# **THÔNG TIN CHUNG CỦA NHÓM**

* Link YouTube video của báo cáo (tối đa 5 phút):   
  https://youtu.be/4bxITdHNF\_s
* Link slides (dạng .pdf đặt trên Github của nhóm):   
  *https://github.com/thonglb19/CS2205.CH190/blob/main/Th%C3%B4ng%20L%C3%AA%20B%C3%A1%20%20CS2205.CH190.DeCuong.FinalReport.Template.Slide.pdf*
* *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
* *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
* *Lớp Cao học, mỗi nhóm một thành viên*

|  |  |
| --- | --- |
| * Họ và Tên: Lê Bá Thông * MSSV: 240101078 | * Lớp: CS2205.CH190 * Tự đánh giá (điểm tổng kết môn): 7.5/10 * Số buổi vắng: 0 * Số câu hỏi QT cá nhân: 3 * Số câu hỏi QT của cả nhóm: 0 * Link Github: <https://github.com/thonglb19/CS2205.CH190/> |

# **ĐỀ CƯƠNG NGHIÊN CỨU**

|  |
| --- |
| **TÊN ĐỀ TÀI (IN HOA)**  TẠO SINH DỮ LIỆU CÓ KIỂM SOÁT CHO PHÁT HIỆN ĐỐI TƯỢNG ÍT MẪU BẰNG MÔ HÌNH DIFFUSION |
| **TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**  CONTROLLED DATA GENERATION FOR FEW-SHOT OBJECT DETECTION USING DIFFUSION MODELS |
| **TÓM TẮT** *(Tối đa 400 từ)*  Phát hiện đối tượng ít mẫu (few-shot object detection) là một bài toán quan trọng trong thị giác máy tính, nơi số lượng ảnh huấn luyện cho một số lớp đối tượng là rất hạn chế. Điều này ảnh hưởng nghiêm trọng đến hiệu suất của các mô hình học sâu, đặc biệt với các lớp thiểu số. Để khắc phục vấn đề này, các kỹ thuật data augmentation và gần đây là mô hình Diffusion đã được sử dụng nhằm tạo sinh dữ liệu huấn luyện bổ sung. Tuy nhiên, các phương pháp hiện tại vẫn gặp khó khăn trong việc kiểm soát chất lượng, cấu trúc bố cục và tính đa dạng ngữ nghĩa của ảnh sinh ra.  Đề tài này đề xuất một mô hình sinh dữ liệu có kiểm soát, kết hợp giữa Diffusion Models và Graph Neural Networks (GNN). Chúng tôi xây dựng một scene graph từ ba loại điều kiện đầu vào: (1) bố cục hình ảnh (bounding boxes), (2) bản đồ phân đoạn ngữ nghĩa (semantic segmentation maps), và (3) mô tả cảnh chi tiết từ mô hình ngôn ngữ lớn (LLMs). GNN xử lý scene graph để học biểu diễn cấu trúc cảnh, sau đó làm điều kiện cho quá trình tạo sinh ảnh bằng mô hình Diffusion. Ngoài ra, đề tài tích hợp mô hình CLIP để đánh giá mức độ khớp giữa ảnh tạo sinh và mô tả bằng cosine similarity, từ đó chọn lọc ra các ảnh có giá trị huấn luyện cao nhất.  Kết quả kỳ vọng là tăng cường chất lượng và tính đa dạng của dữ liệu huấn luyện, đồng thời cải thiện hiệu suất phát hiện đối tượng trên các bộ dữ liệu benchmark như COCO-FS và PASCAL-FS. |
| **GIỚI THIỆU** *(Tối đa 1 trang A4)*  Phát hiện đối tượng ít mẫu (Few-shot Object Detection – FSOD) là một bài toán đầy thách thức trong lĩnh vực thị giác máy tính, nơi mà mục tiêu là phát hiện chính xác các đối tượng mới chỉ dựa vào một số lượng rất ít mẫu được gán nhãn. Các mô hình học sâu truyền thống thường yêu cầu lượng lớn dữ liệu được gán nhãn đầy đủ để đạt được khả năng khái quát tốt. Tuy nhiên, trong nhiều ứng dụng thực tế như y học, robot, hoặc nhận diện đối tượng hiếm, việc thu thập đủ dữ liệu – đặc biệt là với thông tin vị trí đối tượng (bounding box) – là vô cùng tốn kém và khó khả thi.  