ is the element-wise multiplication, and $\mathbf{M}^i$ is a confidence map:

$$\mathbf{M}^i = \mathrm{SoftMax}\Big(\mathrm{MLP}\big(\mathbf{f}^i\big)\Big), \tag{4}$$

where $\mathbf{f}^i$ is the output from the channel relation extractor:

$$\mathbf{f}^i = \mathcal{R}_c\left(\mathrm{Conv}(\phi(\mathbf{s}^i)), \frac{1}{k}\sum_{i=1}^{k}\mathrm{Conv}(\phi(\mathbf{s}^i))\right). \tag{5}$$

Note that to include both "global" (all shots) and "local" (single shot) features, the inputs of the channel relation extractor in (5) include both the feature from that shot and the averaged features over all shots.

Remark 3: Unlike prior work [16, 39–41] relying on fine-tuning with more support data for performance gain, AirDet can extract stronger prototype to achieve improvement on more shots without fine-tuning.

4.4 Prototype Relation Embedding (PRE) for Location Regression

It has been demonstrated that a multi-relation network [6] is effective for the classification branch. Inspired by its success, we further build a prototype relation embedding (PRE) network for the location regression branch. Given a prototype

exemplar $\mathbf{e} \in \mathbb{R}^{C \times a \times a}$, we utilize the spatial relation (1) to embed information from the exemplar to the proposal features $\mathbf{p}^j$, $j = 1, 2, \cdots, p$ as:

$$\mathbf{l}^j = \mathbf{p}^j + \mathcal{R}_{\mathrm{s}}(\mathbf{p}^j, \mathbf{e}), \tag{6}$$

where we take 3×3 convolution layer in (1) for spatial feature extraction. The proposal features $\mathbf{l}^j$ is then used for bounding box regression through an MLP module following Faster-RCNN [28].

Remark 4: Class-related feature $\mathbf{l}^j$ contains information from support objects, which turns out more effective for location regression even if the objects have never been seen in the training set.

5 Experiments

5.1 Implementation

We adopt the training pipeline from [6]. To maintain a fair comparison with other methods [28,36,39,40], we mainly adopt ResNet101 [14] pre-trained on ImageNet [4] as backbone. The performance of other backbones is presented in Appendix B. For fair comparison [6, 28, 36, 39, 40], we utilized their official implementation, support examples, and models (if provided) in all the experiments. AirDet and the baseline [6] take the **same** supports in all the settings. We use 4 NVIDIA GeForce Titan-X Pascal GPUs for experiments. Detailed configuration of AirDet can be found in Appendix A and the source code.

Remark 5: To save human effort, only very few support examples (1-5 samples per class) can be provided during online exploration. Therefore, we mainly focused on $k = 1, 2, 3, 5$-shot evaluation. Since the objects from exploration are usually unseen, we only test novel classes throughout the experiments.

5.2 In-domain Evaluation

We first present the in-domain evaluation on COCO benchmark [22], where the models are trained and tested on the same dataset. Following prior works [2,6,7, 16,31,36,38–41,45], the 80 classes are split into 60 non-VOC base classes and 20 novel classes. During training, the base class images from COCO trainval2014 are considered available. With few-shot samples per novel class, the models are evaluated on 5,000 images from COCO val2014 dataset.

Overall Performance As shown in Table 1, AirDet achieves significant performance gain on the baseline [6]. AirDet without fine-tuning amazingly also achieves comparable or even better results than many fine-tuned methods. With fine-tuning, AirDet outperformed existing SOTAs [2, 6, 28, 36, 39, 40, 44]. Since the results without fine-tuning may be sensitive to support images, we report the averaged performance, and the standard deviation on 3-5 randomly sampled support images, where we surprisingly find AirDet more robust to the variance of support images compared with the baseline [6].

Multi-scale Objects We next report the performance of methods [6,28,36,39,

Table 1. Performance comparison on COCO validation dataset. In each setting, red and green fonts denote the best and second-best performance, respectively. AirDet achieves significant performance gain on baseline without fine-tuning. With fine-tuning, AirDet sets a new SOTA performance. †We randomly sampled 3-5 different groups of support examples and reported the average performance and their standard deviation.

Shots		1			2			3			5		
Method	Fine-tune	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
A-RPN [6]†	✗	4.32 ±0.7	7.62 ±1.3	4.3 ±0.7	4.67 ±0.3	8.83 ±0.5	4.49 ±0.3	5.28 ±0.6	9.95 ±0.8	5.05 ±0.6	6.08 ±0.3	11.17 ±0.4	5.88 ±0.3
AirDet (Ours)†	✗	5.97 ±0.4	10.52 ±0.9	5.98 ±0.2	6.58 ±0.2	12.02 ±0.4	6.33 ±0.2	7.00 ±0.5	12.95 ±0.8	6.71 ±0.7	7.76 ±0.3	14.28 ±0.4	7.31 ±0.4
FRCN [28]	✓	3.26	6.66	3.04	3.73	7.79	3.22	4.59	9.52	4.07	5.32	11.20	4.54
TFA_{fc} [36]	✓	2.78	5.39	2.36	4.14	7.98	4.01	6.33	12.10	5.94	7.92	15.58	7.29
TFA_{cos} [36]	✓	3.09	5.24	3.21	4.21	7.70	4.35	6.05	11.48	5.93	7.61	14.56	7.17
FSDetView [40]	✓	2.20	6.20	0.90	3.40	10.00	1.50	5.20	14.70	2.10	8.20	21.60	4.70
MPSR [39]	✓	3.34	6.11	3.25	5.41	9.68	5.52	5.70	10.54	5.50	7.20	13.55	6.89
A-RPN [6]	✓	4.59	8.85	4.37	6.15	12.05	5.76	8.24	15.52	7.92	9.02	17.29	8.53
W. Zhang *et al.* [44]	✓	4.40	7.50	4.90	5.60	9.90	5.90	7.20	13.30	7.40	-	-	-
FADI [2]	✓	5.70	10.40	6.00	7.00	13.01	7.00	8.60	15.80	8.30	10.10	18.60	11.90
AirDet (Ours)	✓	6.10	11.40	6.04	8.73	16.24	8.35	9.95	19.39	9.09	10.81	20.75	10.27

Table 2. Performance evaluation on multi-scale objects from COCO. Highest-ranking and second-best scores are marked out with red and green, respectively. Without fine-tuning, AirDet can avoid over-fitting and shows robustness on small-scale objects. By virtue of the SCS module, AirDet can achieve higher results than those with FPN.

