ABSTRACT

Trong lĩnh vực xử lý ảnh, việc kết hợp hình ảnh hồng ngoại (IR) và hình ảnh khả kiến (VI) là rất quan trọng đối với các ứng dụng yêu cầu khả năng nhận thức chi tiết và toàn diện như giám sát, theo dõi đối tượng. Hình ảnh IR ghi lại các dấu hiệu nhiệt, đặc biệt hữu ích trong các điều kiện tầm nhìn hạn chế như khói, sương mù hoặc ban đêm. Ngược lại, hình ảnh VIS cung cấp thông tin màu sắc chi tiết, cần thiết cho việc nhận diện và diễn giải các cảnh trong điều kiện ánh sáng bình thường. Tuy nhiên, việc tích hợp hai loại hình ảnh này đặt ra những thách thức đáng kể, vì các phương pháp truyền thống thường không thể kết hợp hiệu quả các đặc điểm riêng biệt của chúng, dẫn đến hình ảnh thiếu chi tiết hoặc không làm nổi bật các thông tin nhiệt quan trọng.

Để vượt qua những thách thức này, nghiên cứu này giới thiệu một mô hình lai mới kết hợp giữa kim tự tháp Laplacian (Laplacian Pyramid) và mô hình học sâu tiên tiến dựa trên NestFuse. Mô hình này tận dụng khả năng phân rã đa tỷ lệ của kim tự tháp Laplacian và khả năng của NestFuse trong việc học, tối ưu hóa các chiến lược kết hợp từ dữ liệu, giúp tăng cường khả năng thích ứng với các đặc điểm độc đáo của hình ảnh IR và VIS.

Quy trình của phương pháp này bao gồm việc sử dụng kim tự tháp Laplacian biến thể để phân rã hình ảnh IR và VIS thành nhiều lớp độ phân giải với các thành phần cơ sở và chi tiết. Thành phần cơ sở được tổng hợp dựa trên phương pháp năng lượng vùng cục bộ cục đại. Thành phần chi tiết được tổng hợp dựa trên mô hình NestFuse được huấn luyện trước với sự thay đổi chiến lược tổng hợp so với mô hình gốc. Phương pháp này giữ lại các chi tiết và thông tin quan trọng từ cả hai hình ảnh, đồng thời cải thiện khả năng xử lý những phức tạp liên quan đến việc kết hợp hình ảnh IR và VIS.

Đóng góp chính của nghiên cứu này là phát triển một mô hình lai tích hợp các kỹ thuật xử lý ảnh truyền thống với học sâu hiện đại. Kết quả thực nghiệm đã cho thấy cách tiếp cận sáng tạo này cải thiện rõ rệt chất lượng của hình ảnh kết hợp, mang lại khả năng thích ứng và hiệu suất vượt trội so với các phương pháp hiện có. Mô hình này đặc biệt hữu ích cho các ứng dụng như an ninh, giám sát, điều hướng xe tự động và chữa cháy, nơi độ rõ nét và chi tiết của hình ảnh được ưu tiên hàng đầu.

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1. Đặt vấn đề

Nghiên cứu này tập trung vào việc kết hợp hình ảnh ánh sáng hồng ngoại (IR) và ánh sáng nhìn thấy (VIS) nhằm nâng cao chất lượng hình ảnh thang xám, đặc biệt giải quyết các thách thức trong ứng dụng giám sát vào ban đêm và trong điều kiện sương mù. Hình ảnh IR rất có giá trị trong các tình huống ánh sáng yếu vì nó ghi lại sự biến đổi nhiệt, điều rất quan trọng để phát hiện các thực thể sống và các vật thể ấm khác. Mặt khác, hình ảnh VIS cung cấp thông tin kết cấu chi tiết trong điều kiện tầm nhìn bình thường và thấp như sương mù.

Vấn đề này đặc biệt quan trọng trong lĩnh vực giám sát, nơi khả năng phát hiện và nhận diện các đặc điểm trong mọi điều kiện tầm nhìn là tối quan trọng. Việc kết hợp hiệu quả hình ảnh IR và VIS thành một hình ảnh thang xám chi tiết duy nhất có thể nâng cao đáng kể khả năng giám sát, đảm bảo rằng các hệ thống giám sát có thể hoạt động hiệu quả bất kể các ràng buộc về tầm nhìn môi trường như bóng tối ban đêm hoặc sương mù. Việc nâng cao khía cạnh xử lý hình ảnh này sẽ dẫn đến các hệ thống giám sát mạnh mẽ hơn, cung cấp dữ liệu hình ảnh rõ ràng và giàu thông tin hơn, từ đó cải thiện các biện pháp an ninh và an toàn tổng thể.

1.2. Các giải pháp hiện tại và hạn chế

Hiện nay, lĩnh vực kết hợp hình ảnh hồng ngoại và hình ảnh khả kiến (IVIF - Infrared and Visible Image Fusion) đang thu hút sự chú ý đáng kể từ các nhà nghiên cứu, đánh dấu sự nổi lên của nó như một lĩnh vực nghiên cứu nổi bật. Mối quan tâm này được thúc đẩy bởi khả năng của IVIF trong việc tích hợp các đặc tính riêng biệt của hình ảnh hồng ngoại (IR) và hình ảnh nhìn thấy (VIS), cung cấp cái nhìn toàn diện mà từng loại hình ảnh riêng lẻ không thể mang lại.

Các phương pháp trong IVIF có thể chia thành hai nhóm: các phương pháp truyền thống và các phương pháp dựa trên học sâu. Các phương pháp truyền thống thường được áp dụng như biến đổi Wavelet rời rạc (Discrete Wavelet Transform – DWT), biến đổi kim tự tháp Laplacian Pyramid… Quy trình thực hiện chung của các phương pháp truyền thống này gồm 3 giai đoạn chính: phân rã hình ảnh với các phương pháp biến đổi thành thành phần tần số thấp và thành phần tần số cao, sử dụng các phương pháp tổng hợp để tổng hợp các thành phần và biến đổi ngược để thu được hình ảnh tổng hợp. Các phương pháp truyền thống này có ưu điểm trong việc tách biệt tốt hơn giữa kết cấu, các cạnh và thông tin không gian giúp giảm các lỗi ánh xạ nhưng đi kèm là độ phức tạp cao, có khả năng mất thông tin trong quá trình biến đổi. Hiệu suất hợp nhất có thể giảm nếu các quy tắc hợp nhất không phù hợp hoặc yếu được chọn.

Các phương pháp học sâu thường dựa trên mạng neural tích chập (Convolutional Neural Networks - CNNs), mạng đối kháng tạo sinh (Generative Adversarial Networks - GANs) và mạng tự mã hóa (Autoencoder – AEs), mang lại khả năng trích xuất đặc trưng mạnh mẽ và tăng cường năng lực kết hợp. Các phương pháp này tự động học cách

tối ưu hóa quy tắc kết hợp từ các tập dữ liệu lớn nhưng đòi hỏi tài nguyên tính toán đáng kể và dữ liệu huấn luyện rộng lớn, điều này có thể là một hạn chế trong các môi trường bị giới hạn tài nguyên.

Ví dụ, phương pháp NestFuse, một phương pháp dựa trên học sâu sử dụng kiến trúc mạng lồng ghép và mô hình chú ý để hợp nhất hình ảnh IR và VIS một cách hiệu quả, thể hiện hiệu suất kết hợp vượt trội so với các phương pháp truyền thống. Tuy nhiên, phương pháp này vẫn phải đối mặt với các thách thức như nhu cầu về các tập dữ liệu huấn luyện lớn và yêu cầu tính toán cao, điều thường thấy ở các phương pháp học sâu. Trong khi đó, các phương pháp truyền thống, mặc dù ít tốn tài nguyên tính toán hơn, lại gặp khó khăn về tính linh hoạt và không phải lúc nào cũng khai thác hết tiềm năng của dữ liệu hình ảnh do những hạn chế của các kỹ thuật chuyển đổi và phân rã được định trước.

1.3. Mục tiêu và định hướng giải pháp

Mục tiêu chính của nghiên cứu này là nâng cao khả năng tổng hợp hình ảnh hồng ngoại và hình ảnh khả kiến bằng cách phát triển một hệ thống lai kết hợp những ưu điểm của phương pháp phân rã kim tự tháp Laplacian truyền thống với các khả năng tiên tiến của kiến trúc học sâu NestFuse. Cụ thể, nghiên cứu thực hiện sử dụng phân rã kim tự tháp Laplacian để phân tích hai hình ảnh ban đầu thành thành phần cơ sở và các thành phần chi tiết. Tổng hợp thành phần cơ sở sử dụng năng lượng vùng cục bộ và sử dụng mô hình NestFuse cho tổng hợp thành phần chi tiết. Cuối cùng, biến đổi Laplacian ngược được thực hiện để thu được hình ảnh tổng hợp.

