Национальный исследовательский
Нижегородский государственный университет им. Н.И. Лобачевского
Институт информационных технологий, математики и механики

Образовательный курс
«Современные методы и технологии
глубокого обучения в компьютерном зрении»

Семантическая сегментация изображений с использованием методов глубокого обучения

При поддержке компании Intel

Гетманская Александра, Кустикова Валентина

Содержание

- Цель лекции
- Постановка задачи семантической сегментации изображений
- Открытые наборы данных
- Показатели качества семантической сегментации изображений
- Глубокие модели для семантической сегментации изображений
- Сравнение моделей семантической сегментации изображений
- Заключение

Цель лекции

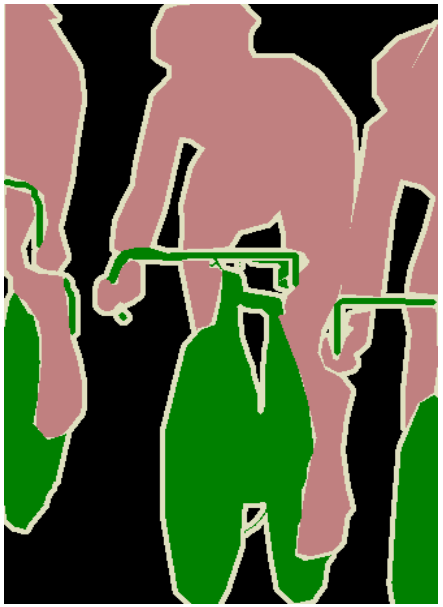
- ***Цель*** – изучить глубокие нейросетевые модели для решения задачи семантической сегментации изображений (изображений естественного мира, медицинских изображений и других)

ПОСТАНОВКА ЗАДАЧИ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ

Постановка задачи (1)

- Задача семантической сегментации состоит в том, чтобы каждому пикселю изображения поставить в соответствие класс объектов, которому этот пиксель принадлежит (разные цвета соответствуют разным классам)

Оригинал



Разметка

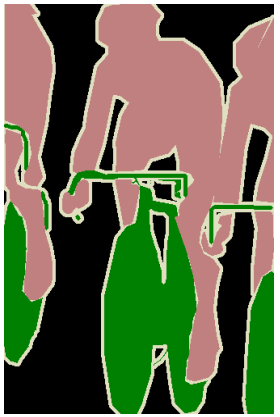
Результат сегментации

* The PASCAL Visual Object Classes Homepage [http://host.robots.ox.ac.uk/pascal/VOC].

Постановка задачи (2)

- Исходное изображение представлено набором интенсивностей пикселей $I = \left(I_{ij}^{k}\right)_{\substack{0 \le i < w \\ 0 \le j < h \\ 0 \le k < 3}}$, где w и h – ширина и высота изображения, k – количество каналов
- Определено множество допустимых классов объектов на изображении $C = \{0, 1, \ldots N-1\}$, 0 соответствует фону, остальное множество идентификаторов однозначно сопоставляются с множеством классов
- Требуется найти отображение

$$\varphi\left(I_{ij}\right) = c$$



ОТКРЫТЫЕ НАБОРЫ ДАННЫХ

Наборы данных (1)

Набор данных	Размер тренировочного множества	Размер тестового множества	Количество классов
Семантическая сегментация объектов реальной жизни			
PASCAL VOC 2012 [http://host.robots.ox.ac.uk/pascal/VOC/voc2012]	9 963	1 447	20
ADE20K [http://groups.csail.mit.edu/vision/datasets/ADE20K]	20 210	2 000	150
MS COCO'15 [http://mscoco.org]	80 000	40 000	80
...			

Наборы данных (2)

Набор данных	Размер тренировочного множества	Размер тестового множества	Количество классов
Семантическая сегментация дорожных объектов			
CamVid [http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid]	468	233	11
Cityscapes [https://www.cityscapes-dataset.com]	2 975	500	19
KITTI [http://www.cvlibs.net/datasets/kitti]	200	200	4
Семантическая сегментация интерьеров			
Sun-RGBD [http://rgbd.cs.princeton.edu]	10 355	2 860	37
NYUDv2 [http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html]	795	645	40

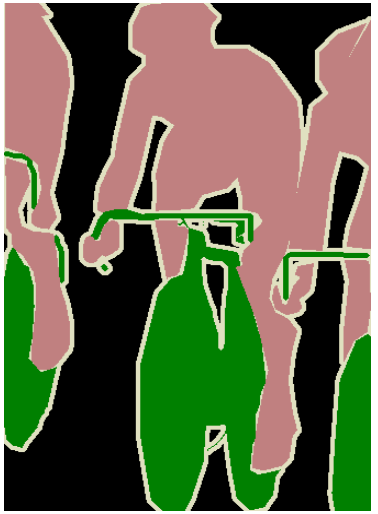
Наборы данных (3)

- MS COCO’15 – самая объемная база изображений для семантической сегментации объектов реальной жизни
- Cityscapes содержит изображения, снятые в 50 городах с видеорегистратора при разных погодных условиях
- Бенчмарк KITTI содержит данные и инструменты для оценки качества решения различных задач на изображениях дорожных сцен (детектирование объектов, семантическая сегментация изображений и объектов, сопровождение объектов, детектирование полос движения и другие)
- Sun-RGBD содержит изображения сцен внутри помещений (дом, офис), для которых решаются задачи классификации изображений (2 категории), семантической сегментации, детектирование трехмерных объектов и оценка их положения, распознавание сцены (scene understanding)

PASCAL VOC 2012

- PASCAL VOC 2012 – наиболее известная база изображений
- 20 классов объектов естественного мира: airplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor

Изображение



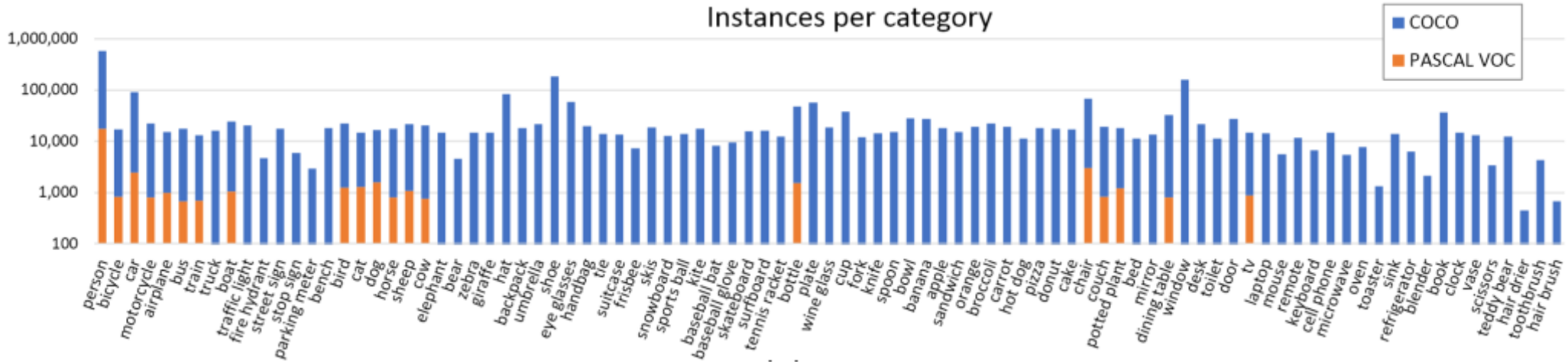
Разметка

(разные цвета – разные классы объектов, отдельно размечены границы)

* The PASCAL Visual Object Classes Homepage [http://host.robots.ox.ac.uk/pascal/VOC].

MS COCO’15

- MS COCO’15 – самая обширная база изображений естественного мира (похожих на PASCAL VOC) с точки зрения количества категорий объектов (80 категорий) и числа изображений, по каждой категории содержится значительное количество изображений (близкое распределение по классам)



* Lin T.Y., et al. Microsoft COCO: Common objects in context // Lecture Notes in Computer Science. – Vol. 8693. – 2014. – P. 740-755. – [https://arxiv.org/pdf/1405.0312].

Cityscapes

- Изображения дорожных сцен, полученных с видеорегистратора
- 5 000 изображений с высококачественной разметкой
- 20 000 изображений с грубой разметкой
- 30 классов, объединенных в 8 групп

Пример точной разметки

Цюрих (Швейцария)

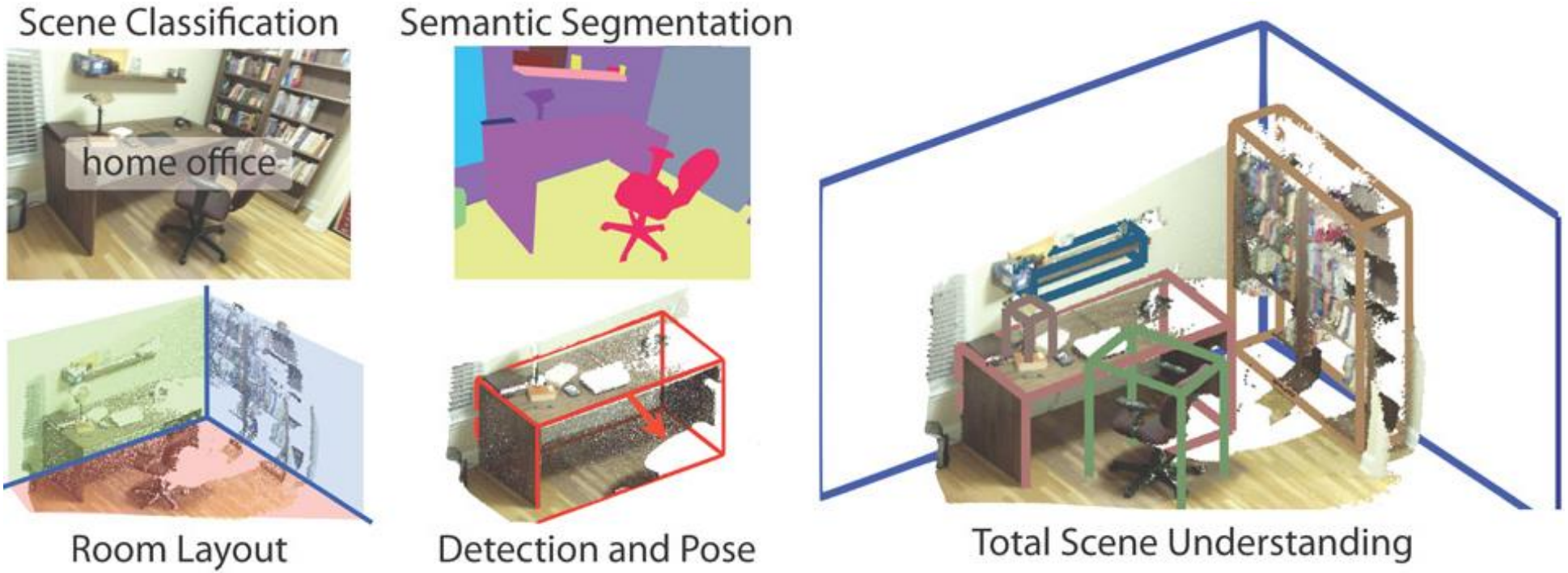
Пример грубой разметки

Саарбрюккен (Германия)

* The Cityscapes Dataset Homepage [https://www.cityscapes-dataset.com/examples].

SUN RGB-D

- SUN RGB-D содержит изображения и разметку сцен внутри помещений для решения нескольких задач (примеры ниже)



* Song S., Lichtenberg S.P., Xiao J. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite [https://3dvision.princeton.edu/projects/2015/SUNrgbd/poster.pdf].

