ĐẠI HỌC BÁCH KHOA HÀ NỘI

Trường Điện – Điện tử

BÁO CÁO THỊ GIÁC MÁY TÍNH

Phân đoạn ảnh khuyết điểm bề mặt thép sử dụng U-Net và Selective Kernels

Nhóm 10:

Lê Doãn Đạt	20202595
Nguyễn Việt Thành	20202524
Nguyễn Xuân Toàn	20204611

Giảng viên hướng dẫn: PGS.TS. Phạm Văn Trường

PGS.TS. Trần Thị thảo

Khoa: Tự Động Hóa

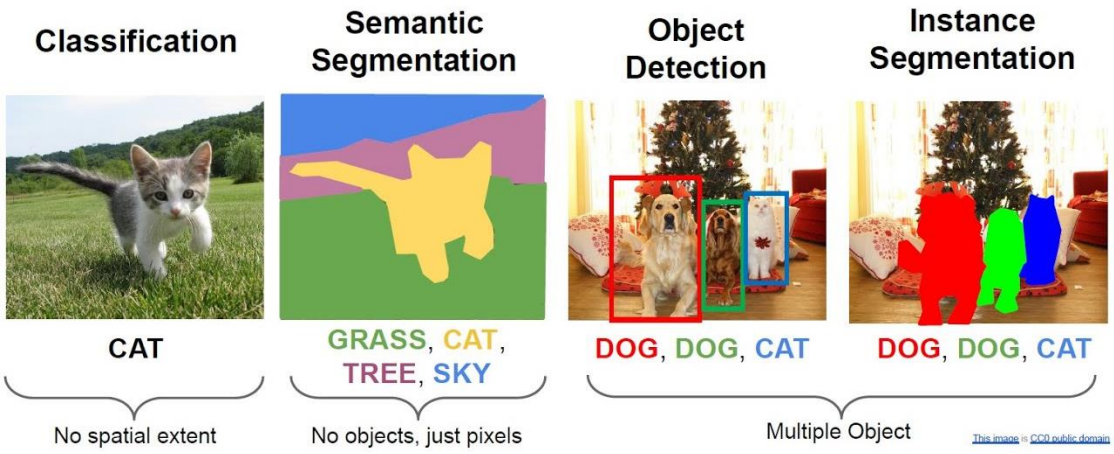
Trường: Điện – Điện tử

HÀ NỘI, 1/2024

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU

Bài toán phân đoạn hình ảnh (image segmentation) là một tác vụ quan trọng trong thị giác máy tính. Nhiệm vụ của phân đoạn là gán nhãn cho từng pixel trong khung hình thuộc về các lớp quy định. Tùy thuộc vào việc gán nhãn mà sẽ được chia thành một số loại tác vụ phân đoạn khác nhau, như theo Hình 1.



Hình 1. Phân loại tác vụ phân đoạn hình ảnh

Lĩnh vực phân đoạn hình ảnh đã trải qua nhiều phát triển và nghiên cứu trong việc tìm kiếm các phương pháp hiệu quả. Trước đó, các kỹ thuật cơ bản như k-means clustering và thresholding được sử dụng để phân đoạn hình ảnh dựa trên giá trị pixel hoặc màu sắc.

Các phương pháp truyền thống như Region Growing, Watershed, và Active Contour Model đã được áp dụng để phân đoạn vùng quan tâm trong hình ảnh. Tuy nhiên, những phương pháp này thường gặp khó khăn khi xử lý các ảnh có độ phức tạp cao hoặc chứa nhiễu.

Sự xuất hiện của các phương pháp dựa vào học máy như Support Vector Machines (SVM), Random Forests, và kỹ thuật phân loại dựa trên đặc trưng (feature-based classification) đã mang lại sự cải tiến trong việc phân đoạn hình ảnh. Tuy nhiên, các phương pháp này thường đòi hỏi quá trình trích xuất đặc trưng được thiết kế một cách tỉ mỉ và có thể không đủ linh hoạt đối với đa dạng các nhiệm vụ phân đoạn.

U-net [1], được giới thiệu bởi Olaf Ronneberger, Philipp Fischer và Thomas Brox vào năm 2015, đã đánh dấu một bước tiến lớn trong lĩnh vực phân đoạn hình ảnh. Mô hình này có cấu trúc encoder-decoder và sử dụng kỹ thuật "skip connections" để kết hợp thông tin cấp cao và cấp thấp. U-net nhanh chóng trở thành một trong những mô hình quan trọng và phổ biến nhất trong việc giải quyết các vấn đề phân đoạn hình ảnh.