1.4. Đóng góp của đồ án

Nghiên cứu này đóng góp một số yếu tố mới trong tổng hợp hình ảnh hồng ngoại và hình ảnh khả kiến bằng cách tích hợp phân rã kim tự tháp Laplacian và mô hình học sâu NestFuse. Các đóng góp chính của nghiên cứu gồm:

- Phương pháp lai sáng tạo kết hợp những ưu điểm của phân rã kim tự tháp Laplacian với các khả năng tiên tiến của mô hình học sâu NestFuse, nâng cao chất lượng và hiệu quả của quá trình hợp nhất.
- Đề xuất một phương pháp phân rã kim tự tháp Laplacian và biến đổi Laplacian ngược mới hạn chế tính toán và sự mất thông tin khi thực hiện các phép Down và Subtract trong phương pháp phân rã cũ. Biến đổi đề xuất này giúp đơn giản hóa quá trình tính toán, bảo toàn các thông tin chi tiết và tăng cường tính toàn vẹn của thành phần cơ sở.
- Chiến lược tổng hợp kết hợp giữa khối bổ sung tính năng CMDAF và cơ chế chú ý không gian (Spatial Attention) trong giai đoạn tổng hợp tính năng của NestFuse nâng cao chất lượng hình ảnh tổng hợp.
- Kết hợp mặt nạ chứa những vùng nổi bật trong hình ảnh hồng ngoại thu được qua mô hình U2Net với quá trình tổng hợp thành phần cơ sở và bổ sung thông tin cho hình ảnh kết quả từ biến đổi Laplacian ngược nhằm giữ lại nhiều hơn thông tin quan trọng trong hình ảnh hồng ngoại.

1.5. Bố cục của đồ án

Trong các phần tiếp theo, tôi sẽ trình bày chi tiết hơn về nền tảng lý thuyết, các nghiên cứu liên quan, phương pháp đề xuất và các kết quả thực nghiệm. Cụ thể như sau:

Chương 2 trình bày về nền tảng lý thuyết và kết quả từ các nghiên cứu liên quan.

Chương 3 sẽ trình bày chi tiết về phương pháp đề xuất bao gồm về phương pháp phân rã kim tự tháp Laplacian biến thể, phương pháp MRE tổng hợp thành phần cơ sở, NestFuse cải tiến cho tổng hợp thành phần chi tiết và biến đổi Laplacian ngược để thu được hình ảnh tổng hợp.

Chương 4 sẽ trình bày chi tiết về các chỉ số đánh giá chất lượng của phương pháp và thực nghiệm, so sánh kết quả với các phương pháp SOTA hiện nay

Chương 5 trình bày về các kết luận và hướng phát triển tiếp theo cho nghiên cứu này.

CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT

Trong chương này, tôi sẽ trình bày chi tiết về các nghiên cứu liên quan kèm theo những ưu điểm, nhược điểm của các nghiên cứu này. Tôi cũng trình bày chi tiết nền tảng của nghiên cứu này là phương pháp phân rã kim tự tháp Laplacian và mô hình học sâu NestFuse.

2.1. Nghiên cứu liên quan

Trong các nghiên cứu về tổng hợp hình ảnh hồng ngoại và hình ảnh khả kiến, có hai hướng tiếp cận chính là các phương pháp truyền thống và phương pháp dựa trên học sâu. Các phương pháp truyền thống thường được áp dụng như biến đổi Wavelet rời rạc (Discrete Wavelet Transform – DWT), biến đổi kim tự tháp Laplacian Pyramid… trong khi các phương pháp học sâu dựa vào CNN, GAN, AEs và Transfomers.

2.1.1. Phương pháp truyền thống

Các phương pháp truyền thống chủ yếu dựa vào các kỹ thuật như biến đổi đa tỷ lệ (multi-scale transforms) và biểu diễn thưa (sparse representation), nhằm tối ưu hóa các quy trình trích xuất và kết hợp đặc trưng. Trong đó, biến đổi đa tỷ lệ là một phương pháp trọng tâm trong tổng hợp hình ảnh IR và VIS. Các thao tác cụ thể được áp dụng trên từng cấp độ để đạt được sự kết hợp tối ưu. Hình ảnh cuối cùng được tái tạo thông qua quá trình biến đổi ngược từ các hình ảnh đã được kết hợp ở từng cấp độ. Các phương pháp phổ biến trong nhóm này bao gồm kim tự tháp Laplacian cho phép thao tác chi tiết hình ảnh theo từng lớp, biến đổi sóng rời rạc (Discrete Wavelet Transform) được biết đến với khả năng xử lý các thành phần tần số khác nhau, biến đổi contourlet không lấy mẫu (NSCT - Non-Subsampled Contourlet Transform) và biến đổi shearlet không lấy mẫu (NSST - Non-Subsampled Shearlet Transform) được đề xuất để cải thiện việc kết hợp hình ảnh hồng ngoại và nhìn thấy, bằng cách nắm bắt các chi tiết định hướng và đặc điểm dị hướng (anisotropic) một cách hiệu quả hơn.

Một cách tiếp cận quan trọng khác là biểu diễn thưa (Sparse Representation). Kỹ thuật này biểu diễn mỗi hình ảnh dưới dạng tổ hợp tuyến tính của một tập hợp các vector cơ sở được xác định trước. Quá trình kết hợp sau đó sẽ kết hợp tuyến tính các biểu diễn thưa này để xây dựng hình ảnh cuối cùng.

Các phương pháp truyền thống này cung cấp một khung lý thuyết mạnh mẽ để giải quyết các phức tạp của IVIF. Mỗi phương pháp mang lại những lợi thế riêng, từ việc kiểm soát chính xác chi tiết hình ảnh ở nhiều cấp độ đến việc tăng cường các đặc trưng nổi bật. Điều này tạo tiền đề cho các chiến lược kết hợp tinh vi hơn, giúp cải thiện hiệu suất và tính ứng dụng trong nhiều tình huống thực tế khác nhau.

Tuy nhiên, bất chấp những lợi thế, các phương pháp này có những hạn chế cố hữu, chẳng hạn như khả năng mất thông tin trong quá trình biến đổi và sự thiếu linh hoạt trong việc thích nghi với các biến thể mới hoặc không lường trước của dữ liệu hình ảnh. Điều này đã dẫn đến sự quan tâm ngày càng tăng đối với việc khám phá các kỹ

thuật thích ứng hơn, có thể điều chỉnh động theo các điều kiện hình ảnh khác nhau mà không cần tinh chỉnh thủ công quá nhiều.

2.1.2. Phương pháp dựa trên học sâu

Các phương pháp kết hợp dựa trên học sâu (deep learning) đã mang lại những tiến bộ đáng kể trong lĩnh vực kết hợp hình ảnh hồng ngoại (IR) và nhìn thấy (VIS), giải quyết được những thách thức mà các kỹ thuật truyền thống gặp khó khăn, chẳng hạn như việc trích xuất và tích hợp đặc trưng một cách thích ứng. Các phương pháp này tận dụng các kiến trúc mạng nơ-ron tiên tiến, có khả năng học động từ các tập dữ liệu lớn để tối ưu hóa chiến lược kết hợp, nhờ đó cải thiện tính thích nghi và chất lượng tổng thể của hình ảnh sau khi được kết hợp.

Mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs) dẫn đầu trong các phương pháp này. CNNs nổi bật trong việc trích xuất các hệ thống phân cấp đặc trưng không gian thông qua cấu trúc nhiều lớp, đặc biệt hữu ích để duy trì tính nhất quán không gian của hình ảnh được kết hợp. Ví dụ như trong STDFusionNet [], kiến trúc mô hình gồm hai phần chính là mạng trích xuất đặc trưng và mạng tái tạo đặc trưng. Mạng trích xuất đặc trưng sử dụng các ResBlock để tăng khả năng trích xuất và khắc phục vấn đề mất gradient. Mạng tái tạo đặc trưng gồm bốn ResBlock để hợp nhất đặc trưng và tái tạo ảnh. PMGI [] đề xuất một mạng hợp nhất end-to-end, mô hình hóa vấn đề hợp nhất ảnh như một bài toán bảo toàn kết cấu và cường độ điểm ảnh. Phương pháp này sử dụng hai nhánh riêng biệt để trích xuất thông tin phân bố gradient và cường độ từ ảnh nguồn. Để duy trì mối tương quan giữa hai loại thông tin này, đầu vào cho cả hai nhánh là ảnh hồng ngoại và ảnh nhìn thấy, được ghép nối theo một tỷ lệ cố định. Một module hợp nhất kênh được thêm vào trước các lớp tích chập thứ ba và thứ tư để nâng cao khả năng trích xuất thông tin.

Mạng đối kháng tạo sinh (Generative Adversarial Networks - GANs) cũng được sử dụng để cải thiện tính thực tế của hình ảnh kết hợp. MgAN-Fuse [] đề xuất phương pháp mã hóa hai hình ảnh bằng hai bộ mã hóa riêng biệt để huấn luyện được các đặc trưng riêng của từng hình ảnh. Đồng thời, kết hợp thêm một mô-đun chú ý đa tỉ lệ để khai thác toàn diện các đặc trưng của các lớp đa tỉ lệ và buộc mô hình tập trung vào các vùng phân biệt. Mô hình từ cơ sở GAN với cấu trúc SGMD (một Generator và hai Discriminator).

Autoencoders được sử dụng trong các phương pháp như NestFuse [], tích hợp cấu trúc mạng lồng nhau trong khung làm việc của autoencoder. Cách tiếp cận này giúp nắm bắt và tái tạo hiệu quả các đặc trưng nổi bật từ cả hai loại hình ảnh, giảm thiểu đáng kể việc mất mát thông tin quan trọng và đảm bảo quá trình kết hợp hiệu quả và chính xác.

