SegNet: Một sự kết hợp sâu sắc Kiến trúc bộ mã hóa-giải mã cho hình ảnh phân đoạn

Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, Thành viên cấp cao, IEEE,

Tóm tắt—Chúng tôi trình bày một kiến trúc mạng thần kinh tích chập hoàn toàn mới lạ và thiết thực cho phân đoạn theo pixel ngữ nghĩa đư ợc gọi là SegNet. Công cụ phân đoạn có thể đào tạo cốt lõi này bao gồm một mạng bộ mã hóa, một mạng bộ giải mã tư ơ ng ứng, theo sau là lớp phân loại theo pixel. Kiến trúc của mạng bộ mã hóa giống hệt về mặt cấu trúc với 13 lớp tích chập trong mạng VGG16 [1]. Vai trò của mạng bộ giải mã là ánh xạ các bản đồ tính năng của bộ mã hóa có độ phân giải thấp thành các bản đồ tính năng có độ phân giải đầu vào đầy đủ để phân loại theo pixel. Tính mới của SegNet nằm ở cách bộ giải mã lấy mẫu (các) bản đồ tính năng đầu vào có độ phân giải thấp hơ n của nó. Cụ thể, bộ giải mã sử dụng các chỉ số tổng hợp đư ợc tính trong bư ớc tổng hợp tối đa của bộ mã hóa tư ơ ng ứng để thực hiện lấy mẫu phi tuyến tính. Điều này giúp loại bỏ nhu cầu học cách upsample. Các bản đồ đư ợc lấy mẫu lại thư a thớt và sau đó đư ợc tích hợp với các bộ lọc có thể huấn luyện để tạo ra các bản đồ đặc trư ng dày đặc. Chúng tôi so sánh kiến trúc đề xuất của mình với FCN [2] đư ợc áp dụng rộng rãi và cả với kiến trúc DeepLab-LargeFOV [3], DeconvNet [4] nổi tiếng. So sánh này cho thấy sự đánh đổi giữa bộ nhớ và độ chính xác liên quan đến việc đạt đư ợc hiệu suất phân đoạn tốt.

SegNet chủ yếu đư ợc thúc đẩy bởi các ứng dụng hiểu cảnh. Do đó, nó đư ợc thiết kế để hoạt động hiệu quả cả về bộ nhớ và thời gian tính toán trong quá trình suy luận. Nó cũng nhỏ hơ n đáng kể về số lư ợng tham số có thể đào tạo so với các kiến trúc cạnh tranh khác và có thể đư ợc đào tạo từ đầu đến cuối bằng cách sử dụng giảm dần độ dốc ngẫu nhiên. Chúng tôi cũng đã thực hiện một điểm chuẩn có kiểm soát của SegNet và các kiến trúc khác trên cả cảnh trên đư ờng và các tác vụ phân đoạn cảnh trong nhà SUN RGB-D. Những đánh giá định lư ợng này cho thấy SegNet cung cấp hiệu suất tốt với thời gian suy luận cạnh tranh và bộ nhớ suy luận hiệu quả nhất so với các kiến trúc khác. Chúng tôi cũng cung cấp triển khai Caffe cho SegNet và bản trình diễn trên web tại http://mi.eng.cam.ac.uk/projects/segnet/.

Thuật ngữ chỉ mục —Mạng nơ -ron tích chập sâu, Phân đoạn theo pixel theo ngữ nghĩa, Cảnh trong nhà, Cảnh trên đư ờng, Bộ mã hóa, Bộ giải mã, Tổng hợp, Lấy mẫu.

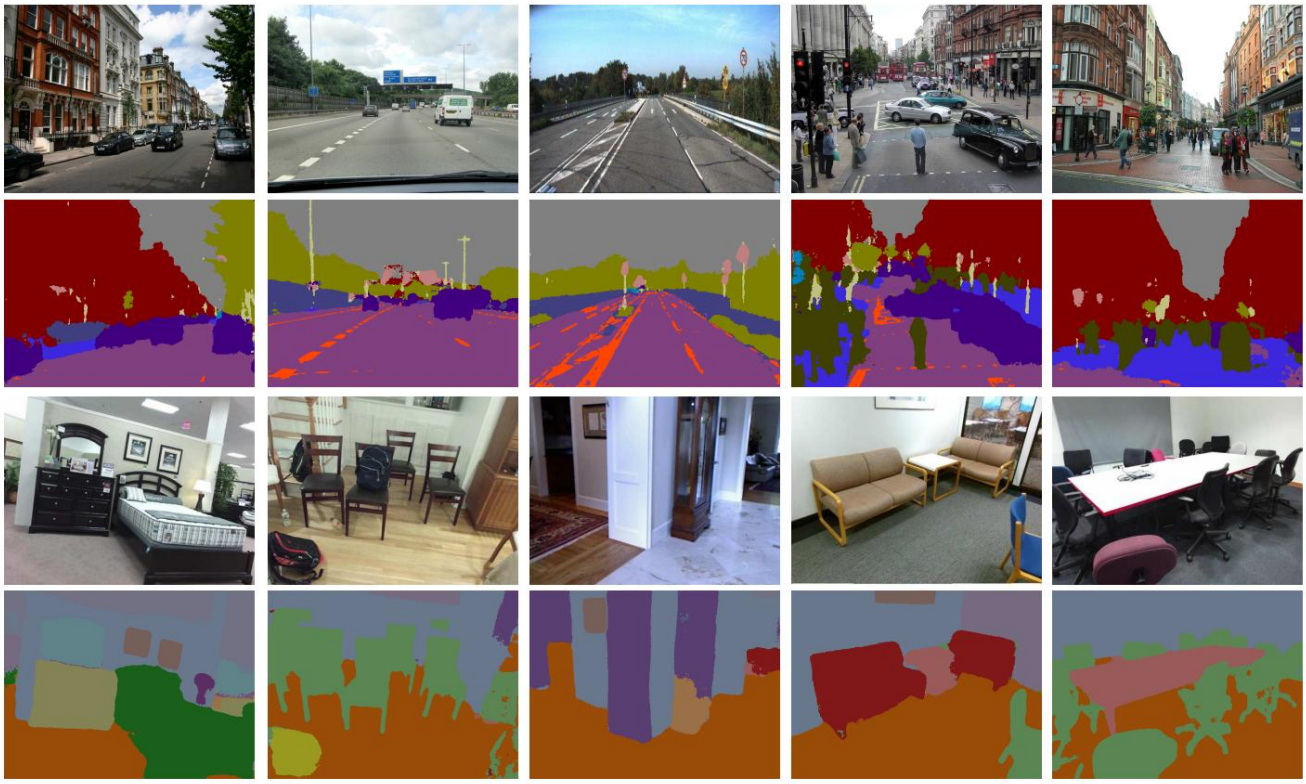
1 GIỚI THIỆU

Phân đoạn ngữ nghĩa có rất nhiều ứng dụng khác nhau, từ hiểu cảnh, suy ra mối quan hệ hỗ trợ giữa các đối tư ợng đến lái xe tự động. Các phư ơ ng pháp ban đầu dựa trên tín hiệu tầm nhìn ở mức độ thấp đã nhanh chóng bị thay thế bởi các thuật toán học máy phổ biến. Đặc biệt, học sâu gần đây đã đạt đư ợc thành công lớn trong nhận dạng chữ viết tay, lời nói, phân loại toàn bộ ảnh và phát hiện các đối tư ợng trong ảnh [5], [6]. Giờ đây, có một mối quan tâm tích cực đối với việc ghi nhãn ngữ nghĩa theo pixel [7] [8], [9], [2], [4], [10], [11], [12], [13], [3] , [14], [15], [16]. Tuy nhiên, một số cách tiếp cận gần đây đã cố gắng áp dụng trực tiếp các kiến trúc sâu đư ợc thiết kế để dự đoán danh mục thành ghi nhãn thông minh theo pixel [7]. Các kết quả, mặc dù rất đáng khích lệ, như ng có vẻ thô [3].
Điều này chủ yếu là do tổng hợp tối đa và lấy mẫu phụ làm giảm độ phân giải của bản đồ đối tư ợng địa lý. Động lực thiết kế SegNet của chúng tôi xuất phát từ nhu cầu ánh xạ các tính năng có độ phân giải thấp thành độ phân giải đầu vào để phân loại pixel-khôn ngoan. Á nh xạ này phải tạo ra các tính năng hữu ích cho việc định vị ranh giới chính xác.

Kiến trúc của chúng tôi, SegNet, đư ợc thiết kế để trở thành một kiến trúc hiệu quả cho việc phân đoạn ngữ nghĩa theo pixel. Nó chủ yếu đư ợc thúc đẩy bởi các ứng dụng hiểu cảnh đư ờng đòi hỏi khả năng mô hình hóa diện mạo (đư ờng, tòa nhà), hình dạng (ô tô, ngư ời đi bộ) và hiểu mối quan hệ không gian (bối cảnh) giữa các lớp khác nhau như đư ờng bộ và lối đi bộ. Trong các cảnh đư ờng điển hình, phần lớn các pixel thuộc về các lớp lớn như đư ờng, tòa nhà và do đó mạng phải tạo ra các phân đoạn mư ợt mà. Động cơ cũng phải có khả năng phân định đối tư ợng dựa trên hình dạng của chúng mặc dù kích thư ớc nhỏ. Do đó, điều quan trọng là giữ lại thông tin ranh giới trong biểu diễn hình ảnh đư ợc trích xuất. Từ góc độ tính toán, mạng cần phải hiệu quả về cả bộ nhớ và thời gian tính toán trong quá trình suy luận. Khả năng đào tạo end-to-end để cùng tối ư u hóa tất cả các trọng số trong mạng bằng cách sử dụng kỹ thuật cập nhật trọng số hiệu quả như giảm độ dốc ngẫu nhiên (SGD) [17] là một lợi ích bổ sung vì nó dễ dàng lặp lại hơ n. Thiết kế của SegNet nảy sinh từ nhu cầu phù hợp với các tiêu chí này.

Mạng bộ mã hóa trong SegNet giống hệt về mặt cấu trúc với các lớp tích chập trong VGG16 [1]. Chúng tôi loại bỏ các lớp VGG16 đư ợc kết nối đầy đủ để làm cho mạng bộ mã hóa SegNet nhỏ hơ n đáng kể và dễ đào tạo hơ n nhiều kiến trúc gần đây khác [2], [4], [11], [18]. Thành phần chính của SegNet là mạng bộ giải mã bao gồm một hệ thống phân cấp các bộ giải mã, một bộ giải mã tư ơ ng ứng với mỗi bộ mã hóa. Trong số này, các bộ giải mã thích hợp sử dụng các chỉ số tổng hợp tối đa nhận đư ợc từ bộ mã hóa tư ơ ng ứng để thực hiện lấy mẫu phi tuyến tính của các bản đồ đặc trư ng đầu vào của chúng. Ý tư ởng này đư ợc lấy cảm hứng từ một kiến trúc đư ợc thiết kế cho tính năng học tập không giám sát [19]. Sử dụng lại các chỉ số tổng hợp tối đa trong quá trình giải mã có một số thực tế

• V. Badrinarayanan, A. Kendall, R. Cipolla làm việc tại Phòng thí nghiệm Trí tuệ Máy móc, Khoa Kỹ thuật, Đại học Cambridge, Vư ơ ng quốc Anh. E-mail: vb292,agk34,cipolla@eng.cam.ac.uk



Hình 1. Dự đoán SegNet về cảnh trên đư ờng và cảnh trong nhà. Để tự mình thử hệ thống của chúng tôi, vui lòng xem bản trình diễn web trực tuyến của chúng tôi tại http:// mi.eng.cam.ac. Anh/dự án/segnet/.

thuận lợi; (i) nó cải thiện việc phân định ranh giới, (ii) nó giảm số lư ợng tham số cho phép đào tạo từ đầu đến cuối và (iii) hình thức lấy mẫu ngư ợc này có thể đư ợc tích hợp vào bất kỳ kiến trúc bộ mã hóa-giải mã nào, chẳng hạn như [2], [10] chỉ với một chút sửa đổi.

Một trong những đóng góp chính của bài báo này là phân tích của chúng tôi về kỹ thuật giải mã SegNet và Mạng tích chập hoàn toàn (FCN) đư ợc sử dụng rộng rãi [2]. Điều này là để truyền đạt những đánh đổi thực tế liên quan đến việc thiết kế kiến trúc phân khúc. Hầu hết các kiến trúc sâu gần đây cho phân đoạn đều có các mạng bộ mã hóa giống hệt nhau, tức là VGG16, như ng khác nhau ở dạng mạng bộ giải mã, đào tạo và suy luận. Một đặc điểm chung khác là chúng có các tham số có thể huấn luyện theo thứ tự hàng trăm triệu và do đó gặp khó khăn trong việc thực hiện huấn luyện từ đầu đến cuối [4]. Khó khăn trong việc đào tạo các mạng này đã dẫn đến đào tạo nhiều giai đoạn [2], nối các mạng vào một kiến trúc đư ợc đào tạo trư ớc như FCN [10], sử dụng các công cụ hỗ trợ như đề xuất khu vực để suy luận [4], đào tạo phân loại rời rạc và mạng phân đoạn [18] và sử dụng dữ liệu đào tạo bổ sung để đào tạo trư ớc [11] [20] hoặc để đào tạo đầy đủ [10]. Ngoài ra, các kỹ thuật xử lý hậu kỳ nâng cao hiệu suất [3] cũng rất phổ biến. Mặc dù tất cả các yếu tố này đều cải thiện hiệu suất trên các điểm chuẩn đầy thách thức [21], như ng thật không may là rất khó từ kết quả định lư ợng của chúng để tháo gỡ các yếu tố thiết kế chính cần thiết để đạt đư ợc hiệu suất tốt. Do đó, chúng tôi đã phân tích quá trình giải mã đư ợc sử dụng trong một số phư ơ ng pháp này [2], [4] và chỉ ra những ư u và như ợc điểm của chúng.

Chúng tôi đánh giá hiệu suất của SegNet trên hai tác vụ phân đoạn cảnh, phân đoạn cảnh trên đư ờng CamVid [22] và phân đoạn cảnh trong nhà SUN RGB-D [23]. Pascal VOC12 [21] đã trở thành thách thức chuẩn cho việc phân đoạn trong nhiều năm.

Tuy nhiên, phần lớn nhiệm vụ này có một hoặc hai tiền cảnh các lớp đư ợc bao quanh bởi một nền rất đa dạng. Điều này hoàn toàn ủng hộ các kỹ thuật đư ợc sử dụng để phát hiện như đư ợc thể hiện trong công việc gần đây trên mạng phân đoạn phân loại tách rời [18] trong đó mạng phân loại có thể đư ợc đào tạo với một tập hợp lớn dữ liệu đư ợc dán nhãn yếu và hiệu suất của mạng phân đoạn độc lập đư ợc cải thiện. Phư ơ ng pháp của [3] cũng sử dụng các bản đồ đặc trư ng của mạng phân loại với kỹ thuật xử lý bài CRF độc lập để thực hiện phân đoạn. Hiệu suất cũng có thể đư ợc tăng cư ờng bằng cách sử dụng các công cụ hỗ trợ suy luận bổ sung như đề xuất khu vực [4], [24]. Do đó, nó khác với việc hiểu cảnh trong đó ý tư ởng là khai thác sự xuất hiện đồng thời của các đối tư ợng và bối cảnh không gian khác để thực hiện phân đoạn mạnh mẽ.

Để chứng minh tính hiệu quả của SegNet, chúng tôi trình bày bản trình diễn trực tuyến thời gian thực về phân đoạn cảnh đư ờng thành 11 loại quan tâm cho lái xe tự động (xem liên kết trong Hình 1). Một số kết quả thử nghiệm ví dụ đư ợc tạo ra trên hình ảnh cảnh đư ờng đư ợc lấy mẫu ngẫu nhiên từ Google và cảnh thử nghiệm trong nhà từ bộ dữ liệu SUN RGB-D [23] đư ợc hiển thị trong Hình 1.

Phần còn lại của bài báo đư ợc tổ chức như sau. Trong giây 2 chúng tôi xem xét các tài liệu gần đây có liên quan. Chúng tôi mô tả kiến trúc SegNet và phân tích của nó trong Sec. 3. Trong giây. 4, chúng tôi đánh giá hiệu suất của SegNet trên bộ dữ liệu cảnh ngoài trời và trong nhà. Tiếp theo là một cuộc thảo luận chung về cách tiếp cận của chúng tôi với các gợi ý cho công việc trong tư ơ ng lai trong Sec. 5. Chúng tôi kết luận trong Sec. 6.

