

ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÀI LUẬN VỀ HƯỚNG NGHIÊN CỨU TUYỂN SINH TRÌNH ĐỘ TIẾN SĨ

Họ tên người dự tuyển: NGUYỄN THÀNH DANH

Ngành dự tuyển: Khoa học Máy tính

Mã số: 9 48 01 01

Tên đề tài:

NGHIÊN CỨU PHƯƠNG PHÁP PHÂN ĐOẠN ẢNH TRONG ĐIỀU KIỆN ÍT DỮ LIỆU HUẤN LUYỆN

Người hướng dẫn chính: TS. Nguyễn Vinh Tiệp

Cơ quan công tác: Trường Đại học Công nghệ Thông tin, ĐHQG TP.HCM

Người hướng dẫn phụ: PGS. TS. Nguyễn Văn Tâm

Cơ quan công tác: Trường Đại học Dayton, Dayton, OH 45469, Hoa Kỳ

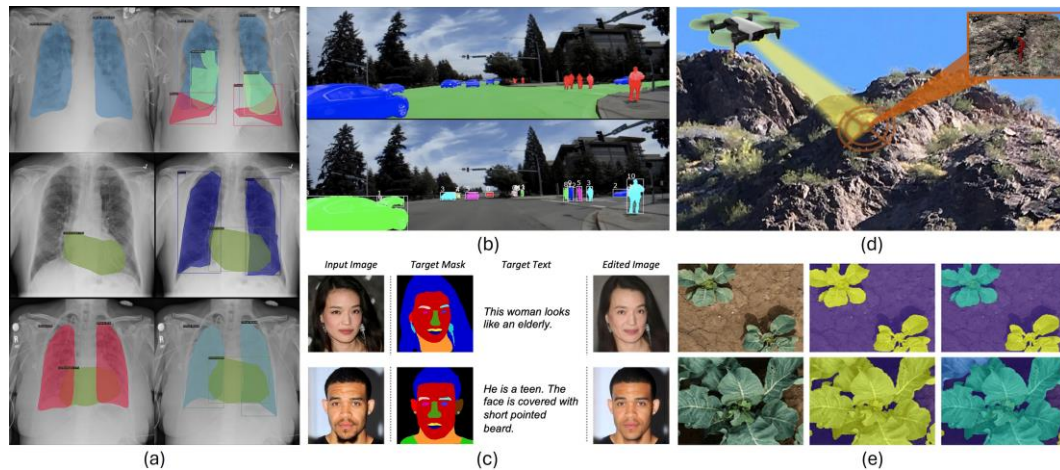
Tp. Hồ Chí Minh, tháng 05 năm 2024

Nội dung

1. Giới thiệu bài toán.....	1
2. Mục tiêu nghiên cứu.....	3
3. Tổng quan tình hình nghiên cứu bài toán phân đoạn ảnh	4
4. Nội dung nghiên cứu.....	7
4.1 Nội dung 1: Giải quyết vấn đề thiếu dữ liệu gán nhãn phân đoạn	7
4.2 Nội dung 2: Hướng tiếp cận phân đoạn ảnh dựa trên mô hình tạo sinh	9
4.3 Nội dung 3: Hướng tiếp cận sử dụng ít dữ liệu huấn luyện.....	11
5. Một số kết quả sơ khởi.....	12
6. Lí do thực hiện đề tài	13
7. Kế hoạch thực hiện	14
8. Tài liệu tham khảo	14

1. Giới thiệu bài toán

Bối cảnh thực tiễn. Phân đoạn ảnh trong lĩnh vực thị giác máy tính có những ứng dụng tiềm năng trong thực tế nhờ vào khả năng phân tích hình ảnh với mức độ chi tiết cao ở cấp độ điểm ảnh. Các kết quả của bài toán phân đoạn ảnh có thể đóng vai trò trung gian hỗ trợ thực thi các bài toán khác. Đơn cử một số ngữ cảnh ứng dụng của bài toán phân đoạn ảnh được thể hiện trong các lĩnh vực như y khoa, xe tự hành, nghệ thuật, nông nghiệp, an ninh hay quân sự (*minh họa tại Hình 1*).

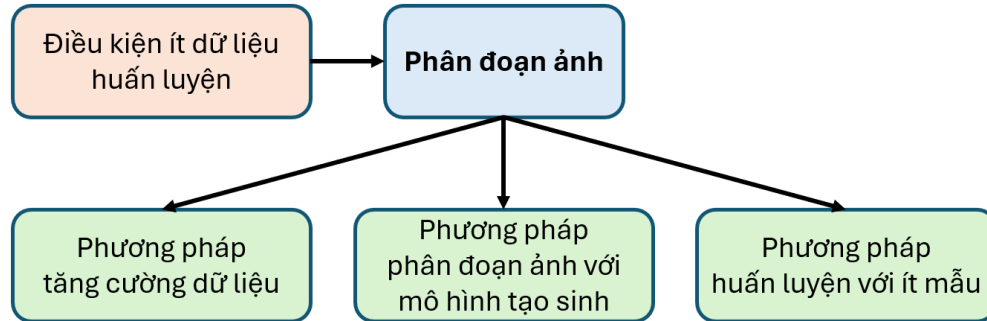


Hình 1. Một số ứng dụng của bài toán phân đoạn ảnh

- Y khoa: Phân đoạn ảnh được sử dụng để phân tích ảnh y khoa, xác định các vùng ảnh nghi ngờ có bệnh, đóng vai trò như một hệ thống khuyến nghị và hỗ trợ ra quyết định giúp các bác sĩ tiết kiệm được thời gian và đưa ra chẩn đoán chính xác hơn. Phân đoạn ảnh không những chỉ ra được các vùng nghi ngờ có bệnh mà còn cho biết được hình dạng của các vùng bệnh này (*Hình 1-a*).
- Xe tự hành: Phân đoạn ảnh là một trong những mô hình thị giác máy tính phục vụ hệ thống hỗ trợ lái thông minh (ADAS), thực thi các tác vụ nhận diện biển báo giao thông, xác định các phương tiện giao thông, làn đường. Mô hình phân đoạn ảnh quan sát môi trường xung quanh và đưa ra cảnh báo, khuyến nghị hoặc can thiệp cần thiết vào quá trình điều khiển phương tiện (*Hình 1-b*).
- Ứng dụng nghệ thuật: Phân đoạn ảnh là bước trung gian của ứng dụng làm đẹp hoặc chỉnh sửa ảnh tự động, cung cấp thông tin ngữ nghĩa và vị trí của các đối tượng trong ảnh, giúp các hệ làm đẹp ảnh tự động thực thi tác vụ chỉnh sửa chính xác, phù hợp và có tính thẩm mỹ hơn (*Hình 1-c*).
- Bên cạnh đó, bài toán phân đoạn ảnh còn có nhiều ứng dụng khác trong các lĩnh vực như quân sự, an ninh với các tác vụ tìm kiếm và giải cứu (*Hình 1-d*), hay phục vụ nông nghiệp (*Hình 1-e*).

Tuy nhiên, sự phát triển của các mô hình học sâu hiện đại đòi hỏi các mô hình phân đoạn tiên tiến cần được huấn luyện với nhiều dữ liệu gán nhãn. Điều này đặt ra các vấn đề liên quan như không đủ dữ liệu huấn luyện, hay tốn kém chi phí gán nhãn, đặc biệt với bài toán cần đến mức độ chi tiết cao ở cấp độ điểm ảnh. Trong đề tài này, ứng viên tập trung phân tích và đưa ra các hướng giải quyết tiềm năng cho các vấn đề trên.