Shots			1			2			3			5		
Method	FPN	Fine-tune	AP_s	AP_m	AP_l	AP_s	AP_m	AP_l	AP_s	AP_m	AP_l	AP_s	AP_m	AP_l
A-RPN [6]†	✗	✗	2.43 ±0.4	5.00 ±1.0	6.74 ±1.1	2.67 ±0.3	5.01 ±0.3	7.18 ±0.4	3.42 ±0.2	6.15 ±0.5	8.77 ±0.8	3.54 ±0.3	6.73 ±0.03	9.97 ±0.2
AirDet (Ours)†	✗	✗	2.85 ±0.3	6.33 ±0.7	9.00 ±0.8	4.00 ±0.3	6.84 ±0.1	9.94 ±0.3	4.13 ±0.1	7.95 ±0.5	11.30 ±0.9	4.22 ±0.2	8.24 ±0.04	12.90 ±0.5
FRCN [28]	✓	✓	1.05	3.68	5.41	0.94	4.39	6.42	1.12	5.11	7.83	1.99	5.30	8.84
TFA_{fc} [36]	✓	✓	1.06	2.71	4.38	1.17	4.02	7.05	1.97	5.48	11.09	2.40	6.86	12.86
TFA_{cos} [36]	✓	✓	1.07	2.78	5.12	1.64	4.12	7.27	2.34	5.48	10.43	2.82	6.70	12.21
FSDetView [36]	✗	✓	0.70	2.70	3.70	0.60	4.00	4.20	1.80	5.10	8.00	3.00	9.70	12.30
MPSR [39]	✓	✓	1.23	3.82	5.58	1.89	5.69	8.73	0.86	4.60	9.96	1.62	6.78	11.66
A-RPN [6]	✗	✓	1.74	5.17	6.96	2.20	7.55	10.49	2.72	9.51	14.74	2.92	10.67	16.08
AirDet (Ours)	✗	✓	3.05	6.40	10.03	4.00	9.65	13.91	3.46	11.44	16.04	3.27	11.20	18.64

40] and AirDet on multi-scale objects in Table 2. Thanks to SCS, AirDet achieves the highest performance for multi-scale objects among all the SOTAs. Especially for small objects, given 5-shots, AirDet can achieve a surprising **4.22** AP_s, nearly doubling the fine-tuned methods with multi-scale FPN features [36, 39].

Comparison of 10-Shot   For a more thorough comparison, we present the 10-shot evaluation on the COCO validation dataset in Table 3. Without fine-tuning, AirDet can surprisingly achieve comparable performance against recent work [36, 39, 41], while all of them require a careful fine-tuning stage. Moreover, our fine-tuned model outperforms most prior methods [2, 3, 6, 12, 15, 16, 19, 31, 36–41, 45] in most metrics, especially average recall rate (AR). Besides, the performance superiority on small objects (AP_s and AR_s) further demonstrates the effectiveness of AirDet on multi-scale, especially the small-scale objects.

Efficiency Comparison   We report the fine-tuning and inference time of AirDet, and the SOTA methods [6, 36, 39, 40] in a setting of 3-shot one class

Table 3. Performance comparison with 10-shot on COCO validation dataset. Red and green fonts indicate best and second-best scores, respectively. AirDet achieves comparable results without fine-tuning and outperforms most methods with fine-tuning, which strongly demonstrates its effectiveness.

Method	Venue	Fine-tune	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR_1	AR_{10}	AR_{100}	AR_s	AR_m	AR_l
LSTD [3]	AAAI 2018	✓	3.2	8.1	2.1	0.9	2.0	6.5	7.8	10.4	10.4	1.1	5.6	19.6
MetaDet [37]	ICCV 2019	✓	7.1	14.6	6.1	1.0	4.1	12.2	11.9	15.1	15.5	1.7	9.7	30.1
FSRW [16]	ICCV 2019	✓	5.6	12.3	4.6	0.9	3.5	10.5	10.1	14.3	14.4	1.5	8.4	28.2
Meta RCNN [41]	ICCV 2019	✓	8.7	19.1	6.6	2.3	7.7	14.0	12.6	17.8	17.9	7.8	15.6	27.2
TFA_{fc} [36]	ICML 2020	✓	9.1	17.3	8.5	-	-	-	-	-	-	-	-	-
TFA_{cos} [36]	ICML 2020	✓	9.1	17.1	8.8	-	-	-	-	-	-	-	-	-
FSDetView [40]	ECCV 2020	✓	12.5	27.3	9.8	2.5	13.8	19.9	20.0	25.5	25.7	7.5	27.6	38.9
MPSR [39]	ECCV 2020	✓	9.8	17.9	9.7	3.3	9.2	16.1	15.7	21.2	21.2	4.6	19.6	34.3
A-RPN [6]	CVPR 2020	✓	11.1	20.4	10.6	-	-	-	-	-	-	-	-	-
SRR-FSD [45]	CVPR 2021	✓	11.3	23.0	9.8	-	-	-	-	-	-	-	-	-
FSCE [31]	CVPR 2021	✓	11.9	-	10.5	-	-	-	-	-	-	-	-	-
DCNet [15]	CVPR 2021	✓	12.8	23.4	11.2	4.3	13.8	21.0	18.1	26.7	25.6	7.9	24.5	36.7
Y. Li et al. [19]	CVPR 2021	✓	11.3	20.3	-	-	-	-	-	-	-	-	-	-
FADI [2]	NIPS 2021	✓	12.2	22.7	11.9	-	-	-	-	-	-	-	-	-
QA-FewDet [12]	ICCV 2021	✓	11.6	23.9	9.8	-	-	-	-	-	-	-	-	-
$FSOD^{up}$ [38]	ICCV 2021	✓	11.0	-	10.7	4.5	11.2	17.3	-	-	-	-	-	-
AirDet	**Ours**	✗	8.7	15.3	8.8	4.3	9.7	14.8	19.1	33.8	34.8	13.0	37.4	52.9
AirDet	**Ours**	✓	13.0	23.9	12.4	4.5	15.2	22.8	20.5	33.7	34.4	9.6	36.4	55.0

Table 4. Efficiency comparison with official source code. We adopt the pre-trained models provided by [36], so their fine-tuning time is unavailable.

