

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN

Phạm Thái Huy - Nguyễn Đức Mạnh -  
Lê Quang Vĩnh Quyền

CẢI TIẾN  
NGHIÊN CỨU HỌC BIỂU DIỄN  
KHÔNG GIAN-THỜI GIAN PHÂN CẤP  
CHO BÀI TOÁN NHẬN DẠNG DÁNG ĐI

BÁO CÁO CUỐI KỲ - MÔN SINH TRẮC HỌC  
CHƯƠNG TRÌNH CHUẨN

Tp. Hồ Chí Minh, Ngày 3 tháng 1 năm 2026

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN

Phạm Thái Huy - 21120081  
Nguyễn Đức Mạnh - 22120204  
Lê Quang Vĩnh Quyền - 22120307

CẨI TIẾN  
NGHIÊN CỨU HỌC BIỂU DIỄN  
KHÔNG GIAN-THỜI GIAN PHÂN CẤP  
CHO BÀI TOÁN NHẬN DẠNG DÁNG ĐI

BÁO CÁO CUỐI KỲ - MÔN SINH TRẮC HỌC  
CHƯƠNG TRÌNH CHUẨN

GIẢNG VIÊN HƯỚNG DẪN  
PGS.TS. Lê Hoàng Thái  
ThS. Dương Thái Bảo

Tp. Hồ Chí Minh, Ngày 3 tháng 1 năm 2026

# Lời cam đoan

Chúng tôi xin cam đoan rằng đây là công trình nghiên cứu của chính chúng tôi. Dữ liệu và kết quả nghiên cứu trình bày trong báo cáo khóa luận là trung thực và không trùng lặp với bất kỳ đề tài nào khác. Chúng tôi cam kết rằng toàn bộ nội dung của báo cáo này không sao chép hay đạo văn từ bất kỳ công trình nghiên cứu nào khác.

Trong suốt quá trình thực hiện khóa luận, chúng tôi luôn chấp hành đầy đủ các quy định về đạo đức nghiên cứu, đảm bảo trích dẫn đúng và đầy đủ tất cả các nguồn tài liệu tham khảo, đồng thời giữ vững tinh thần trung thực học thuật trong mọi khía cạnh của báo cáo.

# Lời cảm ơn

## Có cần không ?

Trước hết, chúng tôi xin bày tỏ lòng biết ơn chân thành và sâu sắc đến PGS.TS. Lê Hoàng Thái và ThS. Dương Thái Bảo vì sự hướng dẫn tận tình, khích lệ và hỗ trợ quý báu, giúp chúng tôi có cơ hội khám phá các hướng nghiên cứu mới.

Trong suốt quá trình thực hiện đề tài, quý Thầy đã chia sẻ những kiến thức giá trị và đóng góp nhiều gợi ý sâu sắc, giúp chúng tôi định hướng được con đường nghiên cứu của riêng nhóm thông qua các buổi trao đổi từ những ngày đầu cho đến khi hoàn thành nghiên cứu.

# Mục lục

<b>Lời cảm ơn</b>	<b>ii</b>
<b>Mục lục</b>	<b>iii</b>
<b>Danh sách hình vẽ</b>	<b>vi</b>
<b>Danh sách bảng</b>	<b>vii</b>
<b>Tóm tắt</b>	<b>vii</b>
<b>1 Giới thiệu</b>	<b>1</b>
1.1 Bối cảnh chung . . . . .	1
1.2 Động lực . . . . .	2
1.3 Phát biểu bài toán . . . . .	2
1.4 Thách thức của bài toán . . . . .	3
1.5 Đóng góp . . . . .	3
1.6 Bố cục của khóa luận . . . . .	4
<b>2 Các công trình liên quan</b>	<b>5</b>
2.1 Sự phát triển của tô màu ảnh độ xám . . . . .	6
2.1.1 Các phương pháp truyền thống . . . . .	6
2.1.2 Sự xuất hiện của mạng nơ-ron tích chập . . . . .	7
2.1.3 Mạng đối kháng tạo sinh . . . . .	7
2.1.4 Kỹ thuật khuếch tán . . . . .	7
2.2 Tô màu bán tự động và tô màu tự động . . . . .	7

2.2.1	Tô màu tự động . . . . .	8
2.2.2	Tô màu bán tự động . . . . .	8
2.3	Mô hình khuếch tán trong không gian tiềm ẩn . . . . .	9
2.4	Kết hợp điều kiện không gian . . . . .	9
<b>3</b>	<b>Phương pháp đề xuất</b>	<b>10</b>
3.1	Tổng quan về mô hình . . . . .	11
3.2	Thiết lập tín hiệu điều khiển . . . . .	12
3.2.1	Bộ mã hóa văn bản của CLIP . . . . .	12
3.2.2	Bộ mã hóa điều kiện . . . . .	13
3.2.3	Mô hình BLIP . . . . .	13
3.3	Mô hình tạo sinh ảnh có kết hợp điều kiện không gian . . . . .	14
3.3.1	Mô hình khuếch tán . . . . .	15
3.3.2	Mô hình khuếch tán trong không gian tiềm ẩn . . . . .	15
3.3.3	ControlNet . . . . .	15
<b>4</b>	<b>Thực nghiệm</b>	<b>18</b>
4.1	Mục tiêu thực nghiệm . . . . .	18
4.2	Chi tiết quá trình huấn luyện . . . . .	19
4.2.1	Tập dữ liệu huấn luyện . . . . .	19
4.2.2	Quá trình tiền xử lý dữ liệu . . . . .	19
4.2.3	Kết quả quá trình huấn luyện . . . . .	20
4.3	Thử nghiệm mô hình tô màu tổng quát . . . . .	21
4.3.1	Tập dữ liệu thử nghiệm . . . . .	21
4.3.2	Kết quả thử nghiệm . . . . .	22
4.4	Phân tích kết quả . . . . .	22
<b>5</b>	<b>Kết luận</b>	<b>24</b>
5.1	Kết luận . . . . .	24
5.2	Bàn luận . . . . .	25
<b>Tài liệu tham khảo</b>		<b>26</b>

<b>A Phụ lục</b>	<b>29</b>
A.1 Bảng đối chiếu thuật ngữ . . . . .	29
A.2 Mã nguồn mẫu . . . . .	29
A.3 Thông tin bổ sung . . . . .	30

# Danh sách hình vẽ

3.1	Tổng quan về mô hình tô màu ảnh độ xám dựa trên mô hình tổng hợp ảnh có điều kiện sử dụng quy trình khuếch tán.	11
3.2	Kiến trúc của mô hình CLIP.	12
3.3	Cấu trúc mạng mã hóa điều kiện $\mathcal{E}_c$ .	13
3.4	Kiến trúc của hỗn hợp các cặp bộ mã hóa-giải mã đa phương thức MED.	14
3.5	Kiến trúc của mô hình khuếch tán trong không gian tiềm ẩn (LDM).	16
3.6	Kiến trúc của ControlNet.	16
4.1	Chất lượng ảnh được tạo sinh trong quá trình huấn luyện. Ở lân cận bước thứ 3200, hiện tượng <b>hội tụ đột ngột</b> xảy ra.	21
4.2	Ví dụ kết quả tô màu. (a) Ảnh độ xám đầu vào. (b) Ảnh được tô màu bởi mô hình đề xuất. (c) Ảnh màu gốc để so sánh.	23