Transformer là một hướng tiếp cận mới nổi bật gần đây. SBIT-Fuse []đề xuất một phương pháp hợp nhất Symmetrical Bilateral Interaction and Transformer đơn giản và hiệu quả để xây dựng mạng tương tác hai luồng. Một module tương tác hai chiều đối xứng (Symmetrical Bilateral Interaction - SBI), bao gồm một số lớp tương tác kích hoạt

giữa các miền (Cross Domain Activation Interaction - CDAI) nối tiếp. Trong đó, thông tin không hoạt động của bộ điều chỉnh ReLU được chuyển từ một luồng này sang luồng khác thay vì bị loại bỏ.

Mặc dù các mô hình học sâu đã mang lại những tiến bộ vượt bậc, vẫn còn tồn tại một số thách thức như yêu cầu lượng dữ liệu huấn luyện lớn, chi phí tính toán cao đặc biệt đối với các mô hình có kiến trúc phức tạp, khó khăn trong việc điều chỉnh mô hình. Hướng nghiên cứu trong tương lai có thể bao gồm việc phát triển các kỹ thuật huấn luyện hiệu quả hơn, khám phá các mô hình học không giám sát (unsupervised learning) hay tạo ra các phương pháp lai (hybrid) kết hợp ưu điểm của cả các phương pháp truyền thống và học sâu để xây dựng hệ thống kết hợp hình ảnh mạnh mẽ, có khả năng mở rộng và hiệu quả.

2.2. Giới thiệu về Laplacian Pyramid và mô hình NestFuse

2.2.1. Laplacian Pyramid

Laplacian Pyramid (Kim tự tháp Laplacian) là một kỹ thuật đa độ phân giải cơ bản, được sử dụng rộng rãi trong kết hợp hình ảnh và đặc biệt được đánh giá cao nhờ hiệu quả trong việc kết hợp hình ảnh từ các mức tiêu điểm khác nhau để tăng cường chi tiết ở nhiều cấp độ. Quy trình của kỹ thuật này bao gồm các bước như sau:

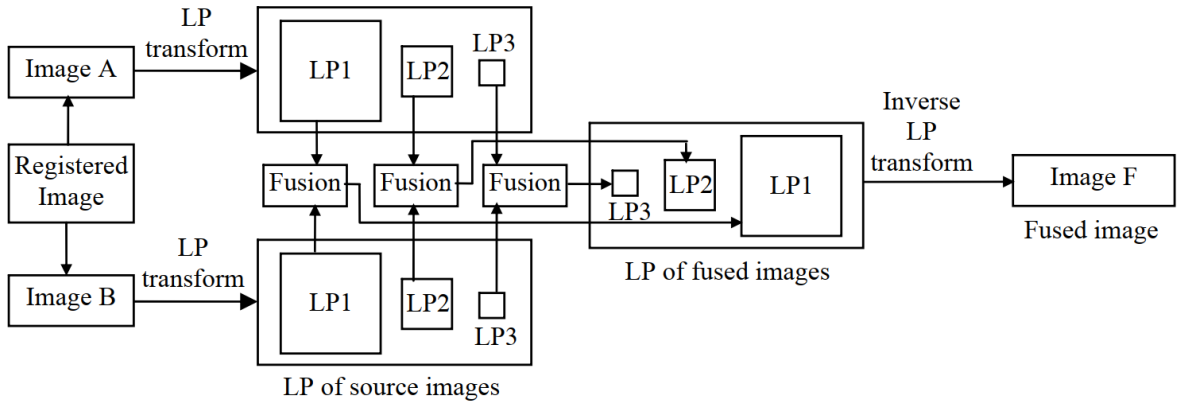
i. Xây dựng Gaussian Pyramid (Kim tự tháp Gaussian): Bắt đầu bằng việc tạo các bản sao của hình ảnh gốc với độ phân giải giảm dần qua các lớp. Đây là bước chuẩn bị để trích xuất các thông tin quan trọng ở từng cấp độ.

ii. Tạo Laplacian Pyramid: Mỗi lớp trong Gaussian Pyramid sẽ được trừ đi từ phiên bản mở rộng (upsampled) của lớp tiếp theo ở cấp độ cao hơn. Kết quả là một Laplacian Pyramid, chứa các dải thông tin tần số cụ thể, đại diện cho các chi tiết ở các mức độ khác nhau.

iii. Quá trình kết hợp: Trong quá trình hợp nhất (fusion), các dải tần số này từ các Laplacian Pyramid của các hình ảnh nguồn được chọn lọc và kết hợp. Điều này đảm bảo rằng thông tin quan trọng từ cả hai hình ảnh được bảo tồn và nhấn mạnh.

iv. Tái tạo hình ảnh: Sau khi kết hợp xong, hình ảnh đầu ra được tái tạo từ Laplacian Pyramid đã hợp nhất, giúp giữ lại các chi tiết tần số cao. Đây là yếu tố quan trọng cho các ứng dụng đòi hỏi độ rõ nét và độ phân giải chi tiết được cải thiện.

Kết quả thu được hình ảnh cuối cùng mang lại chất lượng vượt trội về mặt chi tiết, rất phù hợp cho các ứng dụng yêu cầu độ rõ ràng hình ảnh cao và khả năng tái tạo các chi tiết nhỏ. Hình 1 thể hiện sơ đồ trực quan thể hiện luồng thực hiện của các phương pháp dựa trên Laplacian Pyramid. Công thức chi tiết của Laplacian Pyramid sẽ được trình bày trong chương 3 kèm theo biến thể của nó mà nghiên cứu này đề xuất.

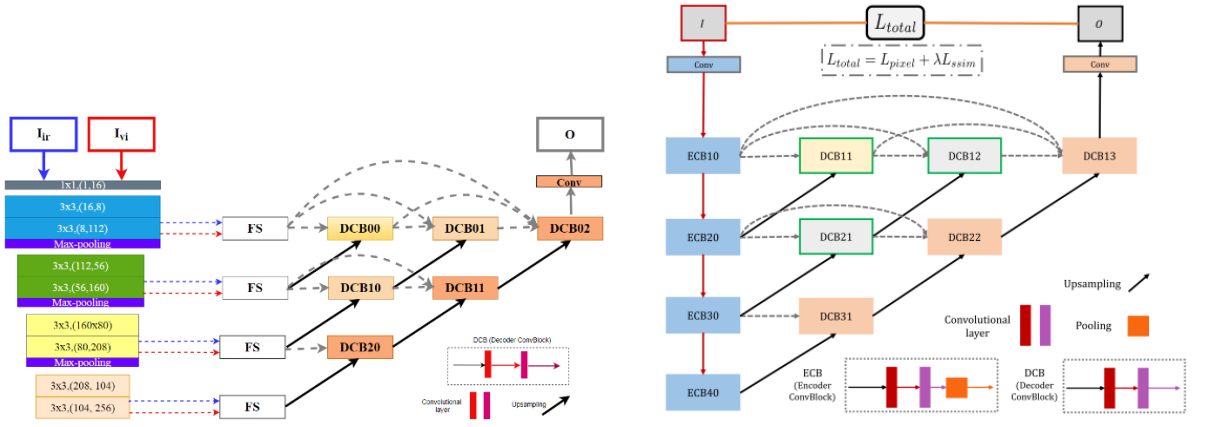


Hình 1. Khung thực hiện tổng hợp hình ảnh dựa trên biến đổi Laplacian Pyramid

2.2.2. NestFuse

NestFuse (mạng hợp nhất dư lồng nhau) là một mô hình học sâu tiên tiến được thiết kế đặc biệt cho nhiệm vụ hợp nhất hình ảnh hồng ngoại và khả kiến ở nhiều tỷ lệ khác nhau. Nó áp dụng một khung học dư (residual learning framework) để tăng cường tích hợp đặc trưng và bảo toàn các chi tiết quan trọng mà không bị suy giảm như thường thấy trong các phương pháp trung bình đơn giản.

Kiến trúc của mô hình NestFuse được thể hiện chi tiết trong hình 2. Trong đó, mỗi block Encoder và Decoder đều chứa các layer Convolution (mỗi khối chứa 2 layer), riêng các khối Encoder có thêm một layer MaxPooling để có được nhiều tỷ lệ khác nhau. NestFuse dành cho pha train không chứa block FS (Fusion Strategy) và được huấn luyện trên bộ dữ liệu MSCOCO [4] để học được một mô hình trích xuất các đặc trưng và tái tạo hình ảnh từ các đặc trưng đó.



a) Mô hình tổng hợp thành phần chi tiết *b) Mô hình cho quá trình huấn luyện*

Hình 2. Kiến trúc mô hình NestFuse

Quá trình huấn luyện của NestFuse[] bao gồm hai giai đoạn chính. Giai đoạn đầu tiên, bộ mã hóa (encoder) và bộ giải mã (decoder) được huấn luyện cùng nhau dưới

dạng một auto-encoder. Mục tiêu của giai đoạn này là tái tạo chính xác các hình ảnh đầu vào, từ đó cải thiện khả năng của mạng trong việc trích xuất và tái tạo các đặc trưng một cách hiệu quả. Giai đoạn tiếp theo, các mô hình chú ý không gian (spatial attention) và kênh (channel attention) được tích hợp và huấn luyện để hợp nhất các đặc trưng sâu đa tỷ lệ (multi-scale deep features) một cách hiệu quả. Các mô hình chú ý này tập trung vào các khía cạnh quan trọng cả về không gian và kênh, đảm bảo rằng các đặc trưng quan trọng nhất được nhấn mạnh trong đầu ra hợp nhất.