ПОКАЗАТЕЛИ КАЧЕСТВА СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ

Рассматриваемые показатели качества

- Попиксельная точность (pixel accuracy)
- Средняя попиксельная точность по классам наблюдаемых объектов (mean pixel accuracy over classes)
- Метрика IoU (Intersection over Union) или индекс Жаккара (Jaccard index)
- Индекс Дайса (Dice index) или F1-score

Попиксельная точность

- ***Попиксельная точность*** (pixel accuracy) определяется следующим образом:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

где $TP + TN$ – количество правильно проклассифицированных пикселей (true positives + true negatives),
$TP + TN + FP + FN$ – общее количество пикселей

		Предсказание	
		True	**False**
Разметка	**True**	TP	FN
	False	FP	TN

Средняя попиксельная точность по классам

- Попиксельная точность отражает количество правильно проклассифицированных пикселей
- Попиксельная точность не является показательной в случае несбалансированности классов

- Поэтому вводится средняя попиксельная точность, вычисленная для каждого класса в отдельности и усредненная по количеству классов, – ***средняя попиксельная точность по классам*** (mean pixel accuracy over classes)

Метрика IoU (1)

- ***Метрика IoU*** (Intersection over Union) или индекс Жаккара (Jaccard index)

$$IoU = \frac{TP}{TP + FP + FN}$$

где TP – количество правильно проклассифицированных пикселей (true positives),
FP – количество пикселей, которые метод проклассифицировал как принадлежащие классу, но они таковыми не являются (false positives),
FN – количество пикселей, которые принадлежат классу, но метод проклассифицировал их как не принадлежащие классу (false negatives)

		Предсказание	
		True	**False**
Разметка	**True**	TP	FN
	False	FP	TN

Метрика IoU (2)

- Обычно вычисляется среднее значение метрики IoU (Mean IoU) по всем классам, на полном наборе данных
- Среднее значение метрики IoU может вычисляться как взвешенное среднее по соответствующим значениям, полученным для отдельных классов. Веса назначаются равными частотам встречаемости пикселей каждого класса
- При вычислении метрики IoU класс «фон» может учитываться, а может не учитываться
- Пиксели на границах объектов могут не учитываться или учитываться с меньшим весом по сравнению с «внутренними» пикселями

Индекс Дайса

- ***Индекс Дайса*** (Dice index) или F1-score определяется следующим образом:

$$DICE = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

- Индекс Дайса отличается от индекса Жаккара одним коэффициентом
- Указанные индексы связаны соотношениями:

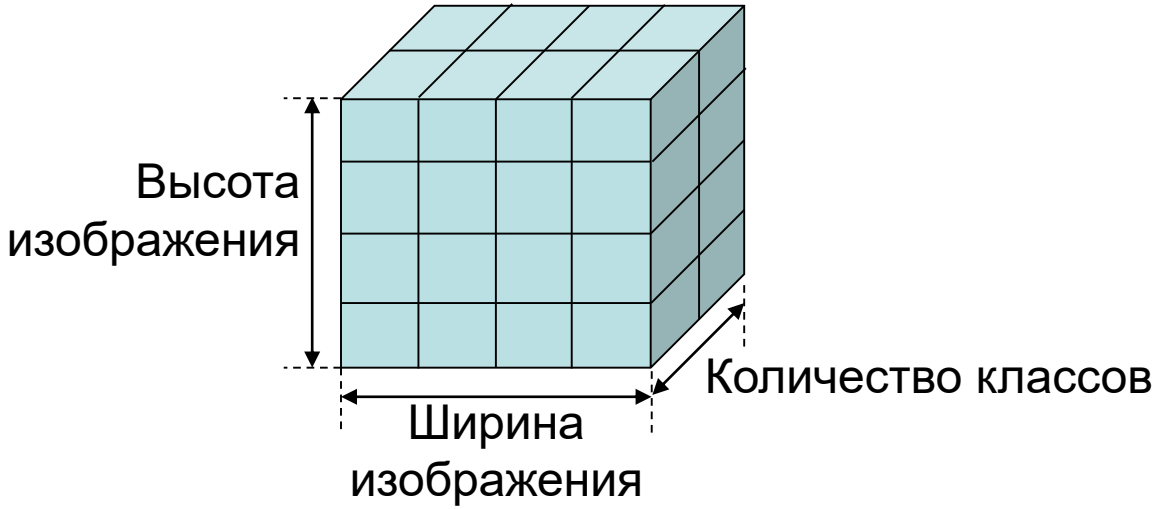
$$IoU = \frac{DICE}{2 - IoU}, \qquad DICE = \frac{2 \cdot IoU}{1 + IoU}$$

- Как следствие, не имеет смысла одновременно определять оба показателя, достаточно вычислять какой-то один показатель

ГЛУБОКИЕ МОДЕЛИ ДЛЯ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ

Проблема применения глубоких моделей для семантической сегментации (1)

- При решении задачи семантической сегментации изображений на выходе модели должен быть трехмерный тензор с элементами, отвечающими достоверности принадлежности каждого пикселя к определенному классу



- ***Каким образом обеспечить на выходе тензор, у которого пространственные размерности совпадают с разрешением входного изображения?***

Проблема применения глубоких моделей для семантической сегментации (2)

- Возможные способы решения проблемы получения выходного тензора, пространственная размерность которого совпадает с разрешением входного изображения:
  - Интерполяция
  - Построение архитектуры «кодировщик-декодировщик» (encoder-decoder architecture)
  - Применение графовых вероятностных методов, в частности, условных случайных полей (Conditional Random Fields, CRF)
- Интерполяция – простой и понятный способ, но не позволяет получить качественный результат, в особенности, для небольших объектов и на границах объектов
- Два оставшихся метода являются более перспективными с точки зрения качества результатов

Рассматриваемые модели (1)

- ***FCNs, SegNet, U-Net (2015)***

Полностью сверточные сети

  - Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965] (опубликованная версия).
  - Badrinarayanan V., Kendall A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. – 2015. – [https://arxiv.org/pdf/1511.00561.pdf], [https://ieeexplore.ieee.org/document/7803544] (опубликованная версия).
  - Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation. – 2015. – [https://arxiv.org/pdf/1505.04597.pdf], [https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28] (опубликованная версия).

Рассматриваемые модели (2)

- ***PSPNet (2016)***
  - Zhao H., Shi J., Qi X., Wang X., Jia J. Pyramid scene parsing network. – 2016. – [https://arxiv.org/pdf/1612.01105.pdf], [https://ieeexplore.ieee.org/document/8100143] (опубликованная версия).
- ***ICNet (2017)***
  - Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25] (опубликованная версия).

Пирамиды признаков с разных масштабов

Рассматриваемые модели (3)

Применение CRF и поиск замены для CRF для ускорения вычислений

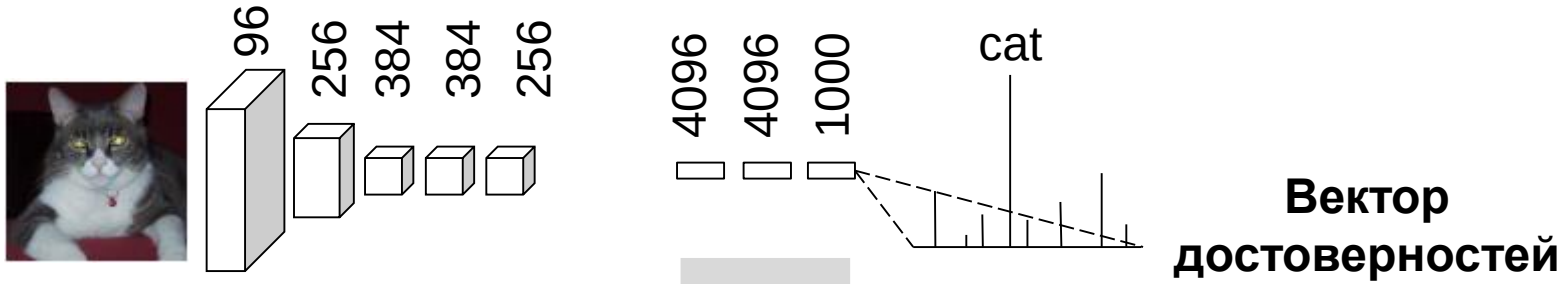
- ***DeepLab-v1,*-v2, *-v3, *v3+ (2014-2018)***
  - Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. – 2014. – [https://arxiv.org/pdf/1412.7062.pdf].
  - Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. – 2017. – [https://arxiv.org/pdf/1606.00915.pdf], [https://ieeexplore.ieee.org/document/7913730] (опубликованная версия).
  - Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].
  - Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

FCN (1)

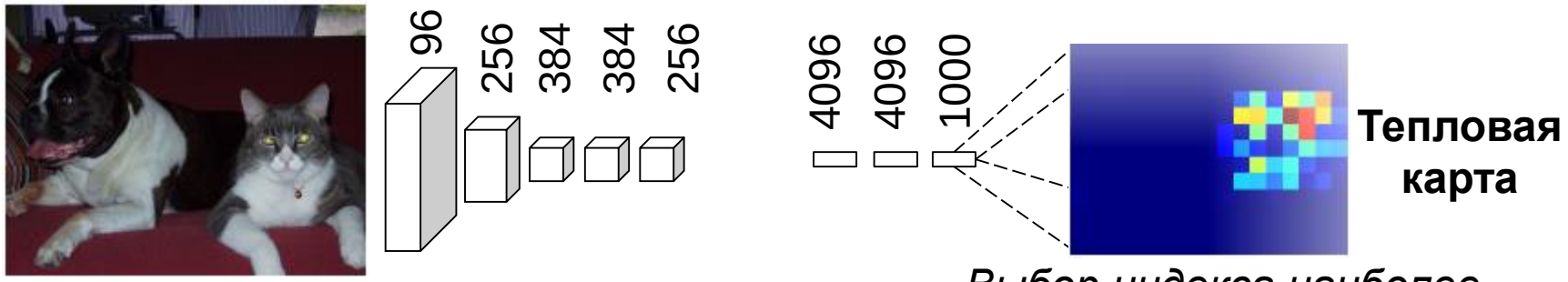
- FCNs (Fully Convolutional Networks) – модели, цель разработки которых адаптировать классификационные сверточные сети (AlexNet, VGG, GoogLeNet) для решения задачи семантической сегментации
  - Классификационные модели принимают на вход изображение фиксированного размера
  - Классификационные модели возвращают вектор достоверностей, отражающих степень принадлежности изображения каждому допустимому классу объектов
  - Заменим полносвязные слои на сверточные, чтобы применять модель к изображениям произвольного размера
  - Таким образом, реализуем «скользящее» окно и для каждого его положения получим вектор достоверностей

* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (2)



«Замена» полносвязных слоев на полностью сверточные (другая интерпретация)



Выбор индекса наиболее достоверного класса

* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (3)

