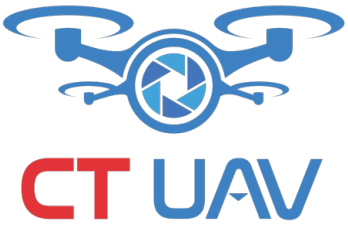


BÁO CÁO KỸ THUẬT

Áp dụng AI & Machine Learning vào bài toán FSOD với ảnh chụp trên không

Người thực hiện: Nhan Hữu Hiếu

Thành phố Hồ Chí Minh, tháng 12 năm 2025

Mục lục

Phần 1

XÁC ĐỊNH BÀI TOÁN

1.1 Tổng quan về Few-Shot Object Detection (FSOD) trong ảnh chụp từ trên không

Few-Shot Object Detection (FSOD) là hướng nghiên cứu nhằm phát hiện đối tượng mới trong điều kiện số lượng mẫu huấn luyện cực kỳ hạn chế. Đặc biệt, trong lĩnh vực ảnh chụp từ trên không (aerial images), bài toán này càng trở nên phức tạp do đặc trưng của những hình ảnh này là chứa các đối tượng nhỏ, đa dạng về hướng, phân bố dày đặc, nền phức tạp và thiếu hụt dữ liệu gán nhãn chất lượng cao [1].

Các phương pháp FSOD truyền thống chủ yếu phát triển trên ảnh tự nhiên (natural images), nhưng khi áp dụng sang ảnh hàng không, hiệu quả giảm mạnh do các yếu tố như tỷ lệ đối tượng nhỏ cao, biến đổi hình học lớn, và sự khác biệt về phân phối dữ liệu [2]. Do đó, các giải pháp FSOD cho ảnh hàng không cần giải quyết đồng thời các vấn đề: học đặc trưng phân biệt từ ít mẫu, thích ứng với đa dạng hướng và tỷ lệ và tận dụng thông tin không gian, ngữ cảnh và dữ liệu không gán nhãn.

1.2 Thách thức đặc thù của FSOD trên ảnh chụp từ trên không

- **Đối tượng nhỏ, đa tỷ lệ và đa hướng:** Trong các bộ dữ liệu ảnh hàng không như DOTA, DIOR, tỷ lệ đối tượng nhỏ (dưới 32×32 pixel) chiếm trên 50% tổng số mẫu [2]. Đối tượng thường có hướng bất kỳ, phân bố dày đặc, có thể bị che khuất và lẫn với nền phức tạp. Điều này gây khó khăn cho việc trích xuất đặc trưng và phân biệt đối tượng.
- **Sự thiếu hụt dữ liệu gán nhãn và mất cân bằng lớp:** Việc thu thập và gán nhãn dữ liệu ảnh hàng không rất tốn kém, dẫn đến số lượng mẫu cho các lớp mới (novel classes) rất ít, thậm chí chỉ 1–5 ảnh/lớp. Ngoài ra, phân phối lớp thường lệch (long-tail), nhiều lớp hiếm xuất hiện, gây khó khăn cho mô hình học được đặc trưng tổng quát [1].
- **Đa dạng về điều kiện chụp và nền phức tạp:** Ảnh hàng không có thể thu nhận từ nhiều nguồn (vệ tinh, UAV, máy bay), với độ phân giải, góc chụp, điều kiện ánh sáng khác nhau. Nền ảnh thường chứa nhiều chi tiết không liên quan, gây nhiễu cho quá trình phát hiện đối tượng [2].

1.3 Liên hệ với sản phẩm của CT UAV

Hiện tại, CT UAV đang nghiên cứu các dòng UAV có nhu cầu phát hiện vật thể từ hình ảnh chụp bằng camera gắn trên thiết bị. Chẳng hạn, UAV An ninh cần nhận diện người lạ; UAV Năng lượng tái tạo phải phát hiện vết nứt, hư hỏng trên các cánh quạt điện gió hoặc tấm pin mặt trời; UAV Kiểm tra xây dựng cần xác định các vật thể trong công trường và đánh giá tình trạng của chúng. Các bài toán này đều đòi hỏi khả năng xử lý ảnh chính xác nhưng thường chỉ có rất ít mẫu huấn luyện.

