

Analyzing the Impact of Low-Rank Adaptation for Cross-Domain Few-Shot Object Detection in Aerial Images

Hicham TALAOUBRID

L2TI & COSE

Université Sorbonne Paris Nord

hicham.talaoubrid1@edu.univ-paris13.fr

Anissa MOKRAOUI

L2TI

Université Sorbonne Paris Nord

anissa.mokraoui@univ-paris13.fr

Ismail BEN AYED

LIVIA, ETS

Montreal, Canada

Ismail.BenAyed@etsmtl.ca

Axel PROUVOST

IMT Mines Alès, France

axel.prouvost@etu.mines-ales.fr

Sonimith HANG

IMT Mines Alès, France

sonimith.hang@etu.mines-ales.fr

Monit KORN

IMT Mines Alès, France

monit.korn@etu.mines-ales.fr

Rémi HARVEY

COSE, Montmagny, France

remi.harvey@cose.fr

Abstract—This paper investigates the application of Low-Rank Adaptation (LoRA) to small models for cross-domain few-shot object detection in aerial images. Originally designed for large-scale models, LoRA helps mitigate overfitting, making it a promising approach for resource-constrained settings. We integrate LoRA into DiffusionDet, and evaluate its performance on the DOTA and DIOR datasets. Our results show that LoRA applied after an initial fine-tuning slightly improves performance in low-shot settings (e.g., 1-shot and 5-shot), while full fine-tuning remains more effective in higher-shot configurations. These findings highlight LoRA’s potential for efficient adaptation in aerial object detection, encouraging further research into parameter-efficient fine-tuning strategies for few-shot learning. Our code is available here: <https://github.com/HichTala/LoRA-DiffusionDet>

Index Terms—Object Detection, Few-Shot Object Detection, Diffusion Models, Cross-Domain, Aerial Images, Low-rank Adaptation.

I. INTRODUCTION

The last few years have seen remarkable improvement in large models. Particularly in natural language processing and computer vision [1] [2]. Parameter-efficient fine-tuning methods have been developed to train these very large models for simpler tasks, without the need to train tens of billions of parameters. One of these is Low Rank Adaptation (LoRA), which, by injecting low rank matrices while freezing the model’s pre-training weights, considerably reduces the number of parameters to be trained in the model. In this way, LoRA helps to limit the overfitting of large models by accelerating their convergence. But its potential in models 100 to 1000 times smaller in terms of parameters remains rather unexplored, particularly in a context of cross domain few-shot object detection, where the overfitting remains the main difficulty.

Few-shot Object Detection (FSOD) is a challenging task that aims to detect objects from novel categories using only a few labeled examples. When applied across

domains, the problem becomes even more complex due to significant distribution shifts between source and target domains. Traditional fine-tuning approaches often struggle in such scenarios as they tend to overfit to the limited training data, especially in smaller models with fewer parameters. This challenge is even greater in cross-domain contexts, where the model needs to generalize to new domains with a minimum of supervision.

Aerial images present an additional complexity for object detection. These images often contain numerous small objects densely distributed over the scene, as well as significant variations in scale between classes, orientation, and illumination [3]. These characteristics make aerial images a particularly demanding domain for cross-domain few-shot object-detection. The main problem remains overfitting in such a scenario. In this work, we investigate the application of LoRA to small models, using the COCO [4] dataset as the source domain and the DOTA [5] and DIOR [6] datasets as target domains. Both datasets are widely recognized references for aerial image analysis [7] [8].

The motivation behind this work is to address overfitting in cross-domain few-shot object detection, particularly for aerial images. We explore Low-Rank Adaptation (LoRA) [9], a technique designed for large models, to improve generalization in smaller architectures like DiffusionDet [10], which has about millions of parameters. DiffusionDet has shown effectiveness in detecting small objects, making it a good choice for aerial images. We compare DiffusionDet with and without LoRA, testing two strategies: (1) direct LoRA application and (2) LoRA after intermediate fine-tuning. Using the DOTA and DIOR datasets, we evaluated LoRA’s ability to reduce overfitting and improve generalization, offering insight into its potential for efficient and robust aerial object detection.

