### Tóm tắt

Bài toán Tìm kiếm người bằng mô tả văn bản (Text-based Person Search - TBPS) nhằm truy xuất hình ảnh người dựa trên mô tả ngôn ngữ tự nhiên. Gần đây, mô hình tiền huấn luyện ngôn ngữ - hình ảnh đa mô thức CLIP (Contrastive Language Image Pretraining) đã thể hiện hiệu suất vượt trội trong nhiều tác vụ nhờ khả năng học ngữ nghĩa liên mô thức mạnh mẽ. TBPS, vốn là một tác vụ truy xuất tinh vi, cũng đang dần chuyển sang hướng nghiên cứu dựa trên CLIP.

Bài viết này là nỗ lực đầu tiên nhằm thực hiện một nghiên cứu thực nghiệm toàn diện về CLIP cho TBPS, từ đó đóng góp một mô hình cơ sở mạnh mẽ – TBPS-CLIP – cho cộng đồng. Chúng tôi xem xét lại các yếu tố thiết kế quan trọng trong CLIP như tăng cường dữ liệu và hàm mất mát. Mô hình với những thiết kế và mẹo huấn luyện đơn giản có thể đạt hiệu suất tốt mà không cần các module phức tạp.

Chúng tôi cũng thực hiện các thí nghiệm đánh giá khả năng khái quát hóa và nén mô hình của TBPS-CLIP, cho thấy hiệu quả của phương pháp từ nhiều góc độ. Công trình này nhằm cung cấp góc nhìn thực nghiệm và mở ra hướng đi cho các nghiên cứu CLIP trong TBPS tương lai.

### 1. Giới thiệu

TBPS là bài toán truy xuất hình ảnh người từ một cơ sở dữ liệu lớn dựa trên mô tả văn bản. Nó ngày càng được quan tâm do có nhiều ứng dụng thực tiễn như tìm nghi phạm, trẻ em mất tích,... Với tính chất truy xuất tinh vi, TBPS đòi hỏi khả năng căn chỉnh và truy xuất dữ liệu liên mô thức hiệu quả.

Các phương pháp hiện tại chủ yếu chiếu thông tin từ văn bản và hình ảnh vào cùng không gian đặc trưng, nhưng lại thường dựa trên các mô hình đơn mô thức (như LSTM/BERT, ResNet-50/ViT), bỏ qua tiềm năng của các mô hình Vision-Language Pretraining (VLP) mạnh mẽ như CLIP.

CLIP, với khả năng mã hóa riêng biệt cho từng mô thức, đã đạt thành tựu trong nhiều bài toán như hỏi-đáp hình ảnh, chú thích hình ảnh, truy xuất ảnh-văn bản,... Gần đây, một số nghiên cứu bắt đầu thử áp dụng CLIP vào TBPS, nhưng lại chủ yếu thêm module phức tạp thay vì khai thác triệt để kiến thức CLIP đã học.

### Đóng góp của bài viết:

Chúng tôi quay về gốc rễ của CLIP và thực hiện nghiên cứu thực nghiệm toàn diện cho TBPS trên hai phương diện:

1. **Tăng cường dữ liệu:**  
   Dù đây là kỹ thuật phổ biến giúp tăng khả năng tổng quát của mô hình, nhưng trong TBPS hầu như chưa được khai thác sâu. Các phương pháp trước chỉ lật ảnh ngẫu nhiên mà không tăng cường văn bản. Bài viết thử nghiệm nhiều chiến lược tăng cường và đề xuất chiến lược hiệu quả nhất cho TBPS.
2. **Hàm mất mát:**  
   Thiết kế hàm mất mát hợp lý là chìa khóa để cải thiện hiệu suất. Chúng tôi thử nghiệm nhiều hàm mất mát phổ biến từ cộng đồng VLP, khác với các hàm mất mát chuyên biệt trong TBPS trước đó.

Bằng cách kết hợp các yếu tố trên và mẹo huấn luyện thực tiễn, chúng tôi xây dựng một mô hình cơ sở mạnh mẽ TBPS-CLIP, gọn nhẹ, dễ huấn luyện và không yêu cầu module phức tạp.