Tuy vậy, đối với từng loại UAV, ta có thể xác định trước một số bối cảnh đặc

trưng. Ví dụ: UAV An ninh thường hoạt động trong khu vực kho bãi, nhà xưởng hoặc nơi đông người; UAV Kiểm tra xây dựng thường bay qua các công trình, cột điện, tòa nhà hay các phương tiện thi công. Việc nhận diện được các bối cảnh đặc thù này cho phép huấn luyện trước mô hình về môi trường hoạt động, từ đó cải thiện hiệu quả phát hiện vật thể khi tích hợp lên UAV.

Từ những đặc điểm thực tiễn và phù hợp với tính chất của bài toán FSOD cho ảnh trên không, các phần tiếp theo của báo cáo sẽ dựa trên cơ sở này để phân tích và đề xuất giải pháp.

Phần 2

Câu hỏi 1: Các giải pháp để phát hiện đối tượng với số lượng mẫu ít trong ảnh chụp trên không (aerial images)

2.1 Phương pháp dựa trên Học Siêu dữ liệu (Meta-learning)

Phương pháp này huấn luyện mô hình để "học cách học" (learning to learn). Meta-learning được thiết kế để giúp mô hình nhanh chóng thích nghi với các tác vụ mới (lớp đối tượng mới) chỉ bằng vài lần lặp lại [3].

Cơ chế hoạt động: Dựa trên việc huấn luyện trên nhiều tasks set khác nhau, mỗi tác vụ bao gồm một tập hỗ trợ (support set - S) và một tập truy vấn (query set - Q) [3] [4].

- **Mục tiêu:** Học một bộ tham số khởi tạo θ_0 tối ưu (Optimization-based) hoặc một hàm khoảng cách hiệu quả (Metric-based) [4].
- **Quá trình thích nghi:** Khi gặp một lớp đối tượng mới (novel class), mô hình sử dụng θ_0 để nhanh chóng điều chỉnh tinh chỉnh chỉ với các mẫu ít ỏi trong tập S, sau đó đánh giá hiệu suất trên Q [4] .
- **Các Kiến trúc Tiêu biểu:**
  - Optimization-based (ví dụ: MAML): Tìm kiếm θ_0 để tối đa hóa tốc độ hội

tụ khi điều chỉnh tinh chỉnh [4].

- Metric/Relation-based (ví dụ: Meta R-CNN): Học không gian nhúng (embedding space) để phân loại các vùng đề xuất dựa trên khoảng cách gần nhất với Prototype (vector đặc trưng trung bình) của các lớp trong tập hỗ trợ [5].

Tính khả thi: Meta-Learning giải quyết trực tiếp những thách thức chính của FSOD trong bối cảnh ảnh hàng không.

- **Giải quyết vấn đề vật thể nhỏ và đặc trưng yếu**: Meta-Learning huấn luyện mạng để tập trung vào việc học các đặc trưng có tính phân biệt cao và khả năng chuyển giao mạnh. Thay vì học các đặc trưng chi tiết cho một lớp cụ thể, mô hình học các mẫu hình chung (ví dụ: hình chữ nhật, độ phản chiếu cao) có thể áp dụng cho nhiều loại phương tiện.
- **Đối phó với biến thể lớn**: Bằng cách trình bày các tác vụ với các mức độ biến thể khác nhau trong quá trình huấn luyện siêu dữ liệu, mô hình buộc phải học cách thích ứng với sự thay đổi về tỷ lệ, độ sáng hay các biến thể khác của vật thể.

2.2 Phương pháp Contrastive Learning trên Proposals (FSCE)

FSCE (Few-Shot Object Detection via Contrastive Learning on Proposals) là một phương pháp tận dụng khả năng của contrastive learning để làm đa dạng và tách biệt các đặc trưng vùng đề xuất, từ đó cải thiện đáng kể hiệu suất FSOD [6].

