**Bài báo: “LP – NestFuse: Phương pháp tổng hợp hình ảnh hồng ngoại và hình ảnh khả kiến dựa trên lai kết hợp giữa phân rã kim tự tháp biến đổi và mô hình học sâu”**

Trịnh Văn Hậu 1, Đinh Phú Hùng 2, Phạm Văn Hải 1

1 Khoa Khoa học máy tính, Trường CNTT và TT, Đại học Bách khoa Hà Nội, 1 Đại Cồ Việt, Hai Bà Trưng, Hà Nội, Việt Nam

2 Khoa Công nghệ thông tin, Trường Đại học Thủy Lợi, 175 Tây Sơn, Đống Đa, Hà Nội, Việt Nam

**Abstract**

Trong lĩnh vực xử lý ảnh, việc kết hợp hình ảnh hồng ngoại (IR) và hình ảnh khả kiến (VIS) 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. Tuy nhiên, theo hiểu biết của chúng tôi, việc tích hợp hai loại hình ảnh này đặt ra những thách thức đáng kể và một số nhược điểm cần được khắc phục. Thứ nhất là một số phương pháp tổng hợp cho ra kết quả hình ảnh có độ tương phản và cường độ sáng trung bình thấp. Lý do cho vấn đề này là các phương pháp sử dụng các phân rã hình ảnh chưa hợp lý và tổng hợp thành phần tần số thấp bằng cách lấy trung bình. Thứ hai là thông tin cạnh chưa được bảo toàn đầy đủ từ hình ảnh nguồn đến hình ảnh tổng hợp. Nguyên nhân cho vấn đề này xuất phát từ quá trình tổng hợp các thành phần tần số cao chưa hiệu quả. Trong bài báo này, chúng tôi đề xuất phương pháp tổng hợp hình ảnh mới nhằm khắc phục hai nhược điểm trên dựa trên việc cải tiến mô hình học sâu NestFuse cho tổng hợp các thành phần tần số cao kết hợp với hàm năng lượng vùng cực đại cho tổng hợp thành phần tần số thấp. Trong đó, các thành phần được tạo ra thông qua phương pháp phân rã hình ảnh Laplacian Pyramid biến thể. Kết quả thực nghiệm cho thấy, phương pháp đề xuất không chỉ hiệu quả trong việc nâng cao chất lượng hình ảnh mà còn bảo toàn thông tin cạnh được truyền từ các hình ảnh đầu vào.

*Key words: Image Fusion, Laplacian Pyramid (LP), Maximum Region Energy (MRE), NestFuse*

**1. Introduction**

Tổng hợp hình ảnh là quá trình tổng hợp các thông tin hữu ích từ các hình ảnh riêng lẻ để tạo ra một hình ảnh duy nhất. Điều này cho phép hình ảnh tổng hợp mang nhiều thông tin hơn, nâng cao chất lượng hình ảnh và phục vụ tốt hơn cho các bài toán liên quan. Việc tổng 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. 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.

Hiện nay, có hai hướng chính trong giải quyết bài toán tổng hợp hình ảnh gồm các phương pháp truyền thống và các phương pháp dựa trên học sâu. Các phương pháp truyền thống được thực hiện dựa trên quy trình ba bước gồm phân rã hình ảnh – tổng hợp các thành phần – biến đổi ngược các thành phần tổng hợp về miền không gian. Trong đó, các phương pháp phân rã hình ảnh chủ yếu dựa trên biến đổi đa tỷ lệ như 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, biến đổi kim tự tháp (Laplacian Pyramid) **[ ]** tăng cường chi tiết ở nhiều cấp độ. Một cách tiếp cận quan trọng khác là phân rã dựa trên các bộ lọc hướng dẫn như Rolling Guidance Filtering **[ ]**, Bilateral Filter **[ ]**, kết hợp giữa Gaussian Filter và Guided Filter **[ ]** … Ưu điểm của cách tiếp cận khả năng phân tách thành công các cấu trúc không gian quy mô lớn, các cạnh quy mô trung bình và các chi tiết kết cấu quy mô nhỏ của hình ảnh. Ngoài ra, chúng chắc chắn giữ lại các chi tiết của hình ảnh gốc giúp giảm hiện tượng răng cưa và hiện tượng quầng sáng trong kết quả hợp nhất.

Các phương pháp truyền thống sử dụng nhiều kỹ thuật khác nhau để tổng hợp các thành phần sau phân rã như phương pháp trung bình **[ ]**, cực đại **[ ]**, dựa trên thông tin vùng **[ ]**, dựa trên năng lượng vùng **[ ]**, dựa trên bản đồ độ nổi bật **[ ]**. 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 hiệu quả hơn, giúp cải thiện hiệu suất và tính ứng dụng trong nhiều tình huống thực tế khác nhau. Tuy nhiên, các phương pháp này gặp phải một số hạn chế, 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. **[Survey 82 trang]**. 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.

Các phương pháp kết hợp dựa trên học sâu đã mang lại những tiến bộ đáng kể trong lĩnh vực kết hợp hình ảnh hồng ngoại (IR) và khả kiến (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. Trong các phương pháp kết hợp dựa trên học sâu có hai nhánh chính là non-end-to-end – tức là các mô hình học sâu chỉ tham gia một phần vào quá trình tổng hợp như tổng hợp thành phần cơ sở [ ] hay tổng hợp thành phần chi tiết **[ ]**, và end-to-end – tức là toàn bộ quá trình tổng hợp là một mô hình học sâu, nhận đầu vào là các hình ảnh và đầu ra là hình ảnh tổng hợp **[ ]**. Các phương pháp non-end-to-end kết hợp giữa phương pháp truyền thống với các mô hình học sâu, cũng gồm ba giai đoạn chính: phân rã hình ảnh – tổng hợp các thành phần kết hợp mô hình học sâu – biến đổi ngược các thành phần tổng hợp về miền không gian. Ví dụ như trong **[ ]**, Wu Wang và cộng sự sử dụng phương pháp kim tự tháp dựa trên Transformer (PSLPT) để phân rã hình ảnh nguồn thành các đặc trưng đa tần số ở nhiều cấp độ, sau đó các cặp thành phần được tổng hợp thông qua bản đồ trọng số tính toán từ module Swin-CNN. Đầu tiên, các phần dư (residuals) từ các đặc trưng của hai hình ảnh nguồn được tính toán. Sau đó, phần dư này được đưa vào module Swin-CNN để trích xuất các đặc trưng ngữ nghĩa. Tiếp theo, một lớp tích chập (convolution) đơn được sử dụng để ánh xạ các đặc trưng ngữ nghĩa đa kênh thành đặc trưng hai kênh. Cuối cùng, phép toán SoftMax được áp dụng để chuẩn hóa các đặc trưng, tạo ra hai mặt nạ (mask). Các mặt nạ này sau đó được nhân từng phần tử với các đặc trưng của hình ảnh nguồn để tạo ra các đặc trưng hợp nhất. Các đặc trưng này thực hiện biến đổi ngược để thu được hình ảnh tổng hợp cuối cùng.

