Phương pháp học bán giám sát dựa vào không gian nhúng giữa các điểm ảnh cho bài toán phân vùng ảnh với giám sát yếu

Lê Phan Minh Đạt
Khoa Công Nghệ Thông Tin,
Đại học Khoa học Tự Nhiên, Việt Nam
Đại học Quốc gia Việt Nam,
Thành phố Hồ Chí Minh, Việt Nam
21120046@student.hcmus.edu.vn

Nguyễn Văn Trí
Khoa Công Nghệ Thông Tin,
Đại học Khoa học Tự Nhiên, Việt Nam
Đại học Quốc gia Việt Nam,
Thành phố Hồ Chí Minh, Việt Nam
21120345@student.hcmus.edu.vn

Ngô Ngọc Vũ
Khoa Công Nghệ Thông Tin,
Đại học Khoa học Tự Nhiên, Việt Nam
Đại học Quốc gia Việt Nam,
Thành phố Hồ Chí Minh, Việt Nam
21120598@student.hcmus.edu.vn

Abstract*—Trong phân vùng ảnh (semantic segmentation), việc huấn luyện đòi hỏi chú thích từng điểm ảnh là một quá trình tốn thời gian và chi phí. Phân vùng ảnh với giám sát yếu (WSSS) đưa ra giải pháp, tận dụng các giám sát yếu hơn như nhãn ảnh toàn ảnh. Tuy nhiên, WSSS gặp phải những thách thức do thông tin hạn chế trong các chú thích yếu, dẫn đến không chính xác về vị trí và khó khăn trong việc tách biệt các đối tượng tiền cảnh khỏi nhiễu loạn nền. Bài báo này đề xuất một phương pháp WSSS giải quyết những hạn chế này bằng cách sử dụng thông tin phân vùng thưa thớt, được gọi là nhãn giả (pseudo-labels), bao gồm nhãn ảnh, khung bao quanh đối tượng, nhãn điểm và nhãn nguệch ngoạc. Bài báo giới thiệu bốn loại mối quan hệ tương phản giữa các điểm ảnh và phân vùng trong không gian đặc trưng: tương đồng hình ảnh cấp thấp, chú thích ngữ nghĩa, đồng xuất hiện và tương đồng đặc trưng. Các mối quan hệ này đóng vai trò như gợi ý, cho phép từng điểm ảnh học hỏi từ dữ liệu huấn luyện với các loại nhãn bán giám sát khác nhau. Đáng chú ý, các điểm ảnh không được dán nhãn không chỉ tham gia vào việc gom cụm dữ liệu trong mỗi ảnh mà còn học các đặc trưng phân biệt giữa các ảnh khác nhau. Phương pháp này đạt được cải thiện hiệu suất đáng kể trên các bộ dữ liệu Pascal VOC [1] và DensePose [2].

I. Giới thiệu

Phân vùng ảnh, một nền tảng của thị giác máy tính, nhằm tự động gán nhãn ngữ nghĩa (ví dụ: "ô tô", "người") cho từng điểm ảnh trong ảnh. Trong khi các phương pháp được giám sát đầy đủ như DeepLabV3+[3] đã đạt được kết quả ấn tượng, chúng yêu cầu một lượng lớn chú thích từng điểm ảnh. Quá trình dán nhãn này thường tốn kém, tốn thời gian và cản trở khả năng mở rộng của các phương pháp này cho các ứng dụng thực tế.

Phân vùng ảnh với giám sát yếu (WSSS) nổi lên như một giải pháp thay thế đầy hứa hẹn, tận dụng dữ liệu được dán nhãn yếu có sẵn, chẳng hạn như thẻ hoặc khung bao quanh ảnh. Sự thay đổi cách tiếp cận này trong dữ liệu huấn luyện cho phép phát triển mô hình hiệu quả hơn và khả năng áp dụng rộng rãi hơn.

Các công trình tiên phong trong WSSS đã khám phá việc tận dụng nhãn ảnh toàn ảnh cho phân vùng từng điểm ảnh. Krahenbuhl et al. [2015][4] giới thiệu Bản đồ kích hoạt lớp (CAM) để xác định vị trí các vùng ảnh phân biệt cho lớp mục tiêu dựa trên bản đồ kích hoạt của mạng nơ-ron tích chập sâu (CNN) được đào tạo trước. Sau đó, Pedro O. Pinheiro [5] đề xuất một khung học tận dụng cả nhãn ảnh toàn ảnh và thông tin cạnh để cải thiện độ chính xác của phân vùng.

