

# TÔ MÀU HÌNH ẢNH THÔNG QUA VIỆC TẠO BẢNG MÀU DỰA TRÊN VĂN BẢN

LÊ THANH PHONG

Khoa Công nghệ thông tin

Bộ môn Khoa học dữ liệu

thanhpheong27092001@gmail.com

MSSV: 19475611

LƯU THỊ YÊN NHI

Khoa Công nghệ thông tin

Bộ môn Khoa học dữ liệu

luuthiyennhi2001@gmail.com

MSSV: 19522491

NGUYỄN VĂN PHÚC NHÂN

Khoa Công nghệ thông tin

Bộ môn Khoa học dữ liệu

nguyennhan.workspace@gmail.com

MSSV: 19440221

PHAN NGUYỄN MAI PHƯƠNG

Khoa Công nghệ thông tin

Bộ môn Khoa học dữ liệu

pnmphuong2001@gmail.com

MSSV: 19469121

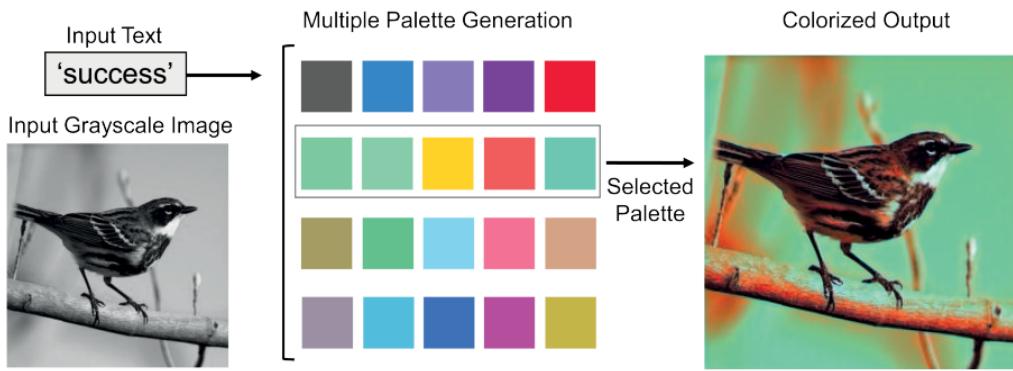
## Tóm tắt nội dung

Ảnh là nơi chứa đựng các thông tin về vật thể hay quang cảnh được chiếu sáng mà con người quan sát, cảm nhận bằng mắt và hệ thần kinh thị giác. Trong kỹ thuật xử lý ảnh (Computer Vision), ảnh được chuyển về các dạng vector số, định nghĩa là hàm hai chiều. Đối với bài báo này, nhóm chúng tôi sẽ tạo ra các bảng màu dựa vào ngữ nghĩa của văn bản đầu vào (text input) sau đó tô màu cho một hình ảnh (dạng gray-image) dựa theo các bảng màu đã tạo. Đồng thời sử dụng tập dữ liệu được sắp xếp theo cách thủ công có tên Palette-and-Text (PAT), bao gồm 10.183 cặp văn bản và bảng màu tương ứng của nó. Với mục đích trên, nhóm chúng tôi đã đề xuất sử dụng mô hình Text2Colors tương ứng với việc tạo văn bản thành bảng màu (The Text-to-Palette Generation Networks (TPN)) rồi tiếp theo sẽ tạo màu dựa trên bảng màu (The Palette-based Colorization Networks (PCN)) và trả về kết quả là hình đã được tô màu từ ảnh xám được đưa vào. Kết quả đánh giá dựa trên việc người quan sát chọn bảng màu phù hợp nhất với đầu vào văn bản.

**Keywords:** Computer Vision , PAT , Text2Colors , TPN , PCN

## 1 Giới thiệu

Tô màu là một bài toán không mới với những nhà mỹ thuật, nhưng lại là một bài toán mới trong lĩnh vực khoa học kĩ thuật. Nếu con người có thể liên kết những chuỗi từ ngữ để tạo ra những bảng màu nhất định thì liệu rằng máy móc có thể làm được điều tương tự như vậy hay không? Có thể thấy rằng phương pháp tô màu được ví như một người thợ sơn, nó giúp những bức ảnh xám trở nên có màu sắc phù hợp và tự nhiên. Và gần đây, sau nhiều thành công nổi bật trong các lĩnh vực xử lý ảnh (Computer Vision), (Deep Learning) cũng mang đến cách tiếp cận mới cho bài toán tô màu ảnh xám. Cụ thể, sau khi được huấn luyện, mô hình của chúng tôi sẽ tạo ra một bảng màu hoàn toàn mới với nhiều màu sắc đa dạng khi được nhập vào một đoạn văn bản có nghĩa. Nổi trội hơn rất nhiều so với phương pháp cũ khi chỉ lấy một từ đơn lẻ đầu vào và ghép với một bảng màu duy nhất trong dữ liệu có sẵn.



Hình 1: Mô hình của chúng tôi có thể tạo ra nhiều sự lựa chọn bảng màu cho đầu vào văn bản, chẳng hạn như 'success'. Người dùng có thể chọn bảng màu nào sẽ được sử dụng cho đầu ra tô màu cuối cùng

Trong nghiên cứu này, chúng tôi đề xuất một phương pháp mới để tạo nhiều bảng màu bằng cách đưa vào mô hình các vector từ có nghĩa và sau đó tô màu một hình ảnh xám thông qua bảng màu đã tạo. Nhận thức về màu sắc vốn đã là một bài toán đa phương thức, nghĩa là khi ta nhập văn bản cụ thể, nó có thể được ánh xạ tới nhiều bảng màu. Để kết hợp tính đa phương thức như vậy vào mô hình, chúng tôi tạo ra các mạng có thể tạo nhiều bảng màu từ một đầu vào văn bản. Tiếp đó, chúng tôi sử dụng bảng màu đã tạo ra từ các mạng trên để sử dụng cho tác vụ tô màu. Việc chúng tôi sử dụng mô hình tô màu này là do người dùng hướng dẫn trước đây sử dụng các gợi ý màu do người dùng đưa ra, từ đó chúng tôi đã nâng cấp thiết kế các mạng tô màu của mình để sử dụng bảng màu trong quá trình tô màu cho ảnh xám.

Tính mới trong nghiên cứu này:

1. Tạo ra được các bảng màu dựa trên kiểu nhập văn bản từ ngôn ngữ tự nhiên thông qua mô hình DeepLearning được đề xuất trong bài báo này.
2. Mô hình này có thể sử dụng bảng màu đã được tạo ra trước đó để tiếp tục tạo ra những màu sắc phù hợp và lại sử dụng những bảng màu đó tô lên những bức ảnh xám.
3. Giới thiệu được tập dữ liệu màu được sắp xếp thủ công mang tên Palette-and-Text (PAT) với 10.183 cặp văn bản và bảng màu tương ứng.

## 2 Nghiên cứu liên quan

**Ngữ nghĩa màu sắc** - Ý nghĩa liên quan đến một màu sắc là vừa bẩm sinh và vừa học được. Ví dụ, màu đỏ có thể khiến chúng ta cảm thấy cảnh giác theo bản năng và màu trắng đôi khi được hiểu theo văn hóa là có liên quan đến sự tinh khiết. Vì màu sắc có mối liên hệ chặt chẽ với các khái niệm ngữ nghĩa cấp cao nên việc tạo bảng màu từ đầu vào văn bản

rất hữu ích trong việc hỗ trợ các nghệ sĩ và nhà thiết kế [1], đồng thời cho phép tô màu tự động từ bảng màu. Một nhược điểm của việc sử dụng văn bản để chọn bộ lọc là tên bộ lọc thường không truyền đạt màu của bộ lọc [2], do đó khiến người dùng khó tìm được bộ lọc phù hợp với sở thích của họ chỉ bằng cách nhìn vào tên bộ lọc. Để khắc phục sự khác biệt này giữa các bảng màu và tên của chúng, liên kết từ màu đã được nghiên cứu từ lâu. Cách tiếp cận dựa trên truy vấn và dựa trên học tập là hai phương pháp điển hình để đề xuất bảng màu dựa trên văn bản đầu vào của người dùng. Các phương thức trước đây sử dụng kiểu nhập văn bản để truy vấn một hình ảnh từ từ điển hình ảnh [2, 3]. Sau đó, màu sắc được trích xuất từ hình ảnh được truy vấn để tạo bảng màu liên quan. Phương pháp này phải đối mặt với một vấn đề lớn, đầu vào văn bản đôi khi được ánh xạ tới nội dung hình ảnh của hình ảnh được truy vấn thay vì màu mà văn bản ngũ ý. Thay vì tìm kiếm mục tiêu một cách trực tiếp, các phương pháp tiếp cận dựa trên học tập khớp các bảng màu với các mô tả ngôn ngữ của chúng bằng cách học liên kết ngữ nghĩa của chúng từ dữ liệu quy mô lớn. Các kỹ thuật học máy [7, 8, 9] đã được sử dụng để đạt được kết quả đầy hứa hẹn. Tuy nhiên, chúng chỉ có thể hỗ trợ đầu vào ở cấp độ từ và cũng không thể tạo bảng màu mới.

**Conditional GANs** - Các mạng đối thủ tạo ra có điều kiện (cGAN) là các mô hình GAN sử dụng thông tin có điều kiện cho bộ phân biệt và bộ tạo [10]. cGAN đã rút ra kết quả đầy hứa hẹn để tạo hình ảnh từ văn bản [11, 12, 13] và dịch từ hình ảnh sang hình ảnh [14, 15, 16]. StackGAN [13] là mô hình đầu tiên sử dụng mэт điều kiện để tổng hợp văn bản thành hình ảnh. Nó đạt được kết quả tốt hơn nhiều so với các mẫu trước đó bằng cách chia nhỏ quá trình tổng hợp hình ảnh thành một tác vụ phác thảo và tinh chỉnh. StackGAN cũng giới thiệu kỹ thuật tăng cường điều hòa cho phép mô hình đưa ra các kết quả đa dạng ngay cả khi được cung cấp cùng một văn bản đầu vào. Tăng cường điều hòa sẽ được xây dựng thêm trong Phần 4.1.