Định nghĩa bài toán. Các bài toán cơ bản trong lĩnh vực thị giác máy tính được biết đến như phân loại ảnh, phát hiện đối tượng hay phân đoạn ảnh đều hướng đến mục tiêu hiểu nội dung của dữ liệu đầu vào thông qua việc dự đoán nhãn. Sự khác biệt là mức độ chi tiết của ngữ nghĩa tăng dần qua các tác vụ kể trên. Bài toán phân đoạn ảnh được định nghĩa với **đầu vào** là ảnh và **đầu ra** là nhãn ngữ nghĩa của từng điểm ảnh trong ảnh đó. Như vậy, mô hình phân đoạn ảnh cần dự đoán nhãn ngữ nghĩa ở mức độ chi tiết từng điểm ảnh của ảnh đầu vào. Như đã đề cập ở trên, trong phạm vi đề tài này, ứng viên giải quyết bài toán phân đoạn ảnh trong điều kiện đặc thù hạn chế về dữ liệu huấn luyện do chi phí gán nhãn cao hay không đủ dữ liệu gán nhãn chi tiết. Vì vậy, các hướng tiếp cận tiềm năng được đề xuất như sau: (1) Sử dụng hướng tiếp cận tăng cường dữ liệu, (2) Sử dụng phương pháp phân đoạn dựa trên mô hình tạo sinh ảnh, và (3) Sử dụng phương pháp học với ít dữ liệu huấn luyện. *Sơ đồ 1* thể hiện tổng quan phương pháp phân đoạn ảnh trong điều kiện ít dữ liệu huấn luyện mà ứng viên đề xuất.



Sơ đồ 1. Tổng quan đề xuất giải quyết bài toán phân đoạn ảnh trong điều kiện ít dữ liệu huấn luyện.

Hướng tiếp cận. Cũng giống như các bài toán cơ bản của thị giác máy tính, các hướng tiếp cận thường được chia thành (1) phương pháp sử dụng đặc trưng cấp thấp, và (2) phương pháp sử dụng đặc trưng học sâu. Hướng tiếp cận (1) bao gồm những phương pháp được đề xuất vào giai đoạn đầu khi bài toán được đề xuất như phân đoạn dựa trên biên cạnh, ngưỡng, vùng, phương pháp gom cụm hay watershed. Các phương pháp này chỉ dựa trên đặc trưng cấp thấp, vì vậy, chúng chưa khai thác được các mối quan hệ về đặc trưng ngữ nghĩa cấp cao để phục vụ quá trình phân đoạn. Hướng tiếp cận sử dụng đặc trưng học sâu (2) khắc phục được vấn đề này và đã có nhiều thành tựu. Nếu xét về khía cạnh kiến trúc mô hình, hướng tiếp cận này có các nhánh mô hình được thiết kế theo hướng 1 giai đoạn [11-16], và mô hình theo hướng 2 giai đoạn [2-9]. Xét về khía

cạnh phương pháp học, có thể chia thành các mô hình dựa trên cơ chế tạo sinh [17-19] (generative) và các mô hình dựa trên cơ chế phân biệt [2-9], [11-16] (discriminative). Trong đề tài này, ứng viên lựa chọn khai thác các mô hình học sâu với cơ chế tạo sinh để phân đoạn ảnh. Hướng tiếp cận này giúp khai thác thông tin từ các mô hình được huấn luyện sẵn với lượng dữ liệu lớn vốn được sử dụng cho tác vụ tạo sinh ảnh để giải quyết bài toán phân đoạn ảnh.

Các thách thức. Phân đoạn ảnh là một trong những bài toán cơ bản của thị giác máy tính, do đó, bài toán này chia sẻ một số thách thức chung của lĩnh vực liên quan đến sự đa dạng của dữ liệu ảnh cũng như giới hạn về phần cứng trong quá trình huấn luyện mô hình. Bên cạnh đó, bài toán này còn đối diện với các thách thức liên quan đến dữ liệu gán nhãn và thách thức liên quan đến phương pháp giải quyết bài toán.

- Dữ liệu ảnh có những thách thức đặc thù liên quan đến sự đa dạng của chúng. Các mô hình thị giác máy tính nói chung và mô hình phân đoạn ảnh nói riêng được huấn luyện với dữ liệu ảnh gặp phải các thách thức liên quan đến sự đa dạng của ảnh như màu sắc, độ sáng, chất lượng ảnh kém, mờ, nhòe hay các đối tượng trong ảnh có nhiều biến thể, bị che khuất, hay không toàn vẹn. Trong đề tài này, ứng viên tiếp cận giải quyết vấn đề này với phương pháp tăng cường dữ liệu, đặc biệt với hướng tiếp cận tạo sinh dữ liệu.
- Với bài toán phân đoạn ảnh, chi phí gán nhãn chi tiết cho dữ liệu luôn là một vấn đề lớn. Việc gán nhãn chi tiết đến từng điểm ảnh gây khó khăn cho người gán nhãn, tốn nhiều thời gian và công sức. Mặc dù đã có các công cụ hỗ trợ gán nhãn, song vẫn cần đến sự hiệu chỉnh thủ công của con người để xác nhận. Khó khăn này đặt ra nhu cầu giải quyết công việc gán nhãn một cách tự động, hoặc khai thác hiệu quả các nhãn dữ liệu đã được gán nhãn sẵn để giải quyết vấn đề thiếu dữ liệu huấn luyện và chi phí gán nhãn cao.
- Các mô hình phân đoạn dựa trên học sâu với kiến trúc CNN đã chứng minh được tính hiệu quả của chúng. Tuy nhiên, các phương pháp này cũng thể hiện hạn chế khi gặp phải vấn đề thiên kiến quy nạp (inductive bias) trong quá trình tạo ra các vùng đề xuất (RPN) hoặc điểm neo (anchor) dẫn đến giới hạn độ chính xác của mô hình phân đoạn. Điều này đặt ra nhu cầu nghiên cứu các phương pháp để giải quyết bài toán phân đoạn ảnh, vượt qua hạn chế của hướng tiếp cận này.

2. Mục tiêu nghiên cứu

Với các thách thức nói trên, mục tiêu tổng quát của đề tài là nghiên cứu, đề xuất các mô hình, các phương pháp tiên tiến giải quyết bài toán phân đoạn ảnh trong điều kiện ít dữ liệu huấn luyện. Các phương pháp đề xuất hướng đến ba mục tiêu chính: (1) nghiên cứu phương pháp tăng cường dữ liệu phục vụ cho bài toán phân đoạn ảnh, (2) nghiên cứu, đề xuất phương pháp phân đoạn ảnh với hướng tiếp cận dựa trên cơ chế tạo sinh, và (3)

nguyên cứu, đề xuất phương pháp khai thác các mẫu dữ liệu gắn nhãn với hướng tiếp cận sử dụng ít dữ liệu huấn luyện.

