ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN

ĐỒ ÁN CHUYÊN NGÀNH

ĐỀ TÀI:

DOCUMENT IMAGE SHADOW REMOVAL BASED ON VISION TRANSFORMERS

Giảng viên hướng dẫn: TS. Nguyễn Duy Khánh

Sinh viên: Nguyễn Quốc Trường

MSSV: 21521604

TP. HỒ CHÍ MINH, NĂM 2024

NHẬN XÉT

……., ngày…….....tháng……năm 2024

Người nhận xét

(Ký tên và ghi rõ họ tên)

MỤC LỤC

DANH MỤC HÌNH ẢNH

CHƯƠNG 1. TỔNG QUAN

1.1. Giới thiệu bài toán

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.

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) 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.

Lợi ích mang lại:

Cải thiện chất lượng ảnh tài liệu số hóa: Việc loại bỏ bóng một cách hiệu quả giúp cải thiện chất lượng của ảnh tài liệu, làm cho các văn bản trở nên rõ ràng và dễ đọc hơn. Điều này rất quan trọng đối với các hệ thống lưu trữ và quản lý tài liệu số hóa.

Tăng độ chính xác của hệ thống OCR: Bóng trên ảnh tài liệu có thể gây ra lỗi nhận dạng trong các hệ thống OCR. Việc loại bỏ bóng sẽ giúp tăng độ chính xác của các hệ thống này, từ đó nâng cao hiệu quả trong việc trích xuất và xử lý thông tin từ tài liệu.

Nâng cao hiệu suất làm việc: Nhờ việc cải thiện chất lượng ảnh và độ chính xác của hệ thống OCR, thời gian và công sức dành cho việc xử lý và chỉnh sửa tài liệu sẽ được giảm thiểu, nâng cao hiệu suất làm việc của các nhân viên văn phòng và các chuyên gia xử lý tài liệu.

Tăng cường khả năng tự động hóa: Phương pháp loại bỏ bóng hiệu quả giúp các hệ thống xử lý ảnh tài liệu hoạt động tự động và chính xác hơn, giảm sự can thiệp của con người và tăng cường khả năng tự động hóa trong các quy trình xử lý và quản lý tài liệu.

Ứng dụng đa dạng: Phương pháp đề xuất có thể được áp dụng trong nhiều lĩnh vực khác nhau như lưu trữ và quản lý tài liệu, số hóa thư viện, quản lý hồ sơ y tế, và các hệ thống thông tin văn phòng khác.

Khả năng mở rộng và phát triển: Với nền tảng là Vision Transformers, phương pháp này có khả năng mở rộng và phát triển để áp dụng cho các bài toán xử lý ảnh tài liệu khác nhau, không chỉ dừng lại ở việc loại bỏ bóng mà còn có thể mở rộng sang các ứng dụng khác như tăng cường chất lượng ảnh, phát hiện và sửa lỗi trong tài liệu.

Đóng góp vào nghiên cứu và phát triển công nghệ: Việc áp dụng Vision Transformers trong xử lý ảnh tài liệu không chỉ mang lại lợi ích thực tiễn mà còn đóng góp vào nghiên cứu và phát triển công nghệ trong lĩnh vực trí tuệ nhân tạo và học sâu. Điều này mở ra nhiều hướng nghiên cứu mới và cơ hội ứng dụng khác trong tương lai.

Phương pháp luận:

Tiền xử lý dữ liệu: Thu thập dữ liệu: Tập hợp một bộ dữ liệu ảnh tài liệu chứa các loại bóng khác nhau. Bộ dữ liệu này có thể bao gồm các ảnh tài liệu từ các nguồn khác nhau với điều kiện ánh sáng và cấu trúc bóng đa dạng. Chuẩn hóa dữ liệu: Thực hiện các bước tiền xử lý như thay đổi kích thước ảnh, chuẩn hóa giá trị pixel và chia ảnh thành các tập huấn luyện, kiểm tra và kiểm định.

Chia ảnh thành các patches: Sử dụng lớp Patch Embedding để chia ảnh tài liệu thành các mảng nhỏ (patches). Các patches này sau đó được nhúng thành các vectors có kích thước cố định.

Sử dụng Vision Transformers: Embedding và Position Embedding: Chuyển đổi các patches thành vectors embedding và thêm thông tin vị trí để duy trì cấu trúc không gian của ảnh tài liệu. Transformer Encoder: Sử dụng các lớp Transformer Encoder để học các đặc trưng từ các vectors embedding. Các lớp này sử dụng cơ chế tự chú ý để nắm bắt thông tin ngữ cảnh toàn cục.

Học đặc trưng và loại bỏ bóng: Mạng Vision Transformer (ViT): Xây dựng và huấn luyện mạng Vision Transformer để học các đặc trưng và loại bỏ bóng trên ảnh tài liệu. ViT sẽ xử lý các vectors embedding và tạo ra các đặc trưng đã được cải thiện. Kết hợp với

Conv2D: Sau khi ViT xử lý xong, kết quả sẽ được chuyển qua một lớp Conv2D cuối cùng để tái tạo lại ảnh tài liệu với các bóng đã được loại bỏ.

Huấn luyện mô hình: Loss Function: Sử dụng hàm mất mát phù hợp để tối ưu hóa mô hình. Optimizer: Sử dụng thuật toán tối ưu hóa như Adam để huấn luyện mô hình. Điều chỉnh các siêu tham số như learning rate, batch size để đạt hiệu suất tối ưu. Validation và Tuning: Sử dụng tập kiểm tra để đánh giá hiệu suất của mô hình trong quá trình huấn luyện và điều chỉnh các siêu tham số nếu cần thiết.

Đánh giá và so sánh: Đánh giá mô hình: Đánh giá hiệu suất của mô hình trên tập kiểm định bằng các chỉ số như RMSE (Root Mean Squared Error), PSNR (Peak Signal-to-Noise Ratio) và SSIM (Structural Similarity Index). So sánh với phương pháp truyền thống: So sánh kết quả của mô hình Vision Transformer với các phương pháp truyền thống để chứng minh tính hiệu quả của phương pháp đề xuất.