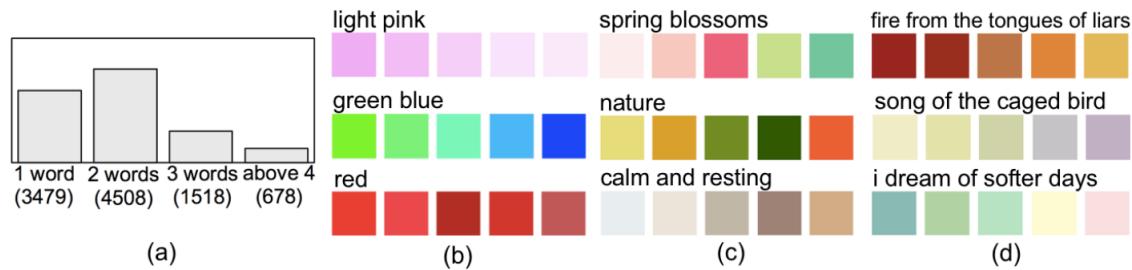
**Tương tác màu sắc** - Tô màu là một nhiệm vụ đa phương thức và kết quả tô màu mong muốn cho cùng một đối tượng có thể khác nhau tùy thuộc theo từng người [4]. Một số nghiên cứu giới thiệu các phương pháp tương tác cho phép người dùng kiểm soát đầu ra màu sắc cuối cùng [5, 17]. Trong các mô hình này, người dùng tương tác trực tiếp với mô hình bằng cách xác định vị trí cần tô màu. Mặc dù các phương pháp này đạt được kết quả khả quan, nhưng một hạn chế là người dùng cần phải có một trình độ nghệ thuật nhất định. Do đó, thay vì khiến người dùng tô màu trực tiếp cho hình ảnh, các nghiên cứu khác khiến người dùng thực hiện một cách tiếp cận gián tiếp hơn bằng cách sử dụng bảng màu để tô màu lại hình ảnh [18, 6]. Các bộ lọc dựa trên bảng màu là một cách hiệu quả để những người không phải là chuyên gia về nghệ thuật có thể thay đổi màu sắc của ảnh [18].

**Sequence-to-Sequence with Attention** - Recurrent Neural Networks (RNNs) là một công cụ phổ biến do khả năng học hỏi vượt trội từ dữ liệu tuần tự. RNN được sử dụng trong các nhiệm vụ khác nhau bao gồm phân loại câu [19], tạo văn bản [20] và dự đoán trình tự theo trình tự [21]. Các hệ thống dịch máy thần kinh lập mô hình xác suất mà một câu

nguồn  $\{x_1, \dots, x_n\}$  được dịch sang một câu đích  $\{y_1, \dots, y_n\}$  dưới dạng  $p(y|x)$ . Việc kết hợp Attention vào một mô hình Sequence-to-Sequence được biết là giúp cải thiện hiệu suất của mô hình [22]. Các mô-đun Attention cho phép các mạng tập trung có chọn lọc vào các phần của câu nguồn. Điều này cho phép một mô hình tìm hiểu mối quan hệ giữa các phương thức khác nhau (ví dụ: văn bản - màu sắc, văn bản - hành động, Anh - Pháp).

### 3 Dữ liệu Palette-and-Text (PAT)

Phần này sẽ giới thiệu tập dữ liệu có tên là Palette-and-Text (PAT). PAT chứa 10.183 văn bản và các cặp bảng màu có năm màu, trong đó tập hợp năm màu trong bảng màu được liên kết với mô tả văn bản tương ứng của nó như trong **Hình 2 (b)-(d)**. Mô tả văn bản được tạo thành từ 4.312 từ duy nhất. Các từ thay đổi tùy theo mối quan hệ của chúng với màu sắc; một số từ là từ chỉ màu sắc trực tiếp (ví dụ: pink, blue, v.v.) trong khi những từ khác gợi lên một tập hợp màu cụ thể (ví dụ: autumn or vibrant). Theo hiểu biết tốt nhất của chúng tôi, không có bộ dữ liệu nào phù hợp với văn bản nhiều từ và bảng màu 5 màu tương ứng của nó. Bộ dữ liệu này cho phép chúng tôi đào tạo các mô hình của mình để dự đoán các bảng màu nhất quán về mặt ngữ nghĩa với các đầu vào dạng văn bản.



Hình 2: Giới thiệu về bộ dữ liệu Palette-and-Text (PAT): **(a)** hiển thị số lượng mục dữ liệu liên quan đến độ dài văn bản của chúng. Bên phải là các ví dụ hiển thị các cặp bảng màu văn bản đa dạng trong PAT. Những mô tả văn bản phù hợp với bảng màu của chúng bao gồm **(b)** tên màu trực tiếp, **(c)** văn bản có mức độ tương đồng thấp về quan hệ ngữ nghĩa với màu sắc, **(d)** những văn bản có ngữ cảnh ngữ nghĩa cấp cao.

**Bộ dữ liệu màu khác** - Khảo sát màu sắc của Munroe [23] là một kho dữ liệu màu quy mô lớn được sử dụng rộng rãi. Dựa trên đánh giá của người dùng có nguồn gốc từ cộng đồng, nó khớp một văn bản với một màu duy nhất. Trong khi cuộc khảo sát về màu sắc của Munroe đề cập đến một màu, thì thang đo hình ảnh màu của Kobayashi [1] là một bộ dữ liệu đa màu được thiết lập tốt hơn. Kobayashi sử dụng 180 tính từ để thể hiện 1170 bảng kết hợp ba màu. Mặc dù cả hai bộ dữ liệu đều được sử dụng tích cực trong nhiều ứng dụng, nhưng nhiệm vụ của chúng tôi yêu cầu một bộ dữ liệu khác khớp văn bản với nhiều màu và đủ lớn để mô hình học sâu học hỏi.

**Thu thập dữ liệu** - Bộ dữ liệu mang tên PAT mà nhóm chúng tôi đã thu thập được từ

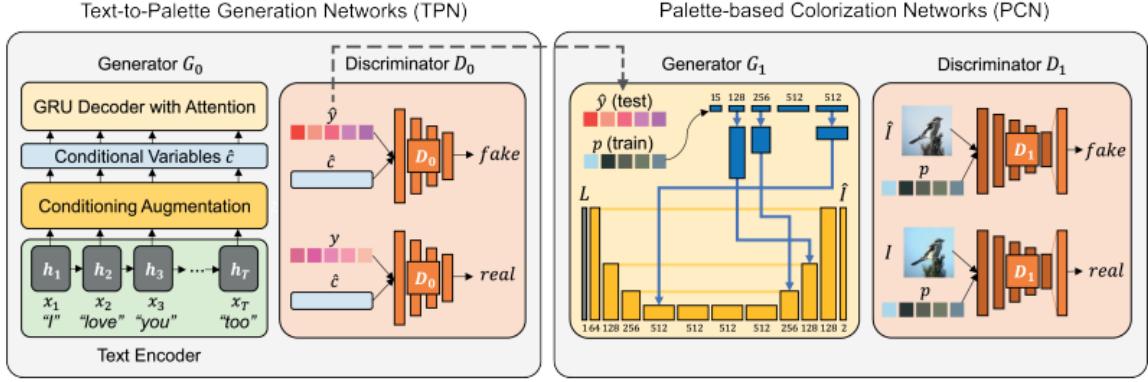
Internet là bộ dữ liệu mở. Với ngôn ngữ đầu vào là tiếng Anh cũng như các dữ liệu bảng màu này đều có thể tinh chỉnh hay đặt tên theo ý thích. Đây là bộ dữ liệu được thu thập từ một trang web cộng đồng có tên color-hex.com - nơi người dùng có thể tải lên các bảng màu tùy chỉnh để người khác sử dụng. Chúng tôi đã thu thập dữ liệu 47.665 cặp văn bản bảng màu và xóa các từ không phải chữ và số cũng như không phải là tiếng Anh. Trong số đó, chúng tôi nhận thấy rằng: đôi khi người dùng gán tên bảng màu một cách tùy tiện, thiếu tính nhất quán về ngữ nghĩa với bảng màu tương ứng của họ. Có một số từ ngữ không hề có nghĩa gì như: "mehmeh" hay "i spilled tea all over my laptop rip" hoặc có lỗi chính tả (ví dụ: 'cause iiiiii see through you boyyyyy' và 'greene gardn'). Do đó, việc sử dụng tên bảng màu thô chưa tinh chỉnh xử lí sẽ cản trở đáng kể hiệu suất của mô hình. Để tinh chỉnh tập dữ liệu thô nhiều như thế này, bốn người chúng tôi đã thảo luận xem những cụm từ ấy nếu được ghép nối với bảng màu được tạo thì có khớp đúng với ngữ nghĩa của nó hay không. Sau đó, chúng tôi chỉ sử dụng các cụm từ mà theo đó đã có ít nhất ba trên tổng số bốn người trong nhóm chúng tôi đồng ý rằng có sự khớp về ngữ nghĩa. Qua đó, lỗi chính tả cũng như nghĩa của từ đã được sửa theo cách thủ công sau khi chúng tôi hoàn tất việc phân loại dữ liệu.