II. RELATED WORK

A. Low-Rank Adaptation (LoRA) for Efficient Fine-Tuning

LoRA [9] is a parameter-efficient fine-tuning method that adapts pre-trained models to new tasks by injecting low-rank decomposition matrices into existing weight layers. Instead of updating all parameters, LoRA introduces small trainable matrices into transformer layers, significantly reducing memory usage and computational costs while preserving the original model weights. Initially developed for natural language processing (NLP), LoRA has been extended to computer vision, enabling efficient adaptation of large-scale models to new domains with limited labeled data.

In the context of aerial images, LoRA has been applied to vision transformers trained on the DOTA [5] dataset, demonstrating its effectiveness in transfer learning for tasks such as object detection [11]. These studies highlight LoRA's ability to adapt large-scale models, such as ViT and Swin Transformer, to new aerial datasets without requiring full fine-tuning. Beyond aerial images, LoRA has also been successfully applied to Vision-Language Models (VLMs) like CLIP [2] and GLIP [12], enabling task-specific adaptation without modifying the backbone network.

However, most of the research on LoRA has focused on large-scale models, where the overall trainable capacity remains high despite parameter reduction. Its application to smaller object detection models, which have significantly fewer parameters, remains largely unexplored. This is particularly relevant in cross-domain few-shot scenarios, where models must generalize from very limited samples and domain shifts are prevalent.

B. Few-Shot and Cross-Domain Object Detection

FSOD focuses on detecting novel object categories using only a small amount of labeled data. Two primary approaches dominate the literature [13]: meta-learning [14] and fine-tuning-based methods. Meta-learning trains models to quickly adapt to new classes by learning transferable representations, while fine-tuning adapts pre-trained models to novel categories using limited labeled data. Although meta-learning excels at learning generalizable features, fine-tuning remains the dominant strategy, particularly in cross-domain scenarios where shifts in data distribution introduce additional complexity [15], [16].

Cross-domain object detection extends FSOD by requiring models to generalize not only to new classes but also to different data distributions. The most common approach is fine-tuning, where a model pre-trained on a source domain is adapted to a target domain using a limited number of labeled examples.

Aerial object detection presents unique challenges due to variations in image resolution, sensor types, and environmental conditions across datasets. To address these domain shifts, researchers have explored techniques such as feature alignment [17] and multi-scale representation learning [18], which aim to bridge the gap between satellite images and drone-captured

images. However, these methods often require extensive adaptation, increasing computational complexity and limiting their practicality in resource-constrained scenarios.

C. DiffusionDet

DiffusionDet [10] is an object detection framework that formulates detection as a denoising diffusion process. Unlike traditional detectors that rely on handcrafted components like anchor boxes or region proposal networks, DiffusionDet directly predicts object bounding boxes and categories by iteratively refining noisy proposals. This approach has proven particularly effective for detecting small objects, making it well-suited for challenging domains such as aerial images, where objects are often densely distributed and vary significantly in scale.

Recent work has explored adapting DiffusionDet to few-shot object detection (FSOD) and cross-domain scenarios. For instance, [19] demonstrated that DiffusionDet can be fine-tuned for few-shot or cross-domain settings by leveraging its iterative refinement process to generalize better to categories with limited labeled data. However, these adaptations often require extensive fine-tuning, which can lead to overfitting, especially in resource-constrained settings with limited labeled data.

Our work builds on these advancements by integrating Low-Rank Adaptation (LoRA) into DiffusionDet, enabling efficient adaptation to cross-domain few-shot object detection. By injecting low-rank matrices into DiffusionDet's architecture, we aim to reduce overfitting while maintaining the model's ability to generalize across domains. This approach extends prior efforts by addressing overfitting challenges associated with fine-tuning DiffusionDet, particularly in the context of aerial images where domain shifts and data scarcity are prevalent.

