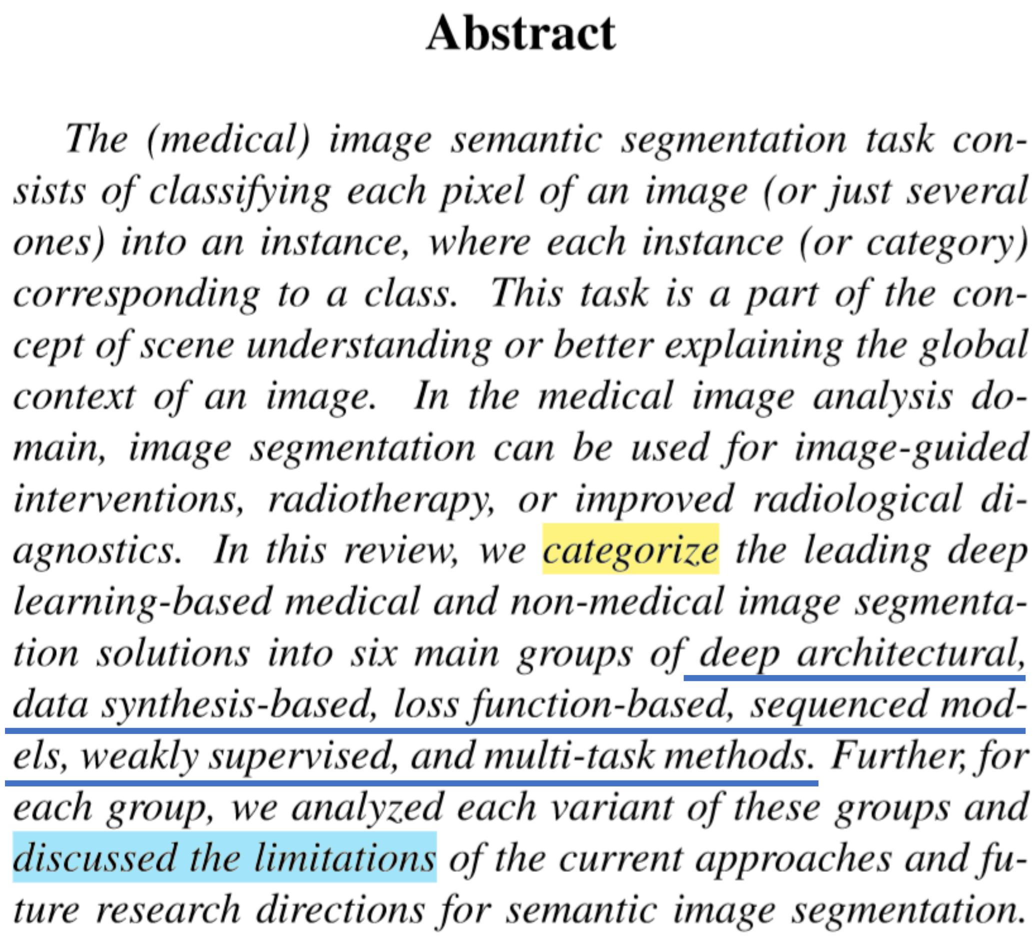
Deep semantic Segmentation of Natural and Medical Images

Index

Abstract



Abstract

The (medical) image semantic segmentation task consists of classifying each pixel of an image (or just several ones) into an instance, where each instance (or category) corresponding to a class. This task is a part of the concept of scene understanding or better explaining the global context of an image. In the medical image analysis domain, image segmentation can be used for image-guided interventions, radiotherapy, or improved radiological diagnostics. In this review, we categorize the leading deep learning-based medical and non-medical image segmentation solutions into six main groups of deep architectural, data synthesis-based, loss function-based, sequenced models, weakly supervised, and multi-task methods. Further, for each group, we analyzed each variant of these groups and discussed the limitations of the current approaches and future research directions for semantic image segmentation.

- **Categorize** the leading deep learning-based medical and non-medical image segmentation solution
  - Deep architectural
  - Data synthesis-based
  - Loss function-based
  - Sequenced models
  - ~~Weakly supervised~~
  - Multi-task methods
- **Discuss the limitations** of the current approaches and future research direction for semantic image segmentation

1. Introduction

- We provide comprehensive coverage of research contributions in the field of natural and medical image semantic segmentation. In terms of imaging modalities, we cover both 2D (RGB and grayscale) and volumetric medical images.

- We group the semantic image segmentation literature into six different categories based on the nature of their contributions - architectural improvements, optimization function based improvements, data synthesis based improvements, weakly supervised models, sequenced models, and multi-task models. Figure 1 indicates the categories we cover in this review.

- Followed by the comprehensive review, we recognize and suggest the important research directions for each of the categories.

