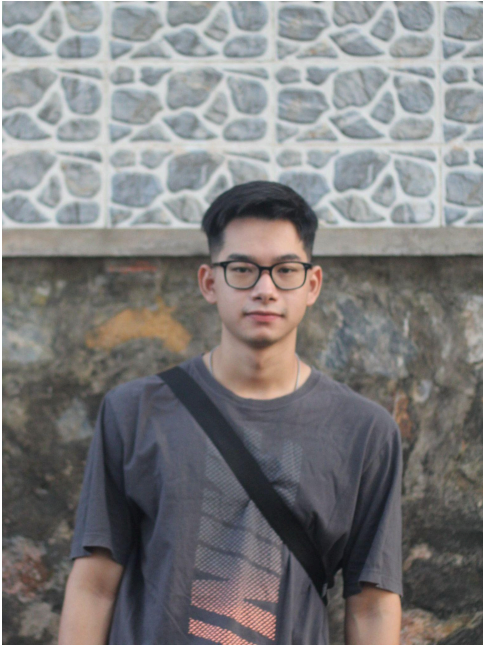


# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://youtu.be/x4w1PmM9kwY>
- Link slides (dạng .pdf đặt trên Github của nhóm):  
<https://github.com/ntien-281/cs519.p11/blob/main/Tiến%20Lý%20Văn%20Nhật%20-%20CS519.P11.DeCuong.FinalReport.Template.Slide.pdf>
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none"><li>● Họ và Tên: Lý Văn Nhật Tiến</li><li>● MSSV: 21521525</li></ul> 	<ul style="list-style-type: none"><li>● Lớp: <b>CS519.P11</b></li><li>● Tự đánh giá (điểm tổng kết môn): 8.5/10</li><li>● Số buổi vắng: 2</li><li>● Số câu hỏi QT cá nhân: 0</li><li>● Số câu hỏi QT của cả nhóm: 0</li><li>● Link Github: <a href="https://github.com/ntien-281/cs519.p11">https://github.com/ntien-281/cs519.p11</a></li><li>● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none"><li>○ Lên ý tưởng đề tài.</li><li>○ Nghiên cứu các công trình liên quan.</li><li>○ Tìm hiểu các phương pháp liên quan.</li><li>○ Viết đề cương, làm slide, poster.</li><li>○ Quay video thuyết trình trên youtube.</li></ul></li></ul>
--	--

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

ỨNG DỤNG SEGMENT ANYTHING MODEL CHO BÀI TOÁN ĐIỀN MÀU  
ẢNH XÁM BA GIAI ĐOẠN CÓ NHẬN THỨC TRÊN VÙNG

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

APPLYING SEGMENT ANYTHING MODEL FOR THREE-STAGE  
SEGMENT-BASED GRAY-SCALE IMAGE COLORIZATION

## TÓM TẮT (Tối đa 400 từ)

Điền màu ảnh xám (*grayscale image colorization*) là một bài toán mang nhiều giá trị ứng dụng cho lịch sử và tri thức, giúp làm sống động các tấm hình đen trắng từ lịch sử, ảnh chụp vệ tinh hay ảnh y khoa,...

Với sự bùng nổ của các mô hình học sâu, đặc biệt là trong lĩnh vực thị giác máy tính, nhiều công trình nghiên cứu đã ra đời nhằm giải quyết bài toán này. Đạt được hiệu quả điền màu chân thực và đồng nhất, các nghiên cứu trên vẫn còn vài điểm yếu nhất định liên quan đến việc hiểu ngữ cảnh cục bộ (*local context*) và ngữ cảnh liên vùng (*cross segment context*) của bức hình.

Nhận thấy rằng các vùng (*segment*) trong một bức hình có thể mang thông tin quan trọng trong việc điền màu cho các vật thể ở trong vùng đó, chúng tôi đề xuất hướng tiếp cận điền màu theo phân vùng ảnh (*image segment*) sử dụng mô hình phân vùng *SAM* (*Segment Anything Model*) [1] kết hợp các phương pháp điền màu

State-of-the-Art hiện nay.

Nghiên cứu này bao gồm:

- (1) Xây dựng một kiến trúc điền màu theo đường dẫn (*pipeline*) gồm 3 bước chính: phân vùng ảnh bằng *SAM*, điền màu cho các vùng bằng mô hình khuếch tán và điều chỉnh màu của bức hình sau khi ghép các vùng lại.
- (2) Sử dụng một bộ dữ liệu về điền màu ảnh xám làm chuẩn (*benchmark*) để đánh giá độ hiệu quả của kiến trúc được đề xuất và so sánh với các phương pháp SOTA hiện

nay.