D. Other Detectors in Cross-Domain Few-Shot Scenarios

Beyond DiffusionDet, several other detectors have been adapted for cross-domain few-shot object detection, each addressing the challenges of domain shifts and limited labeled data. For instance, CD-ViTO [20] leverages vision transformers combined with cross-domain alignment techniques to improve generalization across diverse datasets. By integrating domain-adversarial training and feature alignment modules, CD-ViTO effectively reduces domain discrepancies while maintaining high detection accuracy in few-shot settings. Other approaches include Meta-Det [21]. These methods showcase the variety of strategies available for tackling cross-domain few-shot detection. It would be interesting to extend our studies to include such models, further exploring the potential of LoRA in diverse architectures and settings.

III. CHALLENGE OF CROSS-DOMAIN FEW-SHOT OBJECT DETECTION IN AERIAL IMAGES

Aerial images present a unique set of challenges for object detection [3], making them an especially demanding domain

for few-shot learning and cross-domain adaptation. Unlike ground-level images, aerial scenes often contain numerous small objects, such as vehicles, buildings, or infrastructure components, which are densely distributed across the image. The small size and high density of these objects make them difficult to detect and localize accurately, even for state-of-the-art models. Additionally, aerial images exhibit significant variations in scale, orientation, and lighting conditions, further complicating the detection task. These challenges are exacerbated in few-shot settings, where models must learn to detect novel object categories with only a handful of examples, and cross-domain adaptation introduces even further complexity. In this context, the risk of overfitting is particularly high, as small models may struggle to capture the diverse and intricate patterns present in aerial scenes.

In this context, foundation models, including Vision-Language Models (VLMs) [22] [23], are less suitable for aerial object detection. Beyond the practical limitations of deploying such models in resource-constrained environments, their training data often focuses on horizontal perspectives, resulting in representations that are biased toward ground-level views. For example, in the embedding space of a VLM, the word "car" is more likely to be associated with horizontal representations of cars rather than their aerial counterparts. This misalignment between the model's learned representations and the unique characteristics of aerial images limits their effectiveness in this domain.

Cross-domain few-shot object detection adds an additional level of complexity. In classic few-shot object detection, the model has access to a sufficient number annotated image of base classes from the same domain and aims to generalize to novel classes with limited examples. In contrast, cross-domain few-shot detection involves adapting to a target domain where all classes are novel, using a source domain with abundant data. This scenario requires the model to bridge significant domain shifts while learning from scarce labeled data.

IV. METHODOLOGY

Our methodology evaluates the effectiveness of LoRA [9] for cross-domain few-shot object detection using the DiffusionDet [10] framework. We begin with a pre-trained DiffusionDet model on the COCO [4] dataset, which serves as the source domain, and fine-tune it on reduced subsets of the DOTA [5] and DIOR [6] datasets as target domains. These subsets are carefully curated to simulate a few-shot regime, ensuring that the model is trained with limited labeled data. The backbone used for DiffusionDet is ResNet50 [24], providing a balance between computational efficiency and feature extraction capability.

First, to simulate a few-shot setting, we randomly sampled multiple subsets from the DOTA and DIOR datasets for each shot configuration. Averaging the results across these subsets accounts for selection variability, reduces dataset bias, and ensures a fair evaluation of the model's generalization ability.

To establish a baseline, we first fine-tune the pre-trained DiffusionDet model on the reduced subsets of DOTA and