Để khắc phục vấn đề này, các kỹ thuật tăng cường dữ liệu (data augmentation) đã được áp dụng rộng rãi nhằm làm giàu tập huấn luyện. Gần đây, các mô hình sinh ảnh – đặc biệt là mô hình khuếch tán (diffusion models) – đã cho thấy tiềm năng trong việc tạo ra hình ảnh chất lượng cao và đa dạng. Tuy vậy, các phương pháp hiện tại vẫn thiếu khả năng kiểm soát cấu trúc ảnh sinh ra, dẫn đến hiện tượng mất tính nhất quán ngữ nghĩa, sai lệch bố cục, hoặc không bám sát mô tả đầu vào, đặc biệt là trong các ngữ cảnh yêu cầu phát hiện đối tượng chính xác.  Đề tài này được thúc đẩy bởi các nghiên cứu gần đây trong lĩnh vực sinh ảnh có điều kiện [1], [2], [3], tuy nhiên các phương pháp hiện tại thường dựa trên bố cục được tạo thủ công hoặc tín hiệu điều kiện yếu, và chưa khai thác đầy đủ cấu trúc quan hệ giữa các đối tượng trong ảnh.  Chúng tôi đề xuất một mô hình tổng hợp dữ liệu có kiểm soát, kết hợp ba thành phần bổ sung lẫn nhau:   * Mạng nơ-ron đồ thị (GNN) để mã hóa cấu trúc cảnh thông qua scene graph được tạo từ bounding boxes, bản đồ phân đoạn ngữ nghĩa và mô tả cảnh từ mô hình ngôn ngữ lớn (LLMs). * Mô hình Diffusion có điều kiện, dùng để sinh ảnh bám sát cấu trúc không gian và ngữ nghĩa. * Hàm đánh giá tự động dựa trên CLIP, nhằm lọc và chọn ra các ảnh sinh ra có mức độ phù hợp cao nhất với ngữ cảnh đầu vào.   Đầu vào của hệ thống bao gồm:   * Bố cục đối tượng (bounding boxes). * Bản đồ phân đoạn ngữ nghĩa (semantic segmentation maps). * Mô tả văn bản chi tiết từ LLMs. (Vd: GPT-3.5-turbo thông qua API của OpenAI)   Đầu ra của hệ thống bao gồm:   * Các ảnh tổng hợp có độ phân giải cao, thể hiện rõ tương tác giữa các đối tượng. * Các nhãn đi kèm (bounding boxes, segmentation) được tạo tự động. * Một tập dữ liệu được lọc bằng CLIP để đảm bảo chất lượng và tính khớp ngữ nghĩa. * Hệ thống này hướng tới việc tạo ra dữ liệu huấn luyện tổng hợp có cấu trúc hợp lý, đa dạng về ngữ cảnh, và phù hợp với nhiệm vụ phát hiện đối tượng ít mẫu, đồng thời có khả năng tích hợp vào các pipeline học máy hiện có như COCO-FS và PASCAL-FS, mang lại giá trị cả về lý thuyết lẫn ứng dụng thực tế. |
| **MỤC TIÊU** *(Viết trong vòng 3 mục tiêu)*  Đề tài nghiên cứu này đóng góp 3 điểm chính như sau:   * Đề xuất một mô hình sinh dữ liệu có kiểm soát, kết hợp giữa mô hình khuếch tán (Diffusion Models) và mạng nơ-ron đồ thị (Graph Neural Network – GNN), nhằm tạo ra các hình ảnh tổng hợp có bố cục hợp lý và tương tác giữa các đối tượng chặt chẽ, phục vụ hiệu quả cho bài toán phát hiện đối tượng ít mẫu. * Xây dựng một biểu diễn cảnh thống nhất (scene graph) từ các đầu vào đa dạng như: bounding boxes (vị trí đối tượng), bản đồ phân đoạn ngữ nghĩa (semantic segmentation maps), và mô tả văn bản từ mô hình ngôn ngữ lớn (LLMs). Scene graph này được GNN xử lý để tạo ra embedding làm điều kiện đầu vào cho quá trình sinh ảnh. * Tích hợp cơ chế lọc ảnh tự động dựa trên mô hình CLIP, nhằm đánh giá mức độ khớp giữa ảnh sinh và mô tả văn bản thông qua phép đo cosine similarity. Cách tiếp cận này giúp chọn lọc ra các ảnh chất lượng cao, đúng ngữ nghĩa và có giá trị huấn luyện tốt hơn cho mô hình phát hiện đối tượng. |