1. Nghiên cứu, đề xuất phương pháp tăng cường dữ liệu cho phép khai thác các nhãn mặt nạ phân đoạn được gán sẵn. Mục tiêu tăng cường độ đa dạng cho dữ liệu sẵn có và khai thác các mô hình tạo sinh để làm giàu dữ liệu, giúp giải quyết thách thức về dữ liệu gắn nhãn.
2. Nghiên cứu, đề xuất phương pháp phân đoạn ảnh với hướng tiếp cận tạo sinh cho phép khai thác các mô hình tạo sinh được huấn luyện với số lượng lớn dữ liệu. Việc xem xét bài toán phân đoạn dưới góc nhìn của bài toán tạo sinh cho phép khai thác tri thức từ các mô hình tạo sinh để dự đoán mặt nạ phân đoạn, giải quyết vấn đề thiên kiến quy nạp của các mô hình CNN trước đây.
3. Nghiên cứu, đề xuất phương pháp học với ít dữ liệu huấn luyện để phân đoạn ảnh. Dựa trên quan sát thực tế về quá trình học hỏi tri thức mới của con người, phương pháp huấn luyện sử dụng ít mẫu dữ liệu được đề xuất dựa trên cơ chế dùng tri thức để tiếp thu cái mới. Phương pháp này cho phép mô hình khai thác tri thức tổng quát với các lớp dữ liệu phổ biến để tiến hành phân đoạn với ít dữ liệu gắn nhãn.

3. Tổng quan tình hình nghiên cứu bài toán phân đoạn ảnh

Như đã đề cập ở phần trước, các công trình nghiên cứu giải quyết bài toán phân đoạn ảnh có thể được chia thành 2 hướng tiếp cận chính: dựa trên đặc trưng cấp thấp và dựa trên đặc trưng học sâu. Trong đó, hướng tiếp cận dựa trên đặc trưng học sâu được tập trung phát triển gần đây với các mô hình với kiến trúc 2 giai đoạn hay 1 giai đoạn, và dựa trên cơ chế tạo sinh với các mô hình tạo sinh đối nghịch hay các mô hình khuếch tán. Đặc biệt, trong ngữ cảnh ít dữ liệu huấn luyện, hướng tiếp cận huấn luyện với ít dữ liệu gắn nhãn được đề xuất giải quyết đặc thù trường hợp này. Sau đây, ứng viên trình bày và đưa ra một số thảo luận về các hướng nghiên cứu giải quyết bài toán phân đoạn ảnh hiện nay.

Hướng tiếp cận dựa trên mô hình 2 giai đoạn và 1 giai đoạn.

Các mô hình 2 giai đoạn. Với hướng tiếp cận 2 giai đoạn, các công trình sử dụng một quy trình chung gồm hai bước là phát hiện các vùng quan tâm (ROI) và phân đoạn để tạo ra mặt nạ ngữ nghĩa dựa trên từng khung bao phát hiện được [1]. Các công bố tiêu biểu cho hướng tiếp cận này có thể kể đến như: Mask RCNN [2], Mask Scoring RCNN [3], Cascade Mask RCNN [4], PANet [5], HTC [6], BlendMask [7], Mask Transfiner [8] hay DCNet [9]. Điểm chung của các công trình này là sử dụng một giai đoạn tìm kiếm các vùng đề xuất, sau đó tiến hành phân đoạn trên các vùng đề xuất này với các đề xuất đặc thù của từng phương pháp.

Ưu điểm:

- Kiến trúc tường minh, dễ hiểu: các mô hình 2 giai đoạn được cấu thành từ các mô-đun tác vụ riêng lẻ, do đó, kiến trúc toàn bộ framework rõ ràng và dễ hiểu, dễ giải thích các thức hoạt động.
- Khả năng tùy biến cao: các mô hình 2 giai đoạn cho phép tùy biến bằng cách thêm vào hoặc chỉnh sửa các mô-đun trong kiến trúc mạng một cách dễ dàng mà vẫn có thể tái sử dụng các mô-đun khác trong kiến trúc mạng.
- Độ chính xác cao: các mô hình thuộc nhánh 2 giai đoạn đạt độ chính xác phân đoạn cao trên các tập dữ liệu chuẩn. Vì các tác vụ này được chuyên biệt thành các giai đoạn riêng biệt, mỗi giai đoạn được huấn luyện tối ưu để thực thi từng tác vụ.

Nhược điểm:

- Kiến trúc mô hình cồng kềnh: tuy các mô-đun trong mô hình tường minh nhưng mỗi mô-đun thường đảm nhiệm các tác vụ riêng lẻ, do đó, mô hình cần sử dụng nhiều mô-đun khác nhau dẫn đến số lượng tham số và kích thước mô hình chưa được tối ưu.
- Tốn thời gian thực thi: thời gian thực thi thường tốn nhiều hơn so với các mô hình thuộc hướng tiếp cận 1 giai đoạn. Hướng tiếp cận 2 giai đoạn cần xác định thông tin và phân tích các vùng đề xuất, sau đó mới tiến hành phân đoạn trên các vùng đề xuất này.

Các mô hình 1 giai đoạn. Với hướng tiếp cận một giai đoạn, các mô hình có đặc điểm chung là sử dụng hướng tiếp cận phát hiện đối tượng không sử dụng điểm neo (anchor-free detection) và một số phương pháp tiên tiến gần đây sẽ dựa trên kiến trúc Transformer [10] để xây dựng mô hình. Tiêu biểu cho các phương pháp này có thể kể đến như: YOLACT [11], CondInst [12], SOTR [13], SOLO [14], OSFormer [15], hay SAM [16]. Sau đây là một số nét đặc trưng của các phương pháp trong hướng tiếp cận một giai đoạn để phân đoạn thực thể.

Ưu điểm:

- Thời gian thực thi nhanh: các mô hình thuộc nhánh 1 giai đoạn không dựa trên giai đoạn tìm kiếm các vùng đề xuất để phân đoạn nên tiết kiệm được thời gian thực thi, phù hợp với các ứng dụng cần đến yếu tố thời gian thực thi nhanh hoặc real-time.
- Độ chính xác tương đối tốt: so với các mô hình 2 giai đoạn, độ chính xác của các mô hình 1 giai đoạn còn hạn chế. Tuy nhiên, độ chính xác này là không chênh lệch quá nhiều và chấp nhận được tùy theo các yêu cầu đặt ra của mỗi bài toán.

Nhược điểm:

- Khó tùy biến: so với các mô hình 2 giai đoạn, các mô hình 1 giai đoạn phân tích từ điểm ảnh để tạo ra bản đồ phân đoạn. Do vậy, các mô-đun thường đảm nhiệm

nhiều nhiệm vụ khác nhau khiến cho quá trình tinh chỉnh trên từng mô-đun phức tạp hơn.

- Độ chính xác hạn chế: độ chính xác kém hơn so với các mô hình khác thuộc nhánh 2 giai đoạn tiến hành phân đoạn dựa trên phân tích từ điểm ảnh, không được hỗ trợ và giới hạn không gian bởi các vùng đề xuất.

Hướng tiếp cận dựa trên cơ chế tạo sinh (Generative Approach).

Các mô hình tạo sinh gần đây đã gây được tiếng vang với những ứng dụng tạo sinh ảnh, chỉnh sửa ảnh và video đạt kết quả chân thật và thu hút trên mạng xã hội. Các mô hình này được huấn luyện trên một lượng lớn dữ liệu và có kết hợp cả những mô hình ngôn ngữ lớn để đạt được kết quả nói trên. Thông thường, hướng tiếp cận sử dụng các mô hình tạo sinh được sử dụng để tạo ra dữ liệu (ảnh, video, văn bản...). Tuy nhiên, trong những năm gần đây, cộng đồng nghiên cứu nhận thấy tiềm năng có thể phát triển các mô hình ở nhánh này để phục vụ các tác vụ phân biệt (discriminative task), đơn cử trong đề tài này là bài toán phân đoạn ảnh [36]. Hiện nay đã có một số nghiên cứu sơ khởi để sử dụng các mô hình tạo sinh ảnh phục vụ phân đoạn ảnh dựa trên GAN và cơ chế khuếch tán Diffusion. Đơn cử cho các hướng tiếp cận này là các phương pháp dựa trên GANs [19] và DiffusionInst [17], hay SEMI-DiffusionInst [18].

