THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): https://youtu.be/nhREb3KO-hs
- Link slides (dạng .pdf đặt trên Github của nhóm): https://github.com/dattt19uit/CS2205.FEB2025/*TranTanDat-240101040-CS2205.FEB2025.DeCuong.FinalReport.Doc*
- Thông tin học viên

- Họ và Tên: Trần Tấn Đạt
- MSHV: 240101040

- Lớp: CS2205.FEB2025
- Tự đánh giá (điểm tổng kết môn): 10/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 4
- Link Github: https://github.com/dattt19uit/CS2205.FEB2025

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

TĂNG CƯỜNG KHÔI PHỤC ẢNH BẰNG TÍCH HỢP SÂU MÔ HÌNH NGÔN NGỮ LỚN ĐA MÔ THỨC VÀO KIẾN TRÚC ONERESTORE.

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ENHANCING IMAGE RESTORATION THROUGH DEEP INTEGRATION OF MULTIMODAL LARGE LANGUAGE MODEL INTO ONERESTORE.

TÓM TẮT *(Tối đa 400 từ)*

Các mô hình dựa trên kiến trúc Transformer hiện nay đã chứng minh được tính hiệu quả vượt trội trong bài toán khôi phục ảnh - một bài toán được ứng dụng rộng rãi trong nhiều lĩnh vực như y tế, điều tra…, nhưng vẫn chưa tận dụng đủ ngữ cảnh ngữ nghĩa, dẫn đến hạn chế trong việc khái quát hóa trên các loại suy giảm phức tạp như ảnh vừa bị nhiễu, vừa bị mờ và vừa có vệt mưa. Tính đến thời điểm hiện tại, mặc dù đã có một số nghiên cứu sử dụng mô hình ngôn ngữ lớn LLM hoặc đa mô thức MLLM để tận dụng ngữ cảnh ngữ nghĩa, hỗ trợ điều phối và hướng dẫn các mô hình khôi phục ảnh, cho thấy hiệu quả nhất định trong các tác vụ như khử nhiễu, khử mờ, và khử mưa trên các loại suy giảm đa dạng; cách tiếp cận này làm cho MLLM và mô hình khôi phục ảnh hoạt động độc lập, gây ra độ phức tạp tính toán cao và phụ thuộc nhiều vào khả năng nhận diện suy giảm, khiến hiệu suất chưa tối ưu khi xử lý các loại suy giảm đa dạng. Với những vấn đề trên, liệu rằng có thể tối ưu hiệu suất khôi phục ảnh bằng cách tận dụng hai mô hình có thể hỗ trợ lẫn nhau, khai thác hiệu quả thông tin ngữ nghĩa từ MLLM vào kiến trúc phục hồi ảnh. Với động lực ấy, chúng tôi đề xuất tích hợp sâu MLLM vào pipeline khôi phục ảnh để hỗ trợ ngữ nghĩa, cải thiện chất lượng phục hồi ảnh trong các tình huống suy giảm phức tạp, đồng thời triển khai ứng dụng web minh chứng cho phép tải ảnh, khôi phục và đánh giá trực quan qua PSNR, SSIM (2 chỉ số đặc trưng trong việc so sánh ảnh đã khôi phục và ảnh gốc), mở rộng ứng dụng của MLLM trong xử lý ảnh.

GIỚI THIỆU *(Tối đa 1 trang A4)*

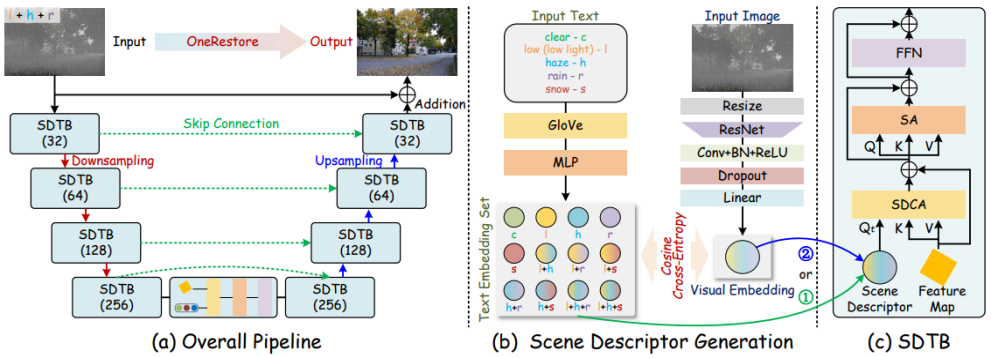
Khôi phục ảnh là một trong những bài toán then chốt trong lĩnh vực Thị giác máy tính, được ứng dụng rộng rãi trong các lĩnh vực như theo dõi đối tượng, y tế, bảo tồn di tích lịch sử và còn nhiều lĩnh vực khác [4, 5, 6].