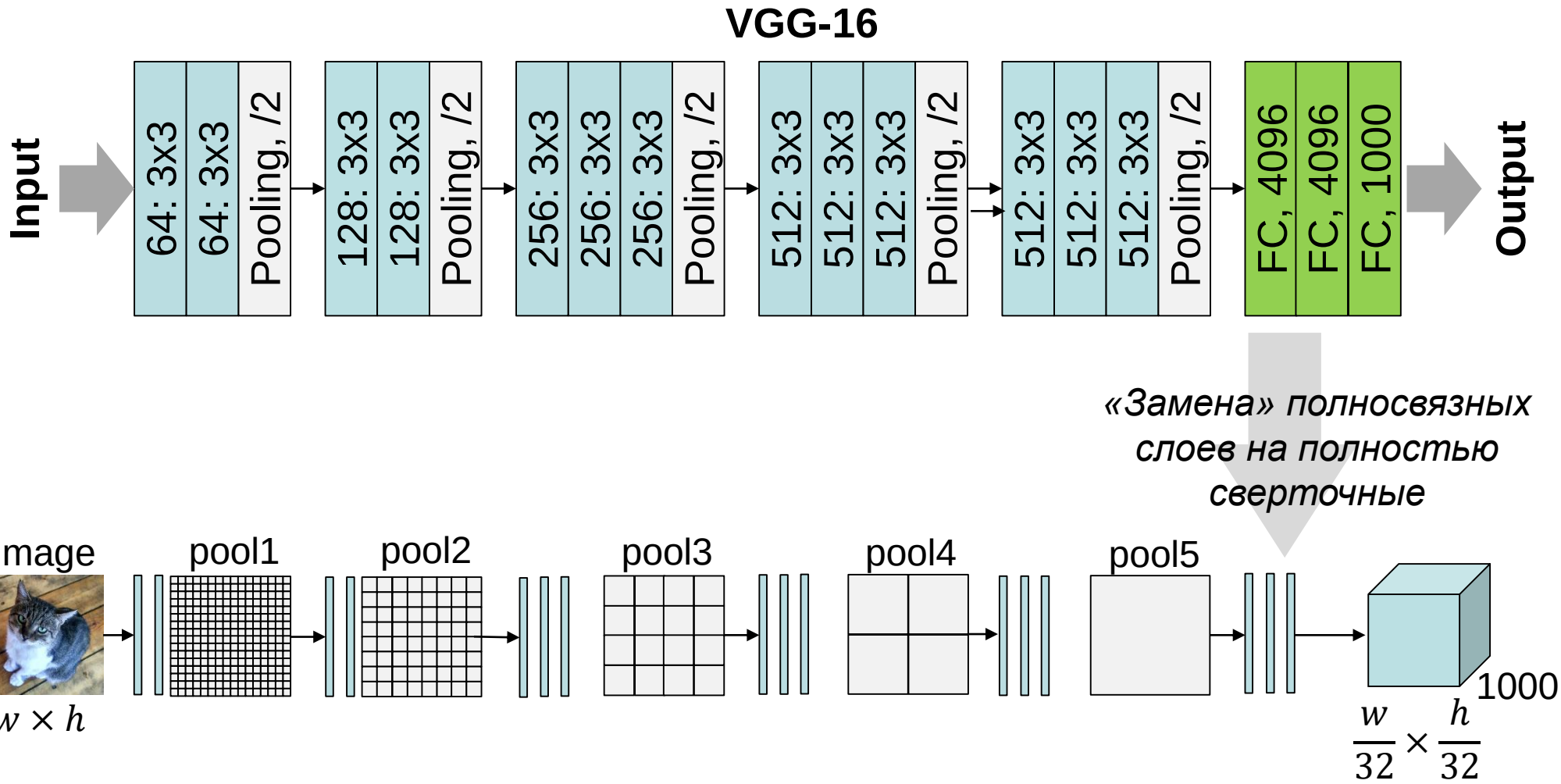
- Полносвязные слои преобразуются в полностью сверточные слои с использованием одномерной свертки (ядро 1x1). Слои остаются теми же, а такая «замена» является эквивалентной
- Входное изображение может быть произвольного разрешения
- На выходе модели формируется трехмерный тензор, в котором количество каналов совпадает с количеством классов объектов, а пространственные размеры соответствуют количеству возможных положений «скользящего» окна на входном изображении
- Выбор индекса класса с максимальным значением достоверности позволяет построить тепловую карту (heatmap), которая рассматривается как результат семантической сегментации, но более низкого разрешения

FCN (4)

- Повышение разрешения карт признаков, включая выходную, выполняется с использованием обратных сверток (deconvolution, backwards или transposed convolution)
- Для улучшения качества результирующей карты предлагается использовать карты признаков, полученные на промежуточных слоях модели, т.е. признаки более низкого уровня
- Авторы FCN* в качестве базовых моделей использовали AlexNet, VGG, GoogLeNet
- VGG-16 показала лучшие результаты, поэтому далее рассматривается FCN, построенная на базе VGG-16

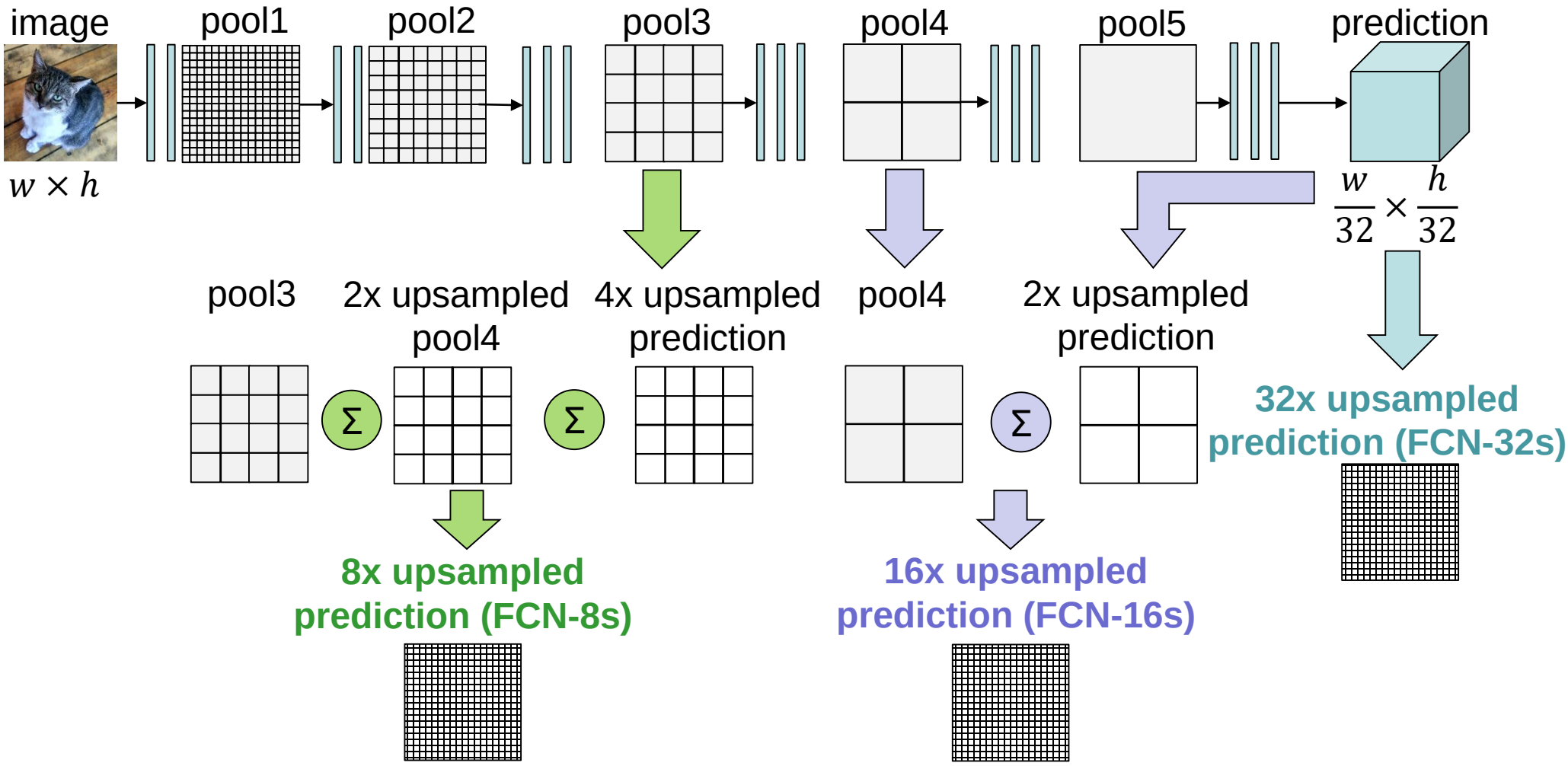
* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (5)



* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (6)



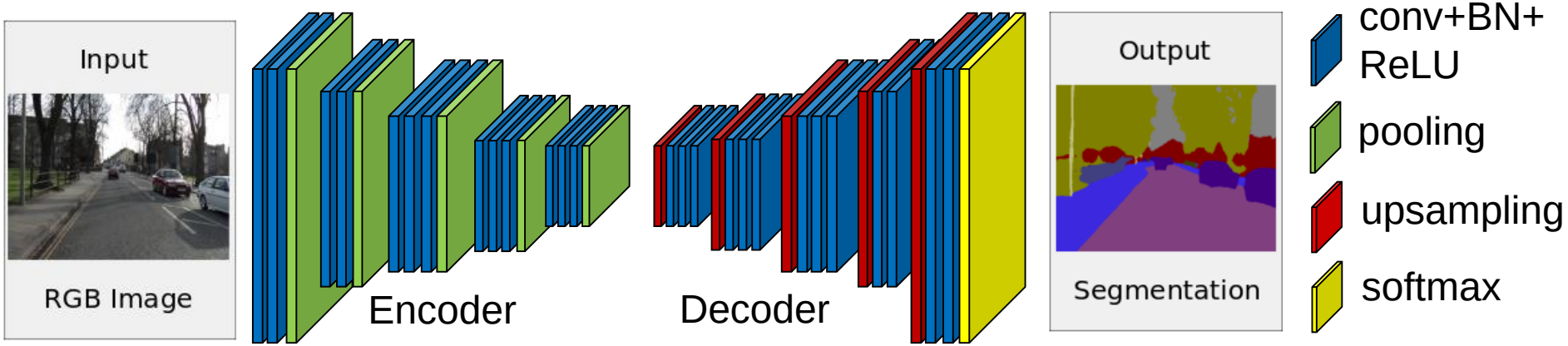
* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (7)

- Замена полносвязных слоев на полностью сверточные в VGG-16:
  - FC 4096 → Conv 4096, 1x1
  - FC 4096 → Conv 4096, 1x1
  - FC 1000 → Conv 1000, 1x1
- После замены на вход модели можно подавать изображение произвольного разрешения $w \times h$, на выходе модели формируется карта достоверностей размера $\frac{w}{32} \times \frac{h}{32} \times 1000$
- Пространственная размерность выходной карты повышается посредством применения обратной свертки с шагом повышающей дискретизации 32 (upsampling stride). Формируется «грубый» результат сегментации (модель FCN-32s)
- Более точные результаты получаются при использовании признаков с промежуточных слоев (модели FCN-16s, FCN-8s)

SegNet (1)

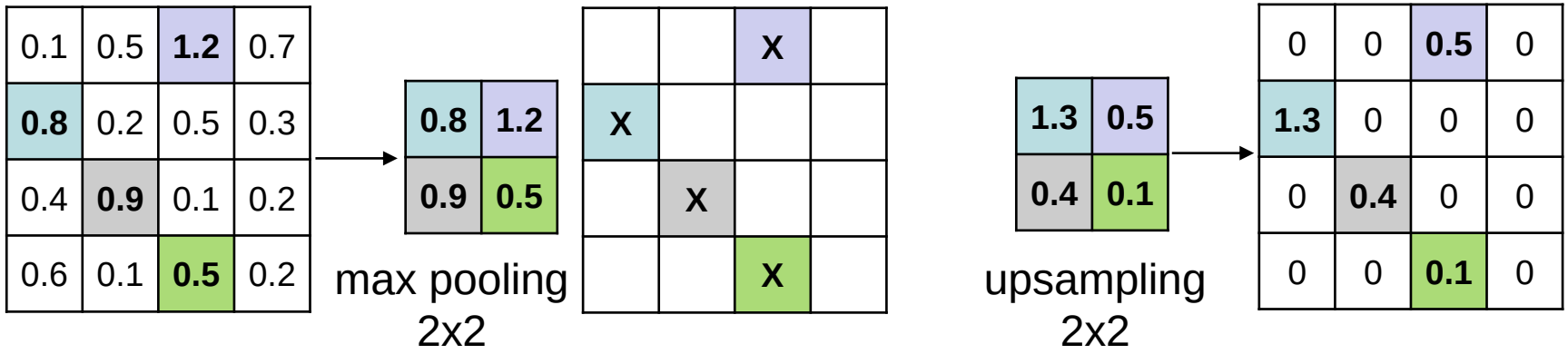
- SegNet – глубокая модель для семантической сегментации, построенная на базе архитектуры «кодировщик-декодировщик» (encoder-decoder)
- Цель разработки – создать сеть для распознавания дорожного движения и интерьеров, эффективную с точки зрения использования памяти и вычислительной сложности



* Badrinarayanan V., Kendall A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. – 2015. – [https://arxiv.org/pdf/1511.00561.pdf], [https://ieeexplore.ieee.org/document/7803544].

