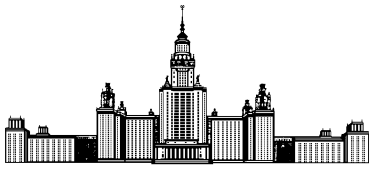
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

ЭССЕ СТУДЕНТА 317 ГРУППЫ

«Сегментация изображений»

Выполнил:
студент 3 курса 317 группы
Афанасьев Глеб Ильич

Москва, 2020

Содержание

Аннотация

Задача сегментации изображений является одной из ключевых не только в области компьютерного зрения, но и в области анализа изображений в целом. С использованием нейросетевых технологий в этой области начались существенные продвижения в точности.

В данной работе будут сначала рассмотрены общие архитектуры и принципы работы нейронных сетей, решающих данную задачу, а во второй части будут приведены несколько архитектур, позволяющих решать данную задачу с хорошей скоростью.

1 Введение

Для начала разберемся с постановкой задачи сегментации. По сути, она является логическим продолжением задачи детекции. Если во втором случае требовалось найти минимальный прямоугольник, содержащий искомый объект и определить его класс, то в задаче сегментации вместо нахождения прямоугольника требуется каждому пикселю присвоить метку того класса, к объекту которого он принадлежит. Поэтому, зачастую, в ней используются все те же функционалы качества.

Рис. 1: Пример решения задачи сегментации.[1]

Выделяют 2 типа сегментации: семантическая сегментация и сегментация объектов. В первом задачей является отделить объекты от фона, во втором-отделить объекты от фона и друг от друга. И та и другая задача весьма трудоемкие и зачастую, обучение подобных нейросетей длится неделями, а предсказания могут занять несколько минут, что уже не подходит для систем реального времени. Далее, мы, постепенно продвигаясь от ранних предложений, рассмотрим некоторые архитектуры, решающие задачу сегментации и в конечном счете дойдем до сетей, подходящих для их применения в системах реального времени.

2 FCN и FastFCN

Для того чтобы понять основные идеи, начнем с рассмотрения самых простых архитектур для сегментации изображений.

2.1 Идея использования сверточных нейросетей в задаче сегментации. FCN

Идея использования сверточных нейросетей проста. В задаче сегментации входными и выходными данными являются тензоры, и в операциях свертки, пулинга и функций нелинейности также и входными и выходными данными являются тензоры(рис 2).

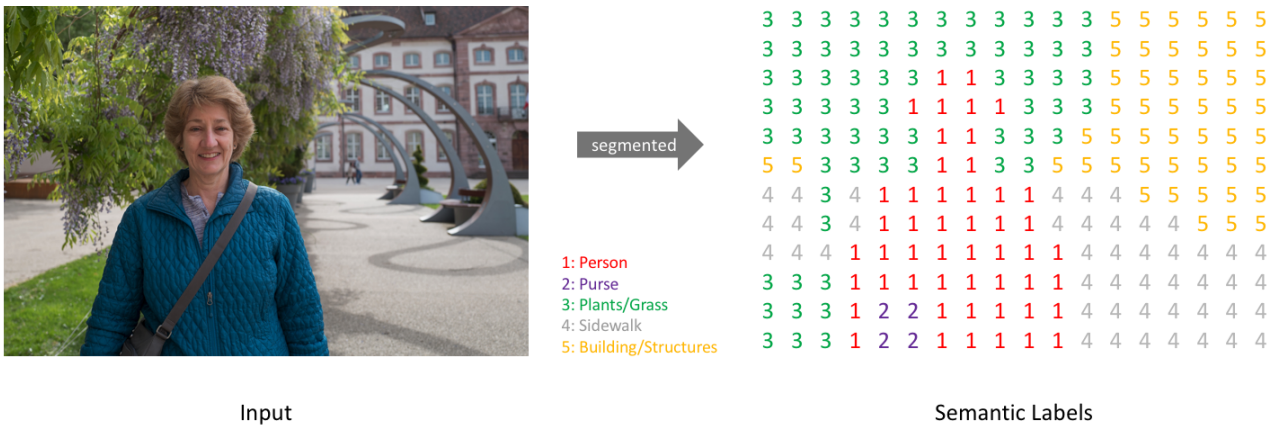


Рис. 2: Пример входных и выходных данных. Слева-трехмерный тензор, справа-двумерный(в некоторых случаях представляется как трехмерный с 5 каналами).[2]

Предлагается решать задачу путем последовательного применения нескольких сверточных слоев и получения на выходе необходимого тензора. Однако, так как свертки зачастую уменьшают размер изображения, то для успешного применения подобного подхода потребуется так называемый upsampling, то есть операция, увеличивающая размер изображения. Таких операций существует несколько типов и выполняются они очень просто(рис 3).