Kết luận và hướng phát triển: Tổng kết kết quả: Đưa ra nhận xét về hiệu suất của mô hình và những cải tiến đạt được trong việc loại bỏ bóng trên ảnh tài liệu. Hướng phát triển: Đề xuất các hướng nghiên cứu và cải tiến trong tương lai để nâng cao hiệu quả và ứng dụng của phương pháp. Thông qua các bước trên, đề tài này sẽ xây dựng một phương pháp hiệu quả và tiên tiến để loại bỏ bóng trên ảnh tài liệu, cải thiện chất lượng ảnh và hiệu suất của các hệ thống xử lý ảnh tài liệu.

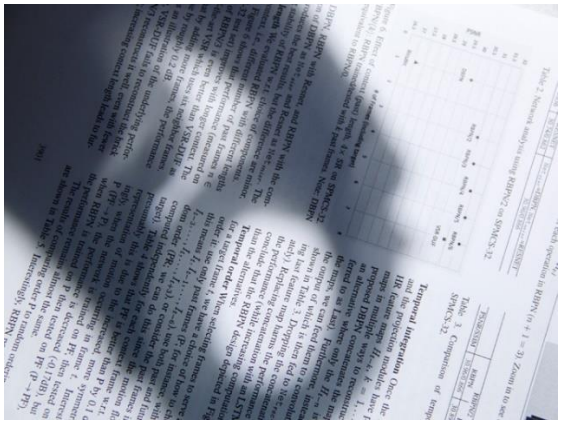
1.1.2 Phát biểu bài toán

Đầu vào

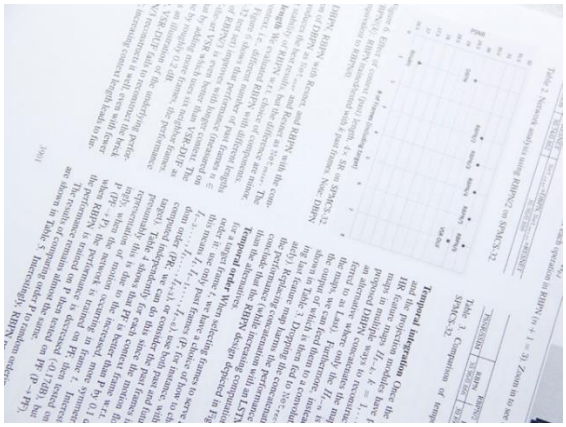
- Ảnh tài liệu chứa các bóng do điều kiện ánh sáng không đồng đều hoặc các vật cản trong quá trình chụp ảnh.
- Các ảnh tài liệu có thể ở các định dạng khác nhau (JPEG, PNG, TIFF, v.v.) và có kích thước, độ phân giải đa dạng.

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.



Input image



Output image

Mục tiêu của hệ thống

Hệ thống sử dụng mô hình dựa trên Vision Transformers để loại bỏ bóng trên ảnh tài liệu, cải thiện chất lượng ảnh và độ chính xác của các ứng dụng xử lý ảnh tài liệu.

1.2. Quá trình thực hiện

Thu thập và xử lý dữ liệu:

- Thu thập dữ liệu: Thu thập một bộ dữ liệu ảnh tài liệu chứa các bóng do điều kiện ánh sáng không đồng đều hoặc các vật cản. Bộ dữ liệu cần đa dạng về nguồn gốc, kích thước, độ phân giải và điều kiện ánh sáng.
- Chú thích dữ liệu: Nếu cần thiết, thêm các chú thích (annotations) cho ảnh tài liệu để đánh dấu các vùng chứa bóng và các vùng không chứa bóng.
- Chuẩn hóa dữ liệu: Thực hiện các bước tiền xử lý như thay đổi kích thước ảnh, chuẩn hóa giá trị pixel, và chia dữ liệu thành các tập huấn luyện, kiểm tra và kiểm định.

Xây dựng mô hình:

- Xây dựng lại mô hình Xóa bóng trên ảnh tài liệu, sau đó điều chỉnh và thay đổi mô hình dựa trên kiến thức về Vision Transformers với 2 thay đổi chính là Patch Embedding và Transformer Encoder.
- Patch Embedding: Chia ảnh tài liệu thành các mảng nhỏ (patches) và chuyển đổi chúng thành vectors embedding.

- Transformer Encoder: Sử dụng các lớp Transformer Encoder để học các đặc trưng từ các vectors embedding thông qua cơ chế tự chú ý.

Huấn luyện mô hình:

- Loss Function: Xác định hàm mất mát phù hợp hoặc các hàm mất mát khác để đo lường sự khác biệt giữa ảnh gốc và ảnh đã loại bỏ bóng.
- Optimizer: Lựa chọn và cấu hình thuật toán tối ưu hóa như Adam để huấn luyện mô hình.
- Huấn luyện mô hình: Chạy quá trình huấn luyện trên tập dữ liệu huấn luyện, sử dụng tập kiểm tra để điều chỉnh siêu tham số và ngăn chặn hiện tượng overfitting.
- Theo dõi và ghi nhận kết quả: Ghi lại các thông số huấn luyện như độ chính xác, loss function và thời gian huấn luyện.

Đánh giá mô hình:

- Đánh giá trên tập kiểm định: Sử dụng tập kiểm định để đánh giá hiệu suất của mô hình bằng các chỉ số như RMSE (Root Mean Squared Error), PSNR (Peak Signal-to-Noise Ratio) và SSIM (Structural Similarity Index).
- So sánh với các phương pháp truyền thống: So sánh kết quả của mô hình Vision Transformer với các phương pháp truyền thống để xác định mức độ cải thiện.

Tối ưu hóa và cải tiến:

- Tối ưu hóa mô hình: Thực hiện các cải tiến và tối ưu hóa mô hình dựa trên kết quả đánh giá, có thể bao gồm việc điều chỉnh siêu tham số, thay đổi cấu trúc mô hình hoặc thử nghiệm các kỹ thuật học sâu khác.
- Kiểm thử trên dữ liệu thực tế: Áp dụng mô hình đã huấn luyện trên các tập dữ liệu thực tế để kiểm tra tính khả thi và hiệu suất của mô hình trong các tình huống thực tế.

