2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

AirShot: Efficient Few-Shot Detection for Autonomous Exploration

Zihan Wang[1,2], Bowen Li[2], Chen Wang[3], and Sebastian Scherer[2]

***Abstract*— Few-shot object detection has drawn increasing attention in the field of robotic exploration, where robots are required to find unseen objects with a few online provided examples. Despite recent efforts have been made to yield online processing capabilities, slow inference speeds of low-powered robots fail to meet the demands of real-time detection-making them impractical for autonomous exploration. Existing methods still face performance and efficiency challenges, mainly due to unreliable features and exhaustive class loops. In this work, we propose a new paradigm AirShot, and discover that, by fully exploiting the valuable correlation map, AirShot can result in a more robust and faster few-shot object detection system, which is more applicable to robotics community. The core module Top Prediction Filter (TPF) can operate on multi-scale correlation maps in both the training and inference stages. During training, TPF supervises the generation of a more representative correlation map, while during inference, it reduces looping iterations by selecting top-ranked classes, thus cutting down on computational costs with better performance. Surprisingly, this dual functionality exhibits general effectiveness and efficiency on various off-the-shelf models. Exhaustive experiments on COCO2017, VOC2014, and SubT datasets demonstrate that TPF can significantly boost the efficacy and efficiency of most off-the-shelf models, achieving up to 36.4% precision improvements along with 56.3% faster inference speed. Code and Data are at: https://github.com/ImNotPrepared/AirShot.**

I. INTRODUCTION

Few-shot object detection (FSOD) [1], [2], [3], [4], [5] aims to detect objects out of the base training set with only a few support examples per novel class. This research area has garnered increasing interest in the robotics community as it plays an essential role in autonomous exploration [6], [7], [8], [9] where robots are expected to detect novel objects in an unknown environment while only a limited number of examples can be provided online by a human operator [9].

However, most FSOD methods [3], [4], [5], [10], [11], [12], [13], [14], [15], [16] cannot be directly applied to real-world robots because they are computationally heavy [8]. One of the reasons is that they require an offline fine-tuning stage on novel classes, which is impractical for *robot online* exploration. Even the models [1], [8] that can work without fine-tuning, still have mainly two drawbacks hindering their effectiveness in robotics. Firstly, two-stage detection frameworks [1], [8] heavily rely on region proposals to produce final predictions, which can easily fall short due to inaccurate proposals. Yet, producing high-quality region proposals for novel classes is difficult since their semantic knowledge is not learned in the feature extractor [1], [17]. This issue becomes even more pronounced in the context of robot exploration, where the few-shot setting typically provides limited information. Secondly, the inference stage of previous designs [1], [2], [8] follows the most exhaustive paradigm, *i.e.*, running loop inference on all potential novel classes. Such a design imposes a significant computational burden, particularly on low-powered robotic platforms, making it inefficient and impractical for real-world applications.

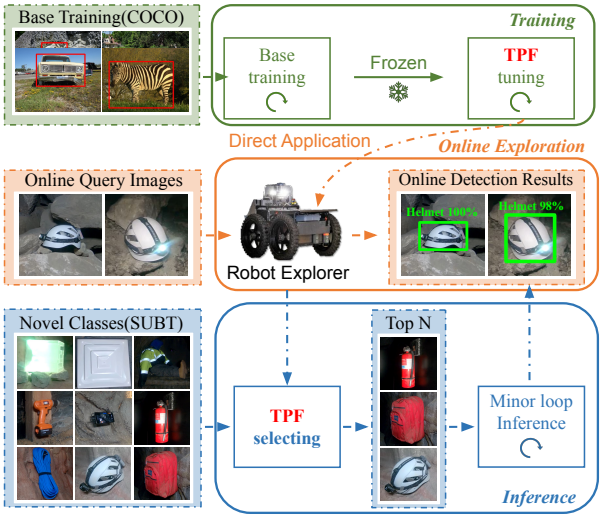

Fig. 1: Application sketch of AirShot. During training, we use TPF to increase the representation capability of correlation maps. When directly applied to Robot Explorer, TPF conducts pre-selection to enable minor loop inference instead of traditional full loops.

In this work, we find that a valuable correlation map can be used to solve the two problems in a unified manner. As shown in Fig. 1, we introduce a simple yet effective module, Top Prediction Filter (TPF) that operates on the correlation maps during both training (green) and inference (blue) stages.

Training Stage The supervision signals of most previous work [1], [3], [4], [5] are provided for generating, classifying, and regressing the region proposals, which are primarily for *local* anchors. However, the semantic cues directly from *global* correlation maps are overlooked. Thus our insight is intermediate supervision on the *global* feature could be beneficial. Specifically, TPF is trained to infer the existence of the category of support image directly from the global correlation maps via contrastive loss [18], which generates a higher quality and more reliable correlation map.