Cơ chế hoạt động: FSCE được xây dựng dựa trên kiến trúc phát hiện đối tượng hai giai đoạn (ví dụ: Faster R-CNN), với một module học tương phản mới được thêm vào [6].

- **Kiến trúc cơ bản**:
  - Giai đoạn 1: Huấn luyện cơ sở (base training): Mô hình được huấn luyện

đầy đủ trên các lớp cơ sở với lượng dữ liệu dồi dào. Giai đoạn này đảm bảo mạng có khả năng trích xuất đặc trưng vùng tốt [7].

  - Giai đoạn 2: Tinh chỉnh (fine-tuning) và học tương phản: Giai đoạn này được thực hiện trên các lớp mới (novel classes) với số lượng mẫu rất ít (K-shot). Đây là giai đoạn thể hiện cơ chế contrastive learning [7].

- **Module học tương phản trên proposals**: Module này tập trung vào việc định hình lại không gian đặc trưng của các vùng đề xuất để tối đa hóa sự khác biệt giữa các đặc trưng positive và negative [7].

Tính khả thi: FSCE là một phương pháp phù hợp và mạnh mẽ cho FSOD vì nó giải quyết ba điểm yếu chính của các phương pháp truyền thống [6] [7]:

- **Cải thiện khả năng phân biệt đặc trưng**: Học tương phản tạo ra sự đối lập mạnh mẽ, phân biệt mạnh giữa các đặc trưng positive và negative. Từ đó làm cho không gian đặc trưng trở nên rời rạc và có cấu trúc tốt hơn, ngay cả khi chỉ sử dụng một lượng nhỏ dữ liệu.

- **Giảm thiểu sự ảnh hưởng của nền**: Bằng cách coi các đặc trưng nền là các mẫu negative mạnh, module tương phản học cách đẩy các đặc trưng đối tượng cần detect ra xa không gian nền. Điều này giúp mô hình tập trung hơn vào các đặc trưng quan trọng của vật thể, nâng cao độ chính xác của phân loại.

2.3 Phương pháp tinh chỉnh, cải tiến (Improved Fine-tuning)

Phương pháp này về cơ bản là một hình thức transfer learning được điều chỉnh, tinh chỉnh theo một hướng cụ thể nhằm sử dụng những kiến thức mà mô hình học được ban đầu để đối phó với dữ liệu huấn luyện ít ỏi [8].

Cơ chế hoạt động: Phương pháp này chia quá trình huấn luyện thành hai giai đoạn rõ ràng, nhằm mục đích truyền tải tri thức đã học từ các lớp cơ sở sang các lớp mới một cách hiệu quả nhất [8] [9].

- **Giai đoạn 1: Huấn luyện Cơ sở**: Huấn luyện một mô hình phát hiện đối tượng tiêu chuẩn (ví dụ: Faster R-CNN) trên bộ dữ liệu lớn của các lớp cơ sở. Có rất nhiều mô hình có sẵn đã được huấn luyện như vậy (pre-train model).
- **Giai đoạn 2: Điều chỉnh, tinh chỉnh, cải tiến (fine-tuning)**: Sử dụng mô hình đã huấn luyện từ Giai đoạn 1 và điều chỉnh nó chỉ với K mẫu cho mỗi lớp mới. Thay đổi lớp cuối cùng của mô hình đã được huấn luyện để phân biệt các vật thể ban đầu để mô hình đó có thể nhận diện thêm các lớp mới tốt hơn mà không cần mất thời gian huấn luyện từ đầu.

