Deep Semantic Segmentation of Natural and Medical Images: A Review

Saeid Asgari Taghanaki*[1], Kumar Abhishek[1], Joseph Paul Cohen[2], Julien Cohen-Adad[3], and Ghassan Hamarneh[1]

[1]School of Computing Science, Simon Fraser University, Canada
[2]Mila, Université de Montréal, Canada
[3]NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Canada

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.

1. Introduction

Deep learning has had a tremendous impact on various fields in science. The focus of the current study is on one of the most critical areas of computer vision: medical image analysis (or medical computer vision), particularly deep learning-based approaches for medical image segmentation. Segmentation is an important processing step in natural images for scene understanding and medical image analysis, for image-guided interventions, radiotherapy, or improved radiological diagnostics, etc. A plethora of deep learning approaches for medical image segmentation have been introduced in the literature for different medical imaging modalities, including X-ray, visible-light imaging (e.g. colour dermoscopic images), magnetic resonance imaging (MRI), positron emission tomography (PET), computerized tomography (CT), and ultrasound (e.g. echocardiographic scans). Deep architectural improvement has been a focus of many researchers for different purposes, e.g., tackling gradient vanishing and exploding of deep models, model compression for efficient small yet accurate models, while other works have tried to improve the performance of deep networks by introducing new optimization functions.

Compared to the other review papers on deep learning-based semantic image segmentation for natural and medical images [35, 41, 47, 52, 71, 140, 179], we make the following contributions:

- 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.

In the following sections, we discuss deep semantic image segmentation improvements under different categories visualized in Figure 1. For each category, we first review the improvements on non-medical datasets, and in a subsequent section, we survey the improvements for medical images.

*Corresponding author: sasgarit@sfu.ca

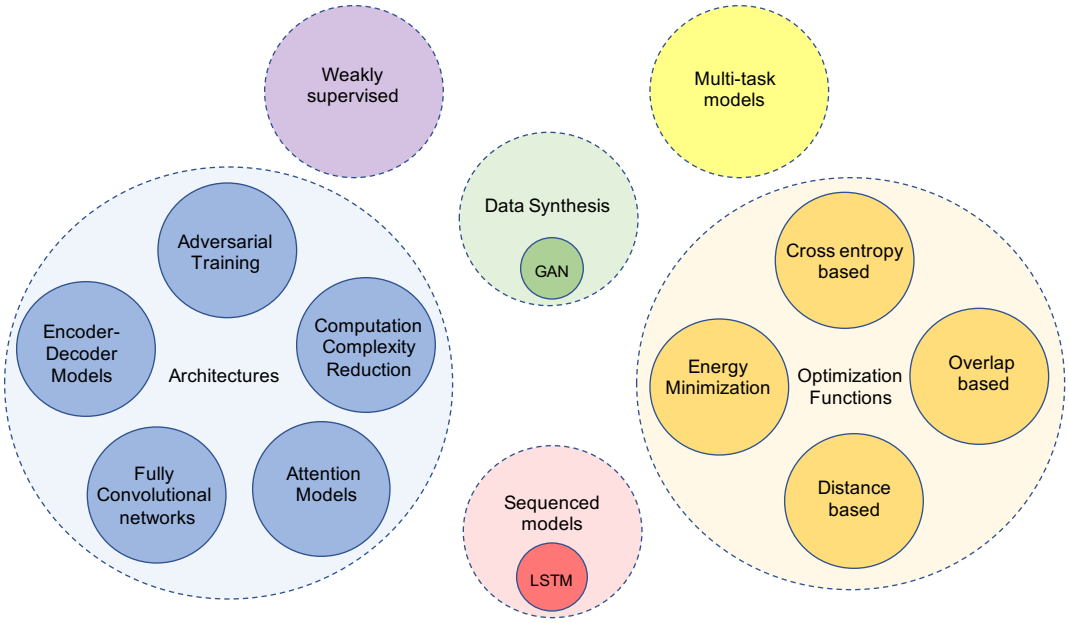

Figure 1: An overview of the deep learning based segmentation methods covered in this review.