Ưu điểm:

- Hạn chế về kiến trúc mạng: các mô hình hiện nay bị giới hạn bởi việc giải mã đặc trưng dựa trên các vùng đề xuất từ mô-đun đề xuất vùng quan tâm mà chưa tận dụng được thông tin từ toàn bộ ảnh, dẫn đến giới hạn về khả năng giải mã tạo ra bản đồ phân đoạn.
- Độ chính xác cao: các mô hình thuộc hướng tiếp cận tạo sinh đạt độ chính xác cao khi so sánh trên các tập dữ liệu chuẩn vì ngoài yếu tố được xây dựng dựa trên các mô hình 2 giai đoạn, các mô hình này còn tận dụng được các mô hình được huấn luyện sẵn, các mô hình ngôn ngữ phục vụ cho các tác vụ tạo sinh ảnh.

Nhược điểm:

- Tốn nhiều thời gian huấn luyện và suy luận: thời gian huấn luyện và thực thi tốn nhiều hơn so với các mô hình thuộc hướng tiếp cận phân biệt. Để đạt được độ chính xác cao, mô hình đòi hỏi lặp lại nhiều lần bước giải mã (denoising).
- Kiến trúc chưa tối ưu: các mô hình tạo sinh vốn được tối ưu cho tác vụ tạo sinh ảnh, khi đưa vào giải quyết tác vụ phân đoạn ảnh, quá trình khử nhiễu (noise-to-filter) chưa được tối ưu hoàn toàn mà vẫn dựa trên thông tin được huấn luyện sẵn từ mô hình tạo sinh.

Hướng tiếp cận sử dụng ít dữ liệu huấn luyện (Few-shot Learning).

Như đã đề cập ở phần trước, thiếu dữ liệu gán nhãn là một trong những thách thức lớn của không chỉ riêng bài toán phân đoạn ảnh, mà còn là thách thức chung của lĩnh vực

học máy. Để giải quyết vấn đề này, cộng đồng nghiên cứu đã đề xuất cách huấn luyện các mô hình sử dụng ít dữ liệu gán nhãn, gọi là few-shot learning. Phương pháp này được phân loại thuộc nhánh meta-learning (học phương pháp học). Theo đó, mô hình phân đoạn ảnh học phương pháp để học phân đoạn trên một tập dữ liệu huấn luyện, sau đó đưa ra dự đoán phân đoạn trên một tập dữ liệu khác, gọi là unrelated sets. Các mô hình tiêu biểu cho hướng tiếp cận này có thể kể đến như: Meta R-CNN [20], CRNet [21], PANet [22], iFS R-CNN [23], DCFS [24], hay DTN [25]. Điểm chung của các mô hình này đều dựa trên các kiến trúc được đề xuất cho bài toán học có giám sát, sau đó tinh chỉnh và xây dựng pipeline cho ngữ cảnh học phương pháp học để phục vụ các bài toán con với ít dữ liệu huấn luyện.

Ưu điểm:

- Giải quyết được tác vụ phân đoạn ảnh sử dụng ít dữ liệu gán nhãn cho trước. Đây chính là mục tiêu đặt ra ban đầu của hướng tiếp cận này, giải quyết phần nào thách thức về dữ liệu gán nhãn.
- Khai thác các mô hình huấn luyện sẵn: ở giai đoạn cơ sở, hướng tiếp cận này cho phép khai thác thông tin và tri thức được huấn luyện trên miền dữ liệu tổng quát của các kiến trúc mạng, giúp tiết kiệm thời gian huấn luyện cho tác vụ trên các miền dữ liệu khác.

Nhược điểm:

- Độ chính xác còn hạn chế ở các mẫu có số lượng rất ít (1 mẫu, hay 2 mẫu): độ chính xác của bài toán phân đoạn với ít dữ liệu huấn luyện được đánh giá trên nhiều cấu hình khác nhau, thông thường, độ chính xác đánh giá tại các cấu hình rất ít dữ liệu còn hạn chế do thông tin chưa được khai thác tốt từ số lượng ít dữ liệu này. Với nhiều mẫu dữ liệu hơn (3 mẫu, hay 5 mẫu trở lên), độ chính xác được cải thiện. Tuy nhiên, đánh đổi kèm theo là bộ nhớ lưu trữ phục vụ quá trình huấn luyện.

4. Nội dung nghiên cứu

4.1 Nội dung 1: Giải quyết vấn đề thiếu dữ liệu gán nhãn phân đoạn

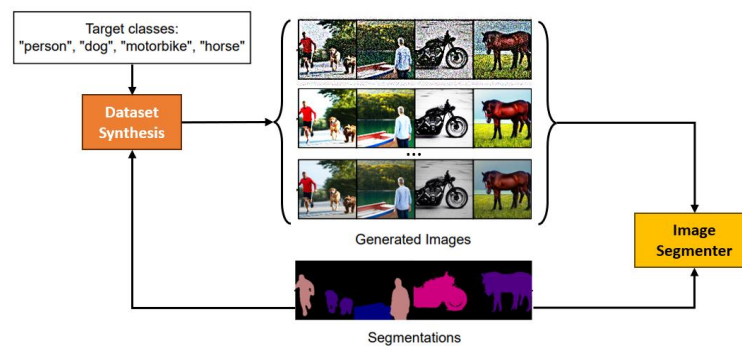
Động lực nghiên cứu: Trong lĩnh vực thị giác máy tính nói riêng và lĩnh vực học sâu nói chung, dữ liệu gán nhãn phục vụ quá trình huấn luyện luôn là vấn đề tiêu tốn nhiều chi phí. Đặc biệt với tác vụ phân đoạn ảnh, nhãn dữ liệu cần được gán chi tiết đến cấp độ điểm ảnh. Cụ thể, mỗi điểm ảnh cần được gán một nhãn thể hiện lớp ngữ nghĩa mà nó thuộc về. Trước đây, quá trình này được thực hiện thủ công và tốn nhiều thời gian và nhân lực. Ngày nay, các công cụ hỗ trợ đã được phát triển phục vụ việc gán nhãn, có thể kể đến như LabelMe, V7Labs, hay DataLoop. Tuy nhiên, việc gán nhãn dữ liệu được hỗ trợ bởi các công cụ này vẫn mang xu hướng thủ công hoặc bán thủ công, nghĩa là vẫn sử dụng khả năng của con người để gán nhãn dữ liệu.

Bên cạnh đó, chúng ta đang chứng kiến sức mạnh vượt trội của các công cụ tạo sinh ảnh sử dụng các mô hình học sâu như các mô hình GAN hay các mô hình sử dụng cơ chế khuếch tán (Diffusion [26]). Các mô hình này đã đạt được những thành tựu nhất định trong việc tạo ra hình ảnh hay video nhân tạo, có thể kể đến như SoRA [27], Mid Journey [28], DALL-E2 [29], ImageGen [30], GliGen [37] hay Mini-GPT4 [31]. Thực trạng này đặt ra câu hỏi, liệu rằng chúng ta có thể tận dụng được sức mạnh của các mô hình tạo sinh kể trên, vốn được huấn luyện đạt độ chính xác cao trên tác vụ tạo sinh ảnh, để tạo ra hình ảnh với điều kiện ràng buộc là mặt nạ phân đoạn để phục vụ bài toán phân đoạn ảnh hay không. Trong thực tế, các công trình này có khả năng tạo ra dữ liệu ảnh, tuy nhiên chưa khai thác yếu tố sử dụng điều kiện là bản đồ phân đoạn để phục vụ tác vụ tăng cường dữ liệu.