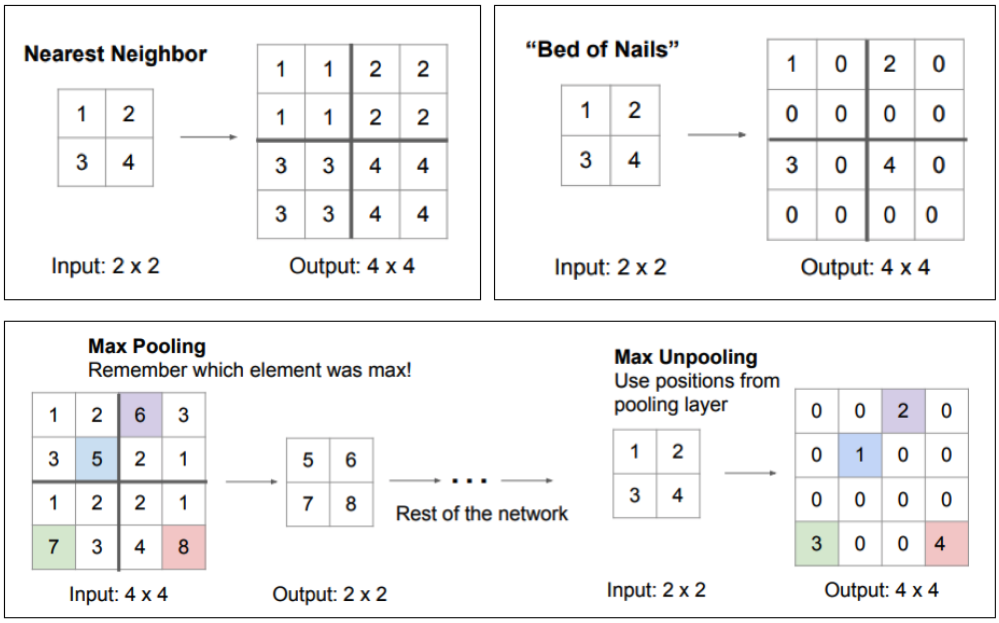


Рис. 3: Операции upsampling.[2]

Также, помимо примитивных операций, есть и более сложная. Так называемый upconvolution, который с помощью ядра также увеличивает размер изображения(рис 4).

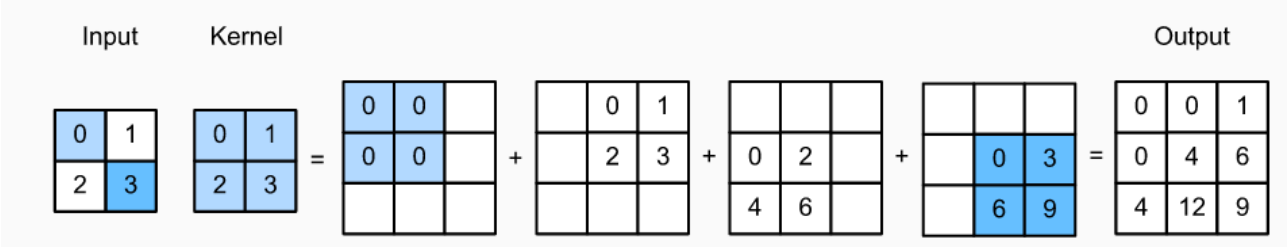


Рис. 4: Upconvolution.[3]

Итак, с введенными выше операциями, архитектура нейросетей, решающих задачу сегментации, будет иметь так называемую encoder/decoder структуру(рис 5).

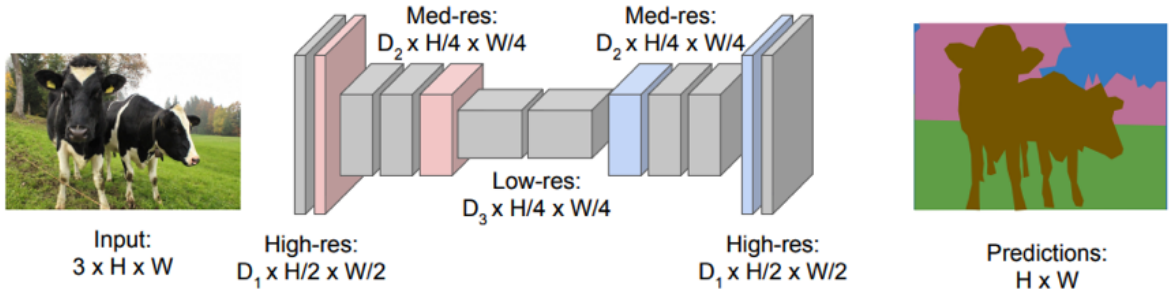


Рис. 5: Примерная архитектура нейросети, решающей задачу сегментации. Первая половина называется encoder, вторая-decoder.[2]

2.2 FastFCN

Часто для восстановления размера, помимо операций upsampling, используют еще и информацию со слоев encoder, что приводит к повышению качества и увеличению вычислительной сложности. Попытка сделать более компактную и вычислительно менее затратную архитектуру предпринята в FastFCN[4]. Ее идея заключается в том, чтобы взять последние 3 сверточных уровня простой FCN и пропустить через модуль, под названием JPU(Joint Pyramid Upsampling)(рис 6).