Việc phân đoạn khuyết tật trong thép đóng một vai trò quan trọng trong ngành công nghiệp sản xuất và chế biến thép. Việc này không chỉ giúp cải thiện chất lượng sản phẩm mà còn đảm bảo an toàn và độ tin cậy của các thành phẩm thép... Điều này không chỉ tối ưu hóa quá trình sản xuất mà còn giúp doanh nghiệp giảm thiểu lãng phí và tăng tính hiệu quả trong quản lý chất lượng. Sự cần thiết của việc

phân đoạn khuyết tật trong ngành công nghiệp thép ngày càng trở nên quan trọng, đóng góp vào sự phát triển bền vững và chất lượng cao trong sản xuất thép.

Với độ hiệu quả đã được kiểm chứng qua nhiều bài toán phân đoạn của U-net, báo cáo này xây dựng mô hình cấu hình chữ U để phân đoạn khuyết tật trong thép. Trong mô hình U-net, khối chuyển tiếp nằm giữa encoder và decoder đóng vai trò rất lớn. Tầm quan trọng của khối bottleneck ở cuối U-Net nằm ở việc nó giúp tăng cường khả năng tái tạo và giữ lại thông tin chi tiết trong quá trình đầu ra. Khối bottleneck giúp mô hình tập trung vào việc tái tạo chi tiết nhỏ và làm cho kết quả đầu ra trở nên chính xác và rõ ràng. Điều này đặc biệt quan trọng trong các ứng dụng như segmentation, nơi mà việc bảo toàn đặc trưng chi tiết có vai trò quan trọng trong việc định vị và phân loại các vùng trong ảnh. Nhằm tăng khả năng thích ứng, tính linh hoạt, báo cáo này xây dựng khối bottleneck sử dụng Selective Kernels, được giới thiệu lần đầu bởi Xiang Li và cộng sự [2].

Việc thay thế khối Bottleneck trong mô hình U-Net bằng selective kernels thay vì convolution thông thường mang lại nhiều lợi ích cho bài toán segmentation. Khối Bottleneck truyền thống thường sử dụng các lớp convolution để giảm kích thước của đặc trưng và tăng độ sâu của mô hình. Tuy nhiên, sử dụng selective kernels có thể cung cấp một cách tiếp cận linh hoạt và hiệu quả hơn trong việc xử lý thông tin.

Thay vì sử dụng convolution với kích thước cố định, selective kernels cho phép mô hình chọn lựa kernel phù hợp dựa trên đặc trưng cụ thể tại từng vùng của ảnh. Điều này giúp tăng cường khả năng tổng quát hóa của mô hình, đặc biệt là khi đối mặt với các vùng ảnh có cấu trúc phức tạp. Thêm vào đó, việc sử dụng selective kernels có thể giảm lượng tham số cần học so với việc sử dụng convolution với kích thước lớn

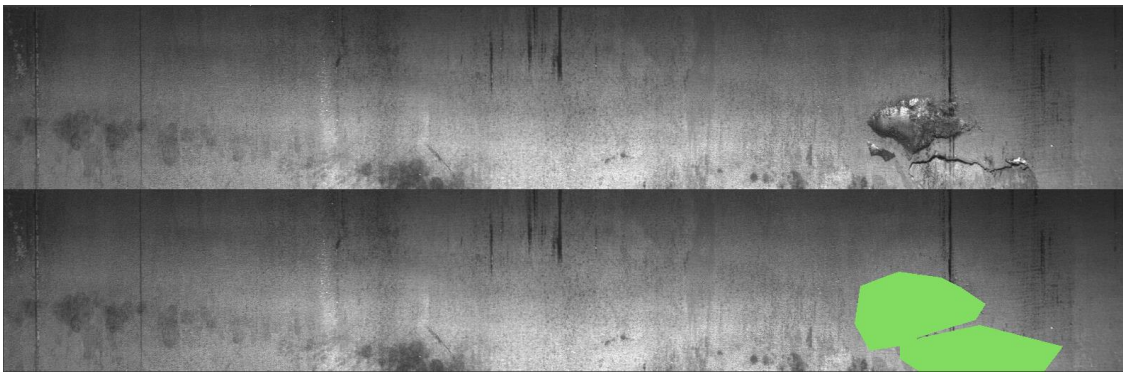
Sự kết hợp của U-net và Selective Kernel mang lại sự hiệu quả trong phân đoạn hình ảnh khuyết điểm bề mặt thép, tốt hơn so với mô hình U-net thông thường.

