SNIDA: Unlocking Few-Shot Object Detection with Non-linear Semantic Decoupling Augmentation
Supplementary Materials

Yanjie Wang[1,2], Xu Zou[1,2,*], Luxin Yan[1,2], Sheng Zhong[1,2], Jiahuan Zhou [3]
[1]Huazhong University of Science and Technology, Wuhan 430074, China
[2]National Key Laboratory of Multispectral Information Intelligent Processing Technology, China
[3]Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China
{aiawyj, zoux, yanluxin, zhongsheng}@hust.edu.cn, jiahuanzhou@pku.edu.cn

The supplementary information is summarised as follows. First, we provide more implementation details of the model and show more visual examples of generated sets. Second, we perform additional experiments to validate the effectiveness of our method. Third, we present a detailed analysis of more qualitative detection results.

1. Implementation Details

Our approach follows the finetuning-based methods of FSOD, which can be divided into two stages. In the first stage, the model is trained with adequate base class annotations. In the second stage, finetuning the novel class with scarce samples to make the whole detector more general.

(1) Base Class Training Stage. In the first stage, the base class with adequate annotations is adopted to train the model, so that the detector is provided with the information that can be transferred to the novel class. We follow the baseline [5, 7] for model training. Thus, the loss $\mathcal{L}_{Training}$ of the base training model is defined as:

$$\mathcal{L}_{Training} = \mathcal{L}_{RPN} + \mathcal{L}_{RCNN} \tag{1}$$

where $\mathcal{L}_{RPN}$ represents the loss of RPN, which generates class-agnostic candidate proposals, and $\mathcal{L}_{RCNN}$ indicates losses of the class-relevant detection result.

(2) Novel Class Finetuning Stage. In this stage, to decouple the foreground/background of few-shot instances, we add a split branch (Mask Head) to our baselines to learn the mask of C_{novel} samples. For the mask head, we added a full convolution mask prediction branch to SAES following the architecture in the previous work [3]. Thus, the loss $\mathcal{L}_{Finetuning}$ of the novel finetuning model is defined as:

$$\mathcal{L}_{Finetuning} = \mathcal{L}_{RPN} + \mathcal{L}_{RCNN} + \mathcal{L}_{Con} \tag{2}$$

where $\mathcal{L}_{Con}$ indicates losses of only novel class mask results. As there are no semantic masks available for few-shot objects, we need to rely on unsupervised semantic segmentation methods. The unsupervised saliency detection [4] is employed to learn the mask regions of novel class objects and apply it as a prior. Pixel-level contrastive loss $\mathcal{L}_{Con}$ is used for the representation of foreground/background. The optimization process of this loss is designed to minimize the distance between embedding vectors belonging to the same classes while pushing foreground and background apart. In this way, a pixel embedding space is introduced as a dense semantic representation to get the semantic segmentation mask of the novel class instances. We adopt the same enhancement set as MaskContrast [6] to generate positive (X, X^+) while ensuring that each image contains at least part of the salient objects ($> 10\%$ of the area). The features of negatives are saved in a memory bank. The masking branch is only applied to finetuning, and it will be removed during inference.

More visualizations of different sets are generated using the base set and novel set of PASCAL VOC in Figure 1. The synthetic set and saliency set are generated by SAES, which augment the number of samples. The diversity would be improved by fusing instances into different backgrounds. The reconstructed set constructed by OREM further implements semantic-guided non-linear enhancement.

2. Additional Expreiments

2.1. Experimental Results of G-FSOD Setting

The generalized few-shot object detection (G-FSOD) aims to detect novel class samples without forgetting previously seen base classes. In this section, to validate the effectiveness of our method in the G-FSOD setting, we present the results of MS COCO following [7].

As shown in Table 1, our report includes metrics such

*Corresponding author.