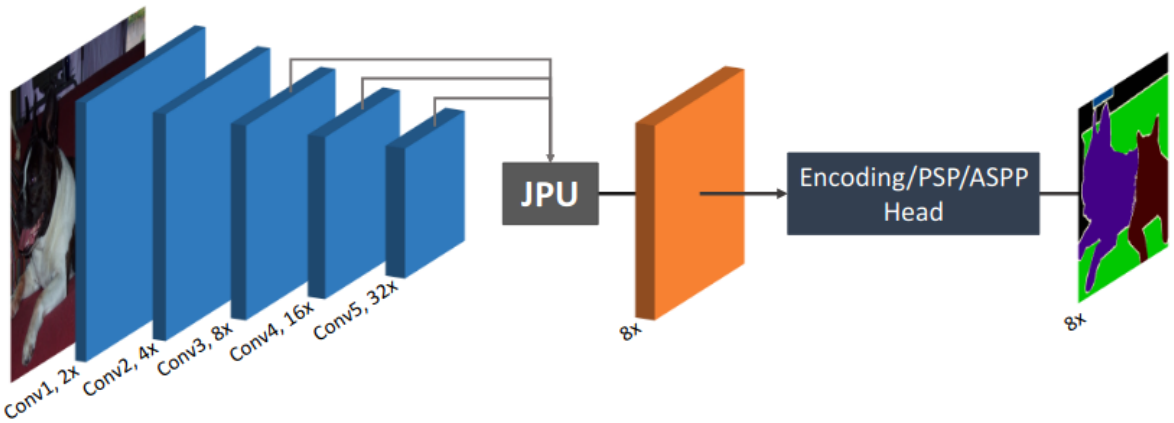


Рис. 6: Архитектура FFCN[4]

Данный модуль устроен следующим образом. Сначала 3 поданных тензора приводятся к одному размеру с помощью операций повышения размера и конкатенируются. После

чего, к тензору применяются 4 так называемых separable convolution. Данная операция заключается в том, чтобы вместо применения свертки nxn по всем каналам одновременно, сначала применить свертку 1x1, после чего отдельно по каждому каналу применять свертку nxn. При этом, в случае JPU, помимо идеи separable convolution, используется еще и dilation rate, заключающийся в том, чтобы применять свертку к большей площади(рис 7).

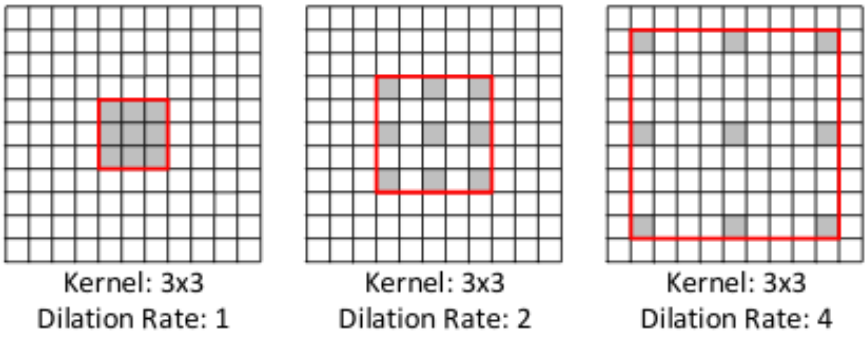


Рис. 7: Использование dilation rate с размером свертки 3 на 3.[5]

Таким образом, к тензору, полученному конкатенацией трех последних слоев FCN, применяется separable convolution с dilation rate 1, 2, 4 и 8. После чего результаты конкатенируются и к ним также применяется свертка(рис 8).