Nghiên cứu này đặt mục tiêu tăng hiệu quả mô hình điền màu ảnh xám. Từ đó, tăng tính ứng dụng vào thực tiễn.

## **GIỚI THIỆU** (Tối đa 1 trang A4)

Điền màu ảnh xám (*grayscale image colorization*) là bài toán hồi quy mà mô hình giải quyết bài toán này cần phải dự đoán được các giá trị trong 3 kênh màu (LAB hoặc RGB) của bức ảnh vốn chỉ có 1 kênh (đen trắng).

Mô tả bài toán:

- **Input:** Ảnh xám (1 kênh màu) kích thước  $a \times b$ .
- **Output:** Bức ảnh đã được điền màu (3 kênh màu) kích thước  $a \times b \times 3$ .

Bên cạnh tính ứng dụng cao như phục hồi ảnh, nén ảnh, cải thiện ảnh y khoa... Bài toán cũng đặt ra các thách thức đặc biệt khó khăn như việc hiểu ngữ cảnh của bức hình (*culture & domain context*) và tính mơ hồ của màu (*color ambiguity*).

Trong quá trình nghiên cứu các phương pháp hiện có, chúng tôi nhận thấy phần lớn tập trung vào:

**(a)** Cải tiến việc điền màu trong 1 lần duy nhất (*single-pass colorization*) bằng các kỹ thuật khác nhau như mạng đối kháng tạo sinh (*Generative Adversarial Network*) và *Transformer* trong *ColorFormer* [2], hay mô hình khuếch tán (*Diffusion*) trong *Palette* [3].

**(b)** Điền màu cho các vật thể trong bức ảnh (*object-based colorization*) như nghiên cứu *Instance-aware Image Colorization* [4].

Các nghiên cứu trên đã giải quyết được vài thách thức chính của bài toán nhờ các mô hình hiện đại. Tuy nhiên, tồn đọng một vài vấn đề cần được giải quyết. Với kỹ thuật khuếch tán, tuy tổng thể bức hình được tô màu hài hòa, bản chất tạo sinh của mô hình đôi khi làm mượt quá mức (*oversmoothing*) khiến vật thể hòa tan (*blend*) vào vùng xung quanh. Trong khi đó, *Transformer* với kết quả điền màu tốt nhờ cơ chế *attention* trên ngữ cảnh là toàn bộ bức hình (*global context*) có thể cho kết quả thiếu chính xác, tiêu biểu là chuyển vùng đột ngột (*abrupt transition*), trên các chi tiết phụ thuộc vào

ngữ cảnh cục bộ (*local context*). Các vấn đề trên đặt ra yêu cầu mô hình phải cân bằng giữa ngữ cảnh toàn cục, cục bộ đồng thời đảm bảo giữ nguyên hiệu quả điền màu trên các chi tiết nhỏ. Với *Transformer*, hướng giải quyết tự nhiên là sử dụng cơ chế *attention* trên các patch ảnh nhỏ. Tuy nhiên, không phải patch nào cũng cần thông tin của các patch ở xa và đôi khi một vùng ảnh có liên quan lại nằm ở nhiều patch, khiến cho hướng tiếp cận này cực kỳ khó thực thi. Hơn nữa, việc chọn kích thước patch tối ưu sẽ rất nhọc nhằn, thường tốn kém tài nguyên tính toán khi huấn luyện và làm cho ứng dụng thực tiễn của bài toán trở nên xa vời. Điều này dẫn đến câu hỏi mà nghiên cứu này được đề xuất để trả lời:

***Làm thế nào để bổ sung ngữ cảnh cục bộ một cách hiệu quả?***

Dựa trên ý tưởng *object-based colorization* và sự tiến bộ của các mô hình phân vùng ảnh (tiêu biểu là *SAM*), chúng tôi đề xuất một hướng tiếp cận cân đối giữa hai hướng đã được giới thiệu: điền màu trên các vùng của bức ảnh sau đó hợp nhất các vùng đã tô màu này.

Đề xuất trên được định hướng nhằm sử dụng thông tin trong vùng và giữa các vùng để điền màu cho chúng, tháo gỡ vấn đề thiếu liên kết thông tin toàn cục dẫn đến kết quả điền màu thiếu sự nhất quán và hài hòa giữa và trong các vùng ảnh.