### Các thí nghiệm bổ sung

**a. Khả năng khái quát hóa mô hình:**  
TBPS-CLIP được thử nghiệm như mô hình cơ sở cho TBPS và chứng minh tính hiệu quả của nó, kể cả trong thiết lập few-shot (dữ liệu huấn luyện rất ít).

**b. Nén mô hình:**  
Phân tích đóng góp của từng thành phần vào hiệu suất truy xuất, từ đó định hướng cho việc giảm kích thước mô hình mà vẫn giữ hiệu quả.

### Kết luận

Bài viết đóng góp mô hình cơ sở TBPS-CLIP cho cộng đồng TBPS – một mô hình hiệu suất cao, nhẹ, tiết kiệm chi phí và dễ dùng. Bên cạnh đó, nghiên cứu cũng cho thấy ưu điểm của TBPS-CLIP về khả năng khái quát hóa và nén mô hình. Các thí nghiệm thực nghiệm trong bài hy vọng sẽ thúc đẩy nhiều hướng nghiên cứu TBPS trong tương lai.

**2. Công trình nghiên cứu liên quan**

**2.1 Tìm kiếm người dựa trên văn bản (TBPS)**  
Các phương pháp TBPS truyền thống sử dụng các mô hình huấn luyện sẵn một mô hình đơn như ResNet-50 hoặc ViT làm bộ mã hóa hình ảnh, và LSTM hoặc BERT làm bộ mã hóa văn bản. Những phương pháp này thiết kế các hàm mất mát đặc biệt để học các biểu diễn phân biệt và trích xuất thông tin chi tiết từ hình ảnh và văn bản để căn chỉnh các tính năng đa mô hình. Các công trình gần đây cũng chia hình ảnh thành các dải và phân tích các cụm danh từ để tăng cường việc căn chỉnh tính năng.

Với sự thành công của Huấn luyện trước Mô hình Thị giác-Văn bản (VLP), các nhà nghiên cứu đã bắt đầu khám phá VLP cho TBPS. Một số phương pháp như CLIP kết hợp Bi-GRU hoặc BERT để cải thiện việc mã hóa văn bản, căn chỉnh hình ảnh với các mô tả đa toàn vẹn, hoặc giới thiệu lý thuyết quan hệ ngầm để cải thiện hiệu suất. Những nỗ lực này nhằm tăng cường TBPS bằng cách đẩy mạnh giới hạn của VLP. Tuy nhiên, bài báo này tập trung vào việc khai thác tối đa tiềm năng của CLIP mà không thêm các mô-đun phức tạp, từ đó mang lại hiệu suất TBPS-CLIP cạnh tranh.

**2.2 Huấn luyện trước Thị giác-Văn bản (VLP)**  
VLP đã trở thành một giải pháp hàng đầu cho các nhiệm vụ đa mô hình như mô tả hình ảnh và trả lời câu hỏi thị giác. Các mô hình VLP học được các biểu diễn đa mô hình mạnh mẽ từ các cặp hình ảnh-văn bản lớn, có thể được tinh chỉnh cho các nhiệm vụ khác nhau. Trong số các phương pháp VLP, CLIP nổi bật với hiệu suất mạnh mẽ trên các nhiệm vụ đa mô hình. Một số cải tiến đã được thực hiện để nâng cao hiệu quả dữ liệu của CLIP.

**3. Nghiên cứu thực nghiệm**

Dựa trên CLIP, trong phần 3.1, chúng tôi giới thiệu một số thủ thuật huấn luyện thực tế để củng cố cơ sở CLIP, và sau đó trình bày các phương pháp tăng cường dữ liệu và hàm mất mát trong phần 3.2 và 3.3. Cuối cùng, chúng tôi nghiên cứu khả năng tổng quát của mô hình và nén mô hình trong các phần 3.4 và 3.5. Tổng quan về mô hình được minh họa trong Hình 1.

