THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): https://youtu.be/Gw5s6sSlZBs
- Link slides (dạng .pdf đặt trên Github của nhóm): https://github.com/thanhtruong-project/CS2205.FEB2025/blob/main/slide.pdf

- Họ và Tên: Trương Lê Mỹ Thanh
- MSSV: 240101073

- Lớp: CS2205.FEB2025
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Link Github: https://github.com/thanhtruong-project/CS2205.FEB2025
-

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT HIỆN VÀ NHẬN BIẾT ĐỐI TƯỢNG TRONG ĐIỀU KIỆN KHÔNG LÝ TƯỞNG KẾT HỢP VỚI TĂNG CƯỜNG NGỮ NGHĨA

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

OBJECT DETECTION AND RECOGNITION IN NON-IDEAL CONDITIONS COMBINED WITH SEMANTIC AUGMENTATION

TÓM TẮT *(Tối đa 400 từ)*

Phát hiện đối tượng không phải là một chủ đề xa lạ và đã có nhiều mô hình cho phép phát hiện và nhận dạng chính xác và với độ chính xác cao, tuy nhiên, khi áp dụng vào các môi trường hợp thực tế, các điều kiện môi trường không lý tưởng làm cho quá trình nhận dạng diễn ra chính xác trở nên khó khăn, tất cả đều bị ảnh hưởng bởi các thay đổi miền, khiến độ chính xác trở nên thấp hơn dự kiến. Nhiều trường hợp thực tế (dữ liệu từ miền mục tiêu không có sẵn) đòi hỏi các mô hình có khả năng tổng quát hóa từ một miền duy nhất (Single Domain Generalization – SDG).

Vì thế một phương pháp SDG mới được đề xuất, bằng cách tận dụng mô hình thị giác-ngôn ngữ **CLIP (backbone:ResNet101)** để đưa các biến thể miền vào quá trình huấn luyện thông qua mỗi **một danh sách prompt văn bản cho một miền chưa biết**. Phương pháp này giúp mô hình học được đặc trưng ảnh ổn định, từ đó tăng khả năng tổng quát hóa sang các miền chưa từng thấy như mưa, sương mù hay ban đêm,... mà không cần bất kỳ dữ liệu nào từ các miền này, mô hình góp phần hỗ trợ triển khai trên các hệ thống giám sát giao thông thông minh và hiệu quả hay trong lĩnh vực robot cứu hộ và dò thám.

GIỚI THIỆU

Trong lĩnh vực phát hiện đối tượng, các mô hình học sâu đã đạt được độ chính xác cao trong môi trường lý tưởng. Tuy nhiên, khi triển khai thực tế, hiệu suất của các mô hình này thường bị suy giảm đáng kể do sự khác biệt giữa miền huấn luyện và miền thực tế, như thay đổi về điều kiện ánh sáng, thời tiết hoặc môi trường. Đặc biệt, trong nhiều trường hợp, dữ liệu từ miền mục tiêu không có sẵn trong giai đoạn huấn luyện, đặt ra yêu cầu về khả năng **tổng quát hóa từ một miền duy nhất**, gọi là **Single Domain Generalization (SDG)**.

Mặc dù SDG đã đạt được thành tựu trong các bài toán phân loại ảnh, nhưng về phát hiện đối tượng vẫn còn hạn chế. Một số phương pháp trước đây như domain adaptation (DA)[1] yêu cầu dữ liệu từ cả miền nguồn và miền đích, trong khi các hướng tiếp cận multi-source domain generalization (DG)[2] đòi hỏi dữ liệu huấn luyện từ nhiều miền khác nhau. Sau đó, một phương pháp nổi bật là Single-DGOD[3], sử dụng kỹ thuật tách đặc trưng miền và tự chưng cất kiến thức, tuy nhiên nó vẫn chưa thực sự hiệu quả trong các điều kiện miền chưa biết để giải quyết vấn đề miền thích nghi trong các bài toán phát hiện đối tượng [4] .

