CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

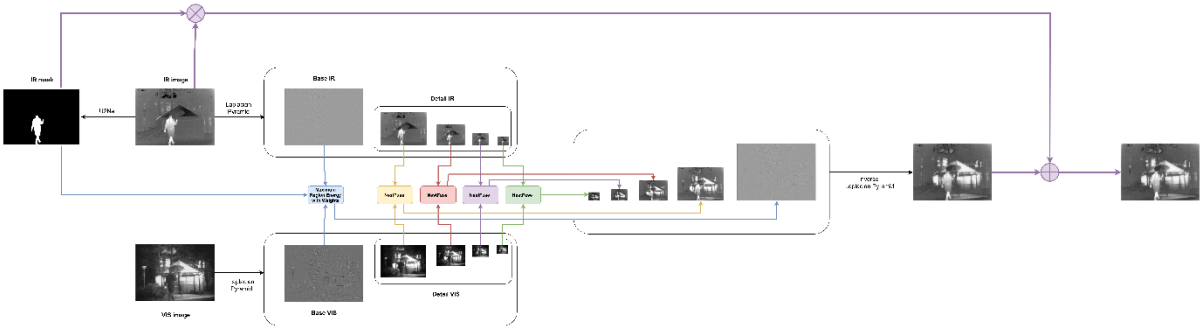
3.1. Tổng quan giải pháp

Trong báo cáo này, kiến trúc tổng quan của phương pháp đề xuất được minh họa trong hình 1. 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 1. 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.

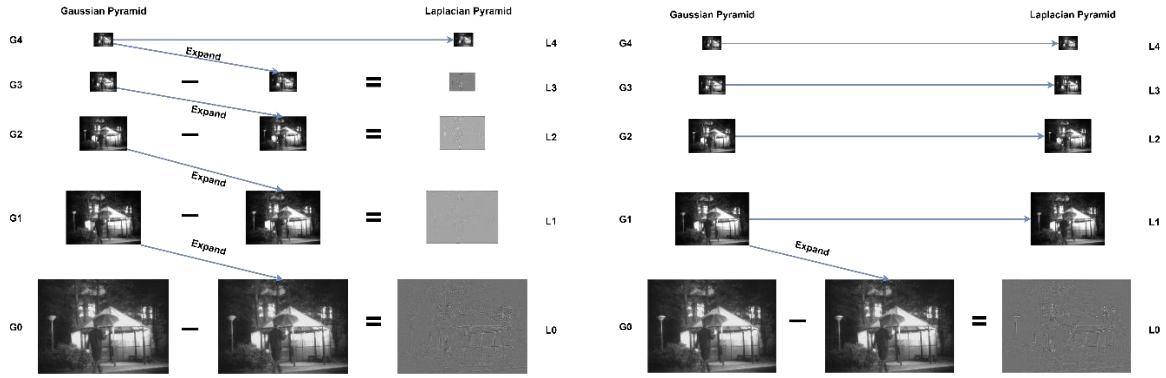
Các thành phần tần số cao của LP truyền thống được xây dựng theo quy trình biểu diễn như sau:

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

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 2 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 2: 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 3 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:

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

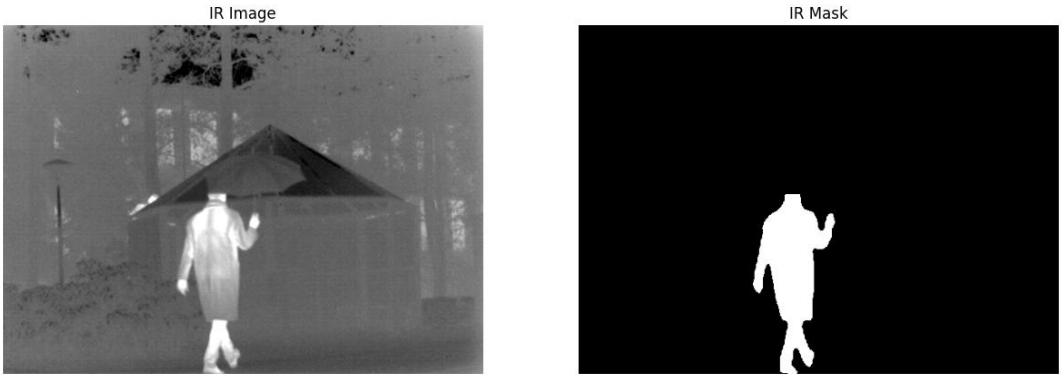
$$RE_{VIS}(m,n) = \sum_{(m',n') \in W} \omega_{m'n'} \left[L_{0,VIS}(m+m',n+n')\right]^2$$