Các phương pháp end-to-end tận dụng các kiến trúc mạng neural tiên tiến như CNN **[ ]**, GAN **[ ]**, Autoencoder **[ ]** hay Transformer **[ ]**, 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. 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. Đầu tiên, các cặp hình ảnh hồng ngoại và khả kiến được đưa vào mô hình, sau đó các đặc trưng được trích xuất và hợp nhất thông qua một mạng hợp nhất hình ảnh cụ thể, và cuối cùng xuất ra hình ảnh tổng hợp. Một hàm mất mát (loss function) được thiết kế tốt đóng vai trò quyết định thông tin nào sẽ được giữ lại trong các hình ảnh hợp nhất cuối cùng **[Overview.7]**. 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. Các đặc trưng thu được từ mạng trích xuất đặc trưng được đưa vào khối ResBlock đầu tiên của mạng tái tạo đặc trưng. 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) **[Paper gốc của Goodfellow]** cũng được sử dụng để cải thiện tính thực tế của hình ảnh kết hợp. Mô hình GAN gồm hai thành phần chính là bộ tạo (Generator – G) và bộ phân biệt (Discriminator – D). Bộ tạo G tạo ra hình ảnh hợp nhất nhằm đánh lừa bộ phân biệt D, nhiệm vụ của D là xác định xem hình ảnh đó là thật hay được tạo ra bởi G. Quá trình huấn luyện có thể được xem như một trò chơi đối kháng giữa G và D. Khi D không còn khả năng phân biệt được hình ảnh từ G hay hình ảnh thực, điều đó cho thấy G đã học chính xác các đặc trưng của hình ảnh nguồn. Trong giai đoạn kiểm tra, các tham số của G được cố định và G được sử dụng để tạo ra hình ảnh hợp nhất **[Overview.7 Ref [40]]**. Ví dụ như 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 kế thừa quy trình hợp nhất hình ảnh truyền thống: trích xuất đặc trưng, hợp nhất đặc trưng và tái tạo đặc trưng. Bộ mã hóa (Encoder) dùng để trích xuất đặc trưng từ hình ảnh và bộ giải mã (Decoder) dùng để tái tạo lại hình ảnh từ các đặc trưng đã trích xuất **[Overview.7]**. Mục tiêu của quá trình huấn luyện là làm cho hình ảnh đầu ra giống nhất có thể với hình ảnh đầu vào. Khi đó, bộ mã hóa đã trích xuất chính xác và đầy đủ các đặc trưng của hình ảnh. Trong giai đoạn kiểm tra, sau khi trích xuất đặc trưng bằng bộ mã hóa, các đặc trưng này được hợp nhất dựa trên một số quy tắc hợp nhất nhất định, sau đó đi qua bộ giải mã để thu được hình ảnh hợp nhất. Ví dụ trong NestFuse **[ ]**, cấu trúc mạng lồng nhau được tích hợp trong khung làm việc của Autoencoder. Trong đó, Encoder và Decoder đều là các CNN-Blocks gồm 2 layer convolutional trong mỗi khối. Quy tắc hợp nhất sử dụng kết hợp giữa cơ chế chú ý kênh và cơ chế chú ý không gian. 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 **[Paper gốc của Goodfellow]** là một hướng tiếp cận mới, nổi bật gần đây, khắc phục nhược điểm của CNN trong việc nắm bắt thông tin toàn cục để tạo ra hình ảnh hợp nhất chất lượng cao dựa vào cơ chế tự chú ý **[Overview.7]**. Một ví dụ điển hình áp dụng Transgormer trong bài toán IVIF là SBIT-Fuse **[ ]**. 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. CNN đòi hỏi việc thiết kế các hàm mất mát phức tạp, giới hạn khả năng cải thiện hiệu suất của mô hình. GAN cho ra kết quả hình ảnh hợp nhất bị nghiêng về phía kênh nhìn thấy, trong khi thông tin từ kênh hồng ngoại chỉ được giữ lại thông qua hàm mất mát nội dung (content loss), dẫn đến sự mất mát lượng lớn thông tin quan trọng từ hình ảnh hồng ngoại. AE gặp khó khăn trong việc cân bằng giữa chất lượng tái tạo hình ảnh và chi phí tính toán. Transformer thường yêu cầu lượng tài nguyên tính toán lớn và do cấu trúc phức tạp, việc tối ưu hóa và điều chỉnh tham số của chúng là một thách thức đáng kể **[Overview.7]**. Bên cạnh đó, các mô hình học sâu thiếu cơ sở lý thuyết mạnh mẽ để khẳng định tính chắc chắn của kết quả đầu ra, nhiều thông tin chi tiết về kết cấu như thông tin cạnh không được đảm bảo.

Ngoài ra, theo hiểu biết hiện tại của chúng tôi, các cách tiếp cận dựa trên phương pháp truyền thống và phương pháp học sâu được trình bày ở trên còn gặp phải một số vấn đề. Thứ nhất là một số phương pháp tổng hợp cho ra kết quả hình ảnh có độ tương phản và cường độ sáng trung bình thấp. Lý do cho vấn đề này là các phương pháp sử dụng các phân rã hình ảnh chưa hợp lý và tổng hợp thành phần tần số thấp bằng cách lấy trung bình. Wencheng Wang **[ ]** tiếp cận giải quyết hạn chế này bằng cách áp dụng phương pháp phân rã kim tự tháp Laplacian Pyramid, tổng hợp thành phần cơ sở dựa trên thông tin vùng và tổng hợp thành phần chi tiết bằng hàm năng lượng vùng cực đại giúp giữ lại các đặc điểm nổi bật từ mỗi hình ảnh. Tuy nhiên, phương pháp phân rã LP này gặp phải vấn đề mất mát thông tin do thực hiện nhiều lần các phép Down và Expand.