Báo cáo và tài liệu hóa:

- Viết báo cáo: Soạn thảo báo cáo chi tiết về toàn bộ quá trình thực hiện, bao gồm các phương pháp, kết quả thực nghiệm và các nhận xét.
- Tài liệu hóa: Cung cấp tài liệu hướng dẫn sử dụng hệ thống và các tài liệu kỹ thuật liên quan để hỗ trợ triển khai và bảo trì hệ thống.

CHƯƠNG 2. MÔ TẢ NGHIÊN CỨU

2.1. Cơ sở lý thuyết

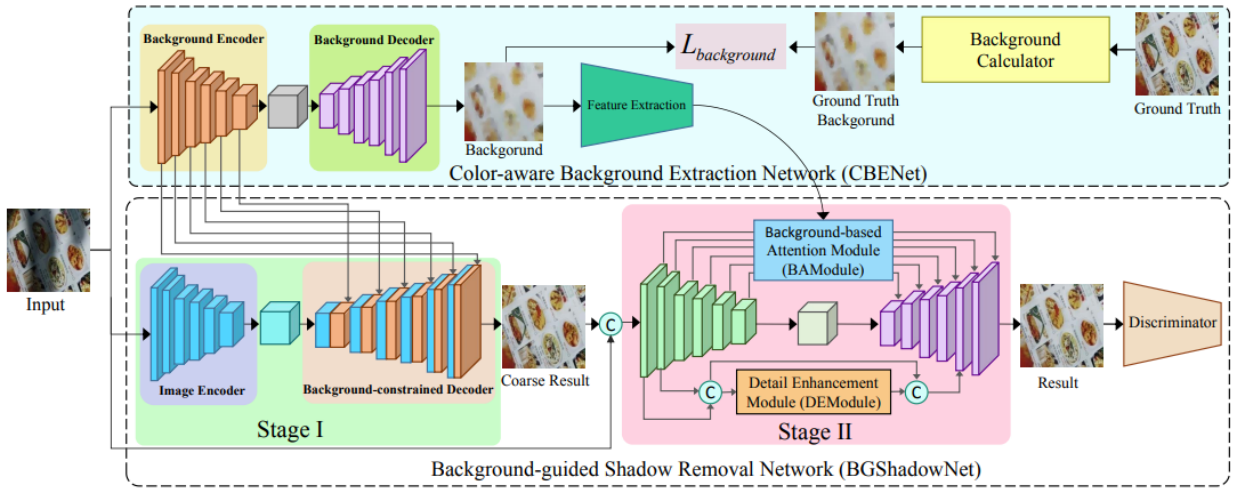
2.1.1 Document image shadow removal

Xóa Bóng Trong Ảnh Tự Nhiên:

- Các phương pháp truyền thống để xóa bóng trong ảnh tự nhiên thường tập trung vào nghiên cứu các đặc tính vật lý khác nhau của bóng. Tuy nhiên, các phương pháp này có thể gây ra hiện tượng viền bóng rõ rệt do sự thay đổi của ánh sáng. 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. Ví dụ, Deshadow-Net trích xuất các đặc trưng đa ngữ cảnh để dự đoán các lớp bóng mờ cho việc xóa bóng. Các mạng GAN có điều kiện xếp chồng để phát hiện và xóa bóng cùng lúc. Hoặc phương pháp sử dụng mạng GAN cho việc xóa bóng bằng cách sử dụng phần dư và chiếu sáng. ARGAN đề xuất một mạng GAN tái phát sinh chú ý cho việc phát hiện và xóa bóng. Gần đây, còn có một số phương pháp như 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 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 sử dụng nội suy lân cận tự nhiên để ước tính ảnh bóng. Jung và cộng sự 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 mờ hoặc vừa, nhưng có xu hướng làm suy giảm màu sắc đối với các cảnh có bóng đậm. Gần đây, Lin và cộng sự 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.

Từ việc xuất hiện nhiều nhược điểm từ các phương pháp trước đó như vậy, một nhóm nhà nghiên cứu đã đề xuất ra phương pháp được đề cập đến trong bài báo "Document image shadow removal guided by Color-aware Background". Phương pháp này cũng chính là phương pháp em sử dụng làm backbone (xương sống) cho phương pháp của em.



Framework của 2 mạng CBENet và BGShadowNet, được đề xuất trong bài báo "Document Image Shadow Removal guided by Color-aware Background"

2.1.2 Color-aware background extraction network (CBENet)

Vì ảnh tài liệu tập trung chủ yếu vào nội dung văn bản, một chiến lược phổ biến 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.

Để giải quyết vấn đề này, các nhà nghiên cứu đã đề xuất "Color-aware background extraction network" (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 để tinh chỉnh. Từ đó thu được nền ảnh hoàn chỉnh mới tiếp tục xử lý.

Nền ảnh thu được sau khi đưa ảnh qua mạng CBENet, bên trái là khi ghép các patch lại, chưa qua tinh chỉnh, bên phải là ảnh nền hoàn chỉnh để đưa đi xử lý ở bước kế tiếp

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.

2.1.3 Background-guided Shadow Removal Network (BGShadowNet)

Như đã đề cập, nền có thể cung cấp thông tin hữu ích để hỗ trợ việc xóa bóng. Từ đó có một phương pháp 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.

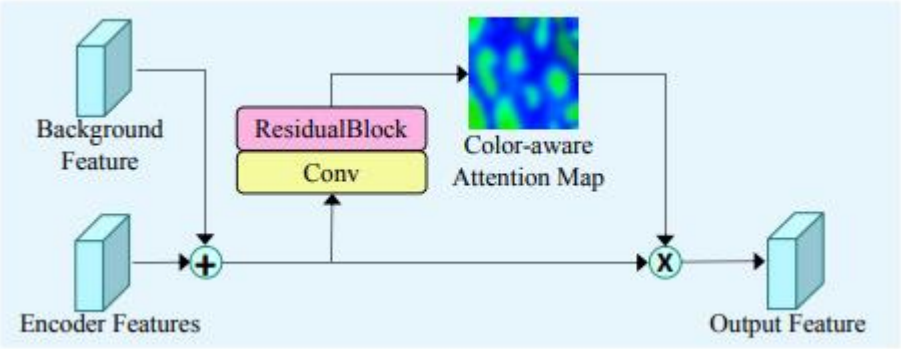
Tại Giai đoạn I, bên cạnh Encoder, một Decoder bị ràng buộc bởi nền được giới thiệu để tạo ra kết quả xóa bóng sơ bộ.