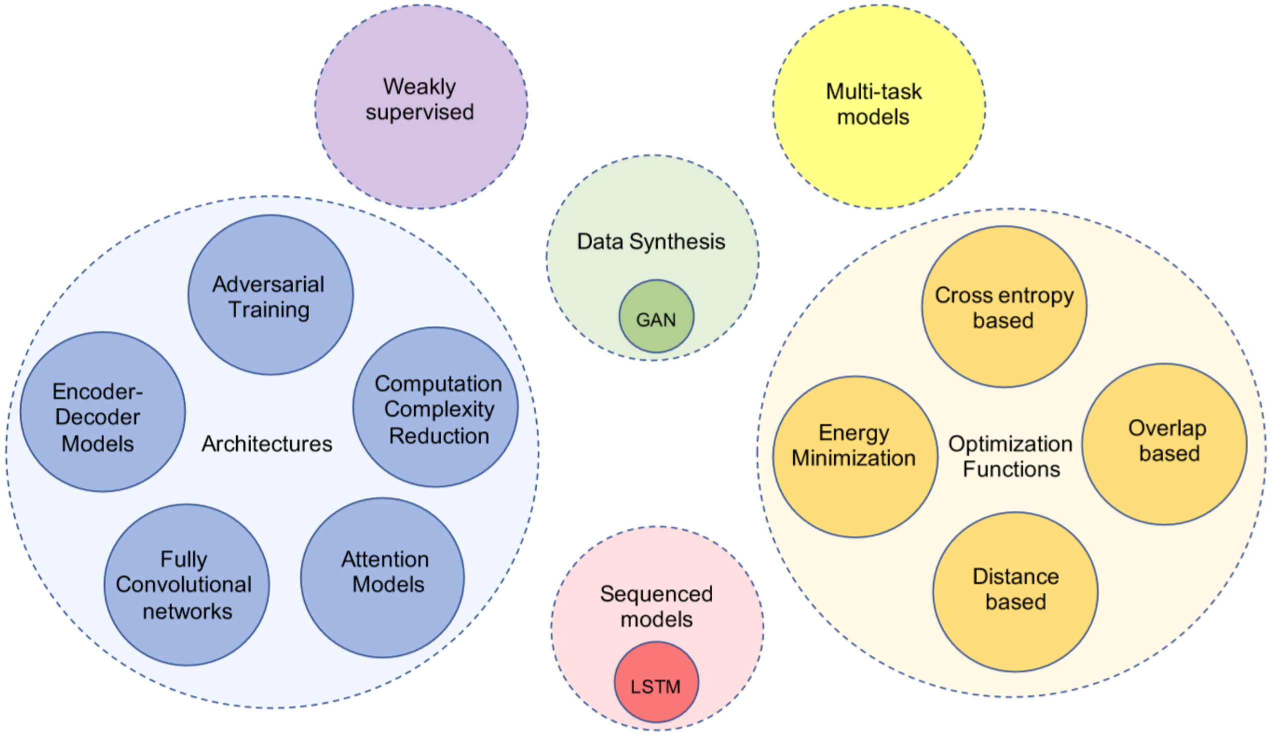


Figure 1: An overview of the deep learning based segmentation methods covered in this review.

→ For each category, we first review the improvements on non-medical datasets, and in a subsequent section, we survey the improvements for medical images.

2. Network Architectural Improvements

2.1. **Fully Convolutional Neural Networks** for Semantic Segmentation

- FCN : 어려운 기법 사용하지 않았음에도 불구하고 semantic segmentation 분야에서 뛰어난 성능
- FCN의 기본 아이디어 : 마지막 Fully Connected Layer → **1x1 Convolution Layer**로 수정 (위치 정보 보존 효과)
  - 하지만 단순하게 1x1 conv layer를 통해 나온 score를 up-sampling 하면, 성능을 기대하기 어려움.
  - 그래서 **skip-connection** 사용! (많은 conv + pooling을 거치면서 작아진 feature map은 detail한 data를 잃었기 때문에 보강하자)
- 한계점 : 미리 정해진 receptive field를 사용 / 해상도 떨어짐

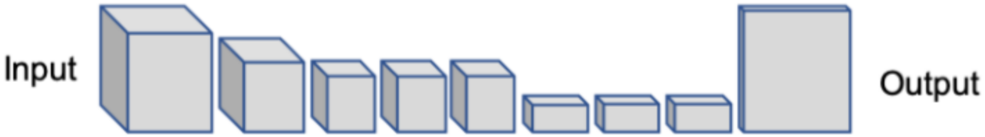


Figure 2: Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation [86].

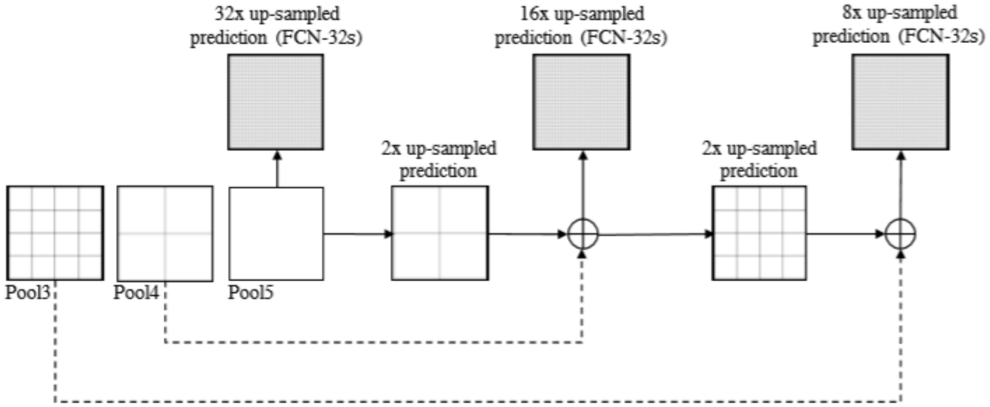

32x up-sampled prediction (FCN-32s)
16x up-sampled prediction (FCN-32s)
8x up-sampled prediction (FCN-32s)
2x up-sampled prediction
2x up-sampled prediction
Pool3
Pool4
Pool5

Figure 3: Upsampling and fusion step of the fully convolution networks [86].