Mục tiêu: Trong nội dung này, ứng viên hướng đến việc sử dụng mô hình tạo sinh ảnh như một công cụ giúp tăng cường dữ liệu cho bài toán phân đoạn. Với bối cảnh dữ liệu gán nhãn phân đoạn còn hạn chế và tốn nhiều chi phí để thực hiện, ứng viên đặt giả thiết rằng, việc sử dụng mô hình tạo sinh để tạo ra dữ liệu ảnh dựa trên một lượng nhỏ dữ liệu gán nhãn sẽ giúp tăng cường số lượng ảnh cũng như độ đa dạng của số lượng dữ liệu được tạo ra. Do đó, mục tiêu của ứng viên là thiết kế một mô hình dựa trên cơ chế tạo sinh để phục vụ quá trình tạo ra ảnh mới mà vẫn tận dụng được các mặt nạ phân đoạn có sẵn và có giới hạn. Hình 2 minh họa ý tưởng tăng cường dữ liệu phục vụ bài toán phân đoạn ảnh.

Sản phẩm khoa học dự kiến và chỉ tiêu đánh giá:

- 01 bài báo hội nghị khoa học thuộc danh mục Scopus hoặc tương đương đề xuất mô hình tạo sinh ảnh dựa trên mặt nạ phân đoạn cho trước
- 01 báo cáo tổng quan kết quả thực nghiệm nghiên cứu



Hình 2. Minh họa ý tưởng sử dụng mô hình Dataset Synthesis để tạo sinh dữ liệu với nhãn phân đoạn.

Phương pháp: (Mô tả chi tiết: phương pháp, quy trình chọn mẫu và cách thức thu thập số liệu có thể có; lý giải việc lựa chọn của tác giả)

- Đề xuất mô hình tạo sinh ảnh với điều kiện ràng buộc là mặt nạ phân đoạn ngữ nghĩa, sử dụng hướng tiếp cận với cơ chế tạo sinh, khuếch tán.

- Thử nghiệm trên một số bộ dữ liệu chuẩn được các phương pháp tiên tiến trong cùng nhánh bài toán thực hiện như MS-COCO, Pascal-VOC hoặc các tập dữ liệu chuẩn phù hợp và sử dụng mô tả từ chính các tập dữ liệu chuẩn đó. Các độ đo đánh giá được cân nhắc sử dụng là SSIM, FID (hoặc tương đương) và độ đo định tính (khảo sát người dùng).

Trước đây, ứng viên cùng nhóm nghiên cứu đã từng đề xuất phương pháp tạo sinh ảnh siêu phân giải trên đa tác nhân thoái hóa thông qua một mô-đun tăng cường ảnh trực tuyến [35]. Với phương pháp này, nhóm nghiên cứu đã thực hiện tăng cường ảnh với nhiều hơn một loại tác nhân thoái hóa để cải thiện khả năng tạo sinh ảnh mới. Việc áp dụng thành công cơ chế tăng cường ảnh vào quá trình tạo sinh dữ liệu đặt tiền đề cho việc mở rộng nghiên cứu khi thực hiện nội dung này. Cụ thể, mô hình được đề xuất ở nội dung này có thể sử dụng một cơ chế tăng cường ảnh trong quá trình tạo sinh, hướng đến tăng sự đa dạng trong quá trình huấn luyện.

Phân tích và diễn giải số liệu thu được:

- Công bố bài báo khoa học quốc tế
- Báo cáo đề tài

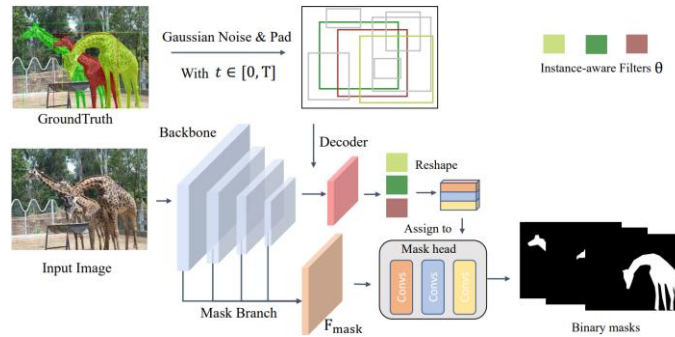
4.2 Nội dung 2: Hướng tiếp cận phân đoạn ảnh dựa trên mô hình tạo sinh

Động lực nghiên cứu: Như đã đề cập ở phần trước, các mô hình tạo sinh dữ liệu đã đạt được nhiều thành tựu trong nhánh bài toán tạo sinh (generative problem) như tạo sinh hình ảnh hay video. Trái lại, các bài toán cơ bản của lĩnh vực thị giác máy tính như phân loại, phát hiện đối tượng hay phân đoạn ảnh được xếp vào nhánh các bài toán phân biệt (discriminative problem). Đồng thời, các mô hình dựa trên CNN gặp phải vấn đề thiên kiến quy nạp (inductive bias) khi sử dụng các vùng đề xuất hoặc các điểm neo. Trong khi đó, các mô hình dựa trên truy vấn như [39, 40] hay cơ chế khuếch tán [26, 41, 42] có thể xử lý vấn đề thiên kiến này. Liệu chúng ta có thể sử dụng hướng tiếp cận này để giải quyết các bài toán phân đoạn hay không? Thông thường, các mô hình sẽ tìm cách rút trích đặc trưng từ dữ liệu đầu vào, tiến hành quá trình huấn luyện và trả về kết quả là một nhãn ngữ nghĩa, một khung bao vật thể, hay một mặt nạ phân đoạn. Tuy nhiên, ở một góc nhìn khác, chúng ta có thể xem việc gán nhãn cho các điểm ảnh tương tự như việc tạo sinh ra bản đồ phân đoạn. Theo đó, thay vì xem xét bài toán ở nhánh bài toán phân biệt, chúng ta có thể sử dụng hướng tiếp cận tạo sinh để giải quyết vấn đề.

Câu hỏi làm cách nào để áp dụng mô hình khuếch tán giải quyết bài toán phân đoạn ảnh vẫn là một vấn đề còn nhiều tranh cãi. Gần đây, DiffusionDet [33] đã được đề xuất để giải quyết bài toán phát hiện đối tượng dựa trên cơ chế khuếch tán, bằng cách tạo sinh các khung bao để xác định đối tượng trong ảnh. Ở giai đoạn huấn luyện, mô hình này thêm nhiều Gaussian vào các khung bao nhãn để tạo ra các khung bao nhiễu. Sau đó, các đặc trưng RoI của các khung bao nhiễu này được đưa vào mô-đun giải mã để dự

đoán khung bao. Toàn bộ mô hình hoạt động theo mô hình denoising – khử nhiễu. Ở giai đoạn suy luận, DiffusionDet [33] tạo ra khung bao bằng cách tuần tự tạo ra các khung bao ngẫu nhiên đưa vào mô-đun giải mã, đây là quá trình đảo ngược của giai đoạn khuếch tán. Trên cơ sở đó, ứng viên đề xuất tiếp tục phát triển phương pháp và ý tưởng được đề xuất của DiffusionDet [33] để thực hiện trên tác vụ phân đoạn ảnh.