Tính khả thi: Phương pháp này là một lựa chọn tốt, hiệu quả cho ảnh trên không vì nó mang lại sự cân bằng giữa hiệu suất và tính ứng dụng [10]:

- **Tính ứng dụng và triển khai cao**: Phương pháp fine-tuning có thể được áp dụng trực tiếp lên bất kỳ kiến trúc phát hiện đối tượng tiêu chuẩn nào (Faster R-CNN, YOLO, SSD) vốn đã phổ biến trong lĩnh vực ảnh hàng không. Từ đó giúp các nhà nghiên cứu và kỹ sư dễ dàng triển khai mà không cần thiết kế lại mạng lưới phức tạp.
- **Khả năng kiểm soát quá khớp**: Việc giữ cố định phần lớn các layer của mô hình giúp mô hình chỉ tập trung học các đặc trưng phân biệt cấp cao (cần thiết cho việc phân loại lớp mới) thay vì học lại các đặc trưng cấp thấp (cạnh, kết cấu) từ đầu.
- Tóm lại, đây là giải pháp thực tế, ít phức tạp về mặt kiến trúc, nhưng vẫn đảm bảo sự ổn định và hiệu suất cao trong việc định vị và phân loại bằng cách kiểm soát quá trình truyền tải tri thức từ lớp cơ sở sang lớp mới, rất cần thiết cho việc xử lý các đối tượng nhỏ và có sự biến đổi cao trong ảnh hàng không.

Phần 3

Câu hỏi 2: Demo giải pháp

Ở đây, chúng ta sẽ demo phương pháp thứ ba, Fine-tuning, và mô hình được lựa chọn để fine-tune là Faster R-CNN vì các lý do sau:

- Như đã trình bày ở trên, phương pháp này đảm bảo cả về hiệu suất lẫn tính ứng dụng, nó không mất quá nhiều thời gian để thiết kế lại nhưng cũng đủ chính xác để giải quyết bài toán.
- Như có phân tích ở đầu báo cáo, các sản phẩm của CT UAV cần đến ứng dụng xác định vật thể thường sẽ hoạt động trong những môi trường đặc thù, có thể xác định trước bối cảnh không gian, vật thể, ... nên việc fine-tuning từ một mô hình có sẵn có lẽ sẽ mang nhiều ý nghĩa và tương thích với tình huống ứng dụng thực tế của sản phẩm UAV.
- Ngoài ra, còn có lý do chủ quan là thời gian và thiết bị không cho phép nên việc lựa chọn phương pháp này là tối ưu và hợp lý nhất.

Model được sử dụng để fine-tune là **Faster R-CNN**.

Dữ liệu được sử dụng để huấn luyện và kiểm thử được trích từ bộ dữ liệu DOTA (Dataset for Object detection in Aerial Images), có được điều chỉnh để phù hợp với model Faster R-CNN.

Model hiện tại được điều chỉnh để phân biệt 4 lớp vật thể bao gồm: plane, ship, large-vihicle và small-vihicle.

Code demo được tổ chức trong một file .ipynb (chi tiết có thể được tìm thấy trong link github bài làm).

Kết quả:

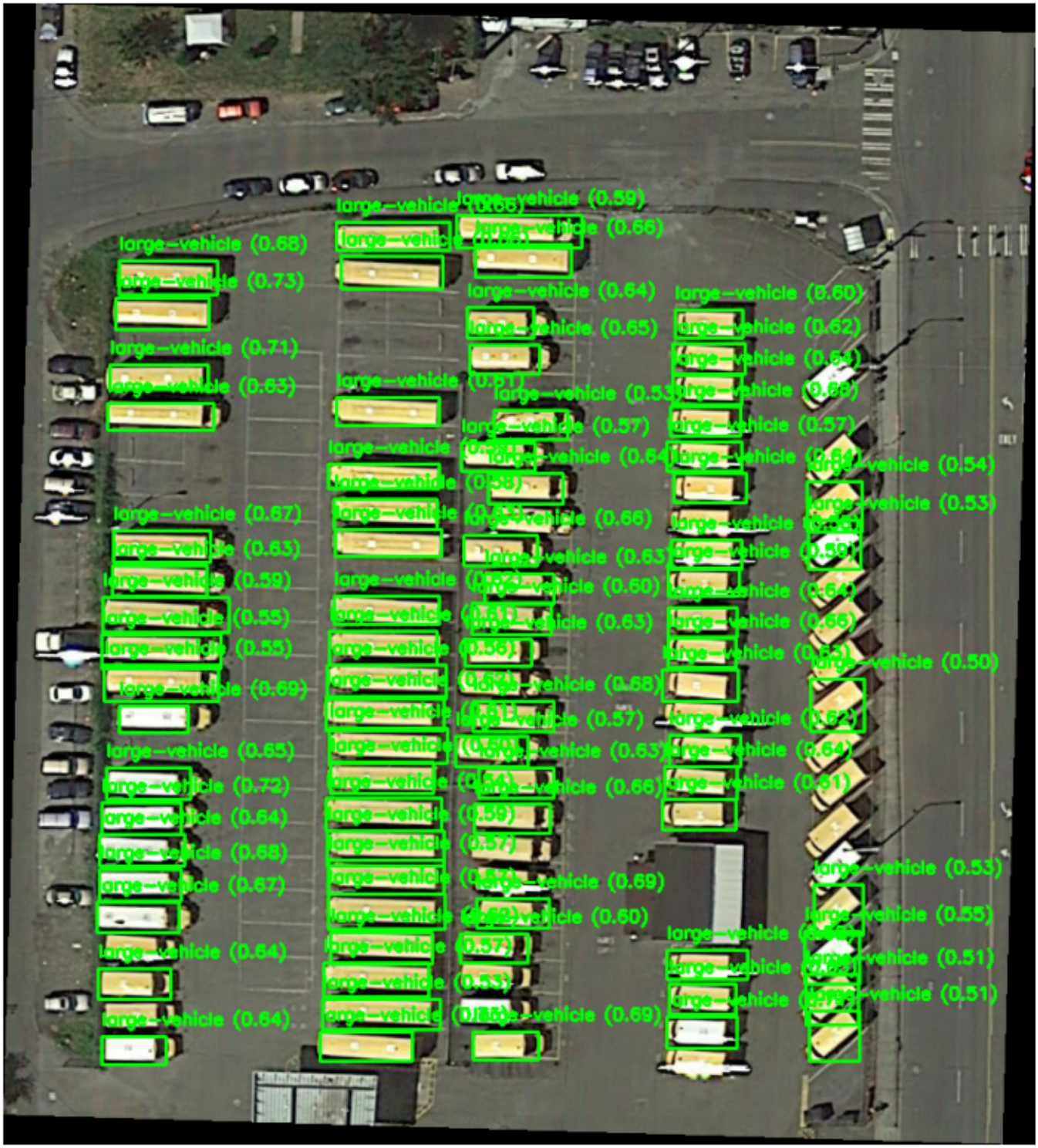
Model được đánh giá dựa trên độ đo mAP với kết quả sau:

```
{'map': tensor(0.1172),
 'map_50': tensor(0.2408),
 'map_75': tensor(0.0704),
 'map_small': tensor(0.0413),
 'map_medium': tensor(0.2201),
 'map_large': tensor(0.0010),
 'mar_1': tensor(0.0021),
 'mar_10': tensor(0.0172),
 'mar_100': tensor(0.1310),
 'mar_small': tensor(0.0435),
 'mar_medium': tensor(0.2432),
 'mar_large': tensor(0.0048),
 'map_per_class': tensor(-1.),
 'mar_100_per_class': tensor(-1.),
 'classes': tensor([1, 3, 4], dtype=torch.int32)}
```

Kết quả khi test thử với một hình ảnh khác như sau (Xem hình dưới):

Nhận xét:

- Model đã nhận diện được các vật thể trong hình. Tuy nhiên vẫn để sót một số vật thể.
- Tuy nhiên, kết quả này hứa hẹn nếu được tiếp tục phát triển, hoặc lựa chọn những mô hình khác, cách fine-tuning khác thì sẽ cho một kết quả khả quan hơn rất nhiều.
- Có thể phát triển theo một số hướng sau:
  - Áp dụng contrastive learning on proposals.
  - Tạo ra các prototype chất lượng cao và ít nhiễu cho các lớp mới.
  - Tạo mẫu đặc trưng Tổng hợp, tăng cường dữ liệu cho các lớp mới trong không gian đặc trưng.