2 ĐÁ NH GIÁ TÀI LIỆU Phân đoạn theo

pixel theo ngữ nghĩa là một chủ đề nghiên cứu tích cực, đư ợc thúc đẩy bởi các bộ dữ liệu đầy thách thức [21], [22], [23], [25], [26]. Trư ớc khi các mạng sâu xuất hiện, các phư ơ ng pháp hoạt động tốt nhất chủ yếu dựa vào các tính năng đư ợc thiết kế thủ công để phân loại các pixel một cách độc lập. Thông thư ờng, một bản vá đư ợc đư a vào một bộ phân loại, ví dụ: Ngẫu nhiên

Forest [27], [28] hoặc Boosting [29], [30] để dự đoán xác suất lớp của pixel trung tâm. Các tính năng dựa trên diện mạo [27] hoặc SfM và diện mạo [28], [29], [30] đã được khám phá cho bài kiểm tra hiểu biết về cảnh đường CamVid [22]. Các dự đoán nhiễu trên mỗi pixel này (thường được gọi là các thuật ngữ đơn nguyên) từ các bộ phân loại sau đó được làm mịn bằng cách sử dụng CRF theo cặp hoặc bậc cao hơn [29], [30] để cải thiện độ chính xác. Các cách tiếp cận gần đây hơn nhằm mục đích tạo ra các đơn vị chất lượng cao bằng cách cố gắng dự đoán nhãn cho tất cả các pixel trong một bản vá thay vì chỉ pixel trung tâm. Điều này cải thiện kết quả của các đơn vị dựa trên Rừng ngẫu nhiên [31] nhưng các lớp có cấu trúc mỏng được phân loại kém. Các bản đồ độ sâu dày đặc được tính toán từ video CamVid cũng đã được sử dụng làm đầu vào để phân loại bằng cách sử dụng Rừng ngẫu nhiên [32]. Một cách tiếp cận khác ủng hộ việc sử dụng kết hợp các tính năng được thiết kế thủ công phổ biến và siêu pixel hóa theo không gian-thời gian để đạt được độ chính xác cao hơn [33]. Kỹ thuật hoạt động tốt nhất trong bài kiểm tra CamVid [30] giải quyết sự mất cân bằng giữa các tần số nhãn bằng cách kết hợp các đầu ra phát hiện đối tượng với các dự đoán của bộ phân loại trong khung CRF. Kết quả của tất cả các kỹ thuật này cho thấy sự cần thiết phải cải thiện các tính năng để phân loại.

Phân đoạn ngữ nghĩa theo pixel RGBD trong nhà cũng đã trở nên phổ biến kể từ khi phát hành bộ dữ liệu NYU [25]. Bộ dữ liệu này cho thấy tính hữu ích của kênh độ sâu để cải thiện phân đoạn. Cách tiếp cận của họ đã sử dụng các tính năng như RGB-SIFT, depth-SIFT và vị trí pixel làm đầu vào cho bộ phân loại mạng thần kinh để dự đoán các đơn vị pixel. Các đơn vị ồn ào sau đó được làm mịn bằng CRF. Các cải tiến đã được thực hiện bằng cách sử dụng bộ tính năng phong phú hơn bao gồm LBP và phân đoạn vùng để đạt được độ chính xác cao hơn [34] sau đó là CRF. Trong công trình gần đây hơn [25], cả mối quan hệ phân đoạn lớp và hỗ trợ đều được suy ra cùng nhau bằng cách sử dụng kết hợp RGB và tín hiệu dựa trên độ sâu. Một cách tiếp cận khác tập trung vào tái cấu trúc khớp thời gian thực và phân đoạn ngữ nghĩa, trong đó Rừng ngẫu nhiên được sử dụng làm bộ phân loại [35]. Gupta và cộng sự. [36] sử dụng phát hiện ranh giới và nhóm theo thứ bậc trước khi thực hiện phân đoạn danh mục. Thuộc tính chung trong tất cả các phương pháp này là sử dụng các tính năng được thiết kế thủ công để phân loại hình ảnh RGB hoặc RGBD.

Sự thành công của các mạng thần kinh tích chập sâu để phân loại đối tượng gần đây đã khiến các nhà nghiên cứu khai thác khả năng học tập tính năng của chúng cho các vấn đề dự đoán có cấu trúc như phân đoạn. Cũng đã có những nỗ lực áp dụng các mạng được thiết kế để phân loại đối tượng vào phân đoạn, đặc biệt bằng cách sao chép các tính năng của lớp sâu nhất trong các khối để phù hợp với kích thước hình ảnh [7], [37], [38], [39]. Tuy nhiên, kết quả phân loại là khối [38]. Một cách tiếp cận khác sử dụng mạng thần kinh tái phát [40] hợp nhất một số dự đoán có độ phân giải thấp để tạo dự đoán độ phân giải hình ảnh đầu vào. Những kỹ thuật này đã là một cải tiến so với các tính năng được thiết kế thủ công [7] nhưng khả năng phân định ranh giới của chúng còn kém.

Các kiến trúc sâu mới hơn [2], [4], [10], [13], [18] được thiết kế đặc biệt để phân đoạn đã nâng cao trình độ tiên tiến bằng cách học cách giải mã hoặc ánh xạ các biểu diễn hình ảnh có độ phân giải thấp thành pixel-khôn ngoan phỏng đoán. Mạng bộ mã hóa tạo ra các biểu diễn có độ phân giải thấp này trong tất cả các kiến trúc này là mạng phân loại VGG16 [1] có 13 lớp tích chập và 3 lớp được kết nối đầy đủ. Trọng số mạng bộ mã hóa này thường được đào tạo trước trên bộ dữ liệu phân loại đối tượng ImageNet lớn [41]. Mạng bộ giải mã khác nhau giữa các kiến trúc này và là phần chịu trách nhiệm tạo ra các tính năng đa chiều cho từng pixel để phân loại.

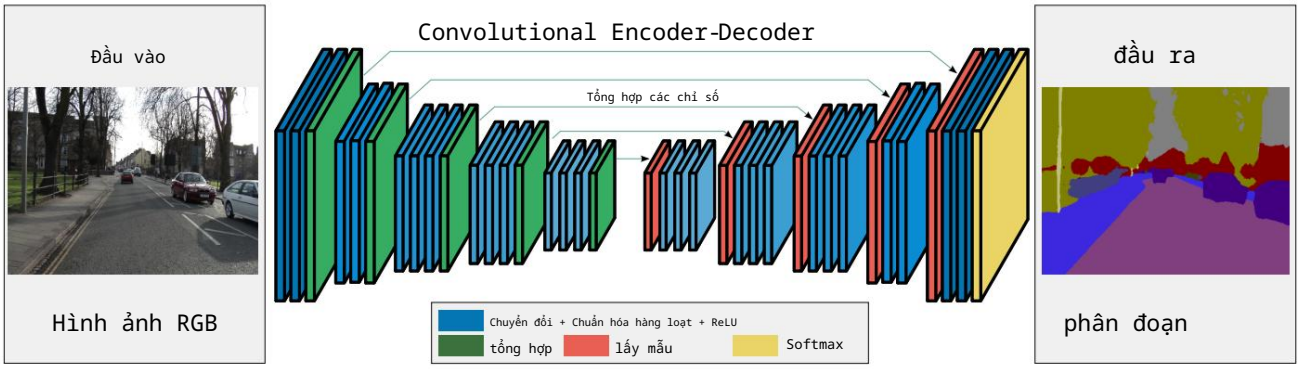
Mỗi bộ giải mã trong Mạng kết hợp hoàn toàn (FCN) kiến trúc [2] học cách lấy mẫu (các) bản đồ tính năng đầu vào của nó và kết hợp chúng với bản đồ tính năng bộ mã hóa tương ứng để tạo đầu vào cho bộ giải mã tiếp theo. Đó là một kiến trúc có số lượng lớn các tham số có thể đào tạo trong mạng bộ mã hóa (134M) nhưng mạng bộ giải mã rất nhỏ (0,5M). Kích thước tổng thể lớn của mạng này khiến việc đào tạo từ đầu đến cuối về một nhiệm vụ có liên quan trở nên khó khăn. Do đó, các tác giả sử dụng quy trình đào tạo theo từng giai đoạn. Ở đây, mỗi bộ giải mã trong mạng bộ giải mã được thêm dần vào mạng được đào tạo hiện có. Mạng được phát triển cho đến khi không quan sát thấy hiệu suất tăng thêm nữa. Sự tăng trưởng này bị dừng lại sau ba bộ giải mã, do đó, việc bỏ qua các bản đồ đặc trưng có độ phân giải cao chắc chắn có thể dẫn đến mất thông tin biên [4]. Ngoài các vấn đề liên quan đến đào tạo, nhu cầu sử dụng lại các bản đồ tính năng của bộ mã hóa trong bộ giải mã khiến nó tốn nhiều bộ nhớ trong thời gian thử nghiệm. Chúng tôi nghiên cứu mạng này chi tiết hơn vì nó là cốt lõi của các kiến trúc khác gần đây [10], [11].

Hiệu suất dự đoán của FCN đã được cải thiện hơn nữa bằng cách nối thêm FCN với mạng thần kinh hồi quy (RNN) [10] và tinh chỉnh chúng trên các bộ dữ liệu lớn [21], [42]. Các lớp RNN bắt chước khả năng phân định ranh giới sắc nét của CRF trong khi khai thác sức mạnh biểu diễn tính năng của FCN. Chúng cho thấy sự cải thiện đáng kể so với FCN-8 nhưng cũng cho thấy sự khác biệt này giảm đi khi sử dụng nhiều dữ liệu huấn luyện hơn để huấn luyện FCN-8. Ưu điểm chính của CRF-RNN được bộc lộ khi nó được đào tạo chung với một kiến trúc như FCN 8. Thực tế là việc đào tạo chung giúp ích cũng được thể hiện trong các kết quả gần đây khác [43], [44]. Thật thú vị, mạng giải chập [4] hoạt động tốt hơn đáng kể so với FCN mặc dù phải trả giá bằng việc đào tạo và suy luận phức tạp hơn. Tuy nhiên, điều này đặt ra câu hỏi liệu lợi thế được nhận thức của CRF-RNN có bị giảm đi khi công cụ phân đoạn chuyển tiếp nguồn cấp dữ liệu cốt lõi được thực hiện tốt hơn hay không. Trong mọi trường hợp, mạng CRF-RNN có thể được thêm vào bất kỳ kiến trúc phân đoạn sâu nào, bao gồm cả SegNet.

Kiến trúc sâu đa quy mô cũng đang được theo đuổi [13], [44]. Chúng có hai loại, (i) những loại sử dụng hình ảnh đầu vào ở một vài tỷ lệ và mạng trích xuất tính năng sâu tương ứng hoạt động, và (ii) những loại kết hợp các bản đồ đặc trưng từ các lớp khác nhau của một kiến trúc sâu duy nhất [45] [11]. Ý tưởng chung là sử dụng các tính năng được trích xuất ở nhiều tỷ lệ để cung cấp cả bối cảnh cục bộ và toàn cầu [46] và việc sử dụng bản đồ tính năng của các lớp mã hóa ban đầu giữ lại nhiều chi tiết tần số cao hơn dẫn đến ranh giới lớp sắc nét hơn. Một số kiến trúc này rất khó đào tạo do kích thước tham số của chúng [13]. Do đó, một quy trình đào tạo nhiều giai đoạn được sử dụng cùng với việc tăng cường dữ liệu. Suy luận cũng tốn kém với nhiều con đường tích chập để trích xuất tính năng. Những người khác [44] gắn CRF vào mạng đa quy mô của họ và cùng huấn luyện chúng. Tuy nhiên, đây không phải là chuyển tiếp nguồn cấp dữ liệu tại thời điểm thử nghiệm và yêu cầu tối ưu hóa để xác định nhãn MAP.

Một số kiến trúc sâu được đề xuất gần đây cho phân đoạn không được chuyển tiếp trong thời gian suy luận [4], [3], [18]. Chúng yêu cầu suy luận MAP qua CRF [44], [43] hoặc hỗ trợ như đề xuất khu vực [4] để suy luận. Chúng tôi tin rằng hiệu suất nhận được tăng lên khi sử dụng CRF là do thiếu các kỹ thuật giải mã tốt trong công cụ phân đoạn chuyển tiếp nguồn cấp dữ liệu cốt lõi của họ. Mặt khác, SegNet sử dụng bộ giải mã để có được các tính năng để phân loại chính xác theo pixel.

Mạng Deconvolutional được đề xuất gần đây [4] và biến thể bán giám sát của nó là Mạng tách rời [18] sử dụng các vị trí tối đa của bản đồ tính năng bộ mã hóa (chỉ số tổng hợp) để thực hiện lấy mẫu phi tuyến tính trong mạng bộ giải mã. Các tác giả của những kiến trúc này, độc lập với SegNet (lần đầu tiên được gửi tới



Hình 2. Minh họa về kiến trúc SegNet. Không có lớp nào được kết nối đầy đủ và do đó nó chỉ là tích chập. Bộ giải mã lấy mẫu đầu vào của nó bằng cách sử dụng các chỉ số nhóm được truyền từ bộ mã hóa của nó để tạo ra (các) bản đồ tính năng thưa thớt. Sau đó, nó thực hiện tích chập với ngân hàng bộ lọc có thể huấn luyện để làm dày bản đồ đặc trưng. Bản đồ tính năng đầu ra của bộ giải mã cuối cùng được đưa đến bộ phân loại soft-max để phân loại theo pixel.

CVPR 2015 [12]), đã đề xuất ý tưởng giải mã này trong mạng bộ giải mã. Tuy nhiên, mạng bộ mã hóa của họ bao gồm các lớp được kết nối đầy đủ từ mạng VGG-16 bao gồm khoảng 90% tham số của toàn bộ mạng của họ. Điều này làm cho việc đào tạo mạng của họ trở nên rất khó khăn và do đó cần có các hỗ trợ bổ sung như sử dụng các đề xuất khu vực để cho phép đào tạo. Hơn nữa, trong quá trình suy luận, các đề xuất này được sử dụng và điều này làm tăng đáng kể thời gian suy luận. Từ quan điểm đo điểm chuẩn, điều này cũng gây khó khăn cho việc đánh giá hiệu suất của kiến trúc (mạng bộ giải mã-mã hóa) mà không có các công cụ hỗ trợ khác. Trong công việc này, chúng tôi loại bỏ các lớp được kết nối đầy đủ của mạng bộ mã hóa VGG16, lớp này cho phép chúng tôi huấn luyện mạng bằng tập huấn luyện có liên quan bằng cách sử dụng tối ưu hóa SGD. Một phương pháp khác gần đây [3] cho thấy lợi ích của việc giảm đáng kể số lượng tham số mà không làm giảm hiệu suất, giảm tiêu thụ bộ nhớ và cải thiện thời gian suy luận.

Công việc của chúng tôi được lấy cảm hứng từ kiến trúc học tập tính năng không giám sát được đề xuất bởi Ranzato et al. [19]. Mô-đun học tập chính là mạng bộ mã hóa-giải mã. Một bộ mã hóa bao gồm tích chập với ngân hàng bộ lọc, phi tuyến tính tanh theo từng phần tử, tổng hợp tối đa và lấy mẫu phụ để thu được các bản đồ đặc trưng. Đối với mỗi mẫu, chỉ số của các vị trí tối đa được tính toán trong quá trình tổng hợp được lưu trữ và chuyển đến bộ giải mã. Bộ giải mã lấy mẫu các bản đồ đặc trưng bằng cách sử dụng các chỉ số gộp được lưu trữ. Nó kết hợp bản đồ được lấy mẫu lại này bằng cách sử dụng ngân hàng bộ lọc bộ giải mã có thể huấn luyện để tái tạo lại hình ảnh đầu vào. Kiến trúc này đã được sử dụng để đào tạo trước không giám sát để phân loại. Một kỹ thuật giải mã tương tự được sử dụng để trực quan hóa các mạng chập được huấn luyện [47] để phân loại. Kiến trúc của Ranzato et al. chủ yếu tập trung vào việc học tính năng theo lớp bằng cách sử dụng các bản vá đầu vào nhỏ. Điều này đã được mở rộng bởi Kavukcuoglu et. al. [48] để chấp nhận kích thước hình ảnh đầy đủ làm đầu vào để tìm hiểu bộ mã hóa phân cấp. Tuy nhiên, cả hai cách tiếp cận này đều không cố gắng sử dụng các mạng bộ giải mã-mã hóa sâu để đào tạo tính năng không giám sát vì chúng loại bỏ bộ giải mã sau mỗi lần đào tạo bộ mã hóa. Ở đây, SegNet khác với các kiến trúc này vì mạng bộ mã hóa-giải mã sâu được đào tạo chung cho nhiệm vụ học tập có giám sát và do đó, bộ giải mã là một phần không thể thiếu của mạng trong thời gian thử nghiệm.