CHƯƠNG 2. MÔ HÌNH HUẤN LUYỆN

2.1 Tập dữ liệu

Tập dữ liệu hình ảnh về khuyết điểm của thép sử dụng cho quá trình huấn luyện mạng và đánh giá chất lượng mô hình được lấy từ tập dữ liệu [2] nhưng không phải toàn bộ. Do tập dữ liệu gốc phục vụ cho nhiều tác vụ khác nhau, báo cáo này chỉ chọn một phần dữ liệu phù hợp cho bài toán phân đoạn hình ảnh.

Tập dữ liệu bao gồm 801 ảnh, đều có độ phân giải 256 x 1600 điểm ảnh. Chúng em phân chia thành 3 tập huấn luyện, kiểm thử, và kiểm tra; lần lượt gồm 701, 50, và 50 ảnh. Mỗi ảnh đều đính kèm một ma trận mã hóa nhãn của từng điểm ảnh là có lỗi hoặc không có lỗi.



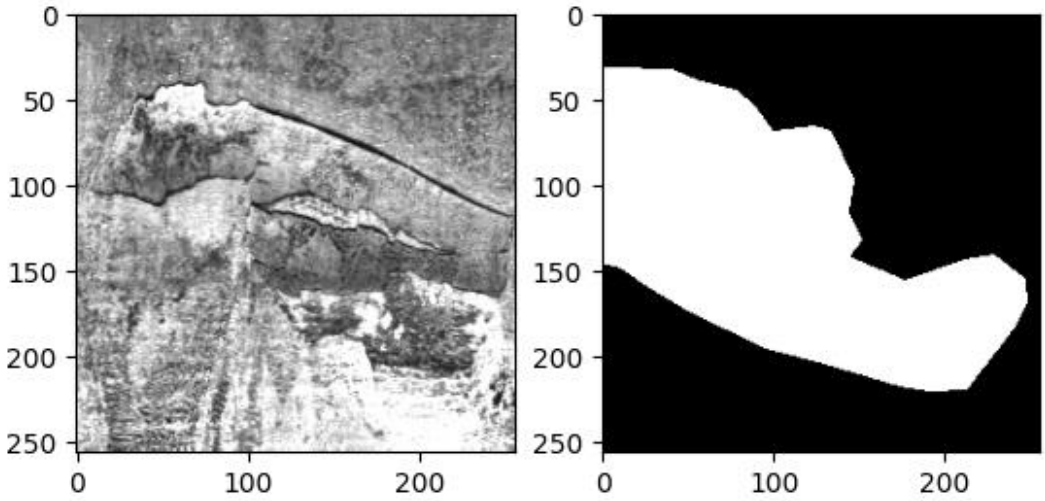
Hình 2. Ảnh chụp bề mặt thép (bên trên) và khuyết điểm (bên dưới)

Do kích thước của ảnh lớn và khuyết điểm chỉ tập trung ở một khoảng nhỏ, chúng em sẽ cắt ảnh gốc thành các ảnh nhỏ kích thước 256 x 256. Đối với tập huấn luyện, các điểm cắt sẽ cách nhau 40 điểm ảnh, còn đối với tập kiểm thử sẽ là 130 điểm ảnh. Do vậy mỗi ảnh ở tập huấn luyện bước đầu được 33 ảnh nhỏ, mỗi ảnh ở tập kiểm thử được chia thành 12 ảnh nhỏ. Tuy nhiên, chúng em chỉ lấy những ảnh có số lượng điểm khuyết tật lớn hơn 12000 đối với tập huấn luyện, 5000 đối với tập kiểm thử, ví dụ ở Hình 3.