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 encode - decode. 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 và bộ phân biệt Markovian làm cấu trúc encode - decode và bộ phân biệt.

Background-constrained decoder: Để tận dụng tối đa các đặc trưng từ ảnh nền, phương pháp này thay thế Decoder thông thường bằng Decoder bị ràng buộc bởi nền tại Giai đoạn I. Cụ thể, các đặc trưng từ Background Encoder được tích hợp vào Decoder 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.

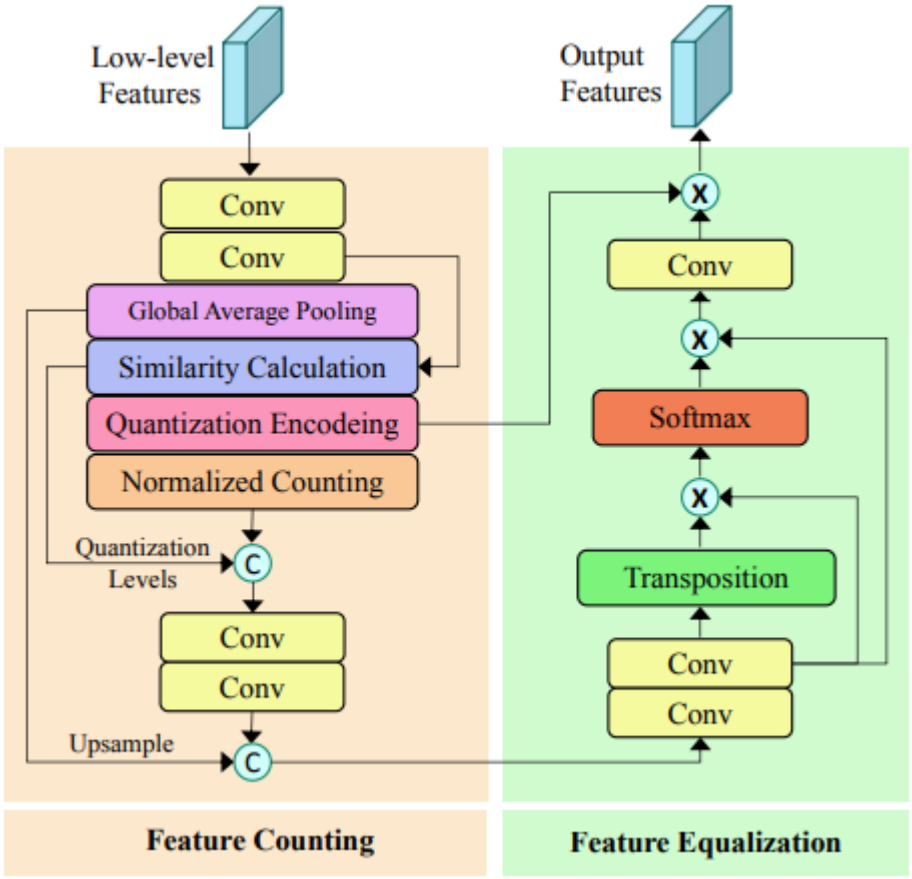
Background-based Attention Module (BAModule): Thông thường, các vùng có nền tương tự nên có sự xuất hiện tương tự (màu sắc và chiếu sáng) trong ảnh. Tuy nhiên, có thể có các hiện tượng ánh sáng hoặc màu sắc trong kết quả xóa bóng sơ bộ. Để giữ sự nhất quán tổng thể của ảnh, phương pháp giới thiệu module 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.



Cấu trúc BAModule

Detail Enhancement Module (DEModule): 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.



Cấu trúc DEModule

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, 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, tính 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, 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, thực hiện một phép chuẩn hóa cho ma trận E. Tiếp đến 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. Sau đó, thực hiện hai phép tích chập 1 × 1 cho C để tăng số kênh, sau đó thực hiện phép nối kênh với 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.

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, thực hiện một phép tích chập 1 × 1 cho H để thu được G. Tiếp đến 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, 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', 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.

2.1.4 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 $\hat{B}$ 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. Sử dụng mô hình VGG19 pretrained.

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

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 $\lambda_1, \lambda_2, \lambda_3$và λ_4 là các tham số.

2.1.5 Vision Transformer

1. Vision Transformer (ViT)
Vision Transformer (ViT) là một mô hình học sâu sử dụng cấu trúc transformer, 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. ViT chuyển đổi ảnh đầu vào thành một chuỗi các patch nhỏ và sử dụng các khối transformer để xử lý chuỗi này. Cấu trúc ViT 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.

2. Ứng Dụng ViT Trong Xóa Bóng Ảnh Tài Liệu
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 ViTs sơ bộ 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 ViT: Huấn luyện mô hình ViT 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 ViT đã huấn luyện để dự đoán và loại bỏ bóng trong ảnh tài liệu.

3. Cơ Sở Lý Thuyết
3.1. Tách Ảnh Thành Các Patch

Việc tách ảnh thành các patch nhỏ giúp mô hình có thể xử lý từng phần của ảnh một cách riêng biệt, học được các đặc trưng cục bộ trước khi tổng hợp thông tin toàn cục.

3.2. Embedding và Position Embedding
Nhúng các patch thành các vector cho phép mô hình xử lý các patch như các token trong xử lý ngôn ngữ tự nhiên. Thêm thông tin vị trí giúp bảo toàn cấu trúc không gian của ảnh, điều này rất quan trọng để giữ được thông tin về sự phân bố bóng và nội dung văn bản.

3.3. Transformer Encoder
Sử dụng transformer encoder giúp mô hình có thể học các mối quan hệ phức tạp giữa các patch, bao gồm các tương tác không gian và ngữ cảnh. Điều này giúp mô hình nhận diện và loại bỏ bóng một cách chính xác hơn.

3.4. Học Máy Có Giám Sát (Supervised Learning)
Mô hình ViT được huấn luyện với một tập dữ liệu gồm các ảnh tài liệu có bóng, ảnh nền không có bóng và ảnh tài liệu không có bóng. Quá trình này cho phép mô hình học và trích xuất được các đặc trưng của ảnh, hỗ trợ cho quá trình xóa bòng của ảnh tài liệu.