# shots	Method	Overall *#80*			Base *#60*	Novel *#20*
		AP	AP50	AP75	AP	AP
1	FRCN+ft [8]	16.2±0.9	25.8±1.2	17.6±1.0	21.0±1.2	1.7±0.2
	TFA [7]	**24.4±0.6**	**39.8±0.8**	26.1±0.8	**31.9±0.7**	1.9±0.4
	DeFRCN [5]	24.0±0.4	36.9±0.6	**26.2±0.4**	30.4±0.4	4.8±0.6
	SNIDA	23.8±0.5 (-0.2)	35.2±0.4 (-0.4)	25.8±0.5 (-0.4)	29.4±0.3 (-1.0)	**6.9±0.5** (+2.1)
2	FRCN+ft [8]	15.8±0.7	25.0±1.1	17.3±0.7	20.0±0.9	3.1±0.3
	TFA [7]	24.9±0.6	**40.1±0.9**	27.0±0.7	**31.9±0.7**	3.9±0.4
	DeFRCN [5]	25.7±0.5	39.6±0.8	28.0±0.5	31.4±0.4	8.5±0.8
	SNIDA	**26.1±0.6** (+0.4)	40.2±0.7 (+0.6)	**28.5±0.4** (+0.5)	31.2±0.3 (-0.2)	**10.9±0.3** (+2.4)
3	FRCN+ft [8]	15.0±0.7	23.9±1.2	16.4±0.7	18.8±0.9	3.7±0.4
	TFA [7]	25.3±0.6	40.4±1.0	27.6±0.7	32.0±0.7	5.1±0.6
	DeFRCN [5]	26.6±0.4	41.1±0.7	28.9±0.4	32.1±0.3	10.7±0.8
	SNIDA	**27.8±0.4** (+1.2)	**42.8±0.6** (+1.7)	**29.4±0.5** (+0.5)	**32.8±0.4** (+0.7)	**12.8±0.4** (+2.1)
5	FRCN+ft [8]	14.4±0.8	23.0±1.3	15.6±0.8	17.6±0.9	4.6±0.5
	TFA [7]	25.9±0.6	41.2±0.9	28.4±0.6	32.3±0.6	7.0±0.7
	DeFRCN [5]	27.8±0.3	43.0±0.6	30.2±0.3	**32.6±0.3**	13.6±0.7
	SNIDA	**28.7±0.4** (+1.9)	**43.7±0.9** (+1.8)	**30.6±0.7** (+1.8)	**32.6±0.4** (0.0)	**14.6±0.6** (+1.0)
10	FRCN+ft [8]	13.4±1.0	21.8±1.7	14.5±0.9	16.1±1.0	5.5±0.9
	TFA [7]	26.6±0.5	42.2±0.8	29.0±0.6	32.4±0.6	9.1±0.5
	DeFRCN [5]	29.7±0.2	46.0±0.5	32.1±0.2	34.0±0.2	16.8±0.6
	SNIDA	**30.2±0.2** (+0.5)	**46.7±0.5** (+0.7)	**32.9±0.2** (+0.8)	**34.3±0.7** (+0.3)	**17.9±0.6** (+1.1)
30	FRCN+ft [8]	13.5±1.0	21.8±1.9	14.5±1.0	15.6±1.0	7.4±1.1
	TFA [7]	28.7±0.4	44.7±0.7	31.5±0.4	34.2±0.4	12.1±0.4
	DeFRCN [5]	31.4±0.1	48.8±0.2	33.9±0.1	34.8±0.1	21.2±0.4
	SNIDA	**32.0±0.6** (+0.6)	**49.1±0.7** (+0.3)	**34.9±0.1** (+1.0)	**35.1±0.5** (+0.3)	**22.7±0.7** (+1.5)

Table 1. Generalized few-shot object detection (*G-FSOD*) performance on COCO dataset. For each metric, we report the average and 95% confidence interval computed over 10 random samples. All comparison results refer from [7].

Method / Shots	Novel Set 1					Novel Set 2					Novel Set 3					Mean
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
Baseline	53.6	57.5	61.5	**64.1**	60.8	30.1	38.1	47.0	53.3	47.9	48.4	50.9	52.3	54.9	57.4	51.9
CutOut [2]	54.7	57.2	**62.3**	64.0	62.5	31.8	39.0	**47.8**	53.5	48.2	49.0	50.7	53.2	55.7	58.3	52.5
GridMask [1]	54.3	**58.0**	62.0	63.7	63.2	**32.0**	**39.1**	47.6	53.2	48.5	**49.3**	**51.2**	**54.6**	55.5	58.5	52.7
CutMix [9]	**55.5**	57.6	61.4	63.9	**63.5**	**32.0**	38.9	47.4	**53.6**	**48.9**	48.8	49.9	53.4	**56.7**	**59.8**	**52.8**
SNIDA	**59.3**	**60.8**	**64.3**	**65.4**	**65.6**	**35.2**	**40.8**	**50.2**	**54.6**	**50.0**	**51.6**	**52.4**	**55.9**	**58.5**	**62.6**	**55.1**

Table 2. The results of Cutout, GridMask, CutMix, and our method on three different splits of PASCAL VOC. **RED**/ **BLUE** indicate the best and the second best performance. Our method significantly improves the performance under different shots and achieves competitive performance compared with other augmentation methods for 3 novel splits.