Hàm mất mát được sử dụng trong quá trình huấn luyện được định nghĩa như sau:

$$L_{total} = L_{pixel} + \lambda L_{ssim}$$

trong đó, $L_{pixel} = \|O - I\|^2$ đo lường lỗi tái tạo theo từng điểm ảnh với O là hình ảnh đầu ra và I là hình ảnh đầu vào. $L_{ssim} = 1 - SSIM(O, I)$ tính toán mất mát dựa trên độ tương đồng cấu trúc (structural similarity loss) giữa hình ảnh đầu ra và hình ảnh đầu vào. λ là tham số điều chỉnh (trade-off parameter) giữa hai thành phần của hàm mất mát, giúp cân bằng giữa tái tạo chi tiết điểm ảnh và bảo toàn cấu trúc hình ảnh.

Tập dữ liệu được sử dụng để huấn luyện là MSCOCO[9], bao gồm 80.000 hình ảnh đã được chuyển đổi thành ảnh xám (grayscale) và thay đổi kích thước về 256 × 256 pixel. Điều này chứng minh khả năng tổng quát hóa của phương pháp từ một tập hợp đa dạng các hình ảnh.

Chiến lược hợp nhất của NestFuse tận dụng sự kết hợp giữa cơ chế chú ý không gian (spatial attention) và chú ý kênh (channel attention) để tích hợp hiệu quả các đặc trưng sâu đa tỷ lệ (multi-scale deep features) từ hình ảnh hồng ngoại và hình ảnh thường.

NestFuse là một mô hình tiên tiến nhờ việc sử dụng các kỹ thuật học sâu và cơ chế chú ý một cách tinh vi. Mô hình này vượt trội so với các phương pháp truyền thống không chỉ ở việc tăng cường chi tiết và chất lượng của hình ảnh hợp nhất mà còn đảm bảo duy trì tính toàn vẹn và tính hữu ích của thông tin từ cả hai nguồn đầu vào. Điều này khiến NestFuse đặc biệt hữu ích trong các ứng dụng đòi hỏi khả năng biểu diễn hình ảnh chất lượng cao, chẳng hạn như trong giám sát an ninh (surveillance), chẩn đoán y tế (medical imaging), và viễn thám (remote sensing).

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

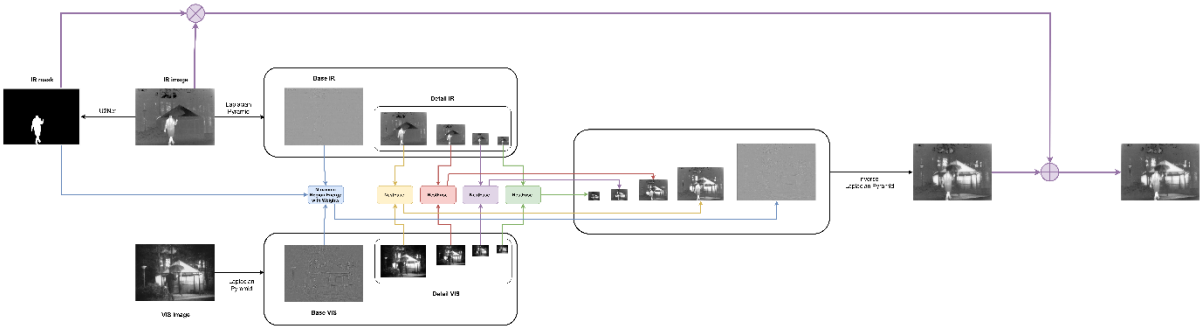
3.1. Tổng quan giải pháp

Trong báo cáo này, kiến trúc tổng quan của phương pháp đề xuất được minh họa trong hình 3. Đầu tiên, hai hình ảnh hồng ngoại và khả kiến được đưa qua Laplacian Pyramid biến thể để tạo ra thành phần cơ sở và các thành phần chi tiết. Ngoài ra, hình ảnh hồng ngoại được đưa qua mô hình U2Net [1] đã được pretrain để trích xuất mặt nạ chứa thành phần nổi bật (như người, xe,…) nhằm phụ trợ cho quá trình hợp nhất thành phần cơ sở và bổ sung thông tin từ ảnh hồng ngoại cho hình ảnh tổng hợp.

Tiếp theo, thành phần cơ sở từ hai hình ảnh được tổng hợp bằng phương pháp năng lượng vùng cực đại kết hợp điều kiện về vùng nổi bật (mặt nạ) của hình ảnh hồng ngoại để đưa ra thành phần cơ sở tổng hợp. Các thành phần chi tiết được tổng hợp dựa trên mô hình NestFuse [2]. Đây là một mô hình theo kiến trúc AutoEncoder, mô hình chi tiết sẽ được trình bày trong các mục tiếp theo. Ngoài ra, để nâng cao chất lượng hình ảnh, một module CMDAF [3] kết hợp với Spatial Attention được đề xuất để tổng hợp các đặc trưng trích xuất được từ Encoder, sau đó đưa qua khối Decoder để có thành phần chi tiết tổng hợp.

Sau khi thu được thành phần cơ sở tổng hợp và các thành phần chi tiết tổng hợp, biến đổi Laplacian Pyramid ngược được áp dụng để thu được hình ảnh tổng hợp từ các thành phần. Đồng thời, nhằm bổ sung các chi tiết nhiệt nổi bật từ hình ảnh hồng ngoại, hình ảnh tổng hợp vừa thu được được kết hợp với vùng nổi bật trong ảnh hồng ngoại (vùng mặt nạ) để thu được hình ảnh tổng hợp cuối cùng.

Phần tiếp theo trình bày chi tiết về các phần trong mô hình bao gồm: Laplacian Pyramid biến thể, MRE (Maximum Region Energy) cho tổng hợp thành phần cơ sở, NestFuse cải tiến cho tổng hợp thành phần chi tiết và phương pháp kết hợp thông tin nổi bật trong ảnh hồng ngoại cho ảnh tổng hợp cuối cùng.



Hình 3. Kiến trúc tổng quan

3.2. Laplacian Pyramid biến thể

Trong phương pháp biến đổi Laplacian Pyramid (LP), với hình ảnh xám đầu vào $I_0 \epsilon \mathbb{R}^{2^n \times 2^n}$, việc xây dựng kim tự tháp Gaussian là bước đầu tiên để tạo ra một LP gồm $n - lớp$. Sử dụng một hạt nhân Gaussian cố định, hình ảnh liên tục được lọc và giảm độ phân giải để tạo ra một kim tự tháp Gaussian $[G_0, G_1, \ldots, G_n]$, trong đó G_0 là hình

ảnh gốc và $G_i \in \mathbb{R}^{2^{n-i} \times 2^{n-i}}$. G_n là hình ảnh có độ phân giải thấp nhất, là thành phần tần số thấp của LP. Quá trình tạo G_{i+1} từ G_i được biểu diễn như sau:

$$G_{i+1} = Down(M * G_i)$$

trong đó M đại diện cho một hạt nhân Gaussian cố định, $*$ là phép tích chập và Down đại diện cho quá trình giảm độ phân giải theo tỷ lệ 2.

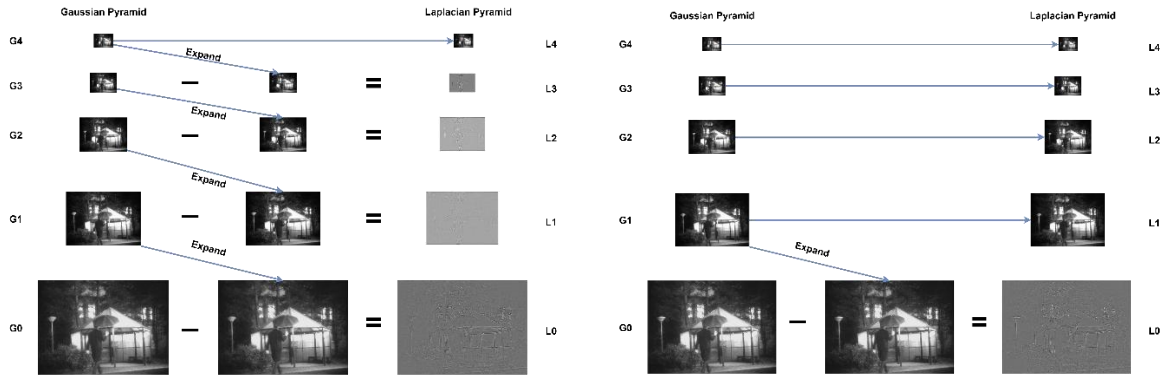
Các thành phần tần số cao của LP truyền thống được xây dựng theo quy trình biểu diễn như sau:

$$H_i = G_i - expand(G_{i+1})$$

trong đó $G_i^* = expand(G_{i+1})$ với hạt nhân M có kích thước $(2k+1) \times (2k+1)$ tuân theo công thức:

$$G_i^*(x,y) = 4 \sum_{m=-k}^{k} \sum_{n=-k}^{k} M(m,n) G_{i+1}\left(\frac{x+m}{2}, \frac{y+n}{2}\right)$$

