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

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

Trương Lê Mỹ Thanh - 240101073
GVHD: **PGS.TS. Lê Đình Duy**

Tóm tắt

- Lớp: CS2205.FEB2025
- Link Github: https://github.com/thanhtruong-project/CS2205.FEB2025
- Link YouTube video: https://youtu.be/Sbd6fxDUuZs

	• Thành viên: Họ và Tên: Trương Lê Mỹ Thanh MSSV: 24010107

Giới thiệu

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 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 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, đòi hỏi yêu cầu về khả năng tổng quát hóa từ một miền duy nhất, gọi là Single Domain Generalization (SDG). 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 trong bài toán phát hiện đối tượng[4].

Vì thế một phương pháp SDG mới được đề xuất, 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**. Nó 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, 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

Phát hiện vật thể (xe, người,...) trong môi trường thời tiết không lí tưởng như: mưa, tối, sương mù, tuyết,...

Ứng dụng trong:	**Khó khăn và thách thức:**
• Hệ thống giám sát an ninh thực tế • Xe tự lái (tránh vật thể) • Robot cứu hộ hoặc do thám	• Thiếu dữ liệu huấn luyện cho môi trường phức tạp • Độ chính xác giảm khi gặp vật thể bị che khuất • Điều kiện ánh sáng kém (ban đêm, mưa, sương mù,...)

Input: Hình ảnh chụp trong điều kiện xấu (ban đêm, mưa, sương,...)

Output: Xác định vị trí và nhận biết vật thể trong hình như: 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 thông tin ngữ nghĩa (semantic information) như mô tả văn bản tương ứng một miền chưa biết để tạo ra các mẫu huấn luyện bổ sung và đa dạng về ngữ ngĩa cho dữ liệu huấn luyện chỉ với 1 miền gốc duy nhất.
- 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, hỗ trợ triển khai tốt trong các hệ thống giám sát giao thông thông minh thực tế hơn

Nội dung và Phương pháp

Sử dụng mô hình **Semantic Augmentation with CLIP**

Tiếp cận khác biệt về cơ bản:

- Tận dụng mô hình được đào tạo trước CLIP (mô hình ngôn ngữ thị giác tự giám sát)
- Tăng cường dữ liệu về mặt ngữ nghĩa bằng cách sử dụng lời nhắc văn bản (prompts).

CLIP xác định rằng "Một hình ảnh chụp vào đêm mưa" có các đặc điểm khác với "Một hình ảnh chụp vào ngày nắng".

Thay vì sửa đổi hình ảnh trực tiếp, Tăng cường ngữ nghĩa biến đổi các đặc điểm của hình ảnh nguồn để phản ánh sự thay đổi này trong không gian đặc điểm không phải không gian hình ảnh (biến đổi cấp độ pixel: blur,..) cho phép thích ứng tốt hơn với các môi trường mới mà không cần sử dụng dữ liệu miền đích.

Nội dung và Phương pháp

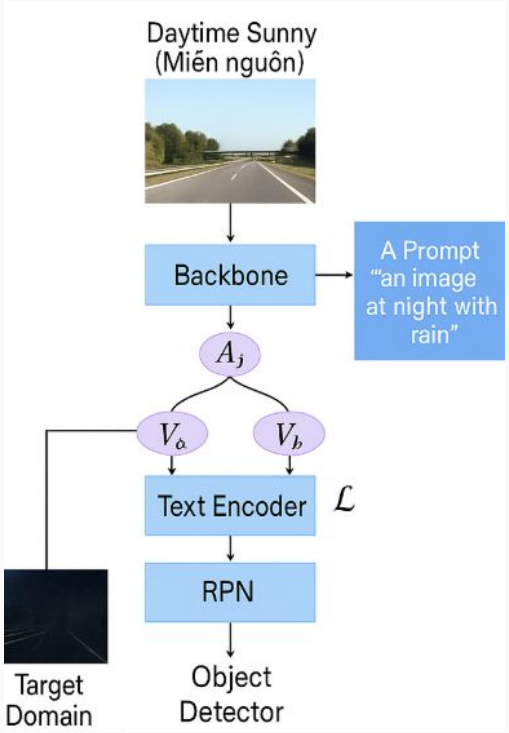
Tăng cường ngữ nghĩa (semantic augmentation)

Dựa vào **embedding không gian chung** của CLIP giữa ảnh và văn bản, các ảnh trong miền source (ví dụ: "ảnh ban ngày trời nắng") sẽ được **dịch chuyển trong không gian đặc trưng** đến các miền mới thông qua 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.