4. Triển Khai Cụ Thể
Trong bối cảnh xóa bóng ảnh tài liệu, ViT có thể được triển khai như sau:
Chuẩn Bị Dữ Liệu: Thu thập một tập dữ liệu lớn các ảnh tài liệu có bóng và không có bóng. Chia ảnh thành các patch nhỏ và gán nhãn cho từng patch.
Huấn Luyện Mô Hình: Sử dụng các phương pháp học có giám sát để huấn luyện mô hình ViT với tập dữ liệu đã chuẩn bị.
Đánh Giá và Tinh Chỉnh: Đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra và thực hiện các tinh chỉnh cần thiết để cải thiện độ chính xác.
Ứng Dụng Thực Tế: Sử dụng mô hình ViT đã huấn luyện để xóa bóng trong các ảnh tài liệu thực tế, đảm bảo chất lượng ảnh được cải thiện và văn bản rõ ràng hơn.
Kết Luận
Vision Transformer (ViT) mang lại một phương pháp mạnh mẽ và hiệu quả cho bài toán xóa bóng trong ảnh tài liệu. Bằng cách tận dụng các đặc trưng không gian và ngữ cảnh mà ViT học được, chúng ta có thể đạt được kết quả xóa bóng chính xác và cải thiện chất lượng ảnh tài liệu một cách đáng kể.

2.2 Xây dựng mô hình

Dựa trên mô hình “document image shadow removal guided by color-aware background”, em đã thay đổi và phát triển lại mô hình này dựa trên Vision Transformer.

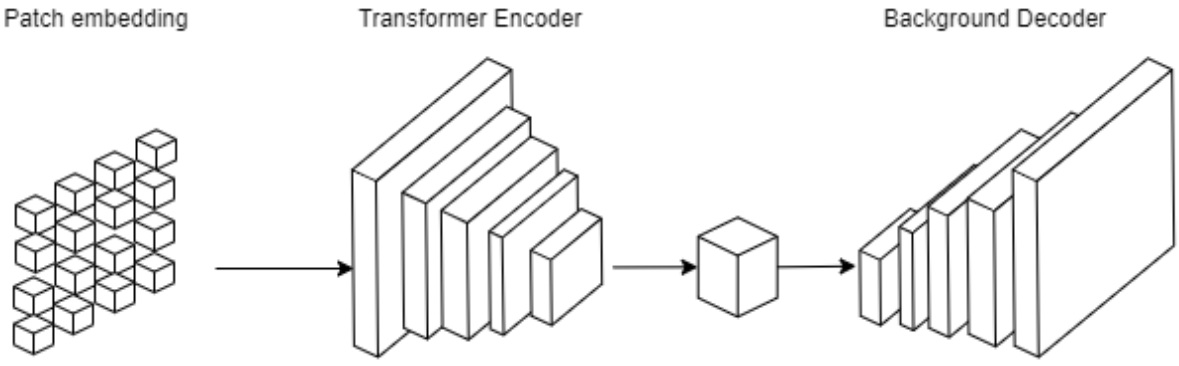
Ý tưởng của 2 mạng CBE và BGShadow là khi trích xuất đặc trưng từ ảnh, sẽ chia ảnh thành 16x16 patch bằng nhau, rồi mới đưa qua tính toán và xử lý. Dựa trên việc chia ảnh thành các patch rồi mới đưa qua Encoder và Decoder đó, em đã nảy ra ý tưởng sử dụng ViTs để thay thế cho các mạng Encoder thông thường (là các lớp conv) trong 2 mạng này.

2.2.1 Color-aware background extraction network based on Vision Transformer (CBETransformer)

Cấu trúc của mô hình:

- Sử dụng các lớp Patch Embedding để chia ảnh thành các mảng nhỏ và nhúng chúng thành các vectors có kích thước cố định.
- Sử dụng các lớp Vision Transformer Encoder để học các đặc trưng từ các vectors embedding. Ở phương pháp gốc, họ đã sử dụng mạng U-net với 5 lớp conv liên tiếp để làm Encoder trích xuất đặc trưng. Trong phương pháp của em, em sẽ thay thế các lớp này thành ViTs Encoder và thực hiện nhiệm vụ tương tự.
- Các ViTs Encoder trong mạng này được cài đặt với mục tiêu chỉ trích xuất các đặc trưng liên quan tới nền ảnh.
- Cuối cùng sử dụng các lớp Decoder để giải mã các đặc trưng thu được từ Encoder nhằm thu được nền ảnh hoàn chỉnh, từ đó đưa vào mạng BGShadowNet tiếp tục xử lý.

Cài đặt mô hình: Xây dựng các lớp cần thiết như Patch Embedding, Transformer Encoder và Decoder để trích xuất nền ảnh. Kết hợp các lớp này để xây dựng mạng CBENet hoàn chỉnh.



Minh họa cấu trúc mạng CBETransformer

PatchEmbedding: Chia ảnh đầu vào thành các patches và nhúng chúng thành các vectors embedding.

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

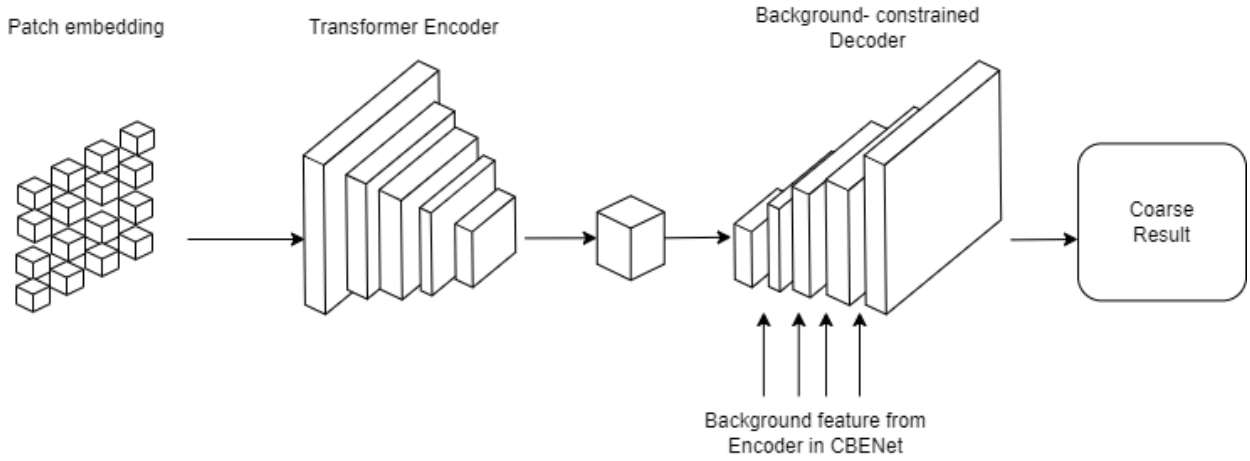
TransformerEncoder: Tập hợp nhiều lớp TransformerEncoderLayer.