SegNet (2)

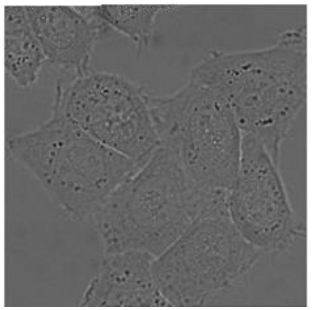
- Кодировщик содержит сверточную часть сети VGG-16
- Декодировщик строится зеркально кодировщику:
  - Каждому сверточному слою в кодировщике соответствует сверточный слой в декодировщике в обратном порядке
  - Каждой операции пространственного объединения (pooling) соответствует операция повышающей дискретизации (upsampling). Индексы на каждом слое максимального объединения в кодировщике сохраняются и используются в декодировщике для соответствующей карты признаков

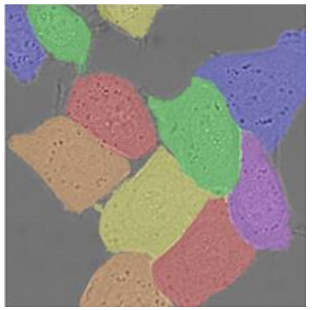


U-Net (1)

- Авторы U-Net предлагают модель и стратегию обучения, которая основана на дополнении данных за счет их трансформации (data augmentation) для более эффективного использования небольшого набора аннотированных образцов
- U-Net – модель, продемонстрировавшая хорошие результаты, в частности, на задаче сегментации нейронных структур в электронно-микроскопических стопках (segmentation of neuronal structures in electron microscopic stacks)

Входное изображение





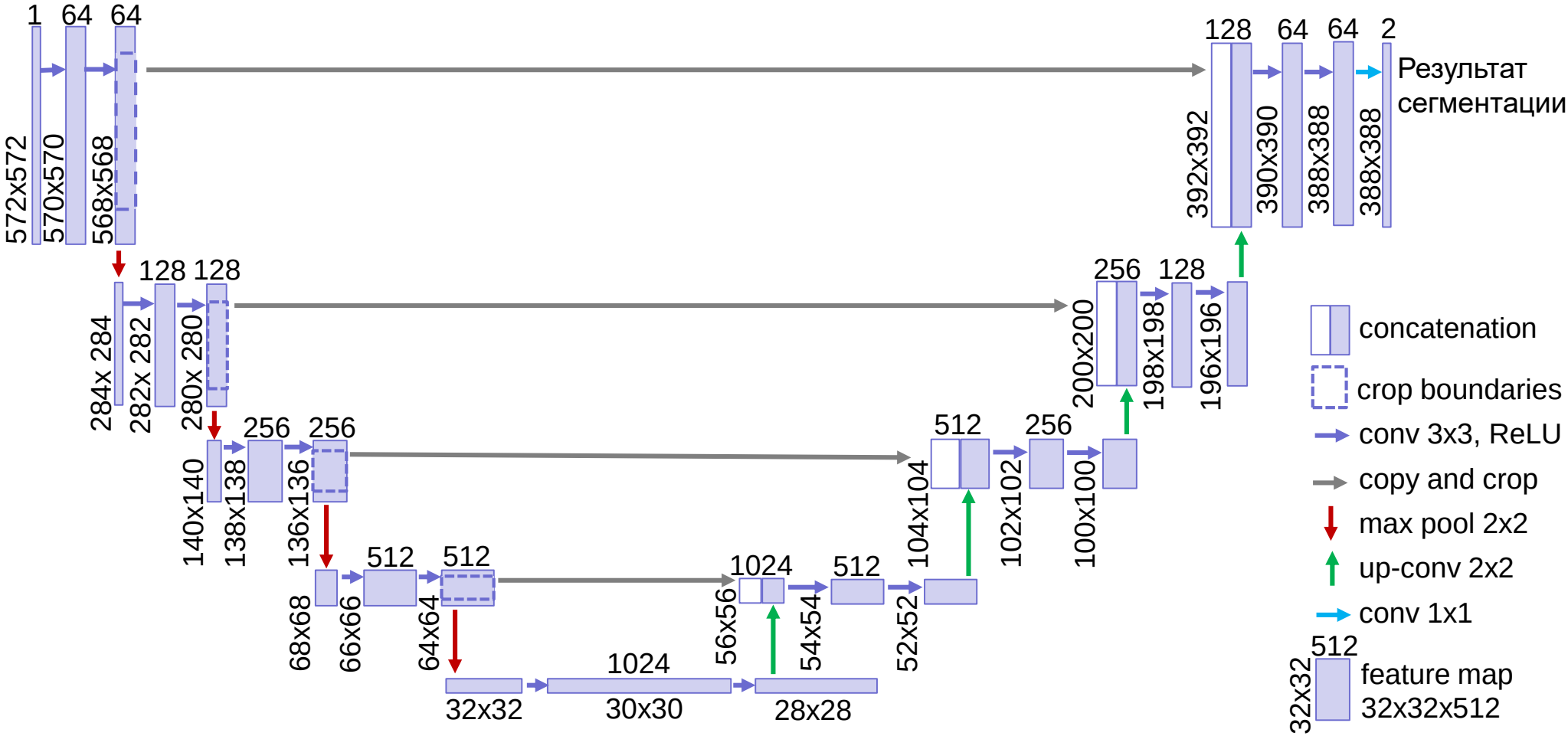
Нейронные структуры

* Ronneberger O., Fischer P., Brox T. U-Net: Convolutional networks for biomedical image segmentation. – 2015. – [https://arxiv.org/pdf/1505.04597.pdf].

U-Net (2)

- Топология U-Net состоит в двух веток:
  - ***«Сжимающий путь»*** (contracting path) – сверточная сеть из последовательности блоков, содержащих две свертки 3x3 (без дополнения краев), после каждой из которых следует «положительная срезка» (ReLU), в конце блока применяется операция пространственного объединения (max pooling) с ядром 2x2 и шагом 2
  - ***«Разжимающий путь»*** (expansive path) включает операцию повышающей дискретизации (upsampling); свертку 2x2, снижающую количество каналов вдвое (up-conv); конкатенацию с соответствующей картой признаков из «сжимающего пути», которая предварительно обрезана; две свертки 3x3, после каждой из которых следует ReLU

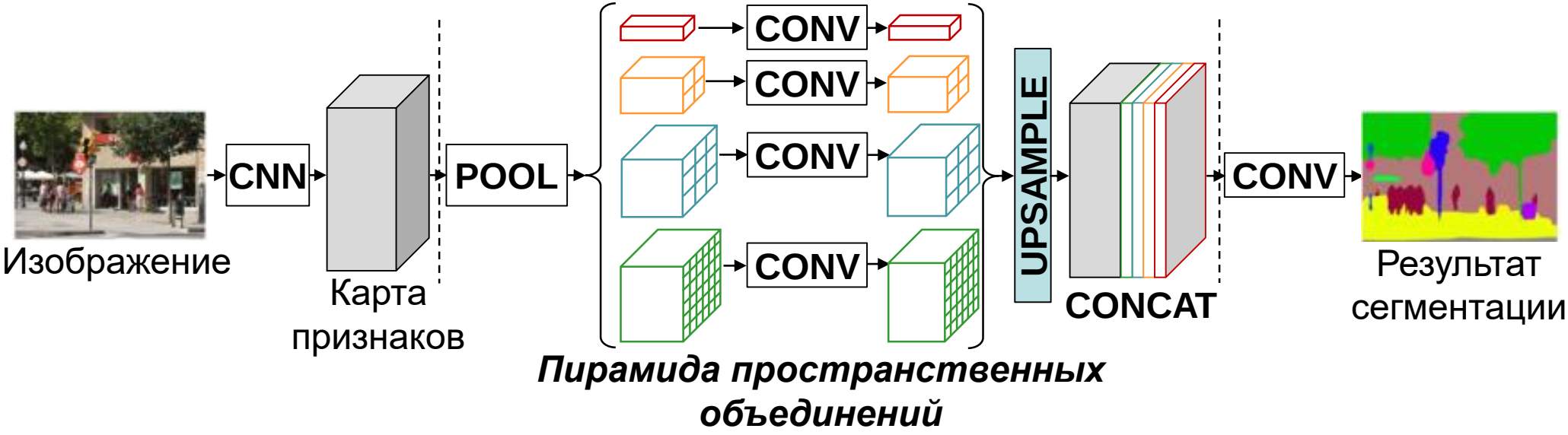
U-Net (3)



* Ronneberger O., Fischer P., Brox T. U-Net: Convolutional networks for biomedical image segmentation. – 2015. – [https://arxiv.org/pdf/1505.04597.pdf].

PSPNet (1)

- PSPNet (Pyramid Scene Parsing) – модель, которая использует построение пирамиды карт признаков разного масштаба для учета информации с разных уровней детализации
- PSPNet показала лучшие результаты на ImageNet Scene Parsing Challenge 2016, PASCAL VOC 2012 и Cityscapes в 2016 году



* Zhao H., Shi J., Qi X., Wang X., Jia J. Pyramid scene parsing network. – 2016. – [https://arxiv.org/pdf/1612.01105.pdf], [https://ieeexplore.ieee.org/document/8100143].