Method	**AirDet**	A-RPN [6]	FSDet [40]	MPSR [39]	TFA_{fc} [36]	TFA_{cos} [36]	$FRCN_{ft}$ [36]
Fine-tuning (min)	**0**	21	11	3	-	-	-
Inference (s/img)	**0.081**	0.076	0.202	0.109	0.085	0.094	0.091

in Table 4, in which the official code and implementation with ResNet101 as the backbone are adopted. Without fine-tuning, AirDet can make direct inferences on novel objects with a comparable speed, while the others methods [6,36,39,40] require a fine-tuning time of about 3-30 minutes, which cannot meet the requirements of online exploration. Note that the fine-tuning time is measured on TITAN X GPU, while such computational power is often unavailable on robots. **Remark 6**: Many methods [2, 3, 6, 12, 15, 16, 19, 31, 36–41, 45] also require an offline process to fine-tune hyper-parameters for different shots. While such *offline* tuning is infeasible for robotic *online* exploration. Instead, AirDet can adopt **the same** base-trained model without fine-tuning for implementation.

5.3   Cross-domain Evaluation

Robots are often deployed to novel environments that have never been seen during training, thus cross-domain test is crucial for robotic applications. In this section, we adopt the same model trained on COCO, while test on PASCAL VOC [5] and LVIS [11] to evaluate the model generalization capability.
PASCAL VOC We report the overall performance on PASCAL VOC-2012 [5] for all methods in Table 5. In the cross-domain setting, even without fine-tuning, AirDet achieves better performance than methods [6,28,36,39,40] that perform relatively well in in-domain test. This means AirDet has a much stronger generalization capability than most fine-tuned prior methods.

Table 5. Cross-domain performance on VOC-2012 validation dataset. Red and green fonts denote the first and second place, respectively. AirDet has been demonstrated strong generalization capability, maintaining obvious superiority against others.

Shots		1			2			3			5		
Method	Fine-tune	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
A-RPN [6][†]	✗	10.45 ±0.1	18.10 ±0.1	10.32 ±0.1	13.10 ±0.2	22.60 ±0.4	13.17 ±0.2	14.05 ±0.2	24.08 ±0.2	14.24 ±0.2	14.87 ±0.08	25.03 ±0.07	15.26 ±0.1
AirDet (Ours)[†]	✗	**11.92 ±0.06**	**21.33 ±0.08**	**11.56 ±0.08**	**15.80 ±0.08**	**26.80 ±0.3**	**16.08 ±0.05**	**16.89 ±0.1**	**28.61 ±0.1**	**17.36 ±0.1**	**17.83 ±0.03**	**29.78 ±0.03**	**18.38 ±0.1**
FRCN [28]	✓	4.49	9.44	3.85	5.20	11.92	3.84	6.50	14.39	5.11	6.55	14.48	5.09
TFA_{cos} [36]	✓	4.66	7.97	5.14	6.59	11.91	6.49	8.78	17.09	8.15	10.46	20.93	9.53
TFA_{fc} [36]	✓	4.40	8.60	4.21	7.02	13.80	6.21	9.24	18.48	8.03	11.11	22.83	9.78
FSDetView [40]	✓	4.80	14.10	1.40	3.70	11.60	0.60	6.60	22.00	1.20	10.80	26.50	5.50
MPSR [39]	✓	6.01	11.23	5.74	8.20	15.08	8.22	10.08	18.29	9.99	11.49	21.33	11.06
A-RPN [6]	✓	9.49	17.41	9.42	12.71	23.66	12.44	14.89	26.30	14.76	15.09	28.08	14.17
AirDet (Ours)	✓	**13.33**	**24.64**	**12.68**	**17.51**	**30.35**	**17.61**	**17.68**	**32.05**	**17.34**	**18.27**	**33.02**	**17.69**

Table 6. Cross-domain performance of A-RPN [6] and AirDet on LVIS dataset. We report the results for 5-shot without fine-tuning on 4 random splits.

Split		1				2				3				4		
Metric	AP	AP_{50}	AP_{75}	AR_{10}	AP	AP_{50}	AP_{75}	AR_{10}	AP	AP_{50}	AP_{75}	AR_{10}	AP	AP_{50}	AP_{75}	AR_{10}
AirDet	**6.71**	**12.31**	**6.51**	**27.57**	**9.35**	**14.23**	**9.98**	**25.42**	**9.09**	**15.64**	**8.82**	**34.64**	**11.07**	**16.90**	**12.30**	**25.76**
A-RPN	5.49	10.04	5.27	26.59	8.85	13.41	9.46	24.45	7.49	12.34	8.13	33.85	10.80	15.46	12.24	25.05

LVIS We randomly sample LVIS [11] to form 4 splits of classes, each of which contains 16 different classes. To provide valid evaluation, the classes that have 20 to 200 images are taken for the test. More details can be found in Appendix E. The averaged performance with 5-shot without fine-tuning is presented in Table 6, where AirDet outperforms the baseline [6] in every split under all metrics. Since the novel categories in the 4 LVIS splits are more (64 classes in total) and rarer (many of them are uncommon) than the VOC 20 classes, the superiority of AirDet in Table 6 highly demonstrate its robustness under class variance.

5.4   Ablation Study and Deep Visualization

In this section, we address the effectiveness of the proposed three modules via quantitative results and qualitative visualization using Grad-Cam [29].

Quantitative Evaluation We report the overall performance on 3-shot and 5-shot for the baseline [6] and AirDet by enabling the three modules, respectively. It can be seen in Table 7 that AirDet outperforms the baseline in all cases. With the modules enabled one by one, the results get gradually higher, which strongly demonstrates the necessity and effectiveness of SCS, GLR, and PRE.