Inference Stage Most existing methods [1], [4], [5], [8] per-

[1]Institute for Imaging, Data and Communications, School of Engineering, The University of Edinburgh, UK zwang114@ed.ac.uk
[2]AirLab, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA zihanwa3@cs.cmu.edu; bowenli2@andrew.cmu.edu; basti@andrew.cmu.edu
[3]Spatial AI & Robotics (SAIR) Lab, Institute for Artificial Intelligence and Data Science, Department of Computer Science and Engineering, University at Buffalo, NY 14260, USA. chenw@sairlab.org

form a full loop over all potential novel classes, which is not only computationally intensive—making them impractical for tasks with low-powered robots—but also leads to slower inference speeds, failing to meet the demands of real-time detection. We observed that not all the potential novel classes will appear on the test images, thus the exhaustive loops have wasted much time on the absent classes. This inspired us to design a lightweight pre-selector to avoid the extra cost during inference. In AirShot, TPF can conduct a rough classification, directly inferring the existence and offering confidence scores of certain categories. For those categories with high confidence scores (top prediction classes), they will be fed to the following networks for finer classification and precise location. In contrast, those with low scores are considered unlikely to appear thus discarded.

Our TPF can actively provide supervision on correlation maps for more reliable and representative features during training, which effectively extract additional valuable information from limited data in few-shot scenarios for robot exploration. Moreover, its pre-selection ability significantly reduces the computational cost of low-powered robot system and further enables time-efficient detection. This dual functionality allows our approach to substantially improve the efficiency and effectiveness of the off-the-shelf models. In general, we summarize our contribution as twofold:

- We proposed a new model AirShot that fully exploits the valuable knowledge in the correlation map. Its core module TPF, a plug-and-play design, also works generally for various FSOD models [1], [8]. The efficiency and effectiveness brought by TPF offer substantial advancements in robot autonomous exploration task.
- We comprehensively test AirShot across two widely used datasets, MS-COCO and Pascal VOC. AirShot yields a 36.4% performance improvement and a 56.28% reduction in computational costs, which demonstrate a new SOTA in the field of FSOD that requires no fine-tuning. To demonstrate the effectiveness for real-world environments, we also tested our system with a challenging FSOD dataset, collected from the DARPA Subterranean (SubT) Challenge by our team.

II. Related Work

A. General Object Detection

Object detection [19], [20], [21], [22], [23] constitutes a pivotal challenge in the computer vision community, which aims to predict the categories and locations of predefined objects in an image. Modern object detection methods are primarily categorized into two classes: anchor-free and anchor-based detectors. Anchor-free detectors utilize a one-stage model structure, avoiding the explicit generation of proposal boxes. These methods either tackle object detection as an end-to-end regression problem, such as YOLO series [22], [24], [25] or utilize pre-defined bounding boxes to address varying object scales, as in the SSD series [21]. Anchor-based detectors [20], first generate class-agnostic region proposals using Region Proposal Network (RPN). These proposals are further refined and classified into different categories through detection heads. By filtering out negative locations using RPN, anchor-based methods often achieve decent results in general detection tasks [23], [26]. Both anchor-based and anchor-free methods require intensive supervision and have a fixed number of object classes post-training, making them unsuitable for tasks such as autonomous exploration, where unseen and novel objects appear dynamically.

B. Few-Shot Object Detection

Few-shot object detection leverages the knowledge from abundant base class data to generalize to novel categories with only a few labeled examples as support. Two main branches are thriving in recent FSOD research, namely meta-learning-based approaches [1], [15], [27], [28] and transfer-learning-based approaches [5], [13], [14], [29], [30].

Transfer-learning [5], [13], [14], [29], [30] aims to identify the best fine-tuning strategy to adapt general object detectors to novel images with limited examples. For instance, Wang *et al.* [5] proposed fine-tuning only the last layer. Wu *et al.* addressed scale scarcity through manually defined positive refinement branches, such as MPSR [15]. Other works also explored semantic relationships between novel and base classes applying contrastive proposal encoding [14].

Meta-learning-based methods aim to train meta-models on individual tasks in an episodic manner, with separate branches for extracting support information and detecting objects within query images. Notable contributions in this area include Meta R-CNN [27] and Meta-DETR [31]. Meta R-CNN [27] focused on support-guided query channel attention, A-RPN [1] with novel attention RPN and a multi-relation classifier. Meta-DETR [31], the current state-of-the-art, introduced a correlation aggregation module to simultaneously aggregate multiple support categories to capture their inter-class correlation. Recent developments include support-query mutual guidance [16], context information aggregation [2], and the construction of heterogeneous graph convolutional networks on proposals [28]. Transformer-based method FCT [32] proposed a model with three interaction stages between query and support in the backbone and one additional interaction stage in the multi-relation detection head, which computes similarities between support and query features to output the final detection results.

Despite recent progress, few-shot learning still faces several challenges. Region-based detection frameworks rely on region proposals to produce final predictions, however, it is not easy to produce high-quality region proposals for novel classes with limited supervision under the few-shot detection setups [31] and the need for re-training with new categories also limits the development. Therefore, a more robust, flexible, and efficient few-shot object detection method that can better adapt to different scenarios and novel categories without fine-tuning stage is needed.