PSPNet (2)

- ***Карта признаков***
  - Для извлечения признаков используется сверточная часть модели ResNet, к которой применены свертки с пропусками (dilated convolutions)
- ***Пирамида пространственных объединений (Pyramid Pooling Module)***
  - Пространственное объединение (POOL)
    - Красная карта: результат глобального среднего объединения по каждому каналу карты признаков (самый «грубый» уровень)
    - Оранжевая карта: результат пространственного объединения по регионам, полученным при разбиении карты признаков на 2x2 блока
    - Голубая карта: результат объединения по регионам, полученным при разбиении карты признаков на 3x3 блока
    - Зеленая карта: результат объединения по регионам, полученным при разбиении карты признаков на 6x6 блоков

PSPNet (3)

- Промежуточные свертки (набор слоев CONV)
  - Свертки с ядрами 1x1 для уменьшения количества каналов, т.е. снижения представления контекста до $\frac{1}{N}$ от исходного, где N – количество уровней пирамиды
  - В представленном примере $N = 4$, если количество каналов входной карты составляет 2048, то на выходе каждого уровня пирамиды количество каналов – 512
- Повышающая дискретизация (UPSAMPLE)
  - Применение билинейной интерполяции для увеличения размерности карт признаков до исходной карты
- Конкатенация карт признаков (CONCAT)
  - Конкатенация исходной карты признаков с картами, полученными в результате повышающей дискретизации

❑ ***Результат сегментации***

- Финальная свертка (CONV)

PSPNet (4)

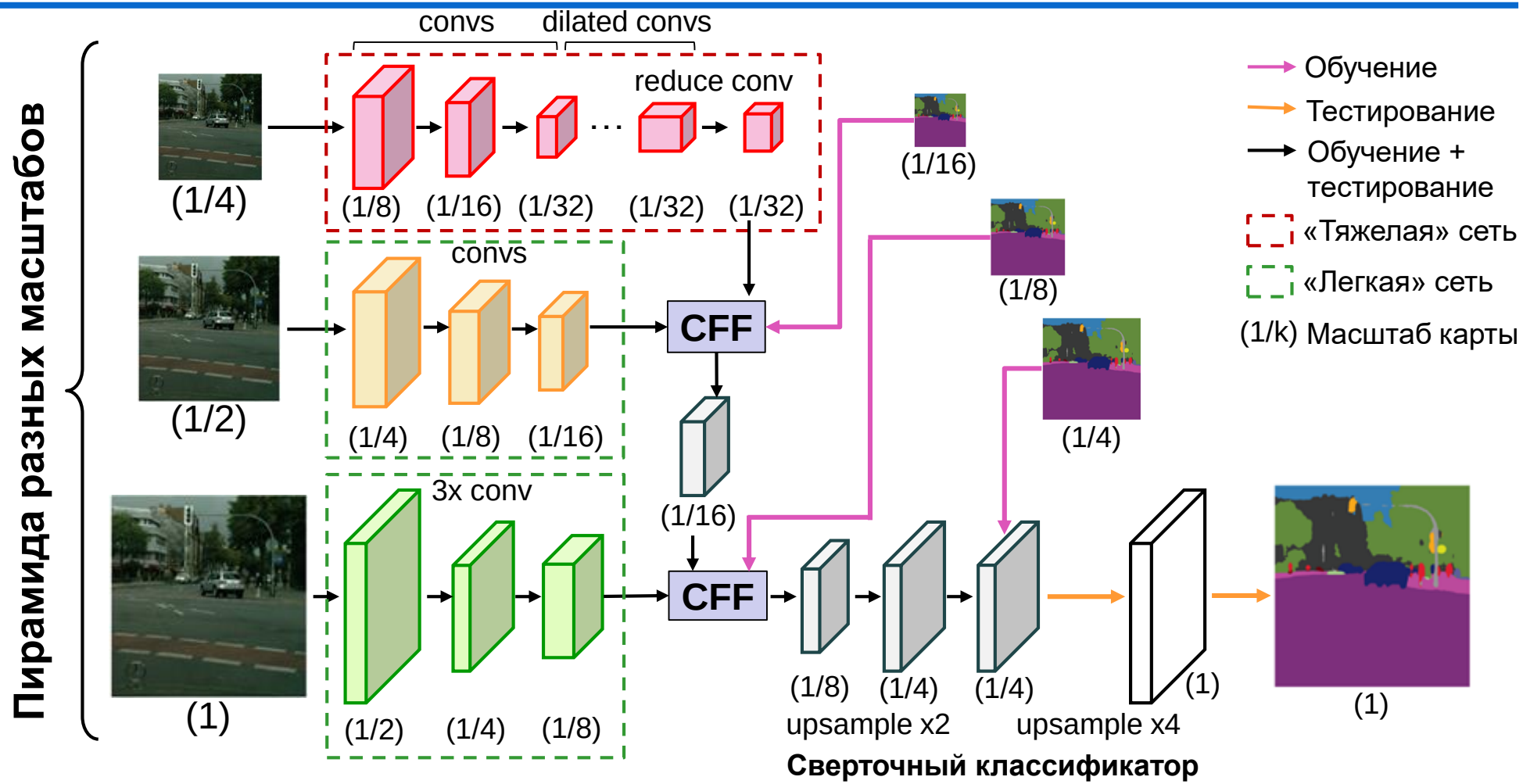
- Особенности обучения:
  - Введена вспомогательная функция потерь (auxiliary loss) с промежуточного слоя модели
  - Вспомогательная функция потерь помогает оптимизировать процесс обучения, в то время как основная функция несет полную ответственность за решение задачи
  - Для балансировки вклада вспомогательной функции потерь вводится весовой коэффициент

ICNet (1)

- ICNet (Image Cascade Network) – модель для семантической сегментации изображений в реальном времени (на одном графическом процессоре), которая основана на построении каскада карт признаков для разных масштабов исходного изображения
- Вход модели – пирамида масштабов исходного изображения
- Для каждого изображения обеспечивается построение карт признаков с использованием сверточных сетей
  - Чем крупнее изображение в пирамиде, тем проще используемая сверточная сеть
  - При построении карты признаков на каждом следующем масштабе используются признаки с предыдущих масштабов

* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

ICNet (2)



* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

ICNet (3)

- Сверточная сеть на каждом слое снижает пространственные размеры карты признаков, либо оставляет их неизменными
- Слияние карт признаков с соседних масштабов обеспечивается с помощью ***модуля слияния каскадных признаков*** (Cascade Feature Fusion, CFF)
- Модуль слияния каскадных признаков позволяет восстанавливать и улучшать результат сегментации с меньшими вычислительными затратами

- Далее рассматривается структура модуля слияния каскадных признаков и схема его работы на этапах обучения и тестирования построенной модели

* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

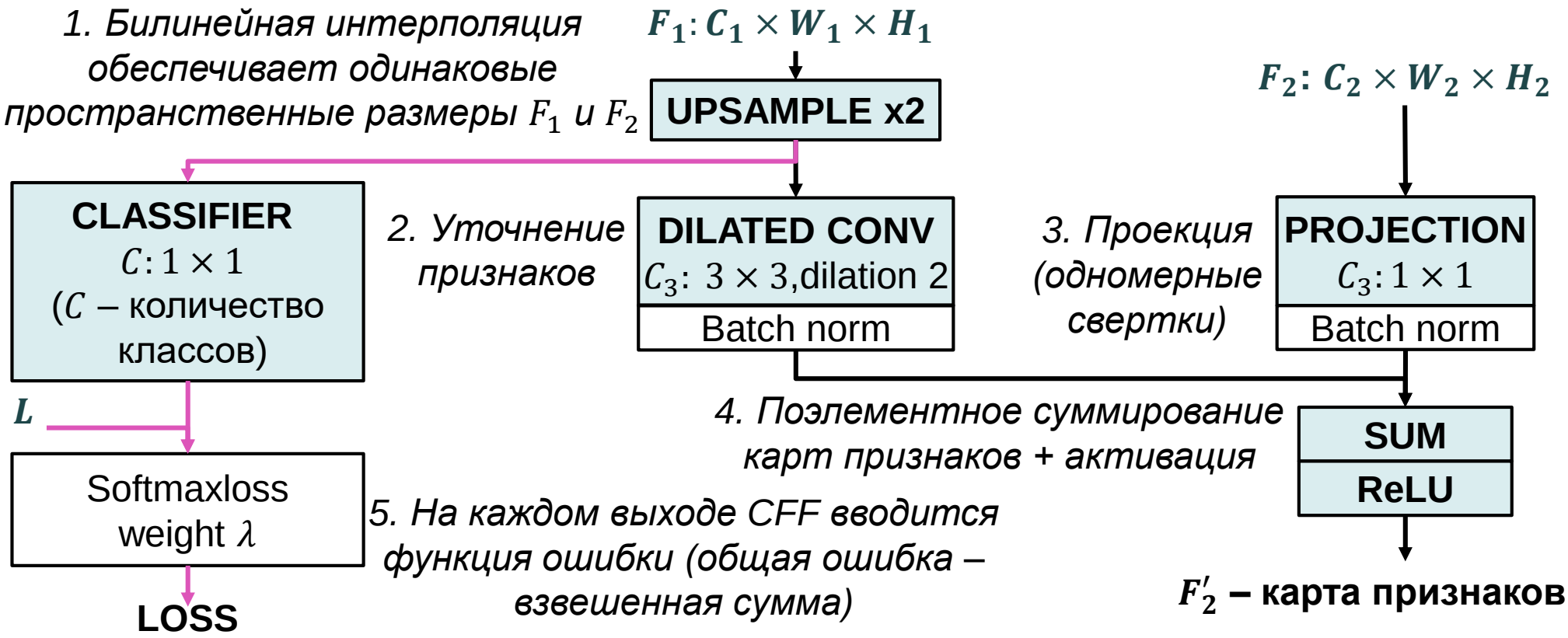
ICNet (4)

- ***Модуль слияния каскадных признаков*** получает на вход три основные компоненты:
  - Карта признаков F_1 размера $C_1 \times W_1 \times H_1$ (используется при обучении и тестировании)
  - Карта признаков F_2 размера $C_2 \times W_2 \times H_2$ (используется при обучении и тестировании). Пространственные размеры F_2 вдвое больше F_1
  - Разметка изображения L размера $1 \times W_2 \times H_2$ (используется при обучении)
- В результате слияния карт F_1 и F_2 формируется объединенная карта признаков F_2', которая учитывается на следующем (более крупном) масштабе

* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

ICNet (5)

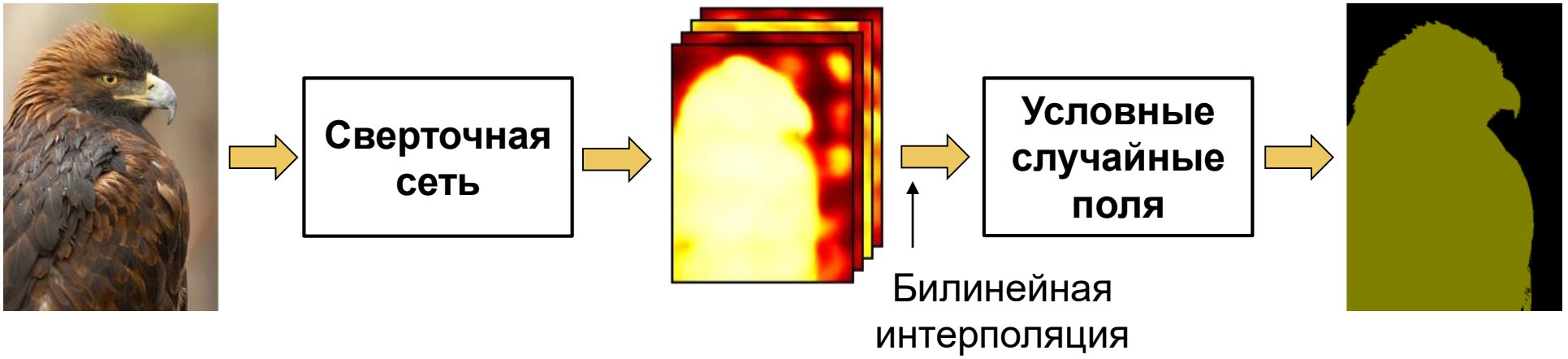
- ***Модуль слияния каскадных признаков:***



* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

DeepLab-v1 (1)

- DeepLab-v1 – один из широко известных методов семантической сегментации, основанный на построении сверточной нейронной сети для получения «грубой» карты сегментов и последующем применении условных случайных полей (Conditional Random Fields, CRF) для уточнения полученных результатов