2. Network Architectural Improvements

2.2. Encoder-decoder Semantic Image Segmentation Networks

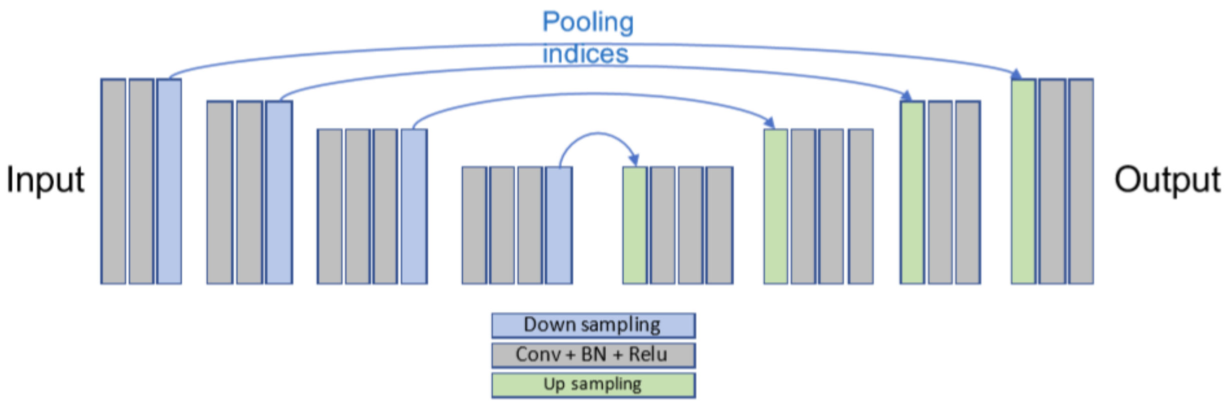
- Such as **SegNet**

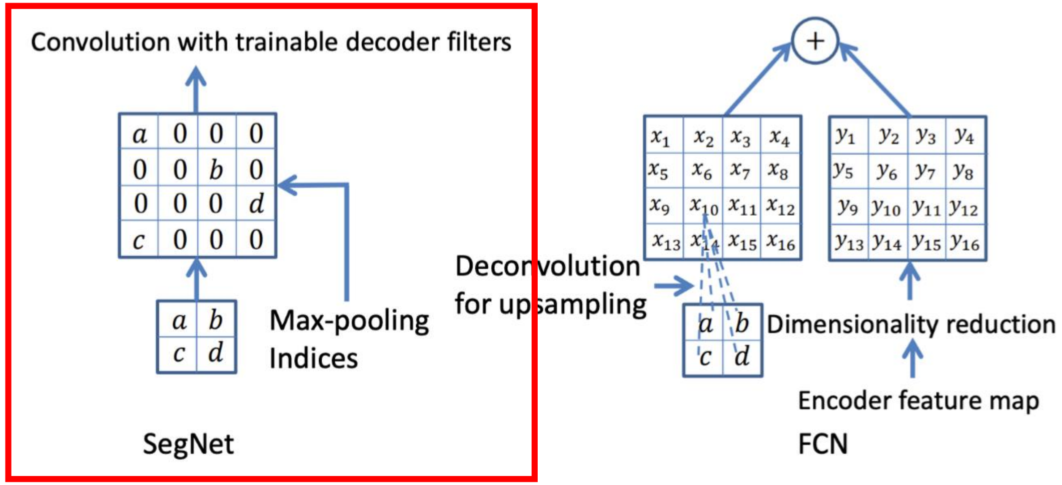
1) **Encoder** Network

- Convolution과 max pooling을 수행하는데, VGG16에서 FCL을 뺀 13개의 Conv later를 그대로 사용한다.
- 2x2 max pooling을 수행하는 동안에 해당하는 **Max Pooling Index (위치 정보를 의미)**를 저장한다.
- 각각의 encode은 하나 이상의 conv layer에 batch normalization과 ReLU가 포함되어있다.

2) **Decoder** Network

- Up-sampling과 Convolution을 수행하며, 마지막 later에는 각 픽셀의 class 예측을 위한 softmax classifier가 존재한다.
- Up-sampling은 Encoder에서 저장한 **Max pooling Indices를 수신 받아** 수행한다.
- 마지막 layer에서는 K-class softmax 분류기를 사용하여 각 픽셀의 클래스를 예측한다.




Convolution with trainable decoder filters
a 0 0 0
0 0 b 0
0 0 0 d
c 0 0 0
a b
c d
Max-pooling Indices
SegNet
Deconvolution for upsampling
x_1 x_2 x_3 x_4
x_5 x_6 x_7 x_8
x_9 x_{10} x_{11} x_{12}
x_{13} x_{14} x_{15} x_{16}
y_1 y_2 y_3 y_4
y_5 y_6 y_7 y_8
y_9 y_{10} y_{11} y_{12}
y_{13} y_{14} y_{15} y_{16}
+
Dimensionality reduction
Encoder feature map
FCN

2. Network Architectural Improvements

2.2. Encoder-decoder Semantic Image Segmentation Networks

- Such as **Unet**
- Add ***skip connections*** to the encoder-decoder
- Unet과 SegNet의 차이점
  - U-net은 주로 bio image에 사용되며,
  - Pooling Index 대신 skip connection을 사용함
    - 전체 feature map이 encoder에서 decoder로 전송
    - 이는 더 많은 메모리를 사용하게 함
  - SegNet이 model의 accurac를 향상시키고,
  - Vanishing gradient 문제를 줄임

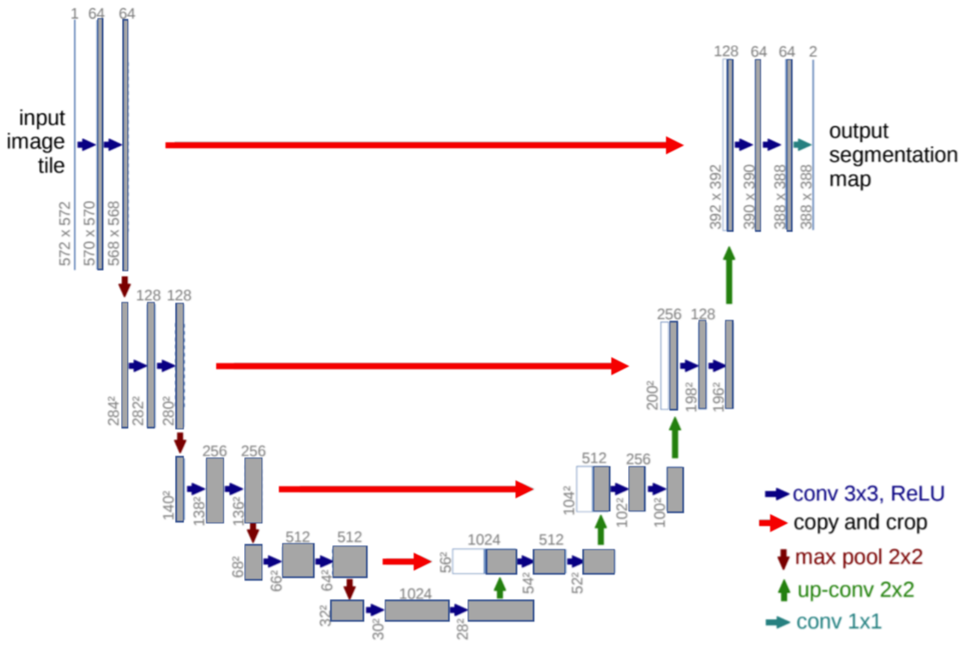


Figure 5: An illustration of the U-Net [119] architecture.