# Danh sách bảng

4.1	Thống kê số lượng ảnh theo từng tập huấn luyện sau khi tiền xử lý. . . . .	20
4.2	Kết quả so sánh các mô hình tô màu tự động trên các tập dữ liệu đánh giá. . . . .	22
A.1	Phụ lục đối chiếu Việt Anh . . . . .	29

# Tóm tắt

## Phần quy định chung khi trình bày báo cáo:

- Tất cả biểu thức toán phải được đánh số.
- Tất cả hình vẽ, bảng biểu phải có chú thích, và đánh dấu để tham chiếu trong mục lục.
- Hạn chế tối đa việc dùng tiếng anh, kể cả các thuật ngữ chuyên môn, có gắng dịch sang tiếng việt với từ ngữ dễ hiểu. Trong trường hợp quá khó để phiên dịch, được phép dùng tiếng anh, nhưng phải bổ sung ở bảng A.1 trong phần **Phụ lục**.
- Các tài liệu đính kèm phải theo một chuẩn duy nhất, không tham khảo loạn xạ. Có nhiều cách để lấy định dạng chuẩn, cách đơn giản nhất là tìm bài báo đó trên trang **arxiv**, tìm dòng **export BibTeX citation**, sao chép định dạng đó vào file **references.bib**.

## Viết theo mẫu bên dưới

Tô màu ảnh độ xám là quá trình ước lượng các giá trị màu RGB cho từng điểm trong ảnh độ xám, nhằm chuyển đổi chúng thành ảnh màu. Tô màu ảnh độ xám được chia thành hai nhóm lớn: tô màu tự động và tô màu bán tự động. Tuy nhiên, hầu hết các mô hình hiện tại chỉ tập trung vào một phương thức duy nhất, điều này hạn chế tính năng mà người dùng mong muốn.

Trong nghiên cứu này, nhóm giới thiệu một mô hình tô màu ảnh độ xám kết hợp cả hai phương thức tự động và bán tự động (dựa trên lời

nhắc văn bản) được phát triển dựa trên kiến trúc mô hình ControlNet. ControlNet là một kiến trúc mạng nơ-ron được thiết kế nhằm tích hợp các điều kiện điều khiển không gian vào các mô hình tạo sinh ảnh từ văn bản đã được huấn luyện trước.

Nhóm đã tiến hành huấn luyện mô hình trên bốn tập dữ liệu cho các mục đích khác nhau. Sau đó nhóm tiến hành thử nghiệm mô hình tô màu chính trên ba tập dữ liệu đánh giá khác nhau. Kết quả đánh giá cho thấy, mô hình của nhóm đã vượt qua hầu hết các phương pháp tô màu tiên tiến trước đó trên phương thức tô màu tự động về chỉ số Colorfulness. Cuối cùng, nhóm xây dựng thêm một giao diện đơn giản để người dùng có thể sử dụng và ứng dụng các mô hình vào các công việc có liên quan đến tô màu ảnh độ xám.

# Chương 1

## Giới thiệu

Mục này được viết dựa trên các bài khảo sát (survey) để giới thiệu tổng quan về chủ đề lớn Nhận dạng đáng đi được nghiên cứu. Cách tổ chức nội dung bên dưới có thể tham khảo, hoặc chỉnh sửa tùy cho phù hợp ngữ cảnh.

### 1.1 Bối cảnh chung

Với sự phát triển của trí tuệ nhân tạo, các ứng dụng xử lý ảnh ngày càng được cải thiện, đặc biệt trong các tác vụ dịch ảnh sang ảnh như tô màu ảnh độ xám. Nhu cầu phục chế màu cho các bức ảnh trắng đen ngày càng gia tăng, đặc biệt trong các ứng dụng liên quan đến lịch sử và xã hội.

Về mặt thực tiễn, nhu cầu tô màu ảnh độ xám đến từ nhiều lĩnh vực như phục dựng ảnh lịch sử, tô màu ảnh phác họa trong thiết kế. Về mặt nghiên cứu, một số vấn đề trong chủ đề tô màu ảnh độ xám vẫn chưa được giải quyết triệt để, bao gồm phương thức tô màu tự động và bán tự động, vấn đề về dữ liệu huấn luyện, và chi phí tính toán.

## 1.2 Động lực

Kỳ vọng của nghiên cứu này là tạo ra một mô hình tô màu có thể đáp ứng các ứng dụng như phục hồi ảnh trắng đen lịch sử. Sản phẩm cuối cùng sẽ cho phép người dùng có thể tô màu theo cả hai phương pháp, tô màu tự động hoặc bán tự động dựa trên nhu cầu.

Bài toán tô màu là một bài toán tạo sinh ảnh không đơn trị, tức là từ một ảnh độ xám có thể có nhiều cách tô màu hợp lý khác nhau. Do đó, một mô hình hiệu quả không chỉ cần tái tạo ảnh màu đẹp mắt mà còn cần hiểu được ngữ cảnh và nội dung ngữ nghĩa trong ảnh.

## 1.3 Phát biểu bài toán

Bài toán tô màu ảnh độ xám là việc **ước tính các giá trị màu của các điểm ảnh trong một ảnh độ xám**. Hệ thống tô màu một ảnh độ xám sẽ gồm các thành phần:

- **Đầu vào:**  $I_g$  là ma trận đại diện cho **ảnh độ xám** cần được tô và  $c$  là điều kiện điều khiển việc tô màu (nếu có).
- **Đầu ra:**  $I_{rgb}$  là ảnh màu trong không gian RGB do hệ thống trả về.

Phương trình dự đoán ảnh màu có thể viết như sau:

$$I_{rgb} = \Phi(I_g). \quad (1.1)$$

Nếu quá trình tô màu được điều khiển bởi điều kiện  $c$ , phương trình 1.1 có thể được viết lại thành:

$$I_{rgb} = \Phi(I_g, c). \quad (1.2)$$

Với  $\Phi(\cdot)$  là hàm số được xây dựng để tạo ra ảnh màu từ ảnh độ xám đầu vào một cách chân thật và phù hợp với điều kiện điều khiển.

## 1.4 Thách thức của bài toán

Thách thức đầu tiên là vấn đề rất khó để có thể kiểm tra được kết quả của quá trình tô màu bằng phương pháp định lượng. Với một bức ảnh độ xám đầu vào, ta gần như có vô số cách tô màu, và kết quả đó có phù hợp hay không chỉ có thể được đánh giá qua cảm nhận của người sử dụng.

Một thách thức khác là việc nhiều bức ảnh cùng chụp 1 đối tượng nhưng có thể bị ảnh hưởng bởi nhiều yếu tố khác nhau như góc chụp, độ chói, độ sáng. Ngoài ra, chúng ta cũng rất khó có thể xác định màu cho một số đối tượng có màu tùy biến như quần áo, phương tiện đi lại, đồ nội thất.

## 1.5 Đóng góp

Trong khóa luận này, mục tiêu là xây dựng một mô hình tô màu ảnh độ xám đa phương thức dựa trên mô hình khuếch tán. Đóng góp bao gồm:

### Về mặt lý thuyết:

- Xây dựng một phương pháp tô màu ảnh độ xám đa phương thức dựa trên mô hình khuếch tán có kết hợp điều kiện.
- Kết hợp bộ giải mã biến dạng để giải quyết vấn đề tràn màu và tô màu sai.
- Kết hợp mô hình ngôn ngữ ảnh BLIP vào quá trình tạo sinh dữ liệu văn bản.