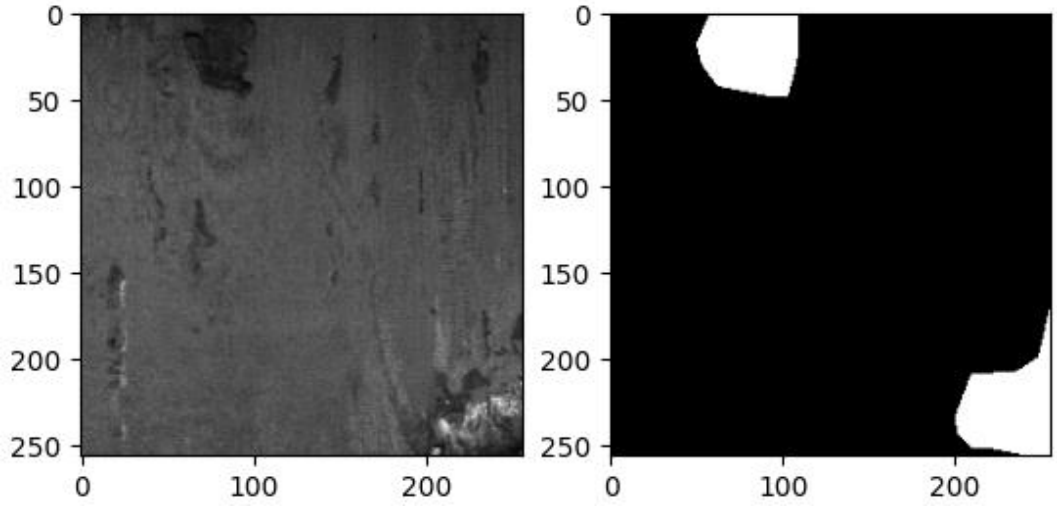
Qua quá trình tiền xử lý, thu được tập huấn luyện gồm 4389 ảnh, tập kiểm thử 165 ảnh, và tập kiểm tra 189 ảnh; đều có kích thước 256 x 256.

2.2 Xây dựng mô hình

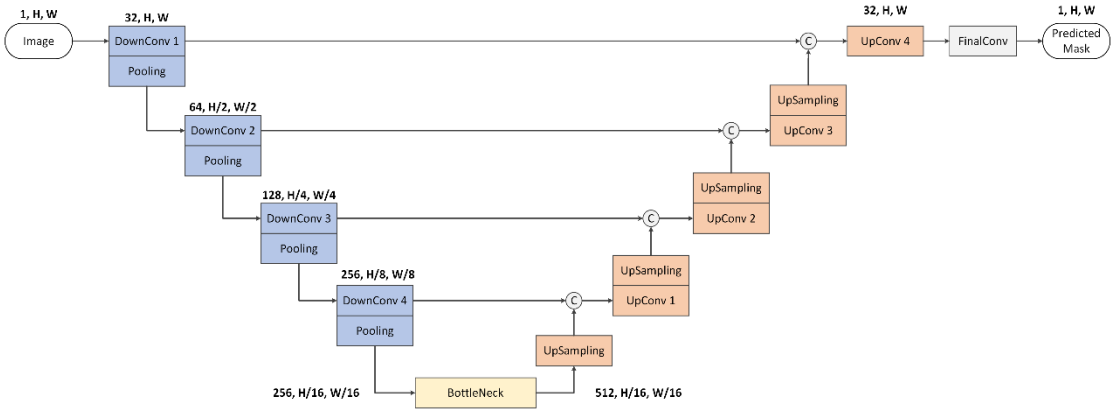
Mô hình được thừa hưởng các đặc trưng của U-net có cấu trúc encoder-decoder với các tầng tích chập giúp trích xuất thông tin về mặt không gian cũng như ngữ nghĩa theo các cấp độ khác nhau, kết hợp nối tắt để giữ được thông tin ảnh xuyên suốt mô hình. Mô hình hoàn chỉnh được mô tả ở Hình 4.



Hình 3a. Ảnh huấn luyện



Hình 3b. Ảnh kiểm thử



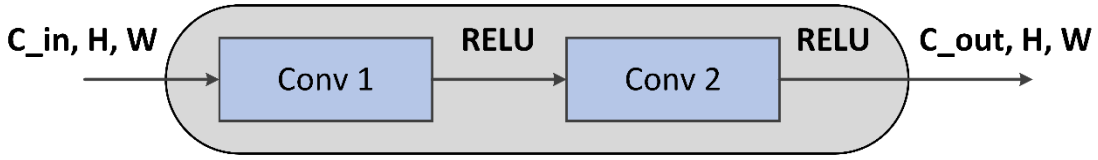
Hình 4. Cấu trúc mô hình đề xuất

Phía encoder với nhiệm vụ dùng các lớp CNN trích xuất đặc trưng ảnh ở các cấp độ khác nhau, kết hợp các lớp Pooling để giảm kích thước đặc trưng. Càng đi sâu qua các tầng encoder, đặc trưng của ảnh càng mang nhiều ngữ nghĩa, trái lại là mất đi thông tin chi tiết về không gian so với những lớp phía trên. Phía decoder của U-

Net đóng vai trò quan trọng trong việc phục hồi chi tiết và cấu trúc của đối tượng trong ảnh, dựa trên thông tin đã được học từ các tầng encoder. Ảnh đặc trưng qua mỗi tầng được tăng kích thước bằng mạng tích chập chuyển vị, đồng thời kết hợp đặc trưng từ phía encoder tương ứng thông qua nối tắt, tiếp tục sử dụng các lớp tích chập để học đặc trưng và chi tiết cụ thể trong quá trình tái tạo. Qua khối decoder, ảnh dự đoán sẽ được tái tạo bằng đúng kích thước của ảnh đầu vào.