Xuất phát từ những hạn chế đó, một phương pháp SDG mới được đề xuất, tận dụng backbone mô hình thị giác-ngôn ngữ **CLIP (backbone:ResNet101)** để đưa các biến thể miền vào quá trình huấn luyện thông qua mỗi **một danh sách prompt văn bản cho một miền chưa biết.** Bằng cách này, mô hình học được các đặc trưng ảnh ổn định và có khả năng tổng quát tốt hơn sang các miền chưa từng thấy như mưa, sương mù hay ban đêm, mà không cần dữ liệu từ các miền này trong quá trình huấn luyện, góp phần hỗ trợ nâng cao hiệu quả trong việc giám sát của các hệ thống thông minh cũng như sử dụng robot cứu hộ và dò thám, giảm thiểu rủi ro hay bỏ sót thông tin trong các môi trường nguy hiểm.

Input: Hình ảnh chụp trong điều kiện xấu (ban đêm, mưa, sương,...)
Output: Xác định vị trí và nhận biết vật thể trong hình: người, xe,...

MỤC TIÊU

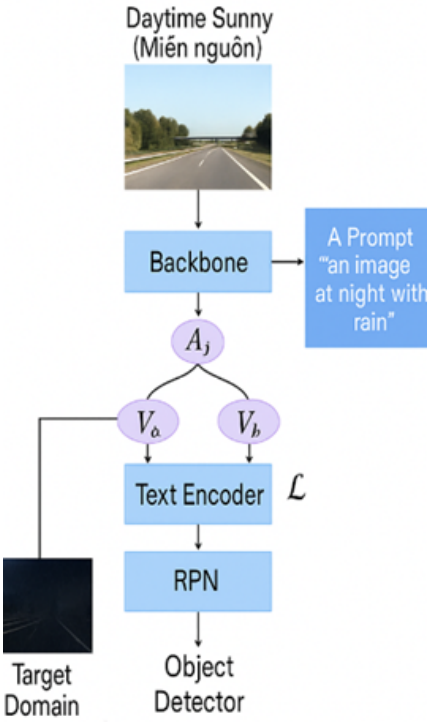
- Tận dụng mô hình CLIP đã huấn luyện trước đó (chỉ lấy backbone: ResNet101)
- Sử dụng mỗi một danh sách các mô tả văn bản mang thông tin ngữ nghĩa để mô phỏng miền chưa biết, tạo ra dữ liệu huấn luyện bổ sung từ chỉ một miền gốc.
- Sử dụng tăng cường ngữ nghĩa + CLIP (chỉ fine-tune backbone) để tăng độ chính xác khi phát hiện vật thể trong điều kiện không lí tưởng trong các miền chưa thấy (như: mưa, sương, đêm,...) chỉ với 1 miền dữ liệu gốc, giúp triển khai tốt trong các hệ thống giám sát giao thông thông minh thực tế hay dò thám.

NỘI DUNG VÀ PHƯƠNG PHÁP

Sử dụng mô hình **tăng cường ngữ nghĩa** kết hợp với **backbone ResNet101 của CLIP**

1. Tăng cường ngữ nghĩa (semantic augmentation)[5]

Dựa trên không gian embedding chung giữa ảnh và văn bản của CLIP, các ảnh từ miền nguồn (ví dụ: "ảnh ban ngày trời nắng") được **dịch chuyển trong không gian đặc trưng** sang miền mới bằng các **prompt văn bản** (ví dụ: "ảnh ban đêm có mưa"). Quá trình này không sửa đổi ảnh gốc, mà tăng cường đặc trưng trích xuất từ ảnh, giúp mô hình học được sự biến thiên miền mà không cần ảnh từ miền mục tiêu.