C. Multi-Scale Feature Extraction

The utilization of multi-scale features in the detection of objects with varying scales has been extensively explored [21], [22], [33], [34], [26], [35], which has been

proved necessary for small object detection. For example, the feature pyramid network (FPN) employed a multi-scale feature map for detection. To exploit information in multi-scale features, FSSD [34] and AirDet [8] proposed a multi-scale feature fusion module. While others [14], [16], [29], [26] employed all scales from FPN and implemented detection on each scale parallel. Granted, a multi-scale feature is necessary to effectively detect small novel classes, but existing methods are cumbersome when running it for all classes. In contrast, our model AirShot can utilize multi-scale features from correlation map more efficiently.

D. Few-Shot Detection without fine-tuning

Although FSOD has made significant progress recently, most existing methods [1], [8] followed the two-stage training paradigm, which requires base training and fine-tuning stages. However, the fine-tuning stage cannot be applied to robot online exploration due to the following concerns: 1) dynamically changing categories, 2) limited onboard computing power of the robot, 3) non-existing evaluation dataset [8]. This task is important for robot exploration and perception in unseen environments, yet, only little progress has been made so far [1], [8]. In this work, the core module of AirDet, TPF, acts as a simple yet effective approach to boost both the efficiency and efficacy of these methods [1], [8].

III. Methodology

A. Preliminary

1) Exploration Task: The exploration task is executed under a few-shot object detection setting. Due to the unseen nature of objects during exploration, it requires model trained on base classes $\mathcal{C}_\mathrm{b}$ can detect novel classes in $\mathcal{C}_\mathrm{n}$, satisfying $\mathcal{C}_\mathrm{n} \cap \mathcal{C}_\mathrm{b} = \emptyset$. To save human effort, only a few annotated images (k-shot samples per novel class) are available during online exploration. Note that fine-tuning is not practicable under this setting. During the exploration, the robot will continuously collect images. The human user needs to annotate the novel objects first and provide them back to the robot explorer. Then the robot explorer should detect unseen objects by observing the surrounding environment. The main challenge of autonomous exploration task exhibits in two main perspectives: one is how to fully utilize limited information in few-shot setting to provide effective supervision, the other is how to run efficient inference when the inherit computation cost is limited for low-powered robots.

2) AirDet Review: AirDet [8] is a meta-learning-based few-shot object detector as shown in Fig. 2. It shows favorable performance without finetuning thanks to its class-agnostic design. With the proposed spatial relation and channel relation, AirDet constructs a support-guided cross-scale (SCS) module as a feature fusion of region proposals, a global-local relation network for shots aggregation, and a prototype relation embedding for precise localization. During exploration, a few prior raw images containing novel objects are sent to a human user for annotation as support images.

The support images and the new query image ($\mathbf{q}$) perceived by the explorer are fed into a shared backbone.

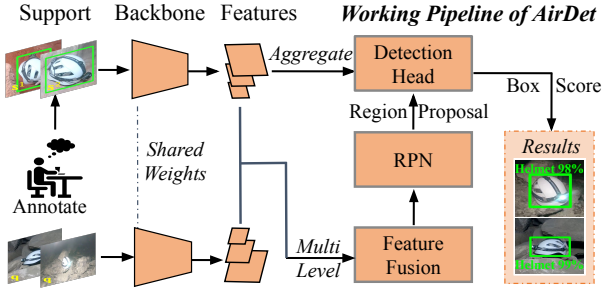

Fig. 2: Working pipeline of AirDet. AirDet includes 3 modules, i.e., the shared backbone, feature fusion module for region proposal and shots aggregation, plus relation-based detection head.

Then SCS accepts extracted multi-level support and query features from different backbone blocks (ResNet [36] 2, 3, and 4 blocks) as input. The SCS module utilizes the relation between query and support multi-scale features to generate fused correlation maps for region proposals. Finally, the region proposal and aggregated class prototype from the shots aggregation module are fed to the relation-based detection head for fine classification and regression.

B. Overview of Top Prediction Filter (TPF)

Despite great performance, we observe that the correlation map in AirDet is not reliable enough. Moreover, like various other FSOD methods [1], [8], [12], it has an efficiency problem, i.e., full inference loops. As in Table I, the backbone feature extraction runs fast, feature fusion (SCS in our case), RPN, and detection head occupy a majority of computational cost. Intuitively, we do not need to exhaustively search every

TABLE I: Time Consumption for Modules

	Backbone	SCS	RPN	Detection Head
Time(s)	0.013	0.099	0.115	0.506
Proportion(%)	1.81	13.55	15.65	68.99

class with full inference loops on these modules. Instead, we can pre-select the classes and reduce the looping iterations to improve efficiency. To this end, we proposed a new module named Top Prediction Filter (TPF) in Fig. 4, which is proven widely effective for various few-shot detectors.

As shown in the top row of Fig. 3, during training, TPF is injected to provide supervision on the fused correlation map to enhance the robustness of proposal generation. As for the inference, we feed *only* the level-4 correlation map into TPF to calculate the score for each class. The score here represents the confidence regarding the existence of a novel class. Then a list of ranked scores is responsible for selecting a reduced number of novel classes. All levels of correlation maps of the remaining novel classes are then sent into the second loop including a feature fusion, proposal generation, and detection head for classification and regression.