2.2.1 DoubleConv

Các khối DownConv, UpConv có tác dụng trích xuất đặc trưng ảnh theo các cấp độ khác nhau, đều có cấu trúc gồm 2 lớp tích chập hai chiều kích thước 3x3 và hàm kích hoạt RELU.

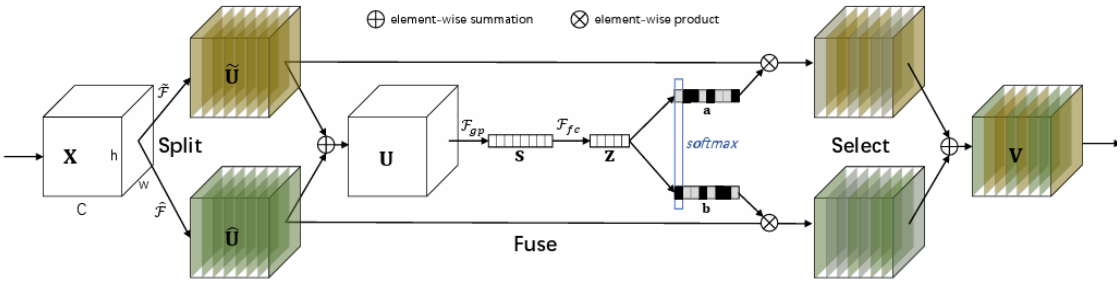


Hình 5. Khối DoubleConv

Trong mô hình của chúng em, ảnh đầu vào là ảnh xám kích thước (1, H, W) từ quá trình chuẩn bị dữ liệu đã nêu. Ảnh đầu vào sẽ qua 4 tầng encoder, mỗi tầng gồm khối DownConv và lớp Max Pooling 2x2 để giảm kích thước ảnh đặc trưng. Qua mỗi tầng, kích thước đặc trưng sẽ giảm một nửa, và số kênh sẽ tăng gấp đôi từ 32 lên đến 256. Tương ứng sẽ có 4 tầng giải mã, tăng kích thước ảnh sử dụng lớp tích chập chuyển vị, kết hợp đặc trưng từ phía encoder và khối tích chập. Cuối cùng, kết quả phân đoạn sẽ là một ảnh kích thước 1 x H x W thu được từ 32 kênh đặc trưng cuối cùng.

2.2.2 Bottleneck

Khối chuyển tiếp bottleneck được xây dựng sử dụng Selective Kernels. Hình 6 mô tả cấu trúc của Selective Kernel



Hình 6. Selective Kernel

Cấu trúc của Selective Kernels sử dụng trong mô hình có 2 nhánh, được chia thành 3 giai đoạn: Split, Fuse, Select.

Split:

- Thiết lập 2 biến đổi $\tilde{F}$ và $\hat{F}$ gồm: depthwise convolution, Batch Normaliztion (BN), và RELU

- Hai biến đổi $\tilde{F}$ (kernel size = 3, dilation = 1) và $\hat{F}$ (kernel size = 3, dilation = 2) thực thi trên đầu vào X, thu được đầu ra $\tilde{U}$ và $\hat{U}$ tương ứng.

Fuse:

- Kết quả từ 2 nhánh biến đổi được kết hợp với nhau bằng element-wise summation:

$$\boldsymbol{U} = \tilde{\boldsymbol{U}} + \hat{\boldsymbol{U}}$$

- "Global average pooling" được tính toán trên $\boldsymbol{U}$ để trích xuất thông tin toàn cục.

$$s_c = F_{gp}(\boldsymbol{U}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{U}_c(i, j)$$

- S được nén thành đặc trưng $\boldsymbol{Z} \in \mathbb{R}^{d \times 1}$ có ít chiều hơn bằng một lớp kết nối đầy đủ (FC), nhằm giữ lại những đặc trưng quan trọng.

$$\boldsymbol{Z} = F_{fc}(\boldsymbol{S}) = \delta\big(\beta(\boldsymbol{W}\boldsymbol{S})\big)$$

trong đó, δ là hàm kích hoạt ReLU, β là batch normalization (BN), $\boldsymbol{W} \in \mathbb{R}^{d \times C}$

