**Abstract**

Trong bài báo này, chúng tôi giới thiệu một **kiến trúc mã hóa-giải mã (encoder-decoder)** khai thác cả **ngữ nghĩa toàn cục và cục bộ** để giải quyết bài toán **tự động tô màu ảnh**. Đối với ngữ nghĩa toàn cục, các đặc trưng mã hóa cấp thấp được tinh chỉnh bằng cách phân loại ngữ cảnh cảnh quan để tích hợp phong cách tổng thể của hình ảnh. Hơn nữa, kiến trúc này xử lý sự không chắc chắn và mối quan hệ giữa các phong cách cảnh quan dựa trên kỹ thuật làm mịn nhãn (label smoothing) và trọng số được huấn luyện trước từ Places365.

Đối với ngữ nghĩa cục bộ, ba nhánh học cách tận dụng lợi ích chung ở cấp độ điểm ảnh. Cụ thể, các phân phối trung bình và đa mô hình lần lượt được tạo ra từ nhánh hồi quy và mã hóa mềm (soft-encoding), trong khi nhánh phân đoạn xác định điểm ảnh thuộc về đối tượng nào.

Các thí nghiệm của chúng tôi, sử dụng bộ dữ liệu **Coco-Stuff** để huấn luyện và đánh giá trên DIV2K, Places365 và ImageNet, cho thấy kết quả thu được rất khả quan.

INTRODUCTION

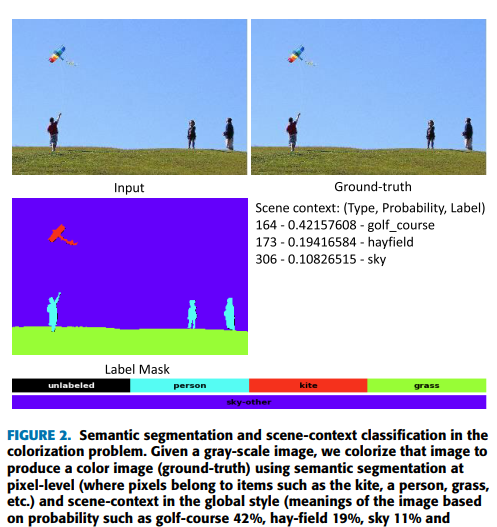
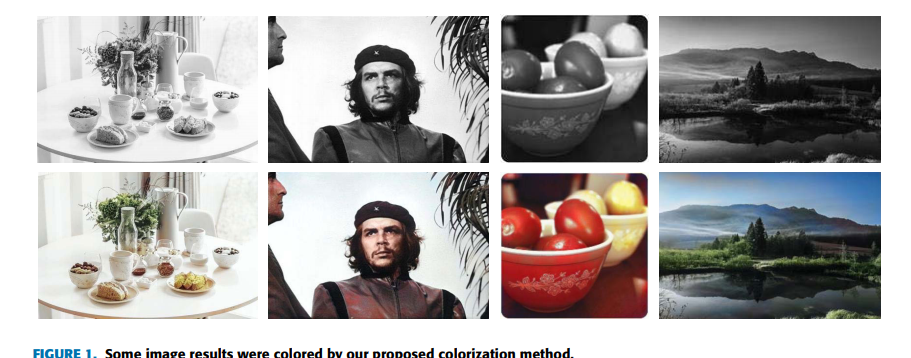
Việc tô màu một hình ảnh xám không chỉ mang lại nhiều ý nghĩa đặc biệt cho hình ảnh đó mà còn giúp nó trở nên sống động và có cảm xúc hơn [1]. Một hình ảnh thường chứa nhiều đối tượng, và con người có thể dễ dàng quyết định màu sắc phù hợp cho từng đối tượng trong ảnh nhờ vào kiến thức về ý nghĩa của đối tượng. Tuy nhiên, hầu hết các đối tượng không chỉ có một màu duy nhất; ví dụ, một chiếc áo có thể có màu đỏ, xanh, vàng hoặc nhiều màu khác. Con người cũng dự đoán màu sắc của đối tượng dựa trên cảm xúc chủ quan của mình, đây là một trong những thách thức lớn nhất đối với một hệ thống máy tính khi cố gắng bắt chước cả kiến thức lẫn cảm nhận của con người.

Vì lý do này, các hệ thống tô màu trước đây thường yêu cầu sự hợp tác với người dùng, chẳng hạn như cung cấp thông tin ban đầu bằng cách nhập các điểm màu [2] hoặc chuyển màu từ các hình ảnh xám tham chiếu. Tuy nhiên, việc chọn một hình ảnh chứa các chi tiết đối tượng chính xác mất quá nhiều thời gian, cũng như việc vẽ các điểm màu.

Do đó, phương pháp tô màu tự động dựa trên mạng nơ-ron tích chập sâu (CNN) mà không cần sự can thiệp của con người đã được đề xuất [3]–[6]. Nhờ vào những lợi ích to lớn của hệ thống tô màu tự động, các nhà nghiên cứu đã tìm ra nhiều cách để giúp máy tính hiểu thông tin ngữ nghĩa của hình ảnh trong quá trình xử lý nhằm đạt được kết quả tô màu hiệu quả.

