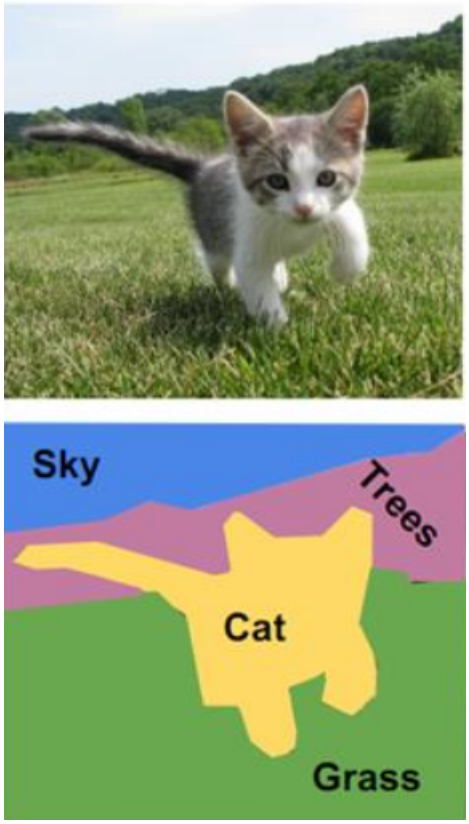
IT4653 HỌC SÂU VÀ ỨNG DỤNG **Bài đọc số : 8**

Phân đoạn hình ảnh - Image segmentation

Người viết : Trịnh Anh Phúc *Người kiểm tra : x*

1 Giới thiệu

Bài toán phân đoạn ảnh (Image segmentation) là một trong như bài toán quan trọng trong lĩnh vực Thị Giác Máy Tính (Computer Vision) ngoài ra cũng là bài toán gần với các bài toán khác như phát hiện đối tượng (object detection). Nó giúp cho người dùng xác định các vùng điểm ảnh (pixel segmentation) tương ứng với lớp phân loại nào. Trong trường hợp này, thường là ảnh tĩnh và bài toán là bài toán học có giám sát với kích thước ảnh đầu vào bằng kích thước ảnh đầu ra sau phân đoạn.



Hình 1: Bài toán phân đoạn hình ảnh, theo ví dụ hình chụp con mèo chạy trên đồng cỏ, nên trời, cây cỏ đều cần phân đoạn. Mọi điểm ảnh (pixel) đều cần được phân loại vào các vùng (segment) tương ứng các lớp được nhận diện.

Cũng như bài toán phát hiện ảnh, bài toán phân đoạn ảnh có ảnh hướng lớn đến ứng dụng của Thị Giác Máy Tính trong các lĩnh vực quan trọng như xây dựng, không ảnh vệ tinh, xe tự lái, lĩnh vực ảnh y tế, nhận dạng hóa đơn, phân vùng ảnh văn bản viết tay. Bài toán phân đoạn khá gần với bài toán phát hiện đối tượng nên hướng tiếp cận vẫn là cửa số trượt và bài toán hồi quy các phân đoạn, tuy nhiên sau đó chuyển sang hướng tiếp cận một giai đoạn do chi phí tính toán cao.

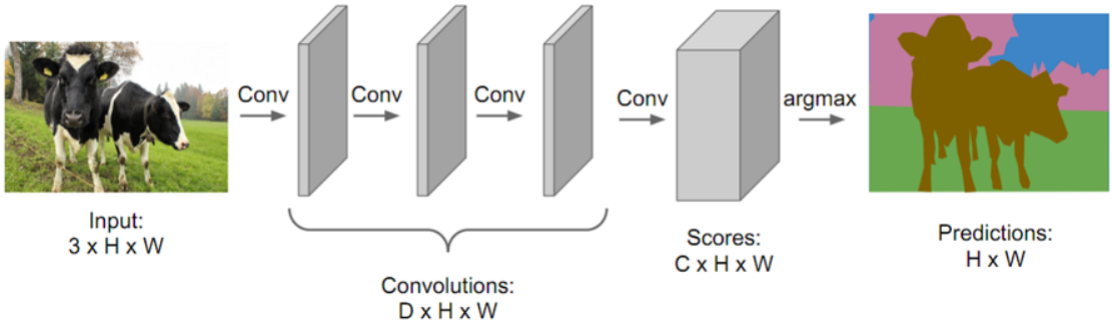
Nội dung bài đọc sẽ được chia tuần tự thành ba phần