How effective is SCS? Given 2-shot per class, we first take the highest ranking proposal from RPN [28] to backpropagate the objectiveness score and resize the gradient map to the original image. Fig. 4 (a) exhibits the heat map from both AirDet and the baseline. We observe that AirDet generally concentrates on objects more precisely than the baseline. Moreover, AirDet can focus better on objects belonging to the support class and is not distracted by other objects

Table 7. Ablation study of the three modules, *i.e.*, PRE, GLR, and SCS in AirDet. With each module enabled, the performance is improved step by step on our baseline. With the full modules, AirDet can amazingly achieve up to **35%** higher results.

Module			3						5			
PRE	GLR	SCS	AP	$\Delta\%$	AP_{50}	$\Delta\%$	AP_{75}	$\Delta\%$	AP	$\Delta\%$	AP_{50}	$\Delta\%$	AP_{75}	$\Delta\%$
Baseline [6]			4.80	0.00	9.24	0.00	4.49	0.00	5.73	0.00	10.68	0.00	5.53	0.00
✓			5.15	+7.29	10.11	+9.41	4.71	+4.90	5.94	+3.66	11.54	+8.05	5.34	-3.43
✓	✓		5.59	+16.46	10.61	+14.83	5.12	+14.03	6.44	+12.39	12.08	+13.11	6.06	+9.58
✓	✓	✓	**6.50**	**+35.41**	**12.30**	**+33.12**	**6.11**	**+36.08**	**7.27**	**+26.78**	**13.63**	**+27.62**	**6.71**	**+21.34**

(a) Deep Attention of SCS   **(b) Deep Attention of Detection Head**
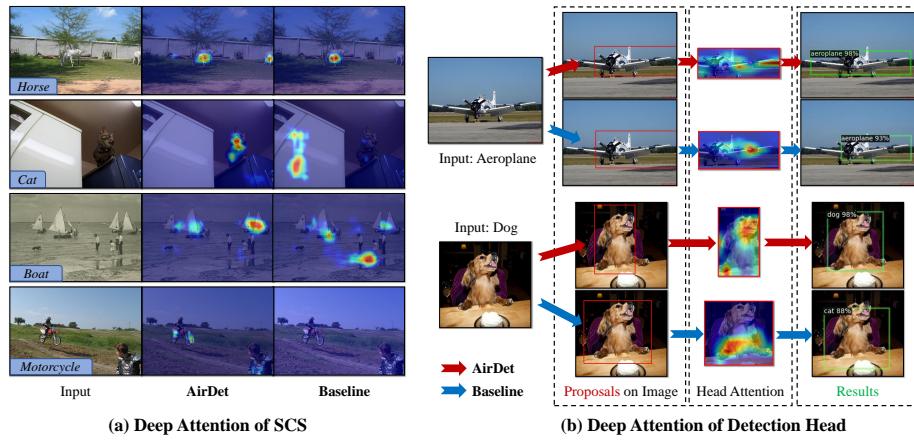
Fig. 4. Deep visualization comparison between AirDet and baseline [6]. In (a), By virtue of SCS, AirDet is capable of finding given support objects effectively. In (b), with similar proposals (red boxes), AirDet can focus on the entire object (aeroplane) and notice the most representative parts (dog), resulting in more precise regression box and correct classification results. More examples are presented in Appendix D.

(2nd and 3rd row). This means that AirDet can generate novel object proposals more effectively.

How effective is GLR and detection head? In Fig. 4 (b), we observe that with similar proposal boxes, AirDet head can better focus on the entire object, *e.g.*, aeroplane is detected with a precise regression box, *e.g.*, the dog is correctly classified with high score. This again demonstrates the effectiveness of our GLR and detection head.

5.5   Real-World Test

Real-world tests are conducted for AirDet and our baseline [6] with 12 sequences that were collected from the DARPA Subterranean (SubT) challenge [1]. Due to the requirements of *online* response during the mission, the models can only be evaluated **without fine-tuning**, which makes existing methods [2, 3, 6, 12,

Table 8. 3-shot real-world exploration test of AirDet and baseline [6]. AirDet can be directly applied without fine-tuning and performs considerably more robust than the baseline by virtue of the newly proposed SCS, GLR, and PRE modules.

Real-world Exploration Test												
Test/#Frames	1/#248		2/#146		3/#127		4/#41		5/#248		6/#46	
Metric	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
AirDet (Ours)	**17.10**	**54.10**	**17.90**	**47.40**	**24.00**	**57.50**	**26.94**	**48.20**	**11.28**	**38.17**	**20.40**	**70.63**
A-RPN [6]	13.56	40.40	14.30	38.80	20.20	47.20	22.41	40.14	6.75	24.10	14.70	59.38
Test/#Frames	7/#212		8/#259		9/#683		10/#827		11/#732		12/#50	
Metric	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
AirDet (Ours)	**5.90**	**16.00**	**15.26**	**43.31**	**7.63**	**27.88**	**13.55**	**23.92**	**15.74**	**34.43**	**21.45**	**45.83**
A-RPN [6]	2.39	7.60	11.27	25.24	6.16	23.40	8.10	14.85	11.54	27.28	18.20	33.98

Table 9. Per class results of the real-world tests. We report the instance number of each novel class along with the 3-shot AP results from AirDet and A-RPN [6]. Compared with the baseline, AirDet achieves higher results for all classes.

Class	Backpack	Helmet	Rope	Drill	Vent	Extinguisher	Survivor
Instances	626	674	723	587	498	1386	205
AirDet	**32.3**	**9.7**	**13.9**	**10.8**	**16.2**	**10.5**	**10.7**
Baseline [6]	26.6	9.7	6	9	14.4	5.6	9.1

15, 16, 19, 31, 36–41, 45] impractical. The environments of SubT challenge also poses extra difficulties, *e.g.*, a lack of lighting, thick smoke, dripping water, and cluttered or irregularly shaped environments, *etc.* To test the generalization capabilities, we adopt the same models of AirDet and the baseline as those evaluated in Section 5.2 and Section 5.3. The performance of 3-shot for each class is exhibited in Table 8, where AirDet is proved better. The robot is equipped with an NVIDIA Jetson AGX Xavier, where our method runs at 1-2 FPS without TensorRT acceleration or other optimizations.

In Table 9, we present the number of instances and the performance on each novel class. To our excitement, AirDet shows smaller variance and higher precision cross different classes. We also present the support images and representative detected objects in Fig. 5. Note that AirDet can detect the novel objects accurately in the query images even if they have distinct scales and different illumination conditions with the supports. We regard this capability to the carefully designed SCS in AirDet. More visualization are presented in Appendix C. The robustness and strong generalization capability of AirDet in the real-world tests demonstrated its promising prospect and feasibility for autonomous exploration.

