Supplementary: When Pixel Difference Patterns Meet ViT: PiDiViT for Few-Shot Object Detection

Hongliang Zhou Yongxiang Liu* Canyu Mo Weijie Li* Bowen Peng Li Liu*

The College of Electronic Science and Technology, National University of Defense Technology, China

honglianqz2022@163.com lyx_bible@sina.com mocanyu@nudt.edu.cn

lwj2150508321@sina.com pbow16@nudt.edu.cn liuli_nudt@nudt.edu.cn

Table 1. nAP50 results on Pascal VOC few-shot benchmark. Results surpassing SOTA are indicated in **bold**. Three Novel Split types total.

Method	Backbone	Novel Split 1					Novel Split 2					Novel Split 3					Avg
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
TFA [10]	RN101	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8	39.9
Multi-Relation Det [4]	RN50	37.8	43.6	51.6	56.5	58.6	22.5	30.6	40.7	43.1	47.6	31.0	37.9	43.7	51.3	49.8	43.1
Retentive RCNN [5]	RN101	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1	41.1
Meta Faster R-CNN [11]	RN101	43.0	54.5	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6	50.5
LVC [8]	ViT-S/8	54.5	53.2	58.8	63.2	65.7	32.8	29.2	50.7	49.8	50.6	48.4	52.7	55.0	59.6	59.6	52.3
CrossTransformer [7]	PVTv2	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7	50.9
HeteroGraph [6]	RN101	42.4	51.9	55.7	62.6	63.4	25.9	37.8	46.6	48.9	51.1	35.2	42.9	47.8	54.8	53.5	48.0
DiGeo [9]	RN101	37.9	39.4	48.5	58.6	61.5	26.6	28.9	41.9	42.1	49.1	30.4	40.1	46.9	52.7	54.7	44.0
NIFF [1]	RN101	62.8	67.2	68.0	70.3	68.8	38.4	42.9	54.0	56.4	54.0	56.4	62.1	61.2	64.1	63.9	59.4
DE-ViT [3]	ViT-L/14	55.4	56.1	68.1	70.9	71.9	43.0	39.3	58.1	**61.6**	**63.1**	58.2	64.0	61.3	**64.2**	67.3	60.2
Ours	ViT-L/14	**57.3**	**56.9**	**68.1**	**73.7**	**73.1**	**43.5**	**44.7**	**61.2**	61.2	62.4	**58.4**	**64.2**	**61.4**	64.1	**67.6**	**61.2**

As shown in Table 1, our method achieves a new SOTA performance on the Pascal VOC few-shot detection benchmark. Specifically, it improves the Avg nAP50 by 1 percentage point compared to the baseline method DE-ViT [3] and outperforms the advanced method NIFF [1] by 1.8 percentage points. Different from DE-ViT [3], we design prior modules targeting the pixel-level feature differences and multi-scale variations in the low-level features of pre-trained ViT: The DCFM achieves differential enhancement of smooth features from the center to the boundary while retaining global information. Meanwhile, the MFFM effectively captures both local details and global contours across multiple scales.

As shown in Table 2, we evaluated the inference time of our method on the COCO few-shot detection benchmark. Our approach outperforms the CNN-based Meta Faster R-CNN [2] while incurring only a marginal increase in inference time (0.22 Secs/Img). Compared to the baseline DE-ViT [3], our method achieves a 4.7-point improvement in AP (10 shots) with only a 0.1 Secs/Img increase in inference time. These results demonstrate that our method maintains efficient inference capabilities while achieving better performance.

Table 2. Inference time comparison on COCO few-shot setting.

Method	Backbone	nAP50 (↑)	Secs/Img (↓)
CrossTransformer [7]	Custom	30.2	3.00
Meta Faster R-CNN [2]	RN101	25.7	**0.61**
LVC [8]	Swin-S	34.1	**-**
DE-ViT [3]	ViT-L/14	52.9	0.83
Ours	ViT-L/14	**57.6**	0.93

References

[1] Guirguis K et al. Niff: Alleviating forgetting in generalized few-shot object detection via neural instance feature forging. In *CVPR*, pages 24193–24202, 2023. 1

[2] Han G et al. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *AAAI*, pages 780–789, 2022. 1

[3] Zhang X et al. Detect every thing with few examples. *arXiv preprint arXiv:2309.12969*, 2023. 1

[4] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, pages 4013–4022, 2020. 1

[5] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *CVPR*, pages 4527–4536, 2021. 1

[6] Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, and Shih-Fu Chang. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In *ICCV*, pages 3263–3272, 2021. 1

[7] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. In *CVPR*, pages 5321–5330, 2022. 1

[8] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Label, verify, correct: A simple few shot object detection method. In *CVPR*, pages 14237–14247, 2022. 1

[9] Jiawei Ma, Yulei Niu, Jincheng Xu, Shiyuan Huang, Guangxing Han, and Shih-Fu Chang. Digeo: Discriminative geometry-aware learning for generalized few-shot object detection. In *CVPR*, pages 3208–3218, 2023. 1

[10] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, pages 9919–9928, 2020. 1

[11] 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*, pages 9577–9586, 2019. 1