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

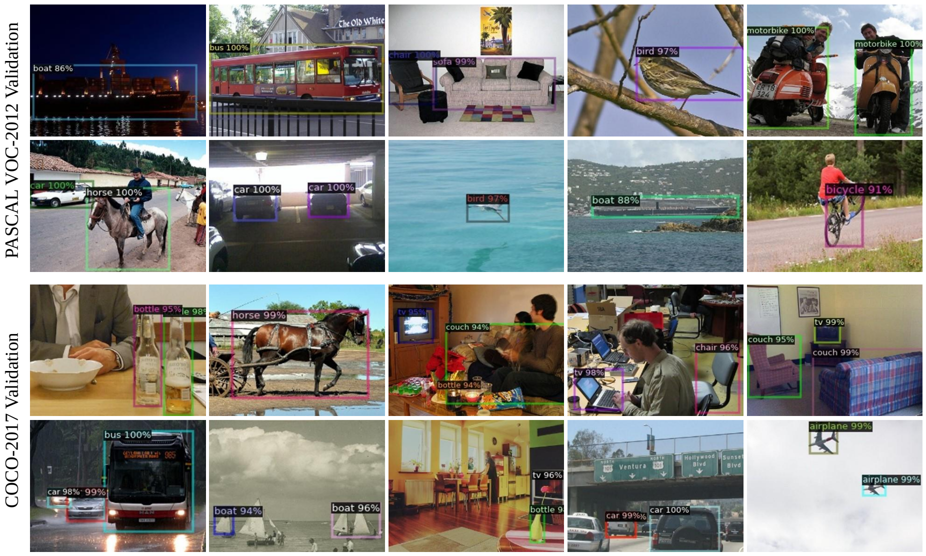

Fig. 1. Representative examples of 3-shot detection of AirDet on VOC-2012 validation and COCO validation dataset. Without fine-tuning, AirDet can robustly detect the novel unseen objects such as boat, bus, sofa, with merely three support images.

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.

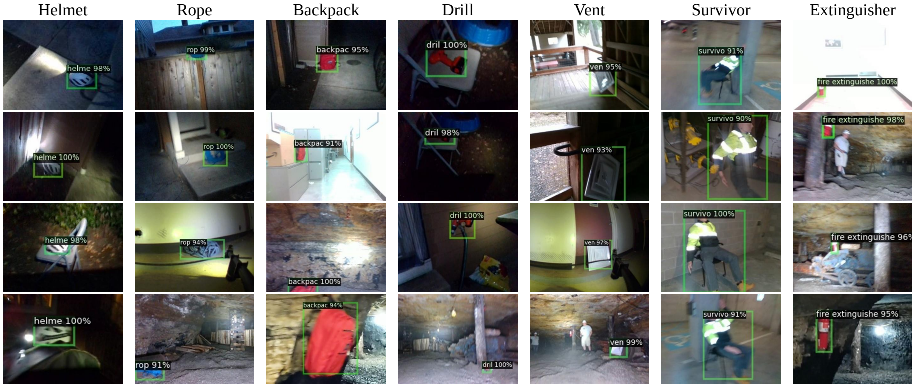

Fig. 2. Representative examples of 3-shot detection of AirDet on the DARPA Subterranean Challenge. Provided with merely three support images for these unseen novel objects, AirDet can directly detect them with scale variation and partial occlusion in distinct environments and illumination conditions.

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

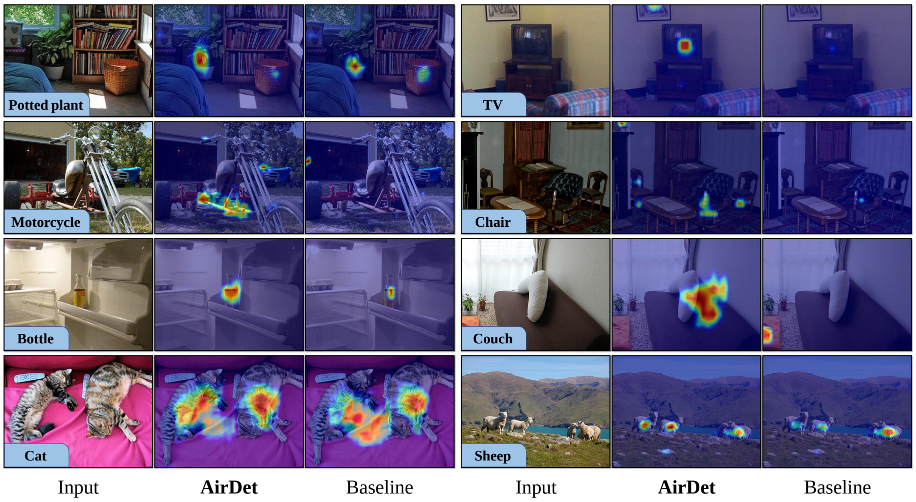

Fig. 3. Deep visualization from the proposal generation module of AirDet and baseline [6]. Compared with baseline, AirDet can better notice and concentrate on and novel object region, which leads to its more effective region proposals.

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-

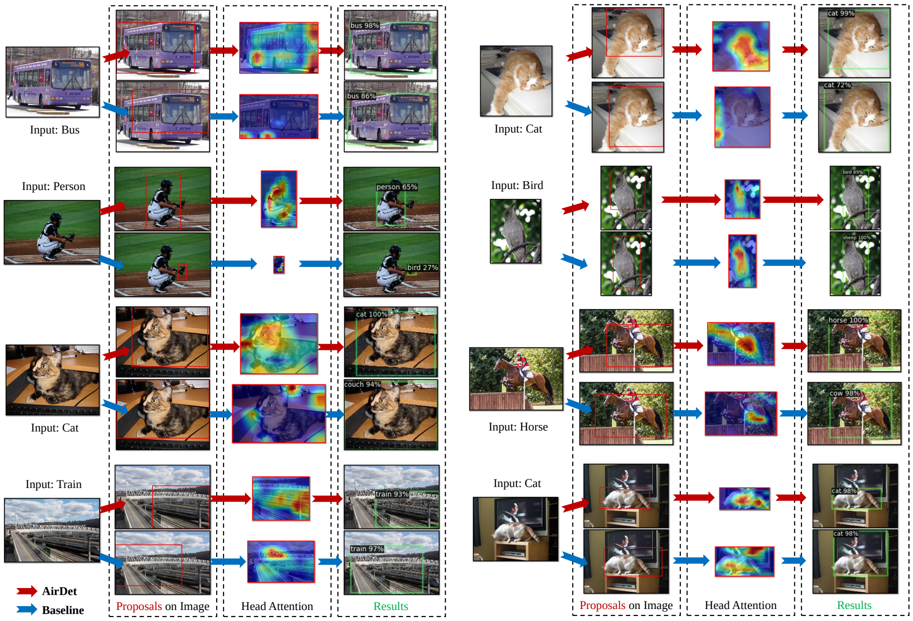

Fig. 4. Deep visualization from the detection head of AirDet and baseline [6]. With similar proposals in red boxes, AirDet can focus more precisely on the most representative part of the object, resulting in more accurate box regression and classification.

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.