**3.1 Thủ thuật huấn luyện**  
Chúng tôi nghiên cứu bốn thủ thuật huấn luyện phổ biến liên quan đến CLIP: lan truyền gradient toàn cục, dropout, khóa các lớp dưới và nhãn mềm. Các thủ thuật này được mô tả chi tiết trong phần Phụ lục.

**3.2 Tăng cường dữ liệu (Data Augmentations)**

Image Augementation

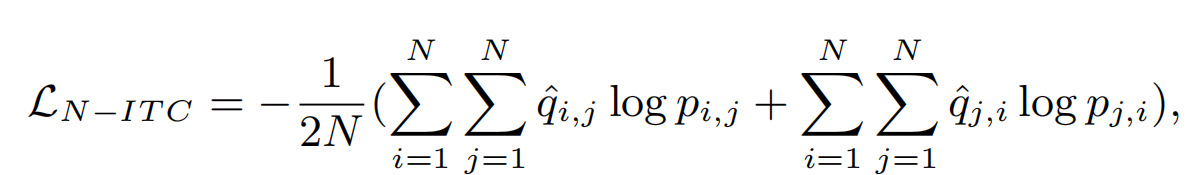
Chúng tôi phân loại tăng cường hình ảnh thành hai nhóm: loại bỏ và thay đổi. Nhóm đầu tiên có các thao tác loại bỏ thông tin từ hình ảnh, bao gồm RandomResizedCrop, RandomErasing, RandomGrayscale và GaussianBlur. Nhóm thứ hai thay đổi màu sắc hoặc hướng của hình ảnh mà vẫn giữ nội dung chính, bao gồm ColorJitter, RandomHorizontalFlip, RandomVerticalFlip và RandomRotation. Các phương pháp tăng cường hình ảnh này được mô tả chi tiết trong Phụ lục.

Ngoài ra, cần lưu ý rằng việc sử dụng đồng thời nhiều phương pháp tăng cường hình ảnh có thể gây ra biến dạng mạnh đối với hình ảnh gốc và làm giảm hiệu suất. Vì vậy, chúng tôi xem xét một loạt các phương pháp tăng cường khác:

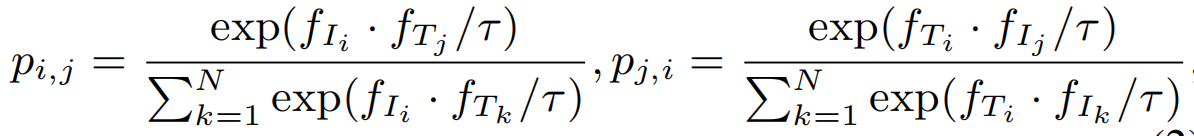
• **AutoAugment** tự động tìm kiếm chính sách tăng cường tốt nhất, sử dụng học tăng cường trong một số chính sách tăng cường.  
• **RandAugment** loại bỏ giai đoạn tìm kiếm trong AutoAugment và chọn ngẫu nhiên từ một tập hợp các thao tác tăng cường.  
• **TrivialAugment** loại bỏ yêu cầu thiết lập tham số trong RandAugment và thay vào đó chọn ngẫu nhiên một thao tác tăng cường và độ mạnh của nó.  
• **An augmentation pool strategy** được thiết kế trong bài báo này, lấy cảm hứng từ các phương pháp tăng cường tự động đã đề cập.

Text Augmentation:  
Khác với tăng cường hình ảnh, có ít phương pháp tăng cường văn bản do tính trừu tượng và rời rạc của ngôn ngữ. Các phương pháp bao gồm dịch ngược, thay thế đồng nghĩa, chèn ngẫu nhiên, hoán đổi ngẫu nhiên, xóa ngẫu nhiên và EDA (trích ngẫu nhiên một phương pháp từ các phương pháp trên và áp dụng vào câu).