Những tiến bộ gần đây hơn đi sâu vào việc khám phá các dạng giám sát yếu phong phú hơn. Ví dụ, sử dụng khung bao quanh để hướng dẫn mô hình bằng cách áp đặt các ràng buộc không gian trên các mặt nạ phân vùng được dự đoán[6]. Mặt khác, đề xuất một phương pháp tận dụng chú thích cấp điểm[7], cho thấy kết quả hứa hẹn với lượng công sức của con người cần thiết cho chú thích là tối thiểu.

Mặc dù WSSS mang lại những lợi thế đáng kể, nó vẫn có thể gặp khó khăn với dữ liệu được dán nhãn hạn chế. Tại đây, các phương pháp học bán giám sát giúp bắc cầu bằng cách kết hợp một lượng lớn dữ liệu không được dán nhãn cùng với các ví dụ được dán nhãn hạn chế. Dữ liệu không được dán nhãn này có thể cung cấp thông tin có giá trị và nâng cao khả năng học biểu diễn mạnh mẽ của mô hình.

Có một điều thú vị trong lĩnh vực này chính là các phương pháp học số liệu bán giám sát theo từng điểm ảnh. Các kỹ thuật này tập trung vào việc học một số liệu khoảng cách trong không gian nhúng ẩn, nơi các điểm ảnh tương tự từ các hình ảnh khác nhau được ánh xạ lại gần nhau hơn, bất kể có hay không có nhãn hiệu rõ ràng. Điều này cho phép mô hình tận dụng cấu trúc vốn có của dữ liệu không được dán nhãn để cải thiện hiệu suất phân vùng.

Bài tổng quan này đi sâu vào phương pháp học số liệu bán giám sát theo từng điểm ảnh, khám phá tiềm năng của chúng để tăng cường độ chính xác phân vùng với lượng dữ liệu được dán nhãn tối thiểu.

II. Các công trình liên quan

Phân vùng ảnh với giám sát yếu (WSSS) đã nổi lên như một chiến lược mạnh mẽ để giải quyết những thách thức trong việc thu thập một lượng lớn chú thích từng điểm ảnh. Phần này khám phá các kỹ thuật WSSS khác nhau và đi sâu vào

lĩnh vực thú vị của học số liệu bán giám sát theo từng điểm ảnh.

Tận dụng các Nhãn Giám Sát Yếu:

- **Nhãn Ảnh Toàn Ảnh**: Các công trình tiên phong của Krahenbuhl et al. [2015] [4] giới thiệu Bản đồ Kích hoạt Lớp (CAM) để xác định vị trí các vùng phân biệt cho lớp mục tiêu dựa trên kích hoạt của mạng CNN được đào tạo trước. Sau đó, Zhang et al. [2018] [5] đề xuất SharpMask, một khung học tận dụng cả nhãn ảnh toàn ảnh và thông tin cạnh để cải thiện độ chính xác của phân vùng.
- **Khung Bao Quanh**: Các kỹ thuật như của Zhao et al. [2019] [6] sử dụng khung bao quanh để hướng dẫn mô hình bằng cách áp đặt các ràng buộc không gian trên các mặt nạ phân vùng được dự đoán. Điều này đưa thông tin không gian có giá trị vào mặc dù thiếu nhãn từng điểm ảnh.
- **Chú Thích Điểm:** Xu et al. [2020] [7] đề xuất một phương pháp tận dụng chú thích cấp điểm, cho thấy kết quả hứa hẹn với lượng công sức của con người cần thiết cho chú thích là tối thiểu. Cách tiếp cận này nhấn mạnh tiềm năng khai thác các dạng giám sát yếu hơn.

Mặc dù WSSS mang lại những lợi thế, dữ liệu được dán nhãn hạn chế vẫn có thể cản trở hiệu suất. Học bán giám sát giúp bắc cầu bằng cách kết hợp một lượng lớn dữ liệu không được dán nhãn cùng với các ví dụ được dán nhãn hạn chế. Phần này tập trung vào những tiến bộ gần đây trong học số liệu bán giám sát theo từng điểm ảnh cho WSSS.

Học Số Liệu cho Dữ Liệu Không Được Giám Sát: Các kỹ thuật như tận dụng triplet loss[8] hoặc các cách tiếp cận tương tự trong không gian nhúng ẩn. Điều này đảm bảo rằng các điểm ảnh tương tự từ các hình ảnh khác nhau được ánh xạ lại gần nhau hơn, ngay cả khi không có nhãn rõ ràng. Mô hình tận dụng cấu trúc vốn có này của dữ liệu không được dán nhãn để cải thiện khả năng phân vùng của nó.

Kết hợp Các Điểm Mạnh: Các công trình gần đây khám phá việc kết hợp các tín hiệu giám sát yếu khác nhau. Ví dụ [9], sử dụng cả khung bao quanh và phân bố nhãn ảnh toàn ảnh để cải thiện hiệu suất .