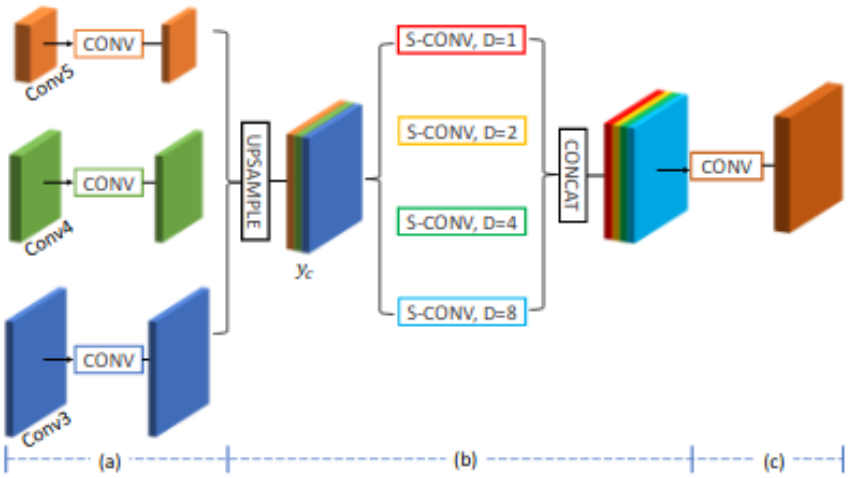


Рис. 8: Структура JPU.[4]

Данная сеть, благодаря более экономной операции separable convolution, уже способна обеспечивать хорошую скорость работы около 30 кадров в секунду на Titan-Xp GPU при входных изображениях размера 512×512[4]. Пример работы данной сети представлен ниже(рис 9).

Рис. 9: работа FastFCN.[4]

3 Instance segmentation, Mask R-CNN

Теперь, когда был описан принцип работы сетей, решающих задачу семантической сегментации, перейдем к рассмотрению задачи сегментации объектов, так как данная задача более релевантна в области компьютерного зрения. Напомним, что в задаче сегментации объектов требуется отделить объекты одного класса не только от фона, но и друг от друга.

Самым очевидным решением данной задачи является следующий алгоритм. Сначала решить задачу детекции, после чего отдельно в каждой выделенной области решать задачу семантической сегментации. Примерно такой алгоритм и предлагает Mask-RCNN[6]. Для получения регионов интереса она использует нейросеть Faster-RCNN[7], после чего с помощью FCN она получает сегментационную маску(рис 10)(рис 11).

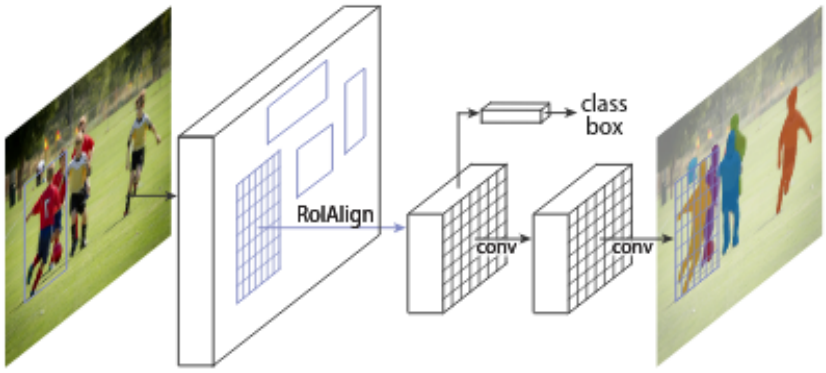


Рис. 10: Mask-RCNN[6]

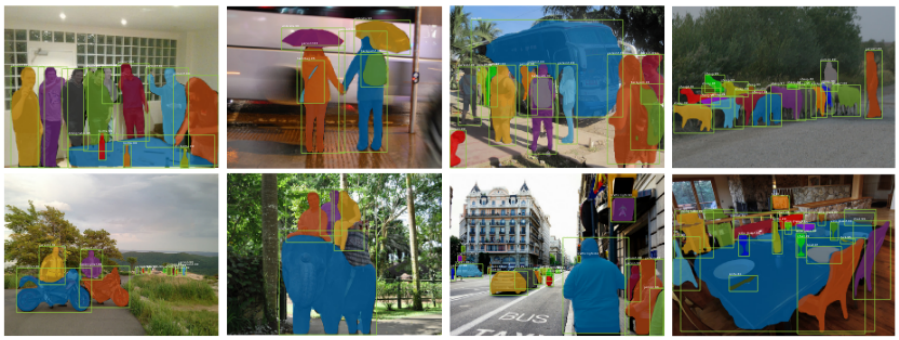


Рис. 11: Работа Mask-RCNN[6]

Кроме того, в статье(ссылка) предлагается восстанавливать границы региона интереса на исходном изображении с помощью метода RoIAlign, который по сути является билинейной интерполяцией(рис 12).