**3.3 Hàm mất mát**  
**Các Hàm Mất Mát**  
CLIP trang bị một hàm mất mát đối kháng giữa hình ảnh và văn bản để kéo các mẫu dương lại gần nhau trong khi đẩy các mẫu âm ra xa. Hơn nữa, chúng ta chuẩn hóa nhãn trong hàm mất mát và thu được hàm mất mát đối kháng hình ảnh-văn bản chuẩn hóa (N-ITC):



Trong đó, q^i,j\hat{q}\_{i,j}q^​i,j​ được chuẩn hóa theo qi,j/∑k=1Nqi,kq\_{i,j} / \sum\_{k=1}^{N} q\_{i,k}qi,j​/∑k=1N​qi,k​, và qi,jq\_{i,j}qi,j​ là nhãn đúng (1 cho cặp dương và 0 cho cặp âm). NNN là số lượng mẫu, và pi,jp\_{i,j}pi,j​ đại diện cho nhãn giả, là xác suất của việc khớp hình ảnh IiI\_iIi​ với văn bản TjT\_jTj​ và ngược lại đối với pj,ip\_{j,i}pj,i​:

Trong đó f\* là các đại diện đã chuẩn hóa ℓ2 của mẫu, và τ là tham số nhiệt độ có thể học.

Ngoài N-ITC, chúng tôi nghiên cứu các hàm mất mát khác theo hai hướng. Một hướng tập trung vào việc nâng cao hiệu quả dữ liệu, và hướng còn lại nhắm tới việc tối ưu hóa mối quan hệ giữa các mẫu.

Cải thiện hiệu quả dữ liệu:  
Hàm mất mát tự giám sát (SS) nhằm tối đa hóa sự tương đồng giữa hai phiên bản khác nhau của hình ảnh và thúc đẩy học các đại diện tính năng mạnh mẽ từ dữ liệu hình ảnh hạn chế. Các phương pháp khác bao gồm giám sát đa góc (multi-view supervision), sử dụng nhiều góc độ khác nhau của dữ liệu để tăng cường giám sát và khả năng khai thác dữ liệu.

Tối ưu hóa cho việc truy xuất:  
Hàm mất mát đối sánh hình ảnh-văn bản ngược (R-ITC) bổ sung cho N-ITC bằng cách tối ưu hóa sự phân tách giữa các cặp tiêu cực và khuyến khích sự tương đồng của các cặp tích cực.

**3.4 Khả năng tổng quát hóa của mô hình**  
Bên cạnh các nghiên cứu thực nghiệm về tăng cường dữ liệu và hàm mất mát, chúng tôi còn chứng minh khả năng tổng quát hóa của mô hình từ hai khía cạnh. Cụ thể, trước tiên chúng tôi áp dụng TBPS-CLIP được đề xuất vào các phương pháp TBPS khác để kiểm tra tính tổng quát của TBPS-CLIP khi dùng làm mô hình nền. Sau đó, chúng tôi tinh chỉnh TBPS-CLIP trên một lượng nhỏ dữ liệu huấn luyện TBPS nhằm kiểm tra khả năng tổng quát trong thiết lập học ít mẫu (few-shot).

Chúng tôi cũng nhận thấy rằng đã có một số biến thể CLIP được đề xuất nhằm nâng cao khả năng học ít mẫu của CLIP. Do đó, chúng tôi thực hiện nghiên cứu thực nghiệm trên hai phương pháp học ít mẫu tiêu biểu là **CoOp** (Zhou et al. 2022) và **CLIP-Adapter** (Gao et al. 2021) để so sánh hiệu năng. Tổng quan về hai phương pháp này được trình bày trong Phụ lục.

**3.5 Nén mô hình**  
Chúng tôi cung cấp cái nhìn sâu hơn về mô hình bằng cách đánh giá vai trò của từng module trong hiệu năng cuối cùng — điều này rất hữu ích cho việc nén mô hình trong các ứng dụng thực tế.

Chúng tôi áp dụng hai tiêu chí đánh giá (Wang và Tu 2020):

Tiêu chí thứ nhất đánh giá mức độ đóng góp của một module bằng cách **loại bỏ module đó** và quan sát mức giảm hiệu năng.

Tiêu chí thứ hai đánh giá tầm quan trọng bằng cách đo **mức độ gần với giá trị khởi tạo** mà các trọng số module có thể đạt được mà vẫn giữ được một mức hiệu năng nhất định.