III. Phương pháp học bán giám sát dựa vào không gian nhúng giữa các điểm ảnh

Ý tưởng của phương pháp này là tạo ra không gian nhúng tốt giữa các điểm ảnh thông qua việc học tương phản giữa các điểm ảnh và phân vùng, giống như hình minh họa dưới đây.

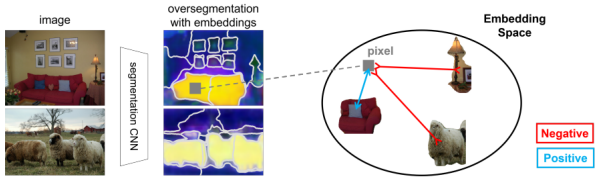


Fig. 1. Tổng quan về phương pháp

Trong tập dữ liệu được dán nhãn yếu, nơi các nhãn không có sẵn cho từng điểm ảnh, một tập các điểm ảnh i có cùng (khác) loại được ký hiệu lần lượt là $C+$ và $C-$. Ví dụ với bối cảnh điểm được dán nhãn yếu, $C+$ và $C-$ sẽ chỉ chứa một vài mẫu theo các nhãn điểm ảnh thưa thớt.

Phương pháp này đề xuất bốn loại mối quan hệ giữa các điểm ảnh và phân vùng để mở rộng các tập $C+$ và $C-$ nhằm cải thiện hiệu quả học đặc trưng.

- **Tương đồng hình ảnh cấp thấp**: Điểm ảnh phân vùng i thuộc về các tín hiệu hình ảnh cấp thấp (như đường viền, cạnh, góc) được coi là phân vùng dương đối với pixel i; còn lại đều là phân vùng âm.
- **Chú thích ngữ nghĩa**: Đầu tiên, phương pháp này tạo ra ảnh nhãn giả (pseudo label) bằng cách sử dụng CAM và các cải tiến của nó. Nhãn của một phân vùng có thể được ước tính bằng cách bỏ phiếu đa số giữa các điểm ảnh; nếu nó giống với nhãn của pixel i, thì phân vùng đó là phân vùng dương đối với i.
- **Đồng xuất hiện ngữ nghĩa**: Giả sử các điểm ảnh trong cùng bối cảnh ngữ nghĩa có xu hướng được nhóm lại với nhau. Nếu một phân vùng xuất hiện trong một ảnh có chia sẻ bất kỳ lớp ngữ nghĩa nào với ảnh của pixel i, thì nó là một phân vùng dương đối với i và ngược lại là phân vùng âm.
- **Tương thích đặc trưng:** Giả sử các điểm ảnh và phân vùng cùng ngữ nghĩa tạo thành một cụm trong không gian đặc trưng.

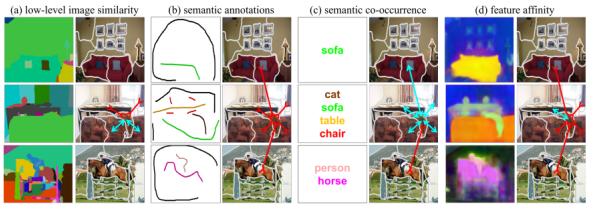


Fig. 2. Bốn loại mối quan hệ giữa các điểm ảnh với phân vùng

A. Mối quan hệ tương phản nhóm giữa điểm ảnh và phân vùng

Sau khi chúng ta đã có một tập phân vùng đã biết từ dữ liệu được gán nhãn sẵn thì chúng ta sẽ cố gắng để có thể gán nhãn các dữ liệu không được gán nhãn.

Tập dữ liệu được gán nhãn là C, tập dữ liệu không được gán nhãn là U.

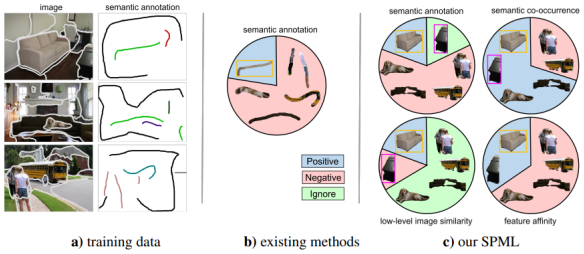
- **Tương đồng hình ảnh cấp thấp:** Để lan truyền nhãn trong các vùng hình ảnh một cách chính xác. Tác giả sử dụng bộ phát hiện biên cạnh `HED`[10] đã được tiền huấn luyện trên tập dữ liệu `BSDS500` và `gPb-owtucm`[11] để tạo ra một phân vùng không có thông tin ngữ nghĩa. Gọi $V+$ và $V-$ lần lượt là các phân vùng dương và âm. Dựa vào các phân vùng không có ngữ nghĩa ta sẽ xem xét các phân vùng trong cùng một hình ảnh so với pixel i để thêm vào $V+$ và $V-$, phương pháp còn căn chỉnh phân vùng đồng mức với các phân vùng được tạo ra bởi

phân cụm K-means để kết quả cho ra tốt hơn. Đây được tác giả gọi là phân vùng thừa mức độ thấp.