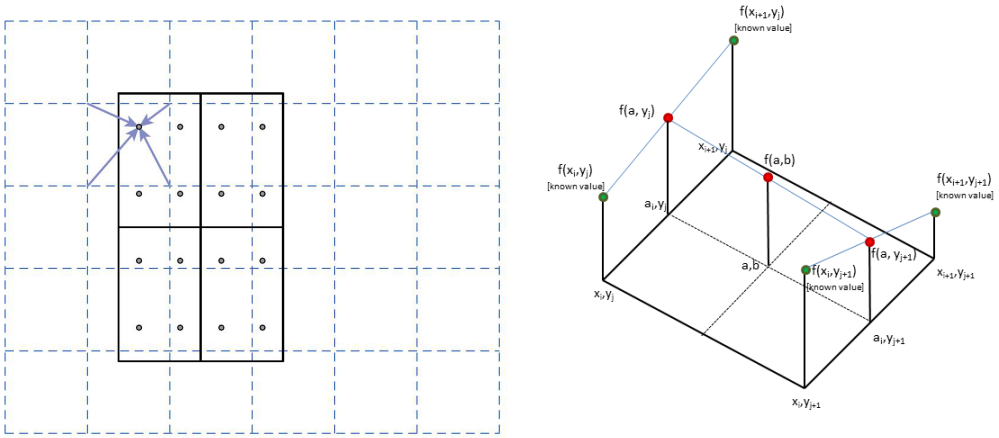


Рис. 12: RoI Align[6] и билинейная интерполяция[8]

На рисунке 12 слева пунктирной сеткой обозначена карта признаков, а непрерывной - отображение на карту признаков границ региона интереса с исходного изображения. В данный регион должно попасть 4 группы по 4 признака, обозначенных на рисунке точками. В предыдущих версиях использовалась операция округления, которая выравнивала регион по целочисленным координатам. RoI Align же оставляет точки в их текущих местах, но вычисляет значения каждой из них при помощи билинейной интерполяции по четырём ближайшим признакам.

Таким образом, Mask-RCNN дает сравнительно неплохие результаты в задаче сегментации объектов, однако она не разрабатывалась для быстрого решения данной задачи. Скорость предсказания составляет от 200 до 400 миллисекунд на Nvidia Tesla M40 GPU[6]

4 Panoptic segmentation, UPSNet

Зачастую возникает необходимость решать задачу не сегментации объектов, а так называемой паноптической сегментации[9]. В отличие от задачи сегментации объектов, данная задача предполагает не только отделение объектов от фона и друг от друга, но и сегментацию самого фона как это делалось в задаче семантической сегментации(рис 13).

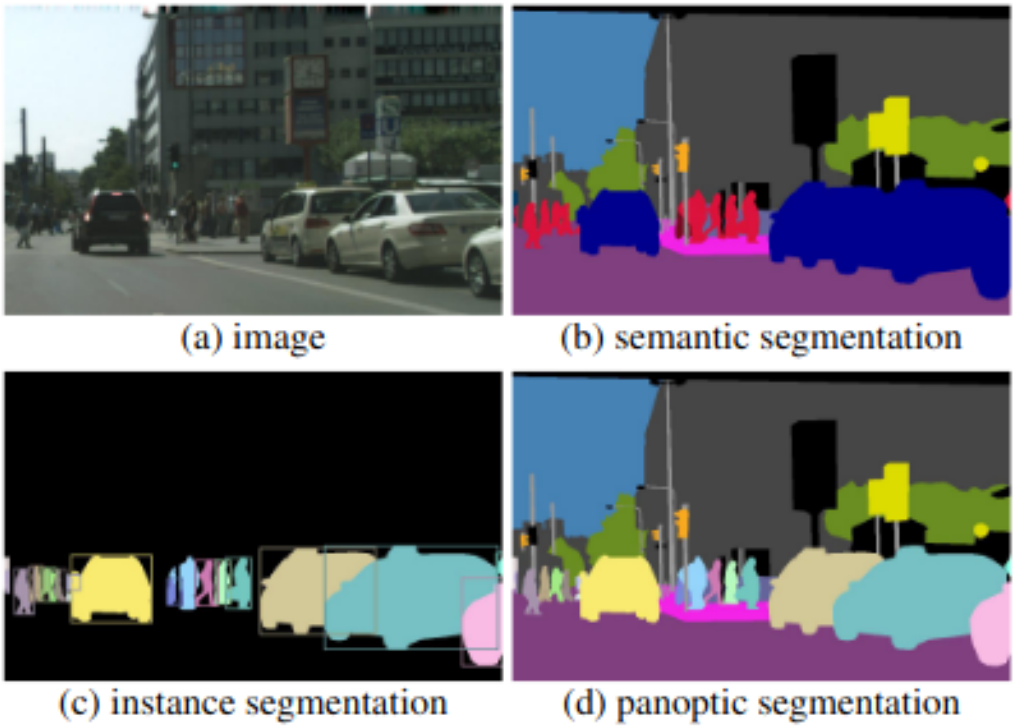


Рис. 13: Задачи сегментации[9]

Для объяснения дальнейших архитектур потребуется объяснить идею Feature Pyramid Network.