Thứ hai là thông tin cạnh chưa được bảo toàn đầy đủ từ hình ảnh nguồn đến hình ảnh tổng hợp. Nguyên nhân cho vấn đề này xuất phát từ quá trình tổng hợp các thành phần tần số cao chưa hiệu quả. NestFuse **[ ]** là một kiến trúc học sâu thuộc lớp mạng Autoencoders, nổi bật trong việc tăng cường chi tiết và chất lượng của hình ảnh hợp nhất, đảm bảo duy trì tính toàn vẹn và tính hữu ích của thông tin từ cả hai nguồn đầu vào. Điều này mang lại ý tưởng cho chúng tôi để áp dụng mô hình NestFuse tổng hợp thành phần chi tiết. Tuy nhiên, chiến lược tổng hợp của NestFuse sử dụng cơ chế chú ý kênh và chú ý không gian tách biệt nhau, tổng hợp bằng phép trung bình làm giảm chất lượng hình ảnh tổng hợp.

Để khắc phục những nhược điểm trên, trong bài báo này, chúng tôi đề xuất một phương pháp tổng hợp lai giữa phương pháp truyền thống và mô hình học sâu với những đóng góp chính như sau:

* Phương pháp lai sáng tạo kết hợp những ưu điểm của phân rã kim tự tháp Laplacian, khả năng giữ lại những đặc điểm nổi bật của hàm năng lượng cục bộ và sự linh hoạt của mô hình học sâu NestFuse nhằm nâng cao chất lượng và hiệu quả của quá trình hợp nhất.
* 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 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.

Phần còn lại của bài báo này được tổ chức như sau: Một số kiến thức nền tảng gồm phương pháp phân rã kim tự tháp Laplacian (Laplacian Pyramid), hàm năng lượng cục bộ theo vùng, mô hình NestFuse, được giới thiệu ngắn gọn trong Phần 2. Phần 3 trình bày những cải tiến được đề xuất và chi tiết quy trình tổng hợp gồm: phân rã hình ảnh dựa trên Laplacian Pyramid biến đổi, hàm năng lượng cục bộ cực đại với bộ trọng số Binomial cho tổng hợp thành phần cơ sở, cải tiến mô hình NestFuse và áp dụng cho tổng hợp thành phần chi tiết. Phần 4 thực nghiệm và đánh giá. Cuối cùng, kết luận và hướng phát triển trong tương lai được đưa ra trong Phần 5.

**2. Background**

Một số kiến thức nền tảng gồm Laplacian Pyramid, Hàm năng lượng theo vùng cực đại, Nest-Fuse sẽ được giới thiệu trong phần này.

**2.1. Laplacian Pyramid**

Laplacian Pyramid **[ ]** (Kim tự tháp Laplacian) là một kỹ thuật đa độ phân giải cơ bản, được sử dụng phổ biến trong phân rã hình ảnh **[danh sách một số bài báo sử dụng LP]** và đặc biệt được đánh giá cao nhờ tính hiệu quả trong việc kết hợp hình ảnh từ các mức tiêu điểm khác nhau để tăng cường chi tiết ở nhiều cấp độ. Chi tiết thuật toán phân rã hình ảnh với Laplacian Pyramid được trình bày trong Algorithm 1. Đầu tiên, xuất phát từ hình ảnh gốc, xây dựng Gaussian Pyramid (Kim tự tháp Gaussian) bằng cách áp dụng bộ lọc Gaussian kết hợp với phép Down Sampling. Tiếp theo, xây dựng 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.

|  |
| --- |
| **Algorithm 1: Phân rã hình ảnh với Laplacian Pyramid** |
| Input: Image  Output: Laplacian Pyramid with and  Constant: number of levels |
| Step 1: Xuất phát từ ảnh gốc , xây dựng Gaussian Pyramid    For to do:  với M là hạt nhân Gauss:  Down đại diện cho quá trình giảm độ phân giải theo tỷ lệ 2.  Step 2: Từ Gaussian Pyramid, xây dựng Laplacian Pyramid    For downto 0 do:  trong đó với hạt nhân M có kích thước tuân theo công thức: |

Hình **[ ]** minh họa trực quan cho kết quả của phân rã hình ảnh với Laplacian Pyramid. Theo sơ đồ minh họa, hình ảnh sau phân rã thu được một thành phần cơ cở là level cao nhất của kim tự tháp chứa các đặc điểm tổng quát và xu hướng toàn cục của hình ảnh, và các thành phần chi tiết là các level còn lại chứa các thông tin về kết cấu và chi tiết của hình ảnh. Phương pháp phân rã này tách biệt thành phần tần số thấp và các thành phần tần số cao, các thành phần ở nhiều mức độ phân giải, tạo điều kiện tăng cường chi tiết đa tỷ lệ.

A black and white image of a building

Description automatically generated

*Hình* ***[ ]****. Phân rã hình ảnh với Laplacian Pyramid truyền thống*

**2.2. Hàm năng lượng theo vùng cực đại (Maximum Region Energy)**

Hàm năng lượng theo vùng cực đại thường được sử dụng để tổng hợp các thành phần sau khi phân rã hình ảnh trong bài toán tổng hợp hình ảnh **[ ].** Ưu điểm của phương pháp này là giữ lại được các đặc trưng và chi tiết nổi bật trong các hình ảnh nguồn cho hình ảnh tổng hợp. Năng lượng theo vùng tại mỗi pixel thường dựa trên kích thước vùng hoặc với các bộ trọng số thể hiện mức độ quan trọng của các pixel lân cận với pixel trung tâm như bộ trọng số một [ ], bộ trọng số Gauss [ ] hay bộ trọng số binomial [ ]. Lấy ví dụ với kích thước vùng và bộ trọng số một , công thức tính năng lượng vùng tại pixel như sau:

Khi đó, hàm năng lượng theo vùng cực đại tổng hợp hai thành phần để thu được như sau:

**2.3. 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. 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. 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 **[ ]**. Trong đó, mỗi block Encoder và Decoder đều chứa các layer Convolutional (mỗi khối chứa 2 layer), riêng các khối Encoder có thêm một layer MaxPooling để thu 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 **[ ]** để học được một mô hình trích xuất các đặc trưng và tái tạo hình ảnh từ các đặc trưng đó. Cấu trúc chi tiết của các layer và các block Encoder và Decoder được thể hiện trong bảng **[ ]**. Conv là layer Convolutional, ECB ký hiệu cho khối Encoder Convolutional gồm layer Convolutional + Max pooling, DCB ký hiệu cho khối Decoder Convolutional chỉ gồm các layer Convolutional mà không có Max pooling.