- **Chú thích ngữ nghĩa:** Các thẻ hình ảnh và các hộp giới hạn không cung cấp việc định vị pixel. Phải suy ra các nhãn giả từ hình ảnh hoặc bản đồ kích hoạt lớp và căn chỉnh chúng với các phân vùng được tạo ra bởi đặc trưng điểm ảnh. Các phân vùng dương hoặc âm của điểm ảnh i là các vùng cùng hoặc khác danh mục ngữ nghĩa, ký hiệu lần lượt là $C+$ và $C-$. Với chú thích ngữ nghĩa thì bỏ qua tất cả các phân vùng không có thẻ hình ảnh hoặc hộp giới hạn.
- **Đồng xuất hiện ngữ nghĩa:** ngữ cảnh ngữ nghĩa mô tả sự xuất hiện cùng lúc của các đối tượng khác nhau, có thể được sử dụng như một tiên đề để nhóm và tách các điểm ảnh. Ngữ cảnh ngữ nghĩa là hợp của các lớp đối tượng trong mỗi hình ảnh. Ngay cả khi không có đặc trưng không gian điểm ảnh của các ngữ nghĩa gán nhãn thì có thể tận dụng ngữ cảnh ngữ nghĩa để áp dụng quy chuẩn toàn cầu cho việc học đặc trưng ẩn. Lưu ý đặc trưng phải tách biệt các hình ảnh mà không có bất kỳ lớp đối tượng nào chồng lên nhau. Cho $O+$ ($O-$) là tập hợp các phân vùng trong các hình ảnh có (không có) các danh mục chồng lên nhau như hình ảnh của điểm ảnh i. Nghĩa là nếu hình ảnh của điểm ảnh i và một hình ảnh khác chia sẻ bất kỳ nhãn ngữ nghĩa nào thì tất cả các đoạn từ hình ảnh đó là các phân vùng dương với i và được bao gồm trong $O+$, ngược lại chúng là phân vùng âm $O-$. Tất cả các phân vùng trong hình ảnh của điểm ảnh i đều thuộc về $O+$ của i. Mối quan hệ ngữ cảnh ngữ nghĩa này không yêu cầu các chú thích không gian nhưng vẫn áp dụng làm quy chuẩn cho việc học đặc trưng điểm ảnh.
- **Tương thích đặc trưng**: Mục tiêu là học một đặc trưng điểm ảnh để chỉ định phân vùng ảnh. Phương pháp này gán một nhãn ngữ nghĩa cho mỗi đoạn không được gán nhãn bằng cách tìm đoạn gần nhất của nó trong không gian đặc trưng. Ký hiệu nhãn mở rộng này bằng $\hat{C}$. Với điểm ảnh i, định nghĩa phân vùng dương, âm của nó lần lượt là là $\hat{C}+$ và $\hat{C}-$ tùy thuộc vào việc một đoạn có cùng nhãn với i hay không. Mối quan hệ tương thích đặc trưng này hoạt động tốt nhất khi : tập hợp nhãn ban đầu đủ lớn để bao phủ không gian đặc trưng, các đoạn được gắn nhãn được phân bố đồng đều trong không gian đặc trưng và đặc trưng theo điểm ảnh đã mã hóa một số thông tin ngữ nghĩa cụ thể. Do đó chỉ nên áp dụng cho DensePose chú thích điểm trong các thí nghiệm, nơi mỗi phần cơ thể được gắn nhãn bằng một điểm.

B. HÀM ĐÁNH GIÁ MẤT MÁT CHO ĐẶC TRƯNG ĐIỂM ẢNH

Giới thiệu một chút về SegSort : SegSort[12] là một mô hình phân đoạn đầy đủ tạo ra một bản đồ đặc trưng điểm ảnh và một phân đoạn kết quả. Giả định các phân phối chuẩn độc lập cho từng đoạn riêng biệt, SegSort tìm kiếm ước lượng hợp lý cực đại của ánh xạ đặc trưng, để phân chia dựa trên đặc trưng trong hình ảnh và phân cụm trên các hình ảnh cung