Bộ dữ liệu PAT có một số đặc điểm. Đầu tiên, nó được tinh chỉnh theo nhận thức của bốn người chúng tôi, vốn mang tính chủ quan. Một cặp bảng màu văn bản hoàn toàn hợp lý với một người nhưng có thể không vừa ý với người khác. Chúng tôi muốn kết hợp tính chủ quan đó bằng cách cho phép lựa chọn đa dạng các cặp bảng màu văn bản. Chỉ bao gồm các cặp bảng màu văn bản trong tập dữ liệu khi cả bốn người chúng tôi đồng ý được coi là phù hợp với ngữ nghĩa, không có nhiều quan điểm cá nhân. Do đó, cặp bảng màu văn bản đã được đưa vào PAT khi có ít nhất ba người trong số bốn người chúng tôi quyết định văn bản phản ánh bảng màu. Thứ hai, tập dữ liệu của chúng tôi bao gồm các văn bản ngắn. Như được hiển thị trong **Hình 2**, khoảng 15% mô tả văn bản trong bộ dữ liệu chứa nhiều hơn bốn từ. Trong quá trình sắp xếp, chúng tôi thấy khó đạt được sự đồng thuận hơn đối với các văn bản dài hơn. Điều này có thể là do khi văn bản trở nên dài hơn, phạm vi các màu phù hợp có thể có của nó trở nên rộng hơn và không rõ ràng. Do đó, các văn bản dài hơn có nhiều khả năng bị xóa khỏi tập dữ liệu cuối cùng hơn.

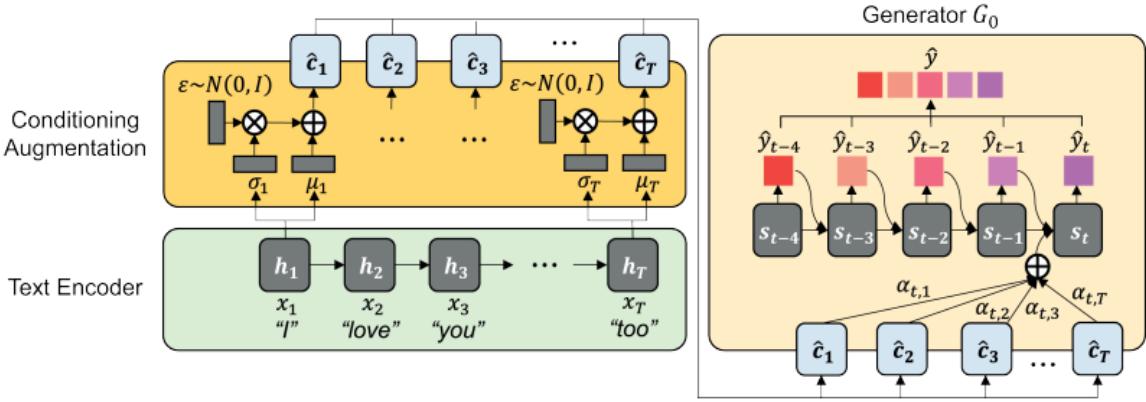
## 4 Phương pháp tiếp cận

Mô hình đề xuất của chúng tôi bao gồm hai mạng: Mạng tạo văn bản thành bảng màu (Text-to-Palette Generation Networks (TPN)) và Mạng tạo màu dựa trên bảng màu (Palette-based Colorization Networks (PCN)). Chúng tôi đào tạo các mạng đầu tiên để tạo các bảng màu cho một văn bản và sau đó đào tạo mạng thứ hai để dự đoán cách tô màu hợp lý cho một hình ảnh thang độ xám với các bảng màu được tạo. Chúng tôi sử dụng GAN có điều kiện (cGAN) cho cả hai mạng.

## 4.1 Text-to-Palette Generation Networks (TPN)



Hình 3: Tổng quan về kiến trúc Text2Colors của chúng tôi. Trong quá trình huấn luyện, trình khởi tạo  $G_0$  học cách tạo bảng màu  $\hat{y}$  cho trước từ một tập hợp các biến có điều kiện  $\hat{c}$  được xử lý từ văn bản đầu vào  $x = \{x_1, \dots, x_T\}$ . Trình khởi tạo  $G_1$  học cách dự đoán đầu ra được tô màu của hình ảnh thang độ xám  $L$  được cung cấp bảng màu  $p$  được trích xuất từ hình ảnh thực. Tại thời điểm thử nghiệm, các trình khởi tạo được huấn luyện  $G_0$  và  $G_1$  được sử dụng để tạo bảng màu từ văn bản đã cho và sau đó tô màu một hình ảnh thang độ xám bảng màu đã được tạo



Hình 4: Kiến trúc mô hình của trình tạo  $G_0$  tạo ra màu thứ  $t$  trong bảng với văn bản đầu vào  $x = \{x_1, \dots, x_T\}$ . Lưu ý rằng tính ngẫu nhiên được thêm vào mỗi vecto trạng thái ẩn  $h$  trong chuỗi trước khi nó được chuyển đến trình tạo

Trong phần này, chúng tôi minh họa mạng tạo Văn bản thành bảng màu được hiển thị trong **Hình 3 và 4**, một trong những mạng đầu tiên tạo ra các bảng màu phù hợp được liên kết với kiểu nhập văn bản. Đặt  $x_i \in R^{300}$  là các vectơ từ được khởi tạo bởi các vectơ được đào tạo trước 300 chiều từ GloVe [24]. Những từ không có trong tập huấn luyện trước được khởi tạo ngẫu nhiên. Sau khi bộ mã hóa văn bản GRU mã hóa  $x$  thành các trạng thái ẩn  $h = \{h_1, \dots, h_T\}$  thêm biến thể vào văn bản được mã hóa bằng cách lấy mẫu các biến ẩn  $\hat{c}$  từ phân phối Gaussian  $N(\mu(h), \Sigma(h))$ . Sử dụng chuỗi các biến điều hòa  $\hat{c} = \{\hat{c}_1, \dots, \hat{c}_T\}$  làm điều kiện, hàm mục tiêu của cGAN đầu tiên có thể được biểu diễn dưới dạng

$$L_{D_0} = E_{y \sim P_{data}} [\log(D_0(\hat{c}, y))] + E_{\hat{y} \sim P_{G_0}} [\log(1 - D_0(\hat{c}, \hat{y}))] \quad (1)$$

$$L_{G_0} = E_{\hat{y} \sim P_{G_0}} [\log(1 - D_0(\hat{c}, \hat{y}))] + \lambda_H L_H(\hat{y}, y) + \lambda_{KL} D_{KL}(N(\mu(h), \Sigma(h)) || N(0, I)) \quad (2)$$

Trong đó, Discriminator  $D_0$  cố gắng tối đa hóa  $L_{D_0}$  ngược lại Generator  $G_0$  cố gắng giảm thiểu  $L_{G_0}$ .  $y$  được lấy từ phân phối bản màu thực  $P_{data}$  trong khi  $\hat{y}$  được lấy từ phân phối màu từ model  $P_{G_0}$ . Từ hàm loss của cGAN có thể thấy là việc train Generator và Discriminator đối nghịch nhau, trong khi D cố gắng maximize loss thì G cố gắng minimize loss.

Các phương pháp trước đây đã được hưởng lợi từ việc trộn mục tiêu GAN với khoảng cách  $L_2$  [25] hoặc khoảng cách  $L_1$  [15]. Chúng tôi nhận thấy rằng so với độ đo  $L_1$  và  $L_2$  thì thấy rằng mất mát Huber (hoặc smooth  $L_1$ ) cho hiệu quả cao nhất trong việc tăng sự đa dạng giữa các màu trong bảng màu đã được tạo. Hàm tổn thất Huber được biểu diễn như sau:

$$L_H(\hat{y}, y) = \begin{cases} \frac{1}{2}(\hat{y} - y)^2 & \text{for } |\hat{y} - y| \leq \delta \\ \delta |\hat{y} - y| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases} \quad (3)$$

Trong đó:

- VỚI  $\frac{1}{2}(\hat{y} - y)^2$  là hàm tính Mean Square Error (MSE)
- VÀ  $\delta |\hat{y} - y| - \frac{1}{2}\delta^2$  tương tự như hàm tính Mean Absolute Error (MAE)

Công thức mất mát này được thêm vào hàm mục tiêu của trình khởi tạo để gần với bảng màu thật trong khi cố gắng đánh lừa trình phân biệt (discriminator).  $\lambda_H$  và  $\lambda_{KL}$  là các hyperparameters để cân bằng ba số hạng trong biểu thức 2. Chúng tôi đặt  $\delta = 1$ ,  $\lambda_H = 100$ ,  $\lambda_{KL} = 0.5$  trong mô hình của mình.

### Kiến trúc của mạng

**Tăng cường điều hòa** Học cách ánh xạ từ văn bản sang màu sắc vốn đã đa phương thức. Chẳng hạn, một văn bản 'autumn' có thể được ánh xạ tới nhiều bảng màu hợp lý. Khi văn bản trở nên dài hơn, chẳng hạn như 'midsummer to autumn' hoặc 'autumn breeze and falling leaves' phạm vi của các bảng màu phù hợp có thể trở nên rộng hơn và đa dạng hơn. Để mô hình hóa một cách thích hợp tính đa phương thức của vấn đề này, chúng tôi sử dụng kỹ thuật tăng điều hòa (CA) [13]. Thay vì sử dụng chuỗi cố định của biểu diễn được mã hóa  $h = \{h_1, \dots, h_T\}$  làm đầu vào cho trình khởi tạo của chúng tôi, việc lấy mẫu ngẫu nhiên các biến tiềm ẩn  $\hat{c}$  từ phân phối Gaussian  $N(\mu(h), \Sigma(h))$  như minh họa trong **Hình 4**. Tính ngẫu nhiên này cho phép mô hình của chúng tôi tạo ra nhiều bảng màu hợp lý với cùng một kiểu nhập văn bản. Chúng tôi đã sử dụng thuật ngữ chính quy hóa phân kỳ Kullback-Leibler (KL) [13]. Đối với bài toán này thì hàm Huber Loss dùng để ước lượng các giá trị của các tập dữ liệu có sẵn và tập dữ liệu tự tạo để buộc tập dữ liệu tự tạo có thể sát với tập dữ liệu có sẵn, kết hợp với việc dùng Kullback-Leibler (KL) divergence để việc

tính toán được dễ dàng và trơn tru hơn, cụ thể:

$$D_{KL} \left( N(\mu(h), \sum(h)) \parallel N(0, I) \right)$$

Sau đó được thêm vào chức năng mục tiêu của trình khởi tạo trong quá trình đào tạo.