C. Top Prediction Filter (TPF)

Our main motivation is to reduce the computational burden caused by novel classes that are unlikely to appear in the

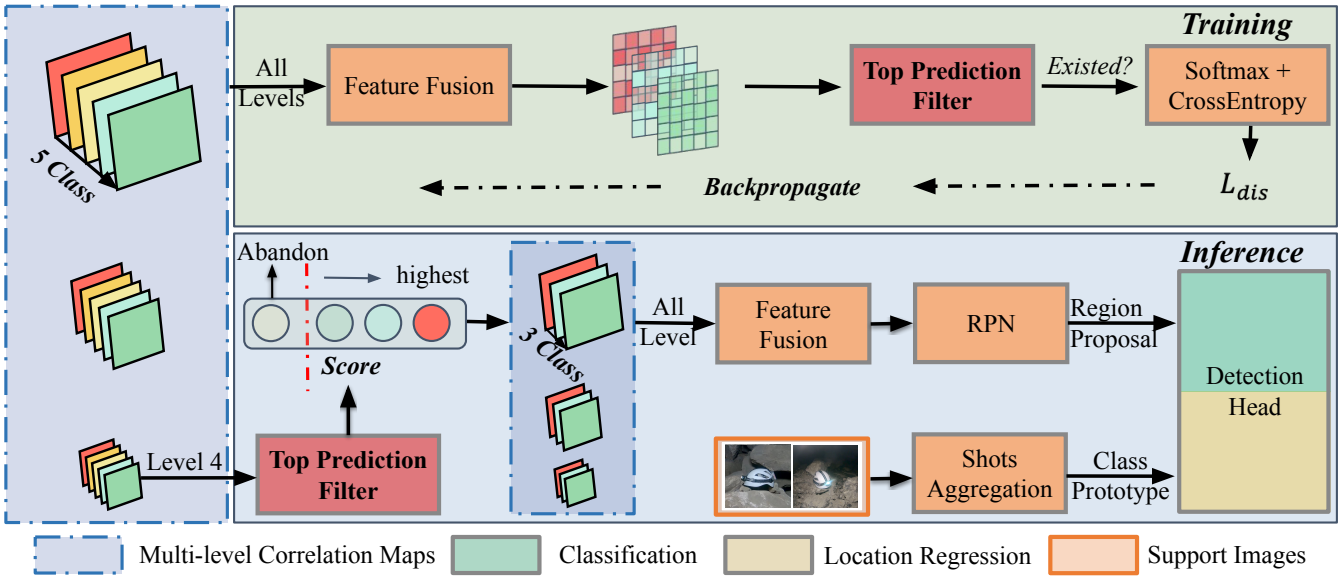

Fig. 3: Detailed working illustration of AirShot in training and inference stage. We adopt the backbone design of AirDet which contains backbone feature extractor, SCS for feature-fusion, relation-based shots aggregation and location regression.

query image. This is particularly relevant in time-consuming modules run in each iteration. The goal of TPF is thus to establish a direct mapping from correlation maps to linearly separable scores or logits. To achieve this, it is desirable to grant the model discrimination ability, which can also guide better correlation map generation during training.

To fulfill the dual functionality, the proposed approach must adhere to the following properties: (1) adaptable to dynamical numbers of novel classes, therefore the module should function as a binary classifier; and (2) limited inherent computational cost, as the module itself needs to run in a full inference loop. The ideal network should strike a good balance between inherent computational cost and performance.

Motivated by the observation that stronger activation of correlation maps corresponds to higher similarity, we propose an approach that effectively leverages the abundant information captured in correlation maps, which directly maps correlation maps to linearly separable scores. To extract the information, we define 2 representations, global representation, and local representation, separately defined as:

1) Global Representation:

$$\mathbf{R_{g,i}} = \mathrm{Avg}\left(\mathrm{ReLU}\left(\mathrm{Norm}\left(\mathbf{c_i}\right)\right)\right), \tag{1}$$

where $\mathbf{c_i}$ is the correlation map of the i_{th} category.

2) Local Representation: Due to the localized nature of the activation strength of correlation map, it is important to extract the local representation to form confidence vector:

$$\mathbf{R_{l,i}} = \mathrm{Avg}\left(\mathrm{MaxPool}\left(\mathbf{c_i}\right)\right). \tag{2}$$

3) Confidence Vector: To better utilize both global and local representation, we simply concatenate two representations to get the fused representation as a confidence vector.

$$\mathbf{V_{con,i}} = \mathrm{Cat}\left(\mathbf{R_{l,i}}, \mathbf{R_{g,i}}\right). \tag{3}$$

Then we pass the confidence vector into a 3-layer MLP (2048, 512, 2) to get the logits for the probability of the existence of the corresponding novel class:

$$\mathbf{p_i} = \mathrm{SoftMax}\left(\mathrm{MLP}\left(\mathbf{V_{con,i}}\right)\right). \tag{4}$$

Similarly, the footnote i above demonstrates the corresponding variables or representations of i_{th} category.