1. Mạng nơ ron tích chập hoàn toàn (Fully Covolutional Neural Networks)
   (a) Kiến trúc chung của mạng nơ ron phân đoạn
   (b) Lớp mạng nơ ron tăng mẫu
   (c) Hàm mục tiêu của mạng phân đoạn
2. Một số mạng phân đoạn tiêu biểu

2 Mạng nơ ron tích chập hoàn toàn (Fully Covolutional Neural Networks)

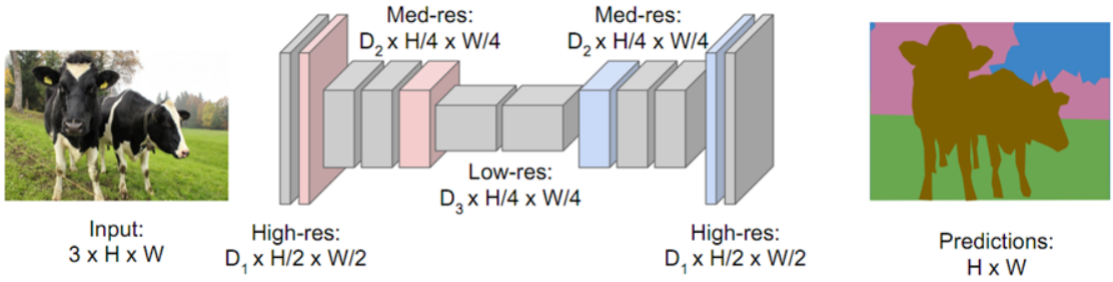
2.1 Kiến trúc chung của mạng nơ ron phân đoạn

Bài toán phân đoạn là bài toán học có giám sát, theo đó, kích thước ảnh đầu ra $H \times W$ bằng kích thước ảnh đầu vào $H \times W$ tuy nhiên tại mỗi vị trí ảnh đầu ra ta cần xác định giá trị softmax của C lớp dự đoán để tạo thành ma trận tích chập ba chiều (tương tự bài toán phát hiện đối tượng).



Hình 2: Bài toán phân đoạn hình ảnh, theo ví dụ hình chụp hình hai con bò nền cỏ , cây cối và trời xanh. Thao tác argmax thực hiện trên từng điểm ảnh sẽ xác định lớp dự đoán của điểm ảnh tương ứng với phân đoạn nào.

Tuy nhiên, với mạng nơ ron trích xuất đặc trưng thông thường sẽ gồm các lớp giảm mẫu, mục đích là co kích thước (H, W) của hình ảnh đầu vào nên các kiến thúc mạng nơ ron phân đoạn nói chung cần các lớp tăng mẫu để tái lập kích thước ảnh đầu ra. Ta cần giới thiệu các lớp tăng mẫu và chuyển tích chập này để đầu ra tích chập $H \times W \times$ có kích thước ảnh cũng như lớp sẽ thường lớn hơn kích thước ban đầu. Cũng như phần đầu, phần sau sẽ có thể tham số hóa để học được quá trình tăng độ phân giải. Đôi khi người ta gọi kiến trúc dạng này là kiến trúc chữ V.



Hình 3: Kiên trúc đối xứng của mạng nơ ron dùng cho bài toán phân đoạn. Phần đầu sẽ làm giảm độ phân giải (resolution) từ cao về thấp. Phần sau sẽ làm tăng độ phân giải từ thấp về cao lại để đảm bảo kích thước đầu ra bằng kích thước đầu vào.

2.2 Các lớp mạng nơ ron tăng mẫu

Cũng như lớp tích chập, lớp giảm mẫu có tham số, không tham số thì các lớp tăng độ phân giải cũng có các lớp có tham số và không tham số.

2.2.1 Lớp chuyển vị - Transposed Convolutional Neural Layer

Thông thường, lớp chuyển vị tạo nên ánh xạ ngược của lớp tích chập, theo đó nó sinh ra đầu ra bản đồ kích hoạt có kích thước lớn hơn kích thước lớp tích chập vào. Lớp chuyển vị được dùng nhiều với các tác vụ học như sinh hình ảnh, tăng độ phân giải, phân đoạn ảnh và các mô hình sinh nói chung. Lớp này đặc biệt hữu dụng trong trường hợp tăng mẫu dữ liệu đầu vào, chẳng hạn tăng độ phân giải dữ liệu đầu vào hoặc sinh ra hình ảnh từ tập nhiều (noise), e.g. mô hình sinh.