**Trình khởi tạo** Để có được tập hợp các vectơ điều hòa  $\hat{c} = \{\hat{c}_1, \dots, \hat{c}_T\}$ , các vectơ từ được đào tạo trước  $x = \{x_1, \dots, x_T\}$  trước tiên được đưa vào bộ mã hóa văn bản GRU để tính toán các trạng thái ẩn  $h = \{h_1, \dots, h_T\}$ . Biểu diễn văn bản này được đưa vào một lớp được kết nối đầy đủ để tạo ra  $\mu$  và  $\sigma$  (các giá trị trong đường chéo của  $\sum$ ) cho phân bố Gaussian  $N(\mu(h), \Sigma(h))$ .  $\hat{c}$  được tính bằng  $\hat{c} = \mu + \sigma \odot \epsilon$ , trong đó  $\odot$  là phép nhân phần tử và  $\epsilon \sim N(0, I)$ . Tập hợp các vectơ  $\hat{c} = \{\hat{c}_1, \dots, \hat{c}_T\}$  sẽ được sử dụng làm điều kiện cho cGAN của chúng tôi.

Chúng tôi thiết kế trình tạo bảng màu của mình như một biến thể của bộ giải mã RNN với cơ chế attention [22, 26]. Bộ giải mã được đào tạo để dự đoán màu tiếp theo  $\hat{y}_t$ , với các vectơ điều kiện  $\hat{c} = \{\hat{c}_1, \dots, \hat{c}_T\}$  và tất cả các màu được dự đoán trước đó  $\{\hat{y}_1, \dots, \hat{y}_{t-1}\}$  của bảng màu  $\hat{y}$ . Nói cách khác, bộ giải mã xác định xác suất trên bảng màu  $\hat{y}$  bằng cách phân tách xác suất chung thành các điều kiện có thứ tự:

$$p(\hat{y}) = \prod_{t=1}^T p(\hat{y}_t | \{\hat{y}_1, \dots, \hat{y}_{t-1}\}, c) \quad (4)$$

Trong đó  $\hat{y} = (\hat{y}_1, \dots, \hat{y}_5)$ . Chúng tôi xác định từng xác suất có điều kiện trong biểu thức (4) như

$$p(\hat{y}_i | \hat{y}_1, \dots, \hat{y}_{i-1}, \hat{c}) = g(\hat{y}_{i-1}, s_i, c_i) \quad (5)$$

Trong đó  $s_i$  là một vectơ trạng thái ẩn GRU cho thời gian  $i$ , được tính bởi

$$s_i = f(s_{i-1}, \hat{y}_{i-1}, c_i)$$

Vectơ ngữ cảnh  $c_i$  phụ thuộc vào chuỗi vectơ điều kiện  $\hat{c} = \{\hat{c}_1, \dots, \hat{c}_T\}$  mà bộ mã hóa văn bản ánh xạ đầu vào văn bản. Mỗi điều kiện  $\hat{c}_i$  chứa thông tin về toàn bộ kiểu nhập văn bản, tập trung chủ yếu vào từ thứ  $i$  với một chút thay đổi. Vectơ ngữ cảnh  $c_i$  được tính bằng tổng trọng số của các điều kiện  $\hat{c}_i$  này, tức là

$$c_i = \sum_{j=1}^T \alpha_{ij} \hat{c}_j \quad (6)$$

Trọng số  $\alpha_{ij}$  của mỗi biến điều kiện  $\hat{c}_j$  được tính bằng

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^T \exp(e_{ik})} \text{ where } e_{ij} = a(s_{i-1}, \hat{c}_j) \quad (7)$$

$$a(s_{i-1}, \hat{c}_j) = w^T \sigma(W^{(1)} s_{i-1} + W^{(2)} \hat{c}_j) \quad (8)$$

Trong đó  $\sigma(\cdot)$  là hàm kích hoạt sigmoid và  $w$  là vectơ trọng số. Attention bổ sung [26]

$a(s_{i-1}, \hat{c}_j)$  tính toán mức độ phù hợp của từ thứ  $j$  của đầu vào văn bản với màu thứ  $i$  của đầu ra bảng màu. Điểm  $\alpha_{ij}$  được tính toán dựa trên trạng thái ẩn GRU  $s_{i-1}$  và điều kiện thứ  $j$   $\hat{c}_j$ . Cơ chế Attention này cho phép mô hình ánh xạ hiệu quả đầu vào văn bản phức tạp sang đầu ra bảng màu.

**Trình phân biệt** Đối với trình phân biệt  $D_0$ , biến điều hòa  $\hat{c}$  và bảng màu được nối và đưa vào một loạt các lớp được kết nối đầy đủ. Bằng cách cùng học các tính năng trên văn bản và bảng màu được mã hóa, trình phân biệt sẽ phân loại xem các bảng màu là thật hay giả.

## 4.2 Palette-based Colorization Networks (PCN)

Mục tiêu của mạng thứ hai là tự động tô màu của hình ảnh với thang độ xám được bằng bảng màu được tạo ra ở mạng thứ nhất dưới dạng biến điều hòa. Đầu vào là một hình ảnh thang độ xám  $L \in R^{H \times W \times 1}$  thể hiện độ sáng trong không gian CIE lab và một bảng màu  $p \in R^{15}$  bao gồm năm màu. Đầu ra  $\hat{I} \in R^{H \times W \times 2}$  tương ứng với các kênh màu ab được dự đoán của hình ảnh. Hàm mục tiêu của mô hình thứ hai có thể được biểu thị bằng

$$L_{D_1} = E_{I \sim P_{data}} [\log(D_1(p, I))] + E_{\hat{I} \sim P_{G_1}} [\log(1 - D_1(p, \hat{I}))] \quad (9)$$

$$L_{G_1} = E_{\hat{I} \sim P_{G_1}} [\log(1 - D_1(p, \hat{I}))] + L_H(\hat{I}, I) \quad (10)$$

$D_1$  và  $G_1$  bao gồm trong phương trình được hiển thị trong Hình 3. Chúng tôi cũng đã thêm tổn thất Huber vào hàm mục tiêu của trình khởi tạo. Nói cách khác, trình khởi tạo học cách làm cho hình ảnh gần với hình ảnh thật cơ bản bằng cách tô màu một cách hợp lý, đồng thời kết hợp các màu sắc từ bảng màu vào hình ảnh đầu ra để đánh lừa trình phân biệt.

### Kiến trúc của mạng

**Trình khởi tạo** Trình khởi tạo bao gồm hai mạng con: mạng tạo màu chính và mạng điều hòa. Các mạng tạo màu chính của chúng tôi áp dụng kiến trúc U-Net [27], đã cho thấy kết quả đầy hứa hẹn trong các nhiệm vụ tạo bảng màu [15, 5]. Các kết nối bỏ qua giúp mô hình khôi phục thông tin [27], vì hình ảnh đầu vào và đầu ra chia sẻ vị trí của các cạnh nổi bật [15].

Vai trò của mạng điều hòa này là áp dụng các màu của bảng màu cho hình ảnh được tạo. Trong quá trình huấn luyện, các mạng được cung cấp một bảng màu  $p \in R^{15}$  được trích xuất từ hình ảnh thật I. Tương tự như nhiệm vụ trước đó [5], bảng điều hòa p được đưa vào một loạt các lớp tích chập Relu  $1 \times 1$  như trong Hình 3. Các đặc trưng trong các lớp 1, 2 và 4 được nhân đôi theo không gian để phù hợp với kích thước không gian của các đặc trưng conv9, conv8 và conv4 trong mạng tô màu chính và được hợp nhất bằng cách bổ sung theo từng phần tử. Biến điều hòa được đưa vào các lớp lấy mẫu với kết nối bỏ qua giống như các mạng ở chính giữa. Điều này cho phép trình khởi tạo phát hiện các điểm nổi bật và áp dụng các màu từ bảng màu cho các vị trí phù hợp của hình ảnh. Trong thời gian thử nghiệm, chúng tôi sử dụng bảng màu được tạo từ mạng đầu tiên (TPN) làm biến

điều hòa, tô màu hình ảnh ở mức thang độ xám bằng các màu từ bảng màu dự đoán.

**Trình phân biệt** Chúng tôi sử dụng một biến thể của kiến trúc mạng DCGAN [28]. Hình ảnh và biến điều hòa  $p$  được nối và đưa vào một loạt các lớp tích chập Leaky Relu để cùng tìm các đặc trưng trên hình ảnh và cũng như bảng màu. Sau đó, nó được đưa vào một lớp được kết nối đầy đủ để phân loại xem hình ảnh là thật hay giả.

### 4.3 Chi tiết triển khai

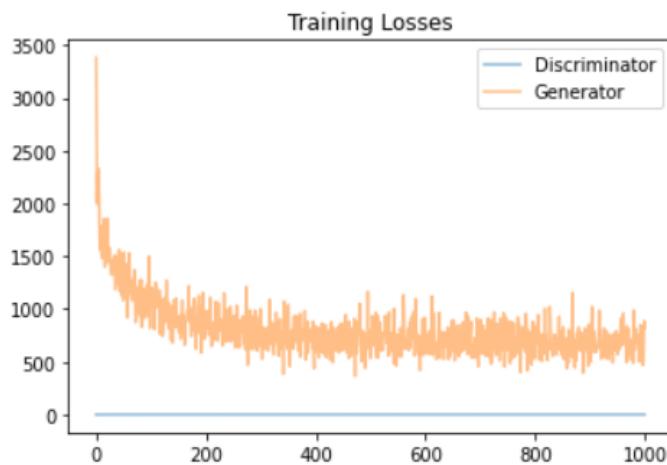
Trước tiên, chúng tôi huấn luyện  $D_0$  và  $G_0$  của TPN trong 1000 epochs bằng cách sử dụng bộ dữ liệu PAT. Sau đó, chúng tôi huấn luyện  $D_1$  và  $G_1$  của PCN trong 100 epochs, sử dụng bảng màu được trích xuất từ hình ảnh thật. Cuối cùng, chúng tôi sử dụng các trình khởi tạo được huấn luyện từ  $G_0$  và  $G_1$  trong thời gian thử nghiệm để tô màu một hình ảnh thang độ xám với bảng màu được tạo từ đầu vào văn bản x. Tất cả các mạng được đào tạo bằng trình tối ưu hóa Adam [29] với tốc độ học tập là 0,0002. Các trọng số được khởi tạo từ phân phối Gaussian với giá trị trung bình bằng 0 và độ lệch chuẩn là 0,05. Chúng tôi đặt các siêu tham số khác là  $\delta = 1$ ,  $\lambda_H = 100$ ,  $\lambda_{KL} = 0.5$ .