Mục tiêu: Trong nội dung này, ứng viên thiết kế một kiến trúc phân đoạn ảnh trên cơ sở sử dụng hướng tiếp cận tạo sinh với mô hình khuếch tán (Diffusion). Điểm mới so với các công trình trước đây chính là việc sử dụng cơ chế tạo sinh (generative approach) để giải quyết bài toán phân biệt (discriminative approach). Điều này cho phép tận dụng sức mạnh của các mô hình tạo sinh được huấn luyện sẵn với các bộ dữ liệu lớn, cải thiện độ chính xác của quá trình huấn luyện, tinh chỉnh và suy luận của mô hình phân đoạn. Hình 3 minh họa ý tưởng của mô hình với cơ chế khuếch tán cho bài toán phân đoạn.



Hình 3. Minh họa ý tưởng sử dụng cơ chế Diffusion [17] trong mô hình phân đoạn ảnh.

Phương pháp: (Mô tả chi tiết: phương pháp, quy trình chọn mẫu và cách thức thu thập số liệu có thể có; lý giải việc lựa chọn của tác giả)

- Đề xuất mô hình phân đoạn ảnh dựa trên hướng tiếp cận tạo sinh với mô hình tạo sinh hoặc khuếch tán Diffusion.
- Thử nghiệm trên một số bộ dữ liệu chuẩn được các phương pháp tiên tiến trong cùng nhánh bài toán thực hiện như MS-COCO, Pascal-VOC hoặc các tập dữ liệu chuẩn phù hợp. Độ đo đánh giá chuẩn được sử dụng là IoU hoặc mIoU, AP – Average Precision (hoặc tương đương), thông thường được đánh giá tại các mức độ độ chính xác như mAP, AP@50, AP@75 hoặc các mức kích thước bản đồ phân đoạn như APs (small), APm (medium), Apl (large).

Với kiến trúc này, bên cạnh việc thừa kế kiến trúc điển hình của các mô hình 2 giai đoạn phân đoạn thực thể (như Mask R-CNN [2], hay Mask Scoring R-CNN [3]), một nhánh mới được thêm vào với mục tiêu cung cấp thêm thông tin dữ liệu tạo sinh. Cụ thể, các khung bao và mặt nạ phân đoạn từ dữ liệu được thêm vào một độ nhiễu nhất định để huấn luyện mô hình. Mô hình phân đoạn cần học được khả năng khử nhiễu để tái tạo lại các bản đồ phân đoạn và khung bao ban đầu của ảnh đầu vào.

Sản phẩm khoa học dự kiến và chỉ tiêu đánh giá:

- 01 bài báo tạp chí khoa học thuộc danh mục ISI-SCI/SCIE hoặc tương đương đề xuất mô hình phân đoạn ảnh dựa trên hướng tiếp cận tạo sinh dữ liệu
- 01 báo cáo tổng quan kết quả thực nghiệm nghiên cứu

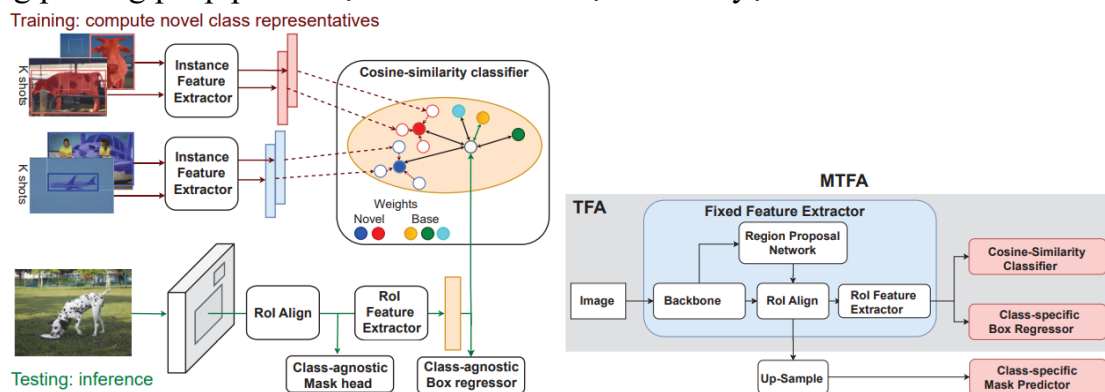
Phân tích và diễn giải số liệu thu được:

- Công bố bài báo tạp chí khoa học quốc tế
- Báo cáo đề tài

4.3 Nội dung 3: Hướng tiếp cận sử dụng ít dữ liệu huấn luyện

Động lực nghiên cứu: Các mô hình học sâu trong lĩnh vực thị giác máy tính hay xử lý ngôn ngữ tự nhiên đạt được độ chính xác cao nhờ vào quá trình huấn luyện trên số lượng lớn dữ liệu gán nhãn. Tuy nhiên, chi phí cho thời gian và nhân lực để gán nhãn thủ công cho các tác vụ trong lĩnh vực này là một vấn đề đang được cộng đồng nghiên cứu quan tâm. Hơn nữa, tác vụ phân đoạn ảnh đòi hỏi độ chính xác của việc gán nhãn ở cấp độ điểm ảnh, làm cho vấn đề liên quan đến chi phí càng tăng cao hơn. Để giải quyết vấn đề này, cộng đồng nghiên cứu đã và đang tìm cách huấn luyện các mô hình học sâu đạt độ chính xác cao mà chỉ cần một số ít mẫu dữ liệu huấn luyện được gán nhãn chi tiết. Dựa trên quan sát thực tế về quá trình học hỏi tri thức mới của con người, phương pháp huấn luyện sử dụng ít mẫu dữ liệu được đề xuất dựa trên cơ chế dùng tri thức để tiếp thu cái mới. Phương pháp này đã đạt được nhiều kết quả khả quan trong các tác vụ phân loại và phát hiện đối tượng trong ảnh. Tuy nhiên, độ chính xác trên tác vụ phân đoạn ảnh còn hạn chế và cần thiết để nghiên cứu và phát triển thêm.

Mục tiêu: Trong nội dung này, ứng viên đề xuất giải quyết bài toán phân đoạn ảnh với ít dữ liệu huấn luyện. Điểm mới của nội dung này bên cạnh việc áp dụng hướng tiếp cận học với ít dữ liệu huấn luyện là việc sử dụng dữ liệu tạo sinh để phục vụ quá trình phân đoạn. Như vậy, mô hình đề xuất có khả năng phân đoạn ảnh với một số lượng ít mẫu dữ liệu thật được gán nhãn, và khai thác được dữ liệu tăng cường thông qua việc sử dụng thêm dữ liệu tạo sinh tham chiếu từ các dữ liệu gốc ban đầu. Hình 4 minh họa ý tưởng phương pháp phân đoạn ảnh với ít dữ liệu huấn luyện.



Hình 4. Minh họa ý tưởng bài toán phân đoạn ảnh với ít dữ liệu huấn luyện [34].