Decoder: Giải mã những thông tin lấy từ lớp Encoder (đặc trưng được trích xuất) thành nền ảnh (không có bóng) cho bước tiếp theo.

2.2.2 Background-guided shadow removal network based on Vision Transformer (BGSTransformer)

Khi điều chỉnh và thay đổi mạng BGShadowNet dựa trên ViTs, em 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, em 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.
Cách xử lý tương tự như ở mạng CBE, tuy nhiên ở đây, ViTs Encoder sẽ trích xuất tất cả đặc trưng có trong ảnh (nền, màu sắc, chữ, ...) khác với ở mạng CBE là chỉ lấy đặc trưng nền ảnh.
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.
Transformer Encoder Layer: Một lớp encoder của Vision Transformer 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.
Transformer Encoder: Tập hợp nhiều lớp Transformer Encoder Layer. Các lớp này kết hợp với nhau để học các đặc trưng phức tạp hơn từ ảnh.
Decoder Layer: Lấy các đặc trưng thu được từ các lớp Encoder, đồng thời kết hợp các đặc trưng nền ảnh thu được từ mạng CBE làm đặc trưng bổ sung để giải mã và cho ra kết quả thô, làm đầu vào cho giai đoạn tiếp theo.



Minh họa cấu trúc Giai đoạn I của mạng BGS

Ở Giai đoạn II, mạng có 2 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 conv là khả quan nhất. Vì thế em thử thay thế lần lượt hoặc toàn bộ các lớp conv bằng ViTs. Cách xây dựng các lớp ViTs này thì tương tự như đã thực hiện ở giai đoạn trước đó.
Kết quả thu được khi điều chỉnh cấu trúc của Giai đoạn II không khả quan, cho thấy sự thay dổi này là không hợp lý, do đó việc cài đặt và chỉnh sửa mạng BGS em chỉ thực hiện ở Giai đoạn I.