### Về mặt thực tiễn:

- Thu thập, xử lý và cung cấp các tập dữ liệu dùng cho quá trình huấn luyện.
- Tạo ra một mô hình tô màu ảnh độ xám chung cho nhiều miền dữ liệu.

- Cung cấp một giao diện đơn giản, dễ sử dụng và có khả năng tùy chỉnh cao.

## **1.6 Bố cục của khóa luận**

Báo cáo khóa luận bao gồm năm chương: Chương 1 trình bày bối cảnh, động lực nghiên cứu và phát biểu bài toán; Chương 2 trình bày các công trình nghiên cứu liên quan; Chương 3 trình bày chi tiết phương pháp đề xuất; Chương 4 trình bày kết quả thực nghiệm; Chương 5 trình bày kết luận và hướng phát triển.

## Chương 2

# Các công trình liên quan

Mục này đi sâu hơn vào các nghiên cứu/phương pháp cho chủ đề Nhập dạng đáng đi. Gợi ý:

- Phần lớn nội dung mục này dựa trên các bài tổng hợp/khảo sát.
- Liệt kê tất cả các phương pháp đến thời điểm hiện tại **2026**. Một số cách trình bày (chắc chắn được dùng trong các bài tổng hợp, nên tham khảo kĩ):
  - Vẽ lược đồ thời gian công bố của các nghiên cứu.
  - Vẽ sơ đồ chia các nghiên cứu theo các nhánh tiếp cận chính.
- Từng nhánh tiếp cận bài toán sẽ trình bày sâu hơn trong từng mục nhỏ:
  - Liệt kê theo trình tự thời gian
  - Trình ý tưởng/nguyên lý của từng nghiên cứu được đề cập.
  - Kết luận bằng điểm mạnh và hạn chế chung của nhóm các phương pháp.
- **Lưu ý:** Khi trình bày đến nhánh tiếp cận chính (**có chứa nghiên cứu được chọn (HSTL)**), nên trình bày nhiều hơn cả về điểm mạnh (ngầm

ý đây là phương pháp được quan tâm) và điểm yếu (ngầm ý đóng góp của nghiên cứu này).

- Đối với các nhánh tiếp cận hoặc nghiên cứu vượt trội nghiên cứu được chọn (**HSTL**) về hiệu năng, hãy nói rõ hạn chế lớn của chúng. Quan trọng hơn hết, hãy trung thực về lý do không lựa chọn đào sâu, ví dụ, chi phí (huấn luyện, lưu trữ dữ liệu) lớn (**gần như là điều hiển nhiên khi chi phí lớn sẽ cho hiệu năng lớn**), **năng lực nghiên cứu hạn chế của nhóm**, ...
- Để người đọc dễ tiếp cận, nên chèn thêm các hình vẽ minh họa về dữ liệu của từng nhánh tiếp cận.
- **Nội dung có sẵn khá lan man, không nên viết theo.**

## 2.1 Sự phát triển của tô màu ảnh độ xám

Tô màu ảnh độ xám đã trải qua nhiều giai đoạn phát triển, đặc biệt với sự xuất hiện của các công nghệ học sâu. Ban đầu, các phương pháp tô màu chủ yếu dựa vào kỹ thuật thủ công như được đề cập tại bài viết của Ethan [1].

### 2.1.1 Các phương pháp truyền thống

Trước khi học sâu trở thành xu hướng, các phương pháp tô màu ảnh chủ yếu dựa vào các thuật toán xử lý ảnh truyền thống. Tuy nhiên như được trình bày trong nghiên cứu của Levin và các đồng nghiệp [2], những phương pháp này thường gặp nhiều khó khăn trong việc tạo ra màu sắc tự nhiên và chân thực.

## 2.1.2 Sự xuất hiện của mạng nơ-ron tích chập

Sự phát triển của mạng nơ-ron tích chập đã đánh dấu một bước ngoặt quan trọng. Nghiên cứu của Zhang và các đồng nghiệp [3] đã giới thiệu mô hình "Colorful Image Colorization", sử dụng mạng nơ-ron tích chập để tự động dự đoán màu sắc. Nghiên cứu của Iizuka [4] cũng đã đóng góp với việc phát triển một hệ thống tô màu có khả năng xử lý các bức ảnh với độ phức tạp cao.

## 2.1.3 Mạng đối kháng tạo sinh

Sự phát triển của các mạng tạo sinh đối kháng đã mở ra những khả năng mới. DeOldify [5] đã áp dụng mạng này để phục hồi và tô màu cho các bức ảnh lịch sử. BigColor [6] và ChromaGAN [7] đại diện cho hai hướng tiếp cận bổ sung lẫn nhau trong lĩnh vực này.

## 2.1.4 Kỹ thuật khuếch tán

Gần đây, kỹ thuật khuếch tán đã được áp dụng vào lĩnh vực tạo sinh ảnh. Các mô hình khuếch tán như DALL-E 2 [8] và Stable Diffusion [9] đã cho thấy khả năng tạo ra hình ảnh chất lượng cao. Nghiên cứu của Ho và các đồng nghiệp [10] đã chỉ ra rằng mô hình khuếch tán có thể tạo ra các bức ảnh với độ chi tiết và chân thực cao.

Palette [11] là một mô hình tiên phong sử dụng kỹ thuật khuếch tán cho các tác vụ dịch ảnh sang ảnh, bao gồm cả tô màu ảnh độ xám. Nghiên cứu này sử dụng kiến trúc U-Net cùng với kỹ thuật tự chú ý để học quá trình loại bỏ nhiễu từ ảnh màu bị nhiễu.

## 2.2 Tô màu bán tự động và tô màu tự động

Các phương pháp tô màu có thể được phân chia thành hai loại chính: tô màu tự động và tô màu bán tự động.

## 2.2.1 Tô màu tự động

Tô màu tự động là một quy trình hoàn toàn tự động, trong đó các thuật toán học sâu được sử dụng để chuyển đổi hình ảnh đen trắng thành màu mà không cần sự can thiệp của người dùng. Mô hình của Zhang và các đồng nghiệp [3] là một nghiên cứu nổi bật trong lĩnh vực này. Các mô hình dựa trên mạng tạo sinh đối kháng như DeOldify [5] hay ChromaGAN [7] cũng đã được phát triển.

## 2.2.2 Tô màu bán tự động

Tô màu bán tự động cho phép người dùng can thiệp vào quá trình tô màu thông qua các điều kiện đầu vào khác nhau. Phương pháp này được chia thành ba loại chính:

**Tô màu dựa trên nét vẽ:** Phương pháp này cho phép người dùng chỉ định màu cho các khu vực cụ thể bằng các nét vẽ. Các mô hình như Icolorit [12], UniColor [13] cho phép sử dụng gọi ý nét vẽ trong tô màu.

**Tô màu dựa trên ảnh mẫu:** Phương pháp này sử dụng một hình ảnh mẫu để điều chỉnh màu sắc cho ảnh độ xám, như trong nghiên cứu của He và các đồng nghiệp [14].