Hoạt động của lớp chuyển vị đối lập với lớp tích chập theo đó, bộ lọc hay nhân tích chập sẽ chạy trên từng phần tử của đầu vào thực hiện phép **nhân** tích chập, trượt trên bản đồ đầu ra, thực hiện **tổng** các bản đồ đầu ra để có bản đồ tích chập cuối cùng.

Đề dễ hình dung ta có ví dụ như sau

1. Lớp tích chập đầu vào $2 \times 2 \times 1$
2. Bộ lọc kích thước $2 \times 2 \times 1$
3. Độ dài bước $S = 2$
4. Viền thêm $P = 0$

Như vậy, trước hết ta phải tính kích thước của bản đồ kích hoạt đầu ra

$$H_{out} = (H-1) \times S + FH - 2 \times P,\ W_{out} = (W-1) \times S + FW - 2 \times P$$

Như vậy, lắp vào công thức trên ta có $H_{out} = (2-1) \times 2 + 2 - 2 \times 0 = 4, W_{out} = (2-1) \times 2 + 2 - 2 \times 0 = 4$ nghĩa là bản đồ kích hoạt đầu ra của lớp tăng mẫu là 4×4

$$\underbrace{\begin{bmatrix}0 & 1\\ 2 & 3\end{bmatrix}}_{\text{Đầu vào x}} * \underbrace{\begin{bmatrix}4 & 1\\ 2 & 3\end{bmatrix}^T}_{\text{Bộ lọc w}} = \begin{bmatrix}0 & 0 & \\ 0 & 0 & \\ & & \end{bmatrix} + \begin{bmatrix} & 4 & 1\\ & 2 & 3\\ & & \end{bmatrix} + \begin{bmatrix} & & \\ & & \\ 8 & 2 & \end{bmatrix} + \begin{bmatrix} & & \\ & & \\ & 12 & 3\\ & 6 & 9\end{bmatrix} = \underbrace{\begin{bmatrix}0 & 0 & 4 & 1\\ 0 & 0 & 2 & 3\\ 8 & 2 & 12 & 3\\ 4 & 6 & 6 & 9\end{bmatrix}}_{\text{Bản đồ kích hoạt y}}$$

Cùng ví dụ này, nhưng ta thay đổi bước nhẩy $S = 1$ thì ta có $H_{out} = 3, W_{out} = 3$ và bản đồ kích hoạt đầu ra sẽ khác

$$\underbrace{\begin{bmatrix}0 & 1\\ 2 & 3\end{bmatrix}}_{\text{Đầu vào x}} * \underbrace{\begin{bmatrix}4 & 1\\ 2 & 3\end{bmatrix}^T}_{\text{Bộ lọc w}} = \begin{bmatrix}0 & 0 & \\ 0 & 0 & \\ & & \end{bmatrix} + \begin{bmatrix} 4 & 1 & \\ 2 & 3 & \\ & & \end{bmatrix} + \begin{bmatrix}8 & 2 & \\ 4 & 6 & \\ & & \end{bmatrix} + \begin{bmatrix} 12 & 3\\ 6 & 9\end{bmatrix} = \underbrace{\begin{bmatrix}0 & 4 & 1\\ 8 & 16 & 6\\ 4 & 12 & 9\end{bmatrix}}_{\text{Bản đồ kích hoạt y}}$$

2.3 Lớp tăng mẫu - Upsampling

Lớp chuyển vị 2D ở trên cũng là lớp tăng mẫu nhưng có tham số, phần này ta chỉ trình bày tăng mẫu không tham số - upooling layer. Trong tình huống ta muốn tăng mẫu, có thể dùng hai hướng tiếp cận

- Láng giềng gần nhất (Nearest Neighbors) trong cùng bộ lọc thì gán cùng giá trị

$$\begin{bmatrix}4 & 1\\ 2 & 3\end{bmatrix} \underbrace{\Longrightarrow}_{\text{Láng giềng gần nhất}} \begin{bmatrix}4 & 4 & 1 & 1\\ 4 & 4 & 1 & 1\\ 2 & 2 & 3 & 3\\ 2 & 2 & 3 & 3\end{bmatrix}$$