Một phương pháp điển hình là nghiên cứu của [7], trong đó sử dụng phân loại cảnh từ cơ sở dữ liệu Places365 để huấn luyện mô hình về ngữ nghĩa toàn cục của hình ảnh. Hơn nữa, nghiên cứu [8] tập trung vào cả bố cục ngữ nghĩa và các đối tượng cục bộ của cảnh để tô màu ảnh bằng kiến trúc mạng VGG-16. Trong phương pháp của Zhang et al. [9], họ bổ sung khả năng tổng hợp ngữ nghĩa cấp đối tượng bằng cách mã hóa liên kênh vào mô hình của mình.

Trong các nghiên cứu trên, ngữ nghĩa chỉ được khai thác ở cấp độ hình ảnh. Ngược lại, các nghiên cứu hiện tại về phân đoạn ngữ nghĩa đã đạt đến cấp độ điểm ảnh bằng cách gán nhãn cho từng điểm ảnh [10]. Theo cách tương tự, tô màu cũng là một nhiệm vụ liên quan đến việc gán màu cho từng điểm ảnh dựa trên một phân phối xác suất màu sắc [11]. Để cải thiện phương pháp tô màu bằng cách kết hợp các khái niệm này, ngoài ý nghĩa cảnh quan, chúng tôi nhận thấy rằng nó một cách hợp lý hơn.



Nhiệm vụ phân đoạn ngữ nghĩa đóng vai trò quan trọng trong việc cung cấp thông tin ngữ nghĩa ở cấp độ điểm ảnh.

Kinh nghiệm sống của chúng ta có vai trò thiết yếu trong việc hình thành mô hình nhận thức thông tin. Thông tin, bao gồm cả loại toàn cục và cục bộ, được lưu trữ trong tiềm thức và dần dần hình thành kiến thức theo thời gian. Khi nhìn vào một hình ảnh, chúng ta có thể phân tích loại cảnh, xác định các đối tượng có thể xuất hiện, chất liệu của chúng, và nhiều yếu tố khác. Kiến thức về màu sắc là một dạng thông tin mà con người tích lũy trong suốt cuộc đời. Do đó, việc tô màu cho một bức ảnh xám là một nhiệm vụ dễ dàng đối với con người.

Tuy nhiên, có nhiều thách thức khi để một máy tính tự động chọn màu sắc chính xác để tô màu cho ảnh đen trắng. Máy tính chỉ có thể tính toán dự đoán dựa trên mối quan hệ giữa ánh sáng và màu sắc mà không xét đến ngữ nghĩa của đối tượng, điều này có thể dẫn đến sự nhầm lẫn về màu sắc. Ngoài ra, do các hình ảnh đầu vào có nhiều kích thước khác nhau, không thể tránh khỏi sự thay đổi tỷ lệ, gây ra hiện tượng lem màu giữa các vùng biên và nhiễu màu.

Một ví dụ về quá trình này được minh họa trong Hình 2, trong đó người xem có thể quan sát một hình ảnh xám và dễ dàng suy ra một số thông tin tổng quan, chẳng hạn như đây là một cảnh ngoài trời có bầu trời và cánh đồng—đây được gọi là thông tin ngữ nghĩa cảnh quan hay thông tin toàn cục. Bên cạnh đó, các đối tượng như con người, diều, cỏ và bầu trời được coi là thông tin cục bộ và có thể được xác định thông qua phân đoạn ngữ nghĩa.

Sự kết hợp giữa kiến thức toàn cục (ngữ nghĩa cảnh quan) và kiến thức cục bộ (phân đoạn ngữ nghĩa) sẽ giúp người quan sát xác định đối tượng trong hình ảnh và tham chiếu mô hình đối tượng để xác định màu sắc phù hợp. Vì lý do này, chúng tôi đề xuất một giải pháp sử dụng cả phân đoạn và ngữ nghĩa cảnh quan để tự động tô màu hình ảnh, như minh họa trong Hình 1.

Ngoài ra, chúng tôi cũng quan tâm đến việc lựa chọn một mạng phù hợp để thúc đẩy sự kết hợp giữa các yếu tố này. Trong bài báo của mình, chúng tôi sử dụng kiến trúc encoder-decoder, như được trình bày trong Hình 3, do tính linh hoạt của nó trong học đa nhiệm (multi-task learning). Các đặc trưng mã hóa cấp thấp sẽ được tinh chỉnh thông qua nhánh phân loại ngữ cảnh cảnh quan, nằm ở giữa hai nhánh mã hóa (encoder) và giải mã (decoder). Điều này có nghĩa là chúng không chỉ chứa dữ liệu mã hóa của hình ảnh đầu vào mà còn tích hợp phong cách tổng thể của hình ảnh, giúp nhánh giải mã thực hiện tô màu một cách hợp lý hơn.

Đề xuất một kiến trúc encoder-decoder sử dụng phân loại ngữ cảnh cảnh quan và phân đoạn cấp điểm ảnh để tô màu hình ảnh.

Hình ảnh sẽ được tô màu chính xác hơn dựa trên ngữ nghĩa toàn cục của ngữ cảnh cảnh quan. Đối với ngữ nghĩa cấp điểm ảnh, đầu ra của nhánh giải mã (decoder) dưới dạng các đặc trưng ánh xạ điểm ảnh 2D sẽ được tinh chỉnh thông qua ba nhánh: phân đoạn, hồi quy (regression), và mã hóa mềm (soft-encoding).