2. Network Architectural Improvements

2.2. Encoder-decoder Semantic Image Segmentation Networks

- Such as **Vnet**
- Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation
- MRI에서 전립선 검출을 위해 V 모양의 네트워크 설계
    - 3차원 볼륨 데이터를 입출력으로 사용
    - Dice coefficient를 사용한 loss function 구성
    - Residual 기반 학습 적용

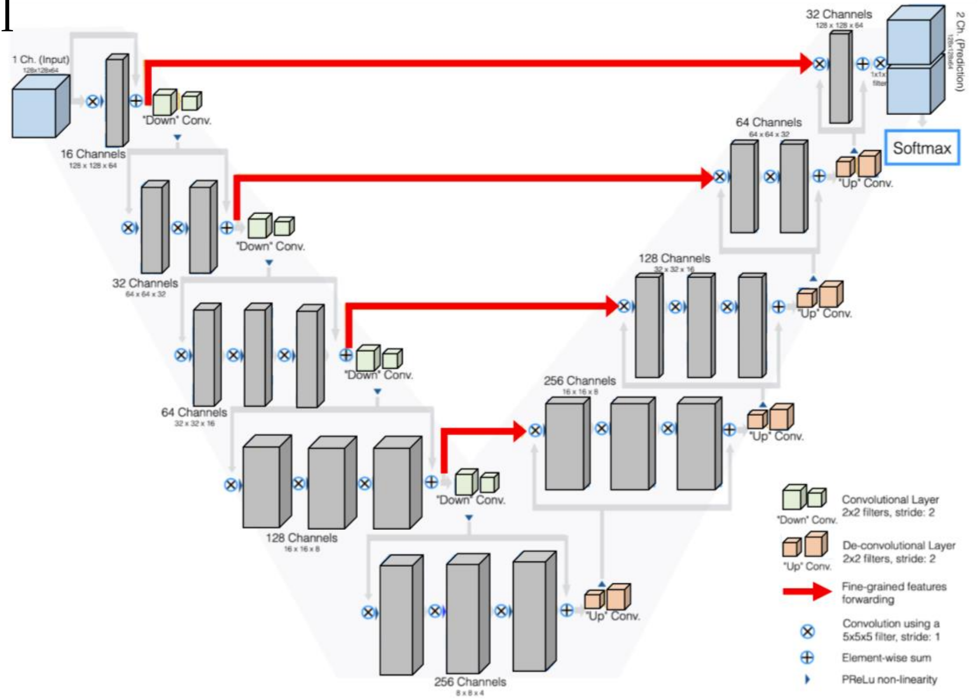


Figure 6: An illustration of the V-Net [91] architecture.

2. Network Architectural Improvements

2.2. Encoder-decoder Semantic Image Segmentation Networks

- Such as **Tiramisu** (Fully Convolutional Densenets for Semantion Segmentation)
- Adapt the U-Net like **encoder-decoder skeleton**.

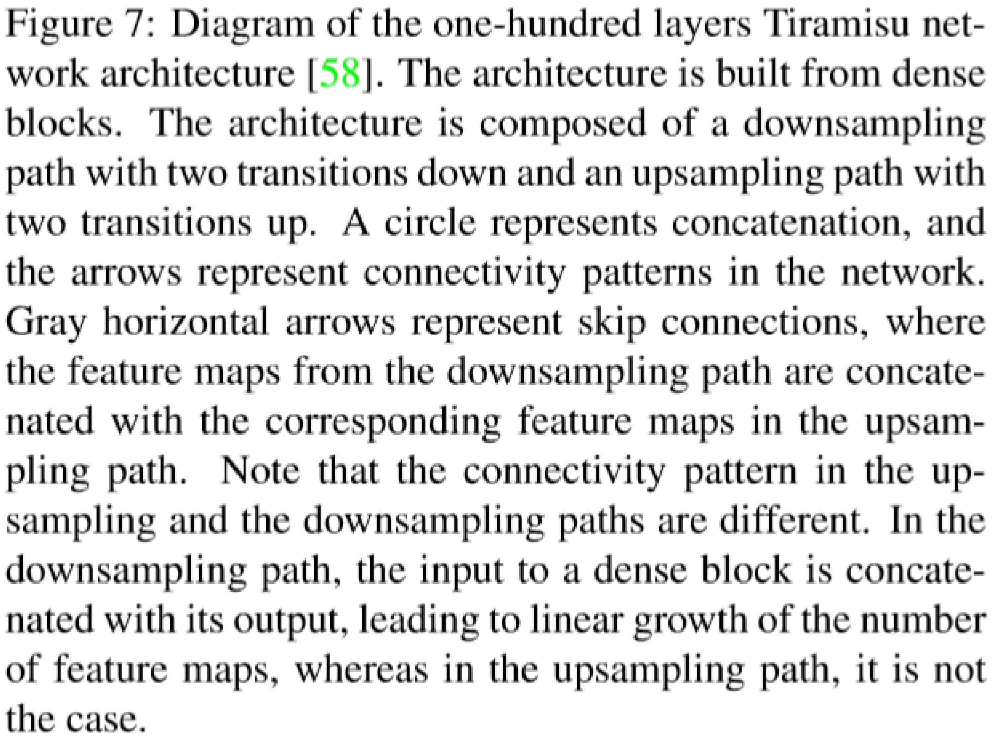
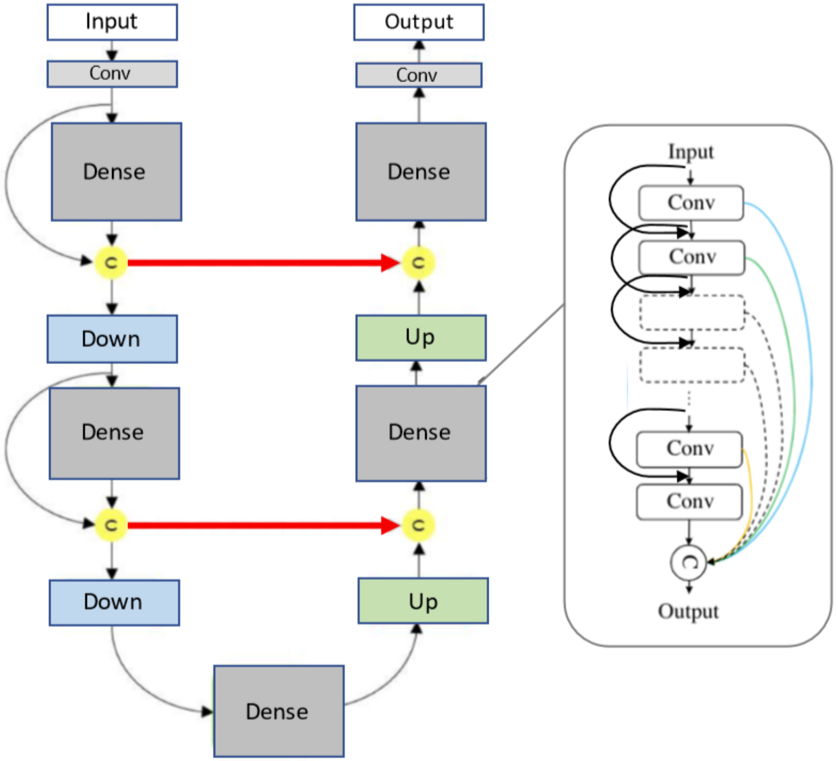
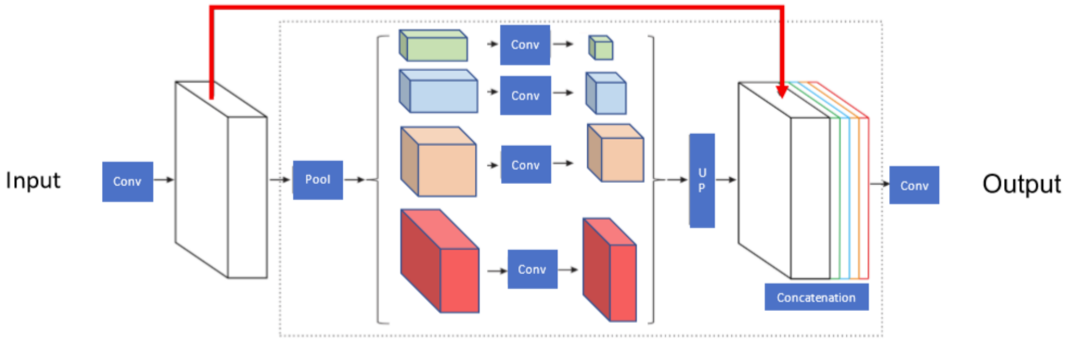