A diagram of a computer flowchart

Description automatically generated A diagram of a computer flowchart

Description automatically generated

|  |  |
| --- | --- |
| *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 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:

trong đó, đo lường lỗi tái tạo theo từng điểm ảnh với O là hình ảnh đầu ra và I là hình ảnh đầu vào. 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 **[ ]**, 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.

A table with numbers and letters

Description automatically generated

Bảng **[ ]**. Cấu trúc chi tiết của các layer và block Encoder, Decoder trong mạng NestFuse

**3. Proposed model**

Trong phần này, chi tiết các thành phần của phương pháp tổng hợp đề xuất được trình bày bao gồm Phương pháp phân rã hình ảnh với Laplacian Pyramid biến thể, Hàm năng lượng vùng cực đại với bộ trọng số binomial cho tổng hợp thành phần cơ sở, cải tiến NestFuse áp dụng cho tổng hợp thành phần chi tiết và toàn bộ quá trình tổng hợp.

**3.1. Phân rã Laplacian Pyramid (LP) biến thể**

Nhiều công bố trước đây đã sử dụng Laplacian Pyramid cho quá trình phân rã hình ảnh như trong **[ ]**. Các phương pháp kết hợp hình ảnh IR và VI dựa trên Pyramid có ưu điểm là thực hiện đơn giản và yêu cầu ít không gian bộ nhớ **[Survey 2017].** Tuy nhiên, nhược điểm của nó là làm mờ một số chi tiết của hình ảnh **[Ref 142 Survey 2017],** kết quả hợp nhất có độ tương phản thấp **[Survey 2023].** Ngoài ra, theo hiểu biết của chúng tôi, 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, Pyramid đơn thuần không thể đạt được chất lượng kết hợp hình ảnh IR và VI tốt. Xuất phát từ những hạn chế trên, nghiên cứu này đề xuất một phương pháp phân rã hình ảnh với Laplacian Pyramid biến thể, trong đó kim tự tháp Laplacian cuối cùng thu được gồm một thành phần tần số thấp và các thành phần tần số cao trong đó *,*  tức là giữ nguyên 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 . 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. Chi tiết thực hiện phân rã của Laplacian Pyramid biến thể được trình bày qua mã giả **Thuật toán 3**. Hình **[ ]** minh họa trực quan kết quả phân rã hình ảnh với phương pháp đề xuất này.

|  |
| --- |
| **Algorithm 3: Phân rã hình ảnh với Laplacian Pyramid biến thể** |
| Input: Image  Output: Laplacian Pyramid with and  Constant: number of levels |
| Step 1: Xuất phát từ ảnh gốc , xây dựng Gaussian Pyramid    For to do:  với M là hạt nhân Gauss:  Down đại diện cho quá trình giảm độ phân giải theo tỷ lệ 2.  Step 2: Từ Gaussian Pyramid, xây dựng Laplacian Pyramid biến thể  For downto 1 do:    trong đó với hạt nhân M có kích thước tuân theo công thức: |

A black and white image of a building

Description automatically generated

*Hình* ***[ ]****: Sơ đồ trực quan phương pháp phân rã với LP đề xuất*

Quá trình phân rã biến đổi cũng dẫn đến sự thay đổi trong quá trình biến đổi ngược về miền không gian từ các thành phần phân rã. Thay vì sử dụng quá trình biến đổi ngược thông qua vòng lặp expand và cộng dần như LP truyền thống, LP biến thể bổ sung thêm trọng số (thông qua sharpness scores) cho mỗi levels trong kim tự tháp. 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**. Thuật toán 4** mô tả chi tiết các bước trong quá trình biến đổi ngược này kèm theo sơ đồ minh họa hình **[ ]**.

|  |
| --- |
| **Algorithm 4: Biến đổi ngược về miền không gian trong Laplacian Pyramid biến thể** |
| Input: Laplacian Pyramid with and  Output: Image  Constant: number of levels |
| Step 1: Khởi tạo    Step 2: Tính sharpness scores của levels trong  trong đó đại diện cho toán tử Laplace  Step 3: Tính tổng có trọng số cho các thành phần chi tiết    Step 4: Kết hợp thành phần cơ sở để thu được hình ảnh tái tổ hợp cuối cùng |

**A diagram of a structure

Description automatically generated with medium confidence**

*Hình* ***[ ]****. Sơ đồ thực hiện biến đổi ngược về miền không gian với Laplacian Pyramid*

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

Phương pháp tổng hợp thành phần cơ sở dựa 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 **[ ]**. Nắm bắt những ưu điểm trên, chúng tôi đã lựa chọn hàm năng lượng theo vùng cực đại cho tổng hợp thành phần cơ sở. Các công bố trước đây sử dụng nhiều bộ trọng số khác nhau như bộ trọng số một [ ], bộ trọng số Gauss [ ], bộ trọng số binomial [ ] … Trong nghiên cứu này chúng tôi sử dụng bộ trọng số binomial, cơ sở định lượng cho điều này sẽ được trình bày rõ ở trong phần thực nghiệm. Chi tiết quy trình tổng hợp thành phần cơ sở dựa trên MRE được thể hiện trong mã giả **thuật toán 5.**

|  |
| --- |
| **Algorithm 5: Maximum Region Energy (MRE) tổng hợp thành phần cơ sở** |
| Input: Thành phần cơ sở của IR và VIS  Output: Thành phần cơ sở tổng hợp  Constant: Trọng số pixel trong vùng cục bộ |
| Step 1: Tính toán năng lượng vùng cục bộ cho hai thành phần cơ sở      Step 2: Tổng hợp thành phần cơ sở |

**3.3. NestFuse cải tiến tổng hợp 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. Tuy nhiên, chiến lược hợp nhất của NestFuse kết hợp giữa cơ chế chú ý kênh và cơ chế chú ý không gian nhưng triển khai độc lập với nhau và tổng hợp lại theo phương pháp trung bình. Chiến lược tổng hợp này làm giảm chất lượng hình ảnh tổng hợp và chưa bảo toàn được các thông tin về chi tiết cạnh.