**Ba nhánh này học cách tận dụng lợi ích chung, trong đó:**

**Phân đoạn** giúp xác định ý nghĩa của từng điểm ảnh,  
 **Hồi quy** tạo mô hình màu trung bình,

**Mã hóa mềm** sử dụng mô hình màu có độ lệch để xây dựng các đặc trưng ánh xạ điểm ảnh 2D mạnh mẽ hơn.

Trong quá trình triển khai, chúng tôi sử dụng mô hình U-Net [12] với cơ chế kết nối bỏ qua (skip connection) để giúp mô hình hội tụ tốt hơn và tránh vấn đề mất dần thông tin (vanishing problem) do kích hoạt chết và độ sâu mạng quá lớn.

Đối với ngữ nghĩa ngữ cảnh cảnh quan, chúng tôi xử lý sự không chắc chắn và mối quan hệ giữa các cảnh bằng kỹ thuật làm mịn nhãn (label smoothing), sử dụng xác suất tiên nghiệm từ trọng số được huấn luyện trước trong Places365.

Đối với ngữ nghĩa cấp điểm ảnh, chúng tôi tích hợp học lợi ích chung từ mô hình phân đoạn cấp điểm ảnh, mô hình màu trung bình và mô hình màu có độ lệch.

Chúng tôi huấn luyện kiến trúc này trên bộ dữ liệu Coco-Stuff [13] và kiểm tra trên DIV2K [14], Places365 [15], ImageNet [16], và các hình ảnh thu thập từ Internet để đảm bảo kết quả đạt chất lượng tốt.

Phần còn lại của bài báo bao gồm năm phần:

**Phần 2**: Trình bày các nghiên cứu trước liên quan đến bài toán tô màu ảnh.

**Phần 3**: Mô tả kiến trúc mạng, quy trình xử lý dữ liệu đầu vào, đầu ra và thông tin ngữ nghĩa, cũng như hàm mất mát đa nhiệm.

**Phần 4**: Mô tả chi tiết quá trình triển khai phương pháp đề xuất.

**Phần 5**: Trình bày chi tiết các thí nghiệm và thảo luận kết quả.

**Phần 6**: Kết luận bài báo.

**RELATED WORKS**

Việc tô màu cho những bức ảnh đen trắng cũ nhằm tái tạo lại các cảnh lịch sử có thể giúp lưu giữ nhiều giá trị văn hóa và kỷ niệm trước sự tàn phá của thời gian, đồng thời mang những người hiện đại đến gần hơn với quá khứ. Tô màu ảnh vẫn là một trong những chủ đề hấp dẫn đối với người dùng phần mềm chỉnh sửa ảnh, điều này được thể hiện qua nhiều hướng dẫn tô màu có trên Internet. Tuy nhiên, tô màu bằng phần mềm chỉnh sửa ảnh rất tốn kém và mất nhiều thời gian. Một nghệ sĩ kỹ thuật số phải bỏ ra nhiều công sức để tô màu ảnh đen trắng, bắt đầu bằng việc tạo nhiều lớp cho các vùng khác nhau của hình ảnh, sau đó gán và điều chỉnh màu sắc cho các vùng này sao cho phù hợp với kiến thức và trí tưởng tượng của họ.

Các nhà nghiên cứu máy tính đã phát triển nhiều ứng dụng tô màu nhằm giải quyết những khó khăn khác nhau liên quan đến quá trình tô màu này.

Các nghiên cứu trước đây thường đề cập đến ba phương pháp tô màu truyền thống. Loại thứ nhất là dựa trên phác thảo màu (scribble-based), trong đó người dùng chú thích màu lên ảnh; loại thứ hai là sử dụng màu từ ảnh tham chiếu để áp dụng vào ảnh mục tiêu; và loại cuối cùng là tô màu tự động dựa trên học sâu (deep learning).

**TÔ MÀU DỰA TRÊN PHÁC THẢO MÀU (Scribble-Based)**

Đây là phương pháp tô màu sử dụng các nét phác thảo màu ban đầu và tự động lan truyền màu sắc đến các pixel lân cận có màu tương tự. Levin et al. [2] đã đề xuất phương pháp này vào năm 2004 dựa trên giả thuyết rằng các pixel xung quanh có cùng độ sáng sẽ có màu tương tự. Tuy nhiên, kết quả của họ cho thấy tình trạng màu bị lan tràn giữa các đối tượng. Sau đó, phương pháp được mô tả trong [17] đã khắc phục tình trạng này bằng cách khai thác thông tin về cạnh ở mức cục bộ để ngăn màu bị lan qua ranh giới giữa các đối tượng. Một số phương pháp như [18] và [19] đã cải thiện hiệu quả bằng cách truyền màu không chỉ đến các pixel lân cận mà còn đến các pixel có cùng mẫu. Mặc dù hình ảnh được tô màu bằng các phương pháp này trông tự nhiên và có màu sắc chi tiết rõ ràng, nhưng chúng đòi hỏi nhiều thời gian và công sức.