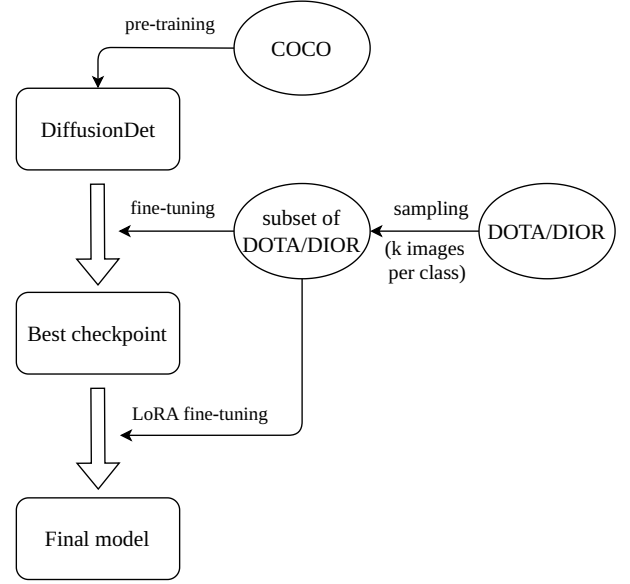


Fig. 1. Training Pipeline for DiffusionDet with LoRA After Intermediate Fine-Tuning.

DIOR without freezing any parameters. The results from this baseline experiment serve as a reference point for evaluating the effectiveness of LoRA in the subsequent steps.

To evaluate the effectiveness of LoRA in cross-domain few-shot object detection, we designed a series of experiments comparing two key approaches, as described below.

1) In the first one, we explore the direct application of LoRA to the pre-trained DiffusionDet model. By injecting low-rank matrices into the model's architecture and freezing the original weights, we significantly reduce the number of trainable parameters. This approach aims to mitigate overfitting while maintaining the model's ability to generalize across domains. The model is then fine-tuned on the reduced subsets of DOTA and DIOR, and its performance is evaluated on a separate validation set.

2) In this second approach, illustrated in Fig. 1, we first fine-tune the pre-trained DiffusionDet model on the reduced subsets of DOTA and DIOR until it reaches a checkpoint with optimal performance on the validation set, just before overfitting occurs. We then apply LoRA to this checkpoint, freezing the original weights and fine-tuning only the low-rank matrices. The idea behind this two-stage approach is to first push the model to its limits, approaching the point of overfitting, and then use LoRA to continue training without the risk of overfitting. By applying LoRA to this checkpoint, we freeze the original weights and fine-tune only the low-rank matrices, allowing the model to adapt further while maintaining generalization. This strategy leverages the benefits of both full fine-tuning and parameter-efficient adaptation, achieving a better balance between performance and robustness.

TABLE I. Object etection results (mAP) of DiffusionDet model pretrained on COCO [4] and fine-tuned on DIOR [6] and DOTA [5] datasets in cross-domain few-shot settings. We compare the baseline (no LoRA), LoRA with different ranks (4, 8, 32, 128), and LoRA applied after intermediate fine-tuning. Best results per shot configuration are bold.

Dataset	Shots	Baseline (no LoRA)	LoRA				LoRA after a Fine-Tuning			
			4	8	32	128	4	8	32	128
DIOR	1	10.66	7.32	6.83	7.51	6.52	11.48	11.58	11.64	11.57
	5	31.29	24.14	24.84	24.02	24.45	32.40	32.2	32.35	32.45
	10	41.50	34.72	34.25	33.91	33.23	40.64	40.68	40.81	41.18
	50	59.71	56.43	53.41	53.24	56.47	57.72	57.74	57.78	57.70
DOTA	1	4.23	1.86	1.81	1.81	1.70	4.89	4.85	4.84	4.97
	5	22.52	15.17	14.83	15.15	14.73	22.75	22.83	22.91	22.85
	10	32.77	25.12	24.54	25.07	25.06	32.23	32.33	32.30	32.14
	50	49.17	42.90	42.07	42.50	42.15	47.90	47.99	48.03	47.94

V. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of LoRA in cross-domain few-shot object detection, we conducted a series of experiments using the DiffusionDet framework. Our experiments were designed to assess the impact of LoRA on model performance, overfitting, and generalization across domains, particularly in the challenging context of aerial images.

We used DiffusionDet model pre-trained on the COCO dataset, using the weights provided by the original authors [10]. For the target domains, we selected the DOTA and DIOR datasets, which were converted to COCO format¹. To simulate a few-shot setting, we randomly selected k images per class for training, where k represents the number of shots. To ensure a fair comparison and account for variability in the selection of images, we repeated each experiment 5 times and reported the average results. Given that DOTA images often contain more than 100 objects, we set the maximum detection threshold to 300 in the pycocoapi evaluation toolkit [25].

