ĐẠI HỌC KHOA HỌC TỰ NHIÊN - ĐẠI HỌC QUỐC GIA TP HCM
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

TRÍ TUỆ NHÂN TẠO NÂNG CAO

Bài báo: Selective frequency network for image restoration

Giảng viên hướng dẫn:

TS. Nguyễn Ngọc Thảo

Sinh viên thực hiện:

24C11051 - Lưu Thiện Đức

24C11061 - Phạm Phú Hoàng Sơn

24C11067 - Nguyễn Anh Tuấn

24C11071 - Phạm Quốc Vương

K34 - Học phần 1

Thành phố Hồ Chí Minh, tháng 03 năm 2025

NỘI DUNG

I. Phân chia công việc

Ghi chú: Bài báo có 4 task khôi phục ảnh chính gồm: Khử mưa, khử sương, khử mờ, khử tuyết.

MSHV	Họ Tên	Công việc thực hiện
24C11051	Lưu Thiện Đức	• **Task chính:** Khử tuyết • Chạy thử code trước khi chọn bài báo • Chạy thực nghiệm với tác vụ khử tuyết với dữ liệu trong và ngoài bài báo • Chuẩn bị nội dung slide và viết báo cáo.
24C11061	Phạm Phú Hoàng Sơn	• **Task chính:** Khử mờ • Chạy thử code trước khi chọn bài báo • Chạy thực nghiệm với dữ liệu khử mờ HIDE trong bài báo và BlueReal_J, BlueReal_R ngoài bài báo. • Viết slide, báo cáo
24C11067	Nguyễn Anh Tuấn	• **Task chính:** Khử sương • Chạy thử code trước khi chọn bài báo • Thực nghiệm tác vụ khử sương mù, fine-tune mô hình trên 2 tập dữ liệu I-Haze và O-Haze. • Chuẩn bị nội dung slide và viết báo cáo.
24C11071	Phạm Quốc Vương	• **Task chính:** Khử mưa • Chạy thử code trước khi chọn bài báo • Chạy thực nghiệm với các tập dữ liệu trong và ngoài bài báo cho task khử mưa • Fine-tune task khử mưa và thực nghiệm • Thực nghiệm thêm: Dùng open-cv cắt các frame ảnh của video từ mạng xã hội để thực nghiệm với pre-trained model và fine-tuned model • Đọc nội dung Identify scientific, nuisance, and fixed hyperparameters (Deep Learning Tuning Playbook by Google) và lấy thông tin hữu ích • Bổ sung một số nội dung phần Methodology trong slide và các slide thực nghiệm khử mưa

II. Giới thiệu bài báo

1. Giới thiệu tổng quát

Khôi phục hình ảnh (Image Restoration) là một bài toán quan trọng trong lĩnh vực thị giác máy tính, nhằm tái tạo hình ảnh chất lượng cao từ những ảnh bị suy giảm do nhiễu, mờ hoặc ảnh hưởng từ môi trường như mưa, tuyết hay sương mù. Bài toán này có nhiều ứng dụng thực tiễn quan trọng trong giám sát an ninh, công nghệ xe tự hành và cảm biến từ xa, giúp nâng cao chất lượng dữ liệu đầu vào cho các hệ thống xử lý tự động.

2. Định nghĩa bài toán

Bài toán khôi phục hình ảnh trong bài báo tập trung vào việc cải thiện chất lượng của những bức ảnh bị suy giảm do nhiều yếu tố khác nhau, chẳng hạn như: hiện tượng mờ do mất nét hoặc chuyển động (defocus/motion blur), sương mù (haze), mưa (rain), tuyết (snow).
Mục tiêu là khôi phục hình ảnh sắc nét, rõ ràng hơn, giảm thiểu các tác nhân làm suy giảm thông tin.
Nghiên cứu này tập trung vào việc nâng cao hiệu suất khôi phục ảnh bằng cách ứng dụng mạng nơ-ron sâu. Cụ thể, công trình đề xuất hai mô-đun mới – MDSF và MCSF – nhằm tối ưu hóa quá trình xử lý trong miền tần số, từ đó cải thiện đáng kể chất lượng ảnh đầu ra.
Phạm vi của nghiên cứu:

- Khử mờ do mất nét hoặc chuyển động (defocus/motion deblurring)
- Loại bỏ sương mù (dehazing)
- Loại bỏ mưa (deraining)
- Loại bỏ tuyết (desnowing)

3. Động lực nghiên cứu

Các phương pháp truyền thống dựa trên giả định thống kê hoặc đặc trưng thủ công không còn phù hợp với các điều kiện phức tạp trong thực tế. Trong khi đó, các mô hình học sâu hiện đại, đặc biệt là CNN và Transformer, vẫn gặp hạn chế như vùng cảm thụ (receptive field) nhỏ hoặc chi phí tính toán cao (Transformer). Bên cạnh đó, các phương pháp dựa trên biến đổi tần số như Fourier/Wavelet cũng chưa tối ưu do khó chọn lọc thông tin quan trọng nhất và yêu cầu biến đổi nghịch phức tạp.

Nếu giải quyết được bài toán này, nghiên cứu sẽ giúp nâng cao chất lượng ảnh trong nhiều tình huống thực tế như giám sát giao thông, phân tích ảnh vệ tinh, hỗ trợ bác sĩ trong chẩn đoán hình ảnh y tế, hay cải thiện trải nghiệm thị giác trong các ứng dụng thực tế ảo.

III. Công trình liên quan

Các nghiên cứu trước đây về khôi phục ảnh có thể chia thành ba nhóm chính:

Đối với các phương pháp truyền thống dựa trên giả định thống kê hoặc đặc trưng thủ công hoặc áp dụng các thuật toán tối ưu hóa để khôi phục ảnh bị suy giảm nhưng đều không đủ linh hoạt khi áp dụng cho các điều kiện thực tế phức tạp.

Đối với các phương pháp dựa trên học sâu (Deep Learning) chia thành 2 nhóm

- Dựa trên CNN nhưng gặp vấn đề với vùng cảm thụ (receptive field) nhỏ, khó nắm bắt thông tin dài hạn.
- Dựa trên Transformer: Được ứng dụng để mở rộng vùng cảm thụ và nắm bắt quan hệ dài hạn trong ảnh (long-range dependencies). Nhưng lại gặp vấn đề với việc phải cần tài nguyên tính toán lớn, làm giảm khả năng ứng dụng thực tế.

Đối với các phương pháp dựa trên miền tần số đã tận dụng được trong miền tần số (frequency domain) bằng cách sử dụng Fourier/Wavelet để biến đổi từ miền không gian(spatial domain) sang miền tần số.Nhưng chúng lại không có cơ chế chọn lọc động thông tin tần số quan trọng nhất, yêu cầu biến đổi nghịch làm tăng chi phí tính toán.

Nghiên cứu này cách khắc phục các hạn chế trên với việc đề xuất hai module mới để tận dụng thông tin tần số một cách linh hoạt và hiệu quả hơn:

- Multi-branch Dynamic Selective Frequency (MDSF):
  - Sử dụng bộ lọc động để tách ảnh thành các thành phần tần số khác nhau.
  - Áp dụng cơ chế attention để chọn lọc thông tin quan trọng nhất.
- Multi-branch Compact Selective Frequency (MCSF):
  - Mở rộng vùng cảm thụ bằng cách kết hợp các kỹ thuật pooling thay vì dùng convolution.
  - Giảm chi phí tính toán so với các phương pháp Transformer và biến đổi Fourier/Wavelet.

Nhờ đó, mô hình không chỉ cải thiện chất lượng ảnh phục hồi mà còn có khả năng ứng dụng thực tế cao hơn do tối ưu hóa hiệu suất tính toán.

IV. Phương pháp đề xuất

1. Kiến trúc tổng quát

Mạng trong SFNet được xây dựng theo kiến trúc encoder-decoder, cho phép học các biểu diễn phân cấp để xử lý dữ liệu đầu vào một cách hiệu quả

Kiến trúc tổng thể:

- SFNet có ba tầng trong cả encoder và decoder với mỗi tầng chứa một ResBlock.
- Với module MDSF chỉ có ở khối dư cuối (last residual block) của mỗi ResBlock, còn module MCSF nằm ở tất cả các khối.