## 5 Kết quả thực nghiệm

Phần này sẽ trình bày các kết quả thu được của mô hình đề xuất của chúng tôi. Chúng tôi đánh giá TPN (Phần 4.1) dựa trên bộ dữ liệu PAT của chúng tôi đã được nêu ở trên. Để đánh giá PCN (Phần 4.2), chúng tôi sử dụng ba bộ dữ liệu khác nhau, CUB-200-2011 (CUB), ImageNet ILSVRC Object Detection (ImageNet dataset) và Graphical Pattern Images.

### 5.1 Text-to-Palette Generation Networks (TPN)

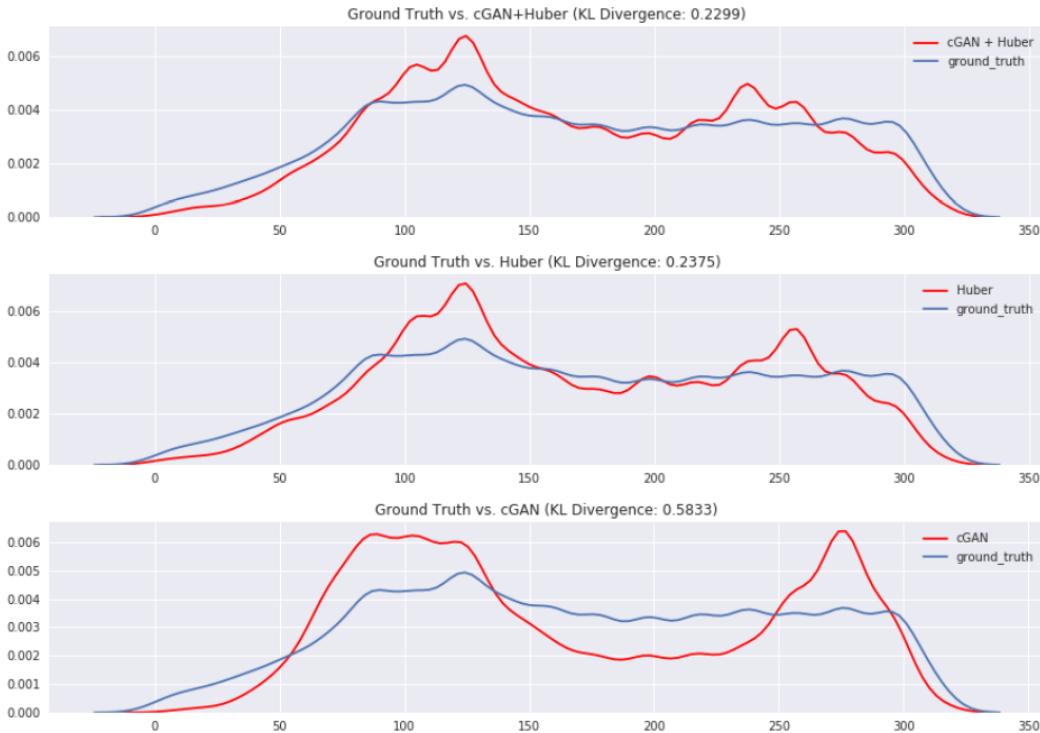
Đối với mô hình tạo bảng màu sau khi chúng tôi huấn luyện mô hình đưa ra được kết quả khá là khả quan. Thông qua Hình 5 thể hiện được độ lỗi trong quá trình huấn luyện mô hình ngày càng giảm dần. Điều đặc biệt ở đây là độ lỗi của Generator (trình khởi tạo) rất cao do chúng tôi định nghĩa ở công thức 2 và các siêu tham số được đặt là cũng khá là cao  $\delta = 1$ ,  $\lambda_H = 100$ ,  $\lambda_{KL} = 0.5$  dẫn đến việc độ lỗi Generator cao hơn rất nhiều so với độ lỗi Discriminator.



Hình 5: Phân tích kết quả sau khi huấn luyện mô hình

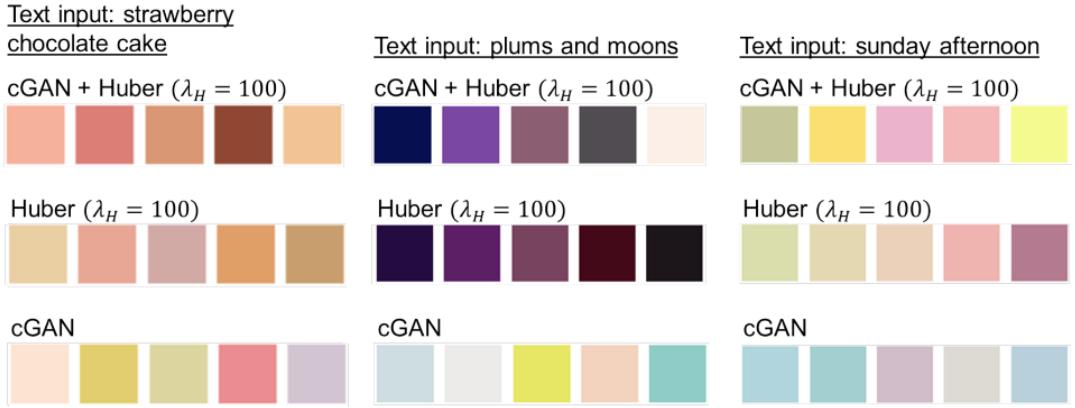
### 5.1.1 So sánh các mô hình từ các biến thể khác nhau

Hình 6 cho thấy sự so sánh về sự phân bố màu giữa các bảng màu thực của dữ liệu đào tạo và các bảng màu được tạo từ dữ liệu thử nghiệm. Đôi với mỗi lần phân bố màu, chúng tôi định lượng các giá trị ab của mỗi màu trong bảng màu thành 313 ngăn màu và trực quan hóa phân bố xác suất của các giá trị ab. Chúng tôi so sánh ba biến thể mô hình của các hàm mục tiêu khác nhau: cGAN+Huber ( $\lambda_H = 100$ ), Huber ( $\lambda_H = 100$ ) và cGAN ( $\lambda_H = 0$ )



Hình 6: Các đường màu đỏ tương ứng với sự phân phối màu của các bảng màu được tạo từ ba biến thể của mô hình. Các đường màu xanh biểu thị phân phối màu thật từ dữ liệu huấn luyện. Phân kỳ KL của ba cặp phân phối được tính theo thứ tự là 0,2299, 0,2375 và 0,5833.

Chúng tôi tính toán phân kỳ Kullback-Leibler (KL) giữa phân phối bảng màu thật của dữ liệu đào tạo và phân phối bảng màu được tạo ra của các biến thể mô hình. Như được thể hiện trong biểu đồ trong của Hình 6, hàm mất mát Huber đóng một vai trò quan trọng trong việc tạo ra màu sắc phù hợp gần giống với hình ảnh gốc từ bộ dữ liệu. Nếu không có hàm tổn thất Huber, mô hình không những không khôi phục được phân bố màu tương tự như dữ liệu gốc. Mặt khác, mô hình có tổn thất cGAN + Huber ( $\lambda_H = 100$ ) ghi lại phân kỳ KL thấp nhất là 0,2299 , trong khi mô hình chỉ có tổn thất Huber ( $\lambda_H = 100$ ) ghi lại tốt thứ hai. Điều này là do thực tế là chỉ sử dụng hàm tổn thất Huber sẽ dẫn đến việc lấy trung bình một cách mù quáng trên nhiều bảng sự thật cơ bản, dẫn đến kết quả bảng màu bị khử bão hòa một chút như thể hiện trong hàng thứ hai của Hình 7. Ngược lại, mô hình có cả cGAN + Huber được học và duy trì các màu gốc khác nhau thay vì chỉ lấy trung bình của chúng, dẫn đến kết quả sáng hơn, có độ bão hòa cao như thể hiện trong hàng đầu tiên của Hình 7.



Hình 7: So sánh kết quả dự đoán bảng màu từ các biến thể mô hình khác nhau

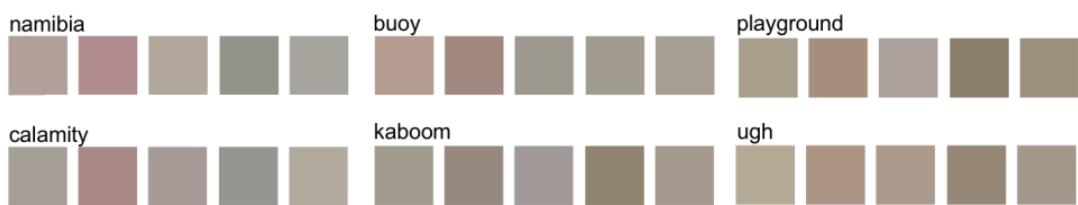
### 5.1.2 Kết quả từ mô hình tạo bảng màu

Các kết quả trong Hình 8 cho thấy kết quả đầu ra của mô hình của chúng tôi so với các bảng màu thật từ bộ dữ liệu. Nếu một từ đầu vào được nhìn thấy ít nhất một lần trong dữ liệu huấn luyện, mô hình của chúng tôi có thể xuất bảng màu liên quan đến từ đầu vào. Chẳng hạn, hãy xem bảng màu có tên 'mango and grapefruit' ở trên cùng bên trái. Từ 'grapefruit' chỉ được đưa vào một lần trong tập huấn luyện. Tuy nhiên, mô hình đã xuất thành công một bảng màu khớp với kiểu nhập văn bản. Ngoài ra, các bảng màu từ bộ dữ liệu được đưa vào để so sánh trực tiếp với các bảng được tạo. Ngay cả khi bảng dự đoán không hoàn toàn giống với bảng từ bộ dữ liệu, cả hai đều có thể được coi là màu hợp lý, điều này cho thấy khả năng khá khả thi của mô hình của chúng tôi.