**TÔ MÀU DỰA TRÊN ẢNH THAM CHIẾU (Example-Based)**

Thay vì vẽ nhiều nét phác thảo, phương pháp tô màu dựa trên ảnh tham chiếu liên quan đến việc chuyển màu từ các hình ảnh tương tự sang ảnh mục tiêu. Một số phương pháp này đo lường sự tương đồng giữa ảnh tham chiếu và ảnh mục tiêu ở mức pixel. Cụ thể, Charpiat et al. [3] đã chọn một số hình ảnh màu liên quan làm đầu vào, sau đó xây dựng xác suất phân bố của các màu hợp lý cho từng pixel ở mức cục bộ, **cuối cùng tối ưu hóa xác suất của ảnh tô màu bằng thuật toán cắt đồ thị (graph-cut).** Với cùng mục tiêu, Welsh et al. [20] đã chuyển màu từ ảnh tham chiếu sang cả ảnh tĩnh và video.

Khác với hai phương pháp trên, Liu et al. [4] đã chuẩn bị ảnh tham chiếu bằng cách tự động tìm kiếm các bức ảnh phù hợp trên Internet. Ở mức đoạn ảnh (segment level), Irony et al. [21] đã tính toán màu sắc trên từng đoạn của ảnh tham chiếu có thể được chuyển sang pixel của ảnh mục tiêu. Hơn nữa, phương pháp trong [22] đã thêm chú thích bằng tay vào quá trình phân đoạn để lọc ra các mẫu tham chiếu phù hợp.

Mặc dù các phương pháp này tạo ra kết quả đẹp và tự nhiên, nhưng chúng vẫn mất nhiều thời gian để tìm kiếm hình ảnh tham chiếu phù hợp, ngay cả khi sử dụng các công cụ tìm kiếm tự động trên web.

**TÔ MÀU DỰA TRÊN HỌC SÂU**

Gần đây, với sự phát triển của các ứng dụng thị giác máy tính, tô màu ảnh ngày càng đóng vai trò quan trọng và liên tục được cải tiến để đáp ứng nhu cầu của người dùng. Nó đã được áp dụng để hỗ trợ các nhiệm vụ nâng cao, chẳng hạn như nghiên cứu của [23], **sử dụng thuộc tính màu sắc để cải thiện hiệu suất phát hiện đối tượng**, và nghiên cứu của [24], **phát triển ứng dụng chỉnh sửa màu để tạo ra kết quả tối ưu hơn.**

Gần đây nhất, trong ngành công nghiệp truyện tranh, **các phương pháp tô màu đã được phát triển mạnh mẽ nhằm giảm đáng kể chi phí và công sức** [9], [25]–[27].

Để giảm bớt công việc thủ công của các phương pháp trước đó, quá trình tô màu đã tận dụng học sâu để học cách dự đoán màu sắc. Ví dụ, nghiên cứu của [6] đã đề xuất một kỹ thuật phân cụm hình ảnh cho một tập dữ liệu lớn và xây dựng một kiến trúc mạng nơ-ron sâu để đào tạo với các bộ mô tả đặc trưng của hình ảnh.

Gần đây, việc đào tạo mạng dựa trên ngữ nghĩa (semantics training) đã được khai thác để cải thiện hiệu suất tốt hơn. Ví dụ, Iizuka et al. [7] đã xây dựng hai nhánh mạng dựa trên mạng nơ-ron tích chập sâu (deep convolutional neural networks) để kết hợp các đặc trưng toàn cục và cục bộ.

Để hỗ trợ ngữ nghĩa cho mạng, Zhang et al. [11] đã sử dụng phương pháp mã hóa chéo kênh (cross-channel encoding), trong khi Larsson et al. [8] đã thiết kế một hệ thống có thể dự đoán biểu đồ màu (color histogram) cho từng pixel và được tiền huấn luyện (pre-trained) cho các nhiệm vụ phân loại. Các phương pháp này đã đạt được kết quả ấn tượng khi tô màu những hình ảnh phức tạp.

Tóm lại, có ba phương pháp chính để tô màu ảnh:

**Tô màu dựa trên phác thảo màu** – Người dùng vẽ lên ảnh dựa vào độ sáng để ánh xạ màu sắc tương tự.

Truyền màu từ các nét vẽ ban đầu đến các pixel lân cận có độ sáng tương tự [2].

Cải thiện bằng cách sử dụng thông tin cạnh để ngăn màu lan giữa các đối tượng [17].

Mở rộng bằng cách truyền màu không chỉ theo pixel lân cận mà còn theo các mẫu tương tự [18], [19].

**Tô màu dựa trên ảnh tham chiếu** – Chuyển màu từ ảnh tham chiếu có sẵn sang ảnh mục tiêu.

Truyền màu từ ảnh tham chiếu sang ảnh mục tiêu bằng cách đo độ tương đồng giữa chúng.