- Để đánh giá sự ảnh hưởng của d lên chất lượng mô hình, sử dụng tham số r để kiểm soát giá trị:

$$d = \max(C/r, L)$$

L là giá trị nhỏ nhất d có thể nhận (chọn $L = 128$)

Select:

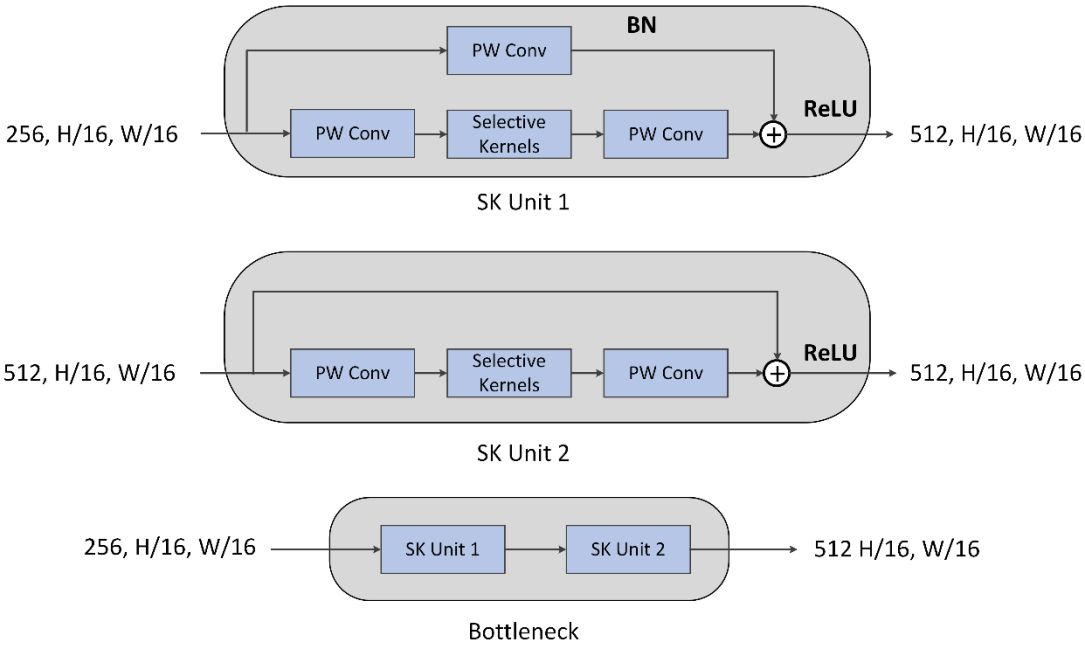
Một cơ chế chú ý trên các kênh được sử dụng để linh hoạt lựa chọn kích thước vùng thông tin. Cụ thể, một toán tử softmax được áp dụng trên các số liệu theo chiều kênh:

$$a_c = \frac{e^{\boldsymbol{A}_c \boldsymbol{z}}}{e^{\boldsymbol{A}_c \boldsymbol{z}} + e^{\boldsymbol{B}_c \boldsymbol{z}}}, b_c = \frac{e^{\boldsymbol{B}_c \boldsymbol{z}}}{e^{\boldsymbol{A}_c \boldsymbol{z}} + e^{\boldsymbol{B}_c \boldsymbol{z}}}$$

Đặc trưng $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_c]$ với $\boldsymbol{V}_c \in \mathbb{R}^{H \times W}$ cuối cùng thu được từ tổ hợp:

$$\boldsymbol{V}_c = a_c \cdot \tilde{\boldsymbol{U}}_c + b_c \cdot \hat{\boldsymbol{U}}_c$$

Cụ thể trong mô hình nhóm xây dựng, Selective Kernels có các đặc tính như sau: $M = 2, G = 32, r = 8, L = 128$. Khối Bottleneck trong mô hình được cấu tạo từ các 2 SK Unit liên tiếp nhau, mỗi SK Unit gồm 3 khối theo thứ tự: 1x1 convolution, SK convolution, 1x1 convolution. Trong mỗi SK Unit, bổ sung thêm nối tắt để giúp thông tin được củng cổ thêm. Đặc biệt ở SK Unit 1, nối tắt cần đi qua thêm 1 lớp PW Conv để tăng chiều phù hợp với đầu ra yêu cầu.