A screenshot of a computer game

Description automatically generated

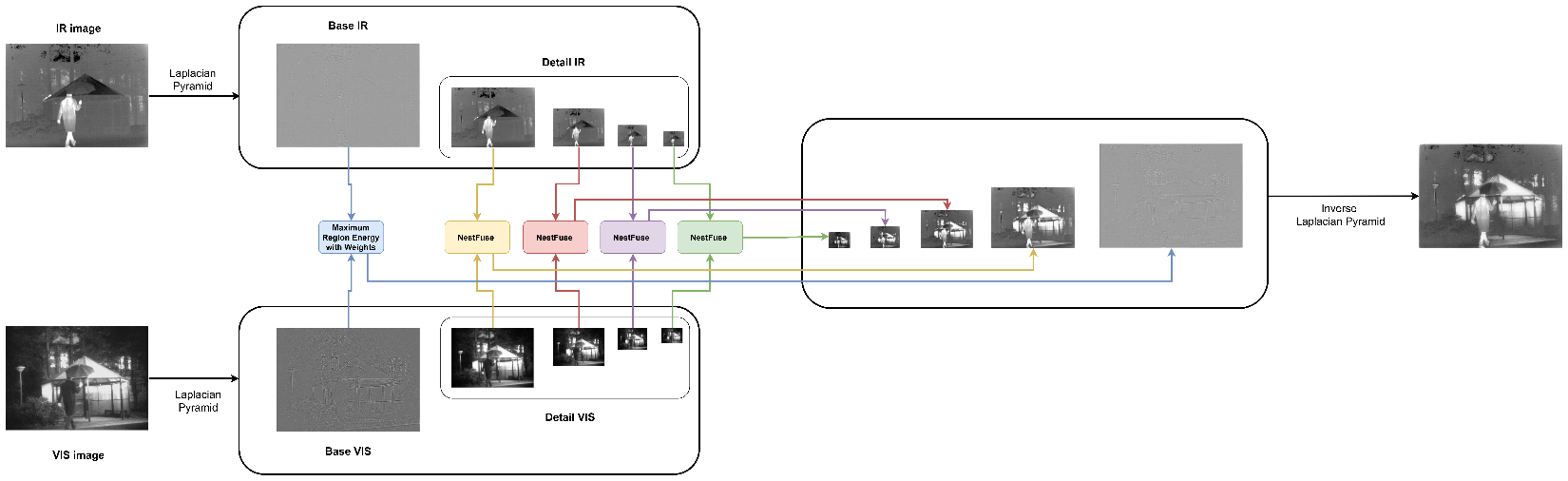
*Hình* ***[ ]****. Chiến lược tổng hợp đặc trưng đề xuất (FS)*

Trong nghiên cứu này, chúng tôi đề xuất một chiến lược tổng hợp mới CMDAF\_SA kết hợp giữa khối CMDAF lấy ý tưởng từ phương pháp PIAFusion **[ ]** và cơ chế Spatial Attention. Hình **[ ]** 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 thực hiện chia sẻ thông qua hàm để tính trọng số cuối cùng cho mỗi thành phần và tổng hợp các thành phần. Chi tiết quy trình tổng hợp thành phần chi tiết dựa trên NestFuse được thể hiện trong mã giả thuật toán 6.

|  |
| --- |
| **Algorithm 6: NestFuse cải tiến tổng hợp thành phần chi tiết** |
| Input: Thành phần chi tiết thứ của IR và VIS  Output: Thành phần chi tiết thứ tổng hợp |
| Step 1: Sử dụng khối Encoder đã huấn luyện để trích xuất đặc trưng      Step 2: Tổng hợp các đặc trưng với chiến lược tổng hợp CMDAF\_SA  Step 2.1. Tính chi tiết tăng cường thông qua Channel Attetion trên tính năng hiệu        Step 2.2. Tăng cường chi tiết cho hai đặc trưng      Step 2.3. Áp dụng cơ chế Spatial Attention  ;    Step 2.3. Tổng hợp tính năng    Step 3: Sử dụng khối Decoder đã huấn luyện để thu được thành phần chi tiết thứ tổng hợp |

**3.4. Tổng thể quy trình**

Phương pháp tổng hợp hình ảnh đề xuất gồm ba bước chính phân rã hình ảnh – tổng hợp các thành phần – biến đổi ngược về miền không gian. Đầu tiên, hình ảnh được phân rã với Laplacian Pyramid biến thể (Algorithm 3) thu được thành phần tần số thấp và các thành phần tần số cao. Tiếp theo, cặp thành phần cơ sở được tổng hợp như trình bày trong Algorithm 5, các cặp thành phần chi tiết được tổng hợp như trình bày trong Algorithm 6. Cuối cùng, biến đổi ngược về miền không gian từ thành phần cơ sở tổng hợp và các thành phần chi tiết tổng hợp dựa trên Algorithm 4. Sơ đồ chi tiết được thể hiện trong hình **[ ]** và mã giả theo thuật toán 7 (LP\_NestFuse).



*Hình* ***[ ]****. Kiến trúc tổng quan của phương pháp*

|  |
| --- |
| **Algorithm 7: LP\_NestFuse** |
| Input: Hình ảnh hồng ngoại IR và hình ảnh khả kiến VIS  Output: Hình ảnh tổng hợp F  Constant: number of levels ( |
| Step 1: Phân rã hình ảnh IR và VIS với Laplacian Pyramid biến thể      Step 2: Tổng hợp thành phần cơ sở    Step 3: Tổng hợp thành phần chi tiết  For to do:    Step 4: Biến đổi ngược về miền không gian |

**4. Experiment**

**4.1. Dữ liệu thực nghiệm**

Trong phần thực nghiệm của nghiên cứu này, chúng tôi sử dụng các cặp hình ảnh được chọn từ tập dữ liệu TNO **[ ]** để đánh giá phương pháp đề xuất và so sánh hiệu quả của phương pháp đề xuất với các phương pháp khác. Cụ thể, chúng 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, 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. Chỉ số đánh giá**

Thực nghiệm đánh giá kết quả trên 7 chỉ số chính được phân thành cách nhóm: chỉ số dựa trên thông tin gồm MI **[ ]**(Mutual Information), chỉ số dựa trên đặc trưng gồm **[ ]**, chỉ số dựa trên chất lượng hình ảnh gồm PSNR **[ ]**(Peak Signal-to-Noise Ratio), EN **[ ]**(Entropy), SD **[ ]**(Standard Deviation), ALI **[ ]**(Average Light Intensity) và chỉ số dựa trên thông tin thị giác gồm VIF **[ ]**(Visual Information Fidelity).