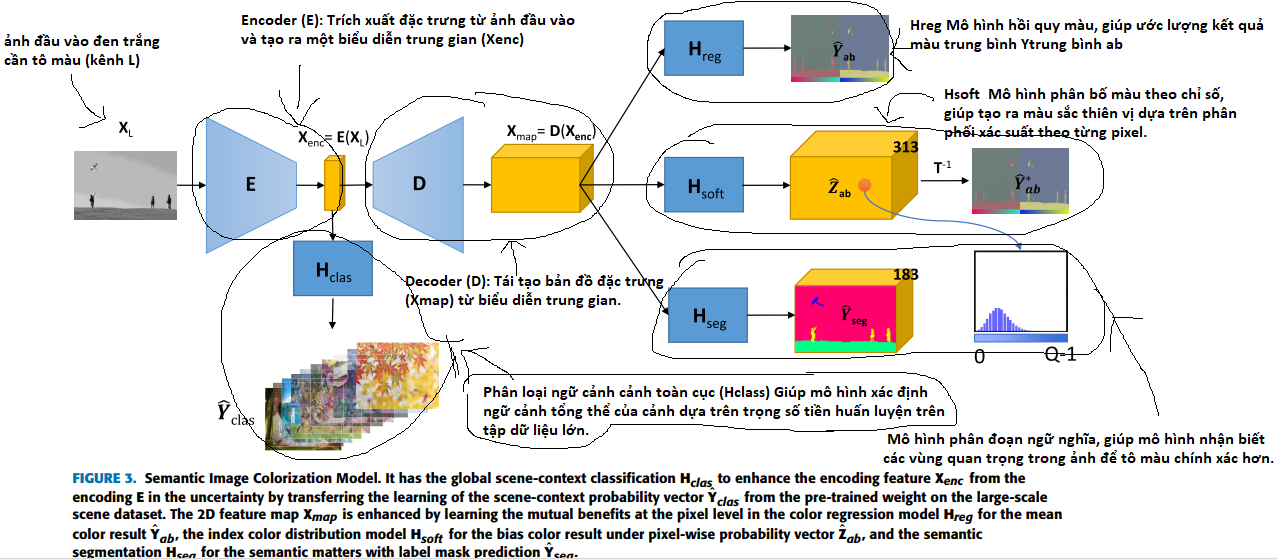
Phương pháp trong [3] sử dụng tối ưu hóa đồ thị (graph-cut) dựa trên xác suất phân phối màu của từng pixel cục bộ.

**Tô màu dựa trên học sâu** – Dùng mạng nơ-ron học cách tô màu tự động mà không cần sự can thiệp của con người.

Mạng hyper-column VGG với hàm mất mát chưa cân bằng [8].

Mạng hai luồng kết hợp đặc trưng toàn cục và cục bộ với phân loại ngữ cảnh cảnh quan [7].

Phương pháp mã hóa kênh chéo để học màu sắc hiệu quả hơn [11].

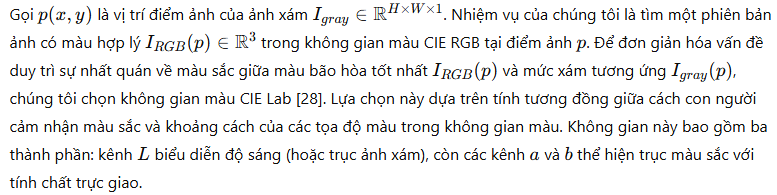


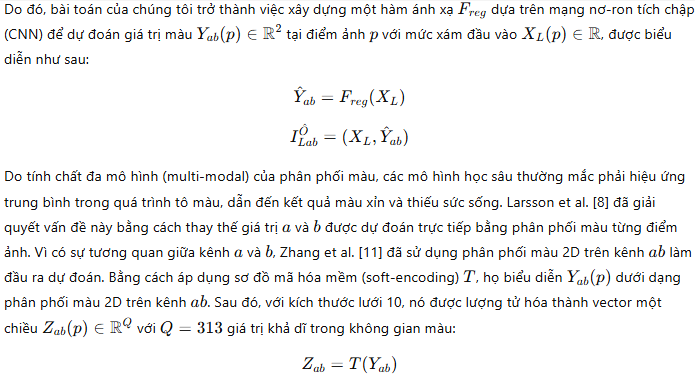
Kết quả cuối cùng là ảnh tô màu được tạo ra thông qua kết hợp các thông tin từ các mô-đun trên.( 3 cái H là 3 cái modun)