D. AirShot during training

The classification supervision of most prior work [1], [2], [3], [4], [5] relies on the region proposals feature extraction, neglecting the rich information from correlation maps directly. To settle unreliable correlation maps, we assume extra supervision is beneficial. Therefore during training, AirShot utilizes the rich information in correlation maps to make direct inferences about the category's existence. We follow a contrastive strategy to determine the existence. Specifically, we feed the fused correlation map into TPF, then receive the output of the final MLP layer, followed by a cross-entropy loss to apply the contrastive learning strategy.

As for training strategy, we first jointly train all modules in AirShot. During training TPF can act as a discriminator to encourage better correlation map generation. We then frozen the rest of the model and fine-tuned TPF separately. This is to achieve a more robust deterministic effect. Empirically the supervision on the fused correlation map is better than the supervision on level 4. We assume that the supervision on fused one can act across all 3 levels. Whereas in the latter setting, supervision can only function on one level. More details regarding feature level can be found in Section V.

E. AirShot during inference

During inference, we first utilize multi-level query features and support features to generate correlation maps, then feed the deepest correlation maps into TPF to get the logit of the

final MLP layer. It is followed by a Softmax layer to normalize the score. Since we do not have positive/negative pairs during inference, we regard the logits of the single positive neuron as the final score to rank following a certain strategy. Once the category set of top predictions is confirmed, all 3 levels of the category will be passed to the feature fusion module (in our case we adopt SCS from AirDet) to generate fused correlation maps. Then the fused correlation maps will be fed to RPN to generate region proposals. Along with the aggregated class prototype, they will be sent to detection head to infer the classification and location. However, the feature fusion module remains for the selected class sent to the detection head for comparable performance concerns.

Noticed that here we feed the deepest correlation map instead of the fused one. We argue that the fused correlation maps are sub-optimal in our pre-select discrimination process. Moreover, this strategy offers two advantages: (1) a performance enhancement due to improved deterministic ability; and (2) the ability to skip inefficient feature fusion or scale-wise parallel computing among all the novel classes.

Regarding strategy, we provide two types, Top N and adaptive method. "Top N" refers to simply picking the categories with the N highest scores (N normally ranges from 5-10). However, given the fact that the number of novel classes on the query images varies, we also introduce the adaptive method, which sets a threshold and selects the class whose score is beyond the threshold. This strategy successfully eliminates the abundant computational cost of the non-existent novel classes. There is little performance drop due to (1) TPF can capture most of the seen classes; and (2) the necessary feature fusion module is preserved for top predictions. Thus AirDet can achieve efficient inference with comparable performance as adopting full loop inference.

IV. Experiments

We adopt the design from AirDet and use it as the baseline. We mainly adopt ResNet101 [36] pre-trained on ImageNet as the backbone. AirShot and the baseline [1], [8] share the same supports in all the settings. We use 4 NVIDIA A100 for both training and evaluation. We mainly present $k = 1, 2, 3, 5$-shot evaluation since there are limited support samples available during real-world applications. Due to the unseen nature of objects during exploration, we only focus on novel classes throughout the experiments. We first show the details of implantation and our evaluation metrics, then we demonstrate the results of conducting a full loop to validate our baseline enhancement, and finally, we show performance of efficient inference strategy of AirShot applying TPF.

A. Implementation Details

We first jointly train all modules in AirShot: the feature fusion part, RPN, detection head with TPF for 80K iterations from scratch. Once the training is done, we froze other parts and fine-tuned TPF separately on the training set. Note that for different evaluation settings, we apply the same trained model directly, without any further fine-tuning stage.

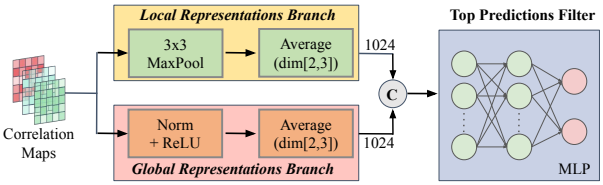

Fig. 4: Network architecture of TPF. We design two branches for global representation and local representation separately. Then the concatenated representation will be fed into a 3-layer MLP.

B. Evaluation Metrics

We defined the strategy going through all classes as *full loop*, and the loop running in our filtered categories selected by TPF as *minor loop*. We compare the performance of in-domain(COCO) and cross-domain (VOC) regarding average precision (AP). To evaluate the performance drop between minor loop utilizing TPF and full loop, we proposed a new metric called **omission rate (OR)**, defined as:

$$OR = -\frac{AP_{\text{full}} - AP_{\text{minor}}}{AP_{\text{full}}} \times 100\%. \tag{5}$$

It is also important to evaluate recall rate of TPF for each class. Thus, we create a class-wise recall bar and show it in Fig. 6, where the footnote n denotes the number of shots.

C. Baseline Enhancement

In this section, we show that utilizing TPF in joint training can achieve better performance. For a fair comparison, all the experiments in this section go through a full inference loop.