Các ứng dụng khác trong đó dự đoán thông minh về pixel được thực hiện bằng cách sử dụng mạng sâu là hình ảnh siêu phân giải [49] và dự đoán bản đồ độ sâu từ một hình ảnh duy nhất [50]. Các tác giả trong [50] thảo luận về nhu cầu học cách lấy mẫu từ các bản đồ đặc trưng có độ phân giải thấp, đây là chủ đề chính của bài báo này.

3 KIẾN TRÚC

SegNet có mạng mã hóa và mạng giải mã tương ứng, theo sau là lớp phân loại theo pixel cuối cùng. Kiến trúc này được minh họa trong Hình 3. Mạng bộ mã hóa bao gồm 13 lớp chập tương ứng với 13 lớp chập đầu tiên trong mạng VGG16 [1] được thiết kế để phân loại đối tượng. Do đó, chúng ta có thể khởi tạo quá trình đào tạo từ các trọng số được đào tạo để phân loại trên các tập dữ liệu lớn [41]. Chúng tôi cũng có thể loại bỏ các lớp được kết nối đầy đủ để giữ lại các bản đồ tính năng có độ phân giải cao hơn ở đầu ra bộ mã hóa sâu nhất. Điều này cũng làm giảm đáng kể số lượng tham số trong mạng bộ mã hóa SegNet (từ 134M xuống 14,7M) so với các kiến trúc gần đây khác [2], [4] (xem Bảng 6). Mỗi lớp mã hóa có một lớp giải mã tương ứng và do đó mạng giải mã có 13 lớp. Đầu ra của bộ giải mã cuối cùng được đưa đến bộ phân loại soft-max nhiều lớp để tạo ra xác suất lớp cho từng pixel một cách độc lập.

Mỗi bộ mã hóa trong mạng bộ mã hóa thực hiện tích chập với ngân hàng bộ lọc để tạo ra một bộ bản đồ đặc trưng. Sau đó chúng được chuẩn hóa hàng loạt [51], [52]). Sau đó, một phi tuyến tính tuyến tính được chỉnh lưu theo từng phần tử (ReLU) max(0, x) được áp dụng. Sau đó, tính năng tổng hợp tối đa với cửa sổ 2 × 2 và sải chân 2 (cửa sổ không chồng lấp) được thực hiện và kết quả đầu ra được lấy mẫu phụ theo hệ số 2. Tính năng tổng hợp tối đa được sử dụng để đạt được tính bất biến dịch chuyển đối với các dịch chuyển không gian nhỏ trong ảnh đầu vào. Lấy mẫu phụ dẫn đến bối cảnh hình ảnh đầu vào lớn (cửa sổ không gian) cho mỗi pixel trong bản đồ tính năng. Mặc dù một số lớp tổng hợp tối đa và lấy mẫu phụ có thể đạt được nhiều bất biến dịch hơn để phân loại mạnh mẽ, tương ứng, có sự mất độ phân giải không gian của các bản đồ đặc trưng. Biểu diễn hình ảnh ngày càng bị mất (chi tiết ranh giới) không có lợi cho việc phân đoạn trong đó việc phân định ranh giới là rất quan trọng. Do đó, cần phải nắm bắt và lưu trữ thông tin ranh giới trong bản đồ đặc trưng của bộ mã hóa trước khi thực hiện lấy mẫu phụ. Nếu bộ nhớ trong quá trình suy luận không bị hạn chế, thì tất cả các bản đồ tính năng của bộ mã hóa (sau khi lấy mẫu phụ) có thể được lưu trữ. Điều này thường không xảy ra trong các ứng dụng thực tế và do đó chúng tôi đề xuất một cách hiệu quả hơn để lưu trữ thông tin này. Nó liên quan đến việc chỉ lưu trữ các chỉ số tổng hợp tối đa, nghĩa là các vị trí của giá trị tính năng tối đa trong mỗi cửa sổ tổng hợp được ghi nhớ cho mỗi bản đồ tính năng bộ mã hóa. Về nguyên tắc, điều này có thể được thực hiện bằng cách sử dụng 2 bit cho mỗi cửa sổ tổng hợp 2 × 2 và do đó lưu trữ hiệu quả hơn nhiều so với việc ghi nhớ (các) bản đồ tính năng ở độ chính xác nổi. Như chúng tôi trình bày

bộ nhớ lưu trữ thấp hơn này làm giảm độ chính xác một chút nhưng vẫn phù hợp với các ứng dụng thực tế.

Bộ giải mã thích hợp trong mạng bộ giải mã lấy mẫu bổ sung (các) bản đồ tính năng đầu vào của nó bằng cách sử dụng các chỉ số gộp tối đa đã ghi nhớ từ (các) bản đồ tính năng bộ mã hóa tương ứng. Bước này tạo ra (các) bản đồ tính năng thưa thớt. Kỹ thuật giải mã SegNet này được minh họa trong Hình 3. Các bản đồ đặc trưng này sau đó được kết hợp với ngân hàng bộ lọc bộ giải mã có thể đào tạo để tạo ra các bản đồ đặc trưng dày đặc.
Sau đó, một bước chuẩn hóa hàng loạt được áp dụng cho từng bản đồ này.
Lưu ý rằng bộ giải mã tương ứng với bộ mã hóa đầu tiên (gần với hình ảnh đầu vào nhất) tạo ra bản đồ tính năng đa kênh, mặc dù đầu vào bộ mã hóa của nó có 3 kênh (RGB). Điều này không giống như các bộ giải mã khác trong mạng tạo ra các bản đồ đặc trưng có cùng số lượng kích thước và kênh như đầu vào bộ mã hóa của chúng. Biểu diễn tính năng chiều cao ở đầu ra của bộ giải mã cuối cùng được đưa đến bộ phân loại soft-max có thể huấn luyện được.
Soft-max này
phân loại từng pixel một cách độc lập. Đầu ra của bộ phân loại soft-max là một hình ảnh kênh K về xác suất trong đó K là số lớp. Phân đoạn được dự đoán tương ứng với lớp có xác suất tối đa tại mỗi pixel.

Chúng tôi nói thêm ở đây rằng hai kiến trúc khác, DeconvNet [53] và U-Net [16] có chung kiến trúc với SegNet nhưng có một số khác biệt. DeconvNet có tham số hóa lớn hơn nhiều, cần nhiều tài nguyên tính toán hơn và khó đào tạo từ đầu đến cuối (Bảng 6), chủ yếu là do việc sử dụng các lớp được kết nối đầy đủ (mặc dù theo cách tích chập). Chúng tôi báo cáo một số so sánh với DeconvNet sau trong giấy Sec. 4.

So với SegNet, U-Net [16] (được đề xuất cho cộng đồng hình ảnh y tế) không sử dụng lại các chỉ số tổng hợp mà thay vào đó chuyển toàn bộ bản đồ đặc trưng (với chi phí nhiều bộ nhớ hơn) tới các bộ giải mã tương ứng và nối chúng với nhau để lấy mẫu (thông qua deconvolution) bản đồ tính năng bộ giải mã. Không có khối conv5 và max-pool 5 trong U-Net như trong kiến trúc mạng VGG. Mặt khác, SegNet sử dụng tất cả các trọng số lớp tích chập được đào tạo trước từ mạng VGG làm trọng số được đào tạo trước.

3.1 Biến thể bộ giải mã

Nhiều kiến trúc phân đoạn [2], [3], [4] chia sẻ cùng một mạng bộ mã hóa và chúng chỉ khác nhau ở dạng mạng bộ giải mã của chúng. Trong số này, chúng tôi chọn so sánh kỹ thuật giải mã SegNet với kỹ thuật giải mã Mạng tích chập hoàn toàn (FCN) được sử dụng rộng rãi [2], [10].

Để phân tích SegNet và so sánh hiệu suất của nó với FCN (các biến thể của bộ giải mã), chúng tôi sử dụng một phiên bản SegNet nhỏ hơn, được gọi là SegNet-Basic có 4 bộ mã hóa[1] và 4 bộ giải mã.
Tất cả các bộ mã hóa trong SegNet-Basic thực hiện lấy mẫu phụ và tổng hợp tối đa, đồng thời các bộ giải mã tương ứng lấy mẫu đầu vào của nó bằng cách sử dụng các chỉ số tổng hợp tối đa nhận được. Chuẩn hóa hàng loạt được sử dụng sau mỗi lớp tích chập trong cả mạng bộ mã hóa và bộ giải mã. Không có độ lệch nào được sử dụng sau khi tích chập và không có tính phi tuyến tính ReLU nào xuất hiện trong mạng bộ giải mã. Hơn nữa, kích thước hạt nhân không đổi là 7 × 7 trên tất cả các lớp bộ mã hóa và giải mã được chọn để cung cấp ngữ cảnh rộng cho việc ghi nhãn trơn tru, tức là một pixel trong bản đồ tính năng lớp sâu nhất (lớp 4) có thể được truy trở lại cửa sổ ngữ cảnh trong hình ảnh đầu vào 106 × 106 pixel. Kích thước nhỏ này của SegNet-Basic cho phép chúng tôi khám phá nhiều biến thể (bộ giải mã) khác nhau và đào tạo chúng trong thời gian hợp lý. Tương tự, chúng tôi tạo FCN-Basic, một phiên bản FCN có thể so sánh được để phân tích chia sẻ cùng một mạng bộ mã hóa như SegNet-Basic nhưng với kỹ thuật giải mã FCN (xem Hình 3) được sử dụng trong tất cả các bộ giải mã của nó.

Ở bên trái trong Hình 3 là kỹ thuật giải mã được sử dụng bởi SegNet (cũng là SegNet-Basic), trong đó không có việc học liên quan đến bước lấy mẫu. Tuy nhiên, các bản đồ lấy mẫu được kết hợp với các bộ lọc giải mã đa kênh có thể huấn luyện để tăng mật độ các đầu vào thưa thớt của nó. Mỗi bộ lọc bộ giải mã có cùng số lượng kênh với số lượng bản đồ tính năng được lấy mẫu. Một biến thể nhỏ hơn là biến thể trong đó các bộ lọc của bộ giải mã là một kênh, tức là chúng chỉ kết hợp bản đồ tính năng được lấy mẫu ngược tương ứng của chúng. Biến thể này (SegNet Basic-SingleChannelDecoder) giảm đáng kể số lượng tham số có thể huấn luyện và thời gian suy luận.

Ở bên phải trong Hình 3 là kỹ thuật giải mã FCN (cũng là FCN-Basic). Yếu tố thiết kế quan trọng của mô hình FCN là bước giảm kích thước của bản đồ đặc trưng bộ mã hóa. Thao tác này sẽ nén các bản đồ tính năng của bộ mã hóa, sau đó được sử dụng trong các bộ giải mã tương ứng. Việc giảm kích thước của bản đồ tính năng bộ mã hóa, chẳng hạn như 64 kênh, được thực hiện bằng cách kết hợp chúng với các bộ lọc có thể huấn luyện 1 × 1 × 64 × K , trong đó K là số lớp. Bản đồ đặc trưng lớp bộ mã hóa cuối cùng của kênh K được nén là đầu vào của mạng bộ giải mã. Trong bộ giải mã của mạng này, việc lấy mẫu tăng được thực hiện bằng cách tích chập nghịch đảo bằng cách sử dụng hạt nhân lấy mẫu đa kênh cố định hoặc có thể đào tạo được. Chúng tôi đặt kích thước hạt nhân thành 8 × 8. Cách lấy mẫu tăng này còn được gọi là giải mã. Lưu ý rằng, để so sánh, SegNet tích chập đa kênh sử dụng các bộ lọc bộ giải mã có thể đào tạo được thực hiện sau khi lấy mẫu lên các bản đồ đặc trưng dày đặc. Bản đồ đối tượng lấy mẫu trong FCN có K kênh. Sau đó, nó được thêm từng phần tử vào bản đồ đặc trưng của bộ mã hóa độ phân giải tương ứng để tạo ra bản đồ đặc trưng của bộ giải mã đầu ra. Các hạt nhân upsampling được khởi tạo bằng cách sử dụng trọng số nội suy song tuyến tính [2].

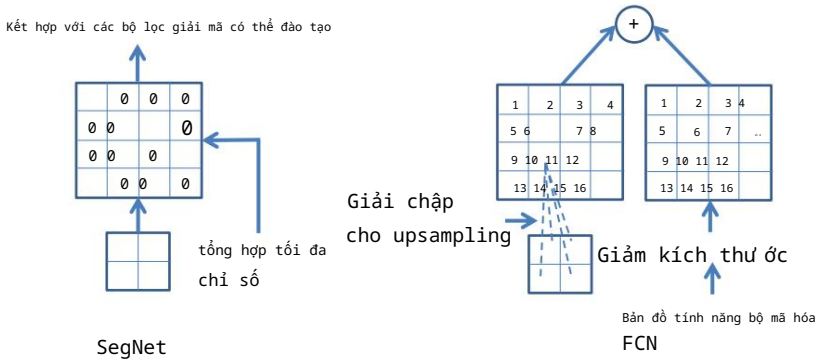
Mô hình bộ giải mã FCN yêu cầu lưu trữ các bản đồ đặc trưng của bộ mã hóa trong quá trình suy luận. Điều này có thể tốn nhiều bộ nhớ cho các ứng dụng nhúng; ví dụ: lưu trữ 64 bản đồ tính năng của lớp đầu tiên của FCN-Basic ở độ phân giải 180 × 240 với độ chính xác dấu phẩy động 32 bit chiếm 11 MB. Điều này có thể được thực hiện nhỏ hơn bằng cách giảm kích thước cho 11 bản đồ tính năng yêu cầu bộ nhớ ≈ 1,9 MB.
Mặt khác, SegNet yêu cầu chi phí lưu trữ gần như không đáng kể đối với các chỉ số gộp (0,17 MB nếu được lưu trữ bằng 2 bit trên mỗi cửa sổ gộp 2 × 2). Chúng tôi cũng có thể tạo một biến thể của mô hình FCN-Basic loại bỏ bước bổ sung bản đồ tính năng bộ mã hóa và chỉ tìm hiểu các hạt nhân lấy mẫu tăng (FCN-Basic-NoAddition).

Ngoài các biến thể trên, chúng tôi nghiên cứu lấy mẫu ngược bằng cách sử dụng các trọng số nội suy song tuyến tính cố định, do đó không cần học cách lấy mẫu ngược (Nội suy song tuyến tính). Ở một thái cực khác, chúng ta có thể thêm 64 bản đồ tính năng bộ mã hóa ở mỗi lớp vào bản đồ tính năng đầu ra tương ứng từ bộ giải mã SegNet để tạo ra một biến thể sử dụng nhiều bộ nhớ hơn của SegNet (SegNet-Basic EncoderAddition). Ở đây, cả hai chỉ số tổng hợp để lấy mẫu tăng đều được sử dụng, tiếp theo là bước tích chập để tăng mật độ đầu vào thưa thớt của nó.
Sau đó, phần tử này được thêm vào bản đồ tính năng của bộ mã hóa tương ứng để tạo ra đầu ra của bộ giải mã.

Một biến thể FCN-Basic khác và sử dụng nhiều bộ nhớ hơn (FCN-Basic-NoDimReduction) là nơi không thực hiện giảm tính chất kích thước cho bản đồ tính năng bộ mã hóa. Điều này ngụ ý rằng không giống như FCN-Basic, bản đồ tính năng bộ mã hóa cuối cùng không được nén thành K kênh trước khi chuyển nó tới mạng bộ giải mã. Do đó, số lượng kênh ở cuối mỗi bộ giải mã giống như bộ mã hóa tương ứng (tức là 64).

Chúng tôi cũng đã thử các biến thể chung khác trong đó các bản đồ đặc trưng được lấy mẫu đơn giản bằng cách sao chép [7] hoặc bằng cách sử dụng một bản đồ cố định (và

1. SegNet-Basic trước đó được gọi là SegNet trong phiên bản lưu trữ của bài báo này [12]



Hình 3. Hình minh họa bộ giải mã SegNet và FCN [2]. a, b, c, d tư ơ ng ứng với các giá trị trong bản đồ đặc trư ng. SegNet sử dụng các chỉ số tổng hợp tối đa để lấy mẫu (không cần học) (các) bản đồ tính năng và kết hợp với ngân hàng bộ lọc bộ giải mã có thể đào tạo. FCN upsamples bằng cách học cách giải mã bản đồ tính năng đầu vào và thêm bản đồ tính năng bộ mã hóa tư ơ ng ứng để tạo đầu ra bộ giải mã. Bản đồ tính năng này là đầu ra của lớp tổng hợp tối đa (bao gồm lấy mẫu phụ) trong bộ mã hóa tư ơ ng ứng. Lư u ý rằng không có bộ lọc giải mã có thể đào tạo nào trong FCN.