6   Limitation and Future Work

Despite the promising prospect and outstanding performance, AirDet still has several limitations. (1) Since abundant base classes are needed to generalize, AirDet needs a relatively large base dataset to train before inference on novel classes. (2) Second, AirDet relies on the quality of support images to work well without fine-tuning. This is because the provided few support images are the
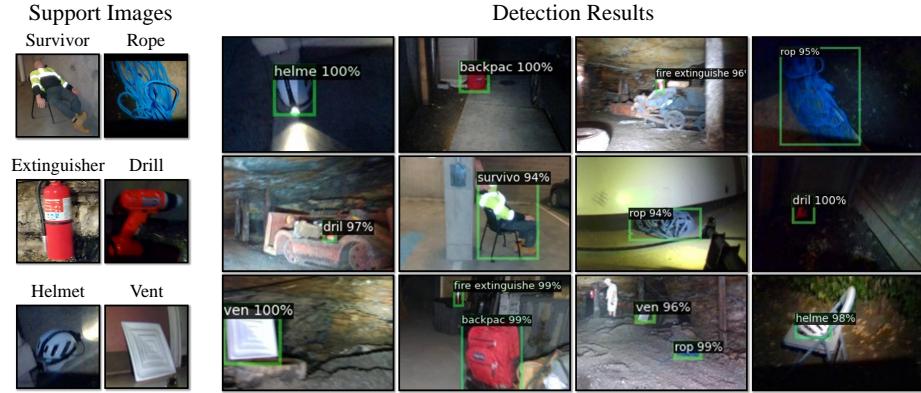
Support Images   Detection Results
Survivor   Rope

Extinguisher Drill

Helmet   Vent

Fig. 5. The provided support images and examples of detection results in the real-world tests. AirDet is robust to distinct object scales and different illumination conditions.

only information for the unseen classes. (3) We observe that the failure cases of AirDet are mainly due to false classification, resulting in a high result variance among different classes in COCO and VOC. (4) Since SCS and the detection head run in loops for multiple novel classes, the efficiency of AirDet will suffer from a large number of novel classes. We provide quantitative results for limitation (1), (2), and (3) in Appendix F.

7   Conclusion

This paper presents a brand new few-shot detector, AirDet, which consists of 3 newly proposed *class-agnostic relation*-based modules and is free of fine-tuning. Specifically, with proposed spatial relation and channel relation, we construct support guided cross-scale feature fusion for region proposals, global-local relation network for shots aggregation, and prototype relation embedding for precise localization. With the strong capability to extract *class-agnostic relation*, AirDet can work comparably or even better than those exhaustively fine-tuned methods in both in-domain and cross-domain evaluation. AirDet is also tested on real-world data with a robotic platform, where its feasibility for autonomous exploration is demonstrated.

References

1. https://subtchallenge.com 12, 20
2. Cao, Y., Wang, J., Jin, Y., Wu, T., Chen, K., Liu, Z., Lin, D.: Few-Shot Object Detection via Association and DIscrimination. Advances in Neural Information Processing Systems (NIPS) **34** (2021) 1, 5, 8, 9, 10, 12
3. Chen, H., Wang, Y., Wang, G., Qiao, Y.: LSTD: A Low-Shot Transfer Detector for Object Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018) 9, 10, 12
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A Large-Scale Hierarchical Image Database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009) 5, 8
5. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision **88**(2), 303–338 (2010) 10, 20
6. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4013–4022 (2020) 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 18, 19, 20, 21, 22, 23
7. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized Few-Shot Object Detection without Forgetting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4527–4536 (2021) 1, 2, 8
8. Farooq, N., Ilyas, U., Adeel, M., Jabbar, S.: Ground Robot for Alive Human Detection in Rescue Operations. In: Proceedings of the International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS). vol. 3, pp. 116–123 (2018) 2
9. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1440–1448 (2015) 3
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). vol. 3, pp. 580–587 (2014) 3
11. Gupta, A., Dollar, P., Girshick, R.: LVIS: A Dataset for Large Vocabulary Instance Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5356–5364 (2019) 10, 11, 22
12. Han, G., He, Y., Huang, S., Ma, J., Chang, S.F.: Query Adaptive Few-Shot Object Detection with Heterogeneous Graph Convolutional Networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3263–3272 (2021) 4, 9, 10, 12
13. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2961–2969 (2017) 3, 18
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016) 8, 18, 19
15. Hu, H., Bai, S., Li, A., Cui, J., Wang, L.: Dense Relation Distillation with Context-aware Aggregation for Few-Shot Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10185–10194 (2021) 4, 9, 10, 12

16. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-Shot Object Detection via Feature Reweighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8420–8429 (2019) 1, 2, 6, 7, 8, 9, 10, 12

17. Kong, T., Yao, A., Chen, Y., Sun, F.: HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 4

18. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4282–4291 (2019) 6

19. Li, Y., Zhu, H., Cheng, Y., Wang, W., Teo, C.S., Xiang, C., Vadakkepat, P., Lee, T.H.: Few-Shot Object Detection via Classification Refinement and Distractor Retreatment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15395–15403 (2021) 1, 2, 9, 10, 12

20. Li, Z., Zhou, F.: FSSD: Feature Fusion Single Shot Multibox DAetector. arXiv preprint arXiv:1712.00960 (2017) 3, 4

21. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2117–2125 (2017) 4

22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 740–755 (2014) 8, 20

23. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single Shot Multibox Detector. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 21–37 (2016) 3, 4

24. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: Defrcn: Decoupled faster r-cnn for few-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8681–8690 (2021) 1, 2, 4, 5

25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-time Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016) 3

26. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7263–7271 (2017) 3, 4

27. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767 pp. 1–6 (2018) 3

28. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2016) 2, 3, 4, 6, 8, 9, 10, 11

29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 618–626 (2017) 11, 20

30. Shen, Z., Liu, Z., Li, J., Jiang, Y.G., Chen, Y., Xue, X.: DSOD: Learning Deeply Supervised Object Detectors From Scratch. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017) 4

31. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding. In: Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR). pp. 7352–7362 (2021) 1, 2, 4, 5, 8, 9, 10, 12

32. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to Compare: Relation Network for Few-Shot Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1199–1208 (2018) 3, 4

33. Tariq, R., Rahim, M., Aslam, N., Bawany, N., Faseeha, U.: Dronaid: A Smart Human Detection Drone for Rescue. In: Proceedings of the 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT). pp. 33–37 (2018) 2

34. Wang, C., Qiu, Y., Wang, W., Hu, Y., Kim, S., Scherer, S.: Unsupervised Online Learning for Robotic Interestingness with Visual Memory. IEEE Transactions on Robotics pp. 1–15 (2021) 2

35. Wang, C., Wang, W., Qiu, Y., Hu, Y., Scherer, S.: Visual Memorability for Robotic Interestingness via Unsupervised Online Learning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 52–68 (2020) 2

36. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly Simple Few-Shot Object Detection. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 1–12 (2020) 1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 23

37. Wang, Y.X., Ramanan, D., Hebert, M.: Meta-Learning to Detect Rare Objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9925–9934 (2019) 1, 2, 6, 9, 10, 12

38. Wu, A., Han, Y., Zhu, L., Yang, Y.: Universal-Prototype Enhancing for Few-Shot Object Detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9567–9576 (2021) 4, 5, 8, 9, 10, 12

39. Wu, J., Liu, S., Huang, D., Wang, Y.: Multi-Scale Positive Sample Refinement for Few-Shot Object Detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 456–472 (2020) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 23

40. Xiao, Y., Marlet, R.: Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 192–210 (2020) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 23

41. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9577–9586 (2019) 1, 3, 4, 5, 6, 7, 8, 9, 10, 12

42. Zhang, L., Zhou, S., Guan, J., Zhang, J.: Accurate Few-Shot Object Detection With Support-Query Mutual Guidance and Hybrid Loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14424–14432 (2021) 1, 2, 3

43. Zhang, L., Zhou, S., Guan, J., Zhang, J.: Accurate Few-Shot Object Detection With Support-Query Mutual Guidance and Hybrid Loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14424–14432 (2021) 4

44. Zhang, W., Wang, Y.X.: Hallucination Improves Few-Shot Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13008–13017 (2021) 8, 9

45. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic Relation Reasoning for Shot-Stable Few-Shot Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8782–8791 (2021) 1, 2, 4, 5, 8, 9, 10, 12

Supplementary Material

Bowen Li[1,2], Chen Wang[1], Pranay Reddy[1,3],
Seungchan Kim[1], and Sebastian Scherer[1]

[1] Robotics Institute, Carnegie Mellon University, USA
chenwang@dr.com, {bowenli2,seungch2,basti}@andrew.cmu.edu
[2] School of Mechanical Engineering, Tongji University, China
[3] Electronics and Communication Engineering, IIITDM Jabalpur, India
2018033@iiitdmj.ac.in

Overview

To ensure reproducibility, we present the detailed configuration in Appendix A and show backbone comparison in Appendix B for thoroughness. More qualitative results from general detection datasets, VOC-2012 validation dataset, and COCO validation dataset, as well as the representative scenes from the DARPA SubT challenge are presented in Appendix C. We also displayed more deep visualization in Appendix D to further validate the effectiveness of SCS and detection head of AirDet. Details about the LVIS dataset splits are in Appendix E. The limitations of AirDet are also more exhaustively studied in Appendix F.

A Detailed Configuration

Training: We follow our baseline [6], where contrastive training pipeline is adopted. We first reconstruct the COCO-2017 training dataset, ensuring that only one class object is annotated for each query image. During training, for one certain query image with objects belonging to class c_1, we provide 20 support images, including ten belonging to class c_1 and ten from another random class c_2, termed as 2-way 10-shot contrastive training. Following our baseline [6], the support images are cropped, resized, and zero-padded to 320×320 pixels.

Inference: Considering N-way K-shot inference, we provide all K-shot support data from N novel classes for one query image. For each novel class, 100 proposals are generated from the support-guided cross-scale fusion (SCS) module, and they are ranked according to the detection confidence. We finally take the top 100 proposals in all the $N \times 100$ candidates for calculating the final performance.

Parameters: The input of SCS are feature maps from ResNet2, ResNet3, and ResNet4 block. We use a global averaged support feature (weights of MLP and Conv in (1)) are all 1) as the 1×1 convolutional kernel for multi-scale query feature. ROI Align strategy [13] is employed for pooling. The default learning rate for both 2 models with ResNet50 and ResNet101 [14] backbone is 0.004. The model employing ResNet50 is trained for a total of $120,000$ iterations with ResNet1, 2, 3 blocks frozen, while the ResNet101 backbone model is trained for

Table 1. Performance comparison of AirDet with different backbones on COCO validation dataset. The model with ResNet101 backbone performs generally better.

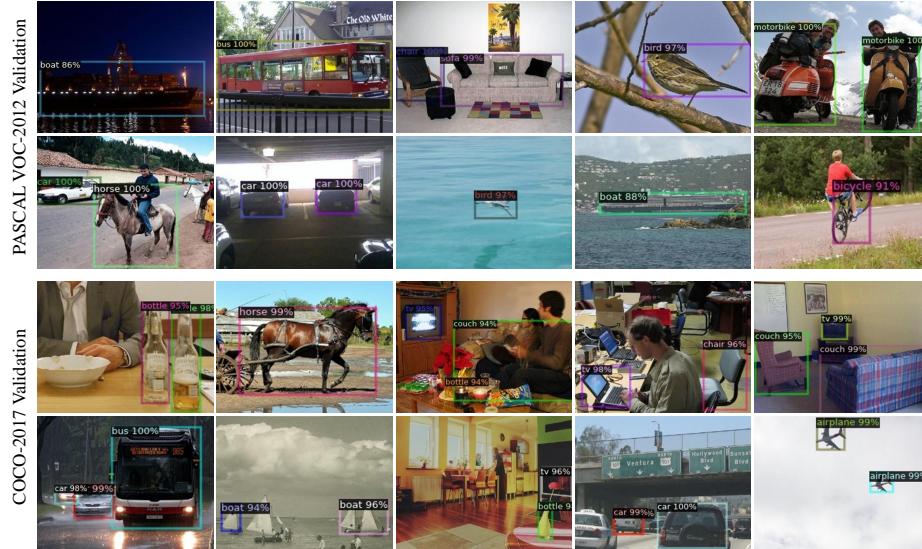| Shots	1				2				3				5			
Metric	AP	AP_{50}	AP_{75}	AP_s	AP	AP_{50}	AP_{75}	AP_s	AP	AP_{50}	AP_{75}	AP_s	AP	AP_{50}	AP_{75}	AP_s
ResNet101	6.00	10.78	5.92	2.77	6.63	12.12	6.37	3.33	6.94	12.96	6.49	4.05	7.63	13.86	7.34	4.32
ResNet50	4.64	9.60	3.97	1.82	5.59	10.81	5.16	2.73	6.38	12.29	5.83	2.83	7.43	13.78	7.17	3.30