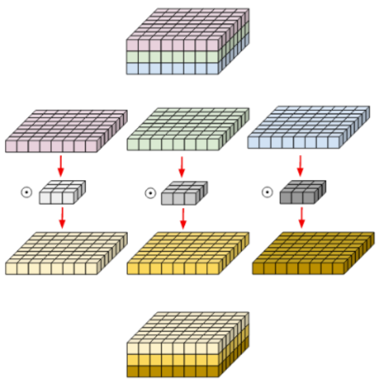
Figure 7: Diagram of the one-hundred layers Tiramisu network architecture [58]. The architecture is built from dense blocks. The architecture is composed of a downsampling path with two transitions down and an upsampling path with two transitions up. A circle represents concatenation, and the arrows represent connectivity patterns in the network. Gray horizontal arrows represent skip connections, where the feature maps from the downsampling path are concatenated with the corresponding feature maps in the upsampling path. Note that the connectivity pattern in the upsampling and the downsampling paths are different. In the downsampling path, the input to a dense block is concatenated with its output, leading to linear growth of the number of feature maps, whereas in the upsampling path, it is not the case.



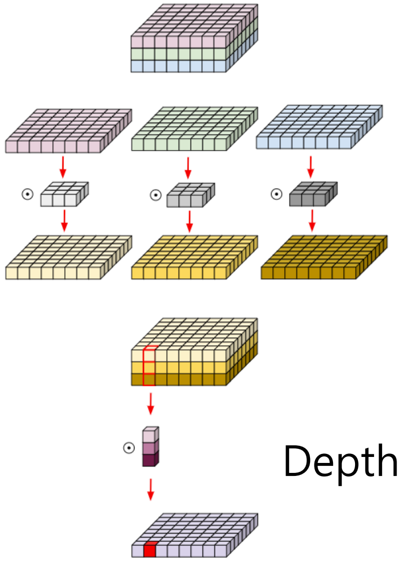
2. Network Architectural Improvements



Pyramid Pooling



Depth-wise Conv



Depth-wise Separable Conv

- **DeepLabV3+**
  - State-of-the-art segmentation networks on PASCAL VOC 2012 and Cityscapes datasets.
  - **Depth-wise Separable Convolution + Atrous Convolution**
    - Depth-wise Separable Convolution
      - Depth-wise Convolution : channel 분리 후 각각 conv 연산
      - Depth-wise Convolution 에서 나온 결과에 1x1xC 크기의 convolution 필터를 적용
      - 즉, 각 channel output이 하나로 합쳐지는 게 특징
    - Atrous Convolution
      - 필터에 빈 공간을 둔 채 작용하는 convolution
      - 기존 conv와 파라미터 수는 동일하지만 field of view가 커짐

2. Network Architectural Improvements

2.3. Computation Complexity Reduction for Image Segmentation Networks

- Several works have been done on **reducing the time and the computational complexity** of deep classification networks.