Phương pháp: (Mô tả chi tiết: phương pháp, quy trình chọn mẫu và cách thức thu thập số liệu có thể có; lý giải việc lựa chọn của tác giả)

- Đề xuất mô hình phân đoạn ảnh sử dụng ít dữ liệu huấn luyện (few-shot learning for segmentation). Cụ thể, quá trình huấn luyện với dữ liệu tổng quát cung cấp tri thức chung cho mô hình phân đoạn. Sau đó, mô hình phân đoạn được huấn luyện với ít mẫu dữ liệu để học phương pháp học đưa ra dự đoán nhãn phân đoạn [34].
- Tiến hành thực nghiệm trên một số bộ dữ liệu chuẩn được các phương pháp tiên tiến trong cùng nhánh bài toán sử dụng như MS-COCO, Pascal-VOC hoặc các tập dữ liệu chuẩn phù hợp. Độ đo đánh giá chuẩn được sử dụng là IoU hoặc mIoU, hay AP – Average Precision, AR – Average Recall (hoặc tương đương) tại các cấu hình số lượng mẫu dữ liệu khác nhau. Các mô hình huấn luyện trên các tập dữ liệu này có thể sử dụng làm mô hình cơ sở để mở rộng nghiên cứu trên các miền dữ liệu khác, đơn cử là dữ liệu nguy trang.
- Để cải thiện độ chính xác phân đoạn, mô hình đề xuất có thể cân nhắc các yếu tố liên quan đến khai thác (1) dữ liệu không gán nhãn hoặc (2) sử dụng dữ liệu tạo sinh. Với hướng tiếp cận khai thác dữ liệu không gán nhãn, các mô hình tạo sinh đã đề cập ở Nội dung 2 có khả năng tạo ra các nhãn giả để phục vụ quá trình huấn luyện, đơn cử là Dataset Diffusion [32], GliGen [37] hay BlendDiff [38].

Sản phẩm khoa học dự kiến và chỉ tiêu đánh giá:

- 01 bài báo tạp chí khoa học thuộc danh mục ISI-SCI/SCIE hoặc tương đương đề xuất mô hình phân đoạn ảnh sử dụng ít dữ liệu huấn luyện
- 01 báo cáo tổng quan kết quả thực nghiệm nghiên cứu

Phân tích và diễn giải số liệu thu được:

- Công bố bài báo khoa học quốc tế
- Báo cáo đề tài

5. Một số kết quả sơ khởi

Ứng viên đã thực hiện nghiên cứu đề tài phân đoạn ảnh và tạo sinh ảnh thuộc lĩnh vực thị giác máy tính trong quá trình thực hiện khóa luận tốt nghiệp đại học và luận văn thạc sĩ. Đồng thời, trong quá trình làm việc tại Trường Đại học Công nghệ Thông tin, ứng viên đã chủ nhiệm và tham gia các đề tài khoa học công nghệ cấp cơ sở và cấp ĐHQG có liên quan đến chủ đề phân đoạn ảnh và tạo sinh ảnh.

Đề tài khóa luận tốt nghiệp đại học của ứng viên là “**Phân đoạn ngữ nghĩa ảnh trong điều kiện thiếu sáng với phương pháp tương thích miền dữ liệu**”. Khóa luận đề xuất phương pháp phân đoạn ảnh trong điều kiện thiếu sáng vào ban đêm mà chỉ sử dụng ảnh gán nhãn vào ban ngày. Mô hình sử dụng mạng tạo sinh đối kháng để tương thích miền dữ liệu ban ngày thành ban đêm để phục vụ quá trình huấn luyện. Khóa luận đã chứng minh tính hiệu quả của phương pháp đề xuất qua từng giai đoạn huấn luyện mô

hình. Kế thừa nghiên cứu về phân đoạn ảnh này, ứng viên tiếp tục phát triển luận văn thạc sĩ với đề tài **“Phân đoạn thực thể nguy trang dựa trên đặc trưng có tính phân biệt cao”**. Luận văn đề xuất hai phương pháp phân đoạn ảnh ở cấp độ thực thể: (1) phương pháp khai thác đặc trưng biên cạnh để nhận diện thực thể nguy trang, và (2) phương pháp sử dụng ít dữ liệu huấn luyện với đề xuất hàm mất mát Instance Triplet Loss và Instance Memory Storage. Các kết quả của khóa luận và luận văn được ứng viên tổng hợp và gửi đăng một bài báo hội nghị quốc tế thuộc danh mục Scopus, và hai bài tạp chí quốc tế thuộc danh mục ISI-SCI/SCIE chờ bình duyệt.

Ứng viên đã chủ nhiệm và tham gia thực hiện các đề tài khoa học công nghệ có liên quan đến bài toán phân đoạn ảnh và bài toán tạo sinh ảnh như sau:

Đề tài “Tạo ảnh siêu phân giải với phương pháp thoái hóa đa tác nhân” (D1-2021-18, chủ nhiệm), nghiệm thu vượt tiến độ năm 2022. Trong đề tài này, nhóm nghiên cứu đề xuất sử dụng mô hình tạo sinh ảnh để tái tạo ảnh với độ phân giải cao. Mô hình đề xuất trong đề tài này khai thác yếu tố đa tác nhân thoái hóa để tăng cường độ đa dạng của thông tin đặc trưng phục vụ quá trình tái tạo ảnh của mô-đun giải mã. Đồng thời, đề tài cũng đề xuất một hàm độ lỗi đa tác nhân thoái hóa để huấn luyện mô hình tái tạo ảnh. Kết quả của đề tài được công bố thành một bài báo khoa học tại hội nghị quốc tế thuộc danh mục Scopus.

Đề tài “Tăng cường đặc trưng biên cạnh cho bài toán phân đoạn thực thể nguy trang với mô hình Transformer” (D1-2023-20, chủ nhiệm), nghiệm thu vượt tiến độ năm 2024. Nhóm nghiên cứu đã đề xuất mô hình phân đoạn thực thể trong ảnh trên nền tảng kiến trúc mô hình Transformer khai thác thông tin đặc trưng phân biệt ở vùng biên cạnh thực thể. Kết quả của đề tài được công bố thành một bài báo khoa học tại hội nghị quốc tế thuộc danh mục Scopus.

Đề tài “Chuyển đổi các miền không gian ảnh trên dữ liệu không ghép cặp sử dụng mạng tạo sinh đối kháng” (C2022-26-01, tham gia), nghiệm thu vượt tiến độ, xếp loại xuất sắc năm 2023. Đề tài đã đề xuất hai bộ dữ liệu ảnh cho bài toán tạo sinh ảnh không ghép cặp, một độ đo đánh giá kết quả mô hình tạo sinh (FISI), hai phương pháp chuyển đổi không gian màu và chuyển đổi họa tiết – đường nét. Kết quả của đề tài được công bố thành hai bài báo trong đó gồm một bài báo tạp chí ISI-SCIE-Q1, và một bài báo hội nghị quốc tế thuộc danh mục Scopus.

6. Lí do thực hiện đề tài

Quá trình thực hiện nghiên cứu trong lĩnh vực thị giác máy tính ở bậc đại học và cao học của ứng viên tập trung vào các bài toán phân đoạn ảnh và tạo sinh ảnh. Do đó, ứng viên mong muốn được kế thừa và tiếp tục phát triển các kết quả nghiên cứu ban đầu cho các công trình ở bậc nghiên cứu sinh. Các nghiên cứu ở giai đoạn này sẽ được thực hiện ở mức độ kết hợp cao hơn, sử dụng các hướng tiếp cận tạo sinh để giải quyết bài toán

phân đoạn ảnh, vốn thuộc nhánh bài toán phân biệt. Như đã đề cập ở các nội dung nghiên cứu chi tiết, ứng viên sẽ thực hiện từng bước để giải quyết các thách thức của bài toán phân đoạn ảnh bằng cách sử dụng mô hình tạo sinh và thông tin ngôn ngữ tự nhiên để tăng sự đa dạng của dữ liệu gán nhãn, khai thác cơ chế tạo sinh phục vụ kiến trúc phân đoạn ảnh, và giải quyết bài toán phân đoạn với ngữ cảnh sử dụng ít dữ liệu huấn luyện được gán nhãn. Đồng thời, xu hướng hiện tại của lĩnh vực khoa học máy tính là sự kết hợp giữa các mô hình trong lĩnh vực thị giác máy tính với thông tin từ các mô hình ngôn ngữ để khai thác các mô hình lớn vốn đã được huấn luyện sẵn với nhiều dữ liệu. Hướng nghiên cứu này đang được cộng đồng nghiên cứu trên thế giới quan tâm, có nhiều công trình nghiên cứu, và còn nhiều vấn đề cần được giải quyết.