1) In-Domain Evaluation: We first present the in-domain evaluation on the COCO benchmark in Table II. Following prior works, we split the 80 classes into 60 non-VOC base classes and 20 novel classes. During training, the available images are the base class images from COCO train2014

TABLE II: In-domain Performance on COCO validation dataset.

	model	AP	AP_{50}	AP_{75}
1 shot	A-RPN	3.32	6.28	3.04
	A-RPN+TPF	**4.53**	**8.05**	**4.55**
	AirDet	5.41	10.19	5.64
	AirShot (Ours)	**6.28**	**11.35**	**6.25**
2 shots	A-RPN	4.10	7.72	3.81
	A-RPN+TPF	**4.81**	**8.70**	**4.89**
	AirDet	5.96	11.22	5.81
	AirShot (Ours)	**7.07**	**12.99**	**6.98**
3 shots	A-RPN	5.47	9.80	5.47
	A-RPN+TPF	**5.91**	**10.53**	**5.88**
	AirDet	6.54	12.26	6.33
	AirShot (Ours)	**8.01**	**14.96**	**7.87**
5 shots	A-RPN	5.87	10.47	5.96
	A-RPN+TPF	**6.35**	**10.93**	**6.28**
	AirDet	7.91	14.61	7.75
	AirShot (Ours)	**8.83**	**15.64**	**8.76**
10 shots	A-RPN	6.01	10.42	6.16
	A-RPN+TPF	**6.49**	**11.54**	**6.64**
	AirDet	8.59	15.15	8.68
	AirShot (Ours)	**9.58**	**16.64**	**9.84**

datasets. Then the trained models are comprehensively evaluated on 5,000 images from the COCO val2014 dataset with few-shot samples per novel class as support images.

2) Cross-Domain Evaluation: Cross-domain performance is crucial for robotic applications as robots are often deployed to novel environments. We adopt the same model trained on COCO as in Section IV-C.1 to evaluate model generalization on the PASCAL VOC dataset as in Table III. Note that the model is directly applied without any fine-tuning process.

TABLE III: Cross-domain performance on VOC-2012 dataset.

	model	AP	AP_{50}	AP_{75}
1 shot	A-RPN	10.03	17.97	10.22
	A-RPN + TPF	**10.73**	**18.96**	**10.41**
	AirDet	10.98	20.36	10.46
	AirShot (Ours)	**11.71**	**22.68**	**11.56**
2 shots	A-RPN	12.98	22.39	13.88
	A-RPN + TPF	**13.29**	**23.92**	**14.15**
	AirDet	13.70	24.84	14.36
	AirShot (Ours)	**15.23**	**27.29**	**15.11**
3 shots	A-RPN	12.60	20.92	13.02
	A-RPN + TPF	**13.87**	**22.96**	**14.09**
	AirDet	15.59	27.27	15.82
	AirShot (Ours)	**16.91**	**29.67**	**17.56**
5 shots	A-RPN	13.19	21.91	13.65
	A-RPN + TPF	**14.48**	**23.96**	**14.57**
	AirDet	16.65	28.67	17.20
	AirShot (Ours)	**18.32**	**31.43**	**18.41**

D. Efficient Inference

In this section, all the experiments run with a minor loop to evaluate the effectiveness of AirShot. We randomly sample the score for each category as the baseline, denoted as *not applying TPF*. We found that the inference time is invariant to the shot settings, so we only report the inference time of different top prediction settings. We first show the result of applying AirShot with the Top 10 strategy in Table IV.

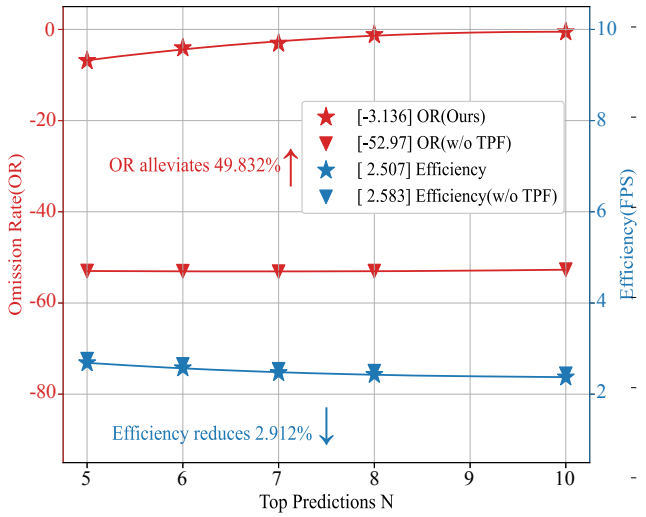

Fig. 5: Ablation of TPF module regarding OR and efficiency (K=3)

TABLE IV: The effect of AirShot applying Top N strategy(N=10)

Dataset	TPF	OR_1	OR_2	OR_3	OR_5	OR_{avg}	T(s)
Base	-	0.00	0.00	0.00	0.00	0.00	0.733
COCO	✗	-55.9	-54.5	-48.7	-51.7	-52.7	0.388
COCO	✓	**-3.12**	**-1.71**	**-1.51**	**-1.50**	**-1.96**	0.392
VOC	✗	-51.07	-49.74	-46.7	-50.9	-49.6	0.382
VOC	✓	**-4.07**	**-3.74**	**-1.95**	**-2.10**	**-2.96**	0.386

We also compare the efficiency without TPF to illustrate the inherent computational cost of TPF is negligible, as in Fig. 5. We use the ground truth category label in COCO to evaluate the recall rate of the TPF module and visualize as Fig. 6. The table shows that AirShot accurately infers the existence among most categories with an average recall rate of nearly 90%. We highlight that any pre-selection methods cannot achieve 100% percent. Since many false predictions are led by low-quality correlation maps, omission caused by this reason will not cause any performance degradation.