All experiments were conducted over 300 epochs, following the training protocol established by the original DiffusionDet authors. We evaluated the model performance using the mean average precision (mAP) at an IoU threshold of 0.5, which is a standard metric for object detection tasks. For the baseline, we fine-tuned the pre-trained DiffusionDet model on the few-shot subsets of DOTA and DIOR without freezing any parameters. As described in section IV, for the LoRA-based experiments, we explored two approaches: applying LoRA directly to the pre-trained model and applying LoRA to the best checkpoint obtained from the baseline fine-tuning. In the latter approach, we selected the checkpoint with the highest validation performance after 300 epochs and fine-tuned it further using LoRA.

To investigate the impact of rank selection on LoRA’s performance, we tested four different ranks: 4, 8, 32, and 128. The results of these experiments are presented in Tab. I, which compares the performance of the baseline, direct LoRA application, and LoRA after intermediate fine-tuning across different ranks. By averaging results across multiple

runs, we ensure a robust evaluation of LoRA’s effectiveness in mitigating overfitting and improving generalization in cross-domain few-shot object detection.

Across both datasets, the baseline outperforms direct LoRA application in all shot configurations. However, LoRA applied after intermediate fine-tuning shows improvements, particularly in low-shot settings. On DIOR, the best mAP of 11.64 (rank 32) is achieved in the 1-shot setting, while on DOTA, the best mAP of 4.97 (rank 128) is achieved. Similarly, in the 5-shot setting, the best mAPs are 32.45 (rank 128) on DIOR and 22.91 (rank 32) on DOTA. In higher-shot settings, the baseline remains competitive, but LoRA after fine-tuning closely matches its performance.

LoRA after intermediate fine-tuning slightly improves performance in low-shot settings, while the baseline remains strong in higher-shot configurations. The choice of rank in LoRA has a moderate impact on performance, with lower ranks (e.g., 4, 8) often performing comparably to higher ranks (e.g., 32, 128).

VI. DISCUSSION

The experimental results demonstrate that LoRA, particularly when applied after intermediate fine-tuning, is a promising approach for cross-domain few-shot object detection. This improvement, although minimal, suggests that efficient parameter fine-tuning could be a viable alternative to full fine-tuning, particularly in resource-constrained environments.

The choice of rank in LoRA has a moderate impact on performance, with lower ranks (e.g., 4, 8) often performing comparably to higher ranks (e.g., 32, 128). This indicates that lower ranks may suffice for many applications. However, the baseline’s strong performance in higher-shot configurations underscores the importance of full fine-tuning when sufficient data is available. These results underline the need for a balanced approach, adjusting the fine-tuning strategy to the specific requirements of the task and dataset, and could be the subject of further study.

While our approach shows promise, it is not without limitations. The performance of LoRA depends on the quality

¹They can be found here: <https://huggingface.co/datasets/HichTala/dota>, <https://huggingface.co/datasets/HichTala/dior>.

of the initial fine-tuning. Additionally, our experiments are limited to DiffusionDet and two aerial datasets; extending this approach to other architectures and domains could yield further insights. Future work could explore combining LoRA with other few-shot learning techniques to enhance its effectiveness.

It is worth noting that the cross-domain scenario explored in this work—adapting from natural images (COCO) to aerial images (DOTA and DIOR), remains particularly challenging due to the significant differences in perspective, scale, and object appearance. A simpler yet equally interesting scenario would involve adapting between aerial images, such as from DOTA to DIOR or vice versa, where the domain shifts are less extreme. One could also explore adapting aerial images across different environments, seasons, or lighting conditions. Such experiments could provide valuable insights into the robustness and versatility of LoRA in less extreme domain shifts.