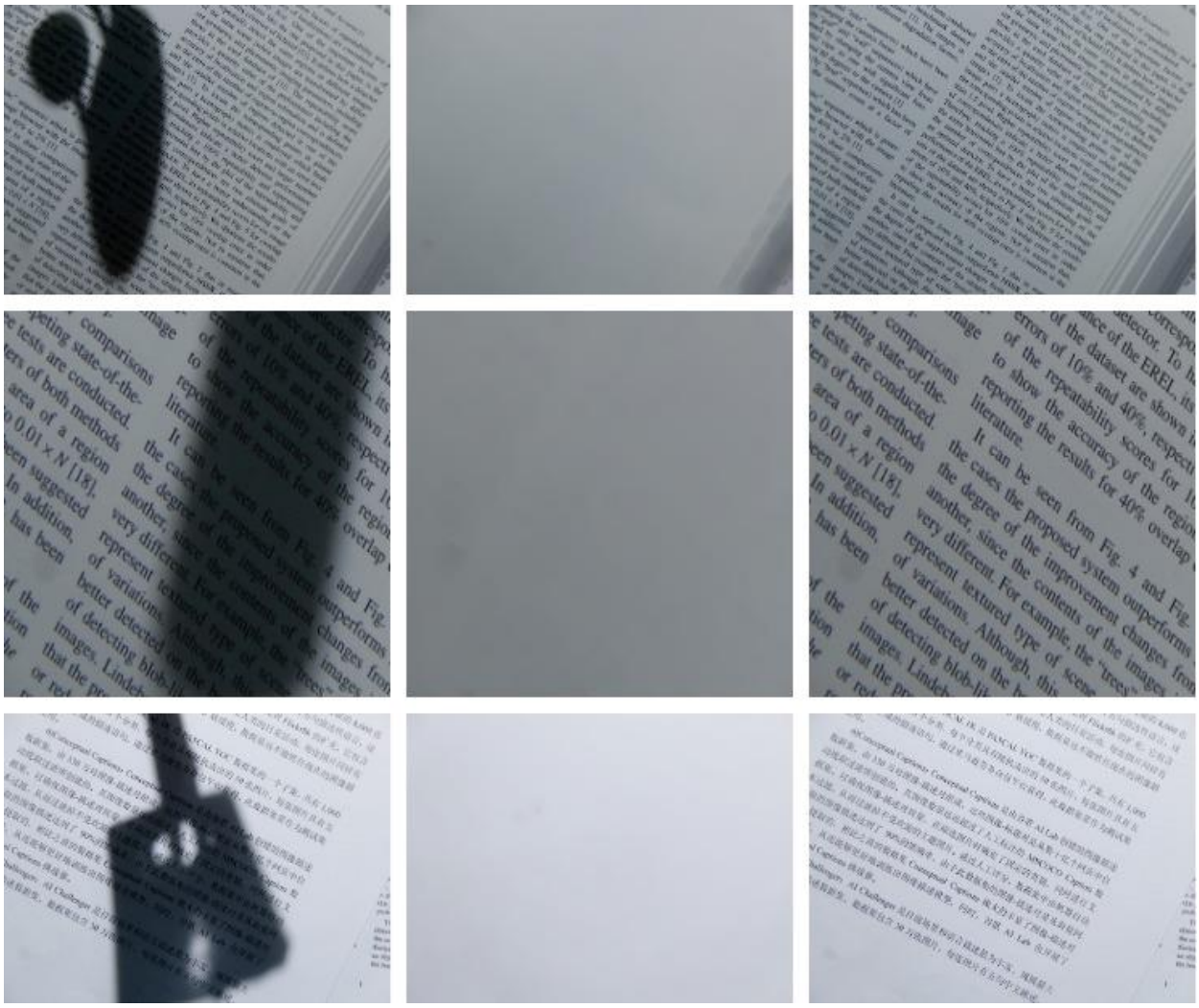
2.2.3 Dataset

RDD dataset

Mặc dù có nhiều bộ dữ liệu loại bỏ bóng trên tài liệu, chẳng hạn như Bako, Kligler, Jung, RDSRD, và SDSRD, nhưng chúng đều có một số hạn chế. Bako, Kligler, Jung và RDSRD là các bộ dữ liệu đánh giá quy mô nhỏ, không phù hợp cho việc huấn luyện mô hình sâu. Trong khi đó, SDSRD là một bộ dữ liệu quy mô lớn nhưng là dữ liệu tổng hợp.

Quá trình tạo ảnh là một quá trình vật lý thông qua sự tương tác giữa ánh sáng và vật liệu. Môi trường ánh sáng trong thế giới thực thường chứa nhiều nguồn sáng khác nhau, điều này khó có thể mô phỏng chính xác trong môi trường tổng hợp. Các đặc trưng thống kê của ảnh tổng hợp và ảnh thực thường khác nhau. Để cải thiện hiệu suất loại bỏ bóng trên ảnh tài liệu, em đã sử dụng một bộ dữ liệu tài liệu thực mới được giới thiệu trong bài báo "Document Image Shadow Removal guided by Color-aware Background", gọi là RDD, dành cho việc loại bỏ bóng. Cụ thể, các nhà nghiên cứu sử dụng các tài liệu như giấy, sách,

tờ rơi quảng cáo, v.v., làm cảnh nền để xây dựng bộ dữ liệu. Đầu tiên, họ chụp ảnh có bóng với ánh sáng bị chặn bởi một vật thể. Sau đó, họ chụp ảnh không có bóng tương ứng bằng cách loại bỏ vật cản. Bộ dữ liệu RDD thu thập được 4916 cặp ảnh có bóng và không có bóng, chia thành hai nhóm, 4371 cho huấn luyện và 545 cho kiểm tra. RDD là bộ dữ liệu tài liệu thực quy mô lớn đầu tiên dành cho việc loại bỏ bóng.



Một số ví dụ trong dataset RDD, với cột bên trái là ảnh tài liệu (có bóng), cột bên phải là ảnh tài liệu (không có bóng) và cột ở giữa là nền ảnh

2.2.4 Các độ đo đánh giá

Một số độ đo (metric) được sử dụng để đánh giá các mô hình.

- 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

- Structural Similarity Index (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.

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

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

2.3 Thử nghiệm và đánh giá mô hình

Ở các mô hình, 2 mạng CBE và BGShadow đều được train song song và riêng biệt với nhau, gồm 200 epoch, với optimizer là Adam và learning rate là 0.0004.

Các tham số trong mô hình là λ_1, λ_2, λ_3 và λ_4 lần lượt mang giá trị 1, 1, 0.05 và 0.01.

2.3.1 CBENet + BGShadowNet (1)

Ở model (1) này, em 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 2 mạng là CBENet và BGShadowNet. Các tham số cũng như loss function giữ nguyên, không thay đổi gì cả.

2.3.2 CBETransformer + BGSTransformer (2)

Ở model (2), em điều chỉnh lại 2 mạng CBE và BGShadow dựa trên ViTs trước khi train. Cụ thể:

CBETransformer 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.

BGSTransformer: Trong quá trình xử lý và điều chỉnh mạng này, em tiến hành thay thế lần lượt các lớp convolutional trong mạng BGShadowNet bằng ViTs rồi tiến hành train. Quá trình này tốn khá nhiều tài nguyên, do đó em đã dùng các ViTs pretrained để thử nghiệm và đánh giá. Cụ thể, em 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, em 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, em chỉ thực hiện chỉnh sửa và thay đổi ở Giai đoạn I, Giai đoạn II không thay đổi.

2.3.4 CBETransformer + BGShadowNet (3)

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

Thay mạng U-Net với các lớp conv thành ViTs, chia ảnh thành các patch rồi chuyển thành các vector embedding rồi tiếp tục xử lý.

Ở mạng BGS thì giữ nguyên cài đặt như phương pháp gốc.

2.3.5 CBENet + BGSTransformer (4)

Tương tự như cách cài đặt các mô hình trước đó.

CBENet cài đặt như ở phương pháp gốc của các nhà nghiên cứu.

Mạng BGS chỉnh sửa lại như đề xuất của em ở các mô hình trước đó (mô hình (2)).

2.3.6 So sánh các mô hình và phương pháp

Các mô hình và phương pháp đều được huấn luyện và thử nghiệm trên dataset RDD, với các độ đo RMSE, PSNR và SSIM.

Models	RMSE	PSNR	SSIM
(1)	2.572	36.301	0.975
(2)	2.477	36.624	0.977
(3)	2.583	36.263	0.972
(4)	2.639	36.077	0.968

Bảng so sánh các mô hình sau khi điều chỉnh với ViTs

Methods	RMSE	PSNR	SSIM
(2)(*)	2.477	36.624	0.977
BGShadowNet(*)	2.572	36.301	0.975
BGShadowNet	2.219	37.585	0.983
BEDSR-Net	2.937	34.928	0.973
Bako	14.648	20.741	0.894
Jung	30.190	14.364	0.861

Bảng so sánh phương pháp đề xuất với các phương pháp SOTA, (*) là các phương pháp do em tự mình huấn luyện và đánh giá

Kết luận:

Qua quá trình cài đặt và thử nghiệm, từ kết quả thu được, cho thấy mô hình có kết quả tốt nhất là mô hình (2) CBETransformer+BGSTransformer. Hiệu quả mà mô hình đạt được thể hiện khá tốt thông qua các độ đo như RMSE, PSNR và SSIM. Có phần tốt hơn so với các mô hình đã có trước đây.