4.1 Feature Pyramid Network

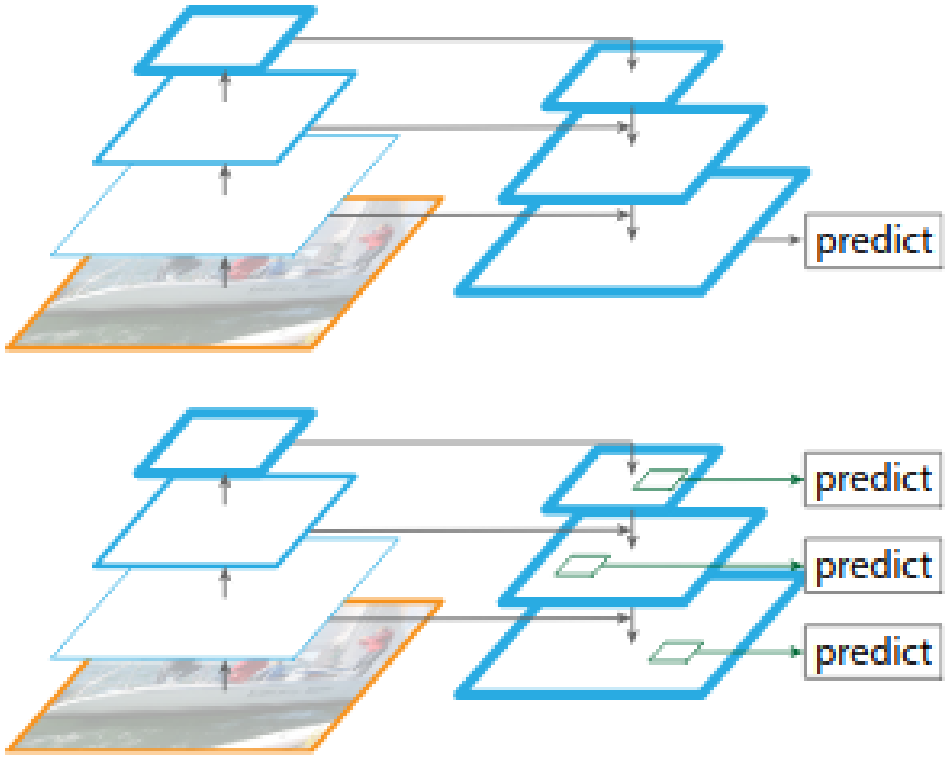


Рис. 14: Feature Pyramid Networks[10]

Общий принцип работы виден из рисунка 14. Сначала FPN с помощью операций свертки генерирует пирамиду карт признаков различного масштаба, после чего самая верхняя карта увеличивается с помощью операций upsampling, и на каждом этапе к ней прибавляется или приконкатенируется соответствующая карта больших масштабов. Данная схема широко применяется как в задачах детекции так и в задачах сегментации.

4.2 UPSNet

Теперь перейдем к описанию архитектуры сети под названием UPSNet[11], которая решающет задачу паноптической сегментации. Архитектура представлена на рисунке 15.

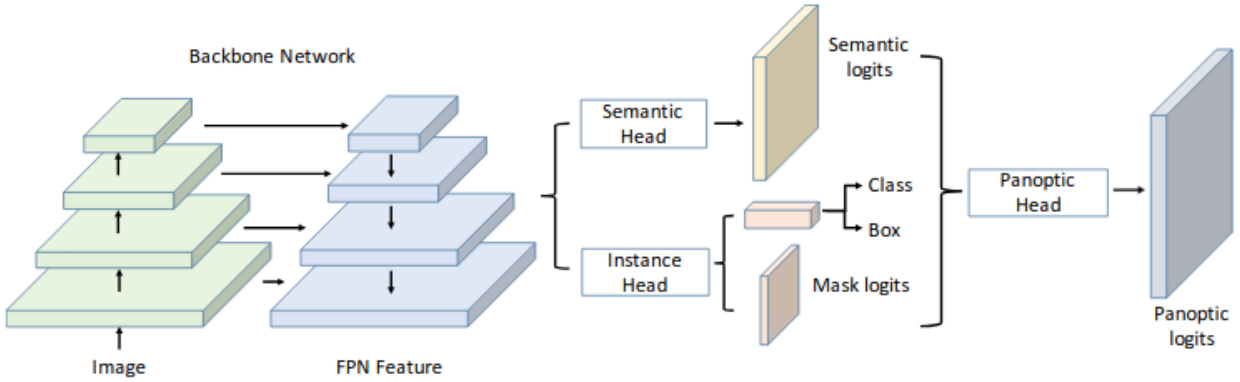


Рис. 15: Feature Pyramid Networks[11]

Сначала генерируются карты признаков с помощью backbone network, в качестве которой авторы сети взяли Mask-RCNN, добавив к ней FPN-надстройку.После чего, уже с помощью классической Mask-RCNN решается задача сегментации объектов, а с помощью FPN генерируют маску семантической сегментации. После чего, на основании наложения 2 выходных карт, создается карта паноптической сегментации.