Để giải quyết bài toán này, một số mô hình dựa trên kiến trúc Transformer, điển hình Restormer [1], với cơ chế attention đã xử lý hiệu quả các tác vụ khôi phục như khử nhiễu, khử mờ, tăng cường ánh sáng yếu; nhưng còn hạn chế trong việc xử lý tình huống suy giảm đa dạng như ảnh vừa nhiễu, vừa mờ và vệt mưa cùng lúc. Để khắc phục, OneRestore [3] sử dụng cơ chế cross-attention để tích hợp thông tin từ mô tả cảnh suy giảm dưới dạng nhãn tĩnh, còn RestoreAgent [2] sử dụng LLM như là tác nhân (agent), tạo ra các embedding động điều phối các mô hình hỗ trợ nhận diện các tình huống suy giảm và hướng dẫn khôi phục. Mặc dù RestoreAgent đạt hiệu suất tốt hơn so với OneRestore, phương pháp thực hiện điều phối bên ngoài, hay cả hai mô hình làm việc độc lập khiến cho chi phí tính toán cao và phụ thuộc vào khả năng nhận diện của MLLM, từ đó làm cho hiệu suất khôi phục ảnh không được đảm bảo.

Với câu hỏi nghiên cứu: **"Làm thế nào để cả 2 mô hình có thể hỗ trợ lẫn nhau, giảm chi phí tính toán và nâng cao hiệu suất khôi phục ảnh?"** Chúng tôi sẽ tận dụng mô hình OneRestore [6], nghiên cứu cơ chế cross-attention để trích xuất các embedding động từ MLLM và tích hợp sâu vào mô hình khôi phục, tạo ra ảnh được khôi phục với chất lượng cao từ đầu vào là ảnh suy giảm, cụ thể:

Input: Ảnh suy giảm (bị nhiễu, mờ, vệt mưa…)

Output: Ảnh được khôi phục chất lượng cao hơn



Hình 1: Hình mô tả pipeline OneRestore gốc khôi phục ảnh

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

1. Chuẩn bị tập dữ liệu thử nghiệm gồm ảnh suy giảm đơn và ảnh suy giảm phức tạp để phục vụ huấn luyện, kiểm định và đánh giá mô hình.
2. Tích hợp E5-V [7] vào pipeline OneRestore bằng cross-attention để tạo embedding động.
3. Huấn luyện lại (pretrain) mô hình tích hợp trên tập LSDIR, tinh chỉnh (finetune) trên tập validation và đánh giá đầu ra trên tập CDD-11 thông qua các chỉ số PSNR, SSIM.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung:

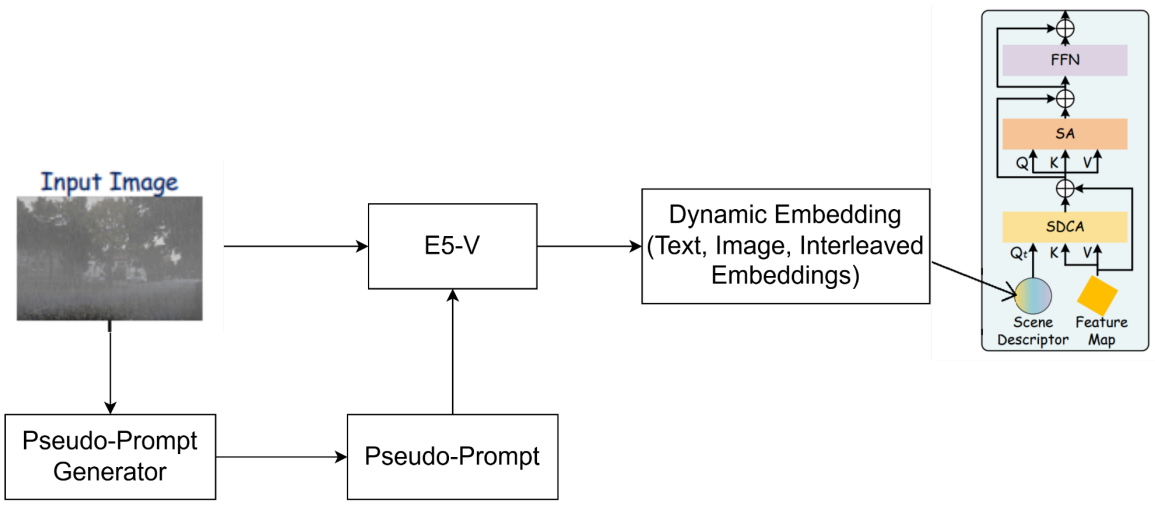
- Chuẩn bị và xử lý bộ dữ liệu gồm ảnh suy giảm đơn và suy giảm phức tạp, phục vụ cho các giai đoạn huấn luyện, kiểm định và đánh giá.
- Phân tích kiến trúc OneRestore, tập trung vào cơ chế tích hợp Scene Descriptor vào Transformer.
- Khảo sát mô hình MLLM E5-V và đề xuất phương án tích hợp vào pipeline OneRestore.
- Thiết kế kiến trúc mô hình mới kết hợp OneRestore với E5-V thay cho khối sinh Scene Descriptor truyền thống.
- Tiến hành huấn luyện lại mô hình tích hợp trên tập ảnh suy giảm đơn và tinh chỉnh trên tập validation.
- Đánh giá chất lượng khôi phục ảnh suy giảm phức tạp bằng các chỉ số PSNR và SSIM.

Phương pháp:

- Tải và xử lý dữ liệu từ Hugging Face:
    - Từ tập LSDIR, chọn 100 ảnh suy giảm đơn làm tập train (từ các shard) và 50 ảnh từ val.tar.gz làm tập validation.
    - Từ CDD-11, chọn 150 ảnh suy giảm phức tạp làm tập test.
    - Ánh xạ thủ công ảnh suy giảm với ground-truth và chuẩn hóa về kích

thước 256 x 256 bằng PIL hoặc OpenCV.

- Đề xuất tận dụng pipeline từ OneRestore, nhưng đầu vào là Input Image sẽ qua một cơ chế Pseudo-Prompt Generator, sau đó kết hợp rule-based prompt, ví dụ "This image has [number] degradations, include [detail degradation] in [location on picture having this degradation],…" trong đó Pseudo-Prompt Generator là một mạng nơ-ron đơn giản (Multi-Layer Perceptron - MLP) dự đoán các thành phần [number], [detail degradation] và [location on picture having this degradation] để tạo hoàn chỉnh Pseudo-Prompt. Cụ thể, MLP nhận Visual Embedding (đặc trưng từ ResNet) làm đầu vào, xử lý qua các lớp ẩn với hàm kích hoạt ReLU, và tạo đầu ra là các xác suất cho loại suy giảm (blur, noise, rain), mức độ (1-5), và vị trí (top-left, bottom-right, v.v.). Ví dụ, nếu MLP dự đoán ảnh có blur mức 3 và noise mức 2, Pseudo-Prompt có thể là "This image has 2 degradations, include blur level 3 in top-left, noise level 2 in bottom-right". Từ đó, Input Image + Pseudo-Prompt là 2 đầu vào vào MLLM E5-V để sinh ra Dynamic Embeddings, tạo ra ngữ nghĩa động cho Scene Descriptor.



Hình 2: Pipeline kết hợp mô hình E5-V từ Input Image và Pseudo-Prompt để tạo ra Dynamic Embeddings giàu ngữ nghĩa cho Scene Descriptor trong khối SDTB.

- Huấn luyện mô hình tích hợp trên tập train bằng các hàm mất mát L1 và

perceptual loss. Tinh chỉnh trên tập validation, sau đó đánh giá mô hình trên tập test bằng các chỉ số PSNR và SSIM, so với ground-truth.

KẾT QUẢ MONG ĐỢI

- Xây dựng được một mô hình khôi phục ảnh suy giảm phức tạp đã tích hợp mô hình ngôn ngữ lớn đa phương thức (E5-V) vào kiến trúc OneRestore thông qua cơ chế hướng dẫn bằng scene descriptor.
- Đánh giá thực nghiệm trên tập CDD-11 cho thấy mô hình tích hợp đã được huấn luyện và validation, đạt PSNR > 30, SSIM > 0.9, và khôi phục ảnh rõ nét hơn 20% so với OneRestore gốc (theo thang đo PSNR trung bình).
- Triển khai mô hình thành một ứng dụng demo chạy trên web với Streamlit, cho phép người dùng tải lên ảnh suy giảm đầu vào (blur, noise, haze, rain,...) và đầu ra là một ảnh chất lượng tốt hơn với mô hình đã được đề xuất, so sánh với ảnh ground-truth để tính các chỉ số PSNR và SSIM nếu người dùng cung cấp.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang: Restormer: Efficient Transformer for High-Resolution Image Restoration. CVPR 2022: 5718-5729

[2]. Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, Lei Zhu: RestoreAgent: Autonomous Image Restoration Agent via Multimodal Large Language Models. NeurIPS 2024

[3]. Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, Shengfeng He: OneRestore: A Universal Restoration Framework for Composite Degradation. ECCV (19) 2024: 255-272

[4]. Zhiwen Yang, Hui Zhang, Dan Zhao, Bingzheng Wei, Yan Xu: Restore-RWKV: Efficient and Effective Medical Image Restoration with RWKV. CoRR abs/2407.11087 (2024)

[5]. Yiran Li: Applications of Diffusion Model Image Restoration in the Field of Heritage Restoration: Overview and Outlook. CAIBDA 2023: 864-876

[6]. Tianyang Xu, Yifan Pan, Zhenhua Feng, Xuefeng Zhu, Chunyang Cheng, Xiao-Jun Wu, Josef Kittler: Learning Feature Restoration Transformer for Robust Dehazing Visual Object Tracking. Int. J. Comput. Vis. 132(12): 6021-6038 (2024)

[7]. Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, Fuzhen Zhuang: E5-V: Universal Embeddings with Multimodal Large Language Models. CoRR abs/2407.12580 (2024)

[8]. Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, Luc Van Gool: LSDIR: A Large Scale Dataset for Image Restoration. CVPR Workshops 2023: 1775-1787