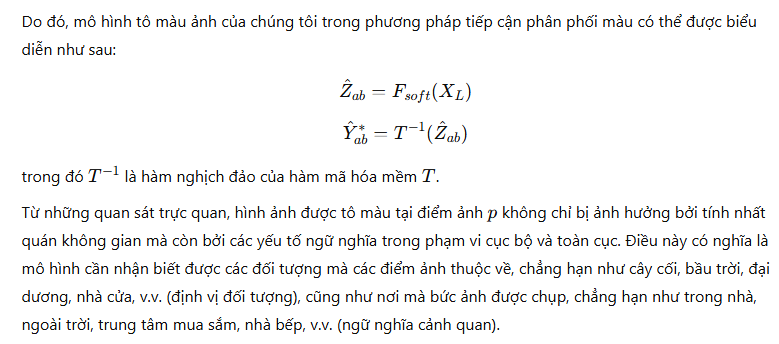
PROPOSED METHOD

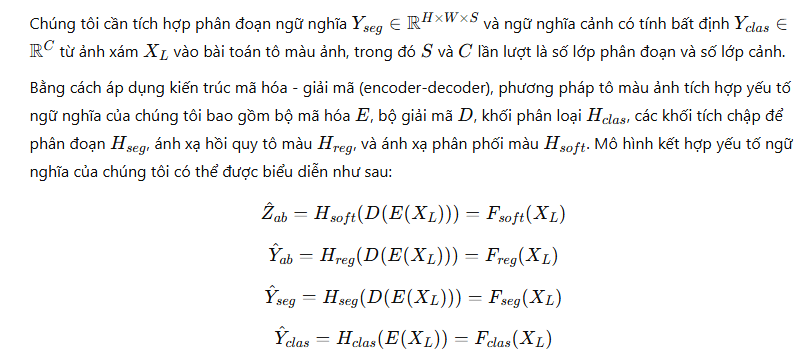
Trong phần này, chúng tôi mô tả mô hình đề xuất của mình, được minh họa trong Hình 3, nhằm giải quyết bài toán tô màu bằng cách sử dụng phân phối màu đa mô hình để lựa chọn màu thiên lệch (bias selection) và hồi quy màu để dự đoán màu trung bình. Đồng thời, mô hình cũng khai thác thông tin ngữ nghĩa bằng cách phân loại ngữ cảnh cảnh quan toàn cục và phân đoạn ngữ nghĩa ở cấp độ điểm ảnh. Sau đó, chúng tôi sẽ mô tả chi tiết phương pháp tô màu ngữ nghĩa, bao gồm: vector phân phối màu chỉ mục với cơ chế mã hóa và giải mã mềm (soft-encoding và decoding), dữ liệu ngữ cảnh cảnh quan có tính bất định, và phân đoạn ngữ nghĩa cấp độ điểm ảnh. Cuối cùng, chúng tôi đề cập đến hàm mất mát đa thành phần trong mô hình đề xuất.