**Tô màu dựa trên lời nhắc văn bản:** Phương pháp này sử dụng lời nhắc bằng văn bản để hỗ trợ quá trình tô màu. Nghiên cứu của Weng và các đồng nghiệp [15] đã tách biệt hai thông tin về đối tượng và màu trong câu hướng dẫn.

UniColor [13] là một mô hình đa phương thức đã kết hợp tất cả ba loại điều kiện bằng cách biểu diễn chúng dưới dạng các điểm gợi ý.

## 2.3 Mô hình khuếch tán trong không gian tiềm ẩn

Mô hình khuếch tán trong không gian tiềm ẩn được nghiên cứu bởi Rombach và các đồng nghiệp [9] là một cải tiến của các mô hình khuếch tán truyền thống, trong đó quá trình khuếch tán được thực hiện trong không gian tiềm ẩn thay vì không gian điểm ảnh thông thường. Bằng cách này, mô hình có thể giảm thiểu chi phí tính toán và tăng cường khả năng sinh ảnh.

## 2.4 Kết hợp điều kiện không gian

ControlNet [16] là một phương pháp tiên tiến cho phép tích hợp các điều kiện vào trong quá trình sinh ảnh của các mô hình tạo sinh ảnh dựa trên văn bản lớn. Phương pháp này cho phép người dùng cung cấp các đầu vào bổ sung, như các bản đồ ngữ nghĩa, ảnh biên cảnh, để kiểm soát quá trình sinh ảnh một cách ổn định hơn. ControlNet [16] hoạt động bằng cách thêm một mạng con vào cấu trúc của mô hình tạo sinh gốc.

## Chương 3

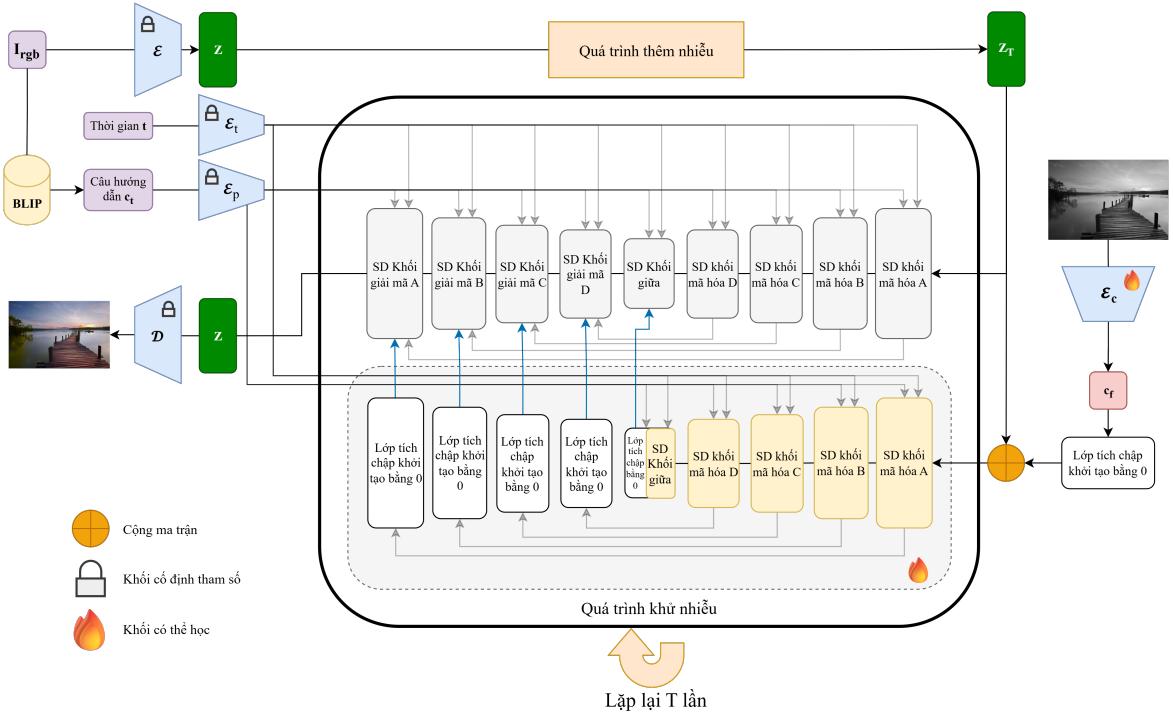
# Phương pháp đề xuất

Mục này có cách viết đơn giản nhất là sẽ trình bày lại gần như đầy đủ các nội dung của phương pháp được chọn, nhưng những phần hạn chế sẽ được thay thế bằng các cải tiến do nhóm đề xuất. Cụ thể:

- Vẽ lại sơ đồ phương pháp đề xuất với các thành phần cải tiến được đề xuất.
- Đối với cải tiến khối **Conv3D**, báo cáo này không trình bày lại, mà trình bày thẳng về khối **Pseudo 3D**. Bên cạnh đó, bổ sung các so sánh lý thuyết giữa kỹ thuật đề xuất so với kỹ thuật gốc ( khác biệt về nguyên lý, độ phức tạp thuật toán, chi phí huấn luyện, tổng tham số của mô hình lớn,... )
- Tương tự với **Circle Loss**.
- Tất cả các hình vẽ về các kỹ thuật hay phương pháp đề xuất, phải được chèn vào.
- Nội dung có sẵn khá lan man, không nên viết theo.

### 3.1 Tổng quan về mô hình

Hình 3.1 thể hiện quy trình huấn luyện tổng quát của mô hình, quy trình này bao gồm 2 giai đoạn:



Hình 3.1: Tổng quan về mô hình tô màu ảnh độ xám dựa trên mô hình tổng hợp ảnh có điều kiện sử dụng quy trình khuếch tán.

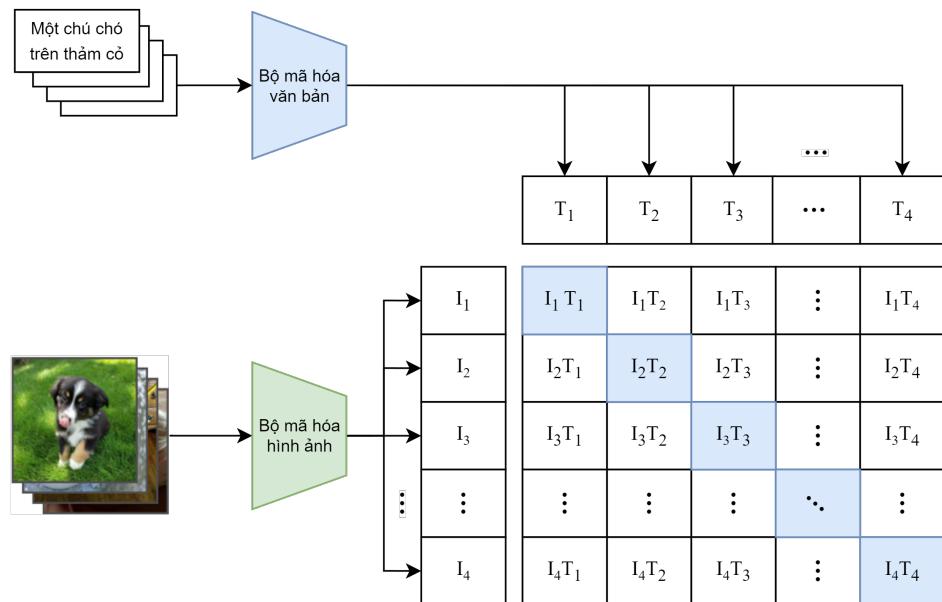
- Giai đoạn 1:** Thiết lập tín hiệu điều khiển thông qua câu hướng dẫn và ảnh độ xám đầu vào. Mô hình sử dụng bộ mã hóa văn bản  $\mathcal{E}_p$  của CLIP [17] để mã hóa các lời nhắc văn bản được tạo ra bởi mô hình BLIP [18]. Mô hình tích hợp một mạng mã hóa điều kiện  $\mathcal{E}_c$  để trích xuất đặc trưng từ ảnh điều kiện trước khi đưa vào ControlNet [16].
- Giai đoạn 2:** Quá trình khuếch tán có kết hợp điều kiện không gian. Ảnh màu mục tiêu được mã hóa vào không gian tiềm ẩn bằng bộ mã