Рис. 16: Результат работы UPSNet[11]

5 SiamMask

Теперь, когда были введены основные методы сегментации изображений, перейдем к рассмотрению модели, работающей в реальном времени, под названием SiamMask[12]. Данная сеть создавалась для задачи object tracking, в которой отслеживается 1 объект. Архитектура данной сети представлена на рисунке 17.

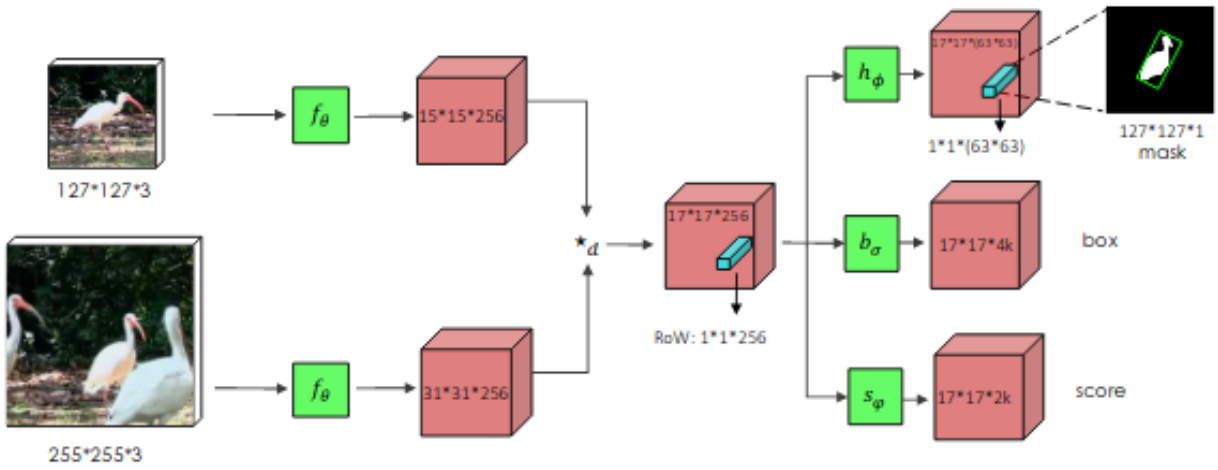


Рис. 17: Архитектура siammask[12]

Для того чтобы данная сеть начала работу, требуется выделить объект интереса на первом кадре. Принцип работы у нее следующий. Сначала исходное изображение и изображение объекта интереса проходят через оду и ту же сверточную нейросеть для получения их признакового описания. После чего, вычисляется кросс-корреляция по каждому уровню в глубину, после чего результат подается на вход 3 нейросетям, задачами которых является построение сегментационной маски, построение ограничивающей рамки как минимальный ограничивающий прямоугольник(стороны необязательно параллельны осям координат). Примеры работы данной сети продемонстрированы на рисунке 18.

Рис. 18: Результаты работы siammask[12]

Данная сеть производит сегментацию изображения на видео с большой точностью и скоростью около 55 кадров в секунду на NVIDIA RTX 2080 GPU[12].

6 CrossVIS

Следующая и последняя сеть, которая будет рассмотрена в данной работе, называется CrossVIS[13]. Она решает задачу video instance segmentation, то есть сегментацию последовательности кадров в реальном времени. В данной сети используется идея так называемого перекрестного обучения, заключающаяся в использовании полученных данных о текущем кадре для предсказания сегментационной маски в следующих кадрах(рис 19).