Masking Ratio	Novel Set 1					Novel Set 2					Novel Set 3					Mean
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
0%	55.5	57.6	61.4	63.9	63.5	32.0	38.9	47.4	53.6	48.9	48.8	49.9	53.4	56.7	59.8	52.8
30%	58.4	59.3	63.0	64.3	64.2	34.1	39.5	49.5	53.9	49.5	50.7	51.3	54.9	57.8	61.8	54.1
50%	58.1	60.2	63.8	64.6	65.0	34.6	39.8	49.8	54.2	49.4	51.3	51.9	55.4	58.2	62.1	54.6
70%	**59.3**	**60.8**	**64.3**	**65.4**	**65.6**	**35.2**	**40.8**	**50.2**	**54.6**	**50.0**	**51.6**	**52.4**	**55.9**	**58.5**	**62.6**	**55.1**

Table 3. Ablation study of different masking ratios in OREM on different splits of PASCAL VOC. Note the masking ratio here is not for all patches of the entire image, but only for the foreground. The BEST performances are high-lighted in **bold**.

Method / Shots	Novel Set 1					Novel Set 2					Novel Set 3					Mean
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
Baseline	57.3	59.9	63.7	64.1	63.7	33.7	39.3	47.1	54.4	49.6	49.4	51.3	55.2	58.0	61.0	53.9
Random	55.9	57.8	62.3	62.5	62.4	33.0	38.1	46.2	53.2	58.5	48.3	49.6	54.1	56.7	59.8	53.0
Misleading	56.5	58.2	63.1	63.0	63.1	33.3	38.3	46.7	53.5	59.1	48.8	49.9	54.7	57.1	60.5	53.5
Consistency	**59.3**	**60.8**	**64.3**	**65.4**	**65.6**	**35.2**	**40.8**	**50.2**	**54.6**	**50.0**	**51.6**	**52.4**	**55.9**	**58.5**	**62.6**	**55.1**

Table 4. Different semantic supervision in Semantic Discriminator of OREM on for 3 novel splits of PASCAL VOC.

as AP, AP50, and AP75 for all classes. Additionally, we provide AP scores specifically for base and novel classes. To augment our data and enhance the model's exposure to novel classes, we paste novel class instances in base class images for data augmentation, reducing the training opportunities of the base class and thus damaging the performance of the base class with only 1-2 shots.

However, for 3-10 shots, the base class performance has been effectively improved, which can be attributed to different degrees of occlusion when pasting novel class instances. It increases the diversity of the base class samples and the robustness of the network to the base class. For AP of novel class, our method also outperforms others to achieve a new state-of-the-art result, which intuitively indicates the effectiveness of our idea.

2.2. Comparison among Augmentation Methods

Table 2 shows the results of CutOut [2], GridMask [1], CutMix [9], and our method on different splits of PASCAL VOC. We adopt DeFRCN [5] as the baseline. Our approach outperforms other augmentation methods and excels across various splits. CutOut and GridMask remove contiguous regions or disconnected pixel sets of input images and encourage the network to use the image context. In contrast, our method attains semantic awareness through foreground/background separation and fusion. The incorporation of semantic-guided non-linear transformations further boosts instance diversity in our approach.

2.3. Ablation Study on All Novel Splits of VOC

Masking Ratio of OREM. We analyzed the impact of varying masking ratios ($k\%$) in OREM across different splits of PASCAL VOC, where random patches are masked on input instances. In Table 3, we applied different $k\%$ values in experiments with various splits. The best OREM performance occurs at a masking ratio of 70%. A lower ratio results in less diverse images, affecting model generalization. Conversely, a higher ratio performs better performance, thanks to accurate high-level semantic supervision from the semantic discriminator loss.

Different Semantic Supervision. To validate the impact of the different semantic supervision, we conducted experiments within the semantic discriminator of OREM, involving three types of semantic supervision. Due to incorrect semantic guidance, the results indicate that both the *Random* and *Misleading* supervision resulted in a slight drop in performance compared to the baseline. While the *Consistency* supervision significantly improved overall performance. Accurate high-level guidance preserves semantic consistency in reconstructed images, while rich non-linear semantic knowledge enhances quality and expressiveness, especially with high masking ratios.