2. Network Architectural Improvements

This section discusses the advancements in semantic image segmentation using convolutional neural networks (CNNs), which have been applied to interpretation tasks on both natural and medical images [36, 84]. The improvements are mostly attributed to exploring new neural architectures (with varying depths, widths, and connectivity or topology) or designing new types of components or layers.

2.1. Fully Convolutional Neural Networks for Semantic Segmentation

As one of the first high impact CNN-based segmentation models, Long et al. [86] proposed fully convolutional networks for pixel-wise labeling. They proposed up-sampling (deconvolving) the output activation maps from which the pixel-wise output can be calculated. The overall architecture of the network is visualized in Fig. 2.

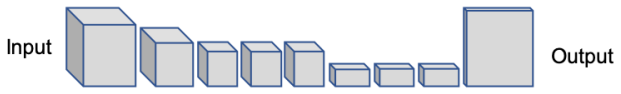

Figure 2: Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation [86].

In order to preserve the contextual spatial information within an image as the filtered input data progresses deeper into the network, Long et al. proposed to fuse the output with shallower layers' output. The fusion step is visualized in Fig. 3.

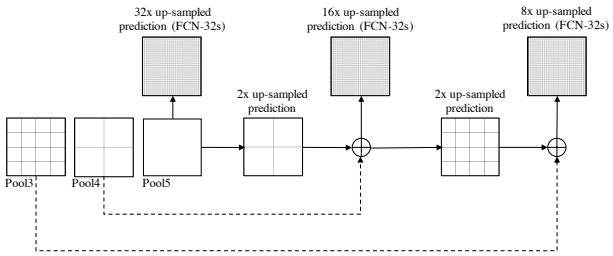

Figure 3: Upsampling and fusion step of the fully convolution networks [86].

2.2. Encoder-decoder Semantic Image Segmentation Networks

Next, encoder-decoder segmentation networks [103] such as SegNet, were introduced [7]. The role of the decoder network is to map the low-resolution encoder feature to full input resolution feature maps for pixel-wise classification. The novelty of SegNet lies in the manner in which the decoder upsamples the lower resolution input feature maps. Specifically, the decoder uses pooling indices (Figure 4) computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. The architecture (Fig. 4) consists of a sequence of non-linear processing layers (encoder) and a corresponding set of decoder layers followed by a pixel-wise classifier. Typically, each encoder consists of one or more convolutional layers with

batch normalization and a ReLU non-linearity, followed by non-overlapping max-pooling and sub-sampling. The sparse encoding due to the pooling process is upsampled in the decoder using the max-pooling indices in the encoding sequence.

Ronneberger et al. [119] proposed an architecture (U-Net) consisting of a contracting path to capture context and a symmetric expanding path that enables precise localization. Similar to the image recognition (He et al. [45]) and keypoint detection (Honari et al. [49]), Ronneberger et al. added *skip connections* (Fig. 5) to the encoder-decoder image segmentation networks e.g. SegNet, which improved the model's accuracy and addressed the problem of vanishing gradients.