VII. CONCLUSION

In this work, we investigated the application of LoRA to DiffusionDet model for cross-domain few-shot object detection, with a focus on the challenging domain of aerial images. Using the DiffusionDet framework, we evaluated the effectiveness of LoRA in mitigating overfitting and improving generalization across the DOTA and DIOR datasets. Our experiments compared three approaches: (1) baseline fine-tuning, (2) direct LoRA application, and (3) LoRA applied after intermediate fine-tuning. The results demonstrated that while the baseline outperformed direct LoRA application, LoRA after intermediate fine-tuning achieved competitive performance, particularly in low-shot settings (e.g., 1-shot and 5-shot). This highlights LoRA’s potential to balance adaptation and generalization, especially when combined with an initial fine-tuning phase.

VIII. ACKNOWLEDGMENT

The authors would like to thank Région Ile de France, COSE company and LabCom IRISER (ANR-21-LCV3-0004) for funding and collaborations, and GENCI-IDRIS (AD011014772R1) for HPC resources.

REFERENCES

- [1] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2302.13971>
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [3] P. Le Jeune and A. Mokraoui, “Improving few-shot object detection through a performance analysis on aerial and natural images,” in *30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 513–517.
- [4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [5] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, “Dota: A large-scale dataset for object detection in aerial images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.
- [6] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.
- [7] Z. Chen, H. Wang, X. Wu, J. Wang, X. Lin, C. Wang, K. Gao, M. Chapman, and D. Li, “Object detection in aerial images using dota dataset: A survey,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 134, p. 104208, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1569843224005648>
- [8] J. Leng, Y. Ye, M. Mo, C. Gao, J. Gan, B. Xiao, and X. Gao, “Recent advances for aerial object detection: A survey,” *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–36, 2024.
- [9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *CoRR*, vol. abs/2106.09685, 2021. [Online]. Available: <https://arxiv.org/abs/2106.09685>
- [10] S. Chen, P. Sun, Y. Song, and P. Luo, “Diffusiondet: Diffusion model for object detection,” *arXiv preprint arXiv:2211.09788*, 2022.
- [11] Y. Wang, R. Zhang *et al.*, “Low-rank adaptation for vision transformers in remote sensing applications,” *arXiv preprint arXiv:2301.09978*, 2023.
- [12] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J. Hwang, K. Chang, and J. Gao, “Grounded language-image pre-training,” *CoRR*, vol. abs/2112.03857, 2021. [Online]. Available: <https://arxiv.org/abs/2112.03857>
- [13] M. Kohler, M. Eisenbach, and H.-M. Gross, “Few-shot object detection: A comprehensive survey,” *IEEE transactions on neural networks and learning systems*, 2021.
- [14] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.
- [15] W. Xiong, “Cd-fsod: A benchmark for cross-domain few-shot object detection,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [16] K. Lee, H. Yang, S. Chakraborty, Z. Cai, G. Swaminathan, A. Ravichandran, and O. Dabeer, “Rethinking few-shot object detection on a multi-domain benchmark,” in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*. Springer, 2022, pp. 366–382.
- [17] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.
- [18] 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 conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [19] P. L. Jeune, H. Talaoubrid, and A. Mokraoui, “Improving few-shot and cross-domain object detection on aerial images with a diffusion-based detector,” in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 8148–8152.
- [20] Y. Fu, Y. Wang, Y. Pan, L. Huai, X. Qiu, Z. Shangquan, T. Liu, Y. Fu, L. V. Gool, and X. Jiang, “Cross-domain few-shot object detection via enhanced open-set object detector,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.03094>
- [21] 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*, 2019, pp. 9577–9586.
- [22] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” 2023.
- [23] J. Chen, J. Yang, H. Wu, D. Li, J. Gao, T. Zhou, and B. Xiao, “Florence-vl: Enhancing vision-language models with generative vision encoder and depth-breadth fusion,” 2024. [Online]. Available: <https://arxiv.org/abs/2412.04424>
- [24] 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.
- [25] cocodataset, “Coco api - dataset.” [Online]. Available: <https://github.com/cocodataset/cocoapi/tree/master/PythonAPI/pycocotools>