**MỤC TIÊU** (*Viết trong vòng 3 mục tiêu*)

**(1)** Xây dựng hệ thống điền màu cho ảnh gồm 3 thành phần:

- Thành phần phân chia ảnh (*segmentation module*): Sử dụng mô hình *SAM*.
- Thành phần điền màu cho các segment: Sử dụng mô hình khuếch tán.
- Thành phần hợp nhất các segment đã được điền màu thành bức ảnh hoàn chỉnh (*fusion module*): Sử dụng phép cộng cơ bản trên các ma trận kênh màu.

**(2)** Đánh giá được hệ thống đã xây dựng và so sánh với các nghiên cứu khác. Hiệu

năng hệ thống được đánh giá dựa trên các tiêu chí: thời gian suy luận, tiêu tốn tài nguyên tính toán. Độ chính xác của hệ thống được đánh giá trên các độ đo: SSIM

(*Structural Similarity Index*), FID (*Fréchet Inception Distance*). Bên cạnh đó, hệ

thống còn được đánh giá bởi con người, do tính mơ hồ trong việc chọn màu, sử dụng

bài kiểm tra ưa thích (*preference test*).

(3) Đánh giá được sự ảnh hưởng của cấu hình các module lên hiệu năng và độ chính xác của hệ thống (*Ablation study*). Đồng thời phát hiện các yếu điểm và đề xuất được hướng phát triển trong tương lai.

## **NỘI DUNG VÀ PHƯƠNG PHÁP**

(1) Tìm hiểu, lựa chọn mô hình và bộ dữ liệu cho module điền màu.

Phương pháp thực hiện:

- Yêu cầu của đề tài này đòi hỏi mô hình sử dụng trong module điền màu phải:  
(a) Có độ chính xác tương đối cao, (b) Có thời gian suy luận thấp, và (c) Có số lượng tham số đủ ít để tối ưu thời gian huấn luyện, suy luận.
- Trước khi được chọn, mỗi mô hình sẽ được đánh giá trên các tiêu chí trên và so sánh với nhau để tiếp tục nội dung tiếp theo với lựa chọn tối ưu nhất. Mô hình đề cử là *Stable Diffusion* [5].
- Với bộ dữ liệu, chúng tôi sẽ tìm hiểu và chọn lọc các bộ có số lượng lớn, ảnh có chất lượng cao và thích hợp. Thực hiện các bước tiền xử lý như cân bằng histogram, điều chỉnh kích thước ảnh, loại bỏ các ảnh mà chất lượng khi chuyển về đen trắng không được tốt,...

(2) Xây dựng module hợp nhất các segment đã được điền màu.

Phương pháp thực hiện:

- Vì tính chất khám phá hướng đi mới cho bài toán điền màu này của nghiên cứu, chúng tôi tập trung vào việc đánh giá tiềm năng và hiệu quả của hệ thống. Cho nên, module hợp nhất sẽ được cài đặt bằng các phương pháp/thuật toán đơn giản để tránh việc module này ảnh hưởng quá nhiều đến độ chính xác của hệ thống.

(3) Finetune mô hình *SAM* trên tập dữ liệu đã được chọn.

Phương pháp thực hiện:

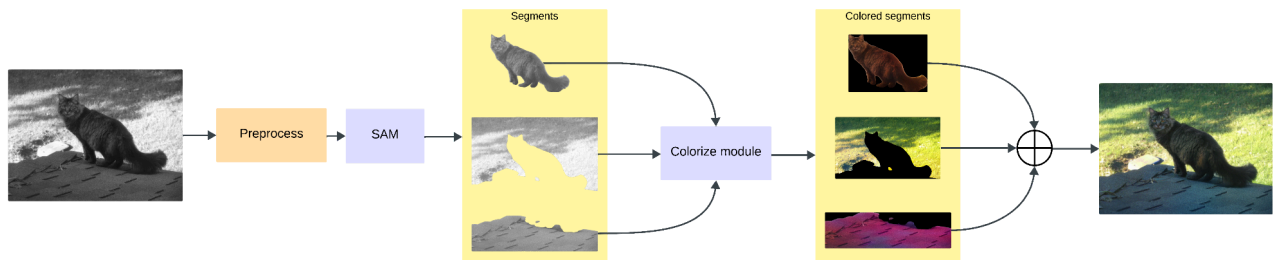
- Bước này nhằm giúp mô hình *SAM* đạt được hiệu quả cao trên ảnh đen trắng (những ảnh đầu vào).