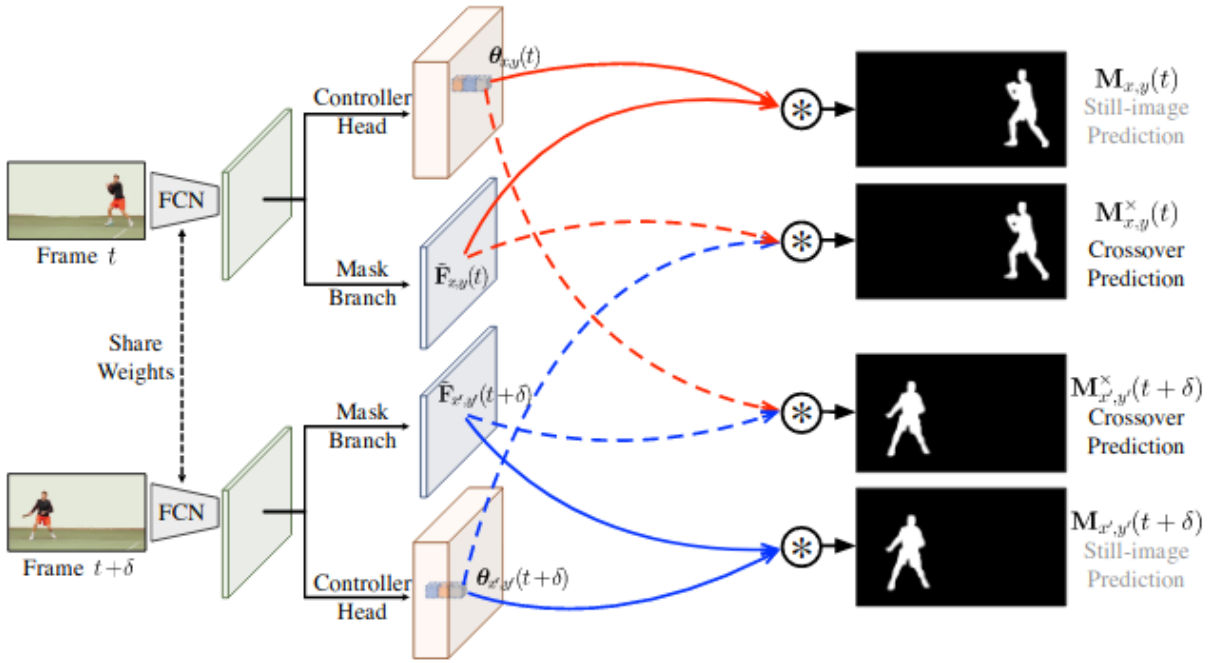


Рис. 19: Архитектура CrossVIS[13]

Для генерации признаков она использует FCN, после чего, на основании информации о внешнем виде отслеживаемых объектов θ и об их местоположении(для того, чтобы различать разные объекты одного класса между собой). На основани иэтих данных, а также на основании карты признаков F, полученной из FCN, делается точное предсказание о местоположении объекта на изображении. На рисунке 19 операциями * обозначено применение 3-х уровневой сверточной нейросети к F с весами θ.

Скорость работы данной сети составляет около 50 кадров в секунду.

7 Заключение

В заключение данной работы можно сделать вывод, что применение нейросетевых технологий в задачах сегментации привело к большому прорыву в данной области. Несмотря на то, что архитектурных элементов как таковых не так много, их различные комбинации дают иногда высокие результаты. Однако, стоит отметить, что несмотря на успешные продвижения в решении задачи сегментации, большинство методов работают не так быстро для внедрения их в системы реального времени, поэтому в данной области еще только предстоит найти успешные решения.

Список литературы

[1] Семантическая сегментация: краткое руководство
https://neurohive.io/ru/osnovy-data-science/semantic-segmention/

[2] An overview of semantic image segmentation.
https://www.jeremyjordan.me/semantic-segmentation/

[3] Transposed Convolution Demystified
https://towardsdatascience.com/transposed-convolution-demystified-84ca81b4baba

[4] FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation (2019) Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, Yizhou Yu
https://arxiv.org/pdf/1903.11816.pdf

[5] Smoothed Dilated Convolutions for ImprovedDense Prediction (2019) Zhengyang Wang and Shuiwang Ji
https://arxiv.org/pdf/1808.08931.pdf

[6] Mask R-CNN (2018) Kaiming HeGeorgia GkioxariPiotr Dollar Ross Girshick
https://arxiv.org/pdf/1703.06870.pdf

[7] Faster R-CNN: Towards Real-Time ObjectDetection with Region Proposal Networks (2016) Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun
https://arxiv.org/pdf/1506.01497.pdf

[8] Understanding Region of Interest — (RoI Align and RoI Warp)
https://towardsdatascience.com/understanding-region-of-interest-part-2-roi-align-and

[9] Panoptic Segmentation (2019) Alexander Kirillov Kaiming He Ross Girshick Carsten Rother Piotr Dollar
https://arxiv.org/pdf/1801.00868.pdf

[10] Feature Pyramid Networks for Object Detection (2017) Tsung-Yi Lin, Piotr Dollar,Ross Girshick,Kaiming He,Bharath Hariharan, Serge Belongie
https://arxiv.org/pdf/1612.03144.pdf

[11] UPSNet: A Unified Panoptic Segmentation Network (2019) Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, Raquel Urtasun
https://arxiv.org/abs/1901.03784

[12] Fast Online Object Tracking and Segmentation: A Unifying Approach (2019) Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, Philip H. S. Torr
https://arxiv.org/pdf/1812.05050.pdf

[13] Crossover Learning for Fast Online Video Instance Segmentation (2021) Authors: Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, Wenyu Liu
https://arxiv.org/pdf/2104.05970.pdf