Kết quả do em tự thử nghiệm và đánh giá có thể không cao bằng kết quả các nhà nghiên cứu đưa ra, dựa trên một vài lý do. Mô hình của em xây dựng dựa trên ViTs pretrained, không đủ thời gian và tài nguyên để train lại hoàn toàn, do đó kết quả thu được sẽ chưa phản ánh chính xác độ hiệu quả hơn của mô hình. Ngoài ra, do máy móc và thiết bị còn hạn chế, không cài đặt, huấn luyện và thử nghiệm chính xác được. Vì thế, kết quả có thể sẽ không cao bằng.

Hiệu quả của mô hình là có cải tiến, tuy nhiên sẽ cần quan sát và điều chỉnh nhiều điểm, nhằm tăng kết quả thử nghiệm và thực tế.

CHƯƠNG 3. TỔNG KẾT

3.1 Kết quả đạt được

Trong quá trình thực hiện đồ án "Document image shadow removal based on Vision Transformers", em đã đạt được những kết quả sau:

- Hiệu quả loại bỏ bóng: Mô hình sử dụng Vision Transformers đã chứng minh khả năng loại bỏ bóng trên ảnh tài liệu một cách hiệu quả, cải thiện rõ rệt chất lượng ảnh so với các phương pháp truyền thống và CNN.
- Độ chính xác và độ rõ nét của ảnh sau khi loại bỏ bóng được nâng cao, hỗ trợ tốt hơn cho các hệ thống nhận dạng ký tự quang học (OCR).
- Cải thiện độ chính xác của OCR: Việc loại bỏ bóng đã giúp tăng độ chính xác của các hệ thống OCR, làm giảm đáng kể các lỗi nhận dạng do bóng gây ra.
- Khả năng xử lý toàn cục: Sử dụng cơ chế tự chú ý của Vision Transformers, mô hình có khả năng nắm bắt thông tin ngữ cảnh toàn cục, giúp xử lý các biến đổi phức tạp của bóng và kết cấu nền một cách hiệu quả.
- Tự động hóa quy trình: Hệ thống có khả năng tự động loại bỏ bóng, giảm thiểu sự can thiệp của con người, tăng hiệu suất làm việc và khả năng ứng dụng thực tế.

3.2 Những hạn chế và khó khăn

Mặc dù đã đạt được nhiều kết quả tích cực, em cũng gặp phải một số khó khăn và hạn chế:

- Yêu cầu tài nguyên tính toán: Vision Transformers yêu cầu tài nguyên tính toán lớn, đặc biệt là GPU có dung lượng bộ nhớ cao để huấn luyện và suy luận. Điều này có thể là một rào cản đối với các tổ chức có tài nguyên hạn chế.
- Thời gian huấn luyện: Quá trình huấn luyện mô hình Vision Transformers đòi hỏi thời gian dài và công suất tính toán cao, điều này có thể làm chậm tiến độ triển khai và phát triển.

- Đa dạng dữ liệu: Mặc dù mô hình hoạt động tốt trên nhiều loại dữ liệu, nhưng vẫn có thể gặp khó khăn với các ảnh tài liệu có cấu trúc và điều kiện ánh sáng quá đặc thù hoặc không phổ biến trong tập huấn luyện.
- Chưa xử lý tốt trên các tài liệu tiếng việt có dấu, hay tài liệu chữ viết tay.
- Khả năng tổng quát hóa: Việc đảm bảo mô hình hoạt động tốt trên nhiều loại tài liệu và điều kiện ánh sáng khác nhau đòi hỏi một tập dữ liệu huấn luyện đa dạng và phong phú, điều này có thể là một thách thức lớn.

3.3 Hướng phát triển tương lai

Dựa trên kết quả và kinh nghiệm thu được, em đề xuất một số hướng phát triển tương lai để cải thiện hệ thống:

- Tối ưu hóa mô hình: Nghiên cứu và áp dụng các kỹ thuật tối ưu hóa để giảm thiểu yêu cầu tài nguyên tính toán và thời gian huấn luyện mà vẫn duy trì hiệu suất cao.
- Thử nghiệm các biến thể của Vision Transformers hoặc kết hợp với các mô hình khác để cải thiện hiệu quả và tốc độ xử lý.
- Mở rộng tập dữ liệu huấn luyện: Thu thập và xây dựng các tập dữ liệu huấn luyện phong phú và đa dạng hơn, bao gồm các loại tài liệu (tiếng việt, chữ viết tay,...) và điều kiện ánh sáng khác nhau, để cải thiện khả năng tổng quát hóa của mô hình.
- Nghiên cứu các kỹ thuật học sâu khác: Khám phá và áp dụng các kỹ thuật học sâu khác như GAN (Generative Adversarial Networks) để cải thiện quá trình loại bỏ bóng và tái tạo ảnh tài liệu.
- Ứng dụng trong các lĩnh vực khác: Mở rộng ứng dụng của phương pháp này sang các lĩnh vực khác như y tế (xử lý ảnh y khoa), bảo tồn di sản (xử lý ảnh tài liệu cổ), và các hệ thống thị giác máy tính khác.
- Phát triển giao diện người dùng: Xây dựng các giao diện người dùng thân thiện để hỗ trợ việc triển khai và sử dụng hệ thống trong các ứng dụng thực tế.

- Tiếp tục theo dõi và cải tiến: Theo dõi hiệu suất của hệ thống sau khi triển khai, thu thập phản hồi từ người dùng và thực hiện các cải tiến liên tục để nâng cao chất lượng và hiệu quả của hệ thống.

Bằng cách thực hiện các hướng phát triển trên, hệ thống loại bỏ bóng trên ảnh tài liệu sử dụng Vision Transformers có thể được hoàn thiện và mở rộng, mang lại nhiều giá trị thực tiễn hơn trong các ứng dụng xử lý ảnh tài liệu và các lĩnh vực liên quan.

TÀI LIỆU THAM KHẢO

[1] Qing Zhang, Zheng Liu, Xiaolong Zhang, Chunxia Xiao, Ling Zhang, Yinghao He. 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.

[3] Y. H. Lin, W. C. Chen, and Y. Y. Chuang. Bedsr-net: A deep shadow removal network from a single document image. In *CVPR*, pages 12905–12914, 2020.

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