Chi tiết về hai tiêu chí này được trình bày trong Phụ lục.

**4. Thí nghiệm**  
Phần này trình bày chi tiết các phân tích thực nghiệm trong các nghiên cứu đã nêu. Dù nghiên cứu của chúng tôi sử dụng các kỹ thuật phổ thông để giữ mô hình đơn giản, chúng tôi cũng cung cấp góc nhìn chuyên biệt cho TBPS, tức là bàn luận lý do vì sao các kỹ thuật này lại hiệu quả trong TBPS.

Chúng tôi tiến hành so sánh với các phương pháp khác trên ba bộ dữ liệu chuẩn: **CUHK-PEDES** (Li et al. 2017b), **ICFG-PEDES** (Ding et al. 2021), và **RSTPReid** (Zhu et al. 2021), trong khi các thí nghiệm tách riêng (ablation study) được thực hiện chủ yếu trên CUHK-PEDES.

Chúng tôi sử dụng các chỉ số đánh giá hiệu năng gồm **Rank-k** và **mean Average Precision (mAP)**. Giới thiệu chi tiết về bộ dữ liệu, tiêu chí đánh giá và thông số cài đặt được trình bày trong Phụ lục.

**Bảng 1: Thử nghiệm tách biệt các mẹo huấn luyện trên CUHK-PEDES**  
(CLIP\* đại diện cho CLIP được áp dụng đầy đủ bốn mẹo huấn luyện)

| **Phương pháp** | **Rank-1** | **Rank-5** | **Rank-10** | **mAP** |
| --- | --- | --- | --- | --- |
| CLIP | 60.67 | 81.99 | 88.87 | 54.72 |
| CLIP + GlobalGrad | 63.66 | 84.08 | 90.11 | 56.46 |
| CLIP + Dropout | 61.06 | 82.00 | 88.95 | 55.10 |
| CLIP + LockBL | 61.27 | 82.23 | 88.79 | 55.43 |
| CLIP + SLabel | 61.53 | 81.99 | 88.71 | 55.22 |
| **CLIP\*** | **64.34** | **84.05** | **90.51** | **57.53** |

**4.1 Thử nghiệm tách biệt các mẹo huấn luyện**  
Xuất phát từ mô hình CLIP gốc, chúng tôi tiến hành các thử nghiệm thực nghiệm với các mẹo huấn luyện được trình bày trong Mục 3.1. Như thể hiện trong Bảng 1, tất cả các mẹo huấn luyện đều mang lại ảnh hưởng tích cực đến hiệu suất của CLIP.

Dưới đây là bản dịch tiếng Việt của đoạn văn từ mục 4.2 đến hết phần so sánh SOTA:

**4.2 Loại bỏ từng yếu tố trong Tăng cường Dữ liệu**  
Chúng tôi trình bày kết quả tối ưu của từng kỹ thuật tăng cường và phân tích tác động của chúng. Thí nghiệm loại bỏ với các siêu tham số khác nhau được trình bày trong Phụ lục.

**Tăng cường Hình ảnh.**  
Bảng 2 phân tích ảnh hưởng của các kỹ thuật tăng cường hình ảnh.  
(1) Với nhóm loại bỏ, các kỹ thuật như RandomResizedCrop và RandomErasing đều giúp tăng hiệu suất. Cả hai đều xóa/cắt ngẫu nhiên một phần ảnh, giúp làm nổi bật chi tiết cục bộ và gián tiếp thúc đẩy học sâu liên kết đa phương thức trong bài toán TBPS. Bất ngờ thay, RandomGrayscale—kỹ thuật loại bỏ hoàn toàn thông tin màu—cũng cải thiện kết quả. Khi hình ảnh không có màu được đưa vào mô hình, mô hình buộc phải chú trọng vào thông tin khác như kết cấu và hình dạng. Dù màu sắc rất quan trọng trong TBPS (Wu et al. 2021; Wang et al. 2022c), kết quả thực nghiệm cho thấy các yếu tố khác ngoài màu cũng có giá trị trong việc truy xuất người. Ngược lại, GaussianBlur—làm mờ chi tiết của ảnh—làm giảm đáng kể hiệu suất do mất thông tin chi tiết cần thiết.