Hình 8: Bảng màu được tạo ra từ bộ dữ liệu kiểm thử

Mặc dù mô hình của chúng tôi có thể tạo ra các bảng màu có ý nghĩa về mặt ngữ nghĩa một cách hiệu quả, nhưng lại gặp khó khăn khi các văn bản đầu vào không được đề cập đến trong tập huấn luyện. Không có gì đáng ngạc nhiên khi mô hình của chúng tôi không thành công và tạo ra các bảng màu có văn bản đầu vào không được đề cập tới trong bộ dữ liệu như trong Hình 9. Mặt khác, mô hình của chúng tôi vẫn có thể tạo ra các bảng màu hợp lý trong trường hợp các tổ hợp từ mới không nhìn thấy được tìm thấy trong tập huấn luyện. Ví dụ: 'bright life' trong Hình 8 được xem riêng là 'bright' và 'life' trong tập huấn luyện nhưng không được xem như là cùng nhau. Do đó, 'bright life' được phân loại là dữ liệu chưa nhìn thấy, mà mô hình của chúng tôi không gặp vấn đề gì trong việc dự đoán bảng màu từ đó.



Hình 9: Bảng màu được tạo ra bị lỗi do việc văn bản đầu vào không có trong bộ dữ liệu huấn luyện

Trong Hình 10, chúng tôi cho thấy cách mô hình của chúng tôi xử lý đầu vào ở mức độ cụm từ. Để so sánh dễ dàng hơn, tất cả các cụm từ đầu vào đều nói về 'love'. Khi xem cách mô hình của chúng tôi chọn để thể hiện sự khác biệt tinh tế về sắc thái có trong văn bản đầu vào. Lưu ý các bảng màu được tạo ra có xu hướng tối hơn đối với các kiểu nhập văn bản là cụm từ 'love' (ví dụ: 'i thought i loved you' và 'where did our love go'). Tất cả các cụm từ đầu vào có trong hình này là dữ liệu chưa có trong bộ dữ liệu huấn luyện ngoại trừ 'i love you'.



Hình 10: Bảng màu được tạo ra từ các cụm từ khác nhau về từ 'love'

Hay thông qua Hình 11. Chúng ta thấy được rằng dựa vào 1 từ trong các cụm từ từ văn bản đầu vào có thể tạo ra được bảng màu. Tuy nhiên, nó sẽ khác xa màu của ảnh gốc. Nhưng nó sẽ không quá khác so với những văn bản đầu vào hoàn toàn không có trong bộ dữ liệu như Hình 9 đã được thể hiện. Hình 11 sẽ cho chúng ta thấy bảng màu được tạo

phù hợp lẫn về từng từ và cụm từ trong văn bản đầu vào để tạo ra được một bảng màu phù hợp với ngữ cảnh và ngữ nghĩa của văn bản

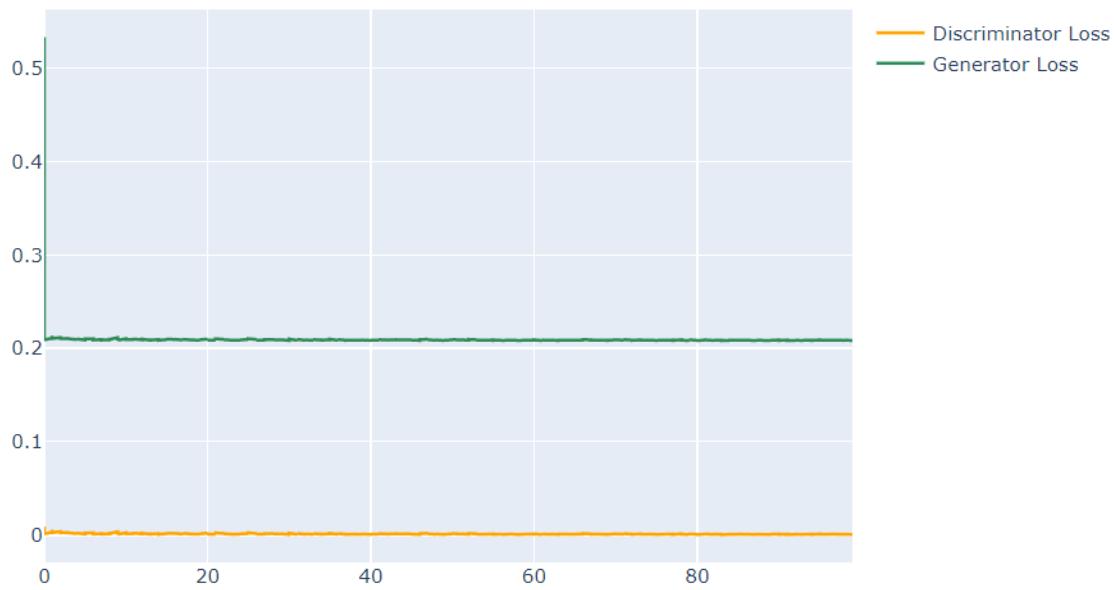
<u>rose</u>	white <u>rose</u>	black <u>rose</u>	wilt <u>rose</u>
<u>rainbow</u>	melancholy <u>rainbow</u>	rainbow in pastel	little bit of <u>rainbow</u>
<u>heart</u>	pastel candy <u>heart</u>	without a <u>heart</u>	furious <u>heart</u>
<u>autumn</u>	romantic <u>autumn</u>	autumnal breeze	warm <u>autumn</u>

Hình 11: Mô hình phản ánh sự khác biệt về sắc thái trong ngữ cảnh ngữ nghĩa của một kiểu nhập văn bản nhất định thông qua các bảng màu. Ngoại trừ cột đầu tiên, tất cả các tổ hợp văn bản có trong hình này đều là dữ liệu không được đề cập đến trong bộ dữ liệu của chúng tôi

## 5.2 Palette-based Colorization Networks (PCN)

Chúng tôi trình bày các kết quả tô màu trên các bộ dữ liệu bao gồm CUB-200-2011 (CUB dataset), ImageNet ILSVRC Object Detection (ImageNet dataset) và Graphical Pattern images (Pattern images). Trong các hình dưới đây, các cột ngoài cùng bên trái hiển thị các hình ảnh với thang độ xám và văn bản đầu vào. Các bảng màu dọc bên cạnh các hình ảnh thang độ xám là những bảng màu được tạo từ văn bản đầu vào. Và ở đây chúng tôi đã tạo ra được năm bảng màu từ văn bản đầu vào. Sau đó, lấy ngẫu nhiên một trong năm bảng màu đã được tạo để tiến hành tô màu lên ảnh xám. Đầu ra đã được tô màu với bảng màu được tạo. Những kết quả này cho thấy mô hình của chúng tôi sử dụng hiệu quả các bảng màu được tạo trong quá trình tô màu. Hình ảnh được tô màu có thể khác với màu của ảnh gốc do mô hình của chúng tôi kết hợp các gợi ý màu bổ sung. Chúng tôi hiển thị hình ảnh gốc ban đầu ở bên phải để so sánh mức độ khác nhau của một hình ảnh sau khi áp dụng các bảng màu.

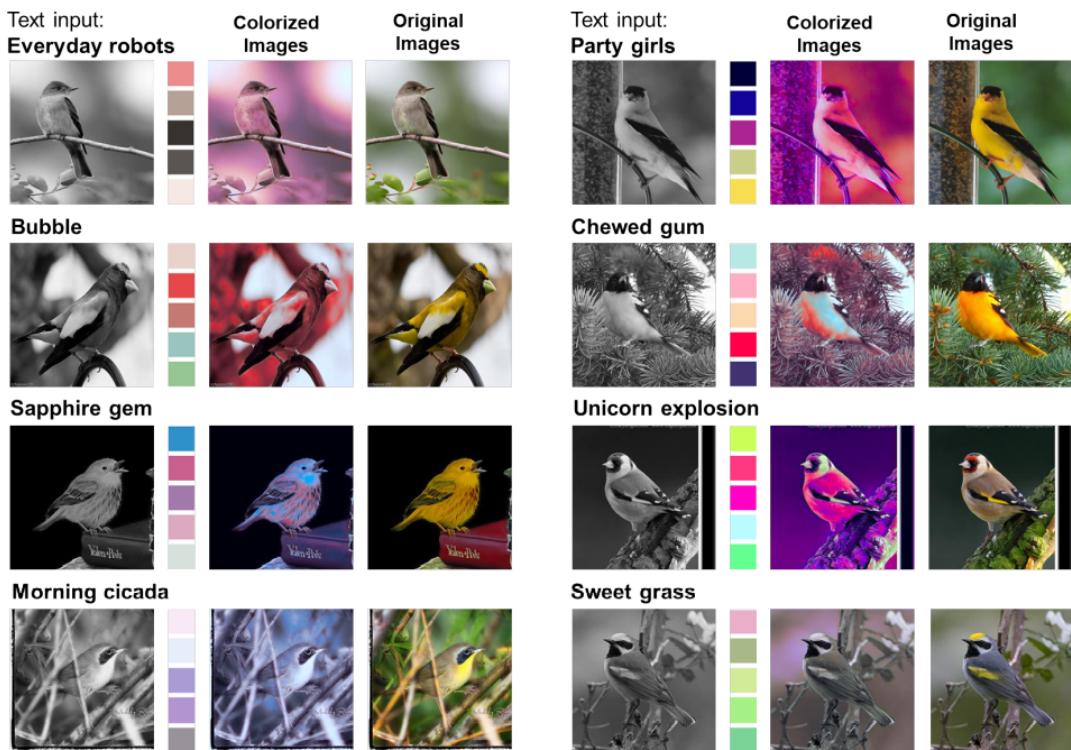
Sau khi huấn luyện mô hình với 100 epochs thì kết quả độ lỗi thu được ở mô hình PCN thông qua Hình 12