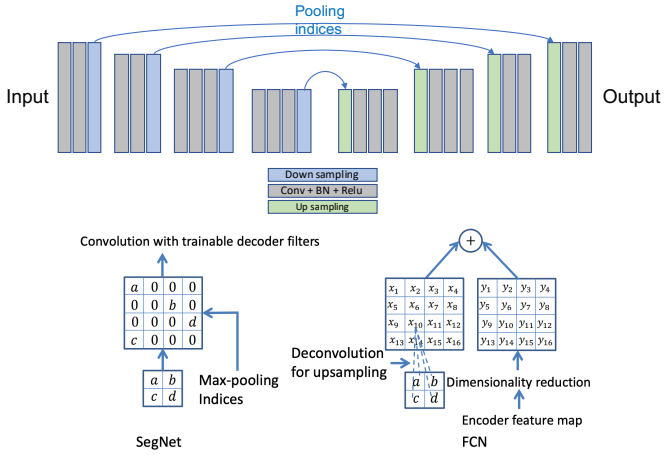

Figure 4: Top: An illustration of the SegNet architecture. There are no fully connected layers, and hence it is only convolutional. Bottom: An illustration of SegNet and FCN [86] decoders. a, b, c, d correspond to values in a feature map. SegNet uses the max-pooling indices to upsample (without learning) the feature map(s) and convolves with a trainable decoder filter bank. FCN upsamples by learning to deconvolve the input feature map and adds the corresponding encoder feature map to produce the decoder output. This feature map is the output of the max-pooling layer (includes sub-sampling) in the corresponding encoder. Note that there are no trainable decoder filters in FCN [7].

Milletari et al. [91] proposed a similar architecture (V-Net; (Fig. 6)) which added residual connections and replaced 2D operations with their 3D counterparts in order to process volumetric images. Milletari et al. also proposed optimizing for a widely used segmentation metric, i.e., Dice, which will be discussed in more detail in the section 4.

Jeugo et al. [58], developed a segmentation version of the densely connected networks architecture (DenseNet) [53] by adapting the U-Net like encoder-decoder skeleton. In

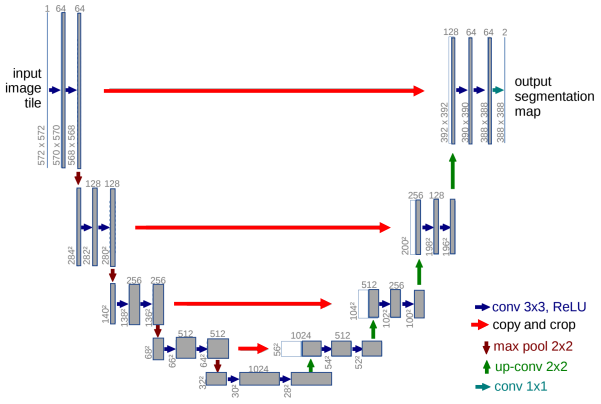

Figure 5: An illustration of the U-Net [119] architecture.

Figure 7, the detailed architecture of the network is visualized.

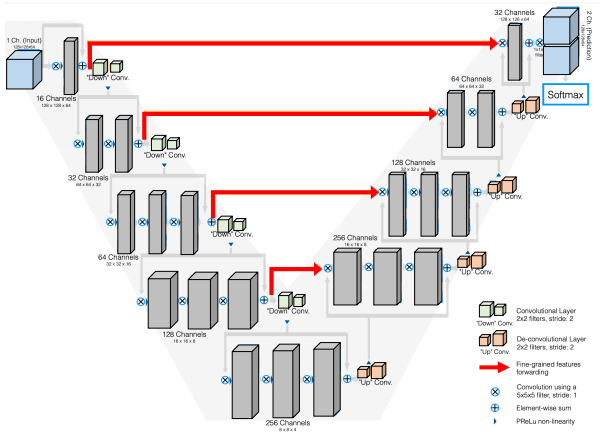

Figure 6: An illustration of the V-Net [91] architecture.

In Figure 8, we visualize the *simplified* architectural modifications applied to the first image segmentation network i.e. FCN.

Several modified versions (e.g. deeper/shallower, adding extra attention blocks) of encoder-decoder networks have been applied to semantic segmentation [5, 32, 82, 107, 113, 155, 170]. Recently in 2018, DeepLabV3+ [23] has outperformed many state-of-the-art segmentation networks on PASCAL VOC 2012 [29] and Cityscapes [177] datasets. Zhao et al. [172], modified the feature fusing operation proposed by [86] using a spatial pyramid pooling module or encode-decoder structure (Figure 9) are used in deep neural networks for semantic segmentation tasks. The spatial pyramid networks are able to encode multi-scale contextual information by probing the incoming features with filters or pooling operations at multiple rates and multiple effective fields-of-view, while the latter networks can capture sharper object boundaries by gradually recovering the spatial infor-