Quá trình biến đổi LP truyền thống cuối cùng thu được kim tự tháp gồm $[H_0, H_1, \ldots, H_{n-1}, G_n]$. Tuy nhiên, phương pháp biến đổi này yêu cầu thực hiện phép expand thông qua nhân tích chập nhiều lần làm tốn tài nguyên tính toán và có thể bị mất thông tin do thực hiện phép Down sau đó là phép trừ cho ảnh đã expand để có thành phần chi tiết. Do đó, báo cáo này đề xuất một phương pháp biến đổi LP biến thể, trong đó kim tự tháp LP cuối cùng thu được gồm $[L_0, G_1, \ldots, G_n]$ trong đó $L_0 = G_0 - expand(G_1)$, tức là giữ nguyên n thành phần cuối cùng của kim tự tháp Gaussian làm thành phần chi tiết còn thành phần cơ sở được xây dựng qua phép trừ ảnh gốc cho ảnh mở rộng từ G_1. Biến thể của LP đề xuất này giúp đơn giản hóa quá trình tính toán, bảo toàn các thông tin chi tiết và tăng cường tính toàn vẹn của thành phần cơ sở. Hình 4 thể hiện sự khác biệt trong phép biến đổi LP truyền thống và LP đề xuất.



a) LP truyền thống *b) LP đề xuất*

Hình 4: So sánh sơ đồ xây dựng LP truyền thống và LP đề xuất

3.3. Maximum Region Energy (MRE) tổng hợp thành phần cơ sở

Phương pháp tổng hợp thành phần cơ sở dựa trên kết hợp năng lượng vùng tối đa, được thiết kế để tối đa hóa năng lượng cục bộ trong các vùng của thành phần cơ sở từ hình ảnh đầu vào. Phương pháp này xác định và hợp nhất các vùng có năng lượng cao nhất, đảm bảo rằng các đặc điểm nổi bật nhất, chẳng hạn như giá trị cường độ cao hơn trong hình ảnh hồng ngoại hoặc kết cấu chi tiết trong hình ảnh nhìn thấy được, được thể hiện nổi bật trong đầu ra hợp nhất.

Ngoài ra, trong báo cáo này, tôi sử dụng thêm một mặt nạ nhị phân được tạo từ hình ảnh hồng ngoại thông qua mô hình U2Net [1] đã được huấn luyện trước. Mô hình này sẽ phát hiện ra những vùng chứa mục tiêu nổi bật như người, xe cộ …, hỗ trợ việc nắm bắt chi tiết các đặc điểm nổi bật trong hình ảnh hồng ngoại mà việc sử dụng MRE có thể bỏ qua. Hình 5 thể hiện mặt nạ được xác định.

Chi tiết quy trình tổng hợp thành phần cơ sở dựa trên MRE và mặt nạ như sau:

Bước 1 – Tính toán năng lượng vùng cục bộ cho hai thành phần cơ sở từ hình ảnh hồng ngoại và khả kiến theo công thức sau:

$$RE_{IR}(m,n) = \sum_{(m\prime,n\prime) \in W} \omega_{m\prime n\prime} \left[L_{0,IR}(m+m',n+n')\right]^2$$

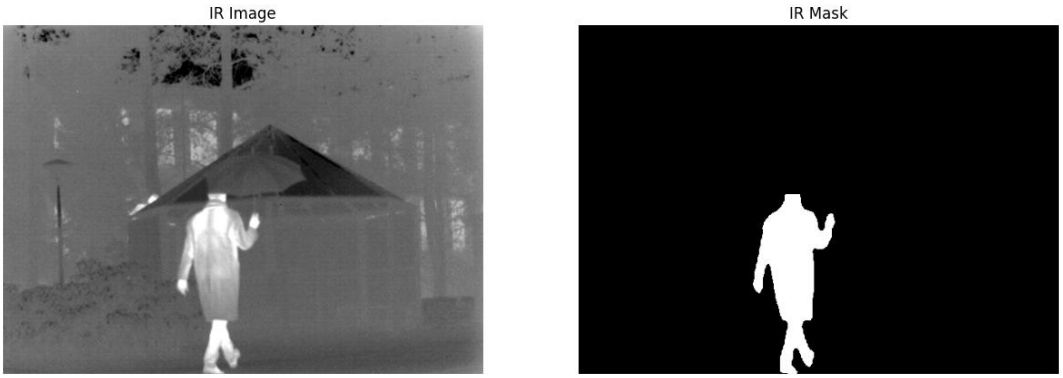
$$RE_{VIS}(m,n) = \sum_{(m\prime,n\prime) \in W} \omega_{m\prime n\prime} \left[L_{0,VIS}(m+m',n+n')\right]^2$$

trong đó W là cửa số vùng cục bộ (có thể là 3×3 hoặc 5×5) và ω là trọng số ứng với các pixel trong vùng cục bộ. Trong báo cáo này, tôi sử dụng cửa sổ W có kích thước 3×3 và bộ trọng số $\omega = \frac{1}{16}\begin{bmatrix}1 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 1\end{bmatrix}$.

Bước 2 – Tổng hợp thành phần cơ sở dựa trên năng lượng vùng và mặt nạ mask

$$L_0^{fused}(i,j) = \begin{cases} L_{0,IR}(i,j) & nếu\ RE_{IR}(i,j) \geq RE_{VIS}(i,j)\ hoặc\ mask(i,j) == 1 \\ L_{0,VIS}(i,j) & nếu\ ngược\ lại \end{cases}$$

Phương pháp này tăng cường hiệu quả hình ảnh hợp nhất bằng cách đảm bảo rằng các vùng có chi tiết nhiệt hoặc hình ảnh trực quan quan trọng nhất được làm nổi bật, đặc biệt có giá trị trong các ứng dụng đòi hỏi độ trung thực chi tiết và độ tương phản cao, chẳng hạn như trong giám sát nâng cao hoặc các hệ thống theo dõi quan trọng.

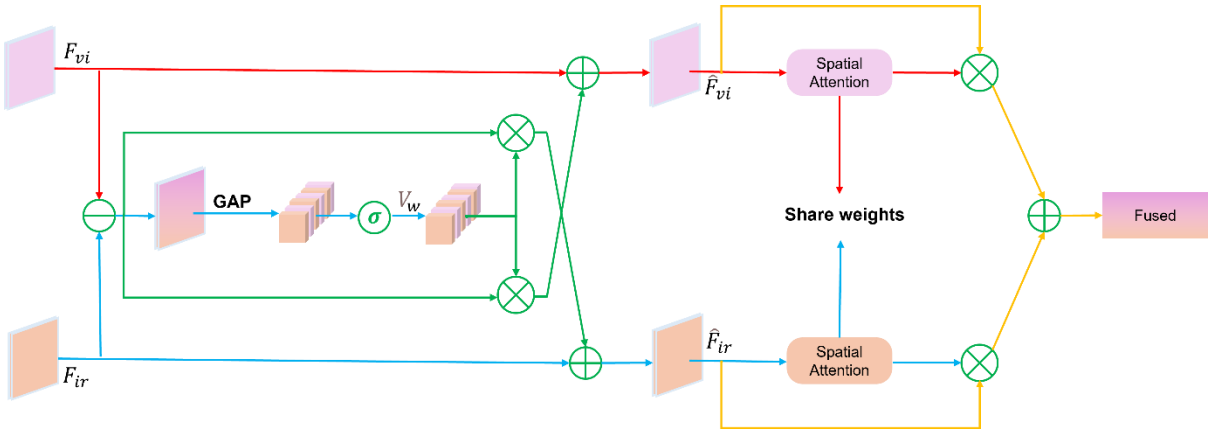


Hình 5. Mô hình U2Net trích xuất mặt nạ chứa các đối tượng nổi bật trong ảnh hồng ngoại

3.4. NestFuse cho tổng hợp các thành phần chi tiết

Đối với các thành phần chi tiết, mô hình NestFuse [] được sử dụng để thực hiện hợp nhất. Mô hình NestFuse hoạt động bằng cách lấy mỗi lớp chi tiết tương ứng từ các kim tự tháp Laplacian của các hình ảnh đầu vào, trích xuất các đặc trưng thông qua khối Encoder đã được huấn luyện, sau đó sử dụng một chiến lược tổng hợp (Fusion Strategy – FS) để tổng hợp đặc trưng, các đặc trưng tổng hợp được đưa qua khối Decoder để thu được thành phần chi tiết tổng hợp. Cách tiếp cận này không chỉ duy trì các chi tiết tần số cao mà còn đảm bảo rằng các sắc thái của cả hai kiểu hình ảnh được nắm bắt phù hợp.

Trong nghiên cứu này, khác với bài báo gốc, tôi thực hiện huấn luyện lại mô hình trên bộ dữ liệu MSRS [5] và thực hiện tăng cường dữ liệu bằng các phép lật, xoay, biến đổi affine, phép cắt thay vì huấn luyện trên bộ dữ liệu MSCOCO. Mục đích nhằm xây dựng một mô hình riêng cho trích xuất đặc trưng của hình ảnh hồng ngoại và khả kiến thay vì một bộ dữ liệu tổng quát. Ngoài ra, tôi đề xuất một chiến lược tổng hợp (Fusion Strategy – FS) mới thay vì tổng hợp dựa trên cơ chế Channel Attention và Spatial Attention như mô hình NestFuse đã thực hiện.