(2) Với nhóm thay đổi, đa số các kỹ thuật đều có lợi cho hiệu suất, gồm ColorJitter-BCS, RandomHorizontalFlip, và RandomRotation. Chúng làm phong phú dữ liệu mà không thay đổi ngữ nghĩa ảnh, tăng độ bền mô hình, dẫn đến kết quả tốt hơn. Ngược lại, ColorJitter-Hue và RandomVerticalFlip lại gây hại vì làm thay đổi màu hoặc hình dạng thô của ảnh, khiến mô hình khó khớp ảnh và mô tả văn bản.

(3) Ngoài việc áp dụng từng kỹ thuật riêng lẻ, chúng tôi còn thử nghiệm tổ hợp nhiều kỹ thuật. Việc kết hợp các kỹ thuật hiệu quả (“Stacking Together”) giúp tăng hiệu suất. Hơn nữa, chúng tôi còn thử các chiến lược tăng cường tự động như AutoAugment (Cubuk et al. 2018), RandAugment (Cubuk et al. 2020) và TrivialAugment (Müller và Hutter 2021). Tuy nhiên, tập tăng cường đề xuất (chọn ngẫu nhiên 2 kỹ thuật hiệu quả mỗi lần) cho kết quả tốt nhất.

**Tăng cường Văn bản.**  
Như Bảng 3 cho thấy, các kỹ thuật tăng cường hiệu quả là dịch ngược và xóa ngẫu nhiên. Dịch ngược làm đa dạng biểu đạt văn bản gốc, còn xóa ngẫu nhiên đóng vai trò điều chuẩn, loại bỏ từ không cần thiết. Kết hợp hai kỹ thuật này giúp tăng 1.38% Rank-1. Ngược lại, các kỹ thuật như thay từ đồng nghĩa, chèn ngẫu nhiên, tráo từ đều làm giảm hiệu suất do phá vỡ cấu trúc câu và ý nghĩa gốc, gây khó cho mô hình. Do đó, EDA (chọn ngẫu nhiên 1 trong các kỹ thuật trên) không mang lại cải thiện rõ rệt.

**Kết hợp các Tăng cường Tối ưu.**  
Theo Bảng 4, khi kết hợp tất cả kỹ thuật tăng cường tối ưu (Augmentation Pool cho ảnh, Stacking Together cho văn bản), độ chính xác Rank-1 tăng đáng kể 2.44%, cho thấy chỉ riêng tăng cường dữ liệu đã mang lại lợi ích đáng kể.

**4.3 Loại bỏ từng yếu tố trong Hàm Mất mát**  
Bảng 5 đánh giá hiệu quả của các hàm mất mát.  
(1) So với CLIP\*+Aug (dùng hàm mất mát so khớp ảnh-văn bản gốc), việc thay bằng hàm N-ITC (phiên bản chuẩn hóa) giúp cải thiện nhẹ 0.13% ở Rank-1.  
(2) Trong nhóm tăng hiệu quả sử dụng dữ liệu, SS-I (tự giám sát ảnh) cho kết quả tốt nhất so với SS-T và SS-IT. Tương tự, MVS-I (giám sát đa góc nhìn ảnh) vượt trội hơn MVS-T và MVS-IT. Điều này cho thấy khai thác dữ liệu hình ảnh hiệu quả hơn văn bản trong TBPS.  
(3) Nhóm tối ưu cho truy xuất: R-ITC và C-ITC đều tăng hiệu suất. Đặc biệt, R-ITC tăng mạnh 1.41% so với CLIP+Aug, nhờ giới hạn đẩy xa các mẫu âm. Cuối cùng, kết hợp tất cả các hàm mất mát hiệu quả giúp tăng đáng kể 2.76% Rank-1, đạt 69.54%.