7. Kế hoạch thực hiện

Nội dung thực hiện	Thời gian thực hiện
Khảo sát, nghiên cứu một số phương pháp phân đoạn ảnh, các phương pháp tạo sinh ảnh có điều kiện. Xác định nội dung nghiên cứu chi tiết.	6 tháng
Nghiên cứu, đề xuất, cải tiến phương pháp tạo sinh ảnh có điều kiện phục vụ bài toán phân đoạn ảnh.	6 tháng
Đề xuất các cải tiến phục vụ bài toán phân đoạn ảnh dựa trên các điểm yếu khai thác được từ những mô hình đã khảo sát, tập trung vào hướng tiếp cận sử dụng cơ chế tạo sinh dữ liệu và sử dụng ít dữ liệu huấn luyện.	12 tháng
Tiến hành cài đặt, thực nghiệm và đánh giá các mô hình đề xuất. Tổng hợp, công bố các kết quả nghiên cứu tại các hội nghị, tạp chí thuộc danh mục theo quy định.	6 tháng
Viết báo cáo và chuẩn bị bảo vệ luận án	6 tháng

8. Tài liệu tham khảo

- [1] Zhi Tian, Chunhua Shen, and Hao Chen. “Conditional Convolutions for Instance Segmentation”. In ECCV. 2020.
- [2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask R-CNN”. In ICCV. 2017, pp. 2961–2969.
- [3] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. “Mask Scoring R-CNN”. In CVPR. 2019.
- [4] Zhaowei Cai and Nuno Vasconcelos. “Cascade R-CNN: Delving into High Quality Object Detection”. In CVPR. 2018.
- [5] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. “Path Aggregation Network for Instance Segmentation”. In CVPR. 2018.

- [6] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. “Hybrid task cascade for instance segmentation”. In CVPR. 2019, pp. 4974–4983.
- [7] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. “Blendmask: Top-down meets bottom-up for instance segmentation”. In CVPR. 2020, pp. 8573–8581.
- [8] Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. “Mask transfiner for high-quality instance segmentation”. In CVPR. 2022.
- [9] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. “Camouflaged Instance Segmentation via Explicit De-Camouflaging”. In CVPR. 2023, pp. 17918–17927.
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need”. In NeurIPS 30 (2017).
- [11] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. “YOLACT: Real-time Instance Segmentation”. In ICCV. 2019.
- [12] Zhi Tian, Chunhua Shen, and Hao Chen. “Conditional Convolutions for Instance Segmentation”. In ECCV. 2020.
- [13] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. “SOTR: Segmenting objects with transformers”. In ICCV. 2021, pp. 7157.
- [14] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. “SOLO: Segmenting Objects by Locations”. In ECCV. 2020.
- [15] Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuanbo Chen, and Luc Van Gool. “OSFormer: One-Stage Camouflaged Instance Segmentation with Transformers”. In ECCV. 2022.
- [16] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R.. “Segment anything”. In ICCV. 2023 (pp. 4015-4026).
- [17] Gu, Z., Chen, H., Xu, Z., Lan, J., Meng, C., & Wang, W. (2022). Diffusioninst: Diffusion model for instance segmentation. arXiv preprint arXiv:2212.02773.
- [18] De Ridder, V., Dey, B., Halder, S., & Van Waeyenberge, B. “SEMI-DiffusionInst: A Diffusion Model Based Approach for Semiconductor Defect Classification and Segmentation”. In 2023 International Symposium ELMAR (pp. 61-66).
- [19] Xun, S., Li, D., Zhu, H., Chen, M., Wang, J., Li, J., ... & Huang, P. “Generative adversarial networks in medical image segmentation: A review”. Computers in Biology and Medicine 2022, 140, 105063.

- [20] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. “Meta R-CNN: Towards general solver for instance-level low-shot learning”. In ICCV. 2019.
- [21] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. “CRNet: Cross-Reference Networks for Few-Shot Segmentation”. In CVPR. June 2020.
- [22] aixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. “PANet: Few-Shot Image Semantic Segmentation with Prototype Alignment”. In ICCV. 2019.
- [23] K. Nguyen and Sinisa Todorovic. “iFS-RCNN: An Incremental Few-shot Instance Segmenter”. In CVPR. 2022, pp. 7010–7019.
- [24] Bin-Bin Gao, Xiaochen Chen, Zhongyi Huang, Congchong Nie, Jun Liu, Jinxiang Lai, Guannan Jiang, Xi Wang, and Chengjie Wang. “Decoupling Classifier for Boosting Few-shot Object Detection and Instance Segmentation”. In NeurIPS. 2022.
- [25] Haochen Wang, Jie Liu, Yongtuo Liu, Subhransu Maji, Jan-Jakob Sonke, and Efstratios Gavves. “Dynamic Transformer for Few-shot Instance Segmentation”. In: ACMMM. 2022, pp. 2969–2977.
- [26] Ho, J., Jain, A., & Abbeel, P. Denoising diffusion probabilistic models. In NeurIPS 33, 6840-6851. 2020.
- [27] Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., ... & Sun, L. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. arXiv preprint arXiv:2402.17177. 2024.
- [28] Midjourney. 2022. Midjourney.com. <https://www.midjourney.com>
- [29] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125. 2022.
- [30] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In NeurIPS 35, 36479-36494.
- [31] Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592. 2023.
- [32] Nguyen, Q., Vu, T., Tran, A., & Nguyen, K. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In NeurIPS 36. 2024.
- [33] Chen, S., Sun, P., Song, Y., & Luo, P. Diffusiondet: Diffusion model for object detection. In ICCV. 2023 (pp. 19830-19843). 2023.
- [34] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. “Incremental Few-Shot Instance Segmentation”. In CVPR. 2021, pp. 1185–1

- [35] Ngoc-Khanh Nguyen, Thanh-Danh Nguyen, and Vinh-Tiep Nguyen. “A More Focus on Multi-degradation Method for Single Image Super-Resolution.” In MAPR, 2021. (Scopus)
- [36] Chen, J., Lu, J., Zhu, X., & Zhang, L. Generative semantic segmentation. In CVPR. 2023. (pp. 7111-7120).
- [37] Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., & Lee, Y. J. Gligen: Open-set grounded text-to-image generation. In CVPR (pp. 22511-22521). 2023.
- [38] Avrahami, O., Lischinski, D., & Fried, O. Blended diffusion for text-driven editing of natural images. In CVPR (pp. 18208-18218). 2022.
- [39] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal imagesegmentation. In CVPR, pages 1280–1289, 2022.
- [40] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In ICCV, pages 6890–6899, 2021.
- [41] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In ICLR, 2021.
- [42] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In NeurIPS, pages 11895–11907, 2019.