hóa  $\mathcal{E}$  và được thêm các nhiễu Gauss qua  $T$  bước, sau đó được đưa vào cả 2 khối ControlNet [16] và Stable Diffusion [9] cho quá trình khử nhiễu có điều kiện.

## 3.2 Thiết lập tín hiệu điều khiển

### 3.2.1 Bộ mã hóa văn bản của CLIP

CLIP là một mô hình ngôn ngữ - thị giác được nghiên cứu bởi Radford và các đồng nghiệp [17], được huấn luyện trên hàng triệu cặp dữ liệu hình ảnh và văn bản. Ý tưởng chính của CLIP [17] là biểu diễn cả hai thành phần văn bản và hình ảnh vào cùng một không gian vec-tơ. Như được minh họa tại hình 3.2, mô hình CLIP [17] được xây dựng bao gồm 2 thành phần chính: một bộ mã hóa văn bản và một bộ mã hóa hình ảnh.

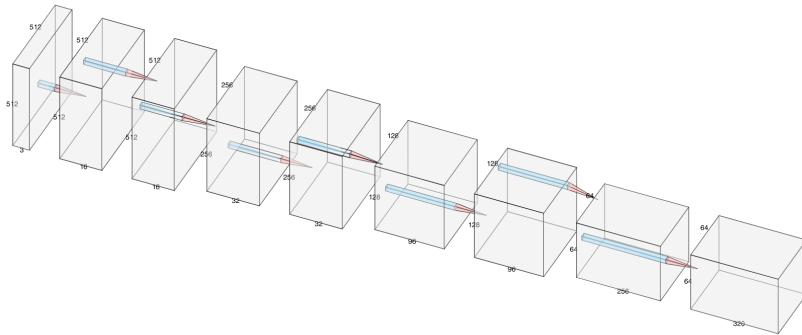


Hình 3.2: Kiến trúc của mô hình CLIP.

Với tác vụ tô màu ảnh độ xám có hướng dẫn văn bản, mô hình sử dụng bộ mã hóa văn bản của CLIP [17] để mã hóa các lời nhắc văn bản thành các vec-tơ, sau đó đưa vào mô hình thông qua cơ chế chú ý chéo.

### 3.2.2 Bộ mã hóa điều kiện

Để mã hóa điều kiện kiểm soát không gian vào không gian tiềm ẩn, mô hình sử dụng một mạng tích chập nhỏ được huấn luyện chung với quá trình đào tạo mô hình chính. Cấu trúc của mạng mã hóa điều kiện  $\mathcal{E}_c$  được thể hiện trong hình 3.3.



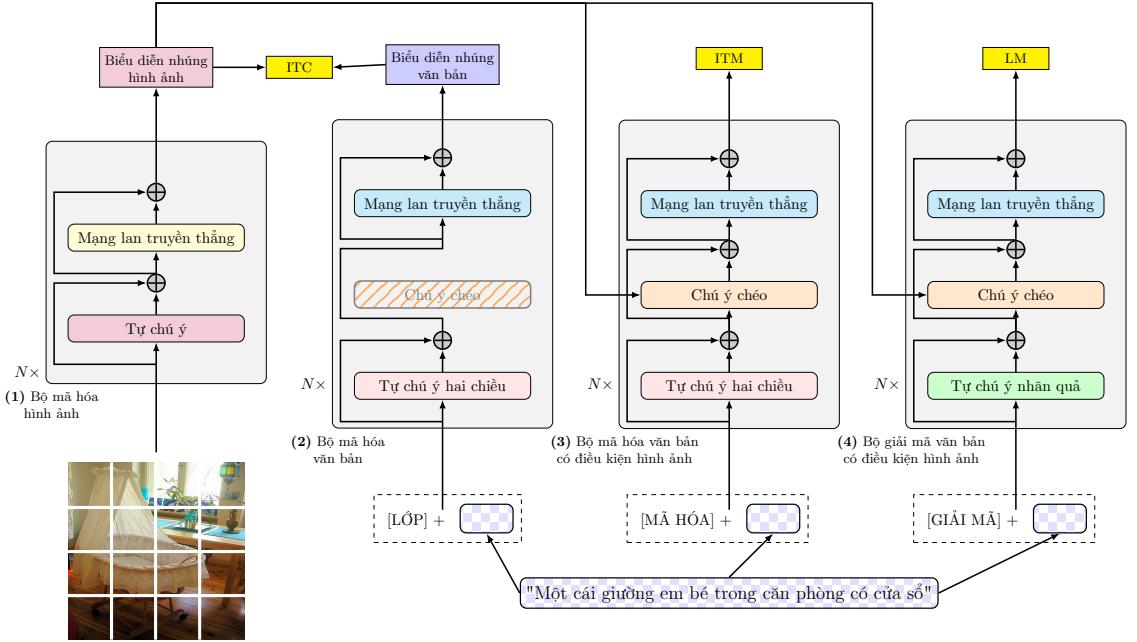
Hình 3.3: Cấu trúc mạng mã hóa điều kiện  $\mathcal{E}_c$ .

### 3.2.3 Mô hình BLIP

BLIP [18] là một mô hình ngôn ngữ - ảnh tiền huấn luyện đa nhiệm. BLIP giải quyết hai hạn chế lớn với hai đề xuất:

- Hỗn hợp các cặp bao gồm bộ mã hóa và giải mã đa phương thức (MED): một kiến trúc có khả năng hỗ trợ quá trình tiền huấn luyện đa nhiệm.
- Chú thích và lọc: một kỹ thuật tạo dữ liệu mới nhằm tăng cường khả năng học từ các cặp dữ liệu hình ảnh - văn bản nhiều.

Trong mô hình, nhóm sử dụng mô hình BLIP [18] để tạo câu mô tả cho ảnh đầu vào, dùng như lời nhắc văn bản cho bộ mã hóa văn bản của CLIP [17].



Hình 3.4: Kiến trúc của hồn hợp các cặp bộ mã hóa-giải mã đa phương thức MED.