a) training data **b)** existing methods **c)** our SPML

cấp sự phân biệt cực đại giữa các đoạn. Trong quá trình suy luận, nhãn đoạn được dự đoán bằng các truy xuất K-Nearest Neighbor. Phân chia dựa trên đặc trưng trong mỗi hình ảnh được tính toán thông qua phân cụm K-Means hình cầu[13]. Đặt e_i là vectơ đặc trưng tại điểm ảnh i, chứa đặc trưng được ánh xạ $\Theta(i)$ và tọa độ không gian của i. Đặt z_i là chỉ số của đoạn mà điểm ảnh i thuộc về, R_s là tập hợp các điểm ảnh trong đoạn s, và μ_s là đặc trưng của đoạn tính toán như trung tâm cụm hình cầu của đoạn s. Trong thủ tục Expectation-Maximization (EM) cho K-means hình cầu, bước E tính toán đoạn mà điểm ảnh i có xác suất lớn nhất thuộc về:

$$z_i = \arg\max_s \{\mu_{0s} e_i\}$$

Đặt s là đoạn kết quả mà điểm ảnh i thuộc về theo phân cụm hình cầu. Xác suất hậu nghiệm của điểm ảnh i trong đoạn s có thể được đánh giá qua tập hợp tất cả các đoạn S như sau:

$$p(z_i = s|e_i, \mu) = \frac{\exp(\kappa \mu_{0s} e_i)}{\sum_{t \in S} \exp(\kappa \mu_{0t} e_i)}$$

trong đó k là một siêu tham số tập trung. SegSort tối thiểu hóa mất mát hàm mật độ xác suất tiêu cực:

$$L_{SegSort}(i) = -\log p(z_i = s|e_i, \mu) = -\log \frac{\exp(\kappa \mu_{0s} e_i)}{\sum_{t \in S} \exp(\kappa \mu_{0t} e_i)}$$

SegSort áp dụng phân bổ vùng mềm[14] để tăng cường việc nhóm các đoạn cùng loại. Đặt $C+$ ($C-$) là tập chỉ số của các đoạn trong cùng (khác) loại với điểm ảnh i ngoại trừ s , đoạn mà i thuộc về. Ta có:

$$\begin{aligned} L_{SegSort^+}(i, C^+, C^-) &= -\log X_{t \in C^+} p(z_i = t|e_i, \mu) \\ &= -\log \frac{\sum_{t \in C^+} \exp(\kappa \mu_{0t} e_i)}{\sum_{t \in C^+ \cup C^-} \exp(\kappa \mu_{0t} e_i)} \end{aligned}$$

Phương pháp học bán giám sát dựa vào không gian nhúng giữa các điểm ảnh tận dụng 4 loại mối quan hệ ngữ nghĩa từ điểm ảnh đối với phân vùng để bổ sung các tập hợp được gán nhãn, bao gồm cả các điểm ảnh không được gán nhãn, hình thành mối quan hệ đối lập động giữa các đoạn (gồm dương, âm, bỏ qua). tổng mất mát tương phản điểm ảnh đối với phân vùng cho điểm ảnh i bao gồm 4 thành phần, mỗi thành phần tương ứng với một trong 4 mối quan hệ thu hút và đẩy đoạn từ pixel đến đoạn.

Công thức hàm mất mát là :

$$L(i) = \lambda_{IL} L_{SegSort^+}(i, V^+, V^-) + \lambda_C L_{SegSort^+}(i, C^+, C^-) + \lambda_{OL} L_{SegSort^+}(i, O^+, O^-) + \lambda_{AL} L_{SegSort^+}(i, \hat{C}^+, \hat{C}^-)$$

IV. THỰC NGHIỆM, PHÂN TÍCH VÀ ĐÁNH GIÁ MÔ HÌNH

A. CHUẨN BỊ DỮ LIỆU

Về dữ liệu, ta sẽ huấn luyện với dataset **Pascal VOC** của Everinghame, ta sẽ có 20 loại object cần học và 1 class nền nhằm để phân loại các object cần phân loại và các object bên ngoài không trong bài toán. Ở đây, chúng ta sẽ training trên tập dữ liệu đã được huấn luyện tăng cường, nghĩa là được được xử lý cũng như có một vài tùy chỉnh nhằm giúp cho việc huấn luyện trở nên hiệu quả và nhanh hơn, với 10582 ảnh trên tập training và 1449 ảnh trên tập validation. Và các tập huấn luyện của chúng ta sẽ được chia ra thành 4 nhãn dán yếu như đã nêu ở trên nhằm hỗ trợ cho việc học giám sát yếu của chúng ta

B. CẤU TRÚC MÔ HÌNH VÀ CÀI ĐẶT CÁC THAM SỐ

Dựa trên paper của Ke, với tập Pascal VOC, ta sẽ sử dụng kiểu kiến trúc với DeepLab[3] và ResNet101[15] là backbone network(là phần của mạng neuron dùng để xử lý và phân tích dữ liệu). Ta chỉ sử dụng các model đã được pre-train trên ImageNet[16] dataset.