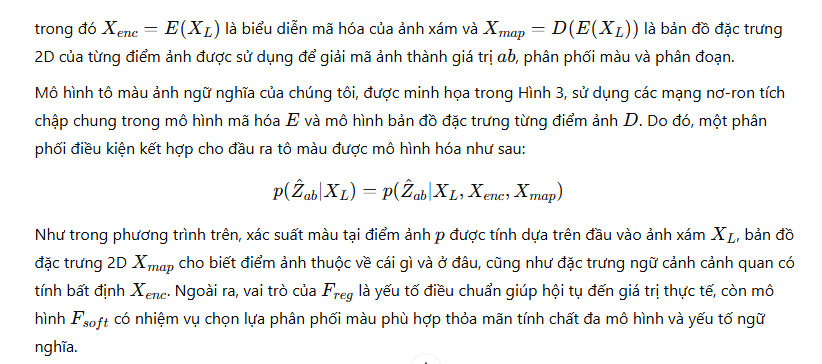
**MÔ HÌNH TÔ MÀU ẢNH NGỮ NGHĨA**

****

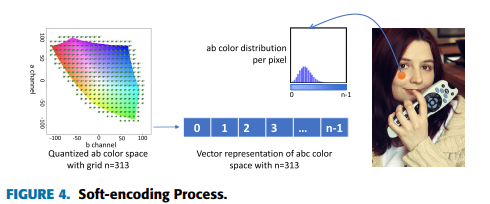
****

****

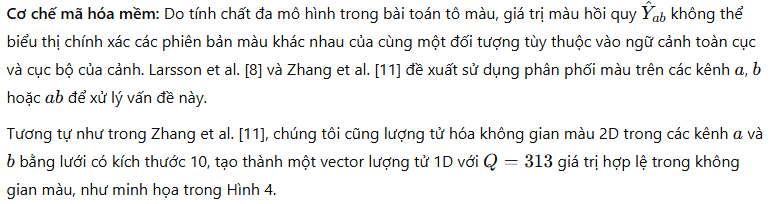
****

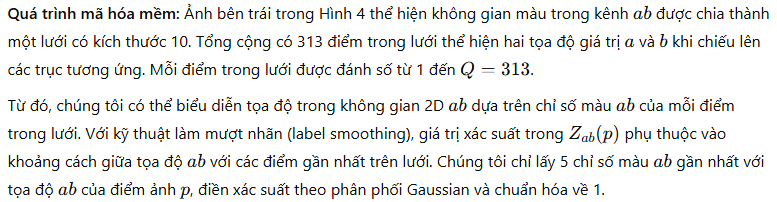


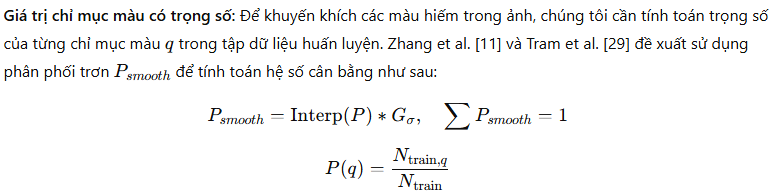
**DỮ LIỆU TÔ MÀU NGỮ NGHĨA**

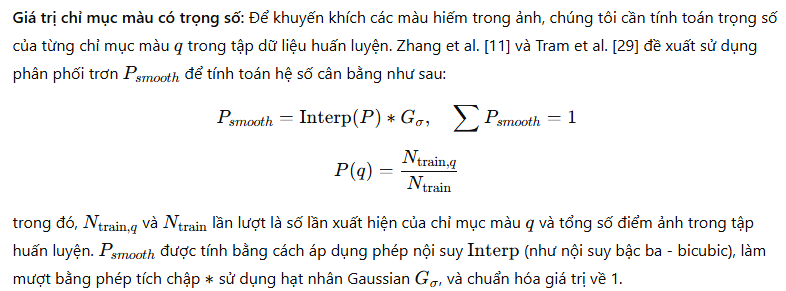


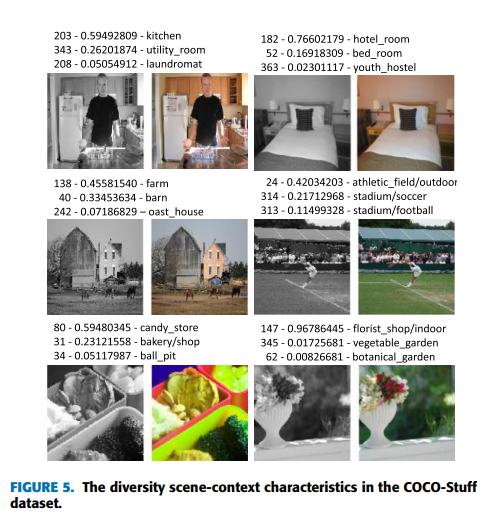
1) Dữ liệu phân phối màu với cơ chế mã hóa mềm