### 3.3 Mô hình tạo sinh ảnh có kết hợp điều kiện không gian

Mô hình được đề xuất dựa trên mô hình khuếch tán có kết hợp điều kiện được phát triển qua các giai đoạn:

- Sự ra đời và phát triển của mô hình khuếch tán trong lĩnh vực tạo sinh ảnh.
- Quá trình điều chỉnh nhằm giảm độ phức tạp tính toán bằng cách đưa các tài nguyên huấn luyện vào không gian tiềm ẩn [9].
- ControlNet [16] ra đời để kết hợp các điều kiện kiểm soát không gian vào các mô hình khuếch tán đã được huấn luyện sẵn.

### 3.3.1 Mô hình khuếch tán

Mô hình khuếch tán là một lớp mô hình sinh dữ liệu, được thiết kế để mô phỏng các phân phối dữ liệu phức tạp thông qua việc đảo ngược quá trình khuếch tán. Lấy cảm hứng từ vật lý thống kê phi cân bằng [19], các mô hình này định nghĩa quá trình sinh dữ liệu như một chuỗi khử nhiễu dần dần.

Quá trình khuếch tán tiến được định nghĩa như sau:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (3.1)$$

Trong đó  $\beta_t$  là lịch trình nhiễu tại bước  $t$ . Mô hình được huấn luyện để dự đoán nhiễu  $\epsilon$ :

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2] \quad (3.2)$$

### 3.3.2 Mô hình khuếch tán trong không gian tiềm ẩn

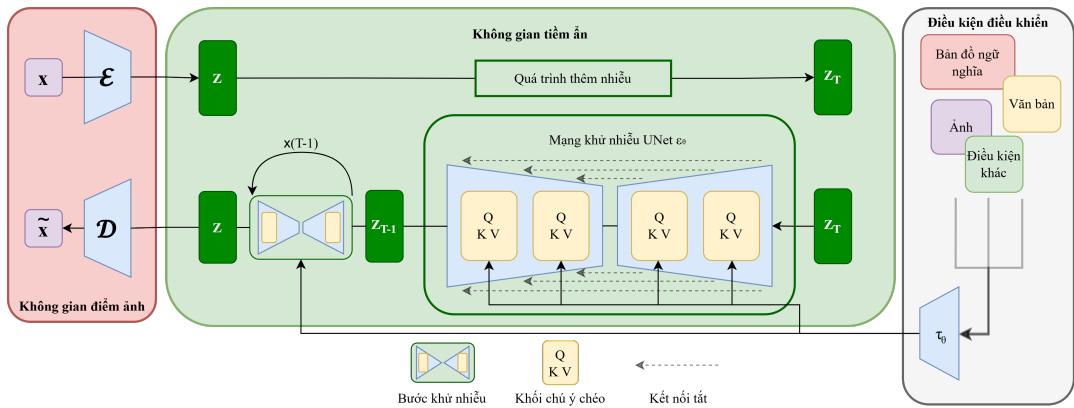
Mô hình khuếch tán trong không gian tiềm ẩn [9] thực hiện quá trình khuếch tán trong không gian tiềm ẩn thay vì không gian điểm ảnh. Quá trình này bao gồm:

1. Mã hóa ảnh gốc  $\mathbf{x}$  vào không gian tiềm ẩn:  $\mathbf{z} = \mathcal{E}(\mathbf{x})$
2. Thực hiện khuếch tán trong không gian tiềm ẩn
3. Giải mã từ không gian tiềm ẩn:  $\mathbf{x} = \mathcal{D}(\mathbf{z})$

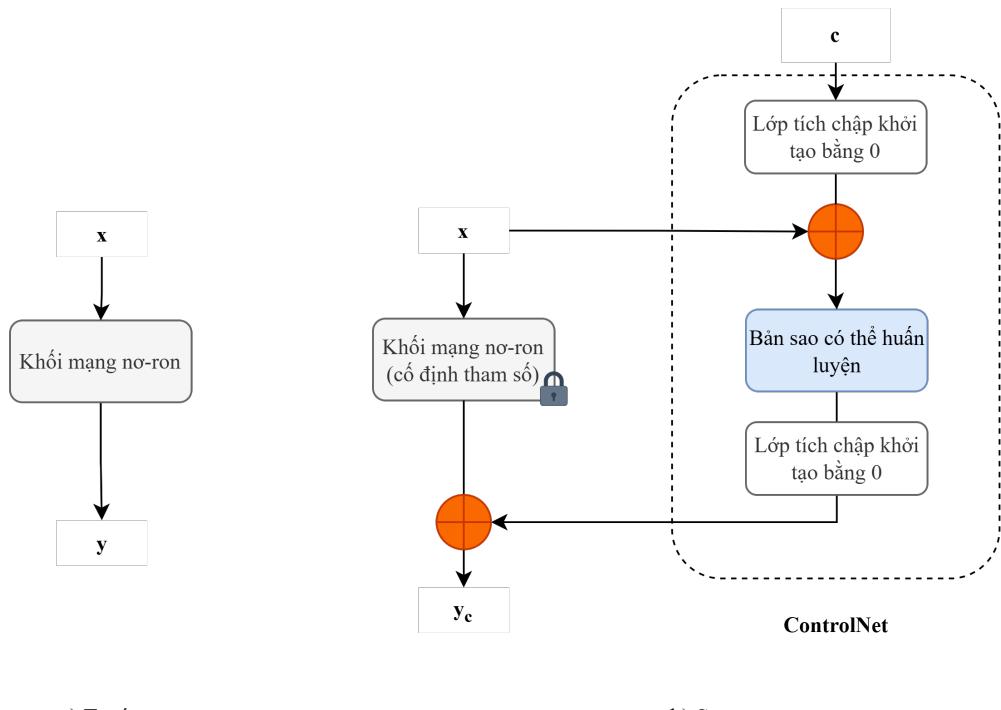
Hình 3.5 minh họa kiến trúc của mô hình LDM.

### 3.3.3 ControlNet

ControlNet [16] cho phép tích hợp các điều kiện không gian vào các mô hình khuếch tán đã được tiền huấn luyện. Kiến trúc của ControlNet được minh họa trong hình 3.6.



Hình 3.5: Kiến trúc của mô hình khuếch tán trong không gian tiềm ẩn (LDM).



Hình 3.6: Kiến trúc của ControlNet.

ControlNet hoạt động bằng cách tạo một bản sao có thể huấn luyện của các khối encoder và middle block của U-Net, trong khi giữ nguyên các tham số của mô hình gốc. Các lớp tích chập khởi tạo bằng không đảm bảo rằng ở giai đoạn đầu của quá trình huấn luyện, ControlNet không ảnh hưởng đến mô hình gốc.

## Chương 4

# Thực nghiệm

Mục này tương đối đơn giản, chỉ cần liệt kê về tập dữ liệu, cài đặt chi tiết cho thực nghiệm, kế hoạch thực nghiệm, kết quả thực nghiệm so với nghiên cứu gốc (baseline).

Ở mục phân tích kết quả, ngoài việc phân tích về sự cải thiện trên kết quả, cần xem xét các khía cạnh khác như:

- Thời gian huấn luyện ra mô hình cuối cùng (100K).
- Chi phí của quá trình con (ví dụ, 100 iterations mất bao lâu).
- Sự biến thiên của hàm mất mát, độ chính xác (dựa trên log, vẽ biểu đồ đường thì càng tốt).
- Phải phân tích các khía cạnh mà mô hình đề xuất bị kém đi, nguyên nhân nào gây ra (do cấu hình chưa tối ưu, hay đó là điều không thể tránh khỏi khi dùng kỹ thuật đó, hay do dữ liệu). Đề xuất một số cách khắc phục các hạn chế đó (dành cho phần hướng nghiên cứu tương lai).