Luồng xử lý của mô hình SFNet được thể hiện qua 4 bước chính sau:

- Ảnh đầu vào bị mờ: SFNet sử dụng cơ chế đa đầu vào (multi-input), tức là ảnh sẽ được giảm kích thước (downsampled) thành nhiều phiên bản có độ phân giải khác nhau để xử lý tốt hơn.
- Encoder: Ảnh đi qua ba tầng encoder, sử dụng lớp down-sampling với strided convolutions để giảm kích thước không gian của ảnh, giúp mô hình tập trung vào các đặc trưng quan trọng.
- Decoder: Ảnh được tái tạo dần qua ba tầng decoder, sử dụng up-sampling với transposed convolution và skip connections để phục hồi chi tiết. Sau mỗi tầng decoder, một ảnh tạm thời được tạo ra do cơ chế đa đầu ra (multi-output), giúp ổn định huấn luyện.
- Ảnh đầu ra đã được khôi phục: Ảnh sau mỗi tầng decoder đều qua convolution 3×3 để cải thiện chất lượng. Ảnh cuối cùng sắc nét hơn, chi tiết rõ hơn và loại bỏ hiệu ứng mờ hiệu quả.

2. Multi-branch Dynamic Selective Frequency (MDSF)

Mô-đun Multi-Branch Dynamic Selective Frequency (MDSF) là một thành phần quan trọng trong mô hình SFNet, nhằm chọn lọc và tăng cường các thành phần tần số quan trọng trong quá trình khôi phục hình ảnh.

MDSF gồm hai thành phần chính:

- Bộ tách tần số (Decoupler): Phân tách đặc trưng thành các thành phần tần số cao và thấp dựa trên các bộ học học được (Learnerable Filters).
- Bộ điều chế tần số (Modulator): Dùng cơ chế chú ý theo kênh (channel-wise attention) để lựa chọn thành phần tần số quan trọng nhất.

Trong phương pháp đề xuất, nhóm tác giả thực hiện phân tách động feature map thành hai thành phần tần số thấp và tần số cao để tăng cường khả năng trích xuất đặc trưng hiệu quả hơn. Điều này được thực hiện thông qua việc sử dụng các bộ lọc học được (learnable filters), giúp mô hình phân biệt các thông tin tổng quát và chi tiết biên trong ảnh.

Cụ thể Low Filter F^l được sinh ra từ lớp sinh bộ lọc (filter-generating layer). Bộ lọc này có kích thước kernel k x k và được chia thành g nhóm kênh để giảm số lượng tham số, giúp mô hình gọn nhẹ hơn.

$$F^l = \text{Softmax}((\mathcal{B}(W(\text{GAP}(X)))))$$

- B là Batch Normalization
- W là trọng số tích chập
- GAP là global averaging pooling

High Filter F^h được tính bằng cách trừ F^l với một identity kernel. Với identity kernel có dạng

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Sau khi có bộ lọc F^l và F^h, feature map đầu vào được tách thành hai phần:

- Thành phần tần số thấp X^l chứa các thông tin tổng quát (cạnh, góc, màu sắc trong ảnh)
- Thành phần tần số cao X^h chứa các thông tin chi tiết (hình dạng, kết cấu)

Công thức thức tính toán của hai thành phần trên như sau

$$X^l_{i,c,h,w} = \sum_{p,q} F^L_{i,p,q} X_{i,c,h+p,w+q}; \quad X^h_{i,c,h,w} = \sum_{p,q} F^H_{i,p,q} X_{i,c,h+p,w+q}$$

- c là chỉ số kênh.
- h, w là tọa độ trong không gian.
- p, q thuộc các giá trị {-1, 0, 1}

Sau khi phân tách, Bộ điều chỉnh tần số (frequency modulator) sẽ nhấn mạnh phần thực sự hữu ích cho việc tái tạo. Quá trình này được thực hiện theo chiều kênh (channel dimension) dựa trên một phiên bản cải tiến của SKNet (Li et al., 2020).