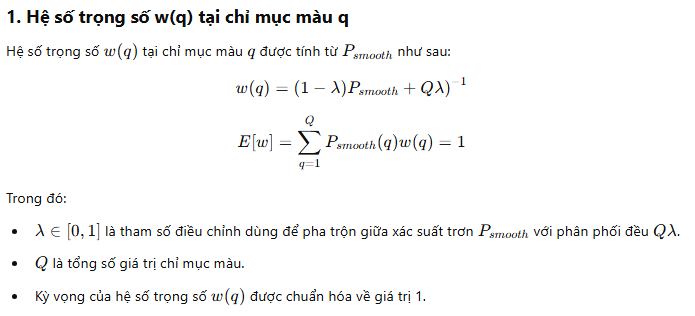


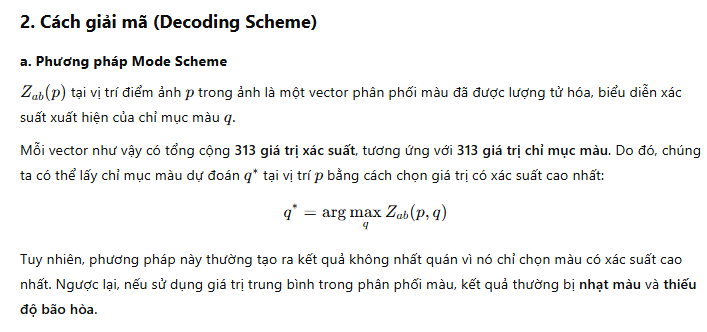


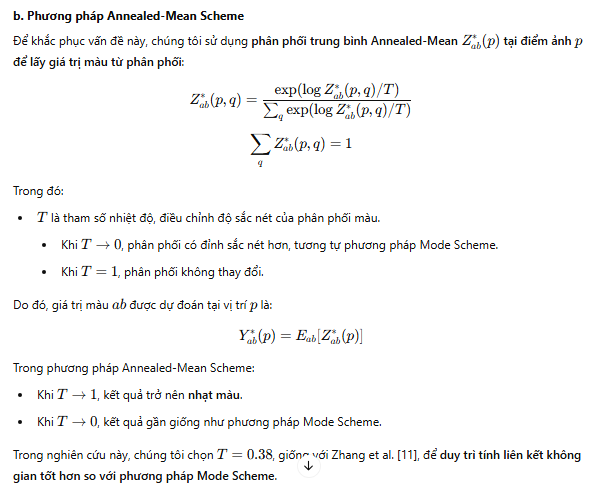


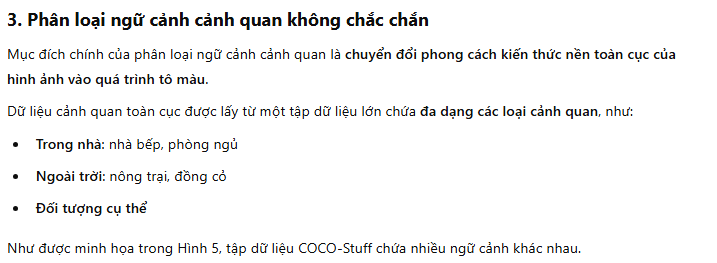


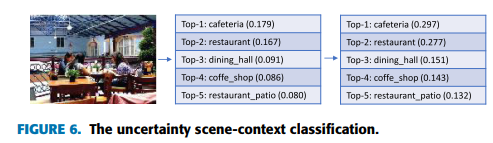


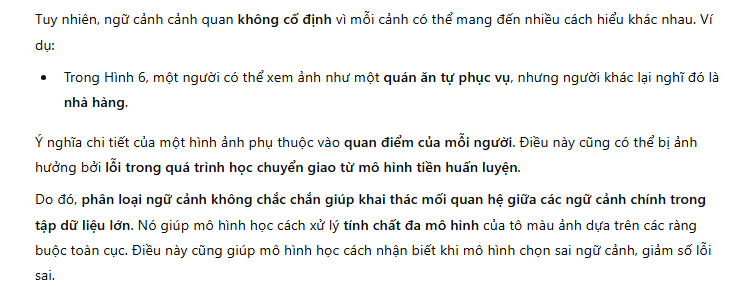


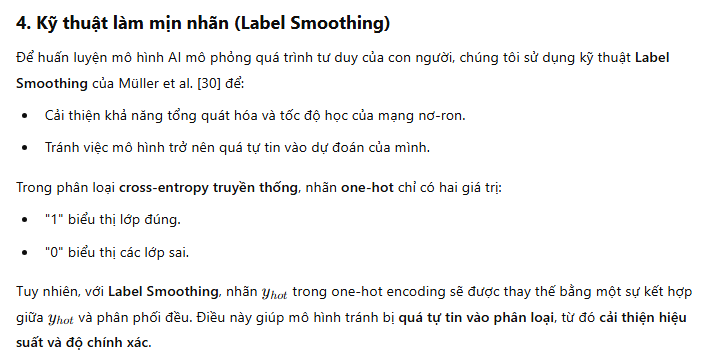






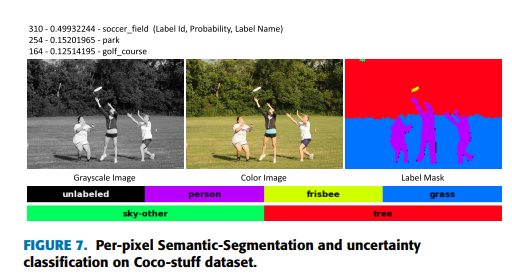






Trong bài báo này, thay vì sử dụng sự kết hợp giữa mã hóa one-hot và phân phối đồng đều như mô tả ở trên, chúng tôi điền vào năm xác suất cao nhất từ mô hình VGG16 đã được huấn luyện trước [31] trên tập Places [15] và chuẩn hóa các giá trị này thành 1, như thể hiện trong Hình 6. Điều này giúp chúng tôi chuyển giao kiến thức từ mạng miền cảnh (scene domain network) sang mô hình tô màu của mình trong điều kiện không chắc chắn, từ đó cải thiện khả năng tổng quát hóa của mô hình.

3) PHÂN ĐOẠN NGỮ NGHĨA Ở CẤP ĐỘ ĐIỂM ẢNH



Ngoài các đặc trưng ngữ nghĩa tổng thể giúp chuyển đổi phong cách trong bối cảnh không chắc chắn, quá trình tô màu còn bị ảnh hưởng bởi yếu tố ngữ nghĩa trong phạm vi cục bộ ở cấp độ điểm ảnh. Nghĩa là mỗi điểm ảnh thuộc về một đối tượng hoặc một loại vật thể nhất định. Trong tầm nhìn của con người, chúng ta sử dụng kinh nghiệm với các đối tượng trong ảnh để tạo ra một hình ảnh được tô màu phù hợp với bối cảnh tổng thể của cảnh.

Trong Hình 7, với việc phân loại theo ngữ cảnh cảnh quan, bối cảnh không chắc chắn được tính toán từ trọng số đã được huấn luyện trước trên Places365 [15] cho một hình ảnh trong tập Coco-Stuff [13]. Hình ảnh này cho thấy ba bối cảnh có thể xảy ra với xác suất cao nhất. Sử dụng kỹ thuật làm mượt nhãn (label smoothing), chúng tôi ngăn chặn lỗi trong quá trình chuyển giao học tập và thiết lập mối quan hệ giữa các loại cảnh như sân bóng đá, công viên và sân golf.

Ngoài ra, trong chi tiết hình ảnh, một mô hình đa phong cách được thể hiện với ba phong cách tô màu khác nhau cho nhãn "người". Đối với các nhãn "vật thể" (stuff), có các nhãn như "cây", "cỏ" và "bầu trời khác" ảnh hưởng đến ngữ cảnh cảnh quan. Cũng có một nhãn đối tượng nhỏ là "đĩa ném (frisbee)". Mặt nạ nhãn nội dung giúp mô hình học được mô tả cục bộ và tăng cường bản đồ đặc trưng giải mã XmapX\_{map}Xmap​ để nâng cao độ chính xác trong các cảnh phức tạp.

Trên thực tế, một hình ảnh có màu có thể được phân đoạn tốt hơn so với một hình ảnh xám [32]. Mặc dù hệ thống đầu vào của chúng tôi chỉ nhận ảnh xám, nhưng mô hình phân đoạn ngữ nghĩa tận dụng mô hình tô màu để cải thiện kết quả phân đoạn. Mặt khác, nó cũng giúp các mô hình khác khai thác các điểm ảnh theo ngữ nghĩa, xác định chúng thuộc về đâu và chúng là đối tượng hoặc vật thể gì.