### 4.1 Mục tiêu thực nghiệm

Nhóm tiến hành huấn luyện để tạo ra 4 mô hình cho các mục tiêu khác nhau. Một mô hình được huấn luyện để tô màu cho các trường hợp tổng

quát, 3 mô hình còn lại được huấn luyện trên các tập dữ liệu chuyên biệt.

Sau đó nhóm sử dụng các độ đo FID và Colorfulness để so sánh trong tác vụ tô màu tự động và sử dụng thêm CLIP score để so sánh trong tác vụ tô màu bán tự động.

## 4.2 Chi tiết quá trình huấn luyện

### 4.2.1 Tập dữ liệu huấn luyện

Nhóm huấn luyện các mô hình trên 4 tập dữ liệu:

- **ImageNet100k:** Tập dữ liệu do nhóm tạo ra bằng cách chọn ngẫu nhiên 100 nghìn ảnh từ tập dữ liệu huấn luyện ImageNet [20]. Tất cả ảnh được điều chỉnh về kích thước  $512 \times 512$ .
- **Fashion Product Images:** Tập dữ liệu bao gồm 44,441 ảnh có kích thước chung  $1800 \times 2400$ .
- **Furniture Image:** bao gồm 15,000 ảnh của 5 loại nội thất riêng biệt.
- **VN-celeb:** Bộ dữ liệu bao gồm 23105 khuôn mặt của 1020 người.

### 4.2.2 Quá trình tiền xử lý dữ liệu

Quy trình tiền xử lý dữ liệu gồm 3 bước:

1. **Lọc ảnh độ xám:** Lọc ra các ảnh độ xám, chỉ giữ lại ảnh màu. Ảnh được xem là ảnh độ xám khi:

$$E(Var(C_i, C_j)) = \frac{1}{3} \sum_{(i,j) \in \{(R,G), (G,B), (B,R)\}} Var(C_i - C_j) < t \quad (4.1)$$

Trong đó  $t = 12$  là ngưỡng độ xám.

2. **Thay đổi độ phân giải:** Tất cả ảnh được điều chỉnh về kích thước  $512 \times 512$ .
3. **Tạo sinh câu mô tả:** Sử dụng mô hình BLIP[18] để tạo sinh câu mô tả cho ảnh màu.

Bảng 4.1 thể hiện số lượng ảnh của 4 tập dữ liệu sau khi tiền xử lý.

Tập dữ liệu	ImageNet100k	Fashion Product	Furniture Image	VN-celeb
Số lượng ảnh	97,590	38,284	13,525	22,284

Bảng 4.1: Thống kê số lượng ảnh theo từng tập huấn luyện sau khi tiền xử lý.

### 4.2.3 Kết quả quá trình huấn luyện

Nhóm trình bày kết quả trực quan của quá trình huấn luyện. Khi sử dụng các lớp tích chập khởi tạo bằng không, ở giai đoạn đầu, các ảnh màu được sinh ra ngẫu nhiên không giống với ảnh độ xám điều kiện. Sau một thời gian học, ở khoảng bước thứ 3100 trở đi, ảnh màu được sinh ra bắt đầu giống với ảnh điều kiện, đây là hiện tượng **hội tụ đột ngột** được giới thiệu bởi Zhang và các đồng nghiệp [16].



Hình 4.1: Chất lượng ảnh được tạo sinh trong quá trình huấn luyện. Ở lân cận bước thứ 3200, hiện tượng **hởi tụ đột ngột** xảy ra.

## 4.3 Thử nghiệm mô hình tô màu tổng quát

### 4.3.1 Tập dữ liệu thử nghiệm

Để đánh giá độ hiệu quả của mô hình, nhóm sử dụng 3 tập dữ liệu đánh giá:

- **Tập val5k:** chứa 5000 ảnh từ bộ dữ liệu đánh giá của ImageNet [20].

- **Tập ctest:** tập 10000 ảnh được đề xuất bởi Larson và các đồng nghiệp [21].
- **Tập COCO-Stuff:** với 5000 ảnh từ bộ dữ liệu COCO-Stuff [22].

### 4.3.2 Kết quả thử nghiệm

Kết quả thực nghiệm đối với các mô hình tô màu tự động được thể hiện trong bảng 4.2. Các điểm số được in đậm thể hiện giá trị tốt nhất, các chỉ số được gạch chân thể hiện chỉ số tốt thứ hai.

Tập dữ liệu	ImageNet (val5k)		ImageNet (ctest)		COCO-Stuff	
	Độ đo	FID↓	Colorfulness↑	FID↓	Colorfulness↑	FID↓
CIC [3]	22.0860	37.0313	12.7651	37.5761	33.3418	37.6487
UGColor [23]	15.1777	27.0966	6.5466	27.8122	21.4010	28.4487
DeOldify [5]	10.5191	26.4827	4.2143	23.1538	13.4318	28.3779
DDColor [24]	<b>5.5726</b>	<u>42.8370</u>	<b>2.6294</b>	42.9575	<b>7.2718</b>	42.2919
Chúng tôi	10.4125	<b>44.2871</b>	6.8341	<b>44.7854</b>	10.4632	<b>45.2589</b>

Bảng 4.2: Kết quả so sánh các mô hình tô màu tự động trên các tập dữ liệu đánh giá.

Kết quả cho thấy, mô hình của nhóm đã vượt qua hầu hết các phương pháp tô màu tiên tiến trước đó trên phương thức tô màu tự động về chỉ số Colorfulness.

## 4.4 Phân tích kết quả

Hình 4.2 minh họa một số kết quả tô màu của mô hình đề xuất.



(a) Ảnh đầu vào



(b) Kết quả tô màu



(c) Ảnh gốc

Hình 4.2: Ví dụ kết quả tô màu. (a) Ảnh độ xám đầu vào. (b) Ảnh được tô màu bởi mô hình đề xuất. (c) Ảnh màu gốc để so sánh.

## Chương 5

# Kết luận

Tổng kết lại đóng góp của nghiên cứu, các cải tiến, và hạn chế, kèm theo hướng nghiên tương lai (đã đề cập ở phần **Phân tích kết quả** của 4).

## 5.1 Kết luận

Trong đề tài này, nhóm đã nghiên cứu và phát triển một mô hình tô màu ảnh độ xám dựa trên mô hình khuếch tán. Nguyên lý thêm nhiễu và khử nhiễu của mô hình khuếch tán, kèm thêm khả năng tùy chỉnh các tham số điều khiển giúp cho mô hình có thể tạo sinh đầu ra đa dạng. Mô hình tiền huấn luyện đã học trên số lượng ảnh đủ lớn, nên tri thức về màu sắc của các đối tượng tự nhiên hay nhân tạo gần như đã được kết xuất.

Việc tinh chỉnh mô hình trên các tập dữ liệu chuyên biệt giúp cho các mô hình chuyên biệt tập trung hơn vào một miền dữ liệu cụ thể, tăng cường chất lượng ảnh được tô màu. Mô hình cũng có khả năng hoạt động trong hai chế độ chính: tô màu tự động hoàn toàn, và tô màu có điều kiện tuân theo hướng dẫn bằng lời nhắc văn bản.