**Bảng 6: So sánh với các phương pháp SOTA trên bộ dữ liệu CUHK-PEDES**  
(Trích đoạn các phương pháp không dùng CLIP và có CLIP.)

Các phương pháp **không dùng CLIP** như ViTAA, DSSL, SSAN... đạt Rank-1 khoảng 55–65%.  
Phương pháp **RaSa (Bai et al. 2023)** là nổi bật nhất trong nhóm này với Rank-1 đạt 76.51%.

Các phương pháp **dùng CLIP** như TBPS-LD... (phần còn lại bị cắt trong ảnh).

**4.4 So sánh với các phương pháp hiện đại nhất**

Chúng tôi so sánh TBPS-CLIP với các phương pháp hiện đại nhất trên các bộ dữ liệu CUHK-PEDES, ICFG-PEDES và RSTPReid, được trình bày lần lượt ở Bảng 6, Bảng 7 và Bảng 8.

(1) So với các phương pháp sử dụng CLIP, TBPS-CLIP đề xuất (với ViT-B/16 làm bộ mã hóa hình ảnh) vượt trội hơn phương pháp hiện đại nhất IRRA trên ICFG-PEDES và RSTPReid, và đạt kết quả tương đương trên CUHK-PEDES. Lưu ý rằng IRRA bổ sung bộ mã hóa tương tác đa mô hình sau CLIP để đạt hiệu suất cao hơn, trong khi TBPS-CLIP giữ nguyên kiến trúc mạng đơn giản ban đầu (kiểu hai luồng của CLIP) nhưng vẫn đạt kết quả đầy hứa hẹn. Bảng 9 cho thấy TBPS-CLIP nhẹ hơn nhiều so với IRRA. TBPS-CLIP khuyến khích CLIP khai thác dữ liệu hiệu quả, cho phép huấn luyện chỉ trong 5 epoch. Việc huấn luyện hiệu quả cao khiến nó trở thành một baseline rất thân thiện.

(2) So với các phương pháp không sử dụng CLIP, có thể thấy RaSa hoạt động rất tốt. Nó dùng ALBEF làm nền tảng và gồm hai mô hình: mô hình online và phiên bản đà (momentum), mỗi cái đều có bộ mã hóa hình ảnh, mã hóa văn bản và mã hóa tương tác đa mô hình. Dù có hiệu suất cao, RaSa lại cồng kềnh và khó mở rộng, như đã thấy trong Bảng 9. TBPS-CLIP với kiến trúc nhẹ, chi phí thấp và hiệu năng tốt, có tiềm năng làm baseline cho các ứng dụng rộng hơn.

(3) Để nâng cao tính tiện lợi của TBPS-CLIP như một baseline, chúng tôi giới thiệu thêm phiên bản rút gọn, chỉ gồm mất mát N-ITC và R-ITC. Phiên bản này vẫn đạt hiệu suất tốt, thậm chí vượt IRRA trên ICFG-PEDES và RSTPReid. TBPS-CLIP rút gọn dễ dàng được ứng dụng làm baseline trong các nghiên cứu tiếp theo.

**Bảng 9:** So sánh với các phương pháp hiện đại gần đây trên CUHK-PEDES. Param.(M) là số lượng tham số (triệu), Epoch là số vòng lặp huấn luyện, Time(s) là thời gian chạy trực tuyến.

| **Phương pháp** | **Nền tảng** | **Param.(M)** | **Epoch** | **Training Time** | **Test Time** |
| --- | --- | --- | --- | --- | --- |
| RaSa (2023a) | ALBEF | 210.2 | 30 | 27967.5s | 869.8s |
| IRRA (2023) | CLIP (ViT-B/16) | 194.5 | 60 | 6110.4s | 31.4s |
| **TBPS-CLIP (chúng tôi)** | CLIP (ViT-B/16) | **149.2** | **5** | **1234.7s** | **31.4s** |

**Bảng 10:** Kết quả khi sử dụng TBPS-CLIP làm baseline cho phương pháp khác trên CUHK-PEDES. Sử dụng chỉ số mINP để so sánh.