  1. **simplify the structure of deep networks** by tensor factorization[69], channel/network pruning [152], or applying sparsity to connections.
  2. **focused on the complexity optimization** of deep image segmentation networks. **proposed a hierarchical neural architecture search for semantic image segmentation** by performing both cell and network-level search and achieved comparable results to the state-of-the-art results on the PASCAL VOC 2012 and Cityscapes datasets.
  3. **focused on searching the much smaller atrous spatial pyramid pooling module using random search.**
  4. Besides network architecture search, Srivastava et al. [132] **modified ResNet in a way to control the flow of information through a connection.**
  5. Lin et al. **adopted one step fusion without filtering the channels.**

2. Network Architectural Improvements

2.4. **Attention**-based Semantic Image Segmentation

- Attention can be viewed as **using information transferred from several subsequent layers/feature maps to select and localize the most discriminative (or salient) part of the input signal.**

  1. a selection mechanism where feature maps are first aggregated using global average pooling and reduced to a single channel descriptor. Then an activation gate is used to highlight the most discriminative features.
  2. added an attention module to the deep residual network (ResNet) for image classification. Their proposed attention module consists of several encoding-decoding layers.
  3. proposed dual attention networks that apply both spatial and channel based attention operations.
  4. proposed a pyramid attention based network, for semantic segmentation.
  5. applied attention to DeepLab which takes multi-scale inputs.

2. Network Architectural Improvements

2.5. **Adversarial** Semantic Image Segmentation

- Goodfellow et al. [37] proposed **an adversarial approach to learn deep generative models.**
- Inspired by adversarial learning, Luc et al. [87] trained a convolutional semantic segmentation network along with an adversarial network that **discriminates segmentation maps coming either from the ground truth or from the segmentation network**. Their loss function is defined as
  - In this setup, the segmentor tries to produce segmentation maps that are close to the ground truth, i.e., which look more realistic.

4. **Optimization Function** based Improvements

4.1. **Cross Entropy**

- The most commonly used loss function for the task of image segmentation is a pixel-wise cross entropy loss.
- This loss examines each pixel individually, comparing the class predictions vector to the one-hot encoded target(or ground truth) vector.

$$\mathrm{CE}(p, \hat{p}) = -(p \log(\hat{p}) + (1 - p) \log(1 - \hat{p})). \qquad (2)$$

- The general form of the equation for multi-region (or multi-class) segmentation can be written as:

$$\mathrm{CE} = -\sum_{classes} p \log \hat{p} \qquad (3)$$

Ex1) 2 class 일 때, label = [0,1]이고, prediction = [0.2, **0.8**] 이면 cross entropy 값은 – (0log(0.2) + 1log(0.8)) = **0.0969 (O)**

Ex2) 2 class 일 때, label = [0,1]이고, prediction = [**0.8**, 0.2] 이면 cross entropy 값은 – (0log(0.8) + 1log(0.2)) = **0.6990 (X)**

4. Optimization Function based Improvements

4.2. Weighted Cross Entropy

- Cross entropy loss can be problematic if the various classes have unbalanced representation in the image, as the most prevalent class can dominate training.

- Long et al.[86] discussed weighting the cross-entropy loss (WCE) for each class in order to counteract a class imbalance present in the dataset.

$$\mathrm{WCE}(p, \hat{p}) = -(\beta p \log(\hat{p}) + (1 - p) \log(1 - \hat{p})). \quad (4)$$

- **To decrease the number of false negatives**, β is set to a value larger than 1, and **to decrease the number of false positives,** β is set to a value smaller than 1.

- To weight the negative pixels as well, the following balanced cross-entropy (BCE) can be used [159].

$$\mathrm{BCE}(p, \hat{p}) = -(\beta p \log(\hat{p}) + (1 - \beta)(1 - p) \log(1 - \hat{p})). \quad (5)$$

4. Optimization Function based Improvements

4.3. **Focal Loss**

- To down-weight the contribution of easy examples so that the CNN focuses more on hard examples, Lin et al., [83] added the term (1− ˆ p)γ to the cross entropy loss as: setting γ = 0 the equation will be equivalent to BCE.

$$\begin{aligned} \mathrm{FL}(p,\hat{p}) = &-(\alpha(1-\hat{p})^{\gamma} p \log(\hat{p}) \\ &+(1-\alpha)\hat{p}^{\gamma}(1-p)\log(1-\hat{p})) \end{aligned} \tag{7}$$

- Focal loss는 object detection의 한 종류인 retinaNet에서 쓰인 loss function
- 잘 찾은 class에 대해서는 loss를 적게 주어 갱신을 거의 하지 못하게 하고, 못 찾은 class에 대해선 loss를 크게 주어서 갱신을 크게 하게 하는 것임. (gamma)

4. Optimization Function based Improvements

4.4. Overlap Measure based Loss Functions

4.4.1 Dice Loss / F1 Score

- Another popular loss function for image segmentation tasks is based on the Dice coefficient, which is essentially a **measure of overlap between two samples** and is **equivalent to the F1 score**.

- This measure ranges from 0 to 1, where a Dice coefficient of 1 denotes perfect and complete overlap. The Dice coefficient (DC) is calculated as: where X and Y are the predicted and ground truth segmentation, respectively.

$$\mathrm{DC} = \frac{2TP}{2TP + FP + FN} = \frac{2|X \cap Y|}{|X| + |Y|}. \qquad (8)$$

- To use this as a loss function the DC can be defined as a Dice loss (DL) function [91]:

$$\mathrm{DL}(p, \hat{p}) = \frac{2\langle p, \hat{p}\rangle}{\|p\|_1 + \|\hat{p}\|_1} \qquad (10)$$

4. Optimization Function based Improvements

4.4. Overlap Measure based Loss Functions

4.4.2 Tversky Loss

- Tversky loss (TL) [122] is a **generalization of the DL**. To **control the level of FP and FN**, TL weights them as the following: setting β = 0.5 simplifies the equation to Eq. 10.

$$\mathrm{DL}(p,\hat{p}) = \frac{2\langle p,\hat{p}\rangle}{\|p\|_1+\|\hat{p}\|_1} \tag{10}$$

$$\mathrm{TL}(p,\hat{p}) = \frac{\langle p,\hat{p}\rangle}{\langle p,\hat{p}\rangle + \beta(1-p,\hat{p}\rangle + (1-\beta)(p,1-\hat{p})} \tag{11}$$