- Làm khó (Beg of nails) trong cùng vị trí bộ lọc thì điền cố định một giá trị vào vị trí tương ứng còn lại là điền không

$$\begin{bmatrix}4 & 1\\ 2 & 3\end{bmatrix} \underbrace{\Longrightarrow}_{\text{Làm khó}} \begin{bmatrix}4 & 0 & 1 & 0\\ 0 & 0 & 0 & 0\\ 2 & 0 & 3 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

Đối với lớp tăng mẫu khi dùng hàm cực đại, dùng với **mạng đối xứng** thì ta có thể ghi nhớ vị trí đạt giá trị lớn nhất với tầng đối xứng (max pooling layer <-> max unpooling layer) còn các vị trí khác sẽ được điền giá trị không

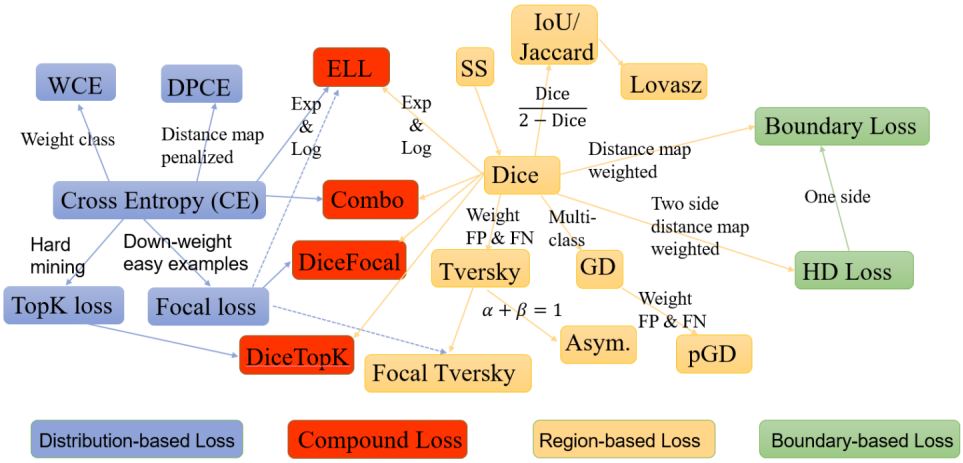
$$\text{Đầu vào} \Rightarrow \begin{bmatrix}1 & 5 & 4 & 8\\ 1 & 6 & 2 & 7\\ 3 & 2 & 0 & 4\\ 1 & 2 & 1 & 3\end{bmatrix} \underbrace{\Longrightarrow}_{\text{Giảm mẫu cực đại}} \begin{bmatrix}6 & 8\\ 3 & 4\end{bmatrix} \cdots\cdots \begin{bmatrix}4 & 1\\ 2 & 3\end{bmatrix} \underbrace{\Longrightarrow}_{\text{Tăng mẫu cực đại}} \begin{bmatrix}0 & 0 & 0 & 1\\ 0 & 4 & 0 & 0\\ 2 & 0 & 0 & 3\\ 0 & 0 & 0 & 0\end{bmatrix} \Rightarrow \text{Đầu ra}$$

Dạng đồ thị đối xứng thường gọi là đồ thị chữ V, có hai nhánh trái mã hoá ảnh còn nhánh phải giải mã ảnh. Được dùng trong cả hai trường hợp học có giám sát ví dụ bài toán phân đoạn ảnh, hoặc không giám sát bài toán tái tạo ảnh. Nhìn chung, nó được dùng nhiều trong lĩnh vực thị giác máy tính.

2.4 Các hàm mất mát

Cũng giống như bài toán phát hiện đối tượng nhưng mức độ cao hơn, hai bài toán toàn hồi quy vùng và phân loại từng điểm ảnh dẫn đến số lượng các hàm mất mát của bài toán phân đoạn ảnh tăng lên đột biến. Ta phân chúng thành các dạng hàm mất mát

- Hàm mất mát dạng phân bố như CrossEntropy, FocalLoss, TopK loss....
- Hàm mất mát dạng kết hợp như Combo loss, DiceFocal, DiceTopK....
- Hàm mất mát dạng vùng như IoU/Jaccard, Lovasz, SS, Dice, ...
- Hàm mất mát dạng biên như Boundary Loss, HD Loss



Hình 4: Các hàm mất mát thường dùng trong bài toán phân đoạn ảnh.