| **Phương pháp** | **Nền tảng** | **Rank-1** | **Rank-5** | **Rank-10** | **mAP** | **mINP** |
| --- | --- | --- | --- | --- | --- | --- |
| IRRA | CLIP | 73.38 | 89.93 | 93.71 | 66.13 | 50.24 |
| IRRA + **TBPS-CLIP** | TBPS-CLIP | **74.97** | 89.82 | 93.80 | **67.84** | **52.53** |
| IRRA + **TBPS-CLIP (rút gọn)** | TBPS-CLIP | 74.56 | 89.26 | 93.52 | 67.52 | 52.58 |

**Hình 3:** Biểu đồ xu hướng hiệu suất khi loại hoặc đóng băng một số tầng của bộ mã hóa văn bản trong TBPS-CLIP và phiên bản rút gọn.

**Bảng 11:** Hiệu suất trong thiết lập học với ít dữ liệu (5% dữ liệu huấn luyện) trên CUHK-PEDES.

| **Phương pháp** | **Rank-1** | **Rank-5** | **Rank-10** | **mAP** |
| --- | --- | --- | --- | --- |
| CLIP | 38.37 | 62.26 | 72.34 | 35.39 |
| CoOp | 11.37 | 24.19 | 32.46 | 10.48 |
| CLIP-Adapter | 11.96 | 25.45 | 33.56 | 10.91 |
| **TBPS-CLIP** | **42.98** | **66.26** | **74.94** | **38.86** |
| **TBPS-CLIP (rút gọn)** | **43.65** | **66.60** | **75.91** | **39.30** |

**4.5 Các thí nghiệm sâu hơn về TBPS-CLIP**

**Khả năng khái quát hóa mô hình:**

(1) Chúng tôi chọn IRRA là đại diện cho phương pháp CLIP tiên tiến nhất hiện nay để kiểm tra khả năng khái quát hóa của TBPS-CLIP. Cụ thể, thay vì dùng CLIP gốc, chúng tôi áp dụng TBPS-CLIP (ViT-B/16) và bản rút gọn của nó làm baseline cho IRRA. Kết quả ở Bảng 15 (Phụ lục) cho thấy rõ hiệu quả và khả năng khái quát hóa của TBPS-CLIP.

(2) Chúng tôi cũng khảo sát khả năng học với ít dữ liệu (5%) ở Bảng 16, thêm kết quả với 1% và 10% được trình bày trong Phụ lục. CLIP và các biến thể như CoOp, CLIP-Adapter hoạt động kém trong bài toán TBPS ít dữ liệu do khác biệt lớn giữa dữ liệu huấn luyện ban đầu và đặc thù bài toán TBPS. Ngược lại, TBPS-CLIP với khả năng học mạnh mẽ vẫn đạt hiệu suất tốt, đặc biệt là bản rút gọn trong thiết lập ít dữ liệu.

**Nén mô hình:**

Chúng tôi tính toán hai chỉ số hướng dẫn nén mô hình TBPS-CLIP (trình bày ở Mục 3.5). Chi tiết cách tính toán và phân tích thuộc tính bên trong được nêu trong Phụ lục. Hình 3 cho thấy đóng băng một phần tầng mã hóa văn bản ít ảnh hưởng đến hiệu suất, trong khi loại bỏ tầng làm giảm hiệu quả. Do đó, có thể nén mô hình bằng cách đóng băng một phần bộ mã hóa văn bản trong quá trình huấn luyện.

**5 Kết luận**

Bài báo này thực hiện nghiên cứu thực nghiệm toàn diện để khai thác tiềm năng của CLIP cho bài toán TBPS. Chúng tôi chứng minh rằng CLIP, khi kết hợp với các tăng cường dữ liệu và hàm mất mát phổ biến, cùng vài thủ thuật huấn luyện đơn giản, có thể đạt hiệu quả tốt mà không cần module phức tạp. TBPS-CLIP được xác thực từ hai khía cạnh: khả năng khái quát hóa và khả năng nén mô hình, mở đường cho nghiên cứu TBPS dựa trên CLIP trong tương lai.