Bộ điều chỉnh tần số hợp nhất 2 thành phần trên với công thức

$$Z = W_{fc}(\mathrm{GAP}(X^l + X^h))$$

- W_{fc} là tham số của fully connected layer(FC)

Sau khi tính toán Z, hai fully connected layer được sử dụng để tính toán trọng số của channel-wise attention cho hai thành phần tần số trên:

$$[W^l, W^h]_c = \frac{e^{[W_l(Z), W_h(Z)]_c}}{\sum_j^{2C} e^{[W_l(Z), W_h(Z)]_j}}$$

- W^l là trọng số của attention cho đặc trưng tần số thấp.
- W^h là trọng số của attention cho đặc trưng tần số cao.
- $[\cdot, \cdot]$ biểu diễn phép nối (concatenation).

Để tận dụng đặc trưng từ nhiều mức độ chi tiết khác nhau, phương pháp này có thể được mở rộng thành kiến trúc đa nhánh (multi-branch structure), trong đó mỗi nhánh sử dụng các bộ lọc có kích thước khác nhau để trích xuất đặc trưng từ nhiều mức độ tần số. Đặc trưng cuối cùng $\widehat{X}$ được tổng hợp từ các nhánh này, được tính như sau:

$$\hat{X} = [\mathcal{M}_1(\mathcal{D}_1(X_1)), ..., \mathcal{M}_m(\mathcal{D}_m(X_m))]$$

- D là bộ phân tách tần số (decoupler)
- M là bộ điều chỉnh tần số (modulator)
- X_m là đặc trưng đã được phân chia đều.

3. Multi-branch Compact Selective Frequency (MCSF)

Các vết mờ/hỏng trên mỗi mẫu ảnh có kích thước khác nhau là một trong những vấn đề của phục hồi ảnh. Điều đó làm cho vai trò của Receptive Field được đề cao để có thể phát hiện các vùng hỏng lớn lẫn nhỏ. Từ đó, bài báo hướng đến việc mở rộng Receptive Field (RF) với module MCSF.

MCSF gồm hai nhánh là global RF và window-based RF. Cả hai đều dựa trên window-based attention. Tuy nhiên, global RF áp dụng lên toàn bộ input thay vì chia window như window-based RF.

Các bước xử lý trên nhánh window-based RF:

1. Chia input thành 4 window
2. Lấy thông tin tần số thấp: Dùng Global Average Pooling
3. Lấy thông tin tần số cao: Lấy các window trừ low-frequency map
4. Chọn các dãy tần hữu ích: Rescale các low-freq và high-freq map bằng learnable weights
5. Chuyển các frequency map về kích thước input

MCSF khác MDSF ở chỗ, nó có sự mở rộng RF, cũng như không dùng Conv Freq-Decoupler và Conv Freq-Modulator. Điều đó giúp giảm chi phí tính toán kéo theo việc có thể đặt được module MCSF nhiều lần trên ResBlock.

4. Hàm mất mát sử dụng

$$L_{\text{spatial}} = \sum_{r=1}^{3} \frac{1}{S_r} \left\| \hat{X}_r - Y_r \right\|_1$$

$$L_{\text{frequency}} = \sum_{r=1}^{3} \frac{1}{S_r} \left\| \mathcal{F}(\hat{X}_r) - \mathcal{F}(Y_r) \right\|_1$$

Với r biểu thị chỉ số của các ảnh đầu vào/đầu ra với các độ phân giải khác nhau;
$\mathcal{F}$ đại diện cho phép biến đổi Fourier nhanh (Fast Fourier Transform - FFT);
S_r là số phần tử dùng để chuẩn hóa; và $\hat{X}_r$, Y_r lần lượt là ảnh đầu ra và ảnh mục tiêu.

Hàm mất mát cuối cùng được xác định là:
$L = L_{\text{spatial}} + \lambda L_{\text{frequency}}$
trong đó λ được đặt là 0.1.