Dưới đây ta liệt kê công thức của một số hàm mất mát quan trọng

- Cross Entropy (CE) là hàm phân loại phổ biến

$$CE(p, \hat{p}) = -(p \log(\hat{p})) + (1 - p) \log(1 - \hat{p})$$

- Weighted Cross Entropy (WCE) là hàm phân loại đánh trọng số vào lớp riêng biệt

$$WCE(p, \hat{p}) = -(\beta p \log(\hat{p})) + (1 - p) \log(1 - \hat{p})$$

  Như vậy xác suất của lớp p tương ứng sẽ được gán trọng số β - siêu tham số cho phép tăng lên, nhẹ đi độ lỗi tương ứng.

- Focal loss (FL) là hàm phân loại thường đánh trọng số thấp vào nền (mẫu dễ) và nổi bật vào các đối tượng cần phát hiện (mẫu khó)

$$FL(p, \hat{p}) = -(\alpha(1 - \hat{p})^{\gamma} p \log(\hat{p})) + (1 - \alpha)\hat{p}^{\gamma}(1 - p) \log(1 - \hat{p})$$

- Dice coeffience là hàm mất mát dựa theo vùng

$$DC = \frac{2TP}{2TP + FP + FN} = \frac{2|X \cap Y|}{|X| + |Y|}$$

trong đó X, Y là các vùng điểm ảnh dự đoán và gán nhãn

- IoU cũng là hàm mất mát dựa theo vùng

$$IoU = \frac{TP}{TP + FP + FN} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

- Dice loss là hàm mất mát dựa theo vùng

$$DL(p, \hat{p}) = 1 - \frac{2p\hat{p} + 1}{p + \hat{p} + 1}$$

- Tversky loss là hàm mất mát dựa theo vùng

$$TI(p, \hat{p}) = \frac{p\hat{p}}{p\hat{p} + \beta(1 - p)\hat{p} + (1 - \beta)p(1 - \hat{p})}$$

- Dice Focal loss là hàm mất mát dạng kết hợp

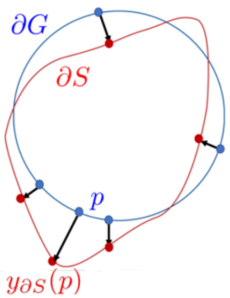
$$DC(p, \hat{p}) + FL(p, \hat{p})$$

- Dice CE loss là hàm mất mát dạng kết hợp

$$CE(p, \hat{p}) + DC(p, \hat{p})$$

- Boundary loss là hàm mất mát dạng biên

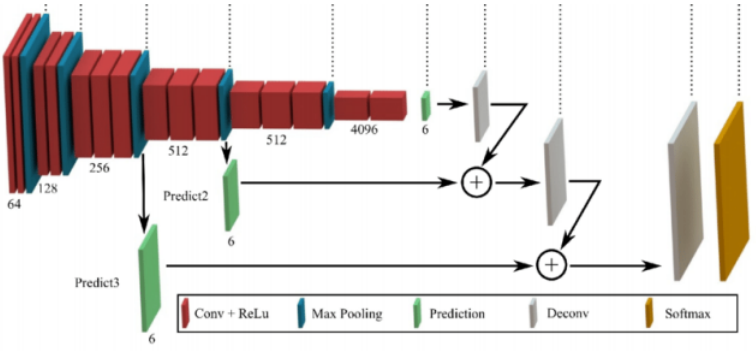
$$Dist(\partial G, \partial S) = \int_{\partial G} ||y_{\partial S}(p) - p|| dp$$



3 Các mạng nơ ron phân đoạn

3.1 Fully Convolutional Network - FCN

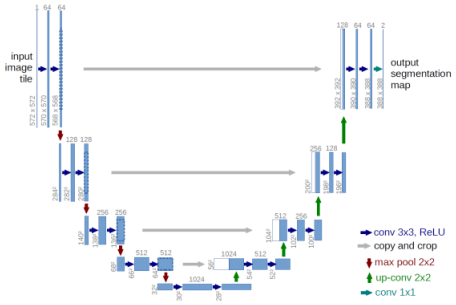
Ý tưởng chính của mạng FCN là sẽ thực hiện tăng mẫu và nhập ở phần cuối của mạng phân loại thông thường. Với mạng phân loại thông thường, các tầng FC cuối cùng sẽ bị loại bỏ và thêm lớp tăng mẫu (Deconvolution) kết hợp phép cộng phần tử (element-wise addition) để giảm thiểu thông tin về không gian bị nén sau các tầng giảm mẫu trước đó. Như vậy với kết quả các tầng giảm mẫu trước đó, ta luôn thực hiện tăng mẫu lại bằng kích thước ảnh đầu vào, cộng phần tử của các lớp tăng mẫu vào để trộn chúng thành đầu ra sự đoán, mục đích là "lấp đầy" khoảng trống do các tầng giảm mẫu gây ra.