**4.2.1. Mutual Information (MI)**

Thông tin lẫn nhau **[ ]**(MI) đo lượng thông tin được truyền từ hình ảnh nguồn đến hình ảnh tổng hợp. Định nghĩa của thông tin lẫn nhau cho hai biến ngẫu nhiên rời rạc U và V theo công thức:

Trong đó, là hàm phân phối xác suất chung của U và V, và là hàm phân phối xác suất biên của U và V. Thông tin lẫn nhau còn có thể được biểu diễn thông qua entropy của hai phân phối U, V và phân phối chung như sau:

MI đo lượng thông tin được truyền từ hai hình ảnh đầu vào A và B tới hình ảnh tổng hợp F theo công thức

Giá trị MI càng cao thể hiện lượng thông tin được truyền tới hình ảnh tổng hợp càng nhiều và chất lượng tổng hợp hình ảnh càng tốt.

**4.2.2. Gradient-Based Fusion Performation ()**

Xydeas và Petrovic **[ ]** đề xuất một thước đo để đánh giá lượng thông tin biên được truyền từ hình ảnh đầu vào vào hình ảnh hợp nhất. Toán tử biên Sobel được áp dụng để lấy cường độ cạnh của ảnh đầu vào là và hướng :

Trong đó là kết quả tích chập theo chiều ngang và chiều dọc với toán tử Sobel. Cường độ tương đối và giá trị định hướng giữa hình ảnh A và ảnh tổng hợp F là:

Các giá trị bảo toàn cường độ và hướng của cạnh có thể được xác định như sau:

Các hằng số và xác định hàm sigmoid được sử dụng để hình thành cường độ cạnh và giá trị bảo toàn định hướng. Khi đó, giá trị bảo toàn thông tin cạnh được định nghĩa:

Đánh giá cuối cùng được lấy từ giá trị trung bình có trọng số của các giá trị bảo tồn thông tin cạnh:

Với trọng số và , L là một hằng số.

Giá trị càng lớn thì thông tin cạnh của ảnh nguồn được giữ lại trong ảnh hợp nhất càng nhiều và hiệu ứng hợp nhất càng tốt.

**4.2.3. Peak Signal-to-Noise Ratio (PSNR)**

PSNR **[ ]** biểu thị tỷ lệ công suất đỉnh và công suất nhiễu trong hình ảnh hợp nhất. Nó có thể đo mức độ biến dạng trong quá trình tổng hợp hình ảnh. PSNR được xác định theo công thức

Trong đó r là giá trị cực đại của ảnh hợp nhất và MSE là lỗi bình phương trung bình tổng hợp giữa hai ảnh nguồn và ảnh hợp nhất. PSNR càng lớn thì ảnh hợp nhất càng gần ảnh nguồn, độ méo càng nhỏ và hiệu ứng hợp nhất càng tốt.

**4.2.4. Entropy (EN)**

Entropy **[ ]** tính toán lượng thông tin có trong hình ảnh, được định nghĩa theo công thức sau:

Trong đó L biểu thị số mức độ xám và pl biểu thị histogram chuẩn hóa của các mức độ xám tương ứng trong ảnh hợp nhất. Giá trị EN dao động trong khoảng từ 0 đến 8, giá trị càng lớn thì lượng thông tin của ảnh càng nhiều, hiệu ứng hợp nhất càng tốt.

**4.2.5. Standard Deviation (SD)**

SD **[ ]** phản ánh sự phân bố và độ tương phản của hình ảnh tổng hợp. SD được định nghĩa theo công thức:

Giá trị SD càng lớn, độ tương phản của hình ảnh hợp nhất càng cao và hiệu ứng hình ảnh của hình ảnh hợp nhất càng tốt.

**4.2.6. Average Light Intensity (ALI)**

ALI **[ ]** đo cường độ sáng trung bình của hình ảnh tổng hợp. ALI càng cao thể hiện hình ảnh có mức sáng càng tốt và được định nghĩa theo công thức:

**4.2.7. Visual Information Fidelity (VIF)**

Độ trung thực thông tin hình ảnh **[ ]** (VIF)đánh giá mức độ mà hình ảnh được hợp nhất giữ lại thông tin hình ảnh từ hình ảnh gốc. VIF trong lĩnh vực hợp nhất hình ảnh được biểu thị dưới dạng tỷ số giữa lượng thông tin hình ảnh có trong hình ảnh hợp nhất và lượng thông tin trong hình ảnh tham chiếu (reference image). Công thức tính VIF như sau:

Trong đó:

* là hệ số khuếch đại (hoặc hệ số suy giảm) được áp dụng cho dải tần con trong khối đối với hình ảnh bị méo.
* biểu thị độ lệch chuẩn cục bộ của hình ảnh gốc/tham chiếu trong dải tần con và khối .
* là một hằng số liên quan đến mật độ phổ công suất của hình ảnh tham chiếu.
* là phương sai của nhiễu biến dạng trong dải tần con và khối .
* là phương sai của nhiễu hình ảnh được mô hình hóa dưới dạng nhiễu Gaussian cộng, được giả định là không đổi trên toàn hệ thống.

**4.3. Thiết lập thực nghiệm**

Một số thí nghiệm khác nhau đã được tiến hành để đánh giá hiệu quả của phương pháp đề xuất. **Thí nghiệm thứ nhất** đánh giá hiệu quả của phương pháp đề xuất (LP – NestFuse) so với các nghiên cứu được công bố gần đây. Các phương pháp sử dụng để so sánh bao gồm: BTSFusion **[ ]**(2024), MPCFusion **[ ]**(2024), CrossFuse **[ ](**2024), PSLPT **[ ]**(2024) và mô hình mà chúng tôi làm nền tảng là NestFuse **[ ]** (2020). Trong đó, BTSFusion sử dụng mạng trích xuất đặc trưng và tái cấu trúc đặc trưng đều dựa trên mạng CNN với cấu trúc RepVGG. MPCFusion trích xuất đặc trưng kết hợp giữa Convolution và Vision Transformer, module chú ý chéo song song và chú ý liên miền tổng hợp đặc trưng và cuối cùng sử dụng mạng kết nối lồng ghép để tái cấu trúc đặc trưng. CrossFuse sử dụng Encoder là các khối DenseBlock, cơ chế chú ý chéo để tổng hợp đặc trưng và Decoder là các khối Convolution. PSLPT phân rã hình ảnh với Laplacian Pyramid dựa trên Transformer và tổng hợp các thành phần bằng khối Frequency Adaptive Fusion với lõi là Transformer dựa trên các quy tắc tổng hợp.