				Cân bằng tần số trung bình				Cân bằng tần số tự nhiên	
Khác nhau	Tham số (M)	Kho thời gian hệ số	Suy luận nhân (ms)	Bài kiểm tra GC	mIoU	Bài kiểm tra BF GC	mIoU	Bài kiểm tra GC	mIoU
Đã sửa lỗi lấy mẫu									
Nội suy song tuyến tính	0,625	0 Upsampling	24,2	77,9	61,1	43,3	20,83	89,1	90,2
sử dụng chỉ số gộp tối đa									
SegNet-Cơ bản	1.425		52.6	82.7	62.0	47.7	35.78	94.7	96. 2
SegNet-Basic-EncoderAddition	1.425	1	35.92	94	3 95.8	92.0	84.2	56.5	47.7
SegNet-Basic-SingleChannelDecoder	0.625		83.5	53.9	45.2	32.45	92.6	68.4	52.8
64 1 Học cách upsample (khởi tạo song tuyến tính)									
FCN-Cơ bản	0,65	11 0,65 n/a	24.2	81.7	62.4	47.3	38.11	92.8	93.6
FCN-Cơ bản-Không bổ sung	1,625	64 FCN-Basic-	31.96	92	5 93.0	87.2	82.3	53.9	44.2
FCN-Cơ bản-NoDimReduction	NoAddition-		83.5	57.3	47.0	37.13	97.2	91.7	84.8
NoDimReduction 1,625	0		95,0	80,2	67,8				

				Cân bằng tần số trung bình				Cân bằng tần số tự nhiên	
Khác nhau	Tham số (M)	Kho thời gian hệ số	Suy luận nhân (ms)	Bài kiểm tra GC	mIoU	Bài kiểm tra BF GC	mIoU	Bài kiểm tra GC	mIoU
Đã sửa lỗi lấy mẫu									
Nội suy song tuyến tính	0,625	0 Upsampling	24,2	77,9	61,1	43,3	20,83	89,1	90,2
sử dụng chỉ số gộp tối đa									
SegNet-Cơ bản	1.425		52.6	82.7	62.0	47.7	35.78	94.7	96. 2
SegNet-Basic-EncoderAddition	1.425	1	35.92	94	3 95.8	92.0	84.2	56.5	47.7
SegNet-Basic-SingleChannelDecoder	0.625		83.5	53.9	45.2	32.45	92.6	68.4	52.8
64 1 Học cách upsample (khởi tạo song tuyến tính)									
FCN-Cơ bản	0,65	11 0,65 n/a	24.2	81.7	62.4	47.3	38.11	92.8	93.6
FCN-Cơ bản-Không bổ sung	1,625	64 FCN-Basic-	31.96	92	5 93.0	87.2	82.3	53.9	44.2
FCN-Cơ bản-NoDimReduction	NoAddition-		83.5	57.3	47.0	37.13	97.2	91.7	84.8
NoDimReduction 1,625	0		95,0	80,2	67,8				

BẢNG 1 So

sánh các biến thể bộ giải mã. Chúng tôi định lư ợng hiệu suất bằng cách sử dụng toàn cầu (G), trung bình loại (C), giá trị trung bình của giao điểm trên liên kết (mIoU) và thư ớc đo đư ờng viền ngữ nghĩa (BF). Độ chính xác của kiểm tra và đào tạo đư ợc hiển thị dư ới dạng phần trăm cho cả chức năng mất đào tạo cân bằng tần số tự nhiên và tần số trung bình. SegNet-Basic hoạt động ở cùng cấp độ với FCN-Basic như ng chỉ yêu cầu lư u trữ các chỉ số tổng hợp tối đa và do đó sử dụng bộ nhớ hiệu quả hơ n trong quá trình suy luận. Lư u ý rằng yêu cầu bộ nhớ lý thuyết đư ợc báo cáo chỉ dựa trên kích thư ớc của bản đồ tính năng bộ mã hóa lớp đầu tiên. FCN-Basic, SegNet-Basic, SegNet-Basic-EncoderAddition đều có điểm BF cao cho thấy như cầu sử dụng thông tin trong bản đồ tính năng của bộ mã hóa để phân định đư ờng viền lớp tốt hơ n. Các mạng có bộ giải mã lớn hơ n và những mạng sử dụng bản đồ đặc trư ng của bộ mã hóa hoạt động tốt nhất, mặc dù chúng kém hiệu quả nhất về thời gian suy luận và bộ nhớ.

thư a thớt) mảng chỉ số để upsampling. Chúng hoạt động khá kém so với các biến thể trên. Một biến thể không có tổng hợp tối đa và lấy mẫu phụ trong mạng bộ mã hóa (bộ giải mã là dự phòng) sẽ tiêu tốn nhiều bộ nhớ hơ n, mất nhiều thời gian hơ n để hội tụ và hoạt động kém. Cuối cùng, xin lư u ý rằng để khuyến khích sao chép kết quả của chúng tôi, chúng tôi phát hành triển khai Caffe cho cả 2 biến thể

.

3.2 Đào tạo Chúng

tôi sử dụng bộ dữ liệu cảnh đư ờng CamVid để đánh giá hiệu suất của các biến thể bộ giải mã. Bộ dữ liệu này nhỏ, bao gồm 367 hình ảnh huấn luyện và 233 hình ảnh RGB thử nghiệm (cảnh ban ngày và hoàng hôn) ở độ phân giải 360×480 . Thách thức là phân đoạn 11 lớp chẳng hạn như đư ờng, tòa nhà, ô tô, ngư ời đi bộ, biển báo, cột điện, vỉa hè, v.v. Chúng tôi thực hiện chuẩn hóa độ tư ơ ng phản cục bộ [54] cho đầu vào RGB.

Tất cả các trọng số của bộ mã hóa và bộ giải mã đều đư ợc khởi tạo bằng kỹ thuật đư ợc mô tả trong He et al. [55]. Để huấn luyện tất cả các biến thể, chúng tôi sử dụng phư ơ ng pháp giảm độ dốc ngẫu nhiên (SGD) với tốc độ học cố định là 0,1 và động lư ợng là 0,9 [17] bằng cách sử dụng triển khai Caffe của SegNet-Basic [56]. Chúng tôi đào tạo các biến thể cho đến khi mất đào tạo hội tụ. Trư ớc mỗi kỷ nguyên, tập huấn luyện đư ợc xáo trộn và mỗi lô nhỏ (12 hình ảnh) sau đó đư ợc chọn theo thứ tự, do đó đảm bảo rằng mỗi hình ảnh chỉ đư ợc sử dụng một lần trong một kỷ nguyên. Chúng tôi chọn mô hình hoạt động cao nhất trên tập dữ liệu xác thực.

Chúng tôi sử dụng tổn thất entropy chéo [2] làm hàm mục tiêu để huấn luyện mạng. Mất mát đư ợc tổng hợp trên tất cả các pixel trong một lô nhỏ. Khi có sự khác biệt lớn về số lư ợng pixel trong mỗi lớp trong tập huấn luyện (ví dụ: pixel đư ờng, bầu trời và tòa nhà chiếm ư u thế trong tập dữ liệu CamVid) thì cần phải cân nhắc mức độ mất mát khác nhau dựa trên lớp thực. Điều này đư ợc gọi là cân bằng lớp. Chúng tôi sử dụng cân bằng tần số trung bình [13] trong đó trọng số đư ợc gán cho một lớp trong hàm mất mát là tỷ lệ giữa tần số trung bình của lớp đư ợc tính trên toàn bộ tập huấn luyện chia cho tần số của lớp. Điều này ngụ ý rằng các lớp lớn hơ n trong tập huấn luyện có trọng số nhỏ hơ n 1 và trọng số của các lớp nhỏ nhất là cao nhất. Chúng tôi cũng đã thử nghiệm đào tạo các biến thể khác nhau mà không cần cân bằng lớp hoặc sử dụng cân bằng tần số tự nhiên tư ơ ng đư ơ ng.

3.3 Phân tích Để

so sánh hiệu suất định lư ợng của các biến thể bộ giải mã khác nhau, chúng tôi sử dụng ba phép đo hiệu suất thư ờng đư ợc sử dụng: độ chính xác toàn cầu (G) đo tỷ lệ phần trăm pixel

2. Xem http://mi.eng.cam.ac.uk/projects/segnet/ để biết mã SegNet và bản demo web của chúng tôi.

được phân loại chính xác trong tập dữ liệu, độ chính xác trung bình của lớp (C) là giá trị trung bình của độ chính xác dự đoán trên tất cả các lớp và giao điểm trung bình trên liên kết (mIoU) trên tất cả các lớp như được sử dụng trong thử thách Pas cal VOC12 [21]. Số liệu mIoU là số liệu nghiêm ngặt hơn so với độ chính xác trung bình của lớp vì nó xử phạt các dự đoán dương sai. Tuy nhiên, số liệu mIoU không được tối ưu hóa trực tiếp thông qua tổn thất entropy chéo đã cân bằng của lớp.

Số liệu mIoU còn được gọi là Chỉ số Jacard là thường được sử dụng trong benchmarking. Tuy nhiên, Csurka et al. [57] lưu ý rằng số liệu này không phải lúc nào cũng tương ứng với các đánh giá định tính của con người (xếp hạng) về phân khúc chất lượng tốt. Họ chỉ ra bằng các ví dụ rằng mIoU ủng hộ độ trơn của vùng và không đánh giá độ chính xác của ranh giới, một điểm gần đây cũng được các tác giả của FCN [58] ám chỉ. Do đó, họ đề xuất bổ sung số liệu mIoU bằng thước đo ranh giới dựa trên điểm phù hợp với đường viền Berkeley thường được sử dụng để đánh giá chất lượng phân đoạn hình ảnh không được giám sát [59]. Csurka và cộng sự. [57] chỉ cần mở rộng điều này sang phân đoạn ngữ nghĩa và chỉ ra rằng phép đo độ chính xác của đường viền ngữ nghĩa được sử dụng cùng với số liệu mIoU phù hợp hơn với xếp hạng của con người về đầu ra phân đoạn.

Ý tưởng chính trong việc tính toán điểm đường viền ngữ nghĩa là đánh giá thước đo F1 [59] liên quan đến việc tính toán độ chính xác và giá trị thu hồi giữa ranh giới lớp chân lý được dự đoán và lớp nền cho một khoảng cách dung sai pixel. Chúng tôi đã sử dụng giá trị 0,75% của đường chéo hình ảnh làm khoảng cách dung sai. Số đo F1 cho mỗi lớp hiện diện trong ảnh kiểm tra độ thật cơ bản được tính trung bình để tạo ra một số đo F1 của ảnh. Sau đó, chúng tôi tính trung bình toàn bộ tập kiểm tra, biểu thị ranh giới-đo F1 (BF) bằng trung bình hình ảnh F1 đo.

Chúng tôi kiểm tra từng biến thể kiến trúc sau mỗi 1000 lần lặp tối ưu hóa trên bộ xác thực CamVid cho đến khi tổn thất huấn luyện hội tụ. Với kích thước lô nhỏ đào tạo là 12, điều này tương ứng với việc kiểm tra khoảng 33 kỷ nguyên (vượt qua) thông qua tập huấn luyện. Chúng tôi chọn phép lặp trong đó độ chính xác toàn cầu là cao nhất trong số các đánh giá trên bộ xác thực. Chúng tôi báo cáo tất cả ba thước đo hiệu suất tại thời điểm này trên bộ thử nghiệm CamVid đã được tổ chức. Mặc dù chúng tôi sử dụng cân bằng lớp trong khi đào tạo các biến thể, điều quan trọng vẫn là đạt được độ chính xác toàn cầu cao để dẫn đến phân đoạn mượt mà tổng thể. Một lý do khác là sự đóng góp của phân đoạn đối với lái xe tự động chủ yếu là để phân định các lớp như đường, tòa nhà, vỉa hè, bầu trời. Các lớp này chi phối phần lớn các pixel trong ảnh và độ chính xác toàn cục cao tương ứng với việc phân đoạn tốt các lớp quan trọng này. Chúng tôi cũng quan sát thấy rằng việc báo cáo hiệu suất bằng số khi mức trung bình của lớp cao nhất thường có thể tương ứng với độ chính xác toàn cầu thấp cho thấy đầu ra phân đoạn bị nhiễu về mặt nhận thức.

Trong Bảng 1, chúng tôi báo cáo các kết quả phân tích bằng số của chúng tôi. Chúng tôi cũng hiển thị kích thước của các tham số có thể huấn luyện và bản đồ tính năng có độ phân giải cao nhất hoặc bộ nhớ lưu trữ chỉ số tổng hợp, tức là của bản đồ tính năng lớp đầu tiên sau khi tổng hợp tối đa và lấy mẫu phụ.
Chúng tôi hiển thị thời gian trung bình cho một lượt chuyển tiếp khi triển khai Caffe, tính trung bình hơn 50 phép đo bằng cách sử dụng đầu vào 360 × 480 trên GPU NVIDIA Titan với khả năng tăng tốc cuDNN v3.
Chúng tôi lưu ý rằng các lớp lấy mẫu trong các biến thể SegNet không được tối ưu hóa bằng cách sử dụng tính năng tăng tốc cuDNN. Chúng tôi hiển thị kết quả cho cả thử nghiệm và đào tạo cho tất cả các biến thể ở lần lặp đã chọn. Các kết quả cũng được lập bảng mà không cần cân bằng lớp (tần số tự nhiên) để đào tạo và kiểm tra độ chính xác. Dưới đây chúng tôi phân tích kết quả với cân bằng lớp.

Từ Bảng 1, chúng ta thấy rằng phép nội suy song tuyến tính dựa trên upsampling mà không có bất kỳ học tập nào hoạt động kém nhất dựa trên tất cả các biện pháp đo lường độ chính xác. Tất cả các phương pháp khác sử dụng tính năng học để lấy mẫu tăng (FCN-Basic và các biến thể) hoặc học các bộ lọc bộ giải mã sau khi lấy mẫu tăng (SegNet-Basic và các biến thể của nó) đều hoạt động tốt hơn đáng kể. Điều này nhấn mạnh sự cần thiết phải học bộ giải mã để phân đoạn. Điều này cũng được hỗ trợ bởi các bằng chứng thực nghiệm được thu thập bởi các tác giả khác khi so sánh FCN với các kỹ thuật giải mã kiểu SegNet [4].

Khi so sánh SegNet-Basic và FCN-Basic, chúng tôi thấy rằng cả hai đều hoạt động tốt như nhau trong bài kiểm tra này trên tất cả các thước đo về độ chính xác. Sự khác biệt là SegNet sử dụng ít bộ nhớ hơn trong quá trình suy luận vì nó chỉ lưu trữ các chỉ số gộp tối đa. Mặt khác, FCN-Basic lưu trữ đầy đủ các bản đồ tính năng của bộ mã hóa , tiêu tốn nhiều bộ nhớ hơn (gấp 11 lần). SegNet-Basic có bộ giải mã với 64 bản đồ tính năng trong mỗi lớp bộ giải mã. So với FCN-Basic, sử dụng giảm kích thước, có ít hơn (11) bản đồ đặc trưng trong mỗi lớp bộ giải mã. Điều này làm giảm số lượng tích chập trong mạng bộ giải mã và do đó FCN-Basic nhanh hơn trong quá trình suy luận (chuyển tiếp). Từ một khía cạnh khác, mạng bộ giải mã trong SegNet-Basic làm cho nó trở thành một mạng lớn hơn so với FCN-Basic. Điều này mang lại cho nó tính linh hoạt hơn và do đó đạt được độ chính xác đào tạo cao hơn so với FCN-Basic cho cùng số lần lặp lại. Nhìn chung, chúng tôi thấy rằng SegNet-Basic có lợi thế hơn FCN-Basic khi bộ nhớ thời gian suy luận bị hạn chế nhưng khi thời gian suy luận có thể bị tổn hại ở một mức độ nào đó.

SegNet-Basic gần giống với FCN-Basic-NoAddition nhất về bộ giải mã của chúng, mặc dù bộ giải mã của SegNet lớn hơn.
Cả hai đều học cách tạo các bản đồ tính năng dày đặc, trực tiếp bằng cách học cách thực hiện giải mã như trong FCN-Basic-NoAddition hoặc bằng cách lấy mẫu trước tiên và sau đó kết hợp với các bộ lọc bộ giải mã được đào tạo. Hiệu suất của SegNet-Basic cao hơn, một phần là do kích thước bộ giải mã lớn hơn. Độ chính xác của FCN-Basic-NoAddition cũng thấp hơn so với FCN-Basic. Điều này cho thấy rằng điều quan trọng là nắm bắt thông tin có trong bản đồ tính năng của bộ mã hóa để có hiệu suất tốt hơn. Đặc biệt, lưu ý sự sụt giảm lớn trong phép đo BF giữa hai biến thể này. Điều này cũng có thể giải thích một phần lý do tại sao SegNet-Basic vượt trội so với FCN-Basic NoAddition.