| **NỘI DUNG VÀ PHƯƠNG PHÁP**  NỘI DUNG  Khảo sát chuyên sâu các mô hình hiện tại về few-shot object detection, diffusion models, GNN và CLIP để xây dựng cơ sở lý luận vững chắc cho đề tài.  Chuẩn bị bộ dữ liệu đầu vào gồm bounding boxes, segmentation maps và mô tả ngữ cảnh bằng LLM, sau đó tổ chức thành scene graph có cấu trúc rõ ràng.  Áp dụng GNN để học biểu diễn cấu trúc cảnh từ scene graph, tạo ra embedding đại diện cho bố cục và ngữ nghĩa cảnh vật.  Tích hợp embedding từ GNN làm điều kiện vào mô hình diffusion, huấn luyện để sinh ảnh có bố cục và nội dung phù hợp với đầu vào.  Sử dụng mô hình CLIP để đánh giá mức độ khớp ngữ nghĩa giữa ảnh sinh và mô tả đầu vào, từ đó lọc giữ các ảnh có chất lượng cao.  Tạo ra tập dữ liệu tổng hợp và sử dụng để huấn luyện các mô hình phát hiện đối tượng ít mẫu hiện có, đánh giá lại hiệu quả và so sánh với các baseline.  PHƯƠNG PHÁP  Để đạt được các mục tiêu nghiên cứu đã đề ra, đề tài được triển khai theo 5 giai đoạn chính. Mỗi giai đoạn tập trung vào một phần quan trọng trong quy trình xây dựng hệ thống sinh dữ liệu có kiểm soát cho bài toán phát hiện đối tượng ít mẫu.  Giai đoạn 1: Khảo sát và xây dựng cơ sở lý thuyết  Trong giai đoạn này, nhóm nghiên cứu tiến hành thu thập và phân tích các tài liệu liên quan đến bài toán few-shot object detection, các phương pháp tăng cường dữ liệu hiện đại (đặc biệt là diffusion models), mạng nơ-ron đồ thị (GNN), và mô hình CLIP. Các công trình tiêu biểu như của Fang et al. (2024), Abdullah et al. (2024) và Sridhar et al. (2024) sẽ được tham khảo kỹ lưỡng. Đồng thời, nhóm cũng xác định bộ dữ liệu sử dụng trong thực nghiệm, chẳng hạn như COCO-FS và PASCAL-FS.  Giai đoạn 2: Tiền xử lý và xây dựng scene graph  Giai đoạn này tập trung xây dựng quy trình tiền xử lý đầu vào, bao gồm: chuẩn hóa bounding boxes, bản đồ phân đoạn ngữ nghĩa (semantic segmentation maps), và sinh mô tả cảnh bằng mô hình ngôn ngữ lớn (LLM). Từ các dữ liệu này, một đồ thị cảnh (scene graph) sẽ được tạo ra, trong đó các nút đại diện cho đối tượng và các cạnh thể hiện mối quan hệ không gian hoặc ngữ nghĩa. Mạng GNN sẽ được huấn luyện để học biểu diễn embedding của cảnh từ đồ thị này.  Giai đoạn 3: Tích hợp mô hình Diffusion có điều kiện  Embedding từ GNN sẽ được sử dụng để điều kiện hóa mô hình khuếch tán (Diffusion Model), giúp kiểm soát quá trình sinh ảnh từ noise ban đầu. Mô hình diffusion có thể được xây dựng dựa trên các kiến trúc như UNet hoặc tích hợp ControlNet để điều kiện hóa đầu vào phức tạp. Giai đoạn này sẽ kiểm thử quá trình sinh ảnh với các điều kiện đầu vào để đảm bảo tính logic về bố cục và nội dung.  