- Để tối ưu quá trình này, chúng tôi đề xuất các phương pháp học chuyển tiếp (*Transfer learning*) như *Zero-shot*, *One-shot* và *Few-shot* với tập dữ liệu con trích từ tập đã được chọn.

#### (4) Xây dựng, huấn luyện và đánh giá hệ thống:

Phương pháp thực hiện:

- Kết nối 3 module trên thành một pipeline hoàn chỉnh.



- Ảnh đen trắng đầu vào sẽ được *SAM* phân vùng. Sau đó, mô hình điền màu sẽ dự đoán màu của các vùng này. Cuối cùng, các vùng đã có màu sẽ được tổng hợp lại thành bức ảnh hoàn chỉnh nhờ module hợp nhất.
- Đầu ra của pipeline sẽ được so với ảnh gốc có màu để tính lỗi và lan truyền ngược. Chúng tôi đề xuất sử dụng hàm lỗi kết hợp giữa lỗi trên điểm ảnh (*Pixel loss*) và lỗi nhận thức (*Perceptual loss*).

$$L = \frac{1}{N} \sum_{i=1}^N \|y_i - \hat{y}_i\|^2 + \sum_l \|\phi_l(y) - \phi_l(\hat{y})\|^2$$

- Hệ thống sẽ được đánh giá dựa trên các độ đo: thời gian suy luận, FID, SSIM và preference test. Kết quả trên sẽ được so sánh với kết quả của các phương pháp hiện có. Ngoài ra, chúng tôi dự định cung cấp thêm biểu đồ lỗi để đánh giá quá trình huấn luyện.

#### (5) Đánh giá sự phụ thuộc của hệ thống vào các module và đề xuất phát triển.

- Nghiên cứu này sẽ thực hiện điều chỉnh cấu hình của các module để đánh giá ảnh hưởng của các lựa chọn này lên hiệu quả của hệ thống.
- Các cấu hình tiêu biểu để thực hiện nội dung này bao gồm kích thước không gian ẩn (*latent space*) của mô hình khuếch tán, lựa chọn phương pháp hợp nhất

các vùng, lựa chọn phương pháp finetune *SAM*. Danh sách trên đã theo thứ tự ưu tiên và có thể bổ sung, thay đổi trong quá trình thực hiện.

## KẾT QUẢ MONG ĐỢI

- (1) Xây dựng, huấn luyện và đánh giá được một hệ thống ổn định.
- (2) Hệ thống đạt kết quả **ít nhất** là ngang ngửa với các phương pháp hiện có trên độ đo FDI và SSIM, điều này thể hiện tiềm năng phát triển của các nghiên cứu theo sau. Bên cạnh đó, thời gian suy luận phải thấp, hướng đến ứng dụng trong thực tiễn.
- (3) Đánh giá được sự phụ thuộc vào cấu hình module, đồng thời đóng gói mã nguồn phục vụ mục đích tham khảo trong nghiên cứu ứng dụng.

## TÀI LIỆU THAM KHẢO (Định dạng DBLP)

- [1]A. Kirillov *et al.*, “Segment Anything,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2023. Accessed: Dec. 29, 2024. [Online]. Available: <https://doi.org/10.1109/iccv51070.2023.00371>
- [2]H. Shafiq, T. Nguyen, and B. Lee, “Colorformer: A Novel Colorization Method Based on a Transformer,” Elsevier BV, 2024. Accessed: Jan. 01, 2025. [Online]. Available: <https://doi.org/10.2139/ssrn.4937072>
- [3]C. Saharia *et al.*, “Palette: Image-to-Image Diffusion Models,” in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, New York, NY, USA: ACM, Aug. 2022, pp. 1–10. Accessed: Jan. 01, 2025. [Online]. Available: <https://doi.org/10.1145/3528233.3530757>
- [4]J.-W. Su, H.-K. Chu, and J.-B. Huang, “Instance-Aware Image Colorization,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 7965–7974. Accessed: Jan. 01, 2025. [Online]. Available: <https://doi.org/10.1109/cvpr42600.2020.00799>
- [5]R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 10674–10685. Accessed: Jan. 12, 2025. [Online]. Available:

<https://doi.org/10.1109/cvpr52688.2022.01042>