Kích thước của FCN-Basic-NoAddition-NoDimReduction mô hình lớn hơn một chút so với SegNet-Basic do các bản đồ tính năng của bộ mã hóa cuối cùng không được nén để phù hợp với số lớp K. Điều này làm cho nó trở thành một sự so sánh công bằng về kích thước của mô hình. Hiệu suất của biến thể FCN này kém hơn so với SegNet-Basic trong thử nghiệm nhưng độ chính xác đào tạo của nó cũng thấp hơn đối với cùng số lượng kỷ nguyên đào tạo. Điều này cho thấy rằng sử dụng bộ giải mã lớn hơn là chưa đủ nhưng điều quan trọng là nắm bắt thông tin bản đồ tính năng của bộ mã hóa để tìm hiểu tốt hơn, đặc biệt là thông tin đường viền chi tiết (chú ý sự sụt giảm trong thước đo BF). Ở đây, thật thú vị khi thấy rằng SegNet-Basic có độ chính xác đào tạo cạnh tranh khi so sánh với các mô hình lớn hơn như FCN Basic-NoDimReduction.

Một so sánh thú vị khác giữa FCN-Basic NoAddition và SegNet-Basic-SingleChannelDecoder cho thấy rằng việc sử dụng các chỉ số tổng hợp tối đa để lấy mẫu nâng cấp và một bộ giải mã tổng thể lớn hơn sẽ dẫn đến hiệu suất tốt hơn. Điều này cũng cho thấy SegNet là một kiến trúc tốt cho phân khúc, đặc biệt khi có nhu cầu tìm kiếm sự thỏa hiệp giữa chi phí lưu trữ, độ chính xác so với thời gian suy luận. Trong trường hợp tốt nhất, khi cả bộ nhớ và thời gian suy luận đều không bị hạn chế, các mô hình lớn hơn như FCN-Basic-NoDimReduction và SegNet-EncoderAddition được

cả hai đều chính xác hơn các biến thể khác. Đặc biệt, việc loại bỏ giảm kích thước trong mô hình FCN-Basic dẫn đến hiệu suất tốt nhất trong số các biến thể FCN-Basic có điểm BF cao. Điều này một lần nữa nhấn mạnh sự đánh đổi giữa bộ nhớ và độ chính xác trong kiến trúc phân đoạn.

Hai cột cuối cùng của Bảng 1 hiển thị kết quả khi không sử dụng cân bằng lớp (tần số tự nhiên). Ở đây, chúng ta có thể quan sát thấy rằng nếu không tính trọng số thì kết quả sẽ kém hơn đối với tất cả các biến thể, đặc biệt là đối với độ chính xác trung bình của lớp và chỉ số mIoU. Độ chính xác toàn cầu là cao nhất mà không tính trọng số vì phần lớn cảnh bị chi phối bởi pixel bầu trời, đường xá và tòa nhà. Ngoài ra, tất cả suy luận từ phân tích so sánh của các biến thể cũng đúng đối với cân bằng tần số tự nhiên, bao gồm các xu hướng đối với thước đo BF. SegNet Basic hoạt động tốt như FCN-Basic và tốt hơn FCN-Basic-NoAddition-NoDimReduction lớn hơn. Các mô hình lớn hơn nhưng kém hiệu quả hơn FCN-Basic-NoDimReduction và SegNet EncoderAddition hoạt động tốt hơn các biến thể khác.

Bây giờ chúng ta có thể tóm tắt các phân tích trên với các điểm chung sau đây.

1) Hiệu suất tốt nhất đạt được khi bản đồ tính năng bộ mã hóa được lưu trữ đầy đủ. Điều này được phản ánh rõ ràng nhất trong thước đo phân định đường viền ngữ nghĩa (BF).
2) Khi bộ nhớ trong quá trình suy luận bị hạn chế, thì có thể lưu trữ và sử dụng các dạng nén của bản đồ tính năng bộ mã hóa (giảm kích thước, chỉ số tổng hợp tối đa) với bộ giải mã thích hợp (ví dụ: loại SegNet) để cải thiện hiệu suất.

3) Bộ giải mã lớn hơn tăng hiệu suất cho một bộ mã hóa nhất định mạng.

4 ĐIỂM CHUẨN

Chúng tôi định lượng hiệu suất của SegNet trên hai tiêu chuẩn phân đoạn cảnh bằng cách sử dụng triển khai Caffe của chúng tôi [3]. Nhiệm vụ đầu tiên là phân đoạn cảnh trên đường, đây là mối quan tâm thực tế hiện nay đối với các vấn đề liên quan đến lái xe tự động khác nhau. Nhiệm vụ thứ hai là phân đoạn cảnh trong nhà, đây là mối quan tâm ngay lập tức đối với một số ứng dụng thực tế tăng cường (AR). Hình ảnh RGB đầu vào cho cả hai tác vụ là 360 × 480.

Chúng tôi đã so sánh SegNet so với một số kiến trúc chuyên sâu được áp dụng tốt khác để phân đoạn, chẳng hạn như FCN [2], DeepLab LargFOV [3] và DeconvNet [4]. Mục tiêu của chúng tôi là hiểu hiệu suất của các kiến trúc này khi được đào tạo từ đầu đến cuối trên cùng một bộ dữ liệu. Để cho phép đào tạo từ đầu đến cuối, chúng tôi đã thêm các lớp [51] chuẩn hóa hàng loạt sau mỗi lớp tích chập. Đối với DeepLab-LargeFOV, chúng tôi đã thay đổi bước gộp tối đa 3 thành 1 để đạt được độ phân giải dự đoán cuối cùng là 45 × 60. Chúng tôi đã giới hạn kích thước đối tượng trong các lớp DeconvNet được kết nối đầy đủ thành 1024 để cho phép đào tạo với cùng kích thước lô như ngươi mẫu khác.
Lưu ý ở đây rằng các tác giả của DeepLab-LargeFOV [3] cũng đã báo cáo một chút mất mát về hiệu suất bằng cách giảm kích thước của các lớp được kết nối đầy đủ.

Để thực hiện một điểm chuẩn có kiểm soát, chúng tôi đã sử dụng cùng một bộ giải SGD [17] với tốc độ học cố định là 10 3 và động lượng là 0,9. Quá trình tối ưu hóa được thực hiện trong hơn 100 kỷ nguyên thông qua tập dữ liệu cho đến khi không thấy hiệu suất tăng thêm nữa. Dropout 0,5 đã được thêm vào cuối sâu hơn các lớp tích chập trong tất cả các mô hình để tránh trang bị quá mức (xem http://mi.eng.cam.ac.uk/projects/segnet/tutorial.html để biết ví dụ về caffe prototxt). Đối với các cảnh trên đường có 11 lớp, chúng tôi sử dụng kích thước lô nhỏ là 5 và đối với các cảnh trong nhà có 37 lớp, chúng tôi sử dụng kích thước lô nhỏ là 4.

4.1 Phân đoạn cảnh đường Một số bộ dữ

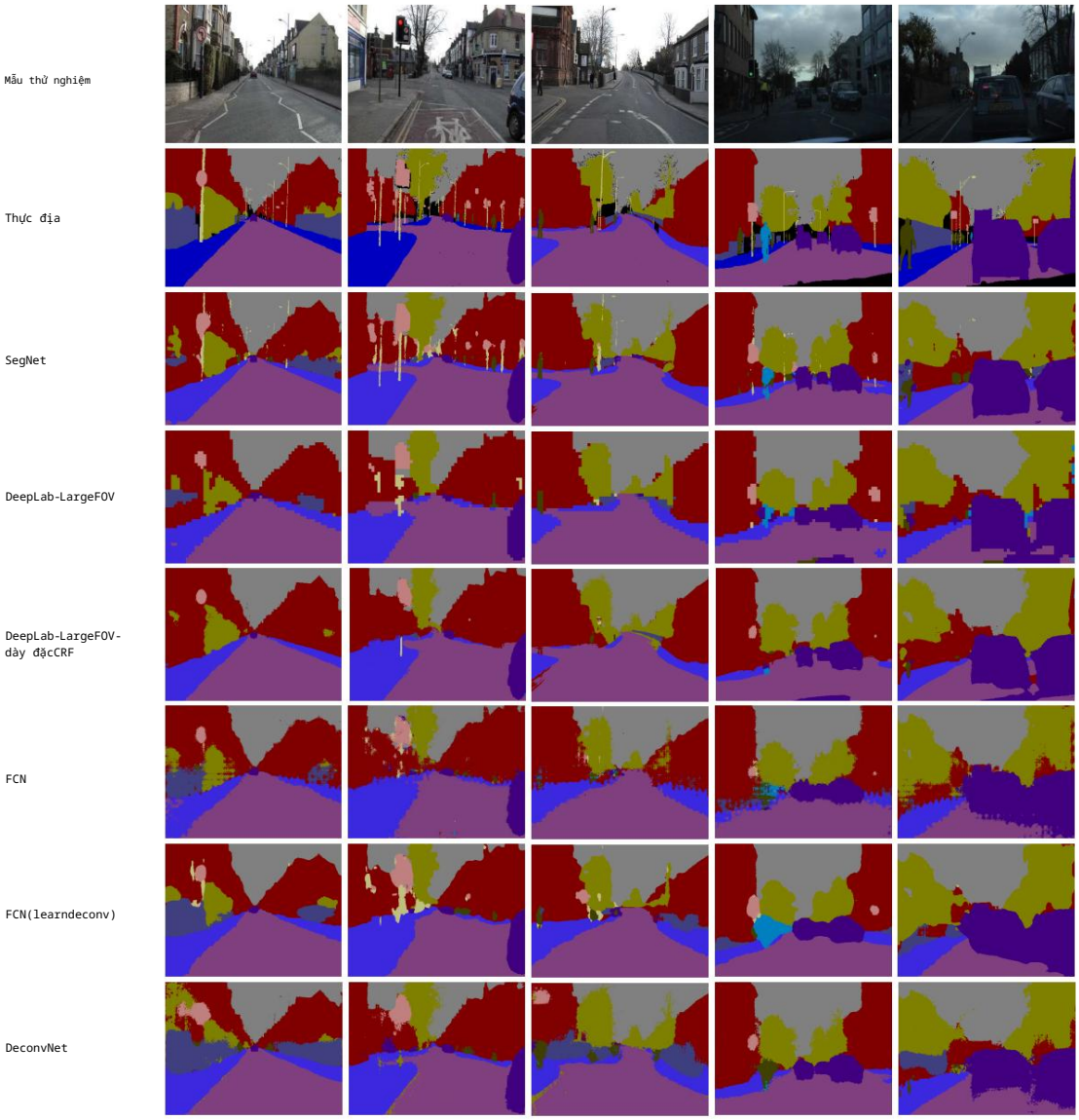
liệu cảnh đường có sẵn để phân tích cú pháp ngữ nghĩa [22], [26], [60], [61]. Trong số này, chúng tôi chọn đánh giá SegNet bằng bộ dữ liệu CamVid [22] vì nó chứa các chuỗi video. Điều này cho phép chúng tôi so sánh kiến trúc đề xuất của mình với kiến trúc sử dụng chuyển động và cấu trúc [28], [29], [30] và phân đoạn video [33].
Chúng tôi cũng kết hợp [22], [26], [60], [61] để tạo thành một tập hợp gồm 3433 hình ảnh nhằm huấn luyện SegNet cho điểm chuẩn bổ sung. Đối với bản trình diễn web (xem chú thích cuối trang 3) về phân đoạn cảnh đường, chúng tôi đưa bộ kiểm tra CamVid vào bộ dữ liệu lớn hơn này. Ở đây, chúng tôi muốn lưu ý rằng một điểm chuẩn phân khúc độc lập và gần đây khác trên các cảnh đường đã được thực hiện cho SegNet và các kiến trúc cạnh tranh khác được sử dụng trong bài viết này [62]. Tuy nhiên, điểm chuẩn không được kiểm soát, nghĩa là mỗi kiến trúc được đào tạo với một công thức riêng biệt với các độ phân giải đầu vào khác nhau và đôi khi có kèm theo một bộ xác thực. Do đó, chúng tôi tin rằng điểm chuẩn được kiểm soát nhiều hơn của chúng tôi có thể được sử dụng để bổ sung cho những nỗ lực của họ.

Có thể thấy các so sánh định tính của các dự đoán SegNet với các kiến trúc sâu khác trong Hình 4. Các kết quả định tính cho thấy khả năng của kiến trúc được đề xuất trong việc phân đoạn các lớp nhỏ hơn trong các cảnh đường trong khi tạo ra sự phân đoạn mượt mà của cảnh tổng thể. Thật vậy, trong cài đặt điểm chuẩn được kiểm soát, SegNet cho thấy hiệu suất vượt trội so với một số mô hình lớn hơn. DeepLab-LargeFOV là mô hình hiệu quả nhất và với xử lý hậu kỳ CRF có thể tạo ra kết quả cạnh tranh mặc dù các lớp nhỏ hơn bị mất. FCN với giải mã đã học rõ ràng là tốt hơn so với lấy mẫu song tuyến tính cố định. DeconvNet là mô hình lớn nhất và kém hiệu quả nhất để đào tạo. Dự đoán của nó không giữ lại các lớp nhỏ.

Trước tiên, chúng tôi cũng sử dụng điểm chuẩn này để so sánh SegNet với một số phương pháp không học sâu bao gồm Random Forests [27], Boosting [27], [29] kết hợp với các phương pháp dựa trên CRF [30]. Điều này được thực hiện để cung cấp cho người dùng góc nhìn về những cải tiến về độ chính xác đã đạt được khi sử dụng mạng sâu so với các kỹ thuật dựa trên kỹ thuật tính năng cổ điển.

Kết quả trong Bảng 2 cho thấy SegNet-Basic, SegNet thu được kết quả cạnh tranh khi so sánh với các phương pháp sử dụng CRF. Điều này cho thấy khả năng của kiến trúc sâu trong việc trích xuất các đặc điểm có ý nghĩa từ hình ảnh đầu vào và ánh xạ nó tới các nhãn phân đoạn lớp chính xác và mượt mà. Kết quả thú vị nhất ở đây là sự cải thiện hiệu suất lớn trong chỉ số trung bình của lớp và mIOU thu được khi tập dữ liệu huấn luyện lớn, thu được bằng cách kết hợp [22], [26], [60], [61], được sử dụng để huấn luyện SegNet. Tương ứng, các kết quả định tính của SegNet (xem Hình 4) rõ ràng là vượt trội so với các phương pháp còn lại. Nó có thể phân chia tốt cả các lớp học nhỏ và lớn. Ở đây, chúng tôi nhận xét rằng chúng tôi đã sử dụng cân bằng lớp tần suất trung bình [50] trong đào tạo SegNet-Basic và SegNet. Ngoài ra, có một chất lượng tổng thể mượt mà của phân đoạn giống như những gì thường thu được với xử lý hậu kỳ CRF. Mặc dù thực tế là kết quả cải thiện với các tập huấn luyện lớn hơn không có gì đáng ngạc nhiên, nhưng phần trăm cải thiện thu được khi sử dụng mạng bộ mã hóa được đào tạo trước và tập huấn luyện này cho thấy rằng kiến trúc này có thể được triển khai cho

3. Bản giới thiệu web và triển khai Caffe của chúng tôi có sẵn để đánh giá tại http://mi.eng.cam.ac.uk/projects/segnet/



Hình 4. Kết quả trên các mẫu thử nghiệm CamVid ngày và hoàng hôn. SegNet cho thấy hiệu suất vượt trội, đặc biệt là với khả năng phân định ranh giới, so với một số mô hình lớn hơn khi tất cả được đào tạo trong môi trường được kiểm soát. DeepLab-LargeFOV là mô hình hiệu quả nhất và với xử lý hậu kỳ CRF có thể tạo ra kết quả cạnh tranh mặc dù các lớp nhỏ hơn bị mất. FCN với giải mã đã học rõ ràng là tốt hơn. DeconvNet là mô hình lớn nhất với thời gian đào tạo dài nhất, nhưng các dự đoán của nó lại bỏ qua các lớp nhỏ. Lưu ý rằng những kết quả này tương ứng với mô hình tương ứng với độ chính xác mIoU cao nhất trong Bảng 3.

ứng dụng thực tế. Thử nghiệm ngẫu nhiên của chúng tôi đối với hình ảnh đô thị và đường cao tốc từ internet (xem Hình 1) chứng minh rằng SegNet có thể hấp thụ một tập huấn luyện lớn và khái quát hóa tốt cho các hình ảnh không nhìn thấy được. Nó cũng chỉ ra rằng sự đóng góp của (CRF) trước đó có thể được giảm bớt khi có đủ lượng dữ liệu đào tạo.