V. Thực nghiệm

Lưu ý: *Xem kết quả thực nghiệm được mô tả trực quan trong slide.*
Độ đo (metric):

- SSIM (chỉ số tương đồng cấu trúc): Đo lường sự tương đồng giữa ảnh khôi phục và ảnh gốc dựa trên các đặc trưng cấu trúc, giá trị càng gần 1 thì hai ảnh càng giống nhau
- PSNR (tỉ số tín hiệu trên nhiễu đỉnh): Đánh giá sự khác biệt về cường độ điểm ảnh giữa hai ảnh, dựa trên độ lệch bình phương trình bình (MSE), đơn vị là dB(Decibel) với giá trị càng cao thì ảnh khôi phục càng giống ảnh gốc

Lưu ý khi chạy code:

- Task khử mờ:
  - Chạy file notebook kèm theo.
  - Tải dữ liệu vào thư mục data huấn luyện của paper hoặc dữ liệu ngoài file .zip vào đường dẫn /content/drive/MyDrive/Dữ liệu .zip
  - Sau đó giải nén vào thư mục /content/SFnet/Motion_deblurring/, lưu ý là khi giải nén ra phải có folder /content/SFnet/Motion_deblurring/test.
  - Tải model GOPrO.pkl đã được tiền huấn luyện vào folder /content/SFNet/Motion_debluring/models
  - Chạy câu lệnh !python main.py --data HIDE --mode test --data_dir "data_path" --test_model "model_path" --save_image True
- Task khử mưa:
  - Chạy file notebook kèm theo.
  - Dataset, source code, output data đã được chuẩn bị và ghi chú trong từng file notebook.
  - Các điểm nên lưu ý đến nếu chạy code bị lỗi (train và evaluate):
    - Đảm bảo thư mục input (ảnh hỏng) và target (ground truth) có số lượng file giống nhau;
    - Mỗi file ở input phải có file cùng tên tương ứng ở target;

  - Chỉnh sửa số mẫu trong dataset bằng bội số của batch size nếu đang train code bị dừng;
  - Batch size lớn hơn 4 có thể khiến cấu hình máy ảo Kaggle không đáp ứng đủ tài nguyên train;
  - Đổi tên các dataset phù hợp trong file code evaluate;
  - Resize ảnh inference được về đúng kích thước ảnh lỗi trước khi chạy code evaluate.
- Task khử tuyết:
  - Chạy file notebook kèm theo.
  - Dataset, Pretrained-model và cách chạy đã được hướng dẫn trong Readme.md của thư mục /desnowing

VI. Nhận xét, Kết luận

- Mô hình overfit mạnh trên các tập test của bài báo (do cùng đặc điểm với dữ liệu train, hoặc đều là các tập dữ liệu con của một tập dữ liệu cha so với dữ liệu train);
- Mô hình underfit trên các tập dữ liệu ngoài dù có một vài đặc điểm tương đồng với dữ liệu train;
- Fine-tune trên dữ liệu ngoài cho kết quả khá tốt. Điều này chứng tỏ phương pháp mà bài báo đề xuất có tính tái sử dụng cao.
- Với dữ liệu khử mờ thì mô hình tiền huấn luyện chạy tốt cho các bộ dữ liệu mờ tự nhiên chưa được nén (bộ RealBlue_R) và yếu trên bộ ảnh đã nén JPED (RealBlur_J). Do mô hình đã được train trên tập GoPro lớn nên với các bộ dữ liệu tương đồng nó hoạt động khá tốt. Tuy kích cỡ mô hình không quá lớn nhưng kết quả ảnh ra rất đẹp , thời gian suy luận cũng nhanh (tập HIDE chạy lại trên kaggle là 1h38p với 2025 iter).
- Bài báo làm với đa tác vụ, suy ra mô hình có khả năng tương thích với các bài toán tương tự cao

VII. Tài liệu tham khảo

[1] H. Zhang, H. Xie, and H. Yao, "Blur-aware Spatio-temporal Sparse Transformer for Video Deblurring", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[2] X. Mao, Q. Li, and Y. Wang, "AdaRevD: Adaptive Patch Exiting Reversible Decoder Pushes the Limit of Image Deblurring," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[3] J. Rim, H. Lee, J. Won, and S. Cho, "Real-World Blur Dataset for Learning and Benchmarking Deblurring Algorithms," in Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[4] Y. Cui et al., "Selective Frequency Network for Image Restoration," in Proceedings of the 11th International Conference on Learning Representations (ICLR), 2023.