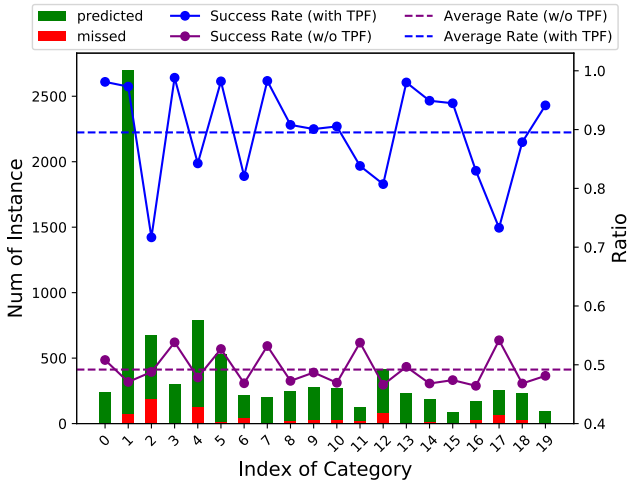

Fig. 6: Recall of TPF module. The **green** bar represents successful prediction by TPF, while the **red** part shows the missed instance. We also report the success rate in class-wise and averaged manner.

E. Inter-Model Comparison

We choose all models that can work without fine-tuning stage for comparison in Fig. 8. We also extend different top prediction settings to show our superior performance. Here we regard N as a hyperparameter where N ranges from 5 to 10. Then we realized that the number of novel classes on one query image varies, thus we empirically set a threshold enabling the adaptive selecting process. Due to unsupported filter characteristics, A-RPN and AirDet run a full inference. AirShot runs a minor inference loop powered by TPF. The result shows that our model significantly beats others in terms of precision and efficiency. One noticeable thing is adaptive method lies beyond the formed line by the Top-N strategy. Although the adaptive method cannot achieve both the best performance and efficiency, it still shows a better balance. In our experiment, the threshold value is set empirically.

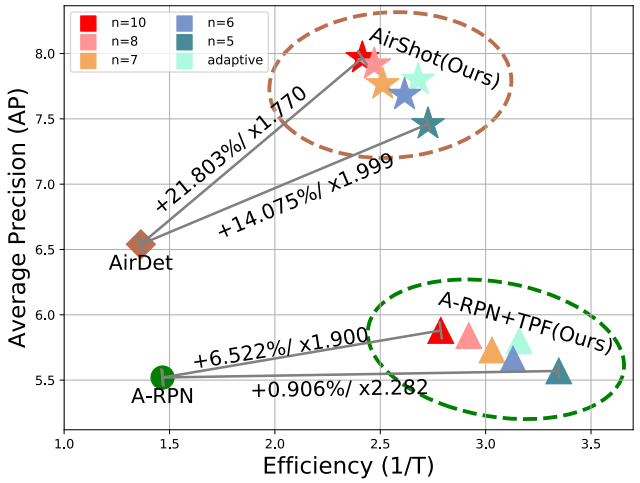

Fig. 7: Inter-model comparison among fine-tuning free models.

F. Dataset Contribution and Real-world Test

We collect and open-source a real dataset from DARPA Subterranean (SubT) challenge [37] The environments poses extra difficulties, e.g., a lack of lighting, dripping water, and cluttered or irregularly shaped environments, etc.

Fig. 8: Preview of our proposed dataset from the DARPA SubT Challenge, showcasing its reality, complexity, and diversity across various scenarios including indoor, outdoor, and cave environments.

To test AirShot in real world, we adopted it on SubT where the robot is equipped with an NVIDIA Jetson AGX Xavier. Dataset preview and qualitative result are presented in supplementary materials. The effectiveness and efficiency of AirShot in the real-world tests demonstrated its promising prospect and feasibility for autonomous exploration.

G. Ablation Study and Deep Visualizations

1) Visualization: We validate the effectiveness of AirShot by showing qualitative visualization with detection results 9.

2) Level of Correlation Maps: The level of correlation maps is crucial during training and inference as they contain different levels of mutual information. Our choices are res2, res3, res4, and the fused correlation maps as in Table V:

Fig. 9: Deep visualization of qualitatively comparison between previous best model **Top:** AirDet (baseline) and **Bottom**: Airshot (ours). AirShot can focus more precisely on the most representative part of the object, resulting in more accurate box regression.

TABLE V: Effect of different level correlation maps tested in coco

method	training	inference	OR
without TPF(K=20)	**fused**	**res4**	**-1.51**
	fused	fused	-2.54
	res4	fused	-6.29
with TPF(K=10)	res4	res4	-4.83
	fused	res3	-26.0
	fused	res2	-58.7

We surprisingly found that the fused correlation map is the sub-optimal during inference stage, the finest feature actually could help the TPF to be more discriminative in determining the logits for existence. Another strength of this property is that it skips the time-consuming feature fusion stage for those classes that are unlikely to appear in the query images, which increases the inference efficiency further.