large-vehicle (0.68)
large-vehicle (0.73)
large-vehicle (0.71)
large-vehicle (0.63)
large-vehicle (0.67)
large-vehicle (0.63)
large-vehicle (0.59)
large-vehicle (0.55)
large-vehicle (0.55)
large-vehicle (0.69)
large-vehicle (0.65)
large-vehicle (0.72)
large-vehicle (0.64)
large-vehicle (0.68)
large-vehicle (0.67)
large-vehicle (0.64)
large-vehicle (0.64)
large-vehicle (0.66)
large-vehicle (0.59)
large-vehicle (0.64)
large-vehicle (0.65)
large-vehicle (0.53)
large-vehicle (0.57)
large-vehicle (0.66)
large-vehicle (0.63)
large-vehicle (0.60)
large-vehicle (0.63)
large-vehicle (0.68)
large-vehicle (0.57)
large-vehicle (0.63)
large-vehicle (0.66)
large-vehicle (0.69)
large-vehicle (0.60)
large-vehicle (0.60)
large-vehicle (0.62)
large-vehicle (0.54)
large-vehicle (0.57)
large-vehicle (0.64)
large-vehicle (0.54)
large-vehicle (0.53)
large-vehicle (0.59)
large-vehicle (0.64)
large-vehicle (0.66)
large-vehicle (0.53)
large-vehicle (0.50)
large-vehicle (0.62)
large-vehicle (0.64)
large-vehicle (0.61)
large-vehicle (0.53)
large-vehicle (0.55)
large-vehicle (0.51)
large-vehicle (0.51)

Tài liệu tham khảo

[1] S. Liu, Y. You, H. Su, G. Meng, W. Yang, and F. Liu, "Few-shot object detection in remote sensing image interpretation: Opportunities and challenges," *Remote Sensing*, vol. 14, no. 18, p. 4435, 2022.

[2] P. Le Jeune and A. Mokraoui, "Improving few-shot object detection through a performance analysis on aerial and natural images," in *Proceedings of the 30th European Signal Processing Conference (EUSIPCO 2022)*, 2022, pp. 513–517, paper ID: 0000513. [Online]. Available: https://eurasip.org/Proceedings/Eusipco/Eusipco2022/pdfs/0000513.pdf

[3] W. Guan, Z. Yang, X. Wu, L. Chen, F. Huang, X. He, and H. Chen, "Efficient meta-learning enabled lightweight multiscale few-shot object detection in remote sensing images," *arXiv preprint*, vol. arXiv:2404.18426, 2024. [Online]. Available: https://arxiv.org/abs/2404.18426

[4] Z. Yang, W. Guan, L. Xiao, and H. Chen, "Few-shot object detection in remote sensing images via data clearing and stationary meta-learning," *Sensors*, vol. 24, no. 12, p. 3882, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/12/3882

[5] D. Zhao, F. Shao, Q. Liu, H. Zhang, Z. Zhang, and L. Yang, "Improved architecture and training strategies of yolov7 for remote sensing image object detection," *Remote Sensing*, vol. 16, no. 17, p. 3321, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/17/3321

[6] A. Wilf, M. Q. Ma, P. P. Liang, A. Zadeh, and L.-P. Morency, "Face-to-face contrastive learning for social intelligence question-answering," 2022. [Online]. Available: https://arxiv.org/abs/2208.01036

[7] 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)*, June 2021, pp. 7352–7362.

[8] Q. Pan, K. Fu, and G. Wang, "Study on few-shot object detection approach based on improved rpn and feature aggregation," *Applied Sciences*, vol. 15, no. 7, p. 3734, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/7/3734

[9] P. Zhou *et al.*, "Fusing adaptive meta feature weighting for few-shot object detection," in *Proceedings of the Machine Learning Research, vol. 245*, 2024, paper ID: peng24a. [Online]. Available: https://proceedings.mlr.press/v245/peng24a.html

[10] Z. Wang, Y. Gao, Q. Liu, and Y. Wang, "Semantic enhanced few-shot object detection," *arXiv preprint*, vol. arXiv:2406.13498, 2024. [Online]. Available: https://arxiv.org/abs/2406.13498