Giai đoạn 4: Tích hợp và đánh giá bằng mô hình CLIP  Các ảnh sinh ra từ diffusion sẽ được đánh giá chất lượng thông qua mô hình CLIP bằng cách mã hóa cả ảnh và mô tả văn bản vào cùng một không gian embedding. Sau đó, cosine similarity được tính để đo mức độ khớp giữa ảnh và mô tả. Chỉ những ảnh có độ tương đồng cao vượt qua ngưỡng định sẵn mới được giữ lại để xây dựng tập dữ liệu huấn luyện cuối cùng.  Giai đoạn 5: Huấn luyện và đánh giá mô hình FSOD  Tập dữ liệu được tổng hợp từ ảnh sinh và được chọn lọc bằng CLIP sẽ được sử dụng để huấn luyện các mô hình phát hiện đối tượng ít mẫu hiện có (ví dụ: Meta R-CNN, TFA). Hiệu quả của mô hình sẽ được đánh giá lại trên các tập kiểm thử chuẩn. Kết quả này sẽ được phân tích và so sánh với các phương pháp baseline không dùng dữ liệu sinh. Cuối cùng, nhóm sẽ hoàn thiện báo cáo nghiên cứu và trình bày kết quả. |
| **KẾT QUẢ MONG ĐỢI**  Dựa trên đề xuất nghiên cứu, chúng tôi kỳ vọng đạt được các kết quả và tiến triển sau, nhằm đánh giá hiệu quả của phương pháp tiếp cận:  Cải thiện Hiệu suất Phát hiện Đối tượng Thiểu số (FSOD):   * Kỳ vọng gia tăng đáng kể độ chính xác (mAP) của các mô hình FSOD trên các tập dữ liệu tiêu chuẩn (COCO-FS, PASCAL-FS) so với các phương pháp baseline. Sự cải thiện này sẽ là bằng chứng trực tiếp cho thấy hiệu quả của việc tăng cường dữ liệu.   Xây dựng quy trình tạo dữ liệu tổng hợp có kiểm soát và thể hiện khả năng tạo sinh ảnh mong muốn   * Phát triển quy trình hiệu quả để tạo ảnh tổng hợp chất lượng cao, có tính logic về bố cục và ngữ nghĩa. Quy trình này sẽ thể hiện khả năng sinh ảnh đa dạng và phù hợp với điều kiện đầu vào (được đánh giá qua phân tích hình ảnh và các thước đo chất lượng ảnh tạo sinh).   Tạo lập Data Augmentation chất lượng cao và đánh giá mức độ đóng góp của nó   * Xây dựng một tập dữ liệu huấn luyện mở rộng, kết hợp dữ liệu thực và ảnh sinh từ Diffusion Models. Tập dữ liệu này sẽ được đánh giá và chọn lọc bởi CLIP để đảm bảo chất lượng, và chứng minh đóng góp vào hiệu suất huấn luyện mô hình FSOD qua các kết quả so sánh.   Chứng minh hiệu quả của GNN và CLIP trong quy trình tạo và đánh giá dữ liệu   * Cung cấp bằng chứng thực nghiệm cho thấy GNN hiệu quả trong việc điều kiện hóa Diffusion Models để tạo dữ liệu tổng hợp có kiểm soát, và CLIP đóng vai trò then chốt trong việc đánh giá, lọc dữ liệu tạo sinh, tối ưu hóa đầu vào cho FSOD. |
| **TÀI LIỆU THAM KHẢO** *(Định dạng DBLP)*  [1] H. Fang, B. Han, S. Zhang, S. Zhou, C. Hu, and W.-M. Ye, “Data Augmentation for Object Detection via Controllable Diffusion Models,” 2024.  [2] A. Abdullah, N. Ebert, and O. Wasenmüller, “Boosting Few-Shot Detection with Large Language Models and Layout-to-Image Synthesis,” 2024. [Online]. Available: https://link.springer.com/conference/accv  [3] D. Sridhar, A. Peri, R. Rachala, and N. Vasconcelos, “Adapting Diffusion Models for Improved Prompt Compliance and Controllable Image Synthesis,” 2024. |