Hình 6. Chiến lược tổng hợp đặc trưng đề xuất (FS)

Chiến lược tổng hợp mới mà tôi đề xuất kết hợp giữa khối CMDAF [6] và cơ chế Spatial Attention. Hình 6 thể hiện chi tiết quy trình thực hiện của chiến lược này. Khối CMDAF thực hiện tăng cường chi tiết và tương quan giữa hai thành phần từ hình ảnh hồng ngoại và hình ảnh khả kiến thông qua thực hiện cơ chế Channel Attention trên thành phần hiệu giữa hai thành phần chi tiết. Kết quả thu được sau khi áp dụng Channel Attention được cộng vào hai thành phần. Tiếp theo, hai thành phần đầu ra từ khối CMDAF thực hiện tính trọng số theo không gian (Spatial Attention) bằng cách lấy giá trị trung bình theo kênh. Hai ma trận trọng số không gian từ hai thành phần (giả sử là sa_{ir} và sa_{vis}) thực hiện chia sẻ thông qua hàm exp theo công thức (1) để tính trọng số cuối cùng cho mỗi thành phần và tổng hợp các thành phần.

$$w_{ir} = \frac{e^{sa_{ir}}}{e^{sa_{ir}} + e^{sa_{vis}}} ; \; w_{vis} = \frac{e^{sa_{vis}}}{e^{sa_{ir}} + e^{sa_{vis}}}$$

3.5. Tổng hợp các thành phần

Mục 3.3 và mục 3.4 đã trình bày chi tiết về các bước thực hiện tổng hợp thành phần chi tiết và thành phần cơ sở. Sau khi thu được kết quả tổng hợp các thành phần trên, tái cấu trúc hình ảnh được thực hiện thông qua Laplacian Pyramid ngược. Giả sử, sau quá trình tổng hợp các thành phần ta thu được kim tự tháp $[L_0, L_1, \ldots, L_n]$, trong đó L_0 là thành phần tổng hợp cơ sở và $L_1, \ldots, L_n$ là thành phần tổng hợp chi tiết. Quy trình tái cấu trúc hình ảnh được xây dựng theo các bước sau và hình 7 thể hiện sơ đồ chi tiết của quá trình tái cấu trúc hình ảnh:

Bước 1 - Khởi tạo cấu trúc: Tại bước này, hình ảnh tái cấu trúc được khởi tạo bằng với thành phần L_n – level cuối cùng của LP

$$I_{reconstruct} = L_n$$

Bước 2 – Tính sharpness scores của n levels cuối: Độ sắc nét của n levels cuối được xác định thông qua giá trị tuyệt đối của kết quả áp dụng toán tử Laplace lên level đó

$$S_i = |\nabla L_i| \; \forall \, i = 1 \ldots n$$

Bước 3 – Tính tổng có trọng số cho các thành phần chi tiết

$$I_{reconstruct} = \sum_{i=1}^{n} \left(\frac{S_i}{\sum_{i=1}^{n} S_i} \times L_i\right) + \left(1 - \frac{S_i}{\sum_{i=1}^{n} S_i}\right) \times Expand(I_{reconstruct})$$

Bước 4 – Kết hợp với thành phần cơ sở

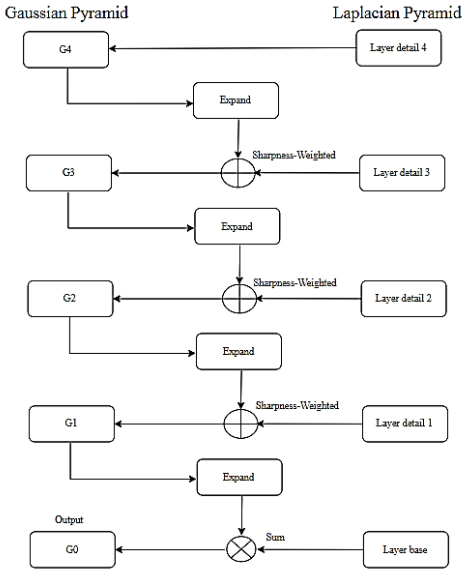
$$I_{reconstruct} = expand(I_{reconstruct}) + L_0$$

Quá trình này đảm bảo rằng hình ảnh cuối cùng giữ lại tất cả thông tin và chi tiết cần thiết từ hình ảnh đầu vào, được tích hợp trên tất cả các mức độ phân giải. Sự kết hợp có trọng số dựa trên độ sắc nét giúp bảo toàn các chi tiết quan trọng nhất, nâng cao chất lượng tổng thể của hình ảnh được tái tạo.

Ngoài ra, tôi bổ sung thông tin vùng nổi bật từ hình ảnh hồng ngoại (trích xuất thông qua mặt nạ như ở mục 3.3) vào hình ảnh thu được từ LP ngược ở trên. Tại những vùng nổi bật, giá trị hình ảnh cuối cùng được bổ sung như sau:

$$I = \gamma I_{reconstruct} + (1-\gamma) I_{IR}$$

Giá trị γ được chọn bằng 0.3. Chi tiết về thông số này được trình bày trong phần thực nghiệm. Điều này giúp hình ảnh tổng hợp giữ nhiều thông tin bổ sung nổi bật hơn từ ảnh hồng ngoại.



Hình 7. Sơ đồ thực hiện tái cấu trúc hình ảnh từ Laplacian Pyramid

CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM

4.1. Tập dữ liệu thực nghiệm

Trong thực nghiệm của nghiên cứu này, tôi sử dụng phương pháp đề xuất trên và mô hình NestFuse cải tiến đã được huấn luyện trước trên bộ dữ liệu MSRS[] kèm theo các phương pháp tăng cường dữ liệu. Để kiểm tra, tôi áp dụng phương pháp các cặp hình ảnh được chọn từ tập dữ liệu TNO[10]. Cụ thể, tôi đã chọn 42 cặp hình ảnh từ tập dữ liệu TNO, tất cả đều đã được chuyển đổi sang dạng ảnh xám. Tập dữ liệu TNO, thường được sử dụng trong IVIF (hợp nhất hình ảnh hồng ngoại và nhìn thấy), bao gồm nhiều tình huống liên quan đến quân sự. Sự phong phú và phổ biến của tập dữ liệu TNO trong đánh giá IVIF làm tiền đề vững chắc cho các kết quả của nghiên cứu này.

4.2. Cài đặt thực nghiệm

Các thí nghiệm được thực hiện trên hệ điều hành Windows 11, tận dụng nền tảng mạnh mẽ và hỗ trợ cho tất cả các công cụ và thư viện được sử dụng. Về phần mềm, cấu hình bao gồm Anaconda phiên bản 24.5.0, đóng vai trò là hệ thống quản lý môi trường và gói, giúp cài đặt chính xác các thư viện cần thiết. Python phiên bản 3.10.14 được sử dụng, được chọn vì tính tương thích với các thư viện xử lý dữ liệu và học máy tiên tiến. Khung học sâu PyTorch phiên bản 2.3.1 đã được áp dụng nhờ tính linh hoạt và hiệu quả trong các thao tác tensor và mô hình hóa mạng nơ-ron. Ngoài ra, NumPy phiên bản 1.24.3 cũng được tích hợp nhờ khả năng tính toán số vượt trội, rất quan trọng trong việc quản lý tập dữ liệu lớn và các phép toán ma trận phức tạp đặc trưng trong các nhiệm vụ xử lý hình ảnh.

Về cấu hình phần cứng, bộ xử lý Intel Core i5-1335U với tốc độ xung nhịp cơ bản 1.3 GHz được sử dụng để đảm bảo hiệu suất chung mạnh mẽ. Việc huấn luyện mô hình học sâu NestFuse được thực hiện trên môi trường Google Colab.

4.3. Các chỉ số đánh giá

4.3.1. Mutual Information (MI)

4.3.2. NCIE

4.3.3. QG

4.3.4. SSIM

4.3.5. PSNR

4.3.6. EN

4.3.7. AG

4.3.8. SD

4.3.9. ALI

4.3.10. VIF

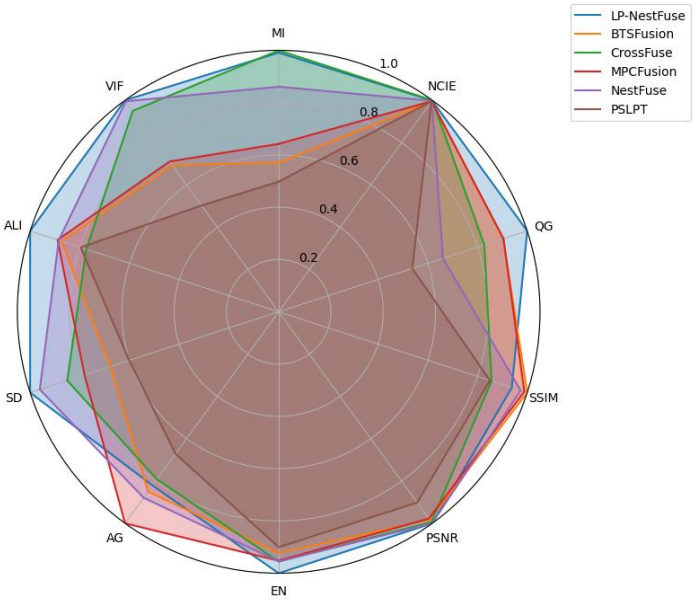
4.4. Kết quả và so sánh

Trong phần này, tôi sẽ trình bày chi tiết về so sánh thực nghiệm phương pháp tôi đề xuất là LP-NestFuse với các phương pháp SOTA được nghiên cứu gần đây gồm **SFDFusion**[](2024), **BTSFusion**[](2024), **MPCFusion**[](2024), **SwinFusion**[](2024), **CrossFuse**[](2024), **DAF-Net**[](2024), **PSLPT**[](2024), và mô hình mà tôi làm nền tảng là **NestFuse**[](2020). Bên cạnh đó, tôi thực hiện so sánh sự tác động của việc thay đổi chiến lược tổng hợp trong NestFuse và việc huấn luyện trên bộ dữ liệu MSRS tăng cường thay vì chiến lược tổng hợp cũ và huấn luyện trên bộ tổng quát MSCOCO.