Trong Bảng 3, chúng tôi so sánh hiệu suất của SegNet với các kiến trúc tích chập hoàn toàn hiện được áp dụng rộng rãi để phân đoạn. So với thử nghiệm trong Bảng 2, chúng tôi đã không sử dụng bất kỳ loại lớp nào để đào tạo bất kỳ kiến trúc chuyên sâu nào, kể cả SegNet. Điều này là do chúng tôi thấy khó đào tạo các mô hình lớn hơn như DeconvNet với cân bằng tần số trung bình. Chúng tôi đánh giá hiệu suất ở các lần lặp 40K, 80K và >80K với kích thước lô nhỏ và kích thước tập huấn luyện xấp xỉ tương ứng với 50, 100 và >100 kỷ nguyên. Đối với điểm kiểm tra cuối cùng, chúng tôi cũng báo cáo số lần lặp lại tối đa (ở đây ít nhất là 150 kỷ nguyên) ngoài số lần lặp lại mà chúng tôi quan sát thấy không có sự cải thiện về độ chính xác hoặc khi cài đặt quá mức phù hợp. Chúng tôi báo cáo các chỉ số ở ba giai đoạn trong giai đoạn đào tạo để tiết lộ cách các chỉ số thay đổi theo thời gian đào tạo, đặc biệt đối với các mạng lớn hơn. Điều quan trọng là phải hiểu nếu thời gian đào tạo bổ sung là hợp lý khi được thiết lập để tăng độ chính xác. Cũng lưu ý rằng đối với mỗi đánh giá, chúng tôi đã thực hiện chạy toàn bộ tập dữ liệu để thu được số liệu thống kê định mức lô và sau đó đánh giá mô hình thử nghiệm với thống kê này (xem http://mi.eng.cam.ac.uk/projects/segnet/tutorial. html cho mã.). Những đánh giá này rất tốn kém để thực hiện trên các tập huấn luyện lớn và do đó chúng tôi chỉ báo cáo số liệu tại ba thời điểm trong giai đoạn huấn luyện.

Từ Bảng 3, chúng ta thấy ngay rằng SegNet, DeconvNet đạt điểm cao nhất trong tất cả các số liệu so với các mô hình khác. DeconvNet có độ chính xác phân định ranh giới cao hơn nhưng SegNet hiệu quả hơn nhiều so với DeconvNet.

Điều này có thể được nhìn thấy từ số liệu thống kê tính toán trong Bảng 6. FCN,

Phương pháp																BF
SfM+Xuất hiện [28]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	61.9	67.3	91.1	71.1	58.5		0,7	60,5	22,5	53.0	37.6	25.8		69.1		n/a
Tăng cường [29]	92.9	49.5	85.3	57.3	95.4	69.2	46.5	98.5	23.8	44.3			77.8	24.7	59.8	22.0	38.1	28.7	55.4	51.4		76.4		n/
Bản đồ độ sâu dày đặc [32]															N/AN/A	56.1	80.7	61.5	88.8	16.4		82.1		a n/
Rừng ngẫu nhiên có cấu trúc [31]																				N/A		72.5		a n/
Rừng quyết định thần kinh [64]													98.0	1.09	4.13	12.4.	-	140K)	89,6	83,4	96,1	87,7	52,7	82.1		a n/
Bộ mô tả nhãn cục bộ [65]	96,4	62,2	53,45	32,1	93,3	36,5	71,20	90,40	60,10	46,84	Phương pháp dựa trên CRF											73.6		a n/
Siêu phân tích cú pháp [33]													17,9									83.3		a n/a
Tăng tốc + CRF theo cặp [29]	70.7	70.8	94.7	74.4	55.9	94.1	45.7	59.9	84,5	72.6	97,5	72.7	34.7	25.3	84.2	34.7	27.7	28.5	59.2	31.5		79,8		n/a
Tăng cường+Thứ tự cao hơn [29]	76.6	96.2	78.7												8.1							83,8		n/a
Tăng tốc+Máy dò+CRF [30]																						83,8		n/a

BẢNG 2

So sánh định lượng SegNet với các phương pháp truyền thống về vấn đề phân đoạn cấp đường CamVid 11 [22]. SegNet vượt trội hơn tất cả các phương pháp khác, kể cả những phương pháp sử dụng độ sâu, video và/hoặc CRF trên phần lớn các lớp. So với các phương pháp dựa trên CRF, dự đoán SegNet chính xác hơn ở 8 trong số 11 lớp. Nó cũng cho thấy sự cải thiện tốt ≈ 10% về độ chính xác trung bình của lớp khi được đào tạo trên một tập dữ liệu lớn gồm các hình ảnh 3,5K. Đặc biệt đáng chú ý là những cải tiến đáng kể về độ chính xác đối với các lớp nhỏ hơn/mỏng hơn. * Lưu ý rằng chúng tôi không thể truy cập dự đoán cho các phương pháp cũ hơn để tính toán chỉ số mIoU, BF.

DeconvNet có các lớp được kết nối đầy đủ (được chuyển thành các lớp tích chập) huấn luyện chậm hơn nhiều và có thời gian chuyển tiếp-lùi tương đương hoặc cao hơn so với SegNet.

Ở đây, chúng tôi cũng lưu ý rằng việc khớp quá mức không phải là vấn đề trong việc đào tạo các mô hình lớn hơn này, vì ở các lần lặp lại tương đương với SegNet, các chỉ số của chúng cho thấy xu hướng ngày càng tăng.

Đối với mô hình FCN, việc học các lớp giải mã thay vì sửa chúng bằng các trọng số nội suy tuyến tính hai chiều, tôi chứng minh hiệu suất, đặc biệt là điểm số BF. Nó cũng đạt được các số liệu cao hơn trong thời gian ngắn hơn nhiều. Thực tế này phù hợp với phân tích trước đó của chúng tôi trong Sec. 3.3.

Đáng ngạc nhiên, DeepLab-LargeFOV được đào tạo để dự đoán nhãn ở độ phân giải 45×60 tạo ra hiệu suất cạnh tranh do đây là mô hình nhỏ nhất về tham số hóa và cũng có thời gian đào tạo nhanh nhất theo Bảng 6. Tuy nhiên, độ chính xác của ranh giới là kém hơn và điều này được chia sẻ bởi các kiến trúc khác. Điểm BF của DeconvNet cao hơn các mạng khác khi được đào tạo trong một thời gian rất dài. Đưa ra phân tích của chúng tôi trong Sec. 3.3 và thực tế là nó chia sẻ kiến trúc kiểu SegNet.

Tác động của quá trình xử lý hậu kỳ CRF dày đặc [63] có thể được nhìn thấy ở thời điểm cuối cùng đối với DeepLab-LargeFOV-denseCRF. Cả toàn cầu và mIoU đều cải thiện nhưng mức trung bình của lớp giảm dần. Tuy nhiên, một sự cải thiện lớn đã đạt được đối với điểm số BF. Lưu ý ở đây rằng các siêu tham số CRF dày đặc đã thu được bằng một quy trình tìm kiếm dạng lưới đắt tiền trên một tập hợp con của tập huấn luyện do không có sẵn tập hợp lệ nào.

4.2 Cảnh trong nhà SUN RGB-D

SUN RGB-D [23] là một tập dữ liệu lớn và đầy thách thức về cảnh trong nhà với 5285 hình ảnh huấn luyện và 5050 hình ảnh thử nghiệm. Các hình ảnh được chụp bởi các cảm biến khác nhau và do đó có nhiều độ phân giải khác nhau. Nhiệm vụ là phân đoạn 37 lớp cảnh trong nhà bao gồm tường, sàn, trần, bàn, ghế, sofa, v.v. Nhiệm vụ này gặp khó khăn bởi thực tế là các lớp đối tượng có nhiều hình dạng, kích cỡ và ở các tư thế khác nhau. Thường xuyên xảy ra hiện tượng che lấp một phần vì thường có nhiều loại khác nhau xuất hiện trong mỗi hình ảnh thử nghiệm. Những yếu tố này làm cho đây trở thành một trong những thách thức phân khúc khó khăn nhất. Chúng tôi chỉ sử dụng phương thức RGB để đào tạo và thử nghiệm. Sử dụng phương thức chiều sâu sẽ đòi hỏi kiến trúc sửa đổi/thiết kế lại [2]. Ngoài ra, chất lượng của hình ảnh độ sâu từ các máy ảnh hiện tại yêu cầu xử lý hậu kỳ cẩn thận để điền vào các phép đo còn thiếu. Chúng cũng có thể yêu cầu sử dụng kết hợp nhiều khung để trích xuất mạnh mẽ các tính năng cho phân đoạn. Do đó, chúng tôi tin rằng việc sử dụng độ sâu để phân đoạn xứng đáng với một nhóm công việc riêng biệt không nằm trong phạm vi của bài viết này. Chúng tôi cũng lưu ý rằng bộ dữ liệu điểm chuẩn trước đó NYUv2 [25] được đưa vào như một phần của bộ dữ liệu này.

Hình ảnh cảnh được có sự thay đổi hạn chế, cả về các lớp quan tâm và sự sắp xếp không gian của chúng. Khi được chụp từ một phương tiện đang di chuyển, vị trí của máy ảnh gần như luôn song song với mặt đường sẽ hạn chế sự thay đổi về điểm xem. Điều này giúp các mạng sâu học cách phân đoạn chúng một cách dễ dàng hơn. Để so sánh, hình ảnh của các cảnh trong nhà phức tạp hơn vì các điểm nhìn có thể khác nhau rất nhiều và ít có tính đều đặn hơn cả về số lớp hiện diện trong một cảnh và sự sắp xếp không gian của chúng. Một khó khăn khác là do kích thước rất khác nhau của các lớp đối tượng trong cảnh. Một số mẫu thử nghiệm từ bộ dữ liệu SUN RGB-D gần đây [23] được hiển thị trong Hình 5. Chúng tôi quan sát thấy một số cảnh có ít lớp lớn và một số cảnh khác có mật độ dày đặc (hàng dưới cùng và bên phải). Hình thức (kết cấu và hình dạng) cũng có thể rất khác nhau ở các cảnh trong nhà. Do đó, chúng tôi tin rằng đây là thách thức khó khăn nhất đối với các kiến trúc và phương pháp phân đoạn trong thị giác máy tính. Các thách thức khác, chẳng hạn như phân đoạn đối tượng nổi bật Pascal VOC12 [21] đã thu hút nhiều nhà nghiên cứu hơn [66], nhưng chúng tôi tin rằng phân đoạn cảnh trong nhà khó khăn hơn và có nhiều ứng dụng thực tế hơn hiện tại như trong AR và robot.

Để khuyến khích nhiều nghiên cứu hơn theo hướng này, chúng tôi đã so sánh các kiến trúc sâu nổi tiếng trên tập dữ liệu SUN RGB-D lớn.

Các kết quả định tính của SegNet trên các mẫu cảnh trong nhà thuộc các loại khác nhau như phòng ngủ, phòng khách, phòng thí nghiệm, phòng họp, phòng tắm được hiển thị trong Hình 5. Chúng tôi thấy rằng SegNet thu được các dự đoán hợp lý khi quy mô của các lớp lớn dưới các điều kiện khác nhau xem điểm. Điều này đặc biệt thú vị vì phương thức đầu vào chỉ là RGB. Hình ảnh RGB cũng hữu ích để phân đoạn các cấu trúc mỏng hơn như chân ghế và bàn, đèn khó đạt được bằng cách sử dụng hình ảnh độ sâu từ các cảm biến hiện có. Có thể thấy điều này từ kết quả của SegNet, DeconvNet trong Hình 5. Việc phân đoạn các đối tượng trang trí như tranh vẽ trên tường cho các tác vụ AR cũng rất hữu ích.

Tuy nhiên so với cảnh ngoài trời thì chất lượng phân đoạn

Mạng/Lặp đi lặp lại	40K				80K				>80K				số lần lặp tối đa
	GC	mIoU	BF	GC	mIoU	BF	GC	mIoU	BF				
SegNet	88.81	59.93	50.02	35.78	89.68	69.82	57.18	42.08	90.40	71.20	60.10	46.84	140K 85.95 60.41
DeepLab-LargeFOV [3]	50.18	26.25	87.76	62.57	53.34	32.04	88.20	62.53	53.88	32.77	140K	89.71	60.67 54.74 40.79
DeepLab-LargeFOV-denseCRF [3]	không đư ợc tính toán									140K	81.97	54.38	46 59 22.86
FCN	82.71	56.22	47.95	24.76	83.27	59.56	49.83	27.99	200K	83.21	56.05	48.68	27.40 83.71 59.64
FCN (deconv đã học) [2]	50.80	31.01	83.14	64.21	51.96	33.18	160K	85.26	46.40	39.69	27.36	85.19	54.08 43.74 29.33
DeconvNet [4]	89.58	70.24	59.77	52.23	260K								

BẢNG 3

So sánh định lư ợng các mạng sâu để phân đoạn ngữ nghĩa trên bộ thử nghiệm CamVid khi đư ợc đào tạo trên kho ngữ liệu gồm 3433 cảnh đư ờng mà không cân bằng lớp. Khi quá trình đào tạo từ đầu đến cuối đư ợc thực hiện với tốc độ học cố định và giống nhau, các mạng nhỏ hơ n như SegNet sẽ học cách hoạt động tốt hơ n trong thời gian ngắn hơ n. Điểm số BF đo lư ờng độ chính xác của việc phân định ranh giới giữa các lớp cao hơ n đáng kể đối với SegNet, DeconvNet so với các mô hình cạnh tranh khác. DeconvNet phù hợp với các số liệu cho SegNet như ng với chi phí tính toán lớn hơ n nhiều. Ngoài ra, hãy xem Bảng 2 để biết độ chính xác của từng lớp đối với SegNet.

Mạng/Lặp đi lặp lại	80K				140K				>140K				số lần lặp tối đa
	GC	mIoU	BF	GC	mIoU	BF	GC	mIoU	BF	70.73	30.82	22.52	9.16 71.66 37.60 27.46
SegNet	11.33	72.63	44.76	31.84	12.66	240K	70.70	41.75	30.67	7.28	71.16	42.71	31.29 7.57 71.90
DeepLab-LargeFOV [3]	42.21	32.08	8.26	240K	66.96	33.06	24.13	9.41	240K	67.31	34.32	24.05	7.88 68.04 37.2
DeepLab-LargeFOV-denseCRF [3]	không đư ợc tính toán									26.33	9.0	68.18	38.41 27.39 9.68
FCN (deconv đã học) [2]	200K	59.62	12.93	8.35	6.50	63.28	22.53	15.14	7.86	66.13	32.28	22.57	10.47 380K
DeconvNet [4]													

BẢNG 4

So sánh định lư ợng các kiến trúc sâu trên bộ dữ liệu SUNRGB-D khi đư ợc đào tạo trên kho dữ liệu gồm 5250 cảnh trong nhà. Lư u ý rằng chỉ phư ơ ng thức RGB đư ợc sử dụng trong các thử nghiệm này. Trong nhiệm vụ phức tạp này với 37 lớp, tất cả các kiến trúc đều hoạt động kém, đặc biệt là do các lớp có kích thư ớc nhỏ hơ n và sai lệch trong phân phối lớp. DeepLab-Large FOV, mô hình nhỏ nhất và hiệu quả nhất có mIoU cao hơ n một chút như ng SegNet có điểm G, C, BF tốt hơ n. Cũng lư u ý rằng khi SegNet đư ợc đào tạo với cân bằng lớp tần số trung bình, nó thu đư ợc 71,75, 44,85, 32,08, 14,06 (180K) làm số liệu.

rõ ràng là ồn ào hơ n. Chất lư ợng giảm đáng kể khi độ lộn xộn tăng lên (xem mẫu kết quả ở cột giữa).

Các kết quả định lư ợng trong Bảng 4 cho thấy rằng tất cả các kiến trúc sâu đều chia sẻ các chỉ số ranh giới và mIoU thấp. Mức trung bình toàn cầu và lớp (tư ơ ng quan tốt với mIou) cũng nhỏ.
SegNet vư ợt trội hơ n tất cả các phư ơ ng pháp khác về chỉ số G, C, BF và có mIoU thấp hơ n một chút so với DeepLab-LargeFOV. Là một thử nghiệm độc lập, chúng tôi đã đào tạo SegNet với cân bằng lớp tần suất trung bình [67] và các số liệu cao hơ n (xem Bảng 4) và điều này phù hợp với phân tích của chúng tôi trong Sec. 3.3. Điều thú vị là, việc sử dụng siêu tham số tối ư u dựa trên tìm kiếm dạng lư ới cho CRF dày đặc đã làm xấu đi tất cả ngoại trừ chỉ số điểm BF cho DeepLab-LargeFOV dày đặcCRF. Có thể tìm thấy nhiều cài đặt tối ư u hơ n như ng quá trình tìm kiếm dạng lư ới quá tốn kém do thời gian suy luận lớn đối với các CRF dày đặc.