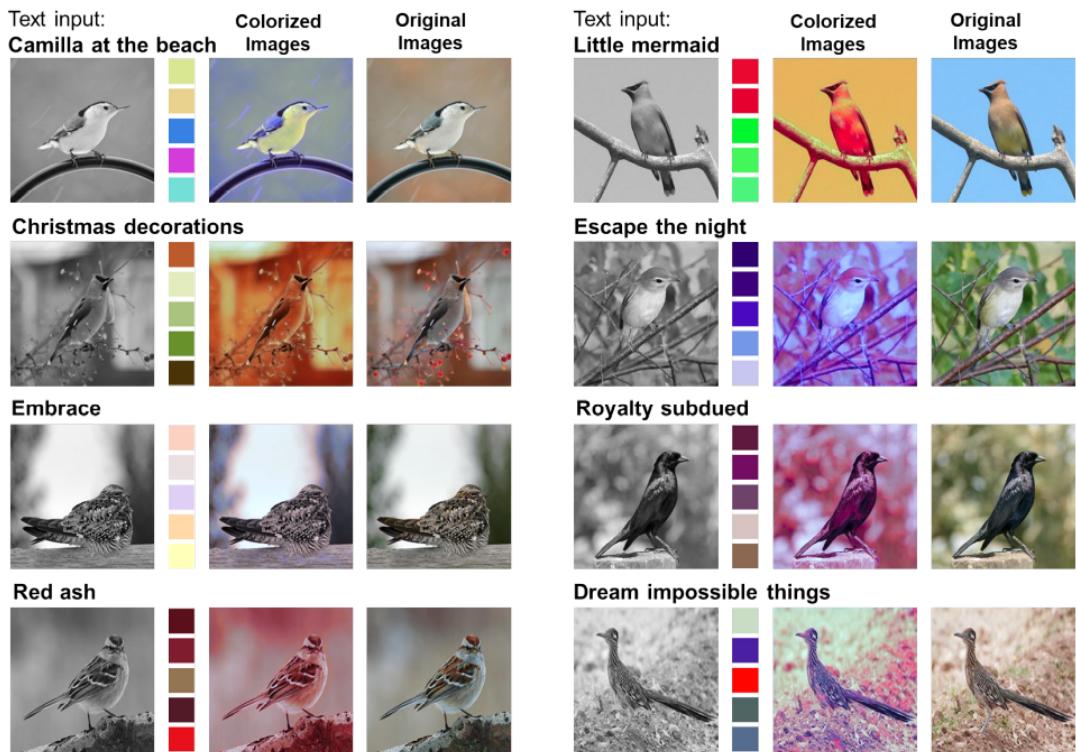
Hình 12: Phân tích kết quả sau khi huấn luyện mô hình

### 5.2.1 CUB-200-2011(CUB dataset)

Hình 13 và 14 hiển thị kết quả tô màu trên bộ dữ liệu CUB.



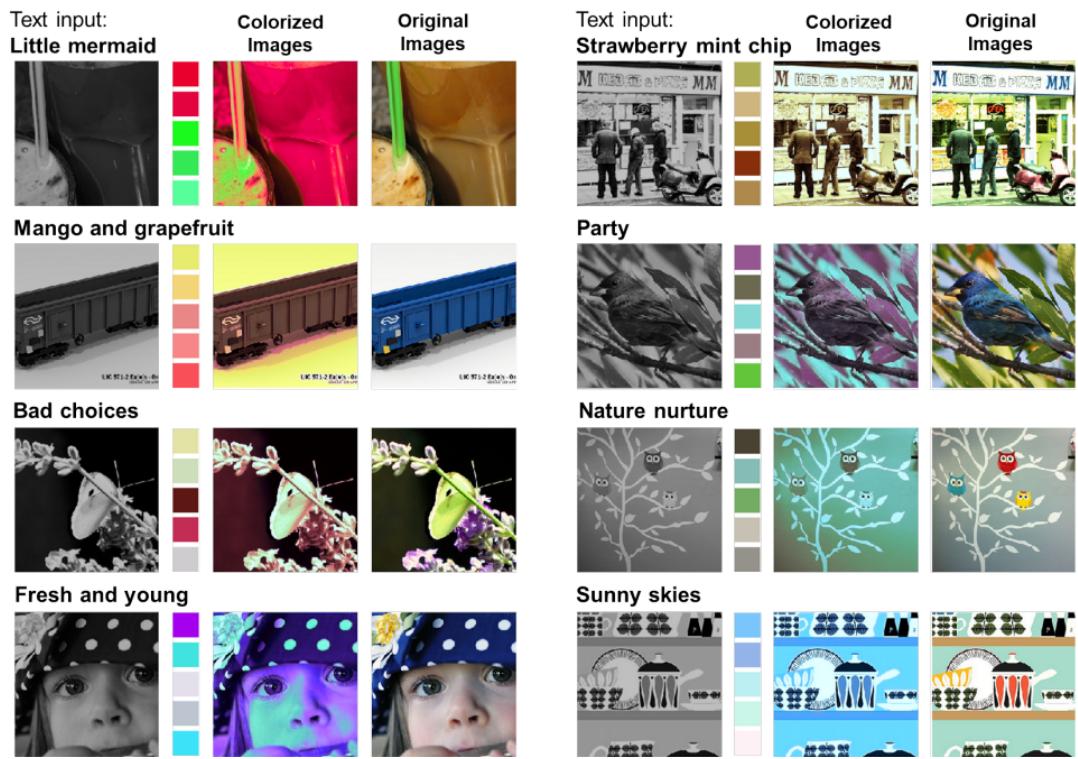
Hình 13: Kết quả trên tập dữ liệu CUB (1)



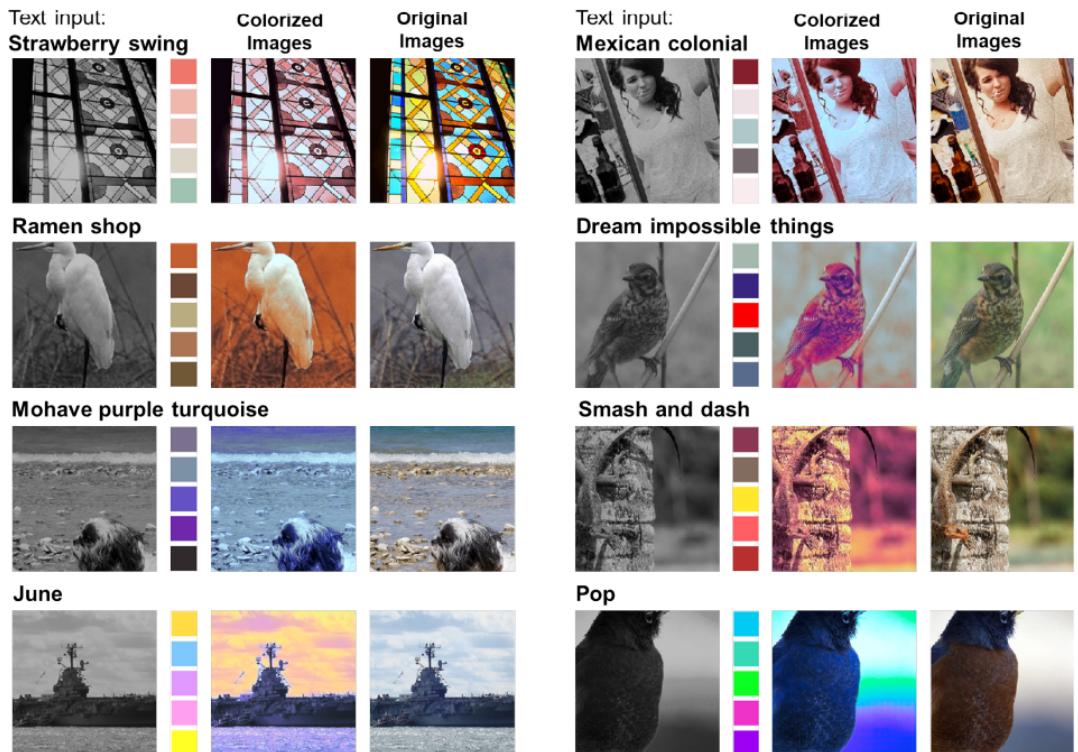
Hình 14: Kết quả trên tập dữ liệu CUB (2)

### 5.2.2 ImageNet ILSVRC Object Detection (ImageNet dataset)

Hình 15 và 16 hiển thị kết quả tô màu trên bộ dữ liệu ImageNet dataset.