Fig. 1. Representative examples of 3-shot detection of AirDet on VOC-2012 validation and COCO validation dataset. Without fine-tuning, AirDet can robustly detect the novel unseen objects such as boat, bus, sofa, with merely three support images.

$80,000$ iterations with only ResNet1 block frozen. We observe that for a deeper backbone, freeing ResNet2 and 3 blocks will help the SCS module generate effective proposals better. For both the two models, the detection head takes a learning rate of 0.008. We have maintained all other parameters the same as our baseline [6]. Please refer to the attached code for more details.

B   Backbone Comparison

The performance comparison of AirDet with different backbones [14] are shown in Table 1. We report the average results of AirDet with ResNet101 and ResNet50 backbone on COCO validation dataset, using the same support examples. We find the model with ResNet101 generally perform better, while the switch to ResNet50 also doesn't result in too severe performance drop, which demonstrates the universal property of AirDet architecture.

Fig. 2. Representative examples of 3-shot detection of AirDet on the DARPA Subterranean Challenge. Provided with merely three support images for these unseen novel objects, AirDet can directly detect them with scale variation and partial occlusion in distinct environments and illumination conditions.

C   More Qualitative Results

More qualitative detection results from the VOC-2012 [5] validation dataset and COCO [22] validation dataset are shown in Fig. 1. Provided with merely 3-shot support images per novel class, AirDet can directly detect unseen objects in various scales and distinct viewpoints from different environments.

We also exhibit more representative 3-shot detection results from the DARPA Subterranean Challenge [1] without fine-tuning in Fig. 2. For each novel class, *i.e.*, helmet, rope, backpack, drill, vent, survivor, and fire-extinguisher, we have selected the objects from distinct environments including in-door, out-door, cave, tunnel, *etc.* We find AirDet can maintain robust when faced with the challenging factors during exploration, *e.g.*, illumination variation (examples from the helmet), partial occlusion (the second and third-row in the backpack), scale variation (examples from drill), and blur (the last row in survivor). Moreover, AirDet generally outputs high classification scores (higher than 0.9) and precise bounding boxes for the novel unseen classes, which demonstrate the promising prospect of AirDet for autonomous exploration tasks.

D   More Deep Visualization

To better demonstrate why the proposed AirDet works well, we present more deep visualization [29] with 5-shot supports for the proposal generation and the detection head. As shown in Fig. 3, we backpropagate the gradient of highest objectiveness score generated by AirDet and our baseline [6] to the whole image. Compared with baseline, AirDet can better notice and concentrate on the novel
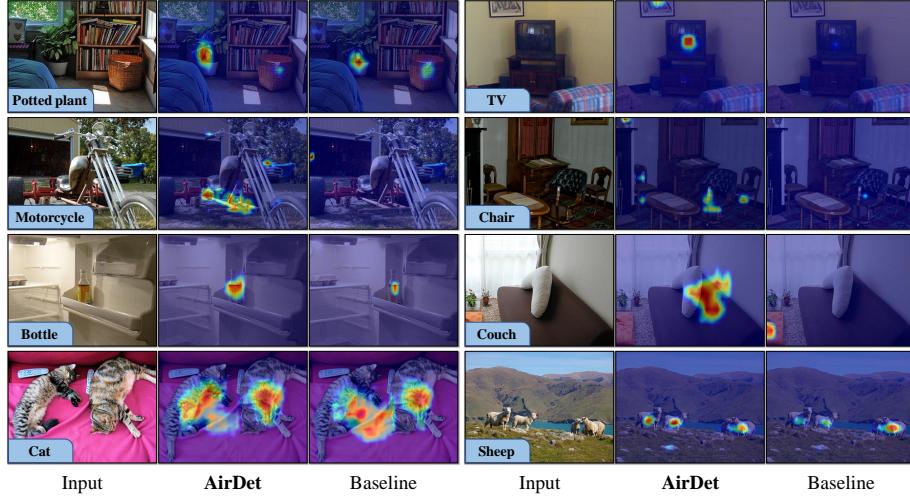
Fig. 3. Deep visualization from the proposal generation module of AirDet and baseline [6]. Compared with baseline, AirDet can better notice and concentrate on and novel object region, which leads to its more effective region proposals.

Table 2. Detailed information about the 4 splits in LVIS dataset.

Split1				Split2			
Class	Instance	Class	Instance	Class	Instance	Class	Instance
bath mat	63	mousepad	66	ashtray	51	billboard	270
birthday cake	74	pan	242	taxi	68	dresser	39
blender	57	paper plate	170	duck	134	figurine	168
blouse	99	printer	59	guitar	52	hair dryer	32
chandelier	66	saddle blanket	94	fume hood	33	polar bear	36
Christmas tree	72	saucer	103	ottoman	48	pajamas	54
grill	90	stool	126	radiator	41	scale	46
mattress	74	tinfoil	210	shoulder bag	61	urinal	237

Split3				Split4			
Class	Instance	Class	Instance	Class	Instance	Class	Instance
blackboard	37	bridal gown	23	bear	116	cistern	182
bullet train	25	doormat	28	paper towel	171	parking meter	282
fire engine	42	fish	92	pickup truck	209	pot	121
hairbrush	28	kettle	31	saddle	320	saltshaker	105
map	36	piano	24	ski parka	428	soccer ball	258
radio	23	teapot	38	statue	204	sweatshirt	427
tongs	44	cover	49	tarp	160	towel	762
tripod	24	wallet	29	vest	168	wine bottle	223

region. For example, in the category "cat", AirDet is not distracted by the carpet while our baseline loses its attention. Also, in "sheep", AirDet can notice all the novel instances well but baseline may miss several instances. Therefore, by virtue of the SCS module, AirDet can generated more effective proposals.