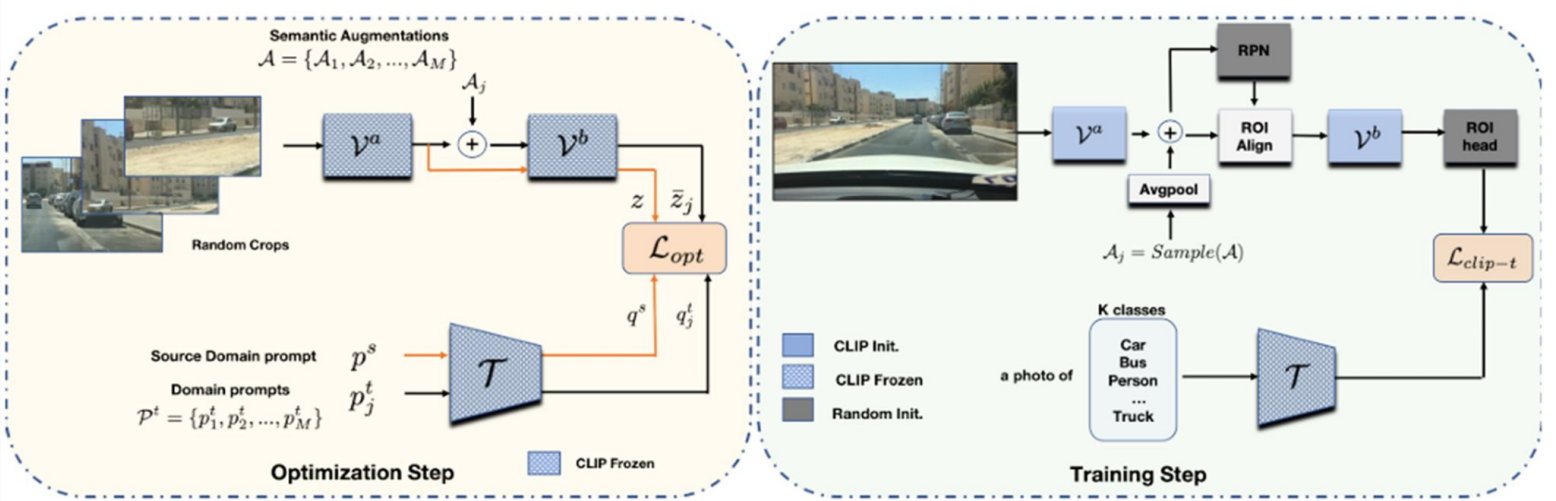
- Thay vì chỉ diễn tả 1 miền với 1 prompt như: “ảnh có sương mù”, dùng **1 danh sách prompt** để mô tả 1 miền:

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



Nội dung và Phương pháp

Hai phần chính: tối ưu hóa và training



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)

- 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.

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

- 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.

Chuẩn bị dữ liệu

- **Training**: 1 miền nguồn duy nhất: ảnh ban ngày, trời nắng
  Ảnh ban ngày: 19,395 ảnh dùng để huấn luyện
  8,313 ảnh khác để chọn mô hình tốt

- **Testing:** 4 miền mục tiêu khác nhau
  Ảnh ban ngày & sương : 3,775 ảnh
  Ảnh đêm & quang đãng : 26,158 ảnh
  Ảnh đêm & mưa : 2,494 ảnh
  Ảnh hoàng hôn & mưa : 3,501 ảnh

Các đối tượng được gắn nhãn để phát hiện: **7 class** (car, bus, truck, motorbike, bike, person, rider)

Kết quả dự kiến

Kết quả trước đó & dự kiến thay đổi

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

Nếu dùng nhiều prompt (prompt ensemble) 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 mAP, đặc biệt ở các miền phức tạp “night rainy”

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.

Tài liệu tham khảo

[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 and Cheng Deng. Single-domain generalized object detection in urban scene 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.