Về mặt dữ liệu, nhóm đã thu thập, xử lý và chuẩn bị các tập dữ liệu phục vụ cho huấn luyện các mô hình tô màu chuyên biệt, bao gồm dữ liệu khuôn mặt, nội thất và thời trang.

## 5.2 Bàn luận

Mô hình tô màu do nhóm đề xuất đã đạt được kết quả khả quan khi tiến hành đánh giá thực nghiệm trên nhiều tập dữ liệu khác nhau. Tuy nhiên, mô hình cũng vẫn còn gặp một số thách thức như:

- Trong những trường hợp mô tả mơ hồ, thiếu ngữ cảnh kết quả vẫn có thể không đúng như mong đợi.
- Khả năng hiểu ngôn ngữ phụ thuộc mạnh vào chất lượng bộ mã hóa văn bản.
- Khi mô tả văn bản xung đột với đặc trưng hình ảnh, mô hình có thể bị lúng túng giữa việc tin ảnh hay tin văn bản.
- Các tập dữ liệu trên các miền chuyên biệt được chuẩn bị chưa quá tốt, dẫn đến kết quả đánh giá của các mô hình chuyên dụng chưa đạt được mong muốn thực tế.

Quá trình thêm điều kiện vào mô hình thông qua các bộ mã hóa đặc trưng và các cơ chế như chú ý chéo góp phần rất quan trọng vào việc cải thiện hiệu suất của mô hình huấn luyện. Nghiên cứu có thể được mở rộng bằng cách thay thế hoặc phát triển các thành phần này để có thể tăng khả năng tô màu của mô hình cũng như tích hợp thêm các điều khiển khác.

Kiến trúc kết hợp điều kiện vào các mô hình khuếch tán cho phép các sự thay đổi linh hoạt trong mô hình. Các hướng nghiên cứu tiếp theo có thể mở rộng sang việc tích hợp thêm các điều kiện không gian khác nhằm tạo ra các mô hình mới trong các lĩnh vực liên quan.

# Tài liệu tham khảo

## Tiếng Anh

- [1] Trex, E., “How (and why) are black and white films colorized,” *Mental Floss*, 2011. [Online]. Available: <https://www.mentalfloss.com/article/26956/how-and-why-are-black-and-white-films-colorized>
- [2] Levin, A., Lischinski, D., and Weiss, Y., “Colorization using optimization,” *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 689–694, 2004.
- [3] Zhang, R., Isola, P., and Efros, A. A., “Colorful image colorization,” *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–11, 2016.
- [4] Iizuka, S., Simo-Serra, E., and Ishikawa, H., “Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification,” *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–11, 2016.
- [5] Antic, J., *Deoldify: A deep learning based project for colorizing and restoring old images*, 2019. [Online]. Available: <https://github.com/jantic/DeOldify>
- [6] Kim, G., Kang, K., Kim, S., et al., “BigColor: Colorization using a generative color prior for natural images,” in *Computer Vision – ECCV 2022*, vol. 13667, Springer Nature Switzerland, 2022, pp. 350–366.

- [7] Vitoria, P., Raad, L., and Ballester, C., *ChromaGAN: Adversarial picture colorization with semantic class distribution*, 2020. arXiv: 1907.09837 [cs.CV].
- [8] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M., *Hierarchical text-conditional image generation with clip latents*, 2022. arXiv: 2204.06125.
- [9] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., *High-resolution image synthesis with latent diffusion models*, 2022. arXiv: 2112.10752 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2112.10752>
- [10] Ho, J., Jain, A., and Abbeel, P., *Denoising diffusion probabilistic models*, 2020. arXiv: 2006.11239 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2006.11239>
- [11] Saharia, C., Chan, W., Chang, H., et al., *Palette: Image-to-image diffusion models*, 2022. arXiv: 2111.05826.
- [12] Yun, J., Lee, S., Park, M., and Choo, J., *iColoriT: Towards propagating local hint to the right region in interactive colorization by leveraging vision transformer*, 2022. arXiv: 2207.06831.
- [13] Huang, Z., Zhao, N., and Liao, J., *UniColor: A unified framework for multi-modal colorization with transformer*, 2022. arXiv: 2209.11223.
- [14] He, M., Chen, D., Liao, J., Sander, P. V., and Yuan, L., *Deep exemplar-based colorization*, 2018. arXiv: 1807.06587.
- [15] Weng, S., Wu, H., Chang, Z., et al., *L-CoDe:language-based colorization using color-object decoupled conditions*, 2022.
- [16] Zhang, L., Rao, A., and Agrawala, M., “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3836–3847.

- [17] Radford, A. et al., *Learning transferable visual models from natural language supervision*, 2021. arXiv: 2103 . 00020 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2103.00020>
- [18] Li, J., Li, D., Xiong, C., and Hoi, S., *BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, 2022. arXiv: 2201 . 12086.
- [19] Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S., *Deep unsupervised learning using nonequilibrium thermodynamics*, 2015. arXiv: 1503 . 03585.
- [20] Russakovsky, O., Deng, J., Su, H., et al., *Imagenet large scale visual recognition challenge*, 2015. arXiv: 1409 . 0575.
- [21] Larsson, G., Maire, M., and Shakhnarovich, G., *Learning representations for automatic colorization*, 2017. arXiv: 1603 . 06668.
- [22] Caesar, H., Uijlings, J., and Ferrari, V., *COCO-Stuff: Thing and stuff classes in context*, 2018. arXiv: 1612 . 03716.
- [23] Zhang, R., Zhu, J.-Y., Isola, P., et al., *Real-time user-guided image colorization with learned deep priors*, 2017. arXiv: 1705 . 02999.
- [24] Kang, X., Yang, T., Ouyang, W., et al., *DDColor: Towards photo-realistic image colorization via dual decoders*, 2023. arXiv: 2212 . 11613.

## Phụ lục A

### Phụ lục

#### A.1 Bảng đổi chiếu thuật ngữ

Bảng A.1 cung cấp bảng đổi chiếu các thuật ngữ Việt-Anh:

Bảng A.1: Phụ lục đổi chiếu Việt Anh

Tiếng Việt	Tiếng Anh
Trí tuệ nhân tạo	Artificial Intelligence
Học sâu	Deep Learning
Mạng nơ-ron tích chập	Convolutional Neural Network
Mạng nơ-ron	Neural Network
Bộ mã hóa	Encoder
Bộ giải mã	Decoder
Chú ý	Attention
Chú ý chéo	Cross-Attention
Bộ biến đổi	Transformer
Mô hình khuếch tán	Diffusion Model

#### A.2 Mã nguồn mẫu

Bạn có thể chèn mã nguồn vào phụ lục:

```
1 def hello_world():
2     print("Hello, World!")
```

```
3     return True
4
5 if __name__ == "__main__":
6     hello_world()
```

Listing A.1: Ví dụ mã Python

## A.3 Thông tin bổ sung

Phụ lục có thể chứa các thông tin bổ sung như:

- Dữ liệu thực nghiệm chi tiết
- Mã nguồn đầy đủ
- Các bảng kết quả bổ sung
- Tài liệu tham khảo bổ sung