Ngoài ra, tôi còn thực nghiệm so sánh các phương pháp tổng hợp khác nhau cho thành phần cơ sở như dựa trên độ lệch chuẩn kết hợp entropy, dựa trên năng lượng vùng cục bộ thay đổi bộ trọng số, dựa trên bản đồ trọng số tổng hợp, dựa trên năng lượng Laplacian cục bộ, dựa trên bộ lọc hướng dẫn kết hợp Laplacian sửa đổi và dựa trên mô hình VGG19 đã pretrain để khẳng định về mặt định lượng cho phương pháp tổng hợp thành phần cơ sở mà tôi đã chọn. Đồng thời, về trọng số λ bổ sung thông tin từ mặt nạ hình ảnh hồng ngoại cho hình ảnh tổng hợp cuối cùng, tôi thực hiện thử nghiệm với các trọng số $\lambda \in [0.1, 0.2, 0.3, 0.4, 0.5]$ để đưa ra kết luận cuối cùng với giá trị $\lambda = 0.3$.

	MI	NCIE	QG	SSIM	PSNR	EN	AG	SD	ALI	VIF
LP-NestFuse	2.8512	0.8068	0.5427	0.4016	7.1866	7.2536	0.0170	0.1829	0.5183	0.9214
BTSFusion	1.6411	0.8038	0.4903	0.4282	7.0667	6.6975	0.0176	0.1240	0.4530	0.6385
CrossFuse	2.8767	0.8079	0.4490	0.3667	7.1442	6.9289	0.0164	0.1556	0.4012	0.8752
MPCFusion	1.8472	0.8043	0.4908	0.4232	7.0263	6.9064	0.0207	0.1429	0.4610	0.6545
NestFuse	2.4761	0.8057	0.3587	0.4169	7.1825	6.9247	0.0182	0.1758	0.4586	0.9167
PSLPT	1.4314	0.8036	0.2917	0.3640	6.4793	6.5320	0.0139	0.1100	0.4128	0.4628

Bảng 1. So sánh giữa phương pháp đề xuất và các phương pháp khác



Hình 8. Biểu đồ radar so sánh giữa phương pháp đề xuất và các phương pháp khác

Kết quả so sánh giữa phương pháp đề xuất LP-NestFuse và các phương pháp khác được thể hiện trong bảng 1 kèm theo biểu đồ radar trực quan về khoảng tỷ lệ $[0,1]$ trong hình 8, là giá trị trung bình các chỉ số được đánh giá trên 42 cặp ảnh của bộ dữ liệu TNO. Trong bảng 1, ô màu xanh lá cây thể hiện giá trị tốt nhất, ô màu xanh lam thể hiện giá trị tốt thứ 2, ô màu nâu thể hiện giá trị tốt thứ 3 và ô màu hồng thể hiện giá trị tốt thứ 4. Ở đây thực hiện so sánh kết quả giữa LP-NestFuse với các phương pháp BTSFusion, CrossFuse, MPCFusion, NestFuse và PSLPT dựa trên 10 chỉ số đánh giá MI, NCIE, QG, SSIM, PSNR, EN, AG, SD, ALI và VIF.

Từ kết quả so sánh cho thấy, các chỉ số đánh giá chất lượng ảnh như PSNR, EN, SD, ALI tốt nhất so với các phương pháp khác, trong đó kết quả EN và ALI có sự vượt trội hẳn thể hiện LP-NestFuse đưa ra hình ảnh tổng hợp chứa nhiều thông tin quan trọng và cường độ sáng, độ tương phản hình ảnh rất tốt. Chỉ số QG cao, tốt hơn gần 11% so với phương pháp tốt thứ 2 là MPCFusion khẳng định khả năng bảo toàn các thông tin về cạnh của phương pháp đề xuất. Điều này rất quan trọng để duy trì tính toàn vẹn thị giác của hình ảnh, đặc biệt trong các ứng dụng như phát hiện vật thể và điều hướng, nơi chi tiết cạnh đóng vai trò quan trọng. Về chỉ số VIF (Visual Information Fidelity), đo lường độ trung thực của hình ảnh hợp nhất so với các hình ảnh gốc từ góc nhìn của hệ thống thị giác con người, LP-NestFuse đạt kết quả ấn tượng trên tập dữ liệu TNO. Nó cho thấy độ trung thực cao nhất, cho thấy chất lượng hình ảnh do LP-NestFuse tạo ra vượt trội so với các mô hình khác.

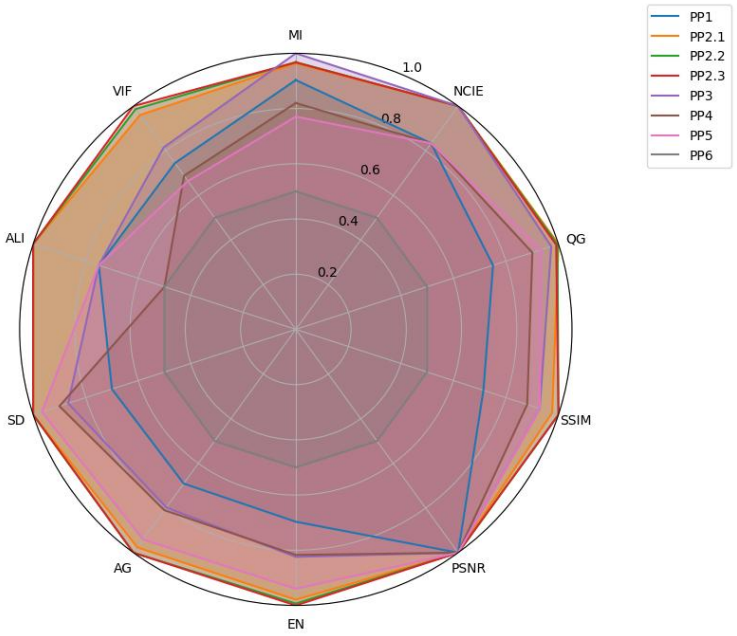
LP-NestFuse cũng thể hiện hiệu suất cạnh tranh trên một số chỉ số như MI, NCIE, SSIM và AG. Phương pháp này bảo toàn tương đối tốt cấu trúc và lượng thông tin được truyền từ hai hình ảnh nguồn đến hình ảnh hợp nhất. Sự tương quan phi tuyến giữa hình ảnh hợp nhất và hai hình ảnh nguồn thông qua chỉ số NCIE cạnh tranh tốt khi chỉ kém 0.1% so với phương pháp tốt nhất là CrossFuse. Tổng thể, LP-NestFuse chứng tỏ là một mô hình hợp nhất hình ảnh hiệu quả và đáng tin cậy, vượt trội ở các khía cạnh quan trọng như bảo toàn chi tiết cạnh, lượng thông tin được truyền, mối tương quan phi tuyến, chất lượng ảnh hợp nhất và chất lượng thị giác.

Bảng 2 kèm theo biểu đồ radar trực quan về khoảng tỷ lệ $[0.5,1]$ trong hình 9 thể hiện sự so sánh giữa các phương pháp tổng hợp thành phần cơ sở. Các phương pháp được ghi chú như sau: dựa trên độ lệch chuẩn kết hợp entropy (PP1), dựa trên năng lượng vùng cục bộ thay đổi bộ trọng số (PP2 với các bộ trọng số khác nhau là PP2.1, PP2.2, PP2.3), dựa trên bản đồ trọng số tổng hợp (PP3), dựa trên năng lượng Laplacian cục bộ (PP4), dựa trên bộ lọc hướng dẫn kết hợp Laplacian sửa đổi (PP5) và dựa trên mô hình VGG19 đã pretrain (PP6). Trong đó, PP2 sử dụng các bộ trọng số cho kích thước cửa sổ vùng cục bộ 3×3 như sau: PP2.1 sử dụng bộ trọng số $\omega = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, PP2.2. sử dụng bộ trọng số theo công thức $\omega = \frac{1}{1+\sqrt{i^2+j^2}}$ với i, j là vị trí so với pixel

tâm (0,0), PP2.3 là phương pháp đề xuất (ô màu cam trong bảng) sử dụng bộ trọng số $\omega = \frac{1}{16}\begin{bmatrix}1 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 1\end{bmatrix}$.