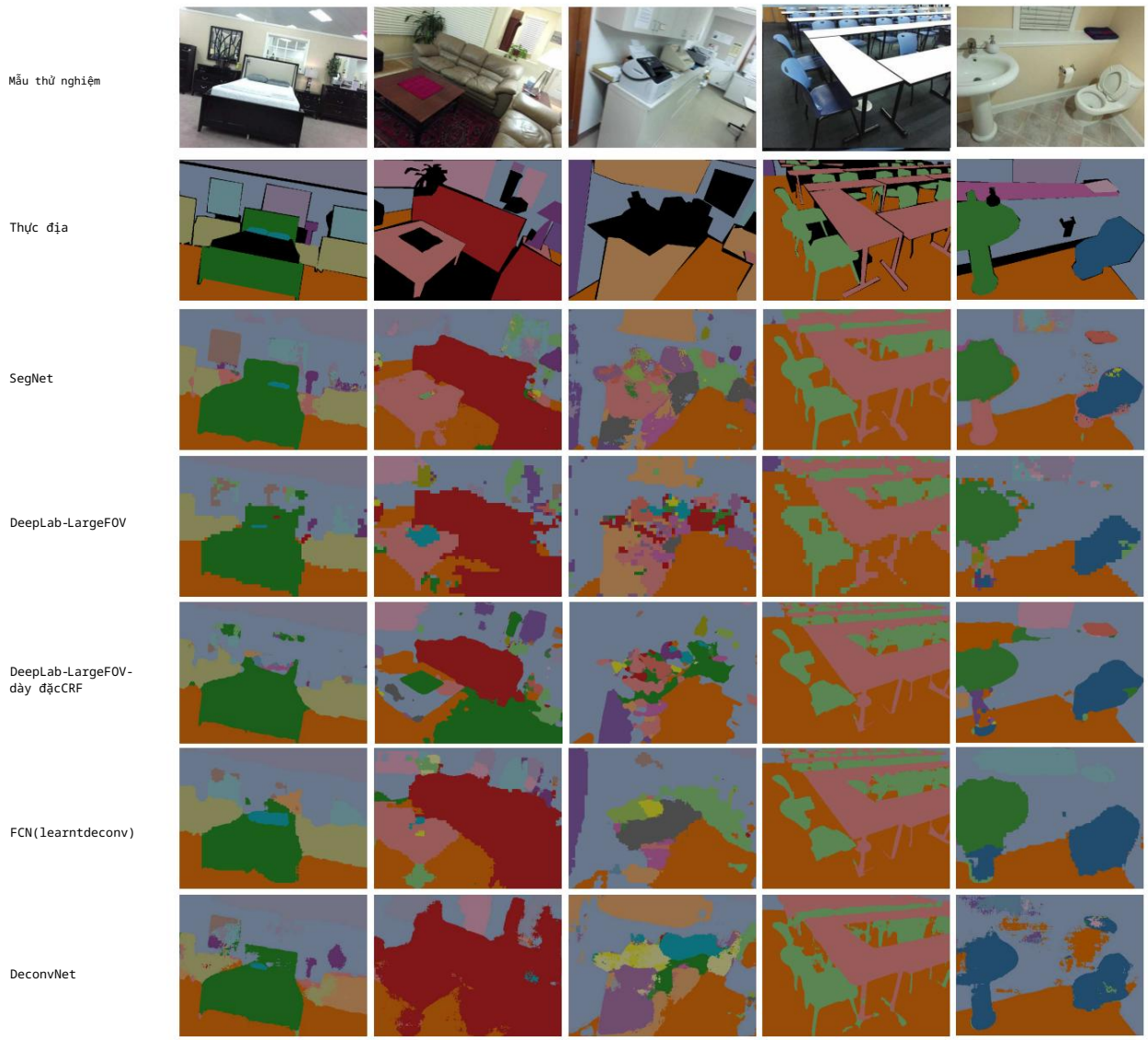
Một lý do cho hiệu suất tổng thể kém là số lư ợng lớn các lớp trong nhiệm vụ phân đoạn này, nhiều lớp trong số đó chiếm một phần nhỏ của hình ảnh và xuất hiện không thư ờng xuyên. Độ chính xác đư ợc báo cáo trong Bảng 5 cho thấy rõ ràng rằng các lớp lớn hơ n có độ chính xác hợp lý và các lớp nhỏ hơ n có độ chính xác thấp hơ n. Điều này có thể đư ợc cải thiện với các bộ dữ liệu có kích thư ớc lớn hơ n và các kỹ thuật đào tạo nhận biết phân phối lớp. Một lý do khác dẫn đến hiệu suất kém có thể nằm ở việc các kiến trúc sâu này (tất cả đều dựa trên kiến trúc VGG [6]) không có khả năng thay đổi lớn trong các cảnh trong nhà. Phỏng đoán này về phần chúng tôi dựa trên thực tế là mô hình nhỏ nhất mà DeepLab-LargeFOV tạo ra độ chính xác tốt nhất về mIoU và khi so sánh, các thông số hóa lớn hơ n trong DeconvNet, FCN không cải thiện hiệu suất ngay cả khi đào tạo lâu hơ n nhiều (DeconvNet). Điều này cho thấy có thể có một lý do chung cho hiệu suất kém trên tất cả các kiến trúc. Cần có nhiều bộ dữ liệu đư ợc kiểm soát hơ n [68] để xác minh giả thuyết này.

5 BÀN LUẬN VÀ CÔNG VIỆC TƯ Ơ NG LAI

Các mô hình học sâu thư ờng đạt đư ợc thành công ngày càng tăng do có sẵn các bộ dữ liệu lớn và mở rộng độ sâu và tham số hóa của mô hình. Tuy nhiên, trong thực tế, các yếu tố như bộ nhớ và thời gian tính toán trong quá trình đào tạo và thử nghiệm là những yếu tố quan trọng cần xem xét khi chọn một mô hình từ một ngân hàng lớn các mô hình. Thời gian đào tạo trở thành một cân nhắc quan trọng đặc biệt khi hiệu suất đạt đư ợc không tư ơ ng xứng với thời gian đào tạo tăng lên như thể hiện trong các thí nghiệm của chúng tôi. Bộ nhớ thời gian thử nghiệm và tải tính toán rất quan trọng để triển khai các mô hình trên các thiết bị nhúng chuyên dụng, chẳng hạn như trong các ứng dụng AR.
Từ quan điểm hiệu quả tổng thể, chúng tôi cảm thấy các mô hình nhỏ hơ n và nhiều bộ nhớ hơ n, tiết kiệm thời gian hơ n cho các ứng dụng thời gian thực như hiểu cảnh đư ờng và AR đã ít đư ợc chú ý hơ n.
Đây là động lực chính đằng sau đề xuất SegNet, nhỏ hơ n và nhanh hơ n đáng kể so với các kiến trúc cạnh tranh khác, như ng chúng tôi đã chứng minh là có hiệu quả đối với các tác vụ như tìm hiểu cảnh đư ờng.

Các thử thách phân đoạn như Pascal [21] và MS-COCO [42] là các thử thách phân đoạn đối tư ợng trong đó một vài lớp có mặt trong bất kỳ hình ảnh thử nghiệm nào. Việc phân đoạn cảnh gặp nhiều thách thức hơ n do tính đa dạng cao của các cảnh trong nhà và nhu cầu phân đoạn một số lư ợng lớn các lớp cùng một lúc. Nhiệm vụ phân đoạn cảnh ngoài trời và trong nhà cũng đư ợc định hư ớng thực tế hơ n với các ứng dụng hiện tại như lái xe tự động, robot và AR.

Các chỉ số mà chúng tôi đã chọn để đánh giá các kiến trúc phân đoạn sâu khác nhau như thư ớc đo F1 ranh giới (BF) đã đư ợc thực hiện để bổ sung cho các chỉ số hiện có thiên về độ chính xác của khu vực. Rõ ràng từ các thử nghiệm của chúng tôi và các điểm chuẩn độc lập khác [62] rằng hình ảnh cảnh ngoài trời đư ợc chụp từ một chiếc ô tô đang di chuyển sẽ dễ dàng phân đoạn hơ n và kiến trúc sâu hoạt động mạnh mẽ. Chúng tôi hy vọng các thí nghiệm của chúng tôi sẽ khuyến khích các nhà nghiên cứu thu hút sự chú ý của họ vào bối cảnh trong nhà đầy thách thức hơ n



Hình 5. Đánh giá định tính các dự đoán SegNet trên các cảnh thử nghiệm RGB trong nhà từ bộ dữ liệu SUN RGB-D đư ợc phát hành gần đây [23]. Trong thử thách khó khăn này, các dự đoán của SegNet phân định rõ ranh giới giữa các lớp cho các lớp đối tư ợng trong nhiều cảnh và quan điểm của chúng. Chất lư ợng phân đoạn tổng thể tốt hơ n khi các lớp đối tư ợng có kích thư ớc hợp lý như ng lại rất nhiễu khi cảnh lộn xộn hơ n. Lư u ý rằng thư ờng các phần của hình ảnh của một cảnh không có nhãn sự thật cơ bản và chúng đư ợc hiển thị bằng màu đen. Những phần này không đư ợc che dấu trong các dự đoán mô hình sâu tư ơ ng ứng đư ợc hiển thị. Lư u ý rằng những kết quả này tư ơ ng ứng với mô hình tư ơ ng ứng với độ chính xác mIoU cao nhất trong Bảng 4.

nhiệm vụ phân vùng.

Một lựa chọn quan trọng mà chúng tôi phải thực hiện khi đánh giá các kiến trúc sâu khác nhau của các tham số hóa khác nhau là cách huấn luyện chúng. Nhiều kiến trúc trong số này đã sử dụng một loạt các kỹ thuật hỗ trợ và công thức đào tạo nhiều giai đoạn để đạt đư ợc độ chính xác cao trên các bộ dữ liệu như ng điều này gây khó khăn cho việc thu thập bằng chứng về hiệu suất thực sự của chúng trong điều kiện thời gian và bộ nhớ hạn chế. Thay vào đó, chúng tôi đã chọn thực hiện đo điểm chuẩn có kiểm soát trong đó chúng tôi sử dụng chuẩn hóa hàng loạt để cho phép đào tạo từ đầu đến cuối với cùng một bộ giải (SGD). Tuy nhiên, chúng tôi lư u ý rằng cách tiếp cận này không thể giải quyết hoàn toàn các tác động của mô hình so với bộ giải (tối ư u hóa) trong việc đạt đư ợc một kết quả cụ thể. Điều này chủ yếu là do việc đào tạo các mạng này liên quan đến truyền ngư ợc gradient không hoàn hảo và việc tối ư u hóa là một vấn đề không lồi trong các kích thư ớc cực lớn. Thừa nhận những thiếu sót này, chúng tôi hy vọng rằng phân tích có kiểm soát này sẽ bổ sung cho các tiêu chuẩn khác [62] và tiết lộ những đánh đổi thực tế liên quan đến các kiến trúc nổi tiếng khác nhau.

Trong tư ơ ng lai, chúng tôi muốn khai thác hiểu biết của mình về kiến trúc phân khúc đư ợc thu thập từ phân tích của mình để thiết kế kiến trúc hiệu quả hơ n cho các ứng dụng thời gian thực. Chúng tôi cũng quan tâm đến việc ư ớc tính độ không đảm bảo của mô hình cho các dự đoán từ kiến trúc phân đoạn sâu [69], [70].

6 KẾT LUẬN Chúng tôi

đã trình bày SegNet, một kiến trúc mạng tích chập sâu dành cho phân đoạn ngữ nghĩa. Động lực chính đằng sau SegNet là nhu cầu thiết kế một kiến trúc hiệu quả để hiểu cảnh đư ờng xá và trong nhà, hiệu quả cả về bộ nhớ và thời gian tính toán. Chúng tôi đã phân tích SegNet và so sánh nó với các biến thể quan trọng khác để tiết lộ những đánh đổi thực tế liên quan đến việc thiết kế kiến trúc để phân đoạn, đặc biệt là thời gian đào tạo, bộ nhớ so với độ chính xác. Những kiến trúc đó

	Tầng	Tủ 63.37		Ghế	Sofa	Bảng	Cửa sổ Giá sách Hình ảnh truy cập Mành						
Tư ơng	93.43	Rèm	Giư ờng 73,18	75.92	59.57	64.18	52.50 57.51 42.05 56.17 37.66 40.29 Thảm trải sàn Quần áo						
83,42	Kệ 11.45	66.56	tủ quần áo	Gối	Gư ơ ng	Trần Sách Tủ lạnh Tivi Giấy							
Bàn 11,92			52,73	43.80	26.30	0.00	34.31	74.11	53,77	29,85 33,76 22,73			
Khăn tắm Màn tắm Hộp Bảng trắng Ngư ời đứng ban đêm 19,83 0,03 27,27						Nhà vệ	Túi đựng bóng đèn bồn tắm 58.10 35.27						
		23.14	60,25		29,88	sinh 76.00	48.86 16.76						

BẢNG 5 Độ

chính xác trung bình của lớp dự đoán SegNet cho 37 lớp cảnh trong nhà trong bộ dữ liệu điểm chuẩn SUN RGB-D. Hiệu suất tư ơ ng quan tốt với kích thư ớc của các lớp học trong các cảnh trong nhà. Lư u ý rằng độ chính xác trung bình của lớp có mối tư ơ ng quan chặt chẽ với chỉ số mIoU.

Mạng	Chuyển tiếp (ms) Chuyển tiếp (ms) Bộ nhớ đào tạo GPU (MB) Bộ nhớ suy luận GPU (MB) Kích thư ớc mô hình (MB)				
SegNet	422.50	488.71	6803	1052	117
DeepLab-LargeFOV [3]	110.06	160.73	5618	1993	83
FCN (deconv đã học) [2]	317.09	484.11	9735	1806	539
DeconvNet [4]	474.65	602.15	9731	1872	877

BẢNG 6 So

sánh thời gian tính toán và tài nguyên phần cứng cần thiết cho các kiến trúc sâu khác nhau. Lệnh caffe time đư ợc sử dụng để tính toán yêu cầu thời gian trung bình trên 10 lần lặp lại với kích thư ớc lô nhỏ 1 và hình ảnh có độ phân giải 360 × 480. Chúng tôi đã sử dụng lệnh nvidia-smi unix để tính toán mức tiêu thụ bộ nhớ. Để đào tạo tính toán bộ nhớ, chúng tôi đã sử dụng một lô nhỏ có kích thư ớc 4 và đối với bộ nhớ suy luận, kích thư ớc lô là 1. Kích thư ớc mô hình là kích thư ớc của các mô hình caffe trên đĩa. SegNet là bộ nhớ hiệu quả nhất trong mô hình suy luận.

lư u trữ các bản đồ tính năng mạng bộ mã hóa hoạt động tốt nhất như ng tiêu tốn nhiều bộ nhớ hơ n trong thời gian suy luận. Mặt khác, SegNet hiệu quả hơ n vì nó chỉ lư u trữ các chỉ số tổng hợp tối đa của bản đồ tính năng và sử dụng chúng trong mạng bộ giải mã của nó để đạt đư ợc hiệu suất tốt. Trên các bộ dữ liệu lớn và nổi tiếng, SegNet hoạt động cạnh tranh, đạt đư ợc điểm số cao về khả năng hiểu biết về cảnh đư ờng. Việc học kiến trúc phân đoạn sâu từ đầu đến cuối là một thách thức khó khăn hơ n và chúng tôi hy vọng sẽ nhận đư ợc nhiều sự chú ý hơ n cho vấn đề quan trọng này.

NGƯ ỜI GIỚI THIỆU

[1] K. Simonyan và A. Zisserman, “Mạng tích chập rất sâu để nhận dạng hình ảnh quy mô lớn,” bản in trư ớc của arXiv arXiv:1409.1556, 2014.

[2] J. Long, E. Shelhamer và T. Darrell, “Mạng tích chập hoàn toàn cho phân đoạn ngữ nghĩa,” trong CVPR, trang 3431–3440, 2015.

[3] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy và A. Yuille, “Phân đoạn hình ảnh theo ngữ nghĩa với mạng tích chập sâu và crfs đư ợc kết nối đầy đủ,” trong ICLR, 2015.

[4] H. Noh, S. Hong và B. Han, “Học giải mã mạng cho phân đoạn ngữ nghĩa,” trong ICCV, trang 1520–1528, 2015.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, và A. Rabinovich, “Đi sâu hơ n với kết chập,” trong CVPR, trang 1–9, 2015.

[6] K. Simonyan và A. Zisserman, “Mạng tích chập rất sâu để nhận dạng hình ảnh quy mô lớn,” CoRR, tập. abs/1409.1556, 2014.

[7] C. Farabet, C. Couprie, L. Najman và Y. LeCun, “Học các tính năng phân cấp để ghi nhãn cảnh,” IEEE PAMI, tập. 35, không. 8, trang 1915–1929, 2013.