3. More Detection Results

As shown in Figure 2, some qualitative results of the proposed 10-shot setting method are illustrated in the Novel Split 1 of PASCAL VOC. The confidence threshold is set to 0.3 when visualizing results. The first four lines show the successful results and the second two lines show the failure scenarios. In the first four lines, our method can accurately locate two objects occluding each other, which indicates that the decoupling of foreground and background is critical for effective learning of objective semantic discriminative features. In the second two lines, we show some failure cases to analyze the cause of the detection error.

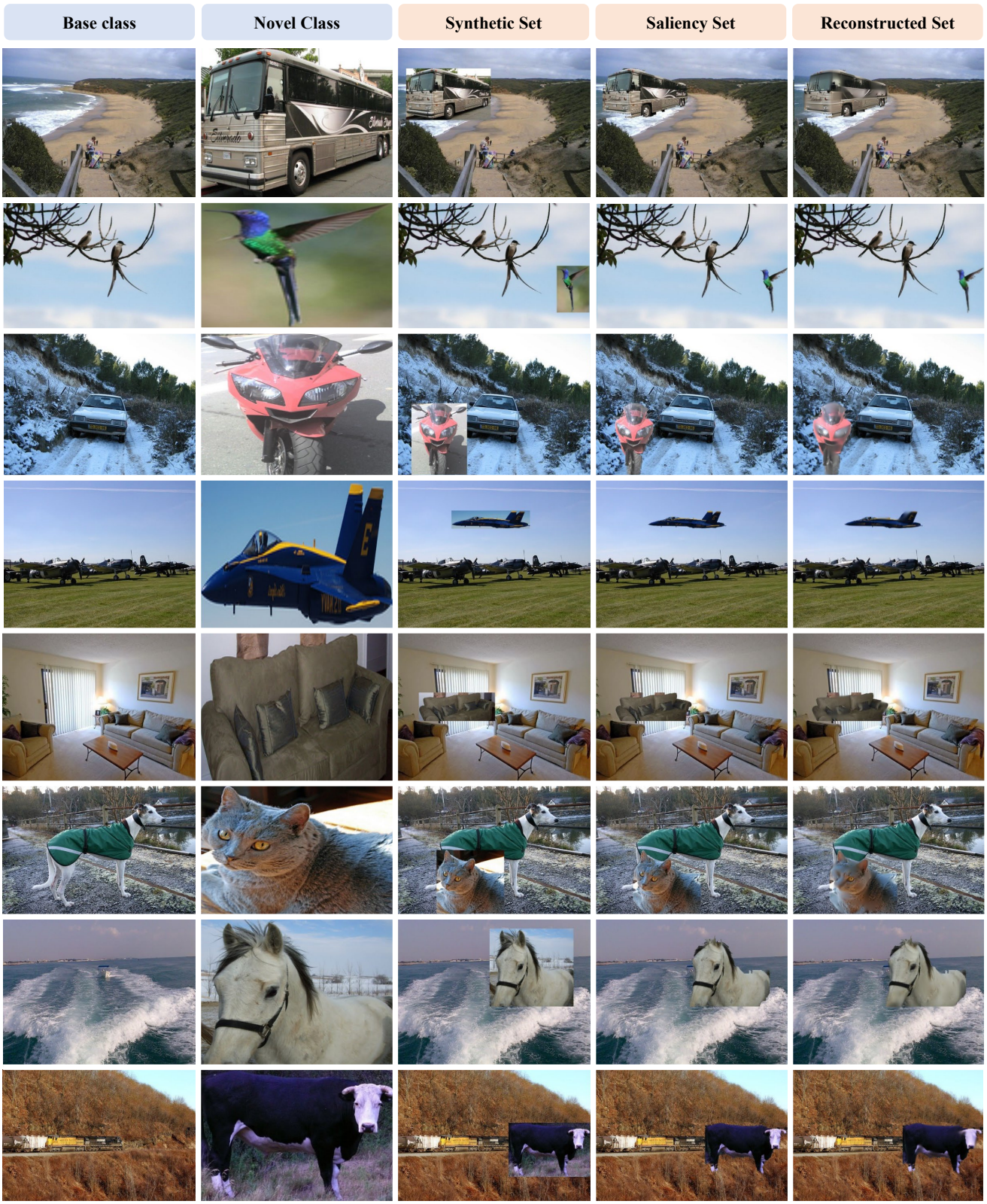

Figure 1. Visualizations of different sets generated using the base set and novel set of PASCAL VOC. The synthetic set and saliency set are generated by SAES, which augment the number of samples. The diversity would be improved by fusing instances into different backgrounds. MAE set constructed by OREM further implements instance-level non-linear enhancement. Best viewed with zoom-in.