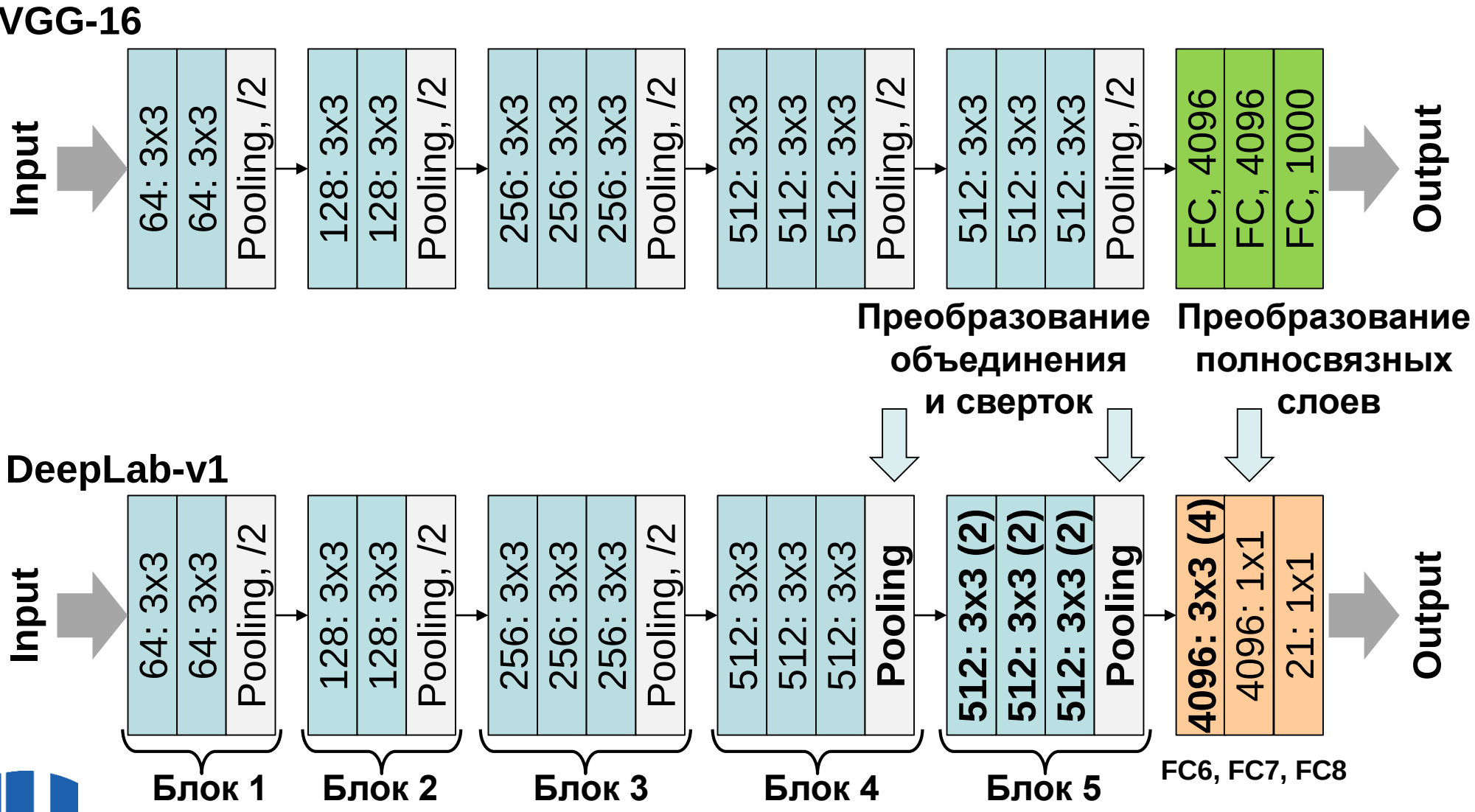


* Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. – 2014. – [https://arxiv.org/pdf/1412.7062.pdf].

DeepLab-v1 (2)

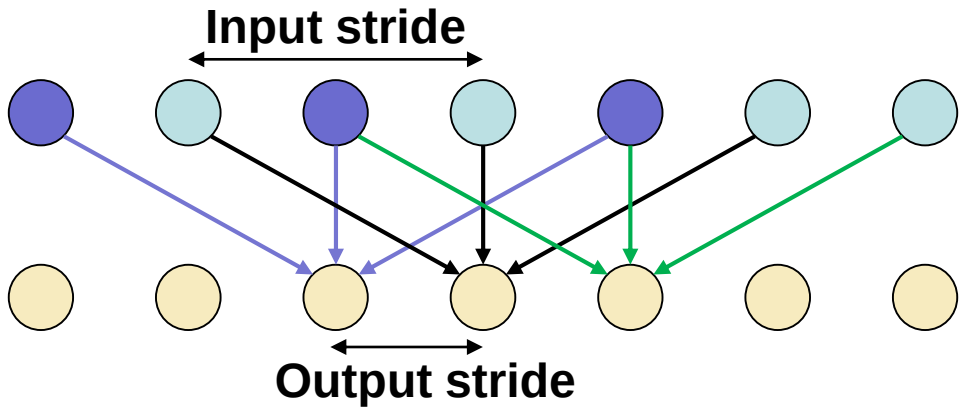
- Сверточная сеть построена на базе VGG-16, обученной для классификации изображений набора данных ImageNet
- Основные отличия:
  - Полносвязные слои преобразованы в полностью сверточные, в результате чего на входе сети может быть передано изображение любого разрешения
  - Вход сети – 513x513 пикселей
  - Выход сети – 21, что соответствует количеству классу в наборе данных PASCAL VOC (вместо 1 000)
  - Для последних двух слоев пространственного объединения по максимуму удаляется понижение дискретизации и модифицируются сверточные слои, следующие за объединениями (3 последние свертки и первый полносвязный слой)

DeepLab-v1 (3)


VGG-16
Input
64: 3x3
64: 3x3
Pooling, /2
128: 3x3
128: 3x3
Pooling, /2
256: 3x3
256: 3x3
256: 3x3
Pooling, /2
512: 3x3
512: 3x3
512: 3x3
Pooling, /2
512: 3x3
512: 3x3
512: 3x3
Pooling, /2
FC, 4096
FC, 4096
FC, 1000
Output
Преобразование объединения и сверток
Преобразование полносвязных слоев
DeepLab-v1
Input
64: 3x3
64: 3x3
Pooling, /2
128: 3x3
128: 3x3
Pooling, /2
256: 3x3
256: 3x3
256: 3x3
Pooling, /2
512: 3x3
512: 3x3
512: 3x3
Pooling
512: 3x3 (2)
512: 3x3 (2)
512: 3x3 (2)
Pooling
4096: 3x3 (4)
4096: 1x1
21: 1x1
Output
Блок 1
Блок 2
Блок 3
Блок 4
Блок 5
FC6, FC7, FC8

DeepLab-v1 (4)

- Модифицированные свертки используют алгоритм «дырки» («atrous» algorithm):
  - Размеры ядер сверток не меняются
  - Ядра накладываются с пропусками («дырками»)
  - Расстояние между элементами ядер для трех сверточных слоев составляет 2, для первого полносвязного – 4
- Пример одномерной свертки с «дыркой» в 2 элемента:



* Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. – 2014. – [https://arxiv.org/pdf/1412.7062.pdf].

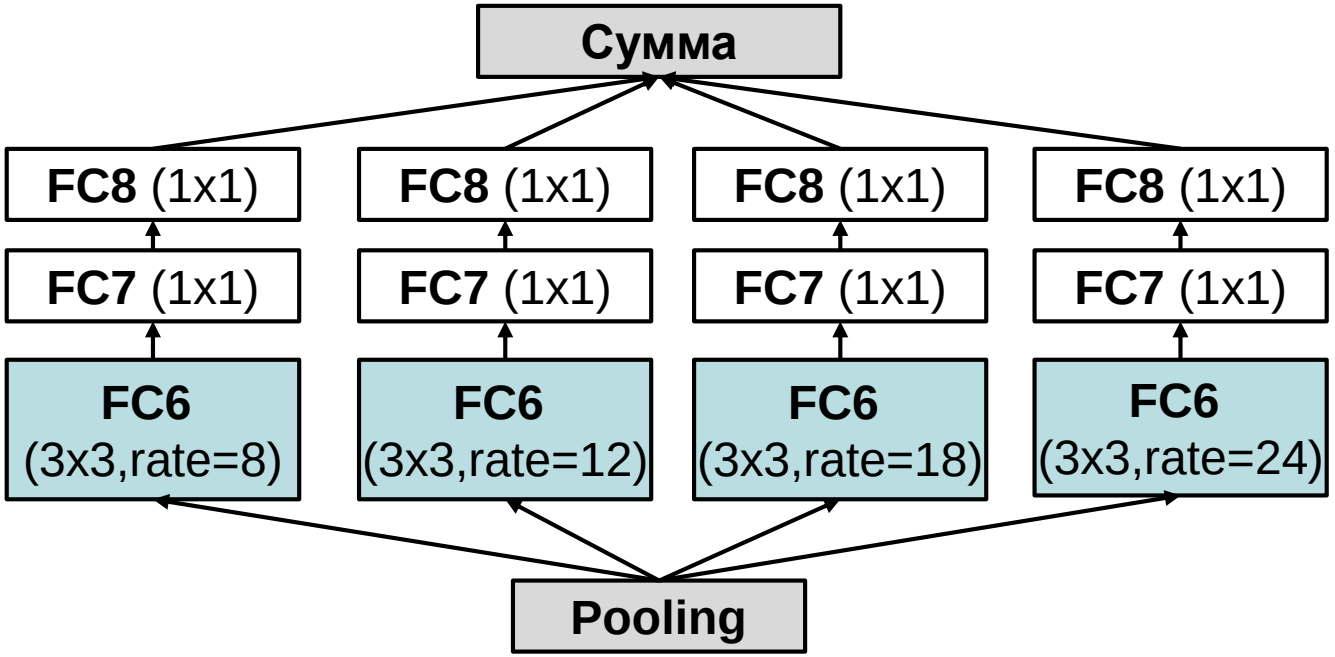
DeepLab-v2 (1)

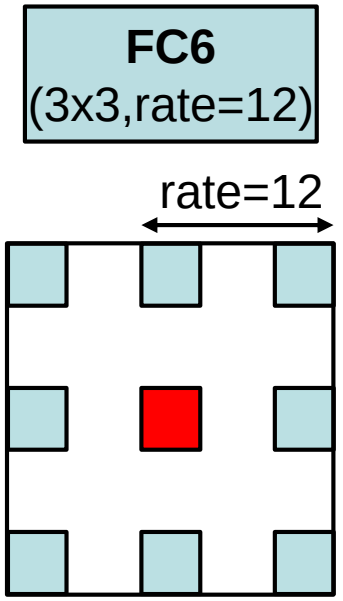
- DeepLab-v2 – модификация DeepLab-v1, разработанная с целью повышения производительности модели
- Решается проблема сегментации объектов, принадлежащих одинаковым классам, но имеющих разный масштаб
- Стандартный подход к решению данной проблемы – масштабирование изображения и агрегация карт признаков, построенных на разных масштабах
- Для реализации вводится ***пространственная пирамида сверток с пропусками*** (Atrous Spatial Pyramid Pooling, ASPP)
- Пространственная пирамида объединяет результаты применения сверток с пропусками для разных размеров «дырок» к некоторой карте признаков

* Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. – 2017. – [https://arxiv.org/pdf/1606.00915.pdf], [https://ieeexplore.ieee.org/document/7913730].

DeepLab-v2 (2)

- Структура пространственной пирамиды сверток с пропусками (FC6, FC7, FC8 – полностью сверточные слои):





Двумерная свертка с пропусками

* Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. – 2017. – [https://arxiv.org/pdf/1606.00915.pdf], [https://ieeexplore.ieee.org/document/7913730].