[8] N. Hft, H. Schulz và S. Behnke, “Phân đoạn ngữ nghĩa nhanh của các cảnh rgb-d với mạng thần kinh sâu đư ợc tăng tốc gpu,” trong KI 2014: Những tiến bộ trong Trí tuệ nhân tạo (C. Lutz và M. Thielscher , eds.), tập. 8736 của Bài giảng về Khoa học Máy tính, trang 80–85, Nhà xuất bản Quốc tế Springer, 2014.

[9] R. Socher, CC Lin, C. Manning và AY Ng, “Phân tích cảnh tự nhiên và ngôn ngữ tự nhiên bằng mạng nơ -ron đệ quy,” trong ICML, trang 129–136, 2011.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang và PH Torr, “Các trư ờng ngẫu nhiên có điều kiện dư ới dạng mạng thần kinh hồi quy,” trong Kỷ yếu của Hội nghị Quốc tế IEEE về Tầm nhìn Máy tính, trang 1529–1537, 2015.

[11] W. Liu, A. Rabinovich, và AC Berg, “Parsenet: Nhìn rộng hơ n để thấy tốt hơ n,” CoRR, tập. abs/1506.04579, 2015.

[12] V. Badrinarayanan, A. Handa và R. Cipolla, “Segnet: Kiến trúc bộ giải mã-mã hóa vòng xoắn sâu để ghi nhãn theo pixel ngữ nghĩa mạnh mẽ,” CoRR, tập. abs/1505.07293, 2015.

[13] D. Eigen và R. Fergus, “Dự đoán độ sâu, quy tắc bề mặt và nhãn ngữ nghĩa với kiến trúc tích chập đa quy mô phổ biến,” trong ICCV, trang 2650–2658, 2015.

[14] G. Papandreou, L.-C. Chen, K. Murphy, và AL Yuille, “Học tập có giám sát yếu và bán giám sát của một dcnn cho phân đoạn hình ảnh ngữ nghĩa,” bản in trư ớc của arXiv arXiv:1502.02734, 2015.

[15] F. Yu và V. Koltun, “Tập hợp ngữ cảnh đa quy mô bằng các cụm từ đư ợc mở rộng,” bản in sẵn arXiv arXiv:1511.07122, 2015.

[16] O. Ronneberger, P. Fischer, và T. Brox, “U-net: Mạng tích chập cho phân đoạn hình ảnh y sinh,” trong MICCAI, trang 234–241, Springer, 2015.

[17] L. Bottou, “Học máy quy mô lớn với độ dốc ngẫu nhiên của mùi,” trong Kỷ yếu của COMPSTAT'2010, trang 177–186, Springer, 2010.

[18] S. Hong, H. Noh và B. Han, “Mạng nơ -ron sâu đư ợc tách rời để phân đoạn ngữ nghĩa bán giám sát,” trong NIPS, trang 1495–1503, 2015.

[19] M. Ranzato, FJ Huang, Y. Boureau và Y. LeCun, “Học tập không giám sát về hệ thống phân cấp tính năng bất biến với các ứng dụng để nhận dạng đối tư ợng,” trong CVPR, 2007.

[20] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Ur tasun, và cộng sự, “Vai trò của ngữ cảnh đối với phát hiện đối tư ợng và phân đoạn ngữ nghĩa trong tự nhiên,” trong Thị giác máy tính và Nhận dạng mẫu (CVPR), Hội nghị IEEE 2014 trên, trang 891– 898, IEEE, 2014.

[21] M. Everingham, SA Eslami, L. Van Gool, CK Williams, J. Winn, và A. Zisserman, “Thách thức các lớp đối tư ợng trực quan pascal: Nhìn lại,” Tạp chí Quốc tế về Thị giác Máy tính, tập. 111, không. 1, trang 98–136.

[22] G. Brostow, J. Fauqueur và R. Cipolla, “Các lớp đối tư ợng ngữ nghĩa trong video: Cơ sở dữ liệu sự thật cơ bản độ nét cao,” PRL, tập. 30(2), trang 88–97, 2009.

[23] S. Song, SP Lichtenberg, và J. Xiao, “Sun rgb-d: Bộ điểm chuẩn hiểu bối cảnh rgb-d,” trong Kỷ yếu của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 567–576, 2015.

[24] CL Zitnick và P. Dollar, “Các hộp cạnh: Định vị đề xuất đối tư ợng từ các cạnh,” trong Computer Vision–ECCV 2014, trang 391–405, Springer, 2014.

[25] N. Silberman, D. Hoiem, P. Kohli và R. Fergus, “Phân đoạn trong nhà và suy luận hỗ trợ từ hình ảnh rgbd,” trong ECCV, trang 746–760, Springer, 2012.

[26] A. Geiger, P. Lenz và R. Urtasun, “Chúng ta đã sẵn sàng cho xe tự hành chư a? bộ chuẩn tầm nhìn KITTI,” trong CVPR, trang 3354–3361, 2012.

[27] J. Shotton, M. Johnson và R. Cipolla, “Rừng văn bản ngữ nghĩa để phân loại và phân đoạn hình ảnh,” trong CVPR, 2008.

[28] G. Brostow, J. Shotton, J. và R. Cipolla, “Phân đoạn và nhận dạng bằng cách sử dụng cấu trúc từ các đám mây điểm chuyển động,” trong ECCV, Marseille, 2008.

[29] P. Sturgess, K. Alahari, L. Ladicky, và PHSTorr, “Kết hợp hình thức và cấu trúc từ các đặc điểm chuyển động để hiểu cảnh đư ờng,” trong BMVC, 2009.

[30] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, và PHS Torr, “Cái gì, ở đâu và bao nhiêu? kết hợp máy dò đối tư ợng và crfs,” trong ECCV, trang 424–437, 2010.

[31] P. Kontschieder, SR Bulo, H. Bischof, và M. Pelillo, “Nhãn lớp có cấu trúc trong các khu rừng ngẫu nhiên để ghi nhãn hình ảnh ngữ nghĩa,” trong ICCV, trang 2190–2197, IEEE, 2011.

[32] C. Zhang, L. Wang và R. Yang, “Phân đoạn ngữ nghĩa của cảnh đô thị sử dụng bản đồ độ sâu dày đặc,” trong ECCV, trang 708–721, Springer, 2010.

[33] J. Tighe và S. Lazebnik, “Superparsing,” IJCV, tập. 101, không. 2, trang 329–349, 2013.

[34] X. Ren, L. Bo và D. Fox, “Gắn nhãn cảnh Rgb-(d): Tính năng và thuật toán,” trong CVPR, trang 2759–2766, IEEE, 2012.

[35] A. Hermans, G. Floros và B. Leibe, “Bản đồ ngữ nghĩa 3D dày đặc của cảnh trong nhà từ hình ảnh RGB-D,” trong ICRA, 2014.

[36] S. Gupta, P. Arbelaez và J. Malik, “Tổ chức tri giác và nhận dạng cảnh trong nhà từ hình ảnh rgb-d,” trong CVPR, trang 564–571, IEEE, 2013.

[37] C. Farabet, C. Couprie, L. Najman và Y. LeCun, “Phân tích cú pháp cảnh với tính năng học tập tính năng đa tỷ lệ, cây thuần khiết và bìa tối ưu,” trong ICML, 2012.

[38] D. Grangier, L. Bottou và R. Collobert, “Mạng tích chập sâu để phân tích cảnh,” trong Hội thảo ICML về Học sâu, 2009.

[39] C. Gatta, A. Romero và J. van de Weijer, “Không kiểm soát phản hồi ngữ nghĩa từ trên xuống lặp đi lặp lại trong các mạng sâu tích chập,” trong Hội thảo CVPR về Tầm nhìn sâu, 2014.

[40] P. Pinheiro và R. Collobert, “Mạng nơ -ron tích chập tái phát để ghi nhãn cảnh,” trong ICML, trang 82–90, 2014.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg và L. Fei -Fei, “Thử thách nhận dạng hình ảnh quy mô lớn ImageNet,” Tạp chí quốc tế về thị giác máy tính (IJCV), trang 1–42, tháng 4 năm 2015.

[42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, và CL Zitnick, “Microsoft coco: Common objects in context,” trong Computer Vision–ECCV 2014, trang 740 –755, Mùa xuân, 2014.

[43] AG Schwing và R. Urtasun, “Mạng có cấu trúc sâu đư ợc kết nối đầy đủ hoạt động,” bản in trư ớc của arXiv arXiv:1503.02351, 2015.

[44] G. Lin, C. Shen, I. Reid, và cộng sự, “Đào tạo từng phần hiệu quả các mô hình có cấu trúc sâu cho phân đoạn ngữ nghĩa,” in lại arXiv arXiv:1504.01013, 2015.

[45] B. Hariharan, P. Arbelaez, R. Girshick và J. Malik, “Hypercolumns for object segmentation and fine-grained localization,” trong CVPR, trang 447–456, 2015.

[46] M. Mostajabi, P. Yadollahpour và G. Shakhnarovich, “Phân đoạn ngữ nghĩa chuyển tiếp về phía trư ớc với các tính năng thu nhỏ,” trong Kỷ yếu của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 3376–3385, 2015.

[47] MD Zeiler, D. Krishnan, GW Taylor, và R. Fergus, “Mạng giải chập,” trong CVPR, trang 2528–2535, IEEE, 2010.

[48] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu và Y. LeCun, “Học hệ thống phân cấp tính năng tích chập để nhận dạng hình ảnh,” trong NIPS, trang 1090–1098, 2010.

[49] C. Dong, C. C. Loy, K. He và X. Tang, “Học một mạng tích chập sâu cho hình ảnh siêu phân giải,” trong ECCV, trang 184–199, Springer, 2014.

[50] D. Eigen, C. Puhrsch và R. Fergus, “Dự đoán bản đồ độ sâu từ một hình ảnh duy nhất sử dụng mạng sâu nhiều tỷ lệ,” trong NIPS, trang 2366–2374, 2014.

[51] S. Ioffe và C. Szegedy, “Chuẩn hóa hàng loạt: Tăng tốc đào tạo mạng sâu bằng cách giảm sự thay đổi đồng biến nội bộ,” CoRR, tập. abs/1502.03167, 2015.

[52] V. Badrinarayanan, B. Mishra, và R. Cipolla, “Hiểu nghĩa cố gắng trong các mạng sâu,”

[53] H. Noh, S. Hong và B. Han, “Học mạng giải chập cho phân đoạn ngữ nghĩa,” CoRR, tập. abs/1505.04366, 2015.

[54] K. Jarrett, K. Kavukcuoglu, M. Ranzato, và Y. LeCun, “Cấu trúc nhiều tầng tốt nhất để nhận dạng đối tư ợng là gì?”, trong ICCV, trang 2146–2153, 2009.

[55] K. He, X. Zhang, S. Ren và J. Sun, “Đi sâu vào bộ chỉnh lư u: Vư ợt qua hiệu suất cấp độ con ngư ời trong phân loại imagenet,” trong ICCV, trang 1026–1034, 2015.

[56] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama và T. Darrell, “Caffe: Kiến trúc tích chập để nhúng tính năng nhanh,” trong Kỷ yếu của hội nghị quốc tế ACM lần thứ 22 về Đa phư ơng tiện, trang 675–678, ACM, 2014.

[57] G. Csurka, D. Larlus, F. Perronnin và F. Meylan, “Đâu là thư ớc đo đánh giá tốt cho phân đoạn ngữ nghĩa?.,” trong BMVC, 2013.

[58] J. Long, E. Shelhamer và T. Darrell, “Mạng chập hoàn toàn cho phân đoạn ngữ nghĩa,” trong https://arxiv.org/pdf/1605.06211v1.pdf, 2016.

[59] DR Martin, CC Fowlkes và J. Malik, “Học cách phát hiện ranh giới hình ảnh tự nhiên bằng cách sử dụng các tín hiệu về độ sáng, màu sắc và kết cấu cục bộ,” Giao dịch của IEEE về phân tích mẫu và trí thông minh của máy, tập. 26, không. 5, trang 530–549, 2004.

[60] S. Gould, R. Fulton, và D. Koller, “Phân tách cảnh thành các vùng nhất quán về mặt hình học và ngữ nghĩa,” trong ICCV, trang 1–8, IEEE, 2009.

[61] BC Russell, A. Torralba, KP Murphy, và WT Freeman, “Labelme: cơ sở dữ liệu và công cụ dựa trên web để chú thích hình ảnh,” IJCV, tập. 77, không. 1-3, trang 157–173, 2008.

[62] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, và B. Schiele, “Bộ dữ liệu cảnh quan thành phố để hiểu bối cảnh đô thị theo ngữ nghĩa, ” bản in trư ớc arXiv arXiv:1604.01685, 2016.

[63] V. Koltun, “Suy luận hiệu quả trong các crf đư ợc kết nối đầy đủ với cạnh gaussian tiềm năng,” trong Trong: NIPS (2011, 2011.

[64] Bulo, S. Rota và P. Kontschieder, “Rừng quyết định thần kinh để ghi nhãn hình ảnh ngữ nghĩa.,” trong CVPR, 2014.

[65] Y. Yang, Z. Li, L. Zhang, C. Murphy, J. Ver Hoeve, và H. Jiang, “Bộ mô tả nhãn cục bộ cho ví dụ về ghi nhãn hình ảnh dựa trên ngữ nghĩa,” trong ECCV, trang 361–375, Mùa xuân, 2012.

[66] Z. Liu, X. Li, P. Luo, C.-C. Loy và X. Tang, “Phân đoạn hình ảnh ngữ nghĩa thông qua mạng phân tích cú pháp sâu,” trong Kỷ yếu của Hội nghị Quốc tế IEEE về Tầm nhìn Máy tính, trang 1377–1385, 2015.

[67] D. Eigen và R. Fergus, “Dự đoán độ sâu, quy tắc bề mặt và nhãn ngữ nghĩa với kiến trúc tích chập đa quy mô phổ biến,” in lại arXiv arXiv:1411.4734, 2014.

[68] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent và R. Cipolla, “Scenenet: Hiểu các cảnh trong nhà trong thế giới thực bằng dữ liệu tổng hợp,” trong CVPR, 2016.

[69] Y. Gal và Z. Ghahramani, “Dropout as a xấp xỉ bayesian: Thông tin chi tiết và ứng dụng,” trong Hội thảo Deep Learning, ICML, 2015.

[70] A. Kendall, V. Badrinarayanan và R. Cipolla, “Bayesian segnet: Mô hình không chắc chắn trong kiến trúc bộ giải mã-mã hóa xoắn sâu để hiểu cảnh,” bản in trư ớc của arXiv arXiv:1511.02680, 2015.

Vijay Badrinarayanan lấy bằng Tiến sĩ tại INRIA Rennes, Pháp vào năm 2009. Ông là cộng tác viên nghiên cứu sau tiến sĩ cấp cao tại Phòng thí nghiệm Ma chine Intelligence, Khoa Kỹ thuật, Đại học Cambridge, Vư ơng quốc Anh. Ông hiện là Kỹ sư chính, Học sâu tại Magic Leap, Inc. ở Mountain View, CA. Mối quan tâm nghiên cứu của anh ấy là về các mô hình đồ họa xác suất, học sâu áp dụng cho các vấn đề về nhận thức dựa trên hình ảnh và video.

Alex Kendall tốt nghiệp Cử nhân Kỹ thuật với Hạng Nhất năm 2013 tại Đại học Auckland, New Zealand. Năm 2014, anh đư ợc trao học bổng Woolf Fisher Scholar để theo học Tiến sĩ tại Đại học Cambridge, Vư ơng quốc Anh. Anh là thành viên của Phòng thí nghiệm Ma chine Intelligence và quan tâm đến các ứng dụng học sâu cho ngư ời máy di động.

Roberto Cipolla lấy bằng Cử nhân (Kỹ sư) của Đại học Cambridge năm 1984, bằng MSE (Kỹ thuật Điện) của Đại học Pennsylvania năm 1985 và bằng D.Phil. (Computer Vision) từ Đại học Oxford năm 1991. Từ năm 1991-92 là thành viên Toshiba và kỹ sư tại Trung tâm Nghiên cứu và Phát triển Tập đoàn Toshiba ở Kawasaki, Nhật Bản. Ông gia nhập Khoa Kỹ thuật, Đại học Cambridge vào năm 1992 với tư cách là Giảng viên và là thành viên của Đại học Jesus. Ông trở thành Độc giả về Kỹ thuật Thông tin năm 1997 và Giáo sư năm 2000.

Anh ấy trở thành thành viên của Học viện Kỹ thuật Hoàng gia (FREng) vào năm 2010. Sở thích nghiên cứu của anh ấy là về thị giác máy tính và ngư ời máy. Ông là tác giả của 3 cuốn sách, biên tập 9 tập và là đồng tác giả của hơ n 300 bài báo.