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

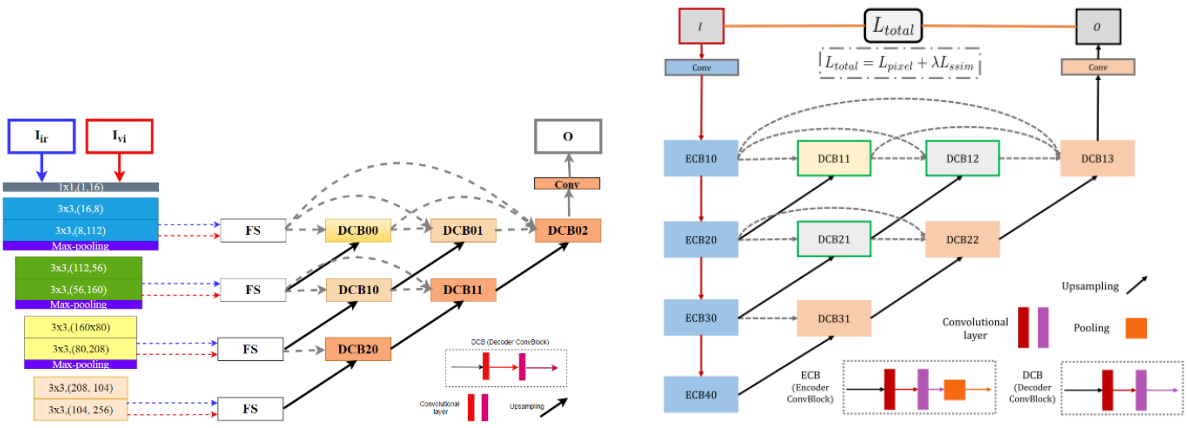
3.4. NestFuse cho tổng hợp các thành phần chi tiết

Đối với các thành phần chi tiết, mô hình NestFuse [2] được sử dụng để thực hiện hợp nhất. 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.

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.

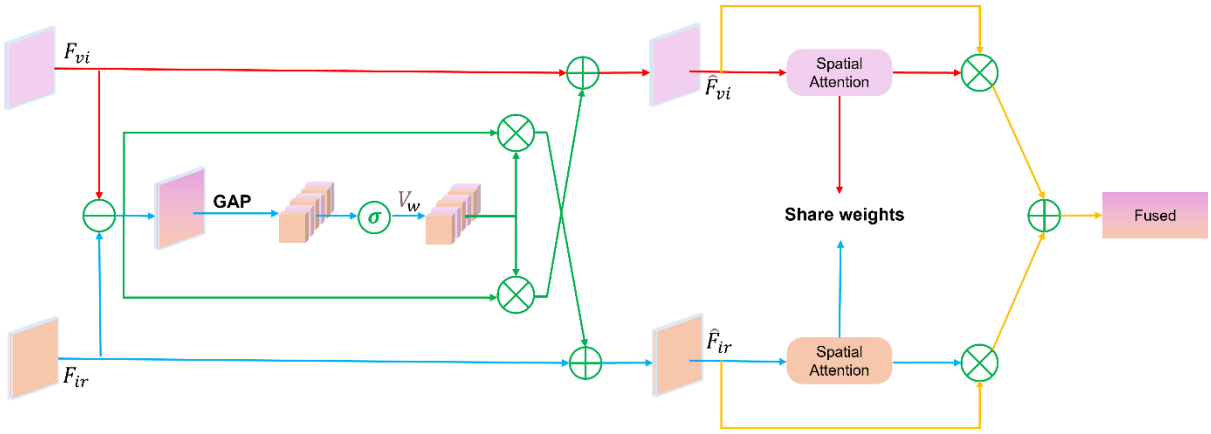
Kiến trúc của mô hình NestFuse được thể hiện chi tiết trong hình 4. Trong đó, mỗi block Encoder và Decoder đều chứa các layer Convolution (mỗi khối chứa 2 layer), riêng các khối Encoder có thêm một layer MaxPooling để có được nhiều tỷ lệ khác nhau. NestFuse dành cho pha train không chứa block FS (Fusion 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.

Trong báo cáo này, 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 (FS) mới thay vì tổng hợp dựa trên cơ chế Channel Attention và Spatial Attention như mô hình NestFuse đã thực hiện.



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

Hình 4. Kiến trúc mô hình NestFuse



Hình 5. 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 5 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 6 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

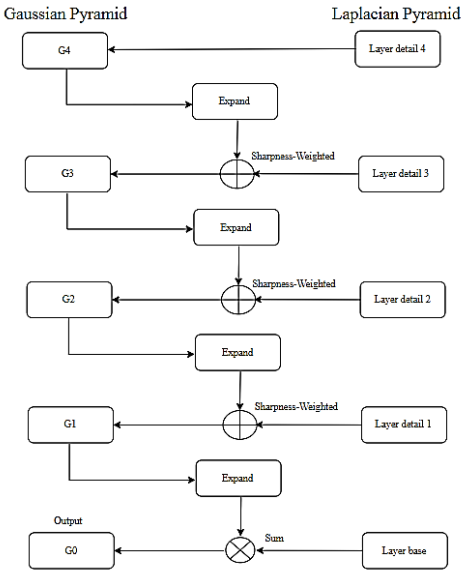
$$I_{reconstruct} = expand(I_{reconstruct}) + L_0$$

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