4. Optimization Function based Improvements

4.4. Overlap Measure based Loss Functions

4.4.3 Exponential Logarithmic Loss

- Wong et al.[156] proposed using a weighted sum of the exponential logarithmic Dice loss(L_{eld}) and the weighted exponential cross-entropy loss(L_{wece}) in order to **improve the segmentation accuracy on small structures for tasks** where there is a large variability among the sizes of the objects to be segmented.

$$\mathcal{L} = w_{\mathrm{eld}}\mathcal{L}_{\mathrm{eld}} + w_{\mathrm{wece}}\mathcal{L}_{\mathrm{wece}}, \qquad (12)$$

$$\mathcal{L}_{\mathrm{eld}} = \mathbf{E}\left[\left(-\ln\left(D_i\right)\right)^{\gamma_D}\right], \text{ and} \qquad (13)$$

$$\mathcal{L}_{\mathrm{wece}} = \mathbf{E}\left[\left(-\ln\left(p_l(\mathbf{x})\right)\right)^{\gamma_{CE}}\right]. \qquad (14)$$

4. Optimization Function based Improvements

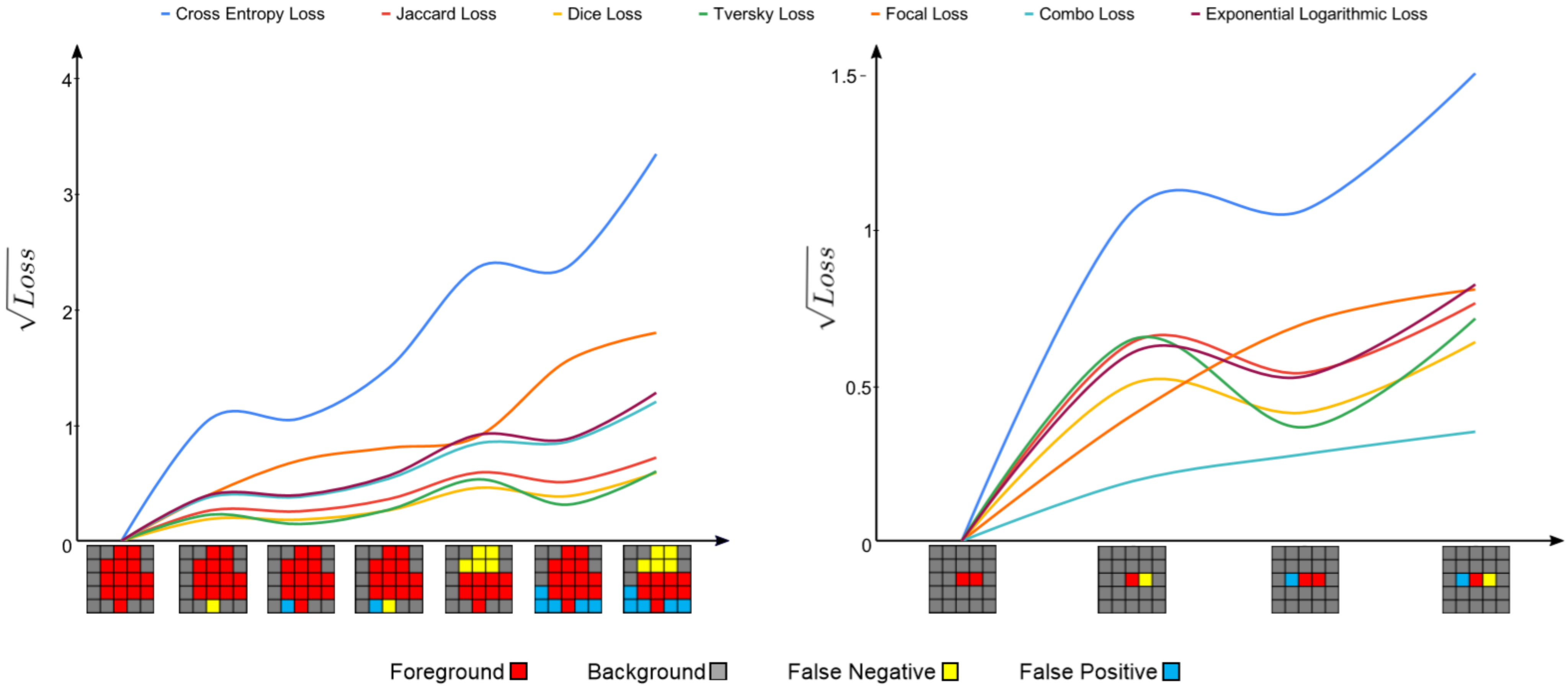


Figure 11: A comparison of seven loss functions for different extends of overlaps for a large (left) and a small (right) object.

8. Multitask Models

- Multi-task learning [16] refers to a machine learning approach where **multiple tasks are learned simultaneously**, and the learning efficiency and the model performance on each of the tasks are improved because of the existing commonalities across the tasks.

- Mask R-CNN has also been used for segmentation tasks in medical image analysis such as automatically segmenting and tracking cell migration in phase-contrast microscopy [142], detecting and segmenting nuclei from histological and microscopic images [60, 144, 145, 149], detecting and segmenting oral diseases [6], segmenting neuropathic ulcers [33], and labeling and segmenting ribs in chest X-rays[154].