**Thí nghiệm thứ hai** đánh giá hiệu quả của việc tổng hợp thành phần cơ sở dựa vào năng lượng vùng cực đại với bộ trọng số binomial bằng cách thay đổi thuật toán tổng hợp thành phần cơ sở và cố định các phần còn lại. Các kỹ thuật tổng hợp thành phần cơ sở khác được so sánh gồm: dựa trên độ lệch chuẩn kết hợp entropy (PP1) **[ ]**, dựa trên năng lượng vùng cực đại thay đổi bộ trọng số **[ ]**, 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, kỹ thuật dựa trên năng lược vùng cực đại thí nghiệm với ba bộ trọng số: bộ trọng số một (PP2.1), bộ trọng số Gauss (PP2.2) và bộ trọng số Binomial (LP-NestFuse).

**Thí nghiệm thứ ba** đánh giá hiệu quả của việc tổng hợp thành phần chi tiết dựa trên mô hình NestFuse cải tiến bằng cách thay đổi thuật toán tổng hợp thành phần chi tiết và cố định các phần còn lại. Các kỹ thuật tổng hợp thành phần chi tiết khác được so sánh bao gồm: dựa trên năng lượng vùng cục bộ (PP1) **[ ]**, dựa trên bản đồ trọng số tổng hợp qua entropy – contrast – visibility (PP2) **[ ]**, dựa trên mô hình VGG19 đã pretrain (PP3) **[ ]**, dựa trên năng lượng Laplacian cục bộ (PP4) **[ ]** và dựa trên trọng số thông qua gradient trung bình (PP5) **[ ]**.

**Thí nghiệm thứ tư** đánh giá hiệu quả của riêng việc thay đổi chiến lược tổng hợp trong mô hình NestFuse tức là thực hiện so sánh giữa việc áp dụng chiến lược tổng hợp cũ và chiến lược tổng hợp mới. Thí nghiệm thực hiện với phương pháp đề xuất LP-NestFuse, phương pháp đề xuất nhưng sử dụng chiến lược tổng hợp cũ (LP-NestFuse\_Old), phương pháp NestFuse **[ ]** và phương pháp NestFuse kết hợp chiến lược tổng hợp mới (NestFuse\_New).

Cuối cùng, trong **thí nghiệm thứ năm**, chúng tôi thực hiện đánh giá thời gian chạy của phương pháp đề xuất LP-NestFuse so với các phương pháp tổng hợp khác trong thí nghiệm thứ nhất và thời gian chạy của phương pháp phân rã Laplacian Pyramid biến thể so với các phương pháp phân rã khác gồm Laplacian Pyramid truyền thống **[ ]**, Constrast Pyramid **[ ]**, Wavelets **[ ]**, Dual – Tree Complex Wavelets **[ ]** và Multi-level Guided Filtering **[ ]**.

**4.4. Cấu hình 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.

**4.5. Kết quả thực nghiệm**

Kết quả của các thực nghiệm lần lượt được thể hiện trong bảng **[ ], [ ], [ ] và [ ]** kèm theo biểu đồ so sánh trong các hình **[ ], [ ], [ ] và [ ].** Trong các bảng **[ ], [ ] và [ ],** ô màu xanh lá thể hiện giá trị tốt nhất, ô màu xanh lam thể hiện giá trị tốt thứ hai, ô màu nâu là thứ ba và ô màu hồng là thứ tư. Trong bảng **[ ]** màu sắc thể hiện giá trị nhiệt theo cột chỉ số, màu sắc càng đậm thể hiện giá trị càng tốt.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MI** | **QG** | **PSNR** | **EN** | **SD** | **ALI** | **VIF** |
| **LP-NestFuse** | 2.8512 | 0.5427 | 7.1866 | 7.2536 | 0.1829 | 0.5183 | 0.9214 |
| **BTSFusion** | 1.6411 | 0.4903 | 7.0667 | 6.6975 | 0.1240 | 0.4530 | 0.6385 |
| **CrossFuse** | 2.8767 | 0.4490 | 7.1442 | 6.9289 | 0.1556 | 0.4012 | 0.8752 |
| **MPCFusion** | 1.8472 | 0.4908 | 7.0263 | 6.9064 | 0.1429 | 0.4610 | 0.6545 |
| **NestFuse** | 2.4761 | 0.3587 | 7.1825 | 6.9247 | 0.1758 | 0.4586 | 0.9167 |
| **PSLPT** | 1.4314 | 0.2917 | 6.4793 | 6.5320 | 0.1100 | 0.4128 | 0.4628 |

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

Kết quả bảng 1, 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.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MI** | **QG** | **PSNR** | **EN** | **SD** | **ALI** | **VIF** |
| **PP1** | 2.8428 | 0.4929 | 7.1866 | 7.2405 | 0.1820 | 0.5182 | 0.9119 |
| **PP2.1** | 2.8507 | 0.5445 | 7.1866 | 7.2527 | 0.1829 | 0.5183 | 0.9199 |
| **PP2.2** | 2.8513 | 0.5437 | 7.1866 | 7.2533 | 0.1829 | 0.5183 | 0.9209 |
| **LP-NestFuse** | 2.8512 | 0.5427 | 7.1866 | 7.2536 | 0.1829 | 0.5183 | 0.9214 |
| **PP3** | 2.8555 | 0.5388 | 7.1866 | 7.2460 | 0.1825 | 0.5182 | 0.9145 |
| **PP4** | 2.8319 | 0.5239 | 7.1866 | 7.2457 | 0.1826 | 0.5181 | 0.9098 |
| **PP5** | 2.8254 | 0.5306 | 7.1866 | 7.2510 | 0.1828 | 0.5182 | 0.9089 |
| **PP6** | 2.7899 | 0.4409 | 7.1865 | 7.2320 | 0.1814 | 0.5181 | 0.9028 |