Hình 5: Mạng nơ ron tích chập dùng phân đoạn ảnh sẽ kết hợp tăng mẫu và cộng các lớp tăng mẫu. Tổng hợp lớp tăng mẫu sẽ dùng hàm softmax để xác định đầu ra sự đoán. Từ trái sang phải, các kết nối tắt FCN-8s, FCN-16s và FCN-32 có độ phân giai khác nhau cần được kết hợp thành đầu ra mong muốn.

3.2 U-Net

Được dùng trong lĩnh vực y tế, ảnh chụp X-Ray, trong đó kiến trúc mạng dạng chữ U thêm các kết nối tắt giữa hai bên nhánh trái-phải của mạng tạo nên hiệu quả cũng như các cải tiến trong dự đoán. Mạng hoàn toàn đối xứng và có chi phí khá cao.

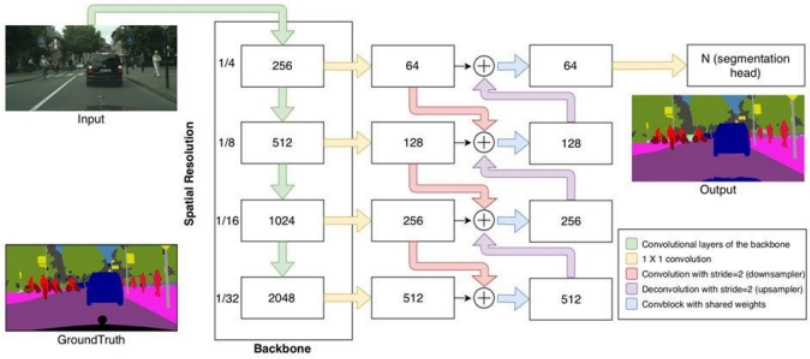


Hình 6: Mạng nơ ron tích chập dùng phân đoạn ảnh y tế, thường là ảnh đen trắng. Kiến trúc hoàn toàn đối xứng, đầu ra gồm hai lớp, có các kết nối tắt giữa mọi tầng để giảm thiểu lỗi khi tăng mẫu trong quá trình phục hồi lại ảnh ban đầu.

Có thêm các Stacked UNets và CUNets có được do kết hợp nhiều UNet với nhau.

3.3 RGPNet

Mạng RGPNet xây dựng các lớp mạng con có độ phân giải khác nhau do các tầng giảm mẫu của ảnh ban đầu sau đó dự đoán, tăng mẫu và tiếp tục thực hiện phân loại hồi quy giống với mạng U-Net ở trên.



Hình 7: Mạng nơ ron tích chập phân đoạn dùng nhiều độ phân giải dự đoán khác nhau, sau đó kết hợp để dự đoán ảnh phân đoạn cuối cùng.

4 Tổng kết

Về cơ bản cũng giống như bài toán phát hiện đối tượng, thay vì dùng phương pháp vét cạn với cửa số trượt, các mô hình phân đoạn hướng đến thiết kế một mạng nơ ron tích chập duy nhất (một giai đoạn) dùng các khung sườn - backborn - dùng để trích chọn đặc trưng của các phân đoạn hình ảnh đồng thời giải hai bài toán phân loại và hồi quy. Nhìn chung, do khi thực hiện trích chọn đặc trưng ta có bị mất đặc tính không gian - spatial features - nên ta phải dự đoán lại thông qua nhiều độ phân giải khác nhau từ ảnh gốc và "điền" lại thông tin các phân đoạn (vùng ảnh) dự đoán sao cho khớp bằng với ảnh ban đầu. Tuy nhiên, một trong nhược điểm của các mô hình này là các vùng bị loang vào nhau và thường phải hậu xử lý.