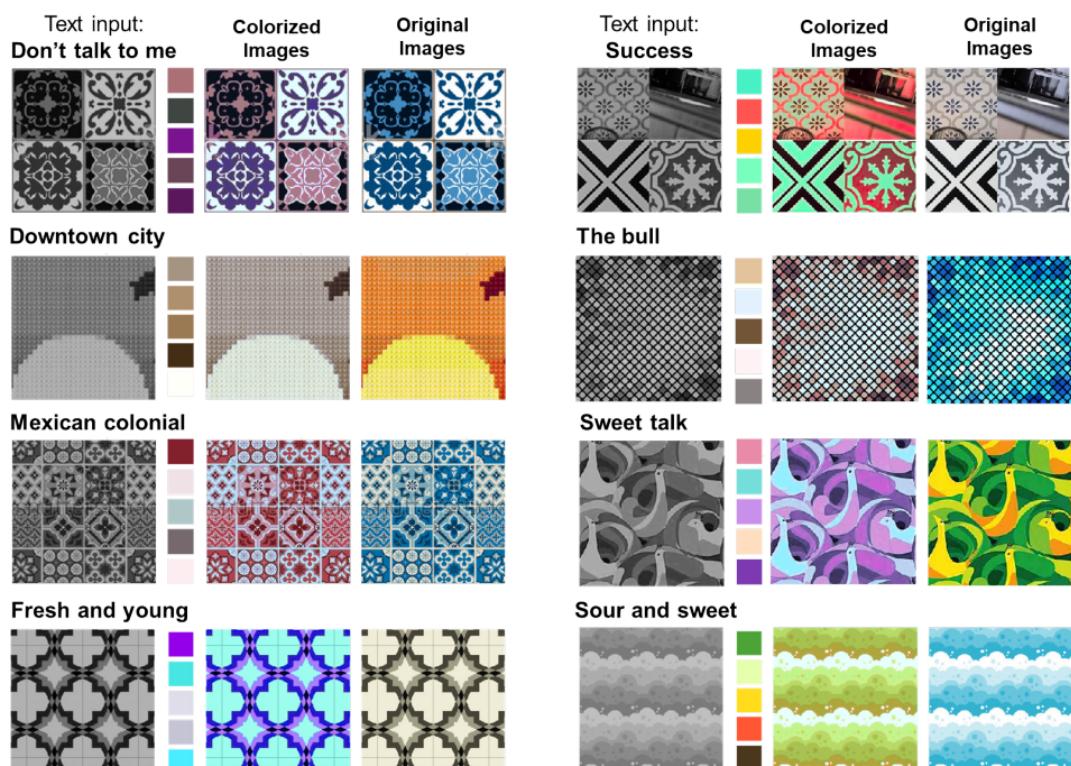
Hình 15: Kết quả trên tập dữ liệu ImageNet (1)



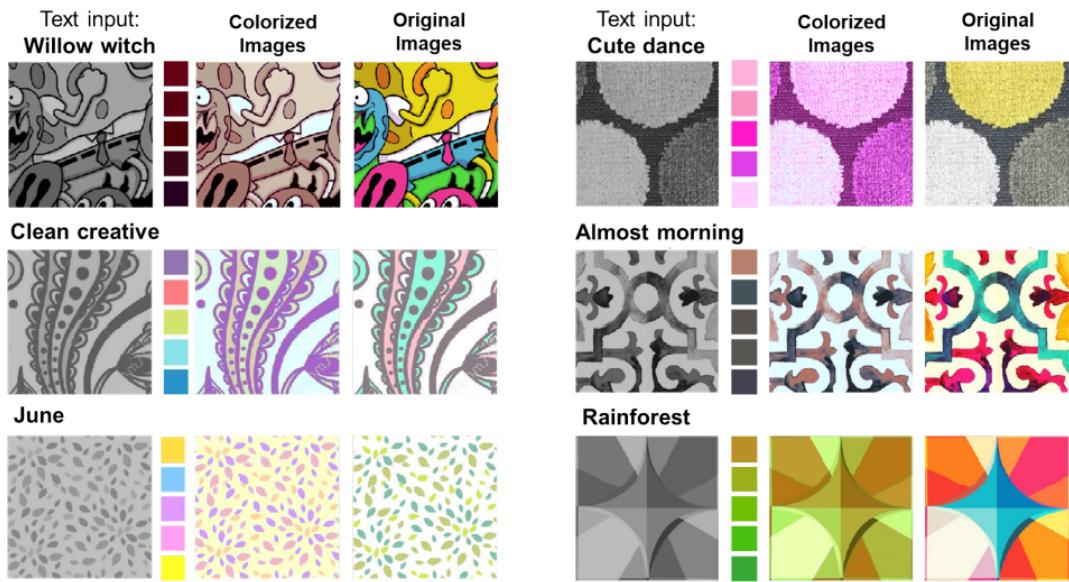
Hình 16: Kết quả trên tập dữ liệu ImageNet (1)

### 5.2.3 Graphical Pattern Images

Mô hình PCN đã đưa ra kết quả một cách đáng ngạc nhiên trên các loại hình ảnh khác. Mô hình đề xuất của chúng tôi được đào tạo trên bộ dữ liệu CUB, phần lớn được tạo thành từ các hình ảnh về loài chim. Thay vì hình ảnh về những loài chim, chúng tôi đã sử dụng mô hình tô màu của mình để tô màu hình ảnh đồ họa. Hình ảnh mẫu đồ họa được thu thập thông tin từ tìm kiếm từ Google với các từ khóa như 'pattern' 'fabric pattern' và 'beautiful patterns'. Như đã thấy trong Hình 17 và 18, hình ảnh đồ họa khá khác so với hình ảnh về các loài chim. Các kết quả đầu ra được tô màu cho thấy rằng mô hình của chúng tôi có thể áp dụng các bảng màu đã tạo cho các hình ảnh có hình dạng và kết cấu đa dạng. Các kết quả cho thấy một cách định tính rằng mô hình tô màu dựa trên bảng màu của chúng tôi có thể chuyển sang các loại hình ảnh khác nhau.



Hình 17: Kết quả trên Graphical Pattern Images (1)



Hình 18: Kết quả trên Graphical Pattern Images (2)

## 6 Kết luận

Chúng tôi đã đề xuất một mô hình tổng quát có thể tạo ra nhiều bảng màu từ kiểu văn bản đầu vào có định dạng và tô màu ảnh xám bằng các bảng màu đã tạo ra. Đánh giá chung thì mô hình TPN của chúng tôi có thể tạo ra bảng màu khá hợp lý từ văn bản đầu vào và có thể kết hợp tính chất đa dạng của màu sắc. Mặc dù đối bảng màu tạo ra có thể không giống với bảng màu từ bộ dữ liệu nhưng nó được coi là hợp lý với văn bản đầu vào.

Kết quả định tính trên mô hình PCN của chúng tôi cũng cho thấy rằng các màu sắc đa dạng trong bảng màu được phản ánh hiệu quả trong kết quả tô màu. Tuy nhiên mô hình tô màu của chúng tôi chưa thực sự cho ảnh có màu sắc tốt nhất dựa trên bảng màu nhưng tổng quát nó hợp lý với bảng màu đưa ra. Hướng phát triển bao gồm mở rộng mô hình của chúng tôi sang phạm vi tác vụ rộng hơn yêu cầu đề xuất màu sắc và tiến hành phân tích chi tiết bộ dữ liệu của chúng tôi.

## Tài liệu

1. Kobayashi, S.: Color image scale. [http://www.ncd-ri.co.jp/english/main\\_0104.html](http://www.ncd-ri.co.jp/english/main_0104.html) (2009)
2. Liu, Y., Cohen, M., Uyttendaele, M., Rusinkiewicz, S.: Autostyle: Automatic style transfer from image collections to users' images. In: Computer Graphics Forum. Volume 33., Wiley Online Library (2014) 21–31
3. Solli, M., Lenz, R.: Color semantics for image indexing. In: Proceedings of Conference on Colour in Graphics Imaging and Vision (CGIV). Volume 2010., Society for Imaging Science and Technology (2010) 353–358
4. Charpiat, G., Hofmann, M., Schölkopf, B.: Automatic image colorization via multimodal predictions. In: Proceedings of the IEEE European Conference on Computer Vision(ECCV), Springer (2008) 126–139
5. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Transactions on Graphics (TOG) (2017)
6. Cho, J., Yun, S., Lee, K., Choi, J.Y.: Palettenet: Image recolorization with given color palette. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2017) 62–70
7. Jahanian, A., Keshvari, S., Vishwanathan, S., Allebach, J.P.: Colors—messengers of concepts: Visual design mining for learning color semantics. ACM Transactions on Computer Human Interaction (TOCHI) 24(1) (2017)
8. Murray, N., Skaff, S., Marchesotti, L., Perronnin, F.: Toward automatic and flexible concept transfer. Computers & Graphics 36(6) (2012) 622–634
9. McMahan, B., Stone, M.: A bayesian model of grounded color semantics. Transactions of the Association of Computational Linguistics (ACL) 3(1) (2015) 103–115
10. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
11. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. Proceedings of the International Conference on Machine Learning (ICML) (2016)
12. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine grained visual descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 49–58
13. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: Text to photo realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017) 5907–5915
14. Kim, T., Cha, M., Kim, H., Lee, J., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. Proceedings of the International Conference on Machine Learning (ICML) (2017)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition (CVPR) (2017)

16. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
17. Li, X., Zhao, H., Nie, G., Huang, H.: Image recoloring using geodesic distance based color harmonization. Computational Visual Media 1(2) (2015) 143–155
18. Chang, H., Fried, O., Liu, Y., DiVerdi, S., Finkelstein, A.: Palette-based photo recoloring. ACM Transactions on Graphics (TOG) 34(4) (2015) 139
19. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). (2015) 1422–1432
20. Sutskever, I., Martens, J., Hinton, G.E.: Generating text with recurrent neural networks. In: Proceedings of the International Conference on Machine Learning (ICML). (2011) 1017–1024
21. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems (NIPS). (2014) 3104–3112
22. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2015)
23. Munroe, R.: Color survey results. Online at <http://blog.xkcd.com/2010/05/03/color-surveyresults> (2010)
24. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP). (2014) 1532–1543
25. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 2536–2544.
26. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Proceedings of the International Conference on Learning Representations (ICLR) (2014).
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Springer (2015) 234–241
28. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. Proceedings of the International Conference on Learning Representations (ICLR) (2015)
29. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Technical report (2011)