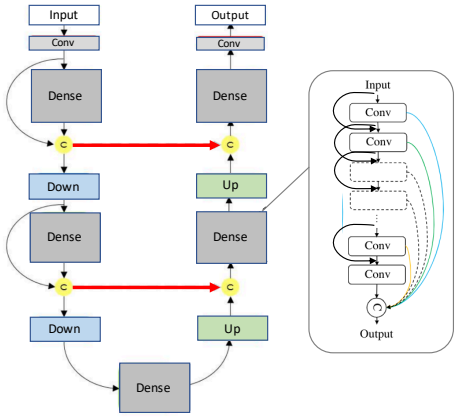

Figure 7: Diagram of the one-hundred layers Tiramisu network architecture [58]. The architecture is built from dense blocks. The architecture is composed of a downsampling path with two transitions down and an upsampling path with two transitions up. A circle represents concatenation, and the arrows 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.

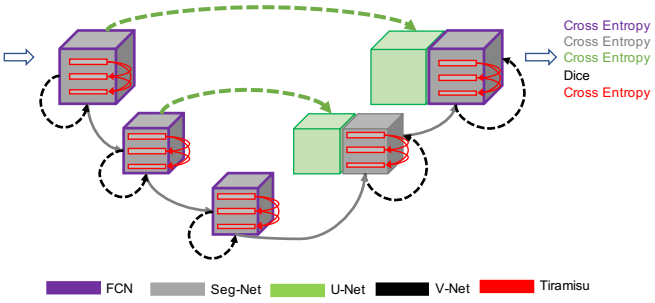

Figure 8: Gradual architectural improvements applied to FCN [86] over time.

mation.

Chen et al. [23] propose to combine the advantages from both dilated convolutions and feature pyramid pooling. Specifically, DeepLabv3+, extends DeepLabv3 [21] by adding a simple yet effective decoder module (Fig. 10) to refine the segmentation results, especially along object boundaries using dilated convolutions and pyramid features.

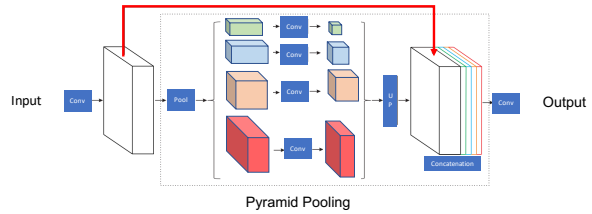

Figure 9: Overview of the pyramid scene parsing networks. Given an input image (a), feature maps from last convolution layer are pulled (b), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (c). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (d)[172].

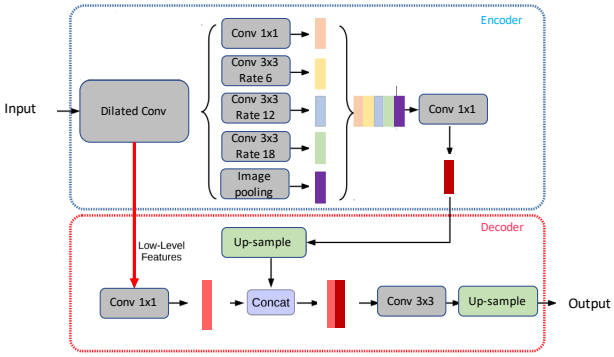

Figure 10: An illustration of the DeepLabV3+; The encoder module encodes multi-scale contextual information by applying atrous (dilated) convolution at multiple scales, while the simple yet effective decoder module refines the segmentation results along object boundaries [23].

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 [50, 74]. A few other works have attempted to simplify the structure of deep networks, e.g., by tensor factorization [69], channel/network pruning [152], or applying sparsity to connections [43]. A few methods have focused on the complexity optimization of deep image segmentation networks. Similar to the work of Saxena et al. [123], Liu et al. [85] 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 [29] and Cityscapes [177] datasets. In contrast, Chen

et al. [20] focused on searching the much smaller atrous spatial pyramid pooling module using random search.

Besides network architecture search, Srivastava et al. [132