Sau đó, ta tiến hành chia các siêu tham số theo từng các nhãn giám sát yếu cho từng các tập dữ liệu. Với tập Pascal VOC, ta sẽ "batchsize" là 16 cho nhãn image tag/bounding box. Với learning rate, ta sẽ đặt learning rate ban đầu là 0.003 nhân với hệ số giảm được tính bằng công thức $1 - \left(\frac{\text{lặp}}{\text{lặp tối đa}}\right)^{0.9}$ với momentum là 0.9. Với siêu tham số cho SegSort, ta sẽ cho các vector được chuẩn hóa về không gian cơ sở về các chiều không gian có độ lớn 64 . Ngoài ra, với thuật toán K-Means, ta sẽ cho lặp ở 10 vòng lặp và tạo ra 36 cụm với tập VOC. Chúng ta cũng sẽ đặt hệ số κ và λ tương ứng với mỗi quan hệ tương phản giữa điểm ảnh và phân vùng như bảng dưới đây (ở đây, ta sẽ xét λ và κ của Affinity là 0):

TABLE I
BẢNG CÀI ĐẶT CÁC THAM SỐ PHÙ HỢP VỚI TỪNG MỐI TƯƠNG QUAN

Tập dữ liệu	Nhãn dữ liệu	λ_I	κ_I	λ_C	κ_C	λ_O	κ_O
VOC	Scribbles	0.1	16	1.0	6	0.5	12
	Points	1.0	16	1.0	6	1.0	8
	Boxes	0.3	16	1.0	6	1.0	8
	Image tags	0.3	16	1.0	6	1.0	8

C. ĐÁNH GIÁ MÔ HÌNH

Ở đây, ta sẽ đánh giá mô hình dựa trên các nhãn yếu mà ta đã quy định ở trên với các mô hình khác. Trước hết, ta sẽ đánh giá với nhãn **Image tags**, với các dấu tick trên cột Saliency nghĩa là các mô hình được đánh dấu có sử dụng các nhãn dán tăng cường nhằm giúp cho các mô hình tập trung vào các điểm cần được training tốt hơn và cho ra các accuracy cao và hiệu quả hơn:

TABLE II
KẾT QUẢ SO SÁNH VỚI NHÃN IMAGE TAGS

Image Tags	Saliency	val	test
DSRG[17]	√	61.4	63.2
FickleNet[18]	√	64.9	65.3
RRM [19]	-	66.3	66.5
SGAN [20]	√	67.1	67.2
CAM+SE [21]	-	66.1	65.9
SPML	-	69.5	71.6

Như ta thấy ở trên, mô hình SPML thể hiện tốt hơn rất nhiều so với các mô hình ở phía trên. Cụ thể, ta xét các mô hình thể hiện tốt nhất, nghĩa là accuracy cao nhất,với các mô hình sử dụng các label tăng cường (có dấu $\surd$ ở saliency), SPML có accuracy cao hơn đến 4.4% vàhơn các mô hình không sử dụng các label tăng cường đến 5.1%.

Bây giờ, ta sẽ đánh giá với nhãn **Bounding Boxes**. Dưới đây sẽ là bảng đánh giá độ chính xác với nhãn dán hộp:

TABLE III
KẾT QUẢ SO SÁNH VỚI NHÃN BOUNDING BOX

Bounding Box	val	test
MCG+GrabCut+ [22]	69.4	-
BCAM[7]	70.2	-
SPML	73.5	74.7

Ta thấy rằng với tập validation, SPML hơn mô hình BCAM 3.3% và mô hình MCG đến 4.1%. Một con số tương đối ấn tượng cho một mô hình.

Tiếp theo, ta sẽ đánh giá mô hình với nhãn dán **Scribble** dựa trên kết quả của từng nhãn dán yếu với các mô hình cũng sử dụng các nhãn dán tương tự. Ở đây, ta sẽ so sánh với các mô hình được và không được xử lý qua phương pháp CRF (Conditional Random Field) nhằm trích xuất và phân tách ra các thông tin trong hình ảnh nhằm giúp các nhãn dán dạng rời rạc như này dễ phân loại và training hơn. Dưới đây là kết quả đánh giá dựa trên việc giám sát tối đa và bán giám sát như sau như sau dựa trên kết quả về độ chính xác của tập val như sau:

TABLE IV
KẾT QUẢ SO SÁNH VỚI NHÃN SCRIBBLE

Scribble	CRF	Full	Weak	WvF
NCL [23]	-	75.6	72.8	96.3
NCL	√	76.8	74.5	97.0
RL [24]	-	75.6	73.0	96.6
RL	√	76.8	75.0	97.7
BPG [25]	-	75.6	73.2	96.8
BPG	√	75.6	73.2	96.8
SPML	-	76.1	74.2	97.5
SPML	√	77.3	76.1	98.4

Với WvF là mối tương quan giữa kết quả khi training với Full Supervision (Full) và Weak Supervision (Weak) được tính bằng cách lấy Weak/Full nhằm giúp ta đánh giá được rằng liệu mô hình khi được huấn luyện bán giám sát có đạt được kết quả khi huấn luyện được giám sát hoàn toàn hay không. Ta có thể thấy được kết quả trước khi sử dụng phương pháp CRF để xử lý, mô hình SPML cho ra kết quả tốt hơn hết các mô hình

trên với độ chính xác cao hơn 1% so với mô hình BPG, cũng như cho ra kết quả tương quan WvF tốt hơn hầu hết các mô hình trên. Còn với khi đã được xử lý qua CRF, ta cũng thấy rằng SPML cũng cho ra kết quả rất tốt ở độ chính xác cũng như sự tương quan khá là cao hơn so với các mô hình ở trên (76.1% so với 75% của mô hình RL).
Cuối cùng, ta sẽ đánh giá với nhãn **Points**. Dưới đây sẽ là bảng đánh giá độ chính xác với nhãn dán hộp:

TABLE V
KẾT QUẢ SO SÁNH VỚI NHÃN POINTS

Points	val
What's the point[26]	46.1
BCAM [7]	70.5
TEL[27]	74.2
SPML	73.2

Ta có thể thấy rằng, độ chính xác của SPML tuy không cao hơn so với TEL (73.2% so với 74.2%), thế nhưng, về độ linh hoạt khi huấn luyện với các nhãn dán yếu, SPML lại cho ra kết quả ổn định hơn rất nhiều so với TEL. Ta cũng có thể thấy được sự linh hoạt khi huấn luyện qua mô hình SPML ở hình dưới đây:

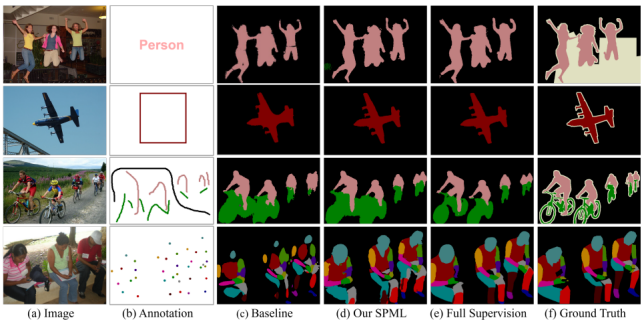


Fig. 3. Kết quả khi sử dụng 4 nhãn dán yếu để phân vùng các hình ảnh

V. TỔNG KẾT

Tóm lại, ta thấy được rằng, với mô hình Semisupervised Pixel-wise Metric Learning trong việc học bán giám sát cho bài toán phân vùng ảnh với giám sát yếu đạt được kết quả khá là tốt, đặc biệt là khi kết hợp với 4 nhãn dán yếu trên. Điều này giúp cho mô hình có thể linh hoạt hơn và đạt hiệu quả tốt hơn khi áp dụng cho các hình ảnh và tập dữ liệu chỉ có 1 hoặc 2 nhãn dán yếu như trên. Ngoài ra, chúng em nghĩ rằng nếu như có thêm thời gian để tìm hiểu cũng như nghiên cứu thêm về mô hình, chúng em có thể cải tiến để có thể cho ra được kết quả tốt hơn bằng cách điều chỉnh các siêu tham số trong mô hình theo một công thức hoặc hàm Loss nào đấy, nhưng chúng em tin rằng, kết quả về độ chính xác của mô hình như trên là hoàn toàn hợp lý và có thể chấp nhận được trong khả năng nghiên cứu của chúng em cũng như theo yêu cầu của đồ án môn học. Chúng em cũng xin cảm ơn thầy đã hướng dẫn chúng em nhiệt tình trong quá trình giảng dạy trên lớp học nhằm giúp chúng em có kiến thức nền tảng tốt để thực hiện đồ án này.

REFERENCES

References

[1] Mark Everingham et al. "The PASCAL Visual Object Classes (VOC) Challenge". English. In: *International Journal of Computer Vision* 88.2 (June 2010), pp. 303–338. ISSN: 0920-5691. DOI: 10.1007/s11263-009-0275-4.