Hình 7. Khối Bottleneck

CHƯƠNG 3. QUÁ TRÌNH HUẤN LUYỆN MẠNG

3.1 Huấn luyện mạng

Tập dữ liệu huấn luyện gồm 4389 ảnh 256x256, được ngẫu nhiên lật theo chiều ngang và dọc, mạng sẽ được kiểm thử trên tập gồm 165 ảnh trong suốt quá trình huấn luyện.

Quá trình huấn luyện được thực hiện qua 100 epochs, số lượng 16 ảnh với một batch, sử dụng thuật toán tối ưu Adam với tốc độ học khởi tạo là 0.0001 cho hàm mục tiêu BCE-Dice Loss:

$$\mathcal{L}(y,p) = \mathcal{L}_{BCE}(y,p) + \mathcal{L}_{Dice}(y,p) \tag{1}$$

$$\mathcal{L}_{BCE}(y,p) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)] \tag{2}$$

$$\mathcal{L}_{Dice}(y,p) = 1 - \frac{2\sum_{i=1}^{N} y_i p_i}{\sum_{i=1}^{N}(y_i + p_i)} \tag{3}$$

trong đó, $y_i \in \{0,1\}$ và $p_i \in (0,1)$ lần lượt là giá trị ground-truth và xác suất tính được của điểm ảnh thứ i.

3.2 Thang đánh giá chất lượng

Dice Similarity Coefficient (DSC) và Intersection over Union (IoU) là hai thang đo được sử dụng để đánh giá chất lượng của mô hình đề xuất. Định nghĩa $s_i \in \{0,1\}$ là nhãn mà mô hình gán cho pixel thứ i. Phương trình của hai thang đo được mô tả như sau:

$$DSC(y,s) = \frac{2\sum_{i=1}^{N} y_i s_i}{\sum_{i=1}^{N}(y_i + s_i) + \varepsilon} \tag{4}$$

$$IoU(y,s) = \frac{2\sum_{i=1}^{N} y_i s_i}{\sum_{i=1}^{N}(y_i + s_i - y_i s_i) + \varepsilon} \tag{5}$$

Hệ số làm mịn được chọn là $\varepsilon = 10^{-5}$

CHƯƠNG 4. KẾT QUẢ

Như đã đề cập ở trên, kết quả của mạng U-Net kết hợp Selective Kernels Bottleneck sẽ được so sánh với U-Net với khối Bottleneck được cấu tạo từ DoubleConv.

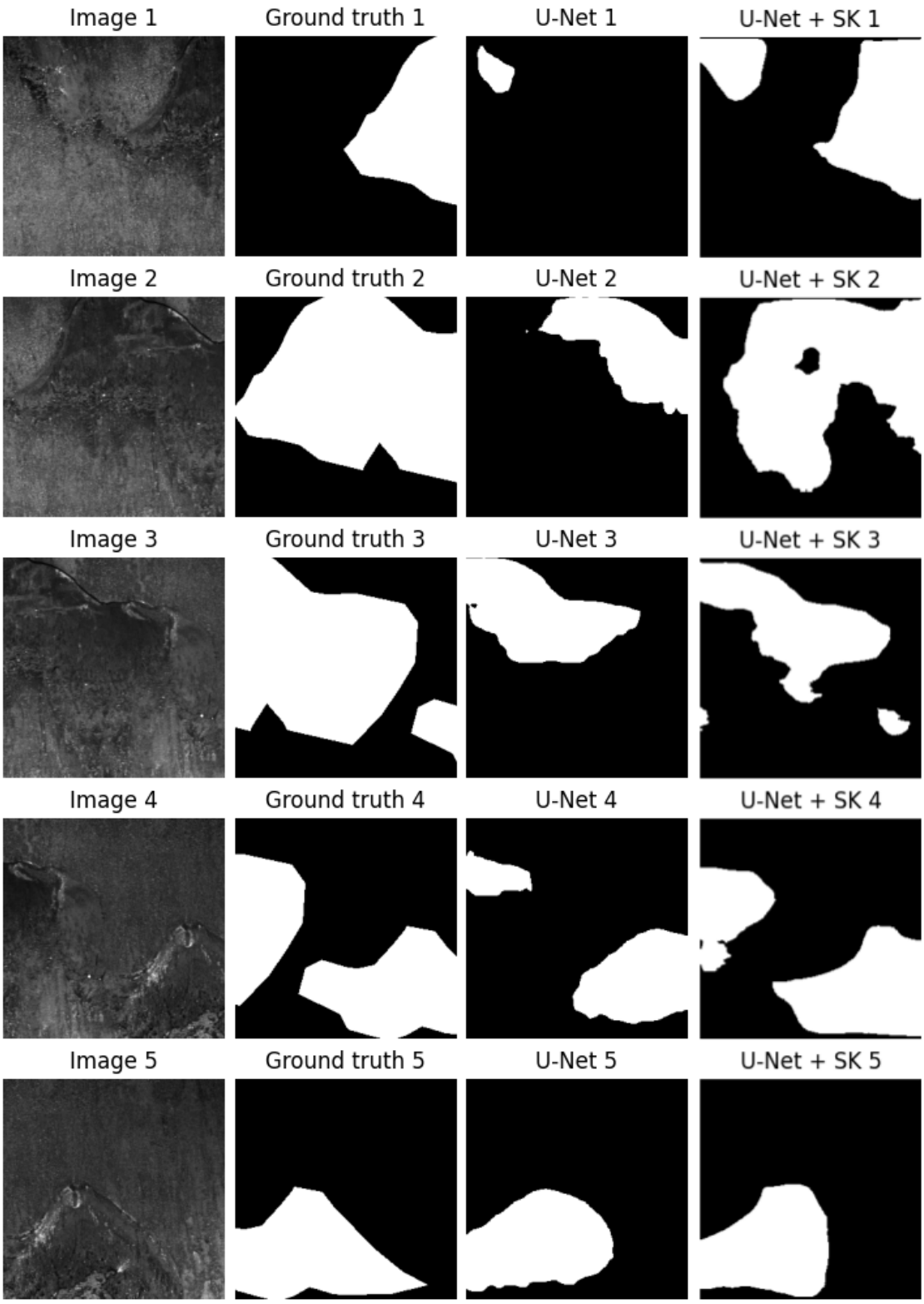
Kết quả phân đoạn của một số ảnh ở tập kiểm tra được thể hiện ở hình 6. Có thể thấy mô hình đề xuất có thể phát hiện và phân đoạn được khuyết điểm trên bề mặt thép ở các vị trí và kích thước khác nhau.