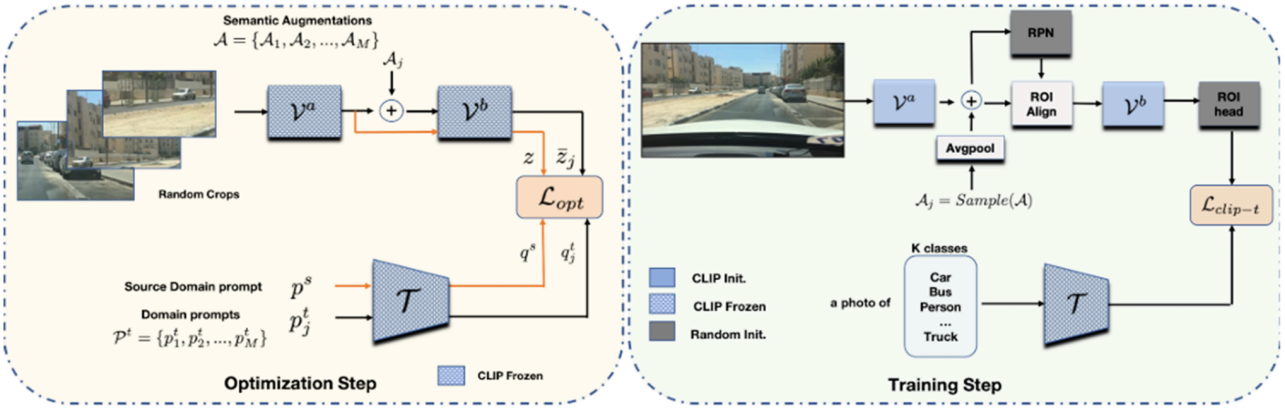


Thay vì dùng một prompt duy nhất để mô tả một miền (“ảnh có sương mù”), sử dụng một **danh sách các prompt đa dạng** để mô tả cùng một miền, ví dụ:

Sương mù = ["một cảnh phố sương mù","một hình ảnh trong điều kiện tầm nhìn kém", "một cảnh bị sương mù che khuất"]

2. Xây dựng mô hình

Hai phần chính: tối ưu hóa và training



- **Tối ưu (bên trái):**
- Dùng ảnh từ miền nguồn và các prompt văn bản về điều kiện thời tiết/ngày đêm.
- Tính vector dịch chuyển đặc trưng trong không gian CLIP.
- Tối ưu A□ sao cho đặc trưng ảnh sau tăng cường gần với đặc trưng miền mục tiêu giả lập.

Mục Tiêu: Tạo ra các vector tăng cường đặc trưng **Aj = {A1, A2,..., A□}** từ Prompt giúp mô hình học được sự thay đổi miền (ví dụ: từ ban ngày sang đêm mưa).

- **Training (bên phải):**
- Ảnh đầu vào được trích đặc trưng, rồi **cộng thêm vector A□** để mô phỏng miền mới.
- Sử dụng **prompt văn bản cho từng class (a photo of a {class} vd:car)** để huấn luyện phân loại theo CLIP.
- Kết hợp với RPN + ROI Align để xác định vị trí và CLIP + prompt văn bản xác định loại đối tượng trong ảnh.

Mục Tiêu: Sử dụng các vector tăng cường **Aj** đã được tối ưu để huấn luyện mô hình CLIP phát hiện đối tượng tổng quá hóa sang miền chưa thấy.

KẾT QUẢ MONG ĐỢI

Vì, trong phần **Semantic Augmentation**, phương pháp trước đó đã cho thấy việc sử dụng prompt phù hợp (so với random prompt) mang lại sự cải thiện rõ rệt.

Phương pháp	**mAP trung bình**
Single-DGOD	33.5 %
CLIP + 1 prompt	38,5 %
Đề xuất: CLIP (backbone) + danh sách prompt	**~39.5 - 40.5 %**

Nếu dùng nhiều prompt 5-10 prompt đa dạng cho mỗi miền chưa biết.
Fog = {"a misty road", "low-visibility weather", "a fog-covered city", ...}
⇒ Nắm bắt nhiều sắc thái ngữ nghĩa và làm đa dạng hóa embedding đặc trưng, tăng khả năng tổng quát hóa miền chưa thấy, dự kiến tăng **1.0 - 2.0 mAP,** đặc biệt ở các miền chưa thấy phức tạp như "night rainy".

TÀI LIỆU THAM KHẢO

[1]. Tianyu Wang, Xiaowei Hu, Chi-Wing Fu, Pheng-Ann Heng: "Single-Stage Instance Shadow Detection With Bidirectional Relation Learning". CVPR 2021: 1-11
[1]. Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou: "Harmonizing transferability and discriminability for adapting object detectors". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8869–8878, 2020
[2]. Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. "Domain generalization via model-agnostic learning of semantic features". Advances in Neural Information Processing Systems, 32, 2019
[3]. Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. "Adversarially adaptive normalization for single domain generalization". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8208–8217, 2021
[4]. Aming Wu, Cheng Deng. "Single-domain generalized object detection in urban scenes via cyclic-disentangled self distillation". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 847–856, 2022.
[5]. Vidit, Engilberge, Martin, Salzmann, Mathieu "CLIP the Gap: A Single Domain Generalization Approach for Object Detection". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) June, 2023, no.3219-3229.