	MI	NCIE	QG	SSIM	PSNR	EN	AG	SD	ALI	VIF
PP1	2.8428	0.8067	0.4929	0.4004	7.1866	7.2405	0.0144	0.1820	0.5182	0.9119
PP2.1	2.8507	0.8068	0.5445	0.4015	7.1866	7.2527	0.0168	0.1829	0.5183	0.9199
PP2.2	2.8513	0.8068	0.5437	0.4016	7.1866	7.2533	0.0170	0.1829	0.5183	0.9209
PP2.3	2.8512	0.8068	0.5427	0.4016	7.1866	7.2536	0.0170	0.1829	0.5183	0.9214
PP3	2.8555	0.8068	0.5388	0.4013	7.1866	7.2460	0.0153	0.1825	0.5182	0.9145
PP4	2.8319	0.8067	0.5239	0.4011	7.1866	7.2457	0.0154	0.1826	0.5181	0.9098
PP5	2.8254	0.8067	0.5306	0.4013	7.1866	7.2510	0.0165	0.1828	0.5182	0.9089
PP6	2.7899	0.8065	0.4409	0.3995	7.1865	7.2320	0.0128	0.1814	0.5181	0.9028

Bảng 2. So sánh giữa phương pháp tổng hợp thành phần cơ sở



Hình 9. Biểu đồ radar so sánh giữa các phương pháp tổng hợp thành phần cơ sở

Kết quả trên cho thấy, phương pháp tổng hợp thành phần cơ sở dựa trên năng lượng vùng cực đại mang lại hiệu quả tốt nhất trên tất cả các chỉ số được đánh giá. Với phương pháp này, ba hướng sử dụng ba bộ trọng số khác nhau là PP2.1, PP2.2, PP2.3 được trình bày ở trên thì phương pháp nghiên cứu sử dụng là PP2.3 cho kết quả tốt hơn về việc bảo toàn thông tin cấu trúc, lượng thông tin trong hình ảnh tổng hợp và chất lượng thị giác của hình ảnh. Trên các phương diện về lượng thông tin truyền từ hình ảnh nguồn đến hình ảnh tổng hợp, mối tương quan tuyến tính, độ tượng phản, cường độ sáng và bảo toàn cạnh, PP2.3 cạnh tranh tốt với hai phương pháp còn lại. Tổng thể, PP2.3 đạt được sự cân bằng tốt nhất trên các chỉ số và có nhiều nhất các chỉ số có kết quả cao nhất so với các phương pháp còn lại.

4.5. So sánh trực quan

Dưới đây là các hình ảnh kết quả tổng hợp cho từng phương pháp hợp nhất, cung cấp sự so sánh trực quan giữa các phương pháp.

a) IR

b) VI

Hình 10. Một cặp hình ảnh trong bộ dữ liệu TNO

a) BTSFusion

b) CrossFuse

c) MPCFusion

d) NestFuse

e) PSLPT *f) Proposed model*

Hình 11. Kết quả hình ảnh tổng hợp của các mô hình

Trong các phương pháp kết hợp ảnh khác nhau, mô hình được đề xuất LP-NestFuse, đã thể hiện sự vượt trội trong việc tích hợp các đặc điểm nhiệt từ ảnh hồng ngoại (IR) với thông tin kết cấu chi tiết từ ảnh nhìn thấy được (VIS). Khả năng này được thể hiện rõ ràng trong Hình 11f. Khác với CrossFuse vốn thường bảo toàn tốt các chi tiết nhìn thấy (xem hình 11b) hay PSLPT thường bảo toàn tốt hơn các chi tiết nhiệt từ hình ảnh hồng ngoại (xem hình 11e), LP-NestFuse duy trì sự cân bằng mà không gây ra những mất mát đáng kể. BTSFusion trong hình 11a hay MPCFusion trong hình 11c có kết quả bảo toàn kết cấu tương đối tốt nhưng gặp phải hiện tượng nhiễu và độ tương phản thấp. NestFuse trong hình 11d cũng tương tự vậy kèm theo là cường độ sáng của hình ảnh thấp và nhiều chi tiết kết cấu về cạnh bị bỏ qua.

Phương pháp LP-NestFuse trong hình 11f khắc phục những nhược điểm này, cho hình ảnh kết quả có độ tương phản và cường độ sáng cao, hạn chế hiện tượng nhiễu và tăng tính trực quan thị giác cho hình ảnh. Hơn nữa, LP-NestFuse nổi bật trong việc bảo toàn các cạnh, đóng góp đáng kể vào sự toàn vẹn cấu trúc tổng thể của các ảnh kết hợp với các đường viền sắc nét và rõ ràng, một khía cạnh quan trọng đối với các ứng dụng yêu cầu phân định đối tượng chính xác. Nhìn chung, LP-NestFuse đạt được sự cân bằng cần thiết giữa các yếu tố thị giác và các tham số chất lượng ảnh, đóng vai trò quan trọng trong các ứng dụng như giám sát, xác định vật thể.

CHƯƠNG 5. KẾT LUẬN

5.1. Tóm tắt và kết luận

Nghiên cứu đã đề xuất một mô hình mới LP-NestFuse tổng hợp hình ảnh hồng ngoại (IR) và hình ảnh khả kiến (VI) dựa trên phương pháp lai kết hợp giữa phương pháp truyền thống và phương pháp tổng hợp dựa trên học sâu. Trong đó, nghiên cứu đã đề xuất một phương pháp phân rã kim tự tháp Laplacian mới như đã trình bày trong chương 3 giúp giảm sự phức tạp tính toán, bảo toàn các thông tin chi tiết và tăng cường tính toàn vẹn của thành phần cơ sở. Thành phần cơ sở được tổng hợp bằng phương pháp năng lượng vùng cục bộ cực đại có trọng số đồng thời kết hợp mặt nạ thu được qua mô hình U2Net được huấn luyện trước nhằm tăng cường thông tin vùng quan trọng trong hình ảnh hồng ngoại cho thành phần cơ sở. Các thành phần chi tiết được tổng hợp dựa trên mô hình học sâu NestFuse có sự thay đổi trong chiến lược tổng hợp. Tổng thể, mô hình đề xuất đưa ra một khung tổng hợp hình ảnh hiệu quả.

LP-NestFuse thể hiện hiệu suất mạnh mẽ trên nhiều tiêu chí khác nhau. Kết quả vượt trội của nó trong các tiêu chí QG, PSNR, EN, SD, ALI và VIF cho thấy mô hình này đặc biệt hiệu quả trong việc bảo toàn thông tin cạnh, hình ảnh tổng hợp chứa nhiều thông tin hơn, độ tương phản và cường độ sáng cao, đồng thời đảm bảo tính trung thực thị giác tốt. Những đặc điểm này khiến LP-NestFuse trở thành lựa chọn tốt cho các ứng dụng yêu cầu độ trung thực cao và bảo toàn chi tiết, chẳng hạn như trong viễn thám và giám sát.

Tuy nhiên, ở một số tiêu chí như MI, NCIE và SSIM, vẫn còn không gian để cải thiện so với các mô hình hiệu suất hàng đầu. Các nghiên cứu trong tương lai có thể tập trung vào việc nâng cao những khía cạnh này, có thể thông qua các kiến trúc mạng sâu hơn hoặc các chiến lược kết hợp tinh vi hơn.

Nhìn chung, mô hình LP-NestFuse nổi bật như một thuật toán hiệu suất cao trong lĩnh vực kết hợp ảnh. Nó liên tục xếp hạng trong số các mô hình hàng đầu trên các tiêu chí đánh giá toàn diện, chứng minh hiệu quả của mình trong việc tạo ra các ảnh kết hợp chất lượng cao. Phân tích xác nhận sự phù hợp của nó cho các ứng dụng quan trọng, nơi việc bảo toàn thông tin và chất lượng hình ảnh là điều cần thiết. Đồng thời, nghiên cứu này cũng cho thấy sự hiệu quả vượt trội của các phương pháp lai kết hợp sức mạnh từ cả các phương pháp truyền thống và các phương pháp dựa trên học sâu. Đây là một hướng phát triển có thể tiếp cận tốt trong tương lai để mang lại kết quả tổng hợp hình ảnh tốt hơn.

5.2. Hướng phát triển trong tương lai

Để tiếp tục phát triển từ nghiên cứu hiện tại, một số hướng đi trong tương lai được đề xuất gồm:

- Mở rộng sang thang màu: Hiện tại, mô hình hoạt động trên các hình ảnh thang độ xám. Công việc tương lai sẽ mở rộng khả năng của nó để xử lý hình ảnh màu, từ đó mở rộng phạm vi ứng dụng và cải thiện khả năng biểu diễn thông tin thị giác.
- Hiệu quả tính toán: Cần có thêm nghiên cứu để tối ưu hóa các thuật toán nhằm giảm tiêu thụ tài nguyên mà không làm giảm chất lượng kết hợp, giúp công nghệ trở nên khả dụng trên các thiết bị tính toán biên trong các ứng dụng IoT.
- Cải tiến mô hình học sâu: Việc tích hợp các chiến lược học sâu tiên tiến hơn, chẳng hạn như học chuyển giao (transfer learning) hoặc các mô hình dựa trên Transformer, có thể cải thiện khả năng thích ứng của mô hình và tăng chất lượng kết cấu của hình ảnh tổng hợp.
- Tổng hợp dựa trên tối ưu hóa: Việc thay đổi phương pháp tổng hợp chẳng hạn như trong tổng hợp thành phần cơ sở dựa trên các thuật toán tối ưu hóa có thể mang lại sự đảm bảo về chất lượng hình ảnh đầu ra
- Phương pháp phân rã bảo toàn kết cấu: Hướng phát triển này có thể nâng cao hơn khả năng truyền thông tin và bảo toàn cấu trúc cho hình ảnh tổng hợp so với các hình ảnh nguồn. Điển hình có thể là các phương pháp dựa trên bộ lọc dẫn hướng bảo toàn cạnh…

Những hướng phát triển này nhằm tinh chỉnh thêm mô hình kết hợp lai và mở rộng khả năng ứng dụng của nó vào các lĩnh vực đa dạng hơn, thúc đẩy sự phát triển trong lĩnh vực xử lý hình ảnh và công nghệ kết hợp hình ảnh.

REFERENCE