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MI** | **QG** | **PSNR** | **EN** | **SD** | **ALI** | **VIF** |
| **LP-NestFuse** | 2.8512 | 0.5427 | 7.1866 | 7.2536 | 0.1829 | 0.5183 | 0.9214 |
| **PP1** | 2.9589 | 0.5385 | 7.1851 | 7.1745 | 0.1757 | 0.5021 | 0.9022 |
| **PP2** | 1.7531 | 0.5531 | 7.1761 | 7.1700 | 0.1579 | 0.4380 | 0.5975 |
| **PP3** | 1.9208 | 0.5315 | 7.1414 | 6.9589 | 0.1426 | 0.4593 | 0.7532 |
| **PP4** | 1.8366 | 0.5611 | 7.1694 | 7.1041 | 0.1524 | 0.4364 | 0.6076 |
| **PP5** | 2.1726 | 0.5612 | 7.1509 | 7.1010 | 0.1603 | 0.4292 | 0.7876 |

*Bảng 3. So sánh giữa các phương pháp tổng hợp thành phần chi tiết*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MI** | **QG** | **PSNR** | **EN** | **SD** | **ALI** | **VIF** |
| **LP-NestFuse\_Old** | 2.4483 | 0.5460 | 7.1849 | 7.1773 | 0.1702 | 0.4958 | 0.8651 |
| **LP-NestFuse** | 2.8512 | 0.5427 | 7.1866 | 7.2536 | 0.1829 | 0.5183 | 0.9214 |
| **NestFuse** | 2.4761 | 0.3587 | 7.1825 | 6.9247 | 0.1758 | 0.4586 | 0.9167 |
| **NestFuse\_New** | 2.5743 | 0.3718 | 7.1828 | 6.9122 | 0.1685 | 0.4650 | 0.9060 |

*Bảng 4. So sánh tác động của LP và chiến lược tổng hợp mới so với mô hình NestFuse gốc*

Trong bảng 2, các phương pháp dựa trên năng lượng vùng cực đại (MRE) tốt nhất hầu hết trên tất cả các chỉ số. Và trong các bộ trọng số cho MRE được thực nghiệm gồm bộ trọng số 1, bộ trọng số Gauss và bộ trọng số Binomial thì bộ trọng số Binomial mà nghiên cứu sử dụng mang lại kết quả tốt nhất. Bảng 3, so sánh các phương pháp tổng hợp thành phần chi tiết, LP-NestFuse vẫn đảm bảo tính toàn diện trên hầu hết các bộ chỉ số. Hai bảng kết quả này chứng minh tính định lượng cho việc lựa chọn hàm năng lượng vùng cực đại cho tổng hợp thành phần cơ sở và NestFuse cải tiến cho tổng hợp thành phần chi tiết.

Kết quả thực nghiệm từ bảng 4 cho thấy, việc kết hợp đồng thời phương pháp phân rã LP biến đổi và chiến lược tổng hợp mới mang lại sự vượt trội so với việc không sử dụng hoặc sử dụng chỉ một thành phần, đặc biệt tốt trong vấn đề bảo toàn thông tin cạnh được truyền đi (QG) và độ tương phản, cường độ sáng của hình ảnh tổng hợp. Khi kết hợp với phương pháp tổng hợp dựa trên phân rã Laplacian Pyramid, bảo toàn thông tin cạnh và cường độ sáng hình ảnh thực sự được cải thiện hơn nhiều. Có thể thấy, giá trị chỉ số QG khi dùng Laplacian Pyramid tốt hơn 50% so với việc không sử dụng, ALI tốt hơn 10%. Các chỉ số như PSNR hay EN cũng được cải thiện.

A diagram of a hexagon

Description automatically generated A diagram of a hexagon with different colored lines

Description automatically generated

|  |  |  |
| --- | --- | --- |
| *Hình* ***[ ].*** *Biểu đồ radar so sánh giữa phương pháp đề xuất và các phương pháp khác* |  | *Hình* ***[ ].*** *Biểu đồ radar so sánh giữa các phương pháp tổng hợp thành phần cơ sở* |

A diagram of a graph

Description automatically generated with medium confidence A diagram of a hexagon

Description automatically generated

|  |  |  |
| --- | --- | --- |
| *Hình* ***[ ].*** *Biểu đồ radar so sánh giữa các phương pháp tổng hợp thành phần chi tiết* |  | *Hình* ***[ ].*** *. Biểu đồ radar so sánh tác động của Laplacian Pyramid và chiến lược tổng hợp mới so với mô hình NestFuse gốc* |

**5. Conclusion**

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 lai kết hợp giữa phương pháp truyền thống và phương pháp tổng hợp bằng mô hình học sâu. Trong đó, nghiên cứu đã đề xuất một phương pháp phân rã kim tự tháp Laplacian mới như đã trình bày trong thuật toán 3 ở phần 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 nhằm tăng cường thông tin vùng quan trọng từ hai hình ảnh 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.

Trong tương lai, chúng tôi dự định sẽ giải quyết một số vấn đề để cải tiến hiệu suất của mô hình hiện tại. Vấn đề đầu tiên về lượng thông tin tương hỗ được truyền từ hình ảnh nguồn tới hình ảnh tổng hợp thể hiện qua chỉ số MI chưa cạnh tranh với các phương pháp khác, tuy nhiên vẫn phải đảm bảo tính bảo toàn cạnh và chất lượng hình ảnh tổng hợp. Ví dụ như việc tổng hợp thành phần cơ sở dựa trên các thuật toán tối ưu hóa như các thuật toán metaheuristic, các thuật toán dựa theo bầy đàn có thể mang lại sự đảm bảo về chất lượng hình ảnh đầu ra. Vấn đề thứ hai về tăng khả năng thích ứng của mô hình và tăng chất lượng kết cấu của hình ảnh tổng hợp, có thể sử dụng các mô hình học sâu tiên tiến chẳng hạn như học chuyển giao (transfer learning) hoặc các mô hình dựa trên Transformer.

**Declarations**

Xung đột lợi ích: Các tác giả tuyên bố rằng họ không có lợi ích cạnh tranh.

**References**