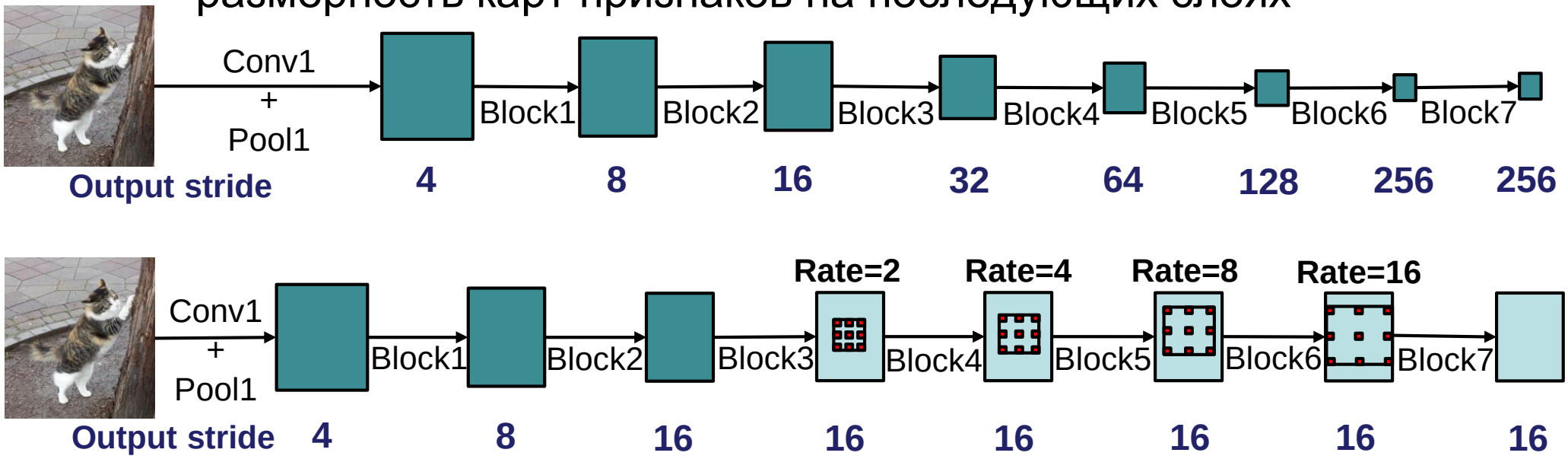
DeepLab-v3 (1)

- DeepLab-v3 – развитие модели DeepLab-v2
- Для решения проблемы сегментации объектов разного масштаба проектируются модули, построенные на свертках с пропусками
- Указанные модули организуются в каскадные или параллельные преобразования, чтобы захватить контекст с разных масштабов посредством введения разных размеров «дырок»
- Модуль с параллельными преобразованиями является расширением пространственной пирамиды сверток с пропусками

* Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].

DeepLab-v3 (2)

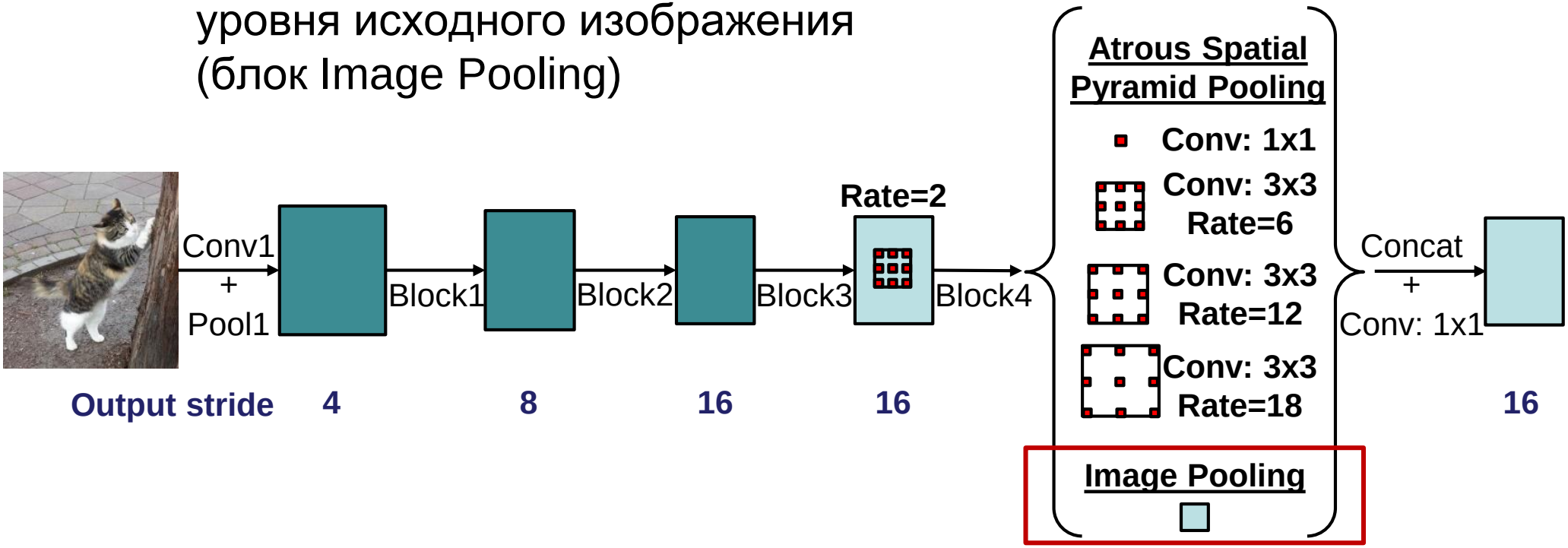
- Структура каскадного модуля:
  - Модель строится из последовательности остаточных блоков
  - Обычные свертки в последних остаточных блоках заменяются на свертки с пропусками, чтобы не снижалась размерность карт признаков на последующих слоях



* Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].

DeepLab-v3 (3)

- Структура параллельного модуля:
  - В пространственную пирамиду сверток с пропусками (Atrous Spatial Pyramid Pooling, ASPP) добавляются признаки уровня исходного изображения (блок Image Pooling)



* Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].

DeepLab-v3 (4)

- Расширение пространственной пирамиды сверток с пропусками (Atrous Spatial Pyramid Pooling, ASPP):
  - Для выделения признаков уровня изображения выполняются следующие преобразования:
    - Вычисление глобального среднего (global average pooling) для последней карты признаков модели
    - Свертка 1x1, 256 фильтров
    - Нормализация по пачке изображений (batch normalization)
    - Билинейная интерполяция карты признаков, чтобы пространственные размеры карт на выходе каждой ветки совпадали
  - Карты со всех ветвей пирамиды конкатенируются, применяются свертки 1x1 (256 фильтров), выполняется нормализация по пачке и финальная свертка 1x1

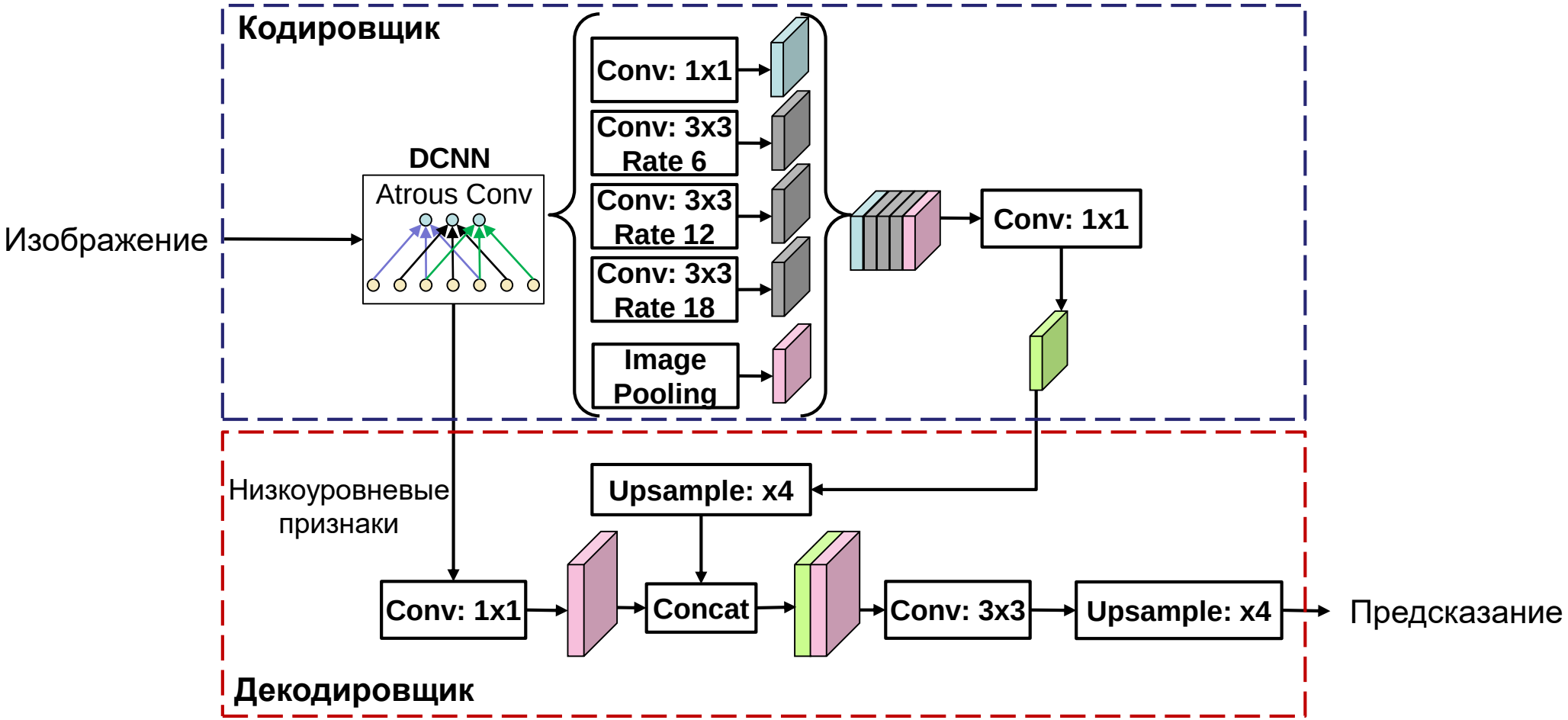
* Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].

DeepLab-v3+ (1)

- DeepLab-v3+ – модификация модели DeepLab-v3, направленная на повышение качества сегментации объектов на границах
- Модель построена на базе архитектуры «кодировщик-декодировщик»
  - Кодировщик представляет собой базовую часть модели DeepLab-v3 (все преобразования до финальной одномерной свертки)
  - Декодировщик составлен из сверток и операций повышающей дискретизации, применяемых к карте признаков уровня изображения и выходу пространственной пирамиды сверток с пропусками

* Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

DeepLab-v3+ (2)



* Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

DeepLab-v3+ (3)

- Для реализации кодировщика используется DeepLab-v3, ResNet-101 или Xception (соответствующие результаты экспериментов приведены в статье*)
- С целью оптимизации вычислений свертки с ядрами 3x3 преобразованы в стандартные отделимые по глубине свертки (depthwise separable convolutions)
  - Каждая свертка представляется пространственной (depthwise) и точечной (pointwise) сверткой
  - Пространственная свертка предполагает разбиение карты признаков на слои, применение свертки 3x3 глубины 1 к каждому слою и конкатенацию результатов сверток
  - Точечная свертка – свертка 1x1x<число слоев>

* Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

СРАВНЕНИЕ МОДЕЛЕЙ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ

Сравнение моделей семантической сегментации (1)

- Задача – семантическая сегментация дорожных сцен
- Тестовый набор данных Cityscapes [https://www.cityscapes-dataset.com]
- Показатель качества – среднее значение IoU (mean Intersection over Union, mean IoU)
- Сравнение* «качество-скорость» качественное, поскольку приведенные результаты экспериментов – результаты из оригинальных статей, полученные на разной тестовой инфраструктуре

- Результаты сравнения на других данных доступно по ссылке**

* Real-Time Semantic Segmentation on Cityscapes test [https://paperswithcode.com/sota/real-time-semantic-segmentation-on-cityscapes].
** Semantic Segmentation [https://paperswithcode.com/task/semantic-segmentation/latest].