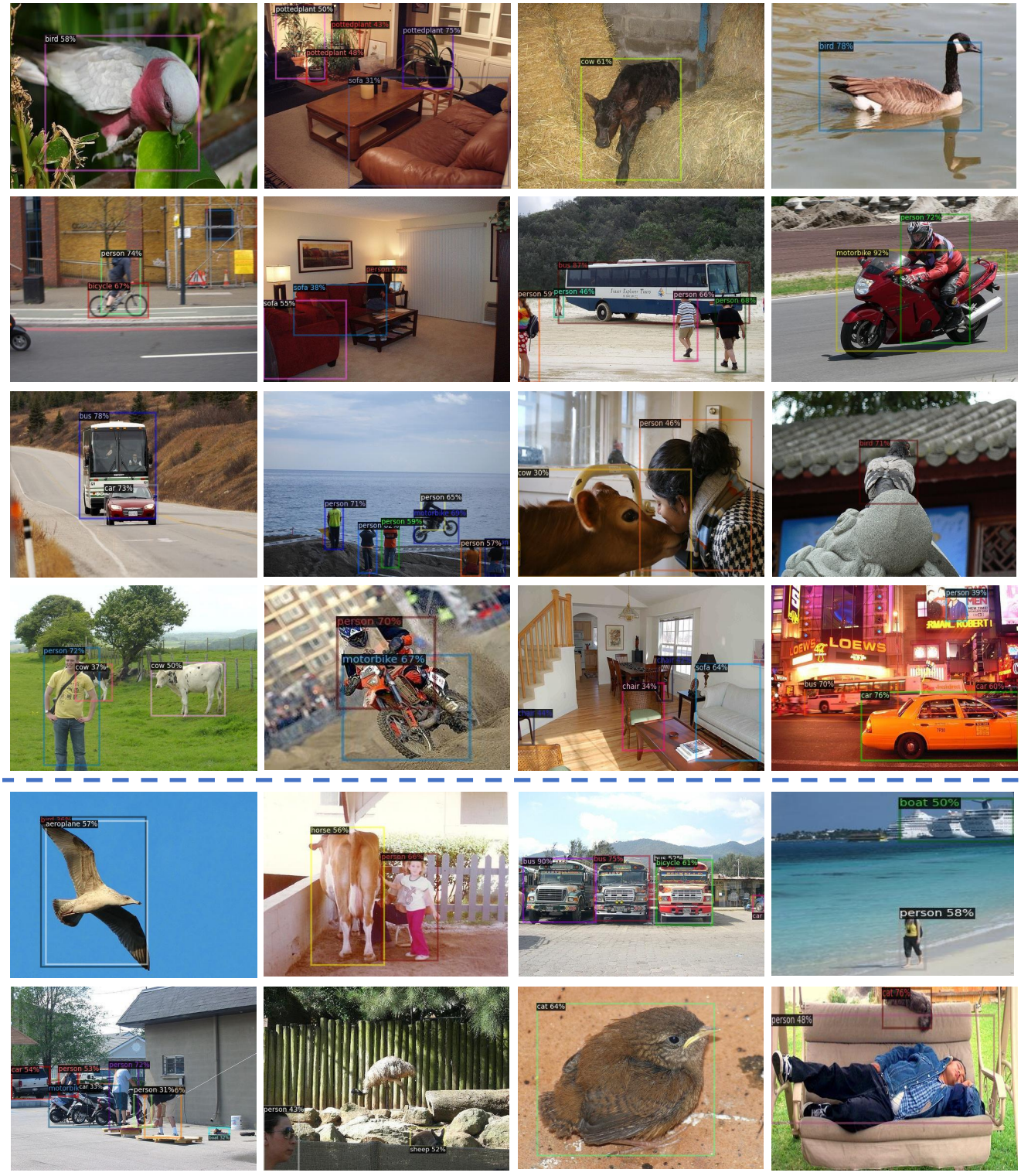

Figure 2. Qualitative detection results of our 10-shot detection method under the Novel Split 1 of PASCAL VOC. The confidence threshold is set to 0.3 when visualizing results. The first four lines show the successful results and the second two lines show the failed scenarios.

References

[1] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 2, 3

[2] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 3

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[4] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In *NeurIPS*, 2019. 1

[5] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *ICCV*, 2021. 1, 2, 3

[6] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, 2021. 1

[7] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, 2020. 1, 2

[8] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *ICCV*, 2019. 2

[9] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 2, 3