Kết quả so sánh giữa hai mô hình đã cho thấy sự cải thiện đáng kể khi tích hợp Selective Kernels (SK) vào kiến trúc U-Net. Đầu tiên, việc chỉ thay SK vào khối Bottleneck đã giảm đáng kể tham số mạng (từ 7.8 triệu còn 5.6 triệu tham số). Mặc dù U-net đã đạt được điểm Dice và IoU tốt với 77.38 và 63.14, nhưng khi kết hợp với Selective Kernels, mô hình đề xuất đã đạt được kết quả cao hơn với 80.96 (Dice) và 68.04 (IoU).

Sự gia tăng này có thể được giải thích bằng cách mà Selective Kernels giúp mô hình tập trung vào việc xử lý các đặc trưng quan trọng và giảm thiểu ảnh hưởng của nhiễu trong quá trình segmentation. Điều này không chỉ làm tăng độ chính xác của việc dự đoán mà còn cải thiện khả năng tổng quát hóa của mô hình đối với dữ liệu mới. Sự so sánh này là một bằng chứng cho hiệu quả của việc tích hợp Selective Kernels nhằm nâng cao khả năng hiệu quả của mô hình U-net trong bài toán segmentation.

Bảng 1. Kết quả đánh giá tập kiểm tra

	U-Net	**U-Net + SK Bottleneck**
Dice	77.38	80.96
IoU	63.14	68.04



Hình 6. Hình ảnh kết quả phân đoạn tập kiểm tra

KẾT LUẬN

Trong báo cáo này, chúng em đã đề xuất mô hình có cấu hình chữ U kết hợp Selective Kernels để xử lí tác vụ phân đoạn ảnh khuyết điểm trên bề mặt thép. Sự hiệu quả của cấu trúc encoder-decoder kết hợp nối tắt giúp đặc trưng ảnh được trích xuất theo các cấp độ và xuyên suốt mô hình. Kết quả thu được rất khả quan, có thể được ứng dụng trong môi trường thực tế. Trong các nghiên cứu tiếp theo, mô hình còn có thể tận dụng nhiều hơn khả năng của SK giúp đồng thời giảm số lượng tham số mạng và tăng hiệu quả phân đoạn ảnh.

TÀI LIỆU THAM KHẢO

[1] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[2] X. Li, W. Wang, X. Hu and J. Yang, "Selective Kernel Networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 510-519, doi: 10.1109/CVPR.2019.00060.

[3] Alexey Grishin, BorisV, iBardintsev, inversion, Oleg. (2019). Severstal: Steel Defect Detection. Kaggle. https://kaggle.com/competitions/severstal-steel-defect-detection