In Fig. 4, we backpropagate the gradient of highest classification score to the corresponding proposal image patch (red box). Thanks to the more repre-
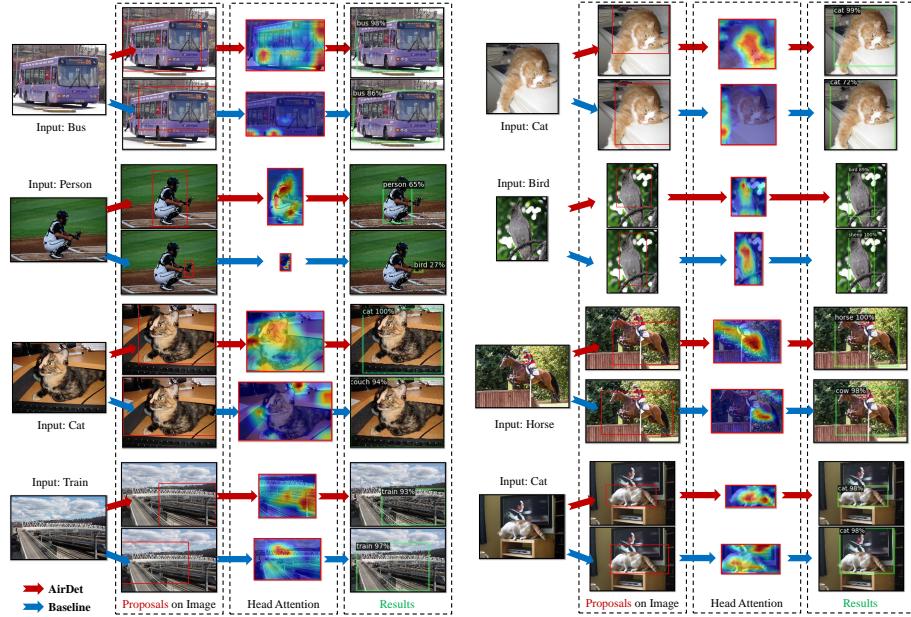
Fig. 4. Deep visualization from the detection head of AirDet and baseline [6]. With similar proposals in red boxes, AirDet can focus more precisely on the most representative part of the object, resulting in more accurate box regression and classification.

sentative class prototype from GLR and the fully relation-based detection head, AirDet is more capable of precisely predicting the category and box of an instance. For example, let's consider the cat at the bottom left. With similar proposal box, AirDet can concentrate better on the object region while baseline is distracted by its context. Consequently, the baseline incorrectly classifies a 'cat' as a 'couch'. Another example is the cat at bottom right, since AirDet can focus better on the object, it produces a more accurate bounding box than baseline.

E   Details About LVIS dataset

We introduce the detailed information for LVIS [11] dataset. The class names and the number of instances for each class are shown in Table 2. It includes 64 classes and is sampled into 4 splits, each of which contains 890, 502, 365, and 1586 images, respectively.

F   Detailed Limitations

Dependence of exhaustive base training: One potential limitation of AirDet is that it requires a large number of base classes during training to

Table 3. Cross-domain performance on VOC-2012 validation dataset. AirDet is pre-trained with different number of classes (left) and instances (right).

Cls/Inst.	50/123,258			55/140,682			60/148,872		
Shots	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
3	6.54	13.13	5.64	11.20	20.74	11.02	**16.89**	**28.61**	**17.36**
5	7.05	14.26	5.85	11.76	21.42	11.53	**17.83**	**29.78**	**18.38**

Table 4. Comparison of average results in the real-world tests using the same (left) or different (right) objects as support examples.

Metric	AP		AP_{50}		AR_1		AR_{10}	
Support	same	diff.	same	diff.	same	diff.	same	diff.
AirDet	**16.4**	**10.4**	**42.3**	**25.2**	**23.6**	**18.0**	**28.6**	**23.2**
A-RPN	12.5	9.4	31.9	21.7	20.3	15.6	23.8	21.4

Table 5. 3-shot per class AP results of AirDet without fine-tuning in COCO validation dataset and VOC-2012 validation dataset.

Category	COCO	VOC	Category	COCO	VOC	Category	COCO	VOC	Category	COCO	VOC
aeroplane	11.5	17.7	bicycle	1.11	7.5	pottedplant	0.3	2.8	sheep	5.2	17.1
boat	2	1.0	bottle	4.1	12.1	train	3.5	11.5	tvmonitor	20.7	21.7
car	13	36.4	cat	15.7	24.7	bird	4.3	21.2	dog	8.2	19.0
cow	2.4	14.0	diningtable	0.4	1.2	bus	30.4	23.0	person	1.3	2.2
horse	11.7	20.2	motorbike	5.4	4.1	chair	2.5	7.4	sofa	9	14.1

generalize. To exhaustively study this, we present the results on VOC-2012 validation dataset, which are obtained using models trained with different numbers of base classes. As shown in Table 3, pre-training with fewer base classes and instances can make the model degrade.

Dependence of high quality support images: The robots in the real-world will utilize the *online* defined objects (supports) to find the specific objects, where the supports are in good quality and appearance. Without this beneficial condition, the performance of AirDet will drop as shown in Table 4. Yet AirDet can still identify the novel objects in the tests compared with A-RPN [6] even with significant appearance change.

Result variance among different classes: As aforementioned, the failure cases of AirDet in COCO and VOC datasets are mainly due to false classification, which results in high variance among different classes. We present the average precision for each novel class with a 3-shot evaluation setting in Table 5. We observe that for novel classes like TV, monitor, and bus, the average precision can be up to 20 and 30. Nevertheless, for some other novel classes, *e.g.*, boat, potted plant, and dining table, the scores are much lower. Such a high variance among different classes indicates the limitation of the classifier in the detection head. We observe that such a limitation also exists in other SOTAs [6,36,39,40], which guides our future work.