9. Summary of the Main Models Tested on Natural Images

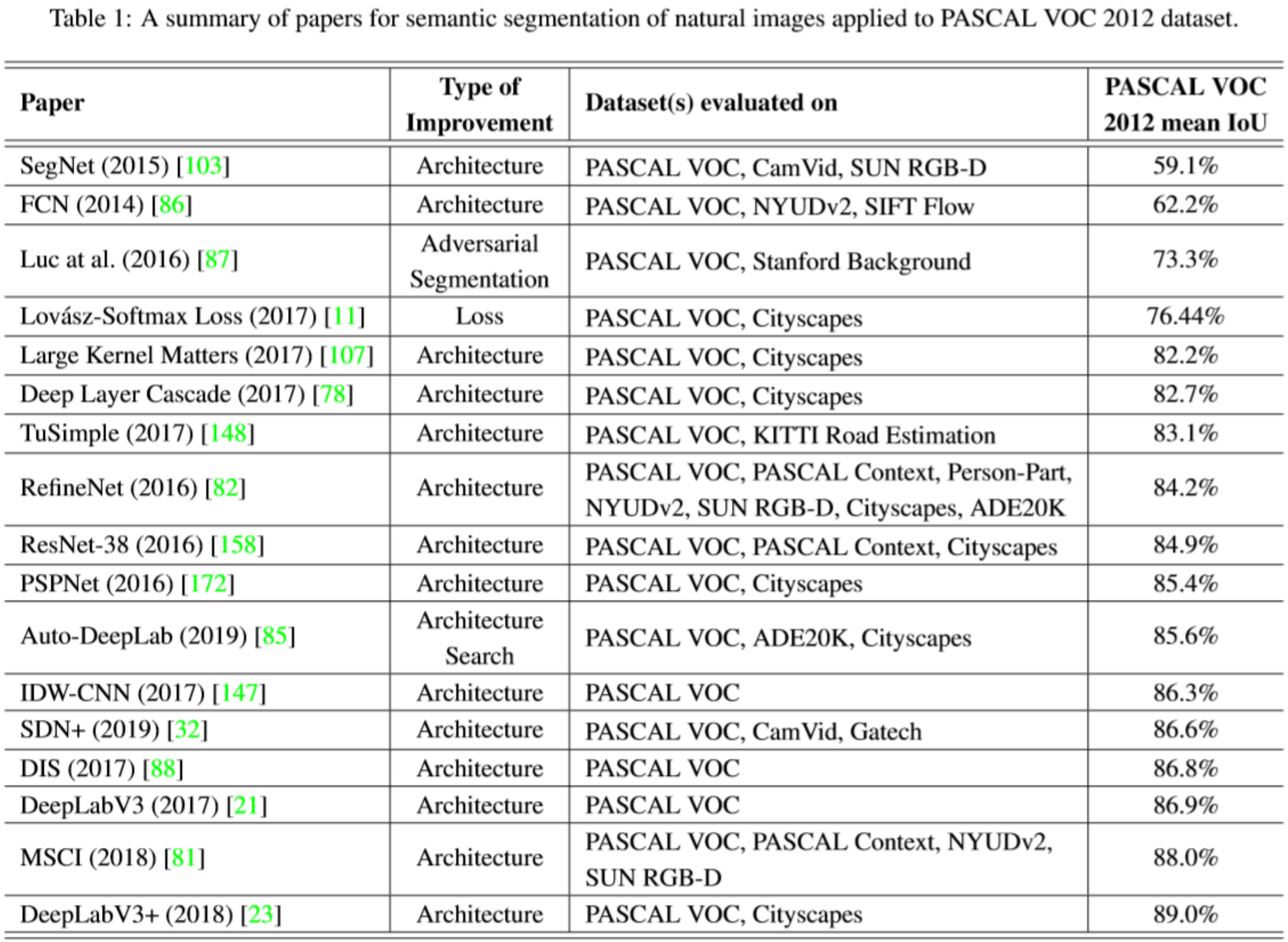


Table 1: A summary of papers for semantic segmentation of natural images applied to PASCAL VOC 2012 dataset.

Paper	Type of Improvement	Dataset(s) evaluated on	PASCAL VOC 2012 mean IoU
SegNet (2015) [103]	Architecture	PASCAL VOC, CamVid, SUN RGB-D	59.1%
FCN (2014) [86]	Architecture	PASCAL VOC, NYUDv2, SIFT Flow	62.2%
Luc at al. (2016) [87]	Adversarial Segmentation	PASCAL VOC, Stanford Background	73.3%
Lovász-Softmax Loss (2017) [11]	Loss	PASCAL VOC, Cityscapes	76.44%
Large Kernel Matters (2017) [107]	Architecture	PASCAL VOC, Cityscapes	82.2%
Deep Layer Cascade (2017) [78]	Architecture	PASCAL VOC, Cityscapes	82.7%
TuSimple (2017) [148]	Architecture	PASCAL VOC, KITTI Road Estimation	83.1%
RefineNet (2016) [82]	Architecture	PASCAL VOC, PASCAL Context, Person-Part, NYUDv2, SUN RGB-D, Cityscapes, ADE20K	84.2%
ResNet-38 (2016) [158]	Architecture	PASCAL VOC, PASCAL Context, Cityscapes	84.9%
PSPNet (2016) [172]	Architecture	PASCAL VOC, Cityscapes	85.4%
Auto-DeepLab (2019) [85]	Architecture Search	PASCAL VOC, ADE20K, Cityscapes	85.6%
IDW-CNN (2017) [147]	Architecture	PASCAL VOC	86.3%
SDN+ (2019) [32]	Architecture	PASCAL VOC, CamVid, Gatech	86.6%
DIS (2017) [88]	Architecture	PASCAL VOC	86.8%
DeepLabV3 (2017) [21]	Architecture	PASCAL VOC	86.9%
MSCI (2018) [81]	Architecture	PASCAL VOC, PASCAL Context, NYUDv2, SUN RGB-D	88.0%
DeepLabV3+ (2018) [23]	Architecture	PASCAL VOC, Cityscapes	89.0%

10. Discussion and Future Directions

10.1. Architectures

- Encoder-decoder networks with **long and short skip connections** are the winning architectures according to the state-of-the-art methods.
- Skip connections in deep networks have improved both segmentation and classification performance by **facilitating the training of deeper network architectures** and **reducing the risks for vanishing gradients.**

10.2. Sequenced Model

- For image segmentation, sequenced models can be used to **segment temporal data such as videos.**
- Ideally, **seeing the whole object of interest in a 3D volume might help to capture the geometrical information** of the object, which might be missed in processing a 3D volume slice by slice.

10. Discussion and Future Directions

10.3. Loss Functions

- In medical image segmentation works, researchers have converged toward using **classical cross-entropy loss functions** along with a second distance or overlap based functions.

10.4. Other Potential Directions

- **Modifying input** instead of the model, loss function, and adding more train data.
- Creating large 2D and 3D publicly available **medical benchmark datasets** for semantic image segmentation such as **the Medical Segmentation Decathlon [127].** This will allow researchers to accurately compare proposed approaches and make incremental improvements for specific datasets/problems.