3) Representations: To illustrate the necessity of our proposed global and local representations, we conduct the ablation study for those two branches as shown in Table VI. The results illustrate that both global and local representations are crucial to a successful mapping building to fully extract information directly from correlation maps.

TABLE VI: Ablation Study of Different Representations

Global	Local	OR	AP	AP50	AP75
✗	✓	-9.75	7.22	13.50	7.11
✓	✗	-6.14	7.51	14.04	7.36
✓	✓	**-1.51**	**8.01**	**14.96**	**7.87**

V. Conclusions & Limitations

We present a novel few-shot object detection system, AirShot for autonomous exploration of mobile robots. Specifically, AirShot fully exploits the correlation map for a more robust and faster FSOD system. It offers a dual functionality that substantially improves the efficiency and effectiveness of most off-the-shelf models. In the experiments, we show that it achieves state-of-the-art performance and efficiency among

current FSOD models that can work without fine-tuning. Additionally, it is proven to be generalized efficiently in various meta-learning-based methods. We expect this method to play an important role in various robotic applications.

However, it also has several limitations: (1) We adopt a relatively simple structure to avoid extra computational cost thus there is still a gap between strictly precise rearranging score order; (2) We use the same model structure in the training stage, while other complicated models can perform better; and (3) When the number of novel classes is close to validation set, the enhancement would be limited.

REFERENCES

[1] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. 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 (CVPR)*, 2020, pp. 4013–4022.

[2] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "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)*, 2021, pp. 10 185–10 194.

[3] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-Shot Object Detection via Feature Reweighting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8420–8429.

[4] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-Learning to Detect Rare Objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9925–9934.

[5] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," *arXiv preprint arXiv:2003.06957*, 2020.

[6] C. Wang, Y. Qiu, W. Wang, Y. Hu, S. Kim, and S. Scherer, "Unsupervised Online Learning for Robotic Interestingness with Visual Memory," *IEEE Transactions on Robotics*, pp. 1–15, 2021.

[7] C. Wang, W. Wang, Y. Qiu, Y. Hu, and S. Scherer, "Visual Memorability for Robotic Interestingness via Unsupervised Online Learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 52–68.

[8] B. Li, C. Wang, P. Reddy, S. Kim, and S. Scherer, "Airdet: Few-shot detection without fine-tuning for autonomous exploration," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*. Springer, 2022, pp. 427–444.

[9] S. Kim, C. Wang, B. Li, and S. Scherer, "Robotic Interestingness via Human-Informed Few-Shot Object Detection," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 1756–1763.

[10] Y. Li, H. Zhu, Y. Cheng, W. Wang, C. S. Teo, C. Xiang, P. Vadakkepat, and T. H. Lee, "Few-shot object detection via classification refinement and distractor retreatment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 395–15 403.

[11] Z. Yang, C. Zhang, R. Li, Y. Xu, and G. Lin, "Efficient few-shot object detection via knowledge inheritance," *IEEE Transactions on Image Processing*, vol. 32, pp. 321–334, 2022.

[12] Z. Fan, Y. Ma, Z. Li, and J. Sun, "Generalized few-shot object detection without forgetting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4527–4536.

[13] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "Defrcn: Decoupled faster r-cnn for few-shot object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8681–8690.

[14] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7352–7362.

[15] J. Wu, S. Liu, D. Huang, and Y. Wang, "Multi-Scale Positive Sample Refinement for Few-Shot Object Detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 456–472.

[16] L. Zhang, S. Zhou, J. Guan, and J. Zhang, "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)*, 2021, pp. 14 424–14 432.

[17] W. Zhang, Y.-X. Wang, and D. A. Forsyth, "Cooperating rpn's improve few-shot object detection," *arXiv preprint arXiv:2011.10142*, 2020.

[18] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[19] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot Multibox Detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.

[22] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[25] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[26] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.

[27] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9577–9586.

[28] G. Han, Y. He, S. Huang, J. Ma, and S.-F. Chang, "Query adaptive few-shot object detection with heterogeneous graph convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3263–3272.

[29] C. Zhu, F. Chen, U. Ahmed, Z. Shen, and M. Savvides, "Semantic Relation Reasoning for Shot-Stable Few-Shot Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8782–8791.

[30] A. Wu, Y. Han, L. Zhu, and Y. Yang, "Universal-prototype enhancing for few-shot object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9567–9576.

[31] G. Zhang, Z. Luo, K. Cui, S. Lu, and E. P. Xing, "Meta-detr: Image-level few-shot detection with inter-class correlation exploitation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[32] G. Han, J. Ma, S. Huang, L. Chen, and S.-F. Chang, "Few-shot object detection with fully cross-transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5321–5330.

[33] T. Kong, A. Yao, Y. Chen, and F. Sun, "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.

[34] Z. Li and F. Zhou, "FSSD: Feature Fusion Single Shot Multibox DAetector," *arXiv preprint arXiv:1712.00960*, 2017.

[35] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning Deeply Supervised Object Detectors From Scratch," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] https://subtchallenge.com.