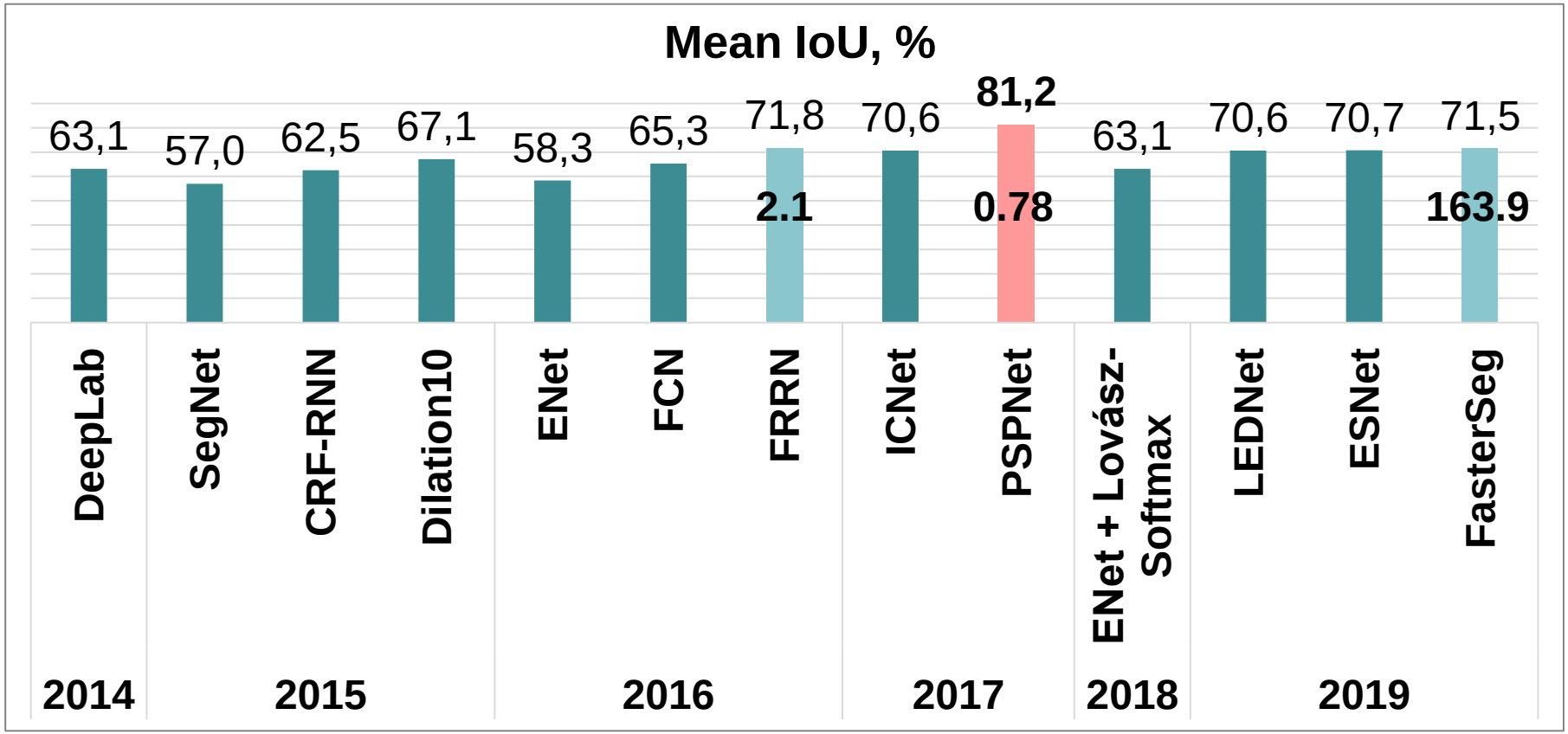
Сравнение моделей семантической сегментации (2)

Модель	Год	Mean IoU, %	FPS	Время, мс
DeepLab	2014	63.1	0.25	4000
SegNet	2015	57.0	16.7	60
CRF-RNN	2015	62.5	1.4	700
Dilation10	2015	67.1	0.25	4000
ENet	2016	58.3	76.9	13
FCN	2016	65.3	2	500
FRRN	2016	**71.8**	2.1	469
ICNet	2017	70.6	30.3	33
PSPNet	2017	**81.2**	**0.78**	**1288**
ENet + Lovász-Softmax	2018	63.1	76.9	13
LEDNet	2019	70.6	71	14
ESNet	2019	70.7	63	16
FasterSeg	2019	**71.5**	163.9	**6.1**

* Real-Time Semantic Segmentation on Cityscapes test [https://paperswithcode.com/sota/real-time-semantic-segmentation-on-cityscapes].

Сравнение моделей семантической сегментации (3)

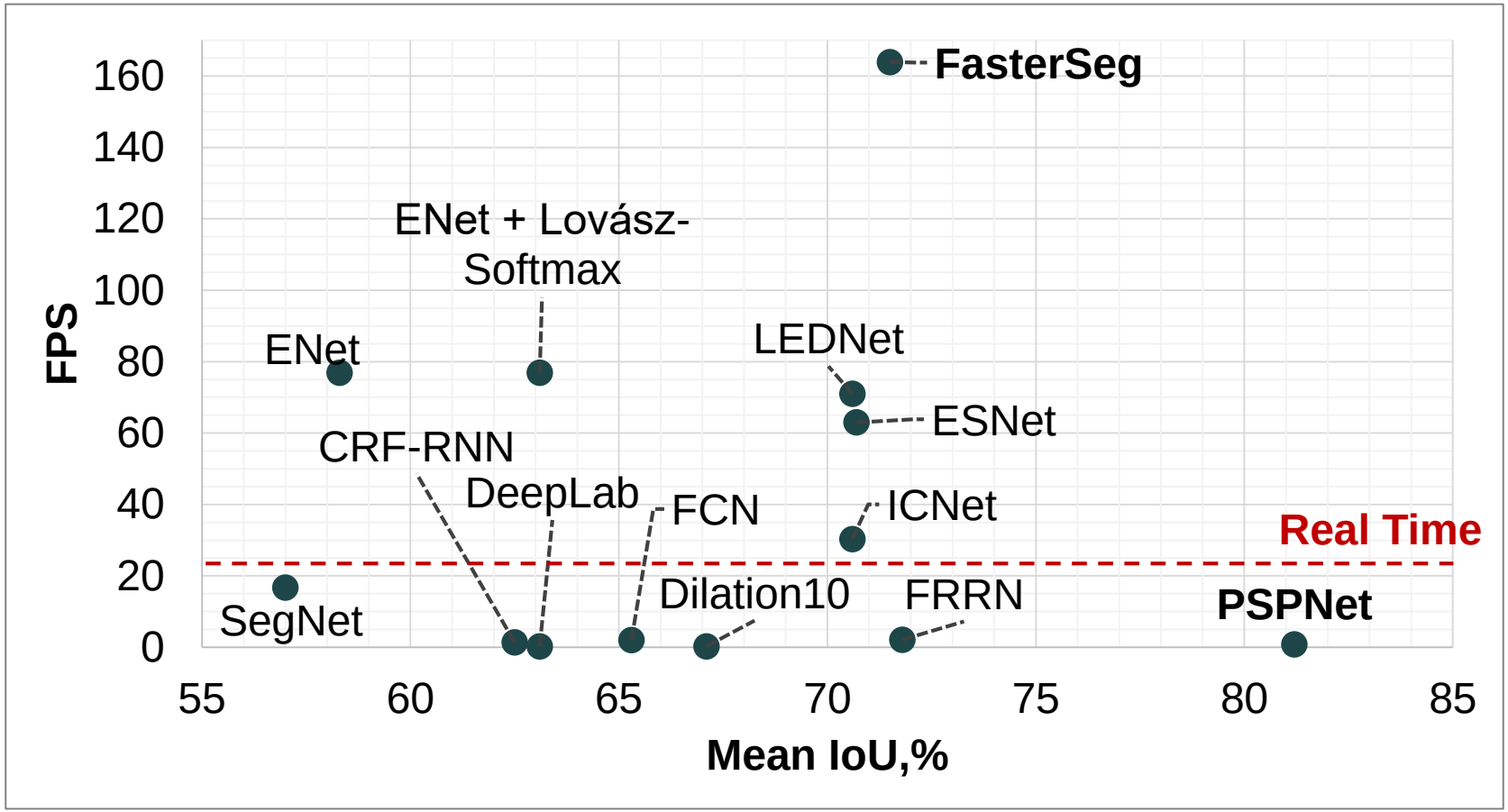
- Изменение среднего значения IoU для избранных моделей:



- ***За 2017-2019 качество варьируется от ~70 до ~81%, при этом лучшая модель является самой «медленной»***

Сравнение моделей семантической сегментации (4)

- ***Выбор эффективной модели – компромисс между качеством и скоростью работы***



Заключение

- Множество глубоких моделей для семантической сегментации изображений не ограничивается рассмотренными в лекции
- Основная проблема при построении моделей – получение выхода, пространственное разрешение которого совпадает с разрешением исходного изображения
- Рассмотренные модели по-разному решают указанную проблему. Как правило, решение в значительной степени влияет на скорость работы
- ***Оптимальная модель – компромисс между качеством и сложностью***
  - Качество определяется требованиями, предъявляемыми к решению практической задачи
  - Сложность определяется доступными вычислительными ресурсами и требованиями ко времени выполнения

Основная литература (1)

- Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].
- Badrinarayanan V., Kendall A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. – 2015. – [https://arxiv.org/pdf/1511.00561.pdf], [https://ieeexplore.ieee.org/document/7803544].
- Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation. – 2015. – [https://arxiv.org/pdf/1505.04597.pdf], [https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28].

Основная литература (2)

- Zhao H., Shi J., Qi X., Wang X., Jia J. Pyramid scene parsing network. – 2016. – [https://arxiv.org/pdf/1612.01105.pdf], [https://ieeexplore.ieee.org/document/8100143].
- Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].
- Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. – 2014. – [https://arxiv.org/pdf/1412.7062.pdf].

Основная литература (3)

- Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. – 2017. – [https://arxiv.org/pdf/1606.00915.pdf], [https://ieeexplore.ieee.org/document/7913730].
- Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].
- Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

Авторский коллектив

- **Турлапов Вадим Евгеньевич**
  д.т.н., профессор кафедры МОСТ ИИТММ ННГУ
  vadim.turlapov@itmm.unn.ru
- **Васильев Евгений Павлович**
  преподаватель кафедры МОСТ ИИТММ ННГУ
  evgeny.vasiliev@itmm.unn.ru
- **Гетманская Александра Александровна**
  преподаватель кафедры МОСТ ИИТММ ННГУ
  alexandra.getmanskaya@itmm.unn.ru
- **Кустикова Валентина Дмитриевна**
  к.т.н., доцент каф. МОСТ ИИТММ ННГУ
  valentina.kustikova@itmm.unn.ru