[2] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. "DensePose: Dense Human Pose Estimation in the Wild". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7297–7306. DOI: 10.1109/CVPR.2018.00762.

[3] Liang-Chieh Chen et al. *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*. 2017. arXiv: 1606.00915 [cs.CV].

[4] Agata Lapedriza Aude Oliva Bolei Zhou Aditya Khosla and Antonio Torralba. *Learning deep features for discriminative localization*. 2015. arXiv: 1512 . 04150 [cs.CV].

[5] Ronan Collobert Piotr Dollàr Pedro O. Pinheiro Tsung-Yi Lin. *Learning to Refine Object Segments*. 2016. arXiv: 1603.08695 [cs.CV].

[6] Wanli Ouyang Chunfeng Song Yan Huang and Liang Wang. *Box- driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation*. 2019. arXiv: 1904.11693 [cs.CV].

[7] R Austin McEver and BS Manjunath. *Weakly supervised semantic segmentation using point supervision*. 2020. arXiv: 2007.05615 [cs.CV].

[8] Gustavo Carneiro Erman Tjiputra Quang D. Tran Thanh-Toan Do Binh X. Nguyen Binh D. Nguyen. *Deep Metric Learning Meets Deep Clustering: An Novel Unsupervised Approach for Feature Embedding*. 2020. arXiv: 2009.04091 [cs.CV].

[9] Han Hu Bin Liu Zhirong Wu and Stephen Lin. *Deep metric transfer for label propagation with limited annotated data*. 2018. arXiv: 1812.08781 [cs.CV].

[10] Saining Xie and Zhuowen Tu. *Holistically-Nested Edge Detection*. 2015. arXiv: 1504.06375 [cs.CV].

[11] Pablo Arbeláez et al. "Contour Detection and Hierarchical Image Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.5 (2011), pp. 898–916. DOI: 10.1109/TPAMI.2010.161.

[12] Jyh-Jing Hwang et al. *SegSort: Segmentation by Discriminative Sorting of Segments*. 2019. arXiv: 1910 . 06962 [cs.CV].

[13] Arindam Banerjee et al. "Clustering on the Unit Hypersphere using von Mises-Fisher Distributions". In: *Journal of Machine Learning Research* 6.46 (2005), pp. 1345–1382. URL: http : / / jmlr . org / papers / v6 / banerjee05a.html.

[14] Jacob Goldberger et al. "Neighbourhood components analysis". In: *Advances in neural information processing systems* 17 (2004). URL: https : / / proceedings .

neurips . cc / paper _ files / paper / 2004 / file / 42fe880812925e520249e808937738d2-Paper.pdf.

[15] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[16] Jia Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[17] Zilong Huang et al. “Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7014–7023.

[18] Jungbeom Lee et al. *FickleNet: Weakly and Semi-supervised Semantic Image Segmentation using Stochastic Inference*. 2019. arXiv: 1902 . 10421 [cs.CV].

[19] Bingfeng Zhang et al. *Reliability Does Matter: An End-to-End Weakly Supervised Semantic Segmentation Approach*. 2019. arXiv: 1911.08039 [cs.CV].

[20] Qi Yao and Xiaojin Gong. *Saliency Guided Self-attention Network for Weakly and Semi-supervised Semantic Segmentation*. 2020. arXiv: 1910 . 05475 [cs.CV].

[21] Yu-Ting Chang et al. *Weakly-Supervised Semantic Segmentation via Sub-category Exploration*. 2020. arXiv: 2008.01183 [cs.CV].

[22] Anna Khoreva et al. *Simple Does It: Weakly Supervised Instance and Semantic Segmentation*. 2016. arXiv: 1603.07485 [cs.CV].

[23] Meng Tang et al. *Normalized Cut Loss for Weakly-supervised CNN Segmentation*. 2018. arXiv: 1804 . 01346 [cs.CV].

[24] Meng Tang et al. *On Regularized Losses for Weakly-supervised CNN Segmentation*. 2018. arXiv: 1803 . 09569 [cs.CV].

[25] Bin Wang et al. “Boundary Perception Guidance: A Scribble-Supervised Semantic Segmentation Approach”. In: *International Joint Conference on Artificial Intelligence*. 2019. URL: https://api.semanticscholar.org/CorpusID:199465856.

[26] Amy Bearman et al. *What’s the Point: Semantic Segmentation with Point Supervision*. 2016. arXiv: 1506. 02106 [cs.CV].

[27] Zhiyuan Liang et al. *Tree Energy Loss: Towards Sparsely Annotated Semantic Segmentation*. 2022. arXiv: 2203.10739 [cs.CV].