ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

NGUYỄN QUỐC TRƯỜNG

KHÓA LUẬN TỐT NGHIỆP

LOẠI BỎ BÓNG ĐỔ TRÊN ẢNH TÀI LIỆU DỰA TRÊN VISION TRANSFORMERS

Document image shadow removal based on Vision Transformers

CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH, 2025

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

NGUYỄN QUỐC TRƯỜNG – 21521604

KHÓA LUẬN TỐT NGHIỆP

LOẠI BỎ BÓNG ĐỔ TRÊN ẢNH TÀI LIỆU DỰA TRÊN VISION TRANSFORMERS

Document image shadow removal based on Vision Transformers

CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH

GIẢNG VIÊN HƯỚNG DẪN

TS. NGUYỄN DUY KHÁNH

TP. HỒ CHÍ MINH, 2025

LỜI CAM ĐOAN

Tôi xin cam đoan:

Những nội dung trong luận văn này là do tôi thực hiện dưới sự hướng dẫn trực tiếp của TS. Nguyễn Duy Khánh.

Mọi tham khảo trong khóa luận đều được trích dẫn rõ ràng tên tác giả, tên công trình, và thời gian công bố.

Mọi sao chép không hợp lệ và vi phạm quy chế đào tạo tôi xin chịu hoàn toàn trách nhiệm.

Nguyễn Quốc Trường

THÔNG TIN HỘI ĐỒNG CHẤM KHÓA LUẬN TỐT NGHIỆP

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số …… ngày …….. của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

1. TS Nguyễn Tấn Trần Minh Khang − Chủ tịch.
2. ThS Nguyễn Thanh Sơn − Thư ký.
3. TS Lê Kim Hùng − Ủy viên.

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc Lập - Tự Do - Hạnh Phúc

TP. HCM, ngày…..tháng…..năm……

NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

CÁN BỘ HƯỚNG DẪN

Tên khóa luận:

LOẠI BỎ BÓNG ĐỔ TRÊN ẢNH TÀI LIỆU DỰA TRÊN VISION TRANSFORMERS

Nhóm SV thực hiện:		**Cán bộ hướng dẫn:**
Nguyễn Quốc Trường	21521604	TS. Nguyễn Duy Khánh

Đánh giá Khóa luận

1. Về cuốn báo cáo:

Số trang	_____	Số chương	_____
Số bảng số liệu	_____	Số hình vẽ	_____
Số tài liệu tham khảo	_____	Sản phẩm	_____

Một số nhận xét về hình thức cuốn báo cáo:

.....

.....

.....

.....

2. Về nội dung nghiên cứu:

.....

.....

.....

3. Về chương trình ứng dụng:

.....

.....

.....

4. Về thái độ làm việc của sinh viên:

.....

.....

.....

Đánh giá chung:

.....

.....

Điểm từng sinh viên:

Nguyễn Quốc Trường**:……../10**

Người nhận xét

(Ký tên và ghi rõ họ tên)

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc Lập - Tự Do - Hạnh Phúc

TP. HCM, ngày…..tháng…..năm……

NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

CÁN BỘ PHẢN BIỆN

Tên khóa luận:

LOẠI BỎ BÓNG ĐỔ TRÊN ẢNH TÀI LIỆU DỰA TRÊN VISION TRANSFORMERS

Nhóm SV thực hiện:		**Cán bộ phản biện:**
Nguyễn Quốc Trường	21521604	TS. Lê Kim Hùng

Đánh giá Khóa luận

5. Về cuốn báo cáo:

Số trang	_____	Số chương	_____
Số bảng số liệu	_____	Số hình vẽ	_____
Số tài liệu tham khảo	_____	Sản phẩm	_____

Một số nhận xét về hình thức cuốn báo cáo:

.....

.....

.....

.....

6. Về nội dung nghiên cứu:

.....

.....

.....

7. Về chương trình ứng dụng:

.....

.....

.....

8. Về thái độ làm việc của sinh viên:

.....

.....

.....

Đánh giá chung:

.....

.....

Điểm từng sinh viên:

Nguyễn Quốc Trường**:……../10**

Người nhận xét

(Ký tên và ghi rõ họ tên)

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc Lập - Tự Do - Hạnh Phúc

ĐỀ CƯƠNG CHI TIẾT

TÊN ĐỀ TÀI TIẾNG VIỆT: LOẠI BỎ BÓNG ĐỔ TRÊN ẢNH TÀI LIỆU DỰA TRÊN VISION TRANSFORMERS

TÊN ĐỀ TÀI TIẾNG ANH: DOCUMENT IMAGE SHADOW REMOVAL BASED ON VISION TRANSFORMERS

Cán bộ hướng dẫn: Tiến sĩ Nguyễn Duy Khánh, Trường đại học Công nghệ thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh

Thời gian thực hiện: Từ ngày 02/09/2024……..đến ngày 01/01/2025……

Sinh viên thực hiện:

Nguyễn Quốc Trường – 21521604

Hệ đào tạo: Tài năng

Nội dung đề tài:

Tổng quan đề tài: Đề tài "Document Image Shadow Removal based on Vision Transformers" (Loại bỏ bóng đổ trên ảnh tài liệu dựa trên Vision Transformers) tập trung vào việc cải thiện chất lượng hình ảnh tài liệu bị ảnh hưởng bởi bóng đổ. Hiện nay, việc chụp ảnh tài liệu bằng thiết bị di động đang ngày càng phổ biến. Tuy nhiên, những yếu tố như ánh sáng không đều, bóng đổ do tay người, vật thể hoặc thiết bị có thể làm giảm chất lượng hình ảnh, ảnh hưởng đến khả năng nhận diện ký tự và xử lý thông tin.

Đầu vào của bài toán: một hình ảnh tài liệu có 1 phần tài liệu bị che bởi bóng đổ từ các vật thể, khiến phần tài liệu bị che khuất có chất lượng thấp, khó đọc hoặc không sử dụng để thực hiện các tác vụ khác được.

Đầu ra của bài toán: một hình ảnh tài liệu sau khi được loại bỏ phần bóng đổ che khuất

tài liệu và được cải thiện chất lượng hình ảnh giúp dễ dàng đọc và thực hiện các tác vụ khác trên ảnh.

Trước đây, các phương pháp loại bỏ bóng đổ trên ảnh tài liệu chủ yếu dựa vào các kỹ thuật xử lý ảnh truyền thống như phân tích histogram, lọc không gian, hay các mô hình học sâu CNN. Tuy nhiên, các phương pháp này vẫn gặp hạn chế trong việc xử lý bóng đổ phức tạp, độ sáng không đồng đều hoặc các nhiễu ảnh khác.

Hiện nay, ngoài các phương pháp truyền thống đó, đã dần dần xuất hiện các phương pháp và nghiên cứu nhằm cải thiện chất lượng và hiệu quả. Tiêu biểu trong số đó là Zhang et al. [1], phương pháp được đề xuất trong bài báo này (2 mạng CBENet và BGShadowNet) có thể xem là phương pháp SOTA (state of the art) hiện giờ, thể hiện hiệu suất cao hơn so với các phương pháp trước đó. Trong các phương pháp loại bỏ bóng đổ trên ảnh tài liệu hiện nay, việc thực hiện đều thông qua 2 quá trình cơ bản, là loại bỏ bóng khỏi ảnh và cải thiện lại chất lượng ảnh. Tuy nhiên, nhằm làm tăng hiệu quả khi thực hiện các quá trình này, phương pháp trong bài báo nói trên thực hiện thêm 1 quá trình nữa là trích xuất nền ảnh dựa trên nhận biết màu sắc, từ đó sử dụng nền ảnh này như là 1 đặc trưng bổ sung cho 2 quá trình còn lại. Các mô hình và mạng neural trong phương pháp này sử dụng mới chỉ là những mô hình thuộc dạng cơ bản như mạng U-net hay các mạng convolution, do đó, phương pháp này có tiềm năng cải tiến nhờ việc sử dụng các mô hình, các công nghệ mới.

Vision Transformers (ViTs) [2] là một kỹ thuật mới trong lĩnh vực thị giác máy tính, đã chứng minh hiệu quả trong việc nhận diện và phân loại hình ảnh một cách chính xác hơn so với các mô hình trước đó. Việc áp dụng ViTs vào bài toán loại bỏ bóng trên ảnh tài liệu có thể mang lại những cải tiến đáng kể về độ chính xác và hiệu suất.

Mục tiêu của đề tài: Mục tiêu của đề tài này là phát triển một mô hình loại bỏ bóng đổ trên ảnh tài liệu dựa trên Vision Transformers, từ đó cải thiện chất lượng hình ảnh đầu ra. Đóng góp vào nghiên cứu và phát triển công nghệ, mở ra nhiều hướng nghiên cứu mới và cơ hội ứng dụng khác trong tương lai. So với các phương pháp truyền thống, mô hình này sẽ:

Cải thiện độ chính xác của các hệ thống nhận diện ký tự quang học nhờ chất lượng ảnh

tốt hơn.

Giảm thiểu sự phụ thuộc vào điều kiện ánh sáng trong quá trình chụp ảnh tài liệu.

Phương pháp thực hiện: Đề tài sẽ sử dụng Vision Transformers (ViTs) làm mô hình chính để xử lý ảnh tài liệu. Các bước thực hiện bao gồm:

- Thu thập dữ liệu: Sử dụng bộ dữ liệu gồm các ảnh tài liệu có bóng đổ, ảnh thể hiện nền ảnh (có màu) và các ảnh đã được làm sạch (bộ dữ liệu RDD [1]).

- Xây dựng mô hình: Sử dụng 2 mạng CBENet và BGShadowNet làm xương sống cho mô hình. Xây dựng và điều chỉnh mô hình dựa trên Vision Transformers.

Huấn luyện mô hình Vision Transformers dựa trên bộ dữ liệu thu thập được.

- Đánh giá: Thực hiện đánh giá mô hình thông qua các chỉ số như RMSE (Root Mean Squared Error), PSNR (Peak Signal-to-Noise Ratio) và SSIM (Structural Similarity Index).

Các nội dung chính và giới hạn của đề tài:

Nội dung chính:

Tìm hiểu, nghiên cứu cơ sở lý thuyết về các phương pháp loại bỏ bóng đổ trên ảnh và ảnh tài liệu trước đây. Từ đó xây dựng và cải tiến các phương pháp đó để thu được kết quả có độ chính xác cao hơn. Xem xét qua nhiều bộ dữ liệu và chọn ra bộ dữ liệu phù hợp để sử dụng cho mô hình, giúp huấn luyện mô hình đạt được tốt nhất có thể.

Xây dựng hệ thống loại bỏ bóng đổ trên ảnh tài liệu bằng cách sử dụng Vision Transformers. Việc xây dựng hệ thống trong đề tài này sử dụng mô hình và phương pháp (CBENet và BGShadowNet) trong Zhang et al. [1] làm backbone (xương sống). Khi xây dựng, tiến hành điều chỉnh và cải tiến các mạng convolution trong 2 mạng CBENet và BGShadowNet bằng Vision Transformers. Sau khi cài đặt, huấn luyện mô hình với bộ dữ liệu RDD có sẵn. Sử dụng mô hình thu được tiến hành đánh giá và điều chỉnh các tham số cho phù hợp với hệ thống, nhằm cải thiện độ chính xác và hiệu quả.

Phương pháp đánh giá hệ thống sẽ sử dụng một số độ đo để đánh giá tương đồng giữa kết quả và ảnh ground truth. Từ các kết quả đánh giá đó, cải tiến hệ thống cho phù hợp, có thể chạy thử hệ thống trên một số bộ dữ liệu khác nhằm đánh giá khả năng tổng quát

hóa của hệ thống.

Giới hạn: Đề tài khử bóng đổ trên ảnh tài liệu chụp từ điện thoại, các thiết bị kỹ thuật số,… chủ yếu tập trung vào các ảnh tài liệu có bóng đổ từ các nguồn ánh sáng phổ biến và không xét đến các trường hợp nhiễu ảnh quá phức tạp như bóng đổ động hay bóng của các vật thể chuyển động.

Phương pháp được sử dụng để khử bóng và cải thiện chất lượng ảnh là mô hình xây dựng dựa trên ViTs dùng 2 mạng CBENet và BGShadowNet từ Zhang et al. [1] làm xương sống.

Ngôn ngữ trong tài liệu chứa trong các ảnh bị giới hạn bởi bộ dữ liệu huấn luyện, do đó, mô hình sẽ hoạt động tốt trên các ngôn như như tiếng Anh, tiếng Trung Quốc và các ngôn ngữ sử dụng chữ la tinh làm chữ viết.

Tập dữ liệu thử nghiệm là tập dữ liệu RDD, bao gồm 3 tập lớn là ảnh có bóng đổ, nền ảnh và ảnh đã được khử bóng, mỗi tập sẽ gồm 4916 ảnh tương ứng với nhau.

Tài liệu tham khảo:

[1] Ling Zhang, Yinghao He, Qing Zhang, Zheng Liu, Xiaolong Zhang, Chunxia Xiao. Document image shadow removal guided by color-aware background. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

Kế hoạch thực hiện: Để đảm bảo hoàn thành đề tài đúng tiến độ và đạt được mục tiêu đề ra, kế hoạch thực hiện sẽ được chia thành các giai đoạn cụ thể như sau:

Giai đoạn 1: Nghiên cứu cơ sở lý thuyết và tổng hợp tài liệu

Tìm hiểu về Vision Transformers (ViTs): Nghiên cứu các nguyên lý hoạt động, cấu trúc của mô hình ViTs, và so sánh với các mô hình truyền thống như CNN.

Khảo sát các nghiên cứu trước: Tìm hiểu các phương pháp hiện có về loại bỏ bóng đổ trên ảnh tài liệu (bao gồm cả các phương pháp học sâu và xử lý ảnh truyền thống).

Xác định công cụ và framework: Chọn các công cụ, thư viện, framework cần thiết (như PyTorch, TensorFlow) để triển khai mô hình.

Tổng hợp tài liệu: Ghi chú, tổng hợp tài liệu tham khảo về các phương pháp xử lý ảnh và công nghệ mới trong lĩnh vực này.

Giai đoạn 2: Thu thập và chuẩn bị dữ liệu

Thu thập dữ liệu: Thu thập bộ dữ liệu ảnh tài liệu có bóng đổ từ nhiều nguồn khác nhau (lấy từ bộ dữ liệu công khai RDD).

Xử lý dữ liệu: Phân loại dữ liệu để chuẩn bị cho quá trình huấn luyện mô hình.

Phân chia dữ liệu: Phân chia bộ dữ liệu thành tập huấn luyện, kiểm thử và đánh giá (training, testing datasets).

Giai đoạn 3: Xây dựng và huấn luyện mô hình

Xây dựng mô hình dựa trên Vision Transformers: Xây dựng kiến trúc mô hình dựa trên Vision Transformers, kết hợp với các phương pháp xử lý ảnh và pre-processing.

Huấn luyện mô hình: Sử dụng bộ dữ liệu đã chuẩn bị để huấn luyện mô hình, theo dõi quá trình huấn luyện và điều chỉnh siêu tham số (learning rate, batch size, epochs).

Tối ưu hóa mô hình: Sử dụng các kỹ thuật tối ưu hóa mô hình (như fine-tuning, regularization) để cải thiện hiệu suất.

Giai đoạn 4: Đánh giá và cải tiến mô hình

Đánh giá kết quả: Sử dụng các chỉ số RMSE, PSNR và SSIM để đánh giá chất lượng mô hình. So sánh với các phương pháp trước đó.

Cải tiến mô hình: Dựa trên kết quả đánh giá, điều chỉnh kiến trúc mô hình hoặc kỹ thuật huấn luyện để tăng độ chính xác và giảm nhiễu.

Chạy thử nghiệm trên các bộ dữ liệu khác: Kiểm thử mô hình trên các bộ dữ liệu chưa được huấn luyện để đánh giá khả năng tổng quát hóa.

Giai đoạn 5: Viết báo cáo và chuẩn bị

Viết báo cáo: Tổng hợp kết quả nghiên cứu, phân tích và viết báo cáo chi tiết về quy trình thực hiện, kết quả đạt được và những đóng góp mới của đề tài.

Chuẩn bị thuyết trình: Chuẩn bị bài thuyết trình cho buổi bảo vệ đề tài, tập trung vào các điểm mạnh, điểm mới và cải tiến của phương pháp.

Giai đoạn 6: Hoàn thiện và bảo vệ khóa luận

Kiểm tra và hoàn thiện: Kiểm tra lại toàn bộ sản phẩm, tài liệu và báo cáo để đảm bảo không có sai sót.

Bảo vệ đề tài: Thực hiện thuyết trình và bảo vệ đề tài trước hội đồng.

Trong suốt quá trình thực hiện, cần có sự đánh giá định kỳ để điều chỉnh kế hoạch và tiến độ nếu cần thiết.

Xác nhận của CBHD (Ký tên và ghi rõ họ tên)	**TP. HCM, ngày 10 tháng 09 năm 2024** **Sinh viên** (Ký tên và ghi rõ họ tên)

LỜI CẢM ƠN

Trước hết, tôi xin gửi lời cảm ơn chân thành và sâu sắc nhất đến tất cả những cá nhân và tổ chức đã hỗ trợ tôi trong suốt quá trình học tập, nghiên cứu và thực hiện khóa luận tốt nghiệp này. Tôi xin bày tỏ lòng biết ơn sâu sắc đến TS. Nguyễn Duy Khánh, người đã tận tâm chỉ bảo, hướng dẫn và đóng góp những ý kiến quý báu trong suốt quá trình thực hiện khóa luận. Những định hướng và sự tận tình của thầy đã giúp tôi vượt qua những khó khăn và hoàn thiện nghiên cứu này một cách tốt nhất. Tôi cũng xin gửi lời cảm ơn chân thành đến các thầy cô trong khoa Khoa học máy tính, những người đã truyền đạt kiến thức và kinh nghiệm quý báu trong suốt thời gian học tập tại trường. Những bài giảng và sự hỗ trợ của các thầy cô đã giúp tôi có nền tảng vững chắc để thực hiện nghiên cứu này. Đồng thời, tôi xin cảm ơn các cơ quan, tổ chức, nơi đã tạo điều kiện cung cấp tài liệu, công cụ và dữ liệu cần thiết, góp phần quan trọng vào việc hoàn thành khóa luận. Tôi cũng muốn gửi lời cảm ơn đến các bạn bè và đồng nghiệp, những người đã luôn sẵn sàng hỗ trợ, chia sẻ và động viên tôi trong suốt quá trình thực hiện nghiên cứu. Cuối cùng, tôi xin bày tỏ lòng biết ơn sâu sắc đến gia đình, những người luôn là nguồn động viên tinh thần to lớn, giúp tôi vượt qua mọi khó khăn để hoàn thành khóa luận này. Một lần nữa, tôi xin gửi lời tri ân sâu sắc nhất đến tất cả những ai đã đồng hành và giúp đỡ tôi trong suốt hành trình vừa qua.

Nguyễn Quốc Trường

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC TỪ VIẾT TẮT

ViTs	**Vi**sion **T**ransformer**s**
CBENet	**C**olor-aware **b**ackground **e**xtraction **net**work
BGShadowNet	**B**ack**g**round-guided **shadow** removal **net**work
Conv	**Conv**olutional
Conv2D	**Conv**olutional **2 d**imension
BAModule	**B**ackground-based **A**ttention **Module**
DEModule	**D**etail **E**nhancement **Module**
RMSE	**R**oot **M**ean **S**quared **E**rror
PSNR	**P**eak **S**ignal-to-**N**oise **R**atio
SSIM	**S**tructural **S**imilarity **I**ndex **M**easurement
CBETransformers	**CBE**Net based on Vision **Transformers**
BGSTransformers	**BGS**hadowNet based on Vision **Transformers**
CNN	**C**onvolutional **N**eural **N**etwork
OCR	**O**ptical **C**haracter **R**ecognition

TÓM TẮT KHÓA LUẬN

Bóng trên ảnh (Image Shadows) là vùng tối xuất hiện khi một nguồn sáng bị cản bởi một vật thể, khiến ánh sáng không thể chiếu trực tiếp vào bề mặt phía sau vật thể đó. Bóng trên ảnh tài liệu (Document Image Shadows) là trường hợp đặc biệt của bóng trên ảnh, xảy ra khi chụp hoặc quét các tài liệu giấy. Bóng đổ (Shadow) là một trong các tác nhân có ảnh hưởng tiêu cực rất lớn tới ảnh và các tác vụ xử lý ảnh. Chúng gây nhiễu và giảm độ chính xác trong qua trình phân tích và trích xuất từ ảnh, giảm chất lượng hình ảnh khiến ảnh khó đọc và bị mất thông tin,... ngoài ra còn có rất nhiều tác động tiêu cực khác. Vì lý do đó, bài toán 'Document Image Shadow Removal' - 'Loại bỏ bóng đổ trên ảnh tài liệu' đã ra đời nhằm phục vụ cho việc loại bỏ bóng đổ và khôi phục chất lượng ảnh, giúp giảm các ảnh hưởng tiêu cực của bóng đổ trên ảnh tài liệu.

Trước đây, đã có nhiều phương pháp được đề xuất để giải quyết vấn đề này. Tuy nhiên, các phương pháp đó thường sử dụng cách tiếp cận chung từ các bài toán xử lý ảnh khác mà không chú trọng đến những đặc điểm riêng của ảnh và bóng trên ảnh tài liệu. Điều này dẫn đến hiệu quả không thực sự tối ưu khi áp dụng trong các trường hợp cụ thể như ảnh tài liệu với bóng phức tạp hoặc đa dạng. Hiện nay, để khắc phục các điểm hạn chế của các phương pháp truyền thống, các nhà nghiên cứu trong lĩnh vực này đã đưa ra nhiều cách tiếp cận mới mẻ, như sử dụng mặt nạ bóng (shadow mask), nền ảnh (background),... hay kết hợp với các mạng nơ-ron mới và mạnh mẽ hoặc là một phương pháp kết hợp với Vision Transformers [1]. Các phương pháp sau này đem lại hiệu quả cao hơn hẳn, và cũng bởi vì ngày càng có nhiều cách khác nhau để tiếp cận bài toán, khả năng phát triển của các phương pháp này vẫn còn là rất lớn, và vẫn có thể cải thiện cho kết quả tốt hơn được.

Trong khóa luận này, chúng tôi đề xuất một phương pháp mới nhằm loại bỏ bóng trên ảnh tài liệu bằng cách ứng dụng Vision Transformers (ViTs) [1], một kiến trúc mạng dựa trên Transformers, kết hợp với phương pháp ‘loại bỏ bóng đổ dựa trên cơ chế nền nhận biết màu sắc’ [2]. Các ViTs sẽ được sử dụng để thay thế cho

các lớp Convolutional (Conv) truyền thống. Cụ thể, ViTs được tích hợp vào hai mạng sâu khá mới và hiệu quả là CBENet và BGShadowNet [2].

Các thí nghiệm được thực hiện trên ba bộ dữ liệu tiêu chuẩn gồm RDD [2], Jung [3] và Kliger [4], đại diện cho nhiều điều kiện và kiểu bóng khác nhau, giúp đảm bảo tính đa dạng và khả năng đánh giá toàn diện của mô hình. Mô hình đề xuất được so sánh với các phương pháp truyền thống và các mô hình gần đây nhất trong lĩnh vực loại bỏ bóng trên ảnh tài liệu. Kết quả thí nghiệm cho thấy phương pháp dựa trên ViTs đạt hiệu quả tốt, cải thiện đáng kể về mặt hiệu suất so với các phương pháp trước, đặc biệt trong việc xử lý các trường hợp bóng phức tạp.

Phương pháp của chúng tôi không chỉ tăng cường độ chính xác mà còn cải thiện khả năng tổng quát hóa, giúp giảm thiểu hiện tượng mất thông tin quan trọng trong tài liệu sau khi loại bỏ bóng. Đây là một bước tiến quan trọng trong việc áp dụng mô hình Vision Transformers vào các bài toán xử lý ảnh, mở ra nhiều tiềm năng ứng dụng trong thực tế.

Chương 1. **GIỚI THIỆU**

1.1. Tổng quan

Trong nhiều lĩnh vực như lưu trữ tài liệu, số hóa sách báo, xử lý văn bản, các hệ thống nhận diện ký tự quang học (OCR), việc đảm bảo chất lượng hình ảnh tài liệu là một yêu cầu quan trọng. Tuy nhiên, quá trình thu thập dữ liệu thường gặp phải nhiều thách thức, trong đó bóng trên ảnh tài liệu là một vấn đề đáng kể. Bóng có thể xuất hiện do điều kiện ánh sáng không ổn định, do sự hiện diện của vật cản trong quá trình chụp ảnh, hoặc từ các yếu tố môi trường khác, gây khó khăn trong việc xử lý và phân tích thông tin trên tài liệu.

Bóng trên tài liệu làm giảm độ tương phản giữa các vùng văn bản và nền, làm mờ nét chữ và đôi khi gây biến dạng hình ảnh, dẫn đến khó khăn trong việc nhận diện và trích xuất thông tin. Điều này ảnh hưởng nghiêm trọng đến các hệ thống nhận diện văn bản tự động, gây sai lệch kết quả và giảm hiệu suất làm việc. Do đó, việc nghiên cứu và phát triển các phương pháp loại bỏ bóng trên ảnh tài liệu đã trở thành một chủ đề quan trọng trong lĩnh vực xử lý ảnh.

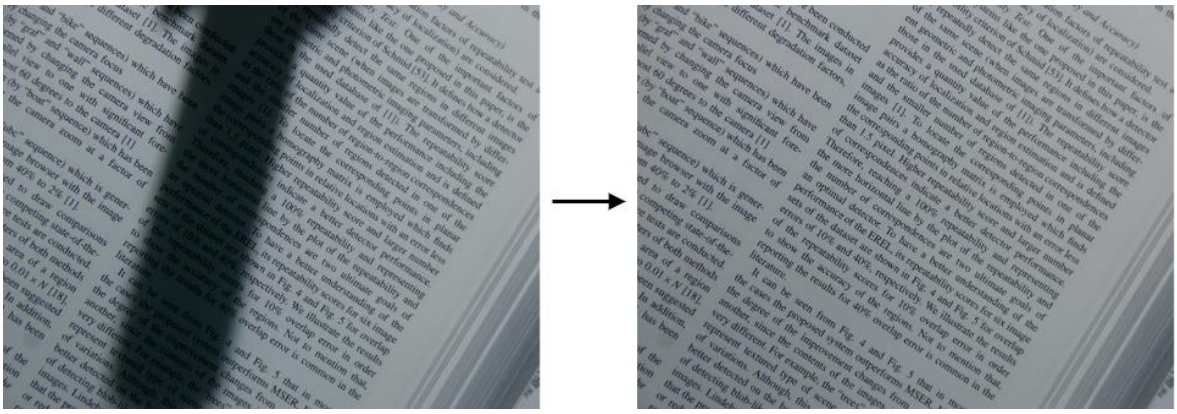
Trước đây, các phương pháp loại bỏ bóng chủ yếu dựa trên các kỹ thuật truyền thống sử dụng mạng nơ-ron tích chập (CNN) và các bộ lọc hình ảnh để phân tách bóng khỏi các vùng văn bản. Tuy nhiên, các phương pháp này vẫn gặp nhiều hạn chế, đặc biệt là khi đối mặt với các trường hợp bóng phức tạp hoặc trong môi trường ánh sáng không đều. Gần đây, sự phát triển của các mô hình Transformer, đặc biệt là Vision Transformers (ViTs) [1], đã mở ra nhiều hướng nghiên cứu mới với tiềm năng cao trong việc giải quyết các vấn đề phức tạp trong xử lý ảnh. ViTs cho phép mô hình hóa thông tin toàn cục (global information) tốt hơn, giúp nhận diện và phân tách các vùng bóng một cách hiệu quả hơn so với các phương pháp truyền thống.

Trong khóa luận này, chúng tôi tập trung nghiên cứu việc tích hợp Vision Transformers vào hai mạng sâu là CBENet và BGShadowNet [2] để nâng cao hiệu

quả loại bỏ bóng trên ảnh tài liệu. Bên cạnh đó, chúng tôi sẽ tiến hành thí nghiệm trên ba bộ dữ liệu tiêu chuẩn để đánh giá hiệu suất của mô hình đề xuất, đồng thời so sánh với các phương pháp trước đây nhằm đưa ra kết luận về tính ưu việt của giải pháp mới này.

1.1.1. Đặt vấn đề

Trong thời đại số hóa hiện nay, việc chuyển đổi các tài liệu giấy sang định dạng số ngày càng trở nên quan trọng nhằm lưu trữ, tìm kiếm và quản lý thông tin một cách hiệu quả. Tuy nhiên, trong quá trình chụp ảnh tài liệu, điều kiện ánh sáng không đồng đều hoặc sự xuất hiện của các vật cản thường gây ra bóng trên ảnh tài liệu, làm giảm chất lượng hình ảnh và ảnh hưởng tiêu cực đến các bước xử lý tiếp theo như nhận dạng ký tự quang học.



Hình 1.1: Loại bỏ bóng đổ trên ảnh tài liệu

Bóng trên ảnh tài liệu không chỉ che khuất thông tin quan trọng mà còn làm biến dạng các ký tự và kết cấu của tài liệu, gây khó khăn cho các hệ thống xử lý ảnh tự động. Việc loại bỏ bóng một cách hiệu quả là cần thiết để đảm bảo tính chính xác và hiệu quả của các ứng dụng xử lý ảnh tài liệu.

Các phương pháp truyền thống dựa trên kỹ thuật xử lý ảnh và mạng neural tích chập (CNN) đã được áp dụng để giải quyết vấn đề này. Tuy nhiên, những

phương pháp này thường gặp khó khăn trong việc xử lý các biến đổi phức tạp của bóng và kết cấu nền. Đặc biệt, khả năng nắm bắt thông tin ngữ cảnh toàn cục của các phương pháp này còn hạn chế.

Trong bối cảnh đó, Vision Transformers (ViTs) [1] đã nổi lên như một công cụ mạnh mẽ trong việc xử lý ảnh với cơ chế tự chú ý, cho phép mô hình nắm bắt và xử lý thông tin ngữ cảnh toàn cục một cách hiệu quả. Việc áp dụng ViTs vào bài toán loại bỏ bóng trên ảnh tài liệu có thể mang lại những cải tiến đáng kể về độ chính xác và hiệu suất.

Mục tiêu của đề tài này là phát triển một phương pháp sử dụng ViTs để loại bỏ bóng trên ảnh tài liệu. Phương pháp đề xuất sẽ chia ảnh tài liệu thành các mảng nhỏ, nhúng các mảng này và sử dụng bộ mã hóa Transformer để học các đặc trưng và loại bỏ bóng. Kết quả thực nghiệm sẽ được đánh giá để chứng minh tính hiệu quả của phương pháp đề xuất so với các phương pháp truyền thống.

1.1.2. Thách thức và động lực

Nghiên cứu về loại bỏ bóng trên ảnh tài liệu gặp phải nhiều thách thức, bao gồm:

- Phát hiện và phân đoạn bóng chính xác: Bóng có thể xuất hiện với hình dạng và kích thước đa dạng, từ bóng nhỏ, mờ nhạt đến các vùng bóng lớn, đậm. Việc phân biệt bóng và các đặc trưng nền của ảnh tài liệu là một bài toán khó, đặc biệt khi có sự thay đổi phức tạp về ánh sáng.
- Xử lý sự không đồng nhất của ánh sáng: Ánh sáng không đồng đều làm thay đổi màu sắc và độ sáng của nền, gây khó khăn trong việc khôi phục ảnh về trạng thái nguyên bản.
- Thiếu dữ liệu chất lượng cao: Các bộ dữ liệu về ảnh tài liệu với bóng (như RDD [2], Jung [3], và Kliger [4]) thường có quy mô nhỏ, dẫn đến nguy cơ overfitting khi huấn luyện mô hình. Bóng tài

liệu trong thế giới thực thường phức tạp hơn so với dữ liệu mô phỏng trong các bộ dữ liệu hiện có.

- Khả năng tổng quát hóa: Các mô hình học sâu hiện tại thường hoạt động tốt trên bộ dữ liệu huấn luyện, nhưng gặp khó khăn khi áp dụng cho các bộ dữ liệu mới với điều kiện ánh sáng hoặc bóng khác biệt.
- Tài nguyên tính toán: Các mô hình học sâu hiện đại, đặc biệt là những mô hình sử dụng Transformer như ShaDocFormer, yêu cầu tài nguyên tính toán lớn và thời gian huấn luyện lâu, gây khó khăn cho việc triển khai thực tế.

Mặc dù gặp nhiều thách thức, nghiên cứu trong lĩnh vực này mang lại nhiều lợi ích thiết thực và mở ra nhiều cơ hội:

- Tầm quan trọng thực tiễn: Bóng trên ảnh tài liệu là vấn đề phổ biến trong các ứng dụng thực tế, như số hóa tài liệu, xử lý văn bản trong nhận dạng ký tự (OCR), và phân tích tài liệu số. Loại bỏ bóng hiệu quả sẽ cải thiện đáng kể chất lượng đầu ra của các ứng dụng này.
- Khả năng áp dụng rộng rãi: Các giải pháp loại bỏ bóng không chỉ hữu ích trong xử lý tài liệu mà còn có thể mở rộng sang các lĩnh vực khác như nhiếp ảnh, thị giác máy tính, và đồ họa.
- Cải tiến công nghệ: Việc tích hợp Transformers vào các mô hình loại bỏ bóng không chỉ giải quyết bài toán hiện tại mà còn mở ra hướng nghiên cứu mới, tận dụng khả năng học toàn cục để giải quyết các vấn đề phức tạp hơn.
- Tạo ra mô hình tiên tiến hơn: Khắc phục được các hạn chế hiện tại, như cải thiện tính tổng quát hóa và giảm độ phức tạp tính toán, sẽ góp phần nâng cao chất lượng và hiệu quả của các hệ thống xử lý ảnh.
- Thúc đẩy nghiên cứu liên ngành: Loại bỏ bóng tài liệu là một bài toán giao thoa giữa các lĩnh vực như thị giác máy tính, học máy, và xử lý

ngôn ngữ tự nhiên. Thành công trong lĩnh vực này sẽ thúc đẩy các nghiên cứu liên ngành khác phát triển.

1.2. Mục tiêu nghiên cứu

Mục tiêu chính của khóa luận là đề xuất và phát triển một phương pháp loại bỏ bóng trên ảnh tài liệu dựa trên Vision Transformers (ViTs) [1], nhằm cải thiện độ chính xác và hiệu quả so với các phương pháp hiện có. Các mục tiêu cụ thể bao gồm:

- Tích hợp ViTs vào các mô hình mạng hiện có: Chúng tôi sẽ thay thế các lớp Convolutional truyền thống trong hai mạng CBENet và BGShadowNet [2] bằng các lớp ViTs để khai thác khả năng mô hình hóa thông tin toàn cục của ViTs.
- Thực nghiệm và đánh giá trên các bộ dữ liệu tiêu chuẩn: Tiến hành thí nghiệm trên ba bộ dữ liệu RDD [2], Jung [3] và Kliger [4], nhằm đảm bảo tính đa dạng của các trường hợp bóng trên tài liệu và đánh giá hiệu quả của phương pháp đề xuất.
- So sánh với các phương pháp trước đây: Đánh giá hiệu suất của mô hình đề xuất so với các phương pháp truyền thống dựa trên CNN và các kỹ thuật xử lý ảnh khác, qua đó chứng minh tính hiệu quả và ưu điểm của Vision Transformers trong bài toán loại bỏ bóng.
- Đưa ra kết luận về khả năng ứng dụng thực tiễn: Dựa trên kết quả thực nghiệm, đưa ra nhận định về tiềm năng áp dụng của phương pháp trong các hệ thống xử lý tài liệu thực tế, đặc biệt là trong các ứng dụng OCR và lưu trữ tài liệu số hóa.

1.3. Phạm vi nghiên cứu

Phạm vi nghiên cứu của khóa luận tập trung vào các khía cạnh sau:

- Loại bỏ bóng trên ảnh tài liệu: Khóa luận tập trung giải quyết vấn đề loại bỏ bóng trên ảnh tài liệu, một trong những bài toán quan trọng

trong lĩnh vực xử lý ảnh tài liệu và chuẩn bị dữ liệu cho hệ thống nhận dạng ký tự quang học (OCR).

- Sử dụng Vision Transformers (ViTs) [1]: Phương pháp chính được sử dụng là Vision Transformers, nhằm thay thế các lớp Conv truyền thống trong các mô hình CBENet và BGShadowNet [2]. Khóa luận sẽ tập trung vào việc tích hợp và điều chỉnh ViTs cho phù hợp với bài toán loại bỏ bóng trên ảnh tài liệu.
- Dữ liệu thực nghiệm: Các thí nghiệm được thực hiện trên ba bộ dữ liệu tiêu chuẩn gồm RDD [2], Jung [3] và Kliger [4]. Đây là các bộ dữ liệu chứa nhiều loại ảnh tài liệu khác nhau với các điều kiện bóng đa dạng, nhằm đánh giá toàn diện hiệu suất của phương pháp đề xuất.
- Phạm vi so sánh: Khóa luận sẽ so sánh phương pháp đề xuất với các phương pháp loại bỏ bóng truyền thống, bao gồm cả các kỹ thuật xử lý ảnh cổ điển và các phương pháp dựa trên mạng CNN. Phạm vi so sánh không chỉ giới hạn ở độ chính xác mà còn bao gồm khả năng tổng quát hóa và tốc độ xử lý.

1.4. Cấu trúc khóa luận

Khóa luận được chia thành 5 chương, với nội dung được tổ chức như sau:

Chương 1: Giới thiệu

Chương này trình bày tổng quan về vấn đề nghiên cứu, bao gồm bối cảnh và tầm quan trọng của việc loại bỏ bóng trên ảnh tài liệu, các thách thức hiện tại và lý do chọn Vision Transformers làm giải pháp thay thế. Ngoài ra, chương này cũng nêu rõ mục tiêu, phạm vi và phương pháp nghiên cứu, đồng thời phác thảo cấu trúc chung của khóa luận.

Chương 2: Cơ sở lý thuyết và các phương pháp liên quan

Chương này cung cấp cái nhìn tổng quan về các khái niệm và lý thuyết nền tảng liên quan đến xử lý ảnh, đặc biệt là bài toán loại bỏ bóng. Các kiến trúc mạng,

mạng CBENet, BGShadowNet, và kiến trúc Vision Transformers (ViTs) cũng được trình bày chi tiết. Ngoài ra, các phương pháp loại bỏ bóng trước đây sẽ được thảo luận và đánh giá để làm cơ sở so sánh với phương pháp đề xuất.

Chương 3: Phương pháp đề xuất

Chương này trình bày chi tiết về phương pháp mà khóa luận đề xuất, trong đó tập trung vào việc tích hợp Vision Transformers vào hai mạng CBENet và BGShadowNet. Cấu trúc của mô hình mới sẽ được mô tả, cùng với cách điều chỉnh các thành phần của ViTs để phù hợp với nhiệm vụ loại bỏ bóng trên ảnh tài liệu. Các thông số kỹ thuật và quy trình huấn luyện mô hình cũng sẽ được trình bày trong chương này.

Chương 4: Thực nghiệm và đánh giá

Chương này mô tả các thí nghiệm được tiến hành để đánh giá hiệu suất của mô hình đề xuất. Các bộ dữ liệu RDD, Jung và Kliger được sử dụng trong các thí nghiệm sẽ được giới thiệu, cùng với các tiêu chí đánh giá như độ chính xác, hiệu suất tổng quát hóa, và tốc độ xử lý. Kết quả so sánh giữa mô hình ViTs với các phương pháp trước đây sẽ được phân tích và thảo luận chi tiết.

Chương 5: Kết luận và hướng phát triển

Chương cuối cùng tóm tắt lại những đóng góp chính của khóa luận và rút ra những kết luận từ các kết quả nghiên cứu. Bên cạnh đó, chương này cũng sẽ nêu ra những hạn chế của phương pháp hiện tại và đề xuất một số hướng nghiên cứu tiềm năng trong tương lai nhằm cải thiện thêm tính hiệu quả và khả năng ứng dụng của mô hình.

Chương 2. **CƠ SỞ LÝ THUYẾT**

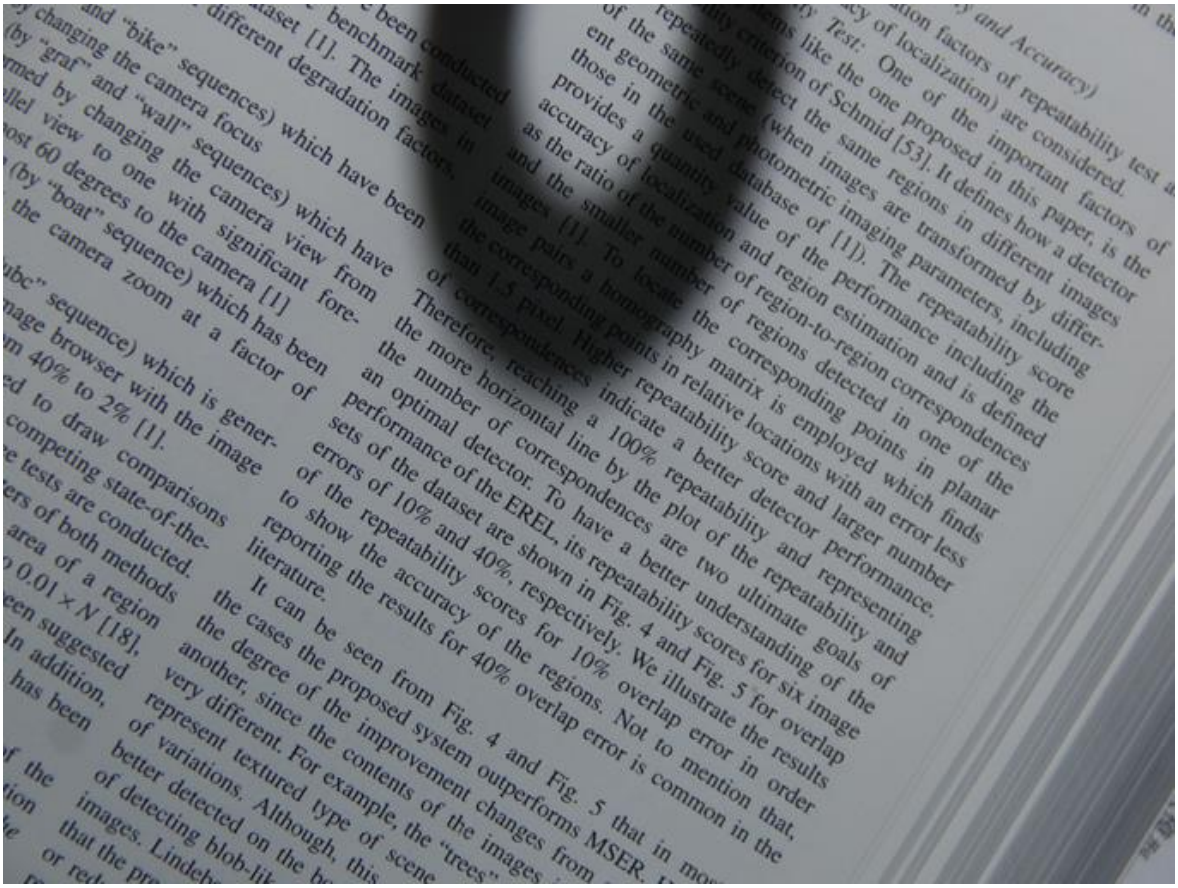
2.1. Tổng quan về chủ đề nghiên cứu

Loại bỏ bóng trên ảnh tài liệu là một thách thức quan trọng trong lĩnh vực xử lý ảnh và thị giác máy tính, đặc biệt đối với các ứng dụng như số hóa tài liệu, nhận dạng ký tự quang học (OCR), và quản lý lưu trữ thông tin. Bóng thường xuất hiện khi tài liệu được chụp hoặc quét trong điều kiện ánh sáng không đồng đều hoặc bị che khuất bởi các vật thể. Những bóng này không chỉ làm biến dạng hình ảnh mà còn làm mất đi các chi tiết quan trọng, gây khó khăn trong việc nhận dạng ký tự và phân tích nội dung, ảnh hưởng trực tiếp đến chất lượng và hiệu quả của các hệ thống xử lý tài liệu.

2.1.1. Đặt vấn đề

Loại bỏ bóng trên ảnh tài liệu là một trong những vấn đề quan trọng trong lĩnh vực xử lý ảnh và thị giác máy tính, đặc biệt trong các ứng dụng liên quan đến số hóa tài liệu, nhận dạng ký tự quang học, và lưu trữ thông tin. Bóng trên ảnh tài liệu thường xuất hiện khi tài liệu được chụp hoặc quét trong điều kiện ánh sáng không đồng đều hoặc bị cản trở bởi các vật thể, dẫn đến việc hình ảnh bị che khuất nội dung, biến dạng, mất chi tiết hoặc khó khăn trong việc nhận dạng các ký tự và hình ảnh trên tài liệu.

Bóng trên tài liệu không chỉ làm giảm chất lượng của hình ảnh mà còn ảnh hưởng nghiêm trọng đến các hệ thống tự động, đặc biệt là các hệ thống OCR, nơi yêu cầu hình ảnh có độ tương phản cao và rõ ràng để nhận dạng chính xác các ký tự. Khi có bóng xuất hiện, các vùng tài liệu bị bóng che khuất thường bị mất thông tin hoặc nhầm lẫn với vùng nền, dẫn đến kết quả nhận dạng sai.



Hình 2.1: Ảnh tài liệu bị bóng đổ che khuất một phần văn bản, khiến việc đọc và xử lý trở nên khó khăn

Hơn nữa, bóng thường có hình dạng và cường độ không đồng nhất, phụ thuộc vào nhiều yếu tố như góc chiếu sáng, độ mạnh của nguồn sáng, và hình dạng của các vật thể xung quanh. Do đó, việc loại bỏ bóng không phải là một nhiệm vụ đơn giản, đặc biệt là khi cần bảo toàn chi tiết của văn bản và các yếu tố quan trọng khác trên tài liệu.

2.1.2. Phát biểu bài toán

Đầu vào

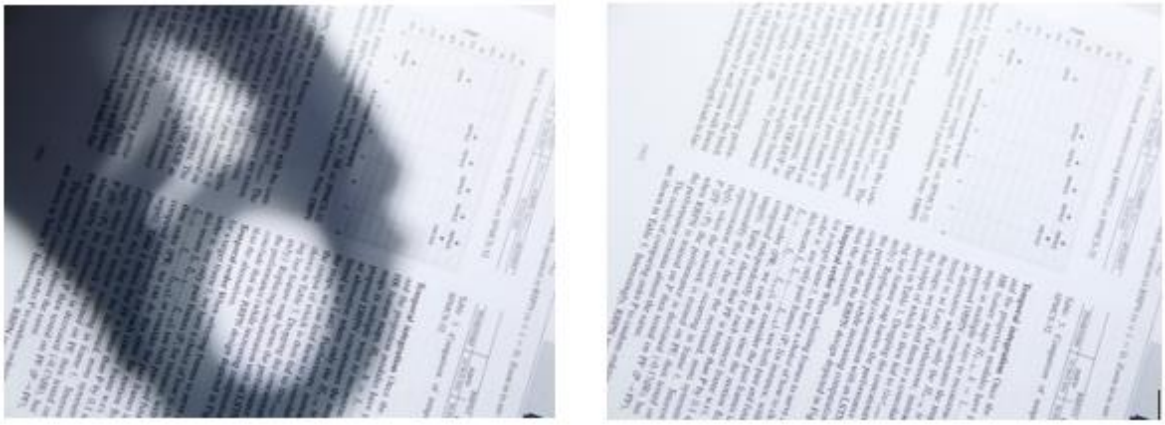
Ảnh tài liệu chứa các bóng do điều kiện ánh sáng không đồng đều hoặc các vật cản trong quá trình chụp ảnh.

Các ảnh tài liệu có thể ở các định dạng khác nhau (JPEG, PNG, TIFF, v.v.) và có kích thước, độ phân giải đa dạng.

Đầu ra

Ảnh tài liệu đã được loại bỏ bóng, cải thiện chất lượng, rõ ràng và dễ đọc hơn.

Các ảnh đầu ra sẽ có cùng kích thước và định dạng với ảnh đầu vào nhưng không còn chứa bóng.



Hình 2.2: Đầu vào (trái) và đầu ra (phải) của bài toán

2.1.3. Vision Transformers

Vision Transformers (ViTs) [1] là một mô hình học sâu sử dụng cấu trúc transformers, ban đầu được thiết kế cho các tác vụ xử lý ngôn ngữ tự nhiên, nhưng đã được áp dụng thành công cho các tác vụ thị giác máy tính. ViTs chuyển đổi ảnh đầu vào thành một chuỗi các patch nhỏ và sử dụng các khối transformers để xử lý chuỗi này. Cấu trúc ViTs bao gồm các bước chính sau:

- Chia Ảnh Thành Các Patch: Ảnh được chia thành các patch nhỏ không chồng chéo, mỗi patch được xem như một token.

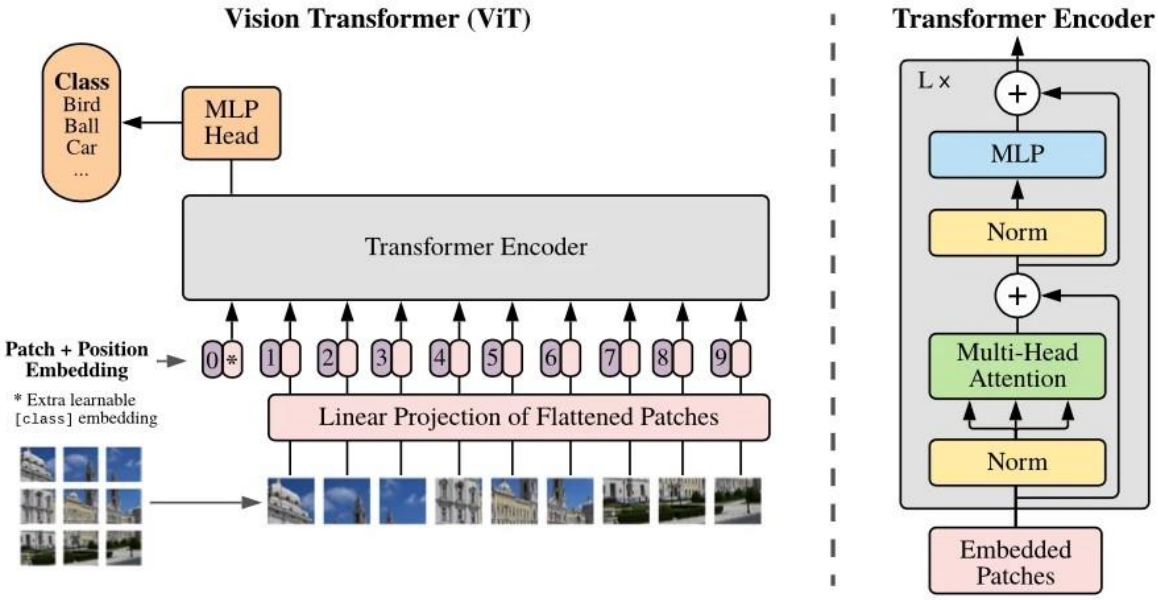
- Tạo Embedding: Mỗi patch được chuyển đổi thành một vector bằng cách sử dụng một lớp nhúng (embedding layer).
- Thêm Vị Trí Nhúng (Position Embedding): Thông tin vị trí được thêm vào mỗi patch để giữ nguyên cấu trúc không gian của ảnh.

Transformer Encoder: Các patch được đưa qua một chuỗi các lớp transformer encoder để xử lý và trích xuất các đặc trưng.

Classification Token: Một token đặc biệt được thêm vào chuỗi các patch để sử dụng cho mục đích phân loại cuối cùng.

Việc sử dụng Vision Transformer trong bài toán xóa bóng ảnh tài liệu có thể mang lại nhiều lợi ích, nhờ khả năng học các đặc trưng không gian và ngữ cảnh mạnh mẽ từ ảnh tài liệu. Các bước triển khai ViT trong bài toán này bao gồm:

- Preprocessing: Chia ảnh tài liệu thành các patch nhỏ.
- Embedding: Chuyển đổi mỗi patch thành vector nhúng và thêm thông tin vị trí.
- Training ViTs: Huấn luyện mô hình ViTs với dữ liệu ảnh tài liệu có bóng và không có bóng để học các đặc trưng phân biệt bóng.
- Prediction: Sử dụng mô hình ViTs đã huấn luyện để dự đoán và loại bỏ bóng trong ảnh tài liệu.



Hình 2.3: Cấu trúc của ViTs [1]

2.2. Tổng quan các nghiên cứu trước đây

2.2.1. Một số các nghiên cứu tiêu biểu

Xóa Bóng Trong Ảnh Tự Nhiên

Các phương pháp truyền thống [5, 6, 7] để xóa bóng trong ảnh tự nhiên thường tập trung vào nghiên cứu các đặc tính vật lý khác nhau của bóng. Finlayson và cộng sự [6] tái tạo các ảnh xóa bóng dựa trên tính nhất quán gradient. Tuy nhiên, các phương pháp này có thể gây ra hiện tượng viền bóng rõ rệt do sự thay đổi của ánh sáng. Shor và cộng sự [8] định nghĩa một mối quan hệ affine giữa các vùng có bóng và không có bóng. Xiao và cộng sự [9, 10] và Zhang và cộng sự [11] loại bỏ bóng bằng cách chuyển đổi chiếu sáng từ vùng không có bóng sang vùng có bóng. Tuy nhiên, các phương pháp này phụ thuộc vào các vùng không có bóng tham chiếu, và thường dẫn đến ánh sáng không đồng nhất khi các vùng tham chiếu không phù hợp.

Nhiều phương pháp dựa trên học máy đã được đề xuất để xóa bóng trong ảnh tự nhiên [12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. Ví dụ, Deshadow-Net [20] trích

xuất các đặc trưng đa ngữ cảnh để dự đoán các lớp bóng mờ cho việc xóa bóng. Wang và cộng sự [22] sử dụng các mạng GAN có điều kiện xếp chồng để phát hiện và xóa bóng cùng lúc. Zhang và cộng sự [23] khám phá sử dụng mạng GAN cho việc xóa bóng bằng cách sử dụng phần dư và chiếu sáng. ARGAN [24] đề xuất một mạng GAN tái phát sinh chú ý cho việc phát hiện và xóa bóng. Liu và cộng sự [19] sử dụng việc tạo bóng cho việc xóa bóng với giám sát yếu. Gần đây, Chen và cộng sự [25] chuyển đổi thông tin ngữ cảnh từ các vùng không có bóng sang các vùng có bóng trong không gian đặc trưng nhúng. Mặc dù các phương pháp này hiệu quả với ảnh tự nhiên, chúng không áp dụng tốt cho xóa bóng trong ảnh tài liệu do các đặc tính khác nhau giữa ảnh tự nhiên và ảnh tài liệu.

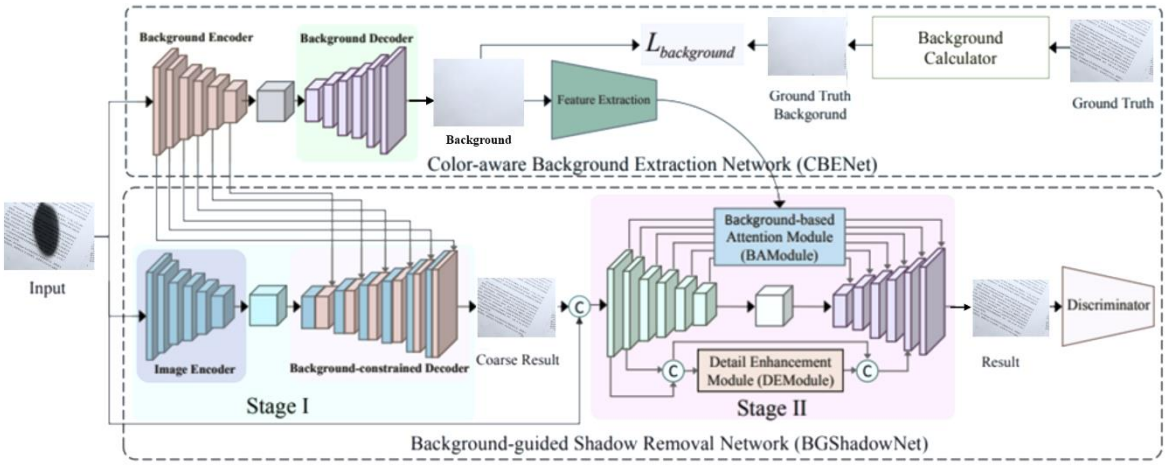
Xóa Bóng Trong Ảnh Tài Liệu

Hầu hết các thuật toán xóa bóng trong ảnh tài liệu hiện có sử dụng các thuật toán dựa trên kinh nghiệm để khai thác các đặc trưng cụ thể của ảnh tài liệu. Bako và cộng sự [26] xóa bóng bằng cách sử dụng bản đồ bóng ước tính. Phương pháp này để lại dấu vết nhẹ ở viền dưới bóng mạnh. Oliveira và cộng sự [27] sử dụng nội suy lân cận tự nhiên để ước tính ảnh bóng. Jung và cộng sự [3] khám phá phương pháp đổ nước để điều chỉnh ánh sáng của ảnh tài liệu bằng cách chuyển đổi ảnh đầu vào thành bề mặt địa hình. Phương pháp này đạt hiệu suất tốt với các bóng yếu hoặc vừa, nhưng có xu hướng làm suy giảm màu sắc đối với các cảnh có bóng nặng. Gần đây, Lin và cộng sự [4] đề xuất BEDSR-Net cho việc xóa bóng trong ảnh tài liệu bằng cách ước tính nền không đổi. Đây là mạng sâu đầu tiên được thiết kế đặc biệt cho việc xóa bóng trong ảnh tài liệu, tận dụng các đặc tính cụ thể của ảnh tài liệu. Tuy nhiên, do bỏ qua một số màu nền khác trong ảnh, phương pháp này có thể gây ra hiện tượng tạo viền bóng hoặc bóng không được xóa hết.

2.2.2. Phương pháp loại bỏ bóng đổ trên ảnh tài liệu sử dụng cơ chế nền nhận biết màu sắc

Vì ảnh tài liệu tập trung chủ yếu vào nội dung văn bản, một chiến lược phổ biến [26, 4] để xóa bóng trong ảnh tài liệu là sử dụng lớp nền được trích xuất từ ảnh, chỉ chứa thông tin màu sắc của ảnh mà không có nội dung văn bản. Các phương pháp này giả định tài liệu có nền màu không đổi (màu của giấy). Tuy nhiên, có thể tồn tại sự khác biệt giữa nền màu không đổi và ảnh.

Từ những yếu tố đó, Zhang và các cộng sự [2] đã đề xuất một phương pháp loại bỏ bóng đổ trên ảnh tài liệu sử dụng cơ chế nền nhận biết màu sắc. Phương pháp này về cơ bản chia việc xử lý ra làm hai phần riêng biệt nhưng liên kết với nhau. Mỗi phần sẽ được xử lý bằng một mạng riêng biệt.



Hình 2.4: Cấu trúc của hai mạng CBENet và BGShadowNet [2] trong phương pháp 'Loại bỏ bóng đổ trên ảnh tài liệu sử dụng cơ chế nền nhận biết màu sắc'

Một phần đề xuất mạng trích xuất nền nhận biết màu sắc (CBENet) để trích xuất nền màu biến đổi không gian cho ảnh tài liệu, bảo tồn các màu nền khác nhau trong ảnh. So với nền không đổi, nền biến đổi không gian có thể cung cấp thông tin màu sắc hữu ích hơn cho mạng xóa bóng tiếp theo. Lưu ý rằng nền này không có bóng, giúp BGShadowNet học được nhiều đặc trưng không bóng hơn,

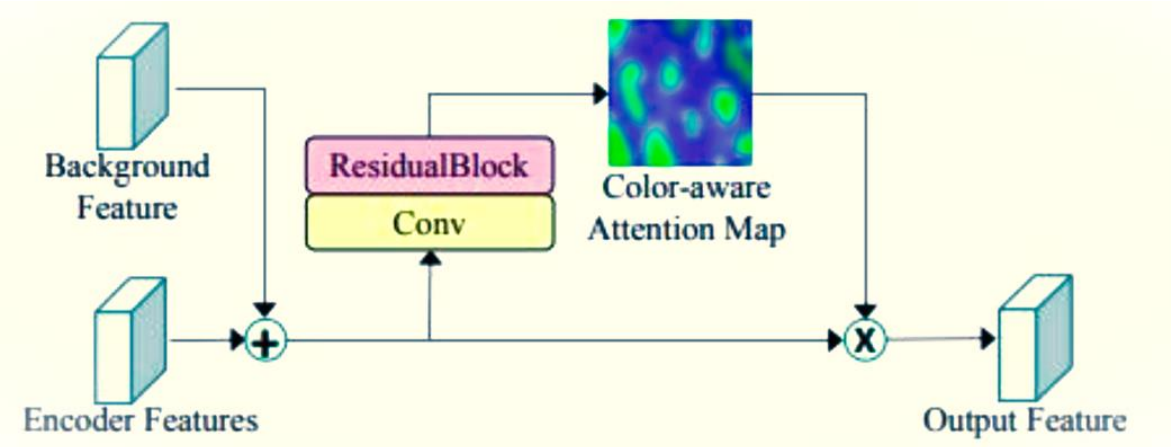
góp phần xóa bóng trong khi tránh tốt hơn các hiện tượng ánh sáng hoặc màu sắc trong ảnh. Khi huấn luyện CBENet, phương pháp này sử dụng chiến lược local-to-global (từ cục bộ tới toàn cục). Ảnh đầu vào khi được đưa vào mô hình sẽ được chia làm 16x16 patch bằng nhau. Từ đây, các patch sẽ được đưa vào mạng để trích xuất các đặc trưng nền ảnh để hình thành nền ở từng patch một. Sau đó, ở các lớp decoder, các patch được ghép lại với nhau thành ảnh hoàn chỉnh rồi sử dụng một operator làm mịn màu [28] để tinh chỉnh. Từ đó thu được nền ảnh hoàn chỉnh mới tiếp tục xử lý. Phương pháp này sử dụng cấu trúc U-Net để triển khai CBENet. U-Net trước tiên áp dụng năm lớp Conv+BN+LReLu để trích xuất đặc trưng từ ảnh. Sau đó, sử dụng năm lớp deconvolutional với chuẩn hóa batch và hàm kích hoạt ReLU để dự đoán ảnh nền. Kết nối bỏ qua được áp dụng giữa các lớp convolutional và deconvolutional, tăng số lượng kênh trong mạng và bảo tồn thông tin ngữ cảnh của lớp phía trước.

Hình 2.5: Nền ảnh thu được sau khi đưa qua CBENet, nền ảnh sau khi ghép các patch lại (trái) và nền ảnh hoàn chỉnh (phải)

Như đã đề cập, nền có thể cung cấp thông tin hữu ích để hỗ trợ việc xóa bóng. Từ đó có một phần còn lại đề xuất một mạng xóa bóng hướng dẫn bởi nền (BGShadowNet) khai thác ảnh nền như thông tin bổ sung. BGShadowNet bao gồm hai giai đoạn. Tại Giai đoạn I, bên cạnh bộ mã hóa ảnh, một bộ giải mã bị

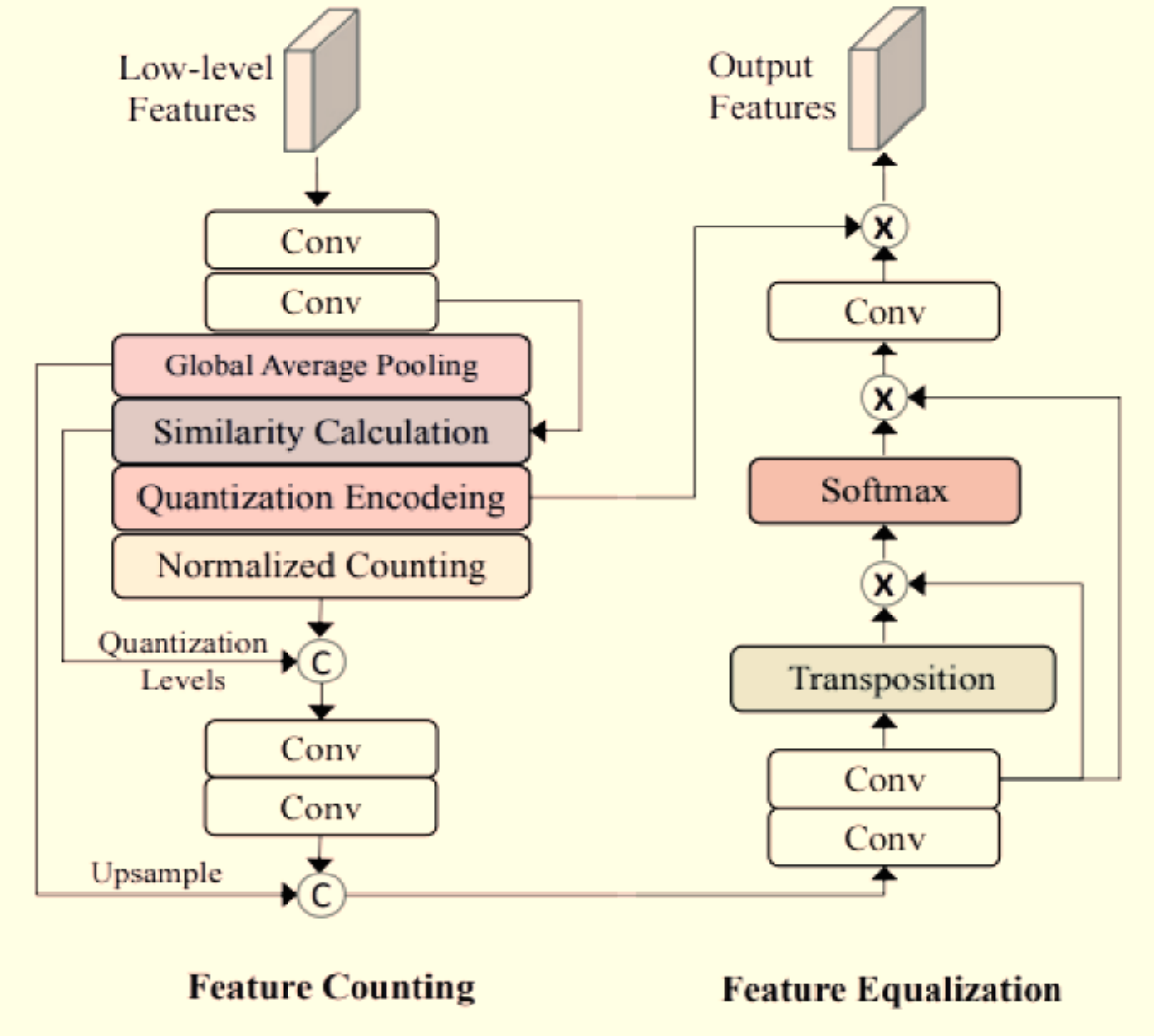
ràng buộc bởi nền được giới thiệu để tạo ra kết quả xóa bóng sơ bộ. Tại Giai đoạn II, để cải thiện kết quả sơ bộ và tạo ra ảnh không có bóng cuối cùng, một module chú ý dựa trên nền (BAModule) và một module tăng cường chi tiết (DEModule) được nhúng vào mạng mã hóa-giải mã. Một bộ phân biệt được xếp chồng ở cuối để phân biệt xem ảnh được tạo ra có thật hay không. Phương pháp này chọn DenseUnet [29] và bộ phân biệt Markovian [30] làm cấu trúc mã hóa-giải mã và bộ phân biệt. Background-constrained decoder: Để tận dụng tối đa các đặc trưng từ ảnh nền, phương pháp này thay thế bộ giải mã thông thường bằng bộ giải mã bị ràng buộc bởi nền tại Giai đoạn I. Cụ thể, các đặc trưng từ bộ mã hóa nền được tích hợp vào bộ giải mã bị ràng buộc bởi nền ở mỗi cấp độ tương ứng. Các đặc trưng tích hợp này có thể bổ sung cho các đặc trưng ảnh và giúp tạo ra kết quả xóa bóng thỏa đáng.

Background-based Attention Module (BAModule) [2]: Thông thường, các vùng có nền tương tự nên có sự xuất hiện tương tự (màu sắc và chiếu sáng) trong ảnh. Tuy nhiên, có thể có các hiện tượng ánh sáng hoặc màu sắc trong kết quả xóa bóng sơ bộ. Để giữ sự nhất quán tổng thể của ảnh, phương pháp giới thiệu module chú ý dựa trên nền (BAModule). Sử dụng các đặc trưng nền đã học và cơ chế chú ý, BAModule giúp loại bỏ sự không nhất quán về sự xuất hiện trong ảnh.



Hình 2.6: Cấu trúc BAModule [2]

Detail Enhancement Module (DEModule) [2]: Do các phép toán chập và giảm mẫu nhiều lần trong mạng, thông tin chi tiết sẽ bị mất ở các lớp cao, dẫn đến kết quả bị mờ chi tiết. So với các đặc trưng cao cấp, các đặc trưng thấp cấp trong các lớp CNN thường chứa nhiều chi tiết kết cấu hơn. Do đó, module tăng cường chi tiết (DEModule) được giới thiệu để khôi phục các chi tiết kết cấu của kết quả sơ bộ bằng cách sử dụng các đặc trưng cấp thấp của mạng. Như chúng ta biết, thông tin kết cấu thống kê của ảnh phản ánh cường độ kết cấu ở một mức độ nào đó. Vì vậy, DEModule được lấy cảm hứng từ cân bằng histogram ảnh, bao gồm hai phần: một là đếm đặc trưng để thu thập thông tin thống kê cho các đặc trưng cấp thấp, và phần kia là cân bằng đặc trưng để tăng cường chi tiết kết cấu. Cụ thể, mô hình kết hợp các đặc trưng của hai lớp cấp thấp đầu tiên từ bộ mã hóa để có được các đặc trưng thấp cấp kết hợp F, sau đó đưa vào DEModule để phân tích thống kê.



Hình 2.7: Cấu trúc DEModule [2]

Đếm Đặc Trưng (Feature Counting): Mục đích của việc đếm đặc trưng là thu được bản đồ mã hóa lượng tử và các đặc trưng thống kê. Đầu tiên, chúng tôi sử dụng hai lớp tích chập 2×2 để tạo ra một bản đồ đặc trưng (M) và thực hiện phép lấy trung bình toàn cục để thu được các đặc trưng trung bình toàn cục cho M (kí hiệu là $\overline{M}$). Tiếp theo, chúng tôi tính toán độ tương quan giữa M và $\overline{M}$ bằng cách sử dụng độ tương đồng cosine, được ký hiệu là S. Để thực hiện việc lượng tử hóa và thống kê hiệu quả, chúng tôi xây dựng một tập hợp các mức lượng tử L, chia phạm vi từ giá trị nhỏ nhất đến giá trị lớn nhất của S thành N phần bằng nhau. Sau đó, ma trận tương quan S có thể được lượng tử hóa thành ma trận mã hóa

lượng tử E bằng cách sử dụng L. Để tránh loại bỏ thông tin gradient, chúng tôi thực hiện một phép chuẩn hóa cho ma trận E. Chúng tôi tích hợp kết quả chuẩn hóa và các mức lượng tử L vào một bản đồ đếm lượng tử C, phản ánh các thống kê tương đối của các đặc trưng đầu vào cấp thấp. Do phép nối kênh, số kênh của C là 2. Do đó, chúng tôi thực hiện hai phép tích chập 1 × 1 cho C để tăng số kênh, sau đó thực hiện phép nối kênh với $\overline{M}$ để thu được thông tin thống kê tuyệt đối H. H biểu thị các đặc trưng thống kê, đóng vai trò như một biểu đồ tần số.

Ma trận mã hóa lượng tử E có thể tính dựa trên ma trận tương quan S và mức lượng tử L:

$$E_{i,n} = \begin{cases} 1 - |L_n - S_i|, & -\frac{0.5}{N} \le L_n - S_i \le \frac{0.5}{N} \\ 0, & \text{còn lại} \end{cases}$$

Trong đó:

i ∈ [1,HW] và n ∈ [1,N]

H và W là chiều cao và chiều rộng của ảnh

L_n là lớp thứ n của mức lượng tử L

S_i là hàng thứ i của ma trận tương quan S

N được thiết lập là 128

Cân Bằng Đặc Trưng (Feature Equalization): Cân bằng đặc trưng được sử dụng để tăng cường chi tiết kết cấu của các lớp cấp thấp bằng cách tái cấu trúc một tập hợp các mức lượng tử mới. Đầu tiên, chúng tôi thực hiện một phép tích chập 1 × 1 cho H để thu được G. Lấy cảm hứng từ cơ chế chú ý (attention), chúng tôi thực hiện phép nhân ma trận của G và ma trận chuyển vị của nó, sau đó là phép softmax, để xây dựng một ma trận kề học X. Ma trận X có thể được coi là một ma trận hệ số tương đồng. Sau đó, chúng tôi có thể tái cấu trúc các mức lượng tử mới L' bằng phép nhân ma trận của X và G. Dựa trên các mức lượng tử tái cấu trúc L', chúng tôi thực hiện cân bằng đặc trưng cho ma trận mã hóa lượng tử gốc E để tăng cường các đặc trưng chi tiết. Các đặc trưng được tăng cường R

có thể thu được bằng phép nhân ma trận của các mức lượng tử L' và ma trận E. Bằng cách sử dụng các chi tiết kết cấu được tăng cường, bộ giải mã có thể dễ dàng nắm bắt thông tin chi tiết.

2.2.3. Một số phương pháp áp dụng ViTs trước đây

Chúng tôi có tham khảo qua một số phương pháp loại bỏ bóng đổ trên ảnh có áp dụng ViTs trước đây, tiêu biểu như:

ShaDocFormer [31] là một kiến trúc dựa trên Transformer, được thiết kế để loại bỏ bóng trên ảnh tài liệu. Mô hình này kết hợp giữa các phương pháp truyền thống và học sâu để cải thiện hiệu suất xử lý. Nó bao gồm hai thành phần chính: Shadow-attentive Threshold Detector (STD): Sử dụng cơ chế attention của Transformer để thu thập thông tin toàn cục và phát hiện chính xác mặt nạ bóng. Cascaded Fusion Refiner (CFR): Áp dụng cấu trúc hợp nhất tầng bậc để khôi phục chi tiết hình ảnh từ thô đến tinh. Mô hình này vượt trội so với các phương pháp hiện đại khác trong cả đánh giá định tính và định lượng trên các bộ dữ liệu

ShaDocNet [32] là một mô hình sử dụng mạng học sâu để tập trung vào xử lý bóng trên tài liệu. Mặc dù chưa có nhiều thông tin chi tiết, mô hình này dường như sử dụng các kỹ thuật tương tự như ShaDocFormer nhưng có thể đơn giản hơn trong cấu trúc, tập trung vào các trường hợp tài liệu có bóng phức tạp.

DocDeshadower [33] là một mô hình dành riêng cho việc loại bỏ bóng khỏi các tài liệu quét và chụp ảnh. Phương pháp này chủ yếu dựa vào việc học biểu diễn từ dữ liệu, tận dụng mạng học sâu để giải quyết các vấn đề liên quan đến ánh sáng không đồng đều và bóng đổ, cải thiện khả năng đọc và hiển thị chi tiết của văn bản.

Ngoài việc tham khảo các phương pháp loại bỏ bóng trên ảnh tài liệu sử dụng ViTs như trên, chúng tôi còn tham khảo qua một số phương pháp loại bỏ bóng dựa trên ViTs như Spa-former [34], Crformer [35], TSRformer [36], ...

2.3. Nhận xét và đánh giá

Các phương pháp truyền thống, như dựa trên phân tích histogram và loại bỏ bóng bằng cách ước lượng ánh sáng, được đánh giá là hiệu quả trong các trường hợp đơn giản. Tuy nhiên, chúng thiếu tính tổng quát và thường không xử lý tốt các ảnh có điều kiện ánh sáng phức tạp.

Các phương pháp học sâu, đặc biệt là những mô hình dựa trên CNN, đã mang lại bước đột phá lớn nhờ khả năng tự động trích xuất đặc trưng và học tập trực tiếp từ dữ liệu. Tuy nhiên, hạn chế của các mô hình này là thường chỉ tập trung vào thông tin cục bộ, dẫn đến khó khăn khi xử lý các vùng bóng lớn hoặc thay đổi ánh sáng phức tạp.

Ngoài ra, khi thực hiện xử lý trên ảnh có nền chứa nhiều màu sắc bằng các phương pháp trước đây, những vùng ảnh bị bóng che khuất sẽ bị loại bỏ các màu sắc của nền ảnh ở khu vực đó, khiến cho chất lượng ảnh giảm sút. Vì thế, phương pháp ‘loại bỏ bóng đổ trên ảnh tài liệu sử dụng cơ chế nền nhận biết màu sắc’ mang lại hiệu quả vượt trội trong các trường hợp này, đem lại kết quả và chất lượng ảnh tốt hơn cả.

ViTs đã và đang dần được áp dụng rộng rãi vào để xử lý bài toán loại bỏ bóng, dựa trên khả năng mạnh mẽ trong lĩnh vực xử lý ảnh. Trước đây đã xuất hiện kha khá các phương pháp tích hợp ViTs nhưng hiệu quả vẫn chưa quá vượt trội so với các mô hình khác.

Từ các yếu tố chủ quan và khách quan đó, nhằm tạo ra một phương pháp mới, có hiệu quả vượt trội và cao hơn các phương pháp trước đây. Chúng tôi đã nảy sinh ra ý tưởng mới để xử lý bài toán loại bỏ bóng đổ trên ảnh tài liệu, đó là sử dụng kết hợp cả hai cơ chế là nền nhận biết màu sắc và sự mạnh mẽ trong lĩnh vực xử lý ảnh của ViTs.

Chương 3. **PHƯƠNG PHÁP ĐỀ XUẤT**

3.1. Phương pháp đề xuất

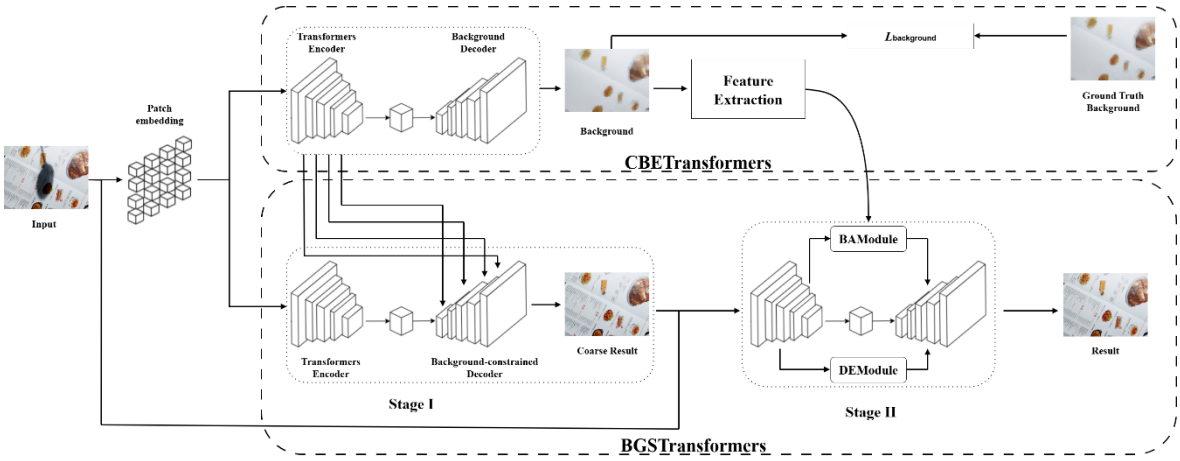
Phương pháp nghiên cứu được sử dụng trong khóa luận, bao gồm việc đề xuất cải tiến các mô hình hiện có bằng cách tích hợp Vision Transformers (ViTs) [1] và quy trình thực nghiệm để đánh giá hiệu quả của mô hình.

Phương pháp nghiên cứu của khóa luận tập trung vào việc cải tiến hai mạng sâu loại bỏ bóng trên ảnh tài liệu là CBENet và BGShadowNet [2] bằng cách thay thế các lớp tích chập (Convolutional Layers) truyền thống bằng Vision Transformers (ViTs). Việc sử dụng ViT giúp mô hình có khả năng xử lý thông tin toàn cục (global information) tốt hơn, từ đó cải thiện hiệu quả loại bỏ bóng, đặc biệt là trong các trường hợp bóng phức tạp.

3.2. Kiến trúc mô hình đề xuất

Dựa trên mô hình được đề xuất trong bài báo ‘Document image shadow removal guided by color-aware background’ [2], chúng tôi đã thay đổi và phát triển lại mô hình này dựa trên Vision Transformers.

Ý tưởng của hai mạng CBE và BGShadow là khi trích xuất đặc trưng từ ảnh, sẽ chia ảnh thành 16x16 patch bằng nhau, rồi mới đưa qua tính toán và xử lý. Dựa trên việc chia ảnh thành các patch rồi mới đưa qua Encoder và Decoder đó, chúng tôi đã nảy ra ý tưởng sử dụng ViTs để thay thế cho các mạng Encoder thông thường trong hai mạng này.



Hình 3.1: Cấu trúc của mô hình dựa trên phương pháp mà chúng tôi đề xuất (gồm hai mạng CBETransformers và BGSTransformers)

3.2.1. CBETransformers

Vì mạng CBETransformers này xây dựng dựa trên CBENet, nên về cơ bản, cấu trúc của hai mạng này là tương tự, chỉ khác biệt ở các thành phần chủ chốt.

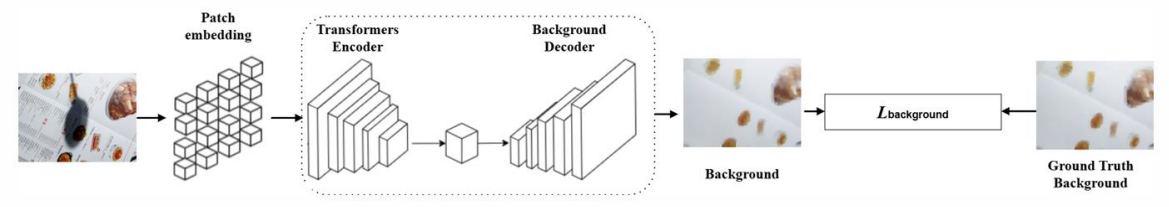
Cấu trúc của mô hình:

Patch Embedding: Chia ảnh đầu vào thành các patches (16x16) và nhúng chúng thành các vectors embedding có kích thước cố định.

Transformers Encoder Layer: Một lớp encoder của Vision Transformers, bao gồm một Multihead Attention và một Feed Forward Network (FFN) với Dropout và Layer Normalization.

Transformers Encoder: Tập hợp nhiều lớp Transformers Encoder Layer, nhằm học đặc trưng từ các vector embedding.

Background Decoder: Gồm nhiều lớp Decoder, thu thập đặc trưng học được từ Transformers Encoder và tái tạo ảnh bằng lớp Conv2D, nhằm thu được nền ảnh hoàn chỉnh.



Hình 3.2: Minh họa cấu trúc của mạng CBETransformers

Cài đặt mô hình:

Xây dựng các lớp cần thiết như Patch Embedding, Transformers Encoder và Decoder để trích xuất nền ảnh.

Kết hợp các lớp này để xây dựng mạng CBETransformers hoàn chỉnh.

3.2.2. BGSTransformers

Khi xây dựng mạng BGSTransformers dựa trên BGShadowNet, chúng tôi nhận ra ở Giai đoạn I của mạng này, có cấu trúc tương tự với cấu trúc của CBENet. Vì thế, quá trình xây dựng và điều chỉnh cũng thực hiện như ở mạng CBETransformers, tuy nhiên vẫn sẽ có vài khác biệt nhỏ. Cụ thể, ảnh đầu vào sẽ được xử lý thông qua các thành phần như sau:

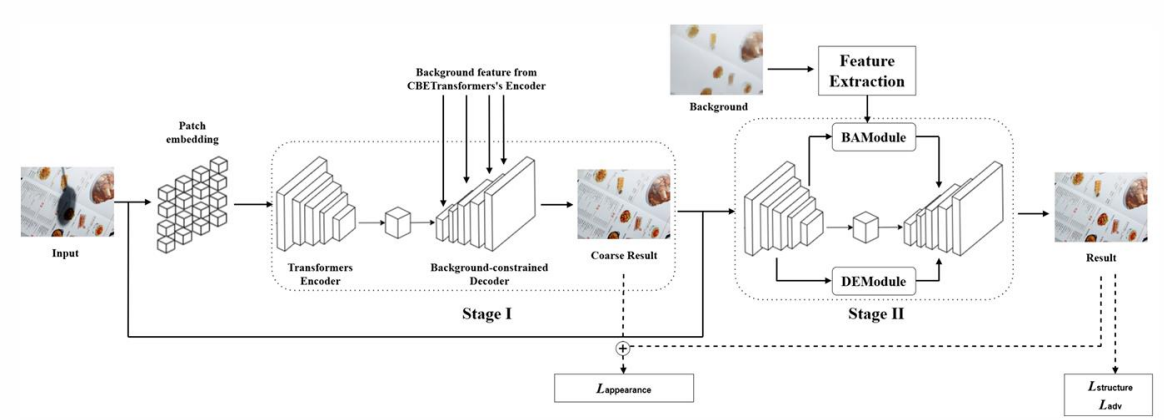
- Patch Embedding: Chia ảnh đầu vào thành các mảng nhỏ (patches) và chuyển đổi chúng thành các vectors embedding. Lớp này sử dụng Conv2D để thực hiện việc chia và nhúng ảnh.
- Transformers Encoder Layer: Một lớp encoder của Vision Transformers bao gồm một Multihead Attention và một Feed Forward Network (FFN) với Dropout và Layer Normalization. Lớp này chịu trách nhiệm học các đặc trưng từ các vectors embedding.
- Transformers Encoder: Tập hợp nhiều lớp Transformers Encoder Layer. Các lớp này kết hợp với nhau để học các đặc trưng phức tạp hơn từ ảnh.

- Tái tạo ảnh: Sau khi học các đặc trưng từ Transformers Encoder, các đặc trưng này được tái tạo lại thành ảnh thông qua lớp Conv2D cuối cùng nằm ở phần Decoder.
- Tại các lớp Decoder này, ngoài việc lấy đặc trưng từ Transformers Encoder, các lớp Decoder cũng tích hợp thêm đặc trưng nền ảnh thu được từ CBETransformers làm các đặc trưng bổ sung để hoàn thiện việc loại bỏ bóng và tái tạo ảnh.

Tương tự với quá trình xây dựng Giai đoạn I, tại giai đoạn II này cũng có các lớp Encoder và Decoder. Do đó việc xây dựng và điều chỉnh thực hiện tương tự như Giai đoạn I. Tuy nhiên, tại giai đoạn II này, ngoài các lớp Encoder và Decoder ra, còn có hai module nhằm hỗ trợ xử lý các đặc trưng (BAModule và DEModule). BAModule sử dụng đặc trưng từ Encoder để hình thành nên một attetion map, giúp loại bỏ sự thiếu nhất quán trong nền ảnh do các yếu tố màu sắc trên nền ảnh gây ra trong quá trình xử lý. DEModule sử dụng đặc trưng từ Encoder trích xuất từ ảnh đầu vào và kết quả thô trong Giai đoạn I để tăng cường chi tiết kết cấu, giúp mô hình tái tạo lại các chi tiết trên và thông tin tài liệu trên ảnh, cũng như cải thiện chất lượng tại các vùng bị bóng đổ che khuất.

Khi điều chỉnh và thay đổi mạng BGShadowNet dựa trên ViTs, chúng tôi đã có một số ý tưởng như thay thế lần lượt hoặc toàn bộ các lớp convolutional bằng ViTs. Cụ thể:

- Ở Giai đoạn I, thay vì chỉ sử dụng các lớp Encoder và Decoder là các lớp convolutional, chúng tôi thay thế và chỉnh sửa chúng bằng các ViTs pretrained, nhằm trích xuất đặc trưng và thu kết quả thô từ ảnh đầu vào rồi mới đưa vào Giai đoạn II tiếp tục xử lý.



Hình 3.3: Minh họa cấu trúc của mạng BGSTransformers

- Ở Giai đoạn II, mạng có hai module là BAM và DEM, trong hai module này chứa khá nhiều phép tính toán, tuy nhiên đa số các phép trong này là nền tảng trong việc giải quyết bài toán. Do đó, không thể tùy tiện biến đổi và thay thế được, chỉ có các lớp Encoder và Decoder (gồm các lớp conv) là khả quan nhất. Vì thế chúng tôi thử thay thế lần lượt các lớp conv bằng ViTs.
- Sau khi thực hiện việc điều chỉnh mô hình bằng ViTs ở cả hai giai đoạn trong BGShadowNet, kết quả cho thấy việc điều chỉnh ở Giai đoạn I mang lại kết quả rất tốt, sau nhiều thử nghiệm khác nhau, chúng tôi nhận thấy việc điều chỉnh ở Giai đoạn I bằng ViTs giúp cho mô hình học được các đặc trưng tốt hơn, do đó giúp mô hình hoạt động tốt hơn, cải thiện kết quả đánh giá mô hình. Tuy nhiên, ngược lại với Giai đoạn I, Giai đoạn II sau khi điều chỉnh lại cho kết quả không tốt, có phần ảnh hưởng tới mô hình và cho hiệu quả kém hơn. Tương tự với việc khi điều chỉnh ở cả hai Giai đoạn, kết quả cũng khá thấp.
- Để tìm hiểu lý do cho vấn đề này, chúng tôi thực hiện nhiều thí nghiệm và nghiên cứu khác nhau. Và cuối cùng rút ra được một số kết luận. Nếu ở Giai đoạn II, thay thế các lớp conv trong Encoder bằng ViTs, các đặc trưng trích xuất ra được phải được xử lý trong một

module phù hợp. Tuy nhiên, tại hai module BAM và DEM, các phép toán trong này khá là phức tạp, do đó chúng tôi không điều chỉnh các module cho phù hợp với các đặc trưng này được. Do sự không phù hợp này, hiệu quả của mô hình bị giảm và khiến kết quả không được tốt.

Vì thế, trong đề tài này, mạng BGSTransformers của chúng tôi chính là từ BGShadowNet chỉ điều chỉnh và cải tiến Giai đoạn I, Giai đoạn II giữ nguyên không đổi gì cả.

Cài đặt mô hình:

Giai đoạn I: Xây dựng các lớp cần thiết như Patch Embedding, Transformers Encoder và Decoder để loại bỏ bóng đổ và tái tạo ảnh. Kết hợp các lớp này để xây dựng Giai đoạn I của mạng BGSTransformers hoàn chỉnh.

Giai đoạn II: Xây dựng tương tự BGShadowNet, Giai đoạn II gồm các lớp Conv và hai module (BAModule và DEModule).

3.2.3. Hàm mất mát (Loss function)

Background reconstruction loss: được sử dụng để ràng buộc CBENet nhằm thu được hình ảnh nền mong muốn. Hàm loss này sử dụng khoảng cách giữa hình ảnh nền $\widehat{B}$ được tạo ra bởi CBENet và hình ảnh nền thực B.

$$\mathcal{L}_{background} = \left\| B - \hat{B} \right\|$$

Appearance consistency loss: đánh giá mất mát dữ liệu giữa các kết quả dự đoán và hình ảnh thực.

$$\mathcal{L}_{appearance} = \lambda_1 \mathcal{L}_{coarse} + \lambda_2 \mathcal{L}_{final} = \lambda_1 \left\| I_{gt} - I_{coarse} \right\| + \lambda_2 \left\| I_{gt} - I_{free} \right\|$$

Structure consistency loss: nhằm mục đích bảo toàn cấu trúc của hình ảnh, với VGG là mạng trích xuất đặc trưng từ mô hình pre-trained VGG19.

$$\mathcal{L}_{structure} = \lambda_3 \left\| VGG(I_{gt}) - VGG(I_{free}) \right\|^2$$

Adversarial loss: được thiết kế cho bộ phân biệt (discriminator) để đánh giá liệu các kết quả được tạo ra là thật hay giả.

$$\mathcal{L}_{adv} = \lambda_4 \mathbb{E}_{(I,I_{free},I_{gt})} \left[\log\left(D\left(I_{gt}\right)\right) + \log\left(1 - D(I)\right) \right]$$

Trong đó λ_1, λ_2, $\lambda_3$và λ_4 là các tham số.

3.3. Quy trình thực hiện

Mô hình được huấn luyện bằng phương pháp tối ưu hóa Adam, với hàm mất mát (loss function) được thiết kế để đánh giá độ tương đồng giữa ảnh đầu ra (ảnh không có bóng) và ảnh mục tiêu.

Học sâu dựa trên GPU được sử dụng để giảm thời gian huấn luyện, và việc kiểm tra thông số được thực hiện thông qua kỹ thuật cross-validation.

Sau khi huấn luyện, mô hình sẽ được thử nghiệm trên một số ảnh tài liệu thực tế không thuộc các bộ dữ liệu được sử dụng trong huấn luyện. Điều này giúp đánh giá khả năng tổng quát hóa của mô hình khi đối mặt với các tình huống thực tế khác nhau.

Mô hình đề xuất được so sánh với các phương pháp loại bỏ bóng trước đây và các kỹ thuật truyền thống. Kết quả được phân tích để chứng minh hiệu quả vượt trội của việc sử dụng ViTs.

Phương pháp nghiên cứu được đề xuất trong khóa luận là sự kết hợp giữa các mô hình loại bỏ bóng hiện có với Vision Transformers, nhằm khai thác lợi thế của ViTs trong việc xử lý thông tin toàn cục. Quy trình huấn luyện và đánh giá chặt chẽ đảm bảo tính chính xác và khả năng áp dụng thực tế của mô hình.

Chương 4. THỬ NGHIỆM VÀ KẾT QUẢ

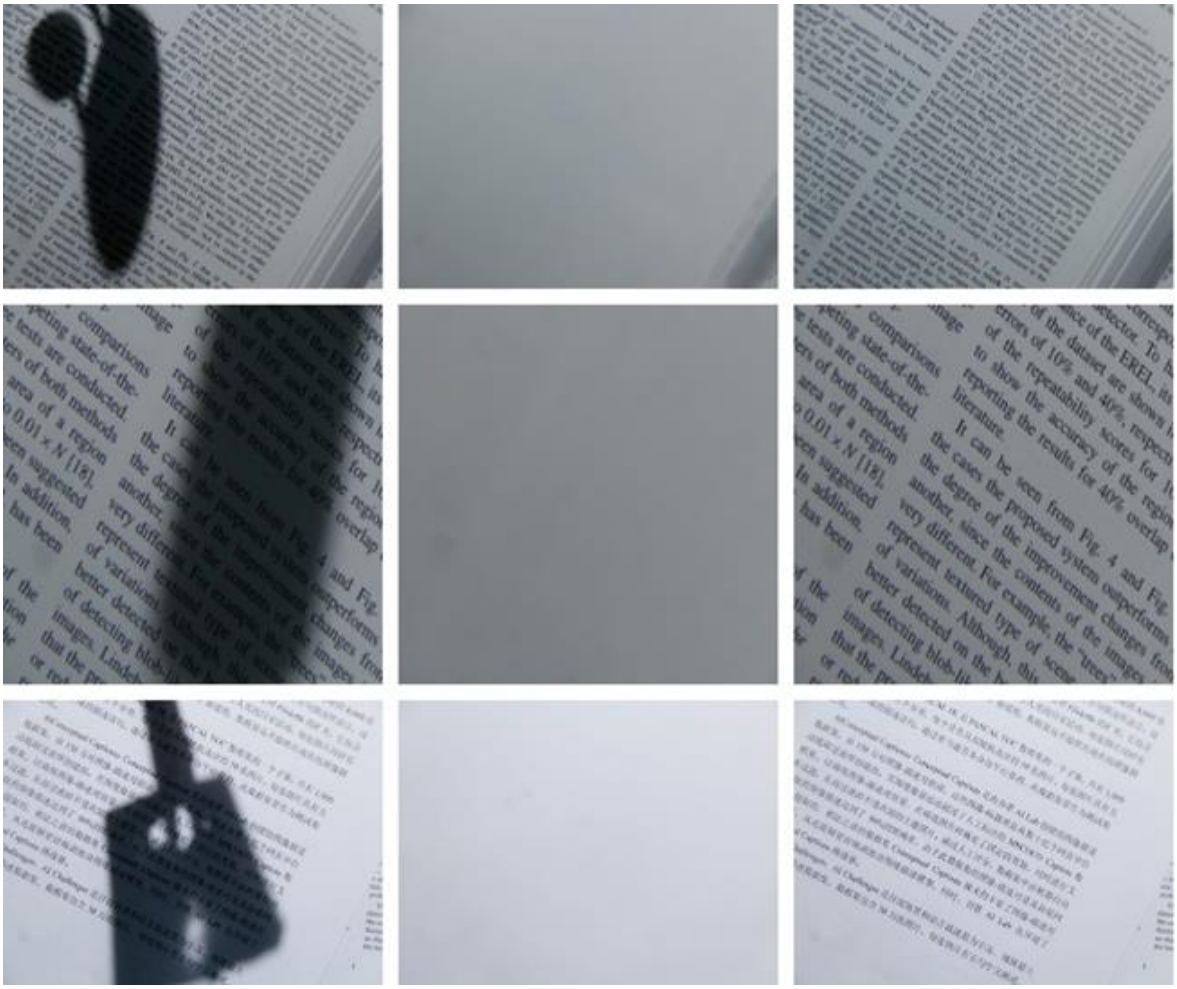
4.1. Thực nghiệm

4.1.1. Bộ dữ liệu

Trong đề tài này, vì để tối ưu kết quả và làm đa dạng dữ liệu cho việc huấn luyện mô hình, chúng tôi sẽ sử dụng ba bộ dữ liệu khác nhau.

Bộ dữ liệu RDD [2] được phát triển nhằm phục vụ cho việc nghiên cứu và đánh giá các phương pháp loại bỏ bóng trên ảnh tài liệu. Bộ dữ liệu này bao gồm 4916 bộ ảnh, trong đó mỗi bộ bao gồm một ảnh chứa bóng, một ảnh nền và một ảnh tài liệu không có bóng. Những bộ ảnh này được thu thập dưới nhiều điều kiện ánh sáng và kiểu bóng khác nhau, giúp đảm bảo tính đa dạng và độ khó trong việc nhận diện và loại bỏ bóng.

- Tập huấn luyện: Bộ dữ liệu RDD có tổng cộng 4371 bộ ảnh dành cho tập huấn luyện, giúp mô hình học được cách phân biệt giữa vùng bóng và vùng tài liệu thông thường.
- Tập kiểm tra: Phần còn lại của bộ dữ liệu, gồm 545 bộ ảnh, được sử dụng cho tập kiểm tra, nhằm đánh giá hiệu quả của mô hình trong việc loại bỏ bóng trên những ảnh chưa từng thấy trước đó.
- Bộ dữ liệu RDD có quy mô lớn và phong phú về điều kiện thu thập dữ liệu, điều này giúp đảm bảo tính khách quan và khả năng tổng quát hóa của mô hình khi triển khai trên các bộ dữ liệu khác.



Hình 4.1: Một số ảnh trong bộ dữ liệu RDD, được chia làm 3 phần, ảnh chứa bóng (trái), nền ảnh (giữa) và ảnh không có bóng (phải)

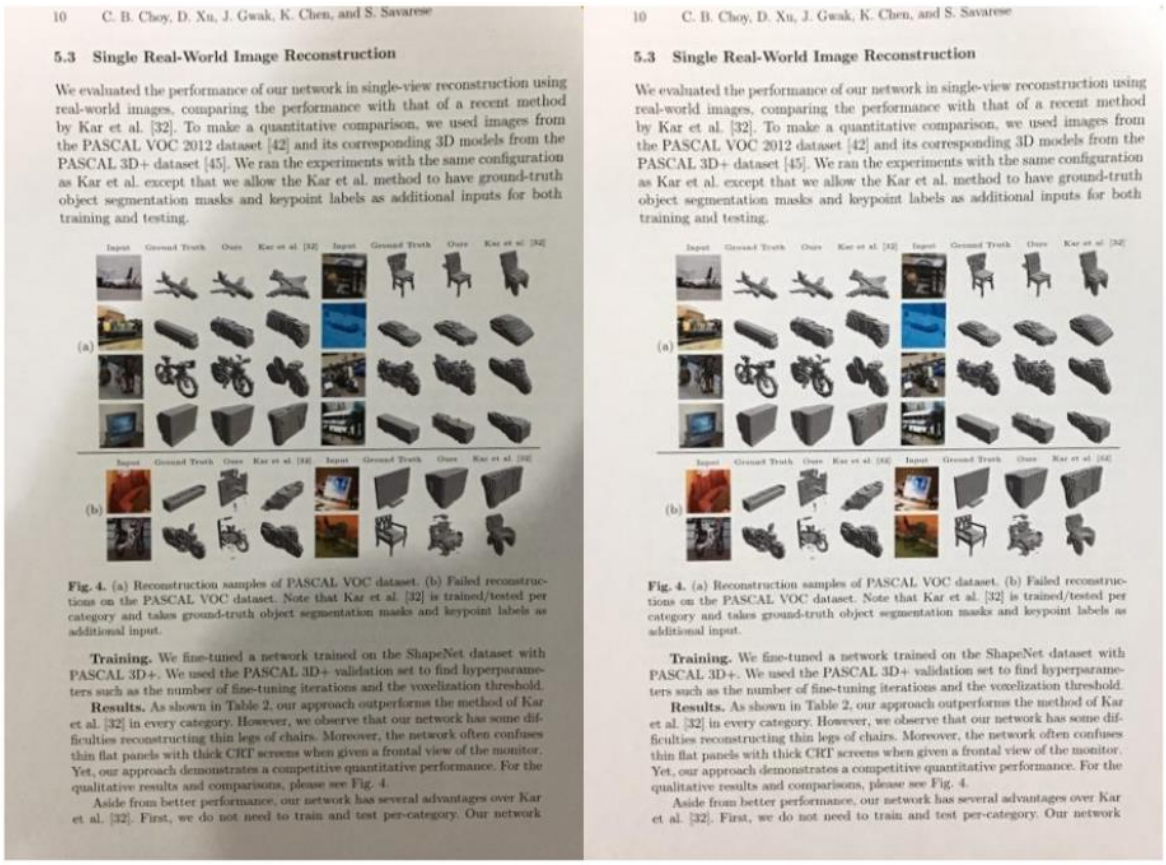
Bộ dữ liệu Jung [3] là một tập hợp gồm 159 ảnh tài liệu bị biến dạng do ánh sáng không đều. Các ảnh này được chụp hoặc quét từ các tài liệu thực tế, do đó, chúng thể hiện một cách chân thực những vấn đề gặp phải trong các ứng dụng xử lý tài liệu hàng ngày, đặc biệt là trong các hệ thống nhận dạng ký tự quang học (OCR).

- Ảnh chụp từ camera: Trong số 159 ảnh, có 109 ảnh được chụp bằng camera của hai điện thoại thông minh khác nhau, điều này giúp mô hình có thể học cách xử lý các loại bóng và điều kiện ánh sáng đa dạng từ các thiết bị phổ biến. Kích thước của các ảnh chụp từ camera

là 3264 × 2448 pixel, đảm bảo độ phân giải cao để giữ lại chi tiết của tài liệu.

- Ảnh quét từ máy scanner: Bên cạnh đó, có 50 ảnh được quét từ tài liệu bằng máy quét với độ phân giải 72 dpi, kích thước ảnh là 3455 × 2464 pixel. Ảnh quét thường có đặc điểm khác với ảnh chụp bởi camera, chẳng hạn như không bị méo hình hoặc ảnh hưởng từ ánh sáng xung quanh, nhưng vẫn có thể bị ảnh hưởng bởi các nguồn sáng trực tiếp hoặc gián tiếp.

Bộ dữ liệu Jung thể hiện rõ những tình huống thực tế mà mô hình loại bỏ bóng cần phải xử lý, bao gồm cả các điều kiện ánh sáng phức tạp và các kiểu bóng khác nhau từ nhiều nguồn sáng.



10 C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese

5.3 Single Real-World Image Reconstruction

We evaluated the performance of our network in single-view reconstruction using real-world images, comparing the performance with that of a recent method by Kar et al. [32]. To make a quantitative comparison, we used images from the PASCAL VOC 2012 dataset [42] and its corresponding 3D models from the PASCAL 3D+ dataset [45]. We ran the experiments with the same configuration as Kar et al. except that we allow the Kar et al. method to have ground-truth object segmentation masks and keypoint labels as additional inputs for both training and testing.

Fig. 4. (a) Reconstruction samples of PASCAL VOC dataset. (b) Failed reconstructions on the PASCAL VOC dataset. Note that Kar et al. [32] is trained/tested per category and takes ground-truth object segmentation masks and keypoint labels as additional input.

Training. We fine-tuned a network trained on the ShapeNet dataset with PASCAL 3D+. We used the PASCAL 3D+ validation set to find hyperparameters such as the number of fine-tuning iterations and the voxelization threshold.

Results. As shown in Table 2, our approach outperforms the method of Kar et al. [32] in every category. However, we observe that our network has some difficulties reconstructing thin legs of chairs. Moreover, the network often confuses thin flat panels with thick CRT screens when given a frontal view of the monitor. Yet, our approach demonstrates a competitive quantitative performance. For the qualitative results and comparisons, please see Fig. 4.

Aside from better performance, our network has several advantages over Kar et al. [32]. First, we do not need to train and test per-category. Our network

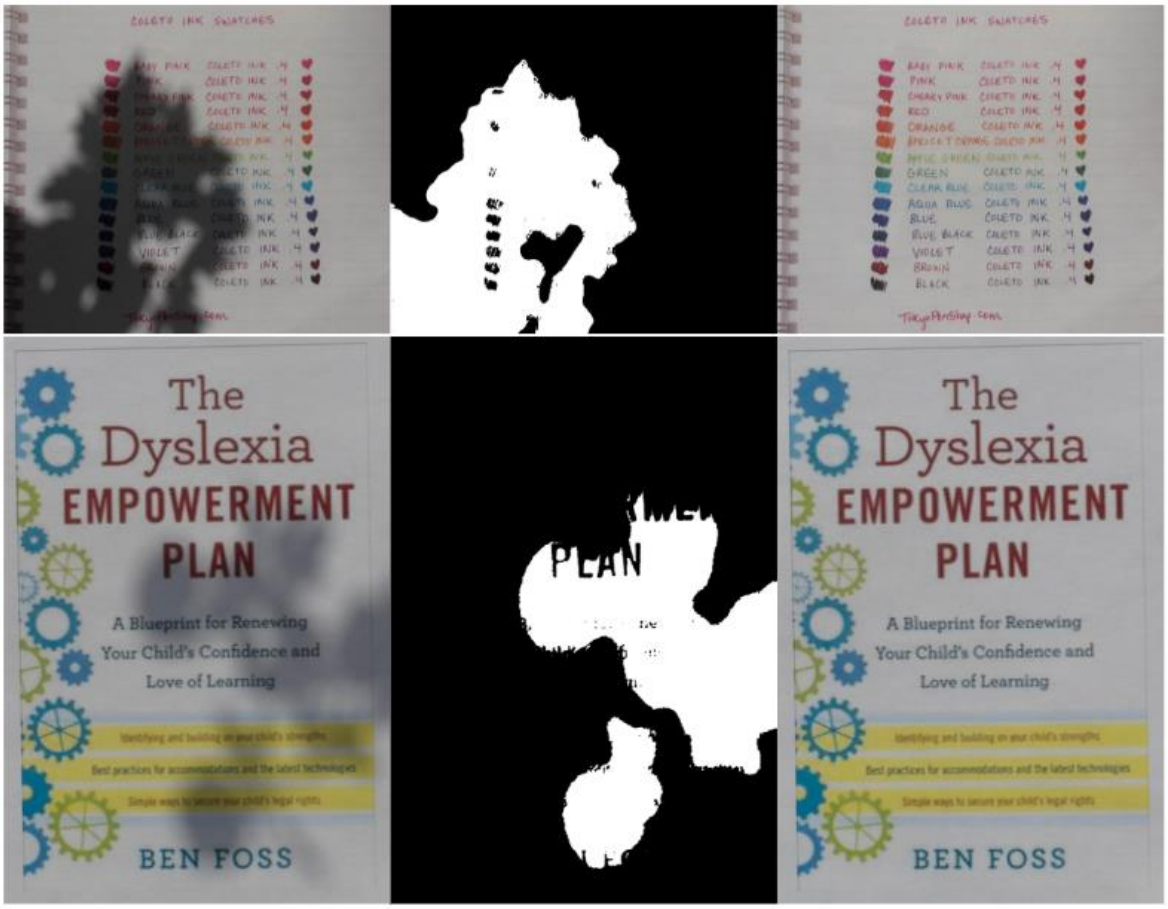
Hình 4.2: Một số ảnh trong bộ dữ liệu Jung, gồm 2 phần, ảnh chứa bóng (phải) và ảnh không chứa bóng (trái)

Bộ dữ liệu Kliger [4] bao gồm 300 cặp ảnh, trong đó mỗi cặp bao gồm một ảnh tài liệu có bóng, một mặt nạ bóng và một ảnh tài liệu đã được loại bỏ bóng. Bộ dữ liệu này chủ yếu tập trung vào các kiểu bóng khác nhau trên ảnh tài liệu, với sự đa dạng về hình dạng, kích thước và độ đậm nhạt của bóng.

- Tập huấn luyện: Bộ dữ liệu Kliger cung cấp 272 cặp ảnh trong tập huấn luyện, giúp mô hình có đủ dữ liệu để học cách phân biệt bóng với các vùng tài liệu không bị ảnh hưởng.
- Tập kiểm tra: Tập kiểm tra gồm 28 cặp ảnh còn lại, được sử dụng để đánh giá hiệu quả của mô hình sau khi huấn luyện. Tập kiểm tra này nhỏ hơn so với tập kiểm tra của RDD, nhưng vẫn đủ để đánh giá khả

năng tổng quát hóa của mô hình đối với các điều kiện bóng khác nhau.

Điểm đáng chú ý của bộ dữ liệu Kliger là việc cung cấp thêm mặt nạ bóng (shadow mask), giúp mô hình có thêm thông tin trực tiếp về vị trí và hình dạng của bóng, điều này có thể giúp cải thiện độ chính xác khi phân tách vùng bóng với vùng không có bóng.



Hình 4.3: Một số ảnh trong bộ dữ liệu Kliger, gồm 3 phần, ảnh chứa bóng (trái), mặt nạ bóng (giữa) và ảnh không chứa bóng (phải)

4.1.2. Tiền xử lý

Trong ba bộ dữ liệu mà chúng tôi sử dụng, chỉ có bộ dữ liệu RDD là đáp ứng đủ yêu cầu để có thể đưa vào huấn luyện trực tiếp cho mô hình (bộ dữ liệu đủ ba phần ảnh có bóng đổ, nền của ảnh và ảnh không có bóng đổ). Do đó, để có thể sử dụng hai bộ dữ liệu còn lại đem đi huấn luyện cho mô hình, chúng tôi đã tiến hành tiền xử lý đối với hai bộ dữ liệu Jung và Kliger.

Để thực hiện việc thu thập nền ảnh nhằm làm nền ảnh ground truth cho hai bộ dữ liệu Jung và Kliger, chúng tôi sử dụng chiến lược local-to-global (từ cục bộ đến toàn cục).

- Ảnh ground truth không có bóng được chia thành 16x16 patch bằng nhau nhằm mục đích tính toán nền cho từng patch.Chúng tôi gọi nền của các patch này là nền cục bộ $\overline{B}$.
- Đối với mỗi patch, mô hình Gaussian Mixture Model (GMM) được sử dụng để chia các patch thành hai cụm dựa trên cường độ sáng (pixel intensity). Hai cụm này thường tương ứng với: Văn bản (text content - những vùng tối hơn) và Nền (background - những vùng sáng hơn).
- Nhận thấy rằng nền của tài liệu thường sáng hơn nội dung văn bản, nhóm có cường độ sáng cao hơn được coi là nền. Sau đó, tính giá trị trung bình màu của nhóm nền này và sử dụng làm màu nền cho toàn bộ các patch.
- Do mỗi patch được xử lý độc lập, màu nền của các patch thường không đồng nhất, tạo ra các đường ranh giới rõ ràng giữa các patch (patch boundaries).
- Để giảm thiểu hiện tượng ranh giới này và tạo ra một nền liên tục, mượt mà, một bộ lọc làm mịn bảo toàn màu sắc [28] (color-preserving smoothing operator) được áp dụng lên $\overline{B}$. Bộ lọc này lấy cảm hứng từ bộ lọc có hướng dẫn (guided filter), được sử dụng để làm mượt hình ảnh mà vẫn bảo toàn các chi tiết cạnh (edge-preserving).

- Sau khi làm mịn, hình ảnh nền chân thực B được tạo ra và được sử dụng làm ground truth để huấn luyện mô hình.
- Giá trị của pixel i trong B có thể biểu thị bằng:

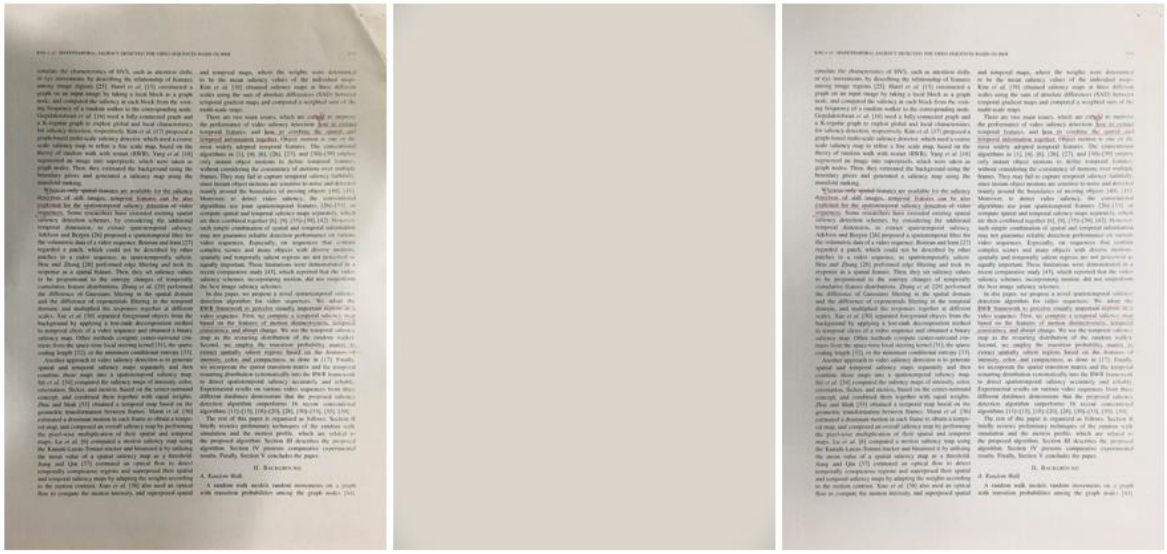
$$B_i = \sum_{j \in N(i)} W_{ij} \bar{B}_j$$

Trong đó:

N(i) là một láng giềng cục bộ (local neighborhood) cảu pixel i

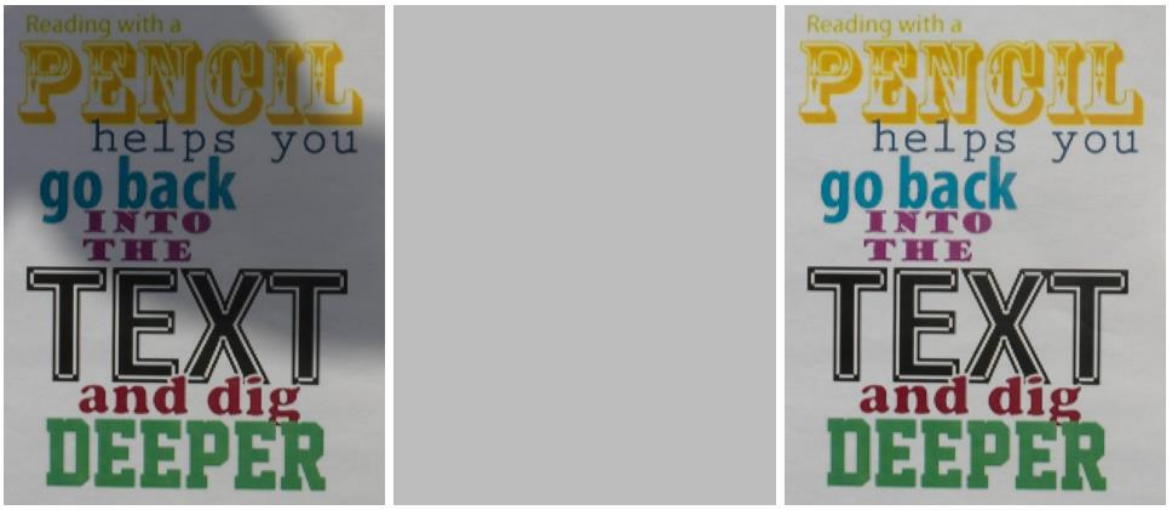
W_{ij} là bộ lọc tính toán độ tương tự màu sắc giữa i và j

Đối với bộ dữ liệu Jung: bộ dữ liệu chỉ gồm có hai phần là ảnh có và không có bóng đổ. Chúng tôi sẽ tiến hành trích xuất nền ảnh từ các ảnh có sẵn trong bộ dữ liệu nhằm cung cấp đủ dữ liệu tiến hành huấn luyện. Việc trích xuất nền ảnh được chúng tôi sử dụng mạng phương pháp như trên để thực hiện. Sau đó chúng tôi tiến hành dò lại thủ công để kiểm tra và đảm bảo rằng tất cả các ảnh trong bộ dữ liệu đều đã được trích xuất nền một cách chính xác, đối với các nền ảnh bị lỗi, chúng tôi sẽ chỉnh sửa thủ công nhằm thu được nên ảnh hoàn chỉnh và chất lượng nhất.



Hình 4.4: Minh họa bộ dữ liệu Jung sau quá trình tiền xử lý, có thêm phần nền ảnh (giữa)

Đối với bộ dữ liệu Kliger: bộ dữ liệu bao gồm ba phần là ảnh có và không có bóng đổ kèm thêm là phần mặt nạ bóng (shadow mask). Để mô hình đáp ứng đủ điều kiện nhằm huấn luyện được mô hình, chúng tôi tiến hành loại bỏ phần dữ liệu mặt nạ bóng, và thay vào đó là trích xuất nền ảnh, việc trích xuất nền ảnh được thực hiện tương tự khi xử lý bộ dữ liệu Jung.



Hình 4.5: Minh họa bộ dữ liệu Kliger sau quá trình tiền xử lý, có thêm phần nền ảnh (giữa) và loại bỏ phần mặt nạ bóng

Qua quá trình tiền xử lý dữ liệu, chúng tôi thu được ba bộ dữ liệu đáp ứng đủ điều kiện để sẵn sàng đưa vào huấn luyện cho mô hình của mình.

4.1.3. Cài đặt và thử nghiệm

Mô hình của chúng tôi được xây dựng và triển khai trên framework là Pytorch.

Sau nhiều tìm hiểu và nghiên cứu, cũng như qua nhiều lần thử nghiệm, chúng tôi đã tiếm hành cài đặt mô hình như sau. Hai mạng CBE và BGShadow được huấn luyện riêng biệt với nhau, sau đó mới tổng hợp thành một mô hình duy nhất. Mạng CBE sẽ được huấn luyện trước với số epoch là 200, sau đó mạng

CBE sẽ được điều chỉnh và chỉnh sửa cho phù hợp để có thể sử dụng dữ liệu đầu ra từ mạng này làm đặc trưng bổ sung nhằm huấn luyện mạng BGShadow. Tiếp đến, mạng BGShadow cũng được huấn luyện với số epoch là 200.

Chúng tôi sử dụng optimizer là Adam nhằm tối ưu mô hình của mình với tốc độ suy giảm (attenuation rate) là betas = (0.5, 0.999).

Tốc độ học (learning rate) ban đầu được chúng tôi cài đặt là 0.0004, sau vài lần thử nghiệm và chỉnh sửa, chúng tôi quyết định giảm learning rate xuống còn 0.00039 và đạt được kết quả tốt hơn. Vì thế, learning rate của mô hình chúng tôi quyết định giữ ở 0.00039.

Một số tham số khác được chúng tôi thiết lập như λ_1, λ_2, λ_3 và λ_4 lần lượt là 1, 1, 0.05 và 0.01.

4.2. Độ đo

Root Mean Squared Error (RMSE): căn bậc hai của MSE, giúp giữ đơn vị đo lường giống với giá trị thực tế và giá trị dự đoán. Trong bối cảnh xử lý ảnh, RMSE đo lường sự khác biệt giữa ảnh gốc và ảnh đã xử lý (ảnh đã loại bỏ bóng). RMSE càng nhỏ, chất lượng phục hồi càng cao.

$$RMSE = \sqrt{MSE}$$

Trong đó:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(I_{pred}(i) - I_{gt}(i))^2$$

I_{pred} là giá trị pixel của ảnh dự đoán.

I_{gt} là giá trị pixel của ảnh thực.

n là tổng số pixel.

Peak Signal-to-Noise Ratio (PSNR): độ đo chất lượng ảnh được sử dụng để so sánh chất lượng của ảnh nén hoặc tái tạo so với ảnh gốc. PSNR tính toán tỉ lệ giữa

giá trị cường độ tối đa của tín hiệu và mức độ nhiễu ảnh hưởng đến chất lượng tín hiệu. PSNR cao cho thấy mức độ nhiễu thấp và chất lượng hình ảnh cao.

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right)$$

Trong đó:

MAX là giá trị tối đa của pixel trong ảnh (thường là 255 đối với ảnh 8-bit)

Structural Similarity Index Measurement (SSIM): độ đo để đánh giá sự tương đồng giữa hai ảnh, xem xét các yếu tố về cấu trúc, độ sáng và độ tương phản. SSIM phản ánh sự tương đồng cấu trúc giữa ảnh gốc và ảnh đã xử lý. SSIM cao cho thấy sự tương đồng lớn giữa ảnh dự đoán và ảnh thực.

$$SSIM(x, y) = [l(x, y)]^{\alpha}.[c(x, y)]^{\beta}.[s(x, y)]^{\gamma}$$

Trong đó:

x, y là ảnh dự đoán và ảnh thực

l(x, y) là thành phần độ sáng

c(x, y) là thành phần độ tương phản

s(x, y) là thành phần cấu trúc

α, β, γ thường được đặt bằng 1

4.3. Đánh giá và phân tích kết quả

4.3.1. Kết quả đánh giá

Để kết quả đánh giá được trực quan nhất, chúng tôi quyết định sẽ thể hiện kết quả đánh giá mô hình ở cả ba bộ dữ liệu RDD, Jung và Kliger. Các mô hình được đem ra đánh giá và so sánh là mô hình của chúng tôi, BGShadowNet [2], BEDSR-Net [4] và Water-Filling [3].

Với độ đo RMSE càng thấp thì hiệu quả càng cao và ngược lại với hai độ đo PSNR và SSIM, hiệu quả càng cao thì hai chỉ số này càng cao.

	RMSE	PSNR	SSIM
BGShadowNet	2.372	36.301	0.975
BEDSR-Net	2.963	34.839	0.971
Water-Filling	29.612	17.835	0.887
Proposed method	**2.277**	**36.624**	**0.977**

Bảng 4.1: Kết quả đánh giá và so sánh các mô hình trên bộ dữ liệu RDD sử dụng ba độ đo RMSE, PSNR và SSIM

	RMSE	PSNR	SSIM
BGShadowNet	5.286	29.022	0.937
BEDSR-Net	6.423	28.792	0.935
Water-Filling	28.379	13.468	0.893
Proposed method	**5.138**	**29.156**	**0.943**

Bảng 4.2: Kết quả đánh giá và so sánh các mô hình trên bộ dữ liệu Kliger sử dụng ba độ đo RMSE, PSNR và SSIM

	RMSE	PSNR	SSIM
BGShadowNet	4.212	33.526	0.965
BEDSR-Net	5.937	32.928	0.955
Water-Filling	30.213	14.665	0.853
Proposed method	**4.133**	**33.895**	**0.971**

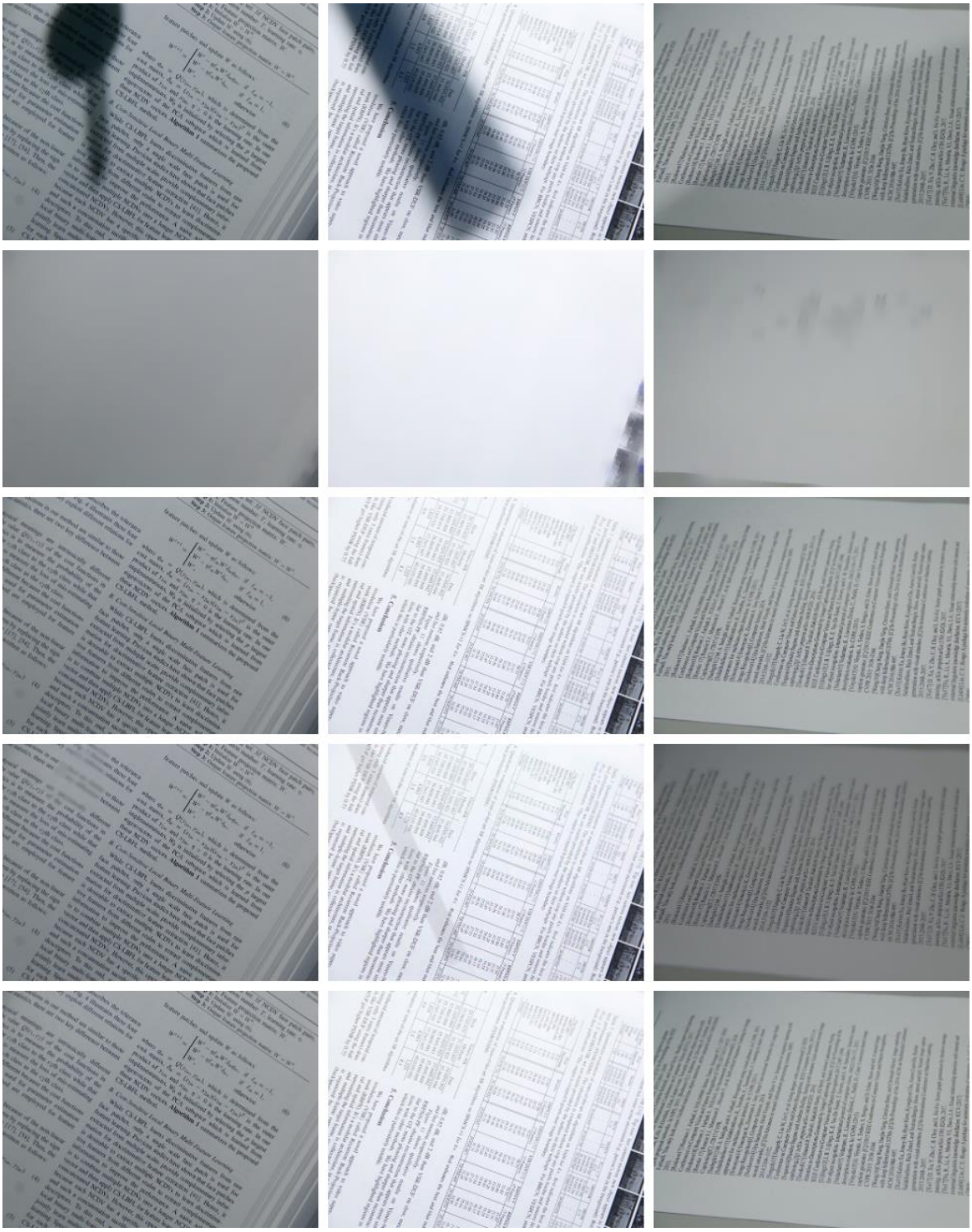
Bảng 4.3: Kết quả đánh giá và so sánh các mô hình trên bộ dữ liệu Jung sử dụng ba độ đo RMSE, PSNR và SSIM

Trong đó, ‘My method’ là mô hình của chúng tôi và 3 mô hình còn lại là các mô hình của các phương pháp đạt kết quả khá tốt trước đây

Qua kết quả sơ bộ cho thấy, phương pháp của chúng tôi (CBENet và BGShadowNet dựa trên ViTs) hoạt động và mang lại hiệu quả cao hơn các phương pháp trước đây.

4.3.2. Một số trường hợp có kết quả được cải thiện

Dưới đây là hình ảnh một số trường hợp mà kết quả từ mô hình cảu chúng tôi đạt hiệu quả cao hơn so với phương pháp được sử dụng làm xương sống (backbone) – ‘loại bỏ bóng đổ trên ảnh tài liệu sử dụng cơ chế nền nhận biết màu sắc’ [2].



Hình 4.6: Một số trường hợp có kết quả được cải thiện so với mô hình BGShadowNet

Ngoài ra còn một số trường hợp khác không liệt kê hết được.

4.3.3. So sánh các mô hình được triển khai

Chúng tôi đã tiến hành điều chỉnh mô hình theo bốn hướng khác nhau nhằm tìm ra phương án tối ưu nhất.

Bộ dữ liệu được chúng tôi sử dụng để so sánh ở đây là bộ dữ liệu RDD [2].

- **CBENet + BGShadowNet (1)**

Ở model (1) này, chúng tôi tiến hành train và đánh giá với mô hình được đề xuất trong bài báo "Document Image Shadow Removal guided by Color-aware Background".

Model gồm có hai mạng là CBENet và BGShadowNet. Các tham số cũng như hàm mất mát (loss function) giữ nguyên, không thay đổi gì cả.

- **CBETransformer + BGSTransformer (2)**

Ở model (2), chúng tôi điều chỉnh lại hai mạng CBE và BGShadow dựa trên ViTs trước khi huấn luyện. Cụ thể:

CBETransformers: là mạng CBENet nhưng được xây dựng lại dựa trên ViTs thay vì U-net. Quá trình phân chia ảnh thành các patchs và trích xuất đặc trưng nền ảnh thực hiện dựa trên Patch Embedding và Transformer Encoder.

BGSTransformers: Trong quá trình xử lý và điều chỉnh mạng này, chúng tôi tiến hành thay thế lần lượt các lớp convolutional trong mạng BGShadowNet bằng ViTs rồi tiến hành huấn luyện. Quá trình này tốn khá nhiều tài nguyên, do đó chúng tôi đã dùng các ViTs pre-trained để thử nghiệm và đánh giá. Cụ thể, phương pháp đã tiến hành điều chỉnh ở cả Giai đoạn I và Giai đoạn II.

Ở Giai đoạn I, cách điều chỉnh thực hiện tương tự với mạng CBE, nhưng thay vì chỉ tập trung vào xử lý nền ảnh, thì ở giai đoạn này sẽ xử lý toàn bộ ảnh đầu vào, với Encoder thay thế bằng Transformer Encoder và Decoder có sự kết hợp các đặc trưng lấy từ Encoder của mạng CBE.

Ở Giai đoạn II, chúng tôi đã thử thay thế lần lượt cũng như toàn bộ các lớp conv bằng ViTs và tiến hành huấn luyện và thử nghiệm. Tuy nhiên, kết quả thu

được lại không khả quan, và cho thấy cách làm này không cải tiến được mô hình mà còn làm hạ thấp hiệu quả. Do đó, ở mạng BGS, chúng tôi chỉ thực hiện chỉnh sửa và thay đổi ở Giai đoạn I, Giai đoạn II không thay đổi gì cả.

- **CBETransformer + BGShadowNet (3)**

Ở model (3), quá trình cài đặt ở hai mạng CBE và BGS thực hiện tương tự như hai model (1) và (2)

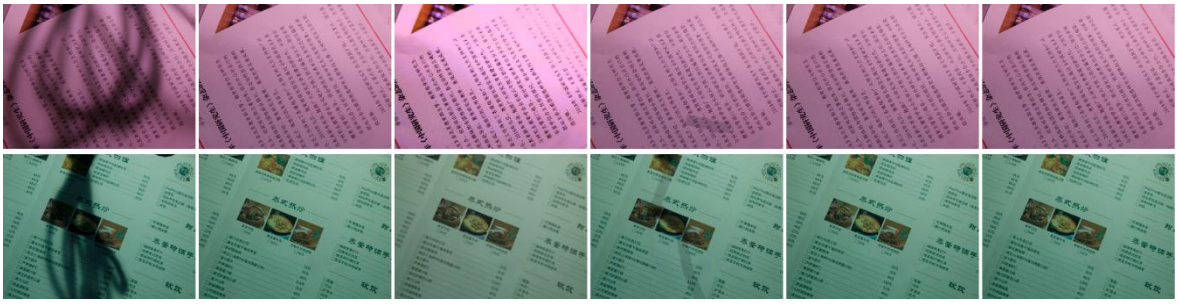
- **CBENet + BGSTransformer (4)**

Ở model (4) này, quá trình cài đặt cảu chúng tôi sẽ tương ứng như các model trước đó, mạng CBE sẽ giữ nguyên, còn mạng BGS sẽ được tiến hành điều chỉnh dựa trên ViTs.

Mô hình	RMSE	PSNR	SSIM
(1)	2.372	36.301	0.975
(2)	**2.277**	**36.624**	**0.977**
(3)	2.583	36.263	0.972
(4)	2.639	36.077	0.968

Bảng 4.4: So sánh hiệu quả mô hình sau khi điều chỉnh các mạng với ViTs

Thông qua kết quả thu được, mô hình số (2) CBETransformer + BGSTransformer đem lại kết quả cao nhất, do đó, chúng tôi sử dụng mô hình này làm mô hình chính cho đề tài, và so sánh với các phương pháp trước đây.



Hình 4.7: So sánh các mô hình do chúng tôi triển khai, lần lượt từ trái qua phải là ảnh chứa bóng, ảnh ground truth, và kết quả các mô hình (3), (4), (1), (2)

4.3.4. Nhận xét và phân tích

Ba bộ dữ liệu RDD [2], Jung [3] và Kliger [4] cung cấp dữ liệu đa dạng gồm nhiều ảnh tài liệu khác nhau với nhiều góc độ và hình dạng bóng đổ, giúp làm đa dạng về dữ liệu huấn luyện, góp phần giúp mô hình được huấn luyện đầy đủ các trường hợp và tránh bị overfitting.

Các độ đo được chúng tôi sử dụng (RMSE, PSNR và SSIM) là các độ đo phổ biến sử dụng trong bài toán loại bỏ bóng trên ảnh tài liệu. Các độ đo này thể hiện các mức độ tương đồng hay khác biệt giữa ảnh kết quả và ảnh tham chiếu (ground truth), hoặc thể hiện mức độ rõ nét, chất lượng hình ảnh của ảnh kết quả.

Trong quá trình tích hợp ViTs với mô hình ‘loại bỏ bóng đổ dựa trên nền nhận biết màu sắc’, chúng tôi thực hiện thử nghiệm trên nhiều trường hợp khác nhau, điều chỉnh và cải tiến các giai đoạn khác nhau của các mạng trong mô hình. Từ kết quả thực nghiệm của chúng tôi thu được cho thấy, kết quả khi thực hiện điều chỉnh ở cả hai mạng CBE và BGShadow (Giai đoạn I) cho kết quả cải thiện rõ ràng nhất. Và ngược lại, khi điều chỉnh bằng ViTs ở giai đoạn II của mạng BGShadow, kết quả thu được rất kém và có phần làm giảm hiệu quả mô hình. Do đó, khi so sánh, chúng tôi chỉ sử dụng một số trường hợp mà chúng tôi đạt kết quả có thể chấp nhận để so sánh với các mô hình của các phương pháp trước đây.

Với hướng điều chỉnh và huấn luyện mô hình ‘loại bỏ bóng đổ dựa trên nền nhận biết màu sắc’ bằng ViTs, hiệu suất của mô hình được cải thiện một cách rõ rệt. Dựa trên kết quả thu được ta có thể thấy .

ViTs góp phần cải thiện quá trình trích xuất nền ảnh, cũng như quá trình loại bỏ bóng đổ. Dựa trên kết quả thực nghiệm, cả hai mạng CBE và BGShadow nếu đều được điều chỉnh dựa trên ViTs thì kết quả mô hình cải thiện được nhiều nhất.

Qua các kết quả thực nghiệm này, đã chứng minh được tiềm năng và khả năng của ViTs trong các mô hình loại bỏ bóng đổ trên ảnh tài liệu, góp phần cải thiện hiệu suất, ngày càng hoàn thiện bài toán hơn.

Chương 5. KẾT LUẬN VÀ ĐỀ XUẤT

5.1. Kết luận

Trong khóa luận này, chúng tôi đã nghiên cứu và đề xuất phương pháp cải tiến hai mô hình loại bỏ bóng trên ảnh tài liệu, CBENet và BGShadowNet [2], bằng cách tích hợp Vision Transformers (ViTs) [1] thay cho các lớp tích chập truyền thống. Phương pháp này tận dụng khả năng xử lý thông tin toàn cục của ViTs, giúp cải thiện hiệu quả trong việc loại bỏ bóng, đặc biệt với các trường hợp bóng phức tạp.

Các thí nghiệm được thực hiện trên ba bộ dữ liệu tiêu chuẩn, bao gồm RDD, Jung, và Kliger, đã chứng minh rằng phương pháp đề xuất đạt được hiệu quả cao hơn so với các phương pháp truyền thống và một số mô hình dựa trên mạng nơ-ron tích chập (CNN). Kết quả được đánh giá thông qua các chỉ số RMSE, PSNR và SSIM, cho thấy mô hình đề xuất không chỉ đạt độ chính xác cao mà còn có khả năng tổng quát hóa tốt đối với các tình huống thực tế.

5.2. Đóng góp và nghiên cứu

Nghiên cứu này đã mang lại những đóng góp chính như sau:

Đề xuất phương pháp tích hợp ViTs: Lần đầu tiên, Vision Transformers được tích hợp vào các mô hình CBENet và BGShadowNet để cải thiện hiệu quả loại bỏ bóng trên ảnh tài liệu. Điều này mở ra hướng nghiên cứu mới trong việc ứng dụng ViTs vào các bài toán xử lý ảnh tài liệu.

Đánh giá toàn diện trên nhiều bộ dữ liệu: Phương pháp được kiểm chứng trên ba bộ dữ liệu tiêu chuẩn với độ phức tạp khác nhau, đảm bảo tính khách quan và khả năng áp dụng thực tế của mô hình.

Tăng cường khả năng tổng quát hóa: Việc sử dụng ViTs giúp mô hình xử lý tốt hơn các trường hợp bóng không đồng nhất và các điều kiện ánh sáng khác nhau, nâng cao độ tin cậy khi triển khai trong thực tế.

5.3. Hạn chế

Mặc dù đạt được những kết quả tích cực, nghiên cứu này vẫn còn một số hạn chế như sau:

Chi phí tính toán cao: Vision Transformers yêu cầu tài nguyên tính toán lớn, đặc biệt trong quá trình huấn luyện, điều này có thể gây khó khăn khi triển khai trên các thiết bị hạn chế về phần cứng.

Hiệu quả trên dữ liệu không đồng nhất: Dù mô hình đạt kết quả tốt trên các bộ dữ liệu tiêu chuẩn, nhưng hiệu quả trên các tài liệu không thuộc các bộ dữ liệu này có thể không ổn định do tính đa dạng trong cấu trúc và điều kiện ánh sáng của dữ liệu thực tế.

Phụ thuộc vào cấu hình siêu tham số: Hiệu suất của mô hình phụ thuộc nhiều vào việc lựa chọn các siêu tham số, điều này đòi hỏi nhiều thời gian và công sức trong quá trình tối ưu hóa.

5.4. Đề xuất hướng phát triển

Để khắc phục các hạn chế và tiếp tục phát triển nghiên cứu, một số hướng đề xuất được đưa ra như sau:

Tối ưu hóa mô hình: Nghiên cứu các phương pháp giảm thiểu chi phí tính toán của Vision Transformers, chẳng hạn như sử dụng các biến thể nhẹ hơn của ViTs hoặc kỹ thuật nén mô hình.

Mở rộng bộ dữ liệu: Thu thập thêm các bộ dữ liệu tài liệu thực tế với đa dạng về cấu trúc và điều kiện ánh sáng để huấn luyện và đánh giá mô hình, đảm bảo khả năng tổng quát hóa tốt hơn.

Kết hợp đa nhiệm (multi-task learning): Tích hợp thêm các nhiệm vụ khác, chẳng hạn như chỉnh sửa biến dạng tài liệu hoặc nâng cao độ phân giải, để tăng giá trị ứng dụng của mô hình.

Nghiên cứu kết hợp mô hình lai: Kết hợp Vision Transformers với các mô hình CNN hoặc các phương pháp khác để tận dụng ưu điểm của cả hai, đồng thời cải thiện hiệu suất và giảm chi phí tính toán.

Triển khai thực tế: Nghiên cứu cách tích hợp mô hình vào các ứng dụng thực tiễn, chẳng hạn như phần mềm OCR hoặc các hệ thống quản lý tài liệu, nhằm kiểm chứng khả năng ứng dụng trong môi trường thực tế.

TÀI LIỆU THAM KHẢO

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[2] Qing Zhang, Zheng Liu, Xiaolong Zhang, Chunxia Xiao, Ling Zhang, Yinghao He, "Document image shadow removal guided by color-aware background," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* June 2023.

[3] Seungjun Jung, Muhammad Abul Hasan, and Changick Kim, "Water-filling: An efficient algorithm for digitized document shadow removal," *Asian Conference on Computer Vision,* pp. 398-414, 2018.

[4] Y. H. Lin, W. C. Chen, and Y. Y. Chuang, "BEDSR-Net: A deep shadow removal network from a single document image," *CVPR,* pp. 12905-12914, 2020.

[5] E. Arbel and H. Hel-Or, "1, 9, 25Shadow removal using intensity surfaces and texture anchor points," *IEEE Transactions on Pattern Analysis Machine Intelligence,* pp. 1202-1216, 2011.

[6] G. D. Finlayson, M. S. Drew, and C. Lu, " On the removal of shadows from images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 59-68, 2006.

[7] Feng Liu and Michael Gleicher, "Texture-consistent shadow removal," *ECCV,* pp. 437-450, 2008.

[8] Yael Shor and Dani Lischinski, "The shadow meets the mask: Pyramid-based

shadow removal," pp. 577-568, 2008.

[9] Chunxia Xiao, Ruiyun She, Donglin Xiao, and Kwan Liu Ma, "Fast shadow removal using adaptive multi-scale illumination transfer," *Computer Graphics Forum,* pp. 207-218, 2013.

[10] Chunxia Xiao, Donglin Xiao, Ling Zhang, and Lin Chen, "Efficient shadow removal using subregion matching illumination transfer," *Computer Graphics Forum,* pp. 421-430, 2013.

[11] Ling Zhang, Qing Zhang, and Chunxia Xiao, "Shadow remover: Image shadow removal based on illumination recovering optimization," *IEEE Transactions on Image Processing,* pp. 4623-4636, 2015.

[12] Xiaodong Cun, Chi-Man Pun, and Cheng Shi, "Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan," *AAAI,* pp. 10680-10687, 2020.

[13] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang, "Autoexposure fusion for single-image shadow removal," *CVPR,* pp. 10571-10580, 2021.

[14] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and PhengAnn Heng, "Direction-aware spatial context features for shadow detection and removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 2795-2808, 2020.

[15] X. Hu, Y. Jiang, C. W. Fu, and P. A. Heng, "Mask-shadowgan: Learning to remove shadows from unpaired data," *ICCV,* pp. 2472-2481, 2019.

[16] Yeying Jin, Aashish Sharma, and Robby T Tan, "Dcshadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," *ICCV,* pp. 5027-5036, 2021.

[17] Hieu Le and Dimitris Samaras, “Shadow removal via shadow image decomposition,” *ICCV*, p. 8578–8587, 2020.

[18] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang, “Shadow removal by a lightness-guided network with training on unpaired data,” *IEEE Transactions on Image Processing*, p. 30.

[19] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang, “From shadow generation to shadow removal,” *CVPR*, pp. 4927-4936, 2021.

[20] L. Qu, J. Tian, S. He, Y. Tang, and Rwh Lau, “Deshadownet: A multi-context embedding deep network for shadow removal,” *CVPR*, pp. 4067-4075, 2017.

[21] O. Sidorov, “Conditional gans for multi-illuminant color constancy: Revolution or yet another approach,” *CVPRW*, p. 1748–1758, 2019.

[22] J. Wang, X. Li, and J. Yang, “Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal,” *CVPR*, pp. 1788-1797, 2018.

[23] Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao, “Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal,” *AAAI*, pp. 12829-12836, 2020.

[24] B. Ding, C. Long, L. Zhang, and C. Xiao, “Argan: Attentive recurrent generative adversarial network for shadow detection and removal,” *ICCV*, pp. 10213-10222, 2020.

[25] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao, “Canet: A context-aware network for shadow removal,” *ICCV*, pp. 4743-4752, 2021.

[26] S. Bako, S. Darabi, E. Shechtman, J. Wang, and P. Sen, “Removing shadows from images of documents,” *ACCV*, pp. 173-183, 2016.

[27] D. M. Oliveira, R. D. Lins, and Gabriel De Frana Pereira E Silva, " Shading removal of illustrated documents," *ICIAR,* 2013.

[28] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1397-1409, 2013.

[29] N Bharath Raj and N Venkateswaran, "Single image haze removal using a generative adversarial network," *CVPR,* pp. 37-42, 2018.

[30] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR,* pp. 1125-1134, 2017.

[31] Weiwen Chen, Yingtie Lei, Shenghong Luo, Ziyang Zhou, Mingxian Li, and Chi-Man Pun, "Shadocformer: A shadow-attentive threshold detector with cascaded fusion refiner for document shadow removal," *2024 International Joint Conference on Neural Networks (IJCNN),* pp. 1-8, 2024.

[32] Xuhang Chen, Xiaodong Cun, Chi-Man Pun, and Shuqiang Wang, "Shadocnet: Learning spatial-aware tokens in transformer for document shadow removal," *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 1-5, 2023.

[33] Shenghong Luo and Ruifeng Xu and Xuhang Chen and Zinuo Li and Chi-Man Pun and Shuqiang Wang, "DocDeshadower: Frequency-aware Transformer for Document Shadow Removal," *ArXiv,* 2023.

[34] Xiao Feng Zhang, Chao Chen Gu, and Shan Ying Zhu, "Spa-former: Transformer image shadow detection and removal via spatial attention," *arXiv,* 2022.

[35] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Zhihao Liu, Song Wang, "CRFormer: A Cross-Region Transformer for Shadow Removal," *arXiv,* 2022.

[36] Hua-En Chang, Chia-Hsuan Hsieh, Hao-Hsiang Yang, I Chen, Yi-Chung Chen, Yuan-Chun Chiang, Zhi-Kai Huang, Wei-Ting Chen, Sy-Yen Kuo, et al., "Tsrformer: Transformer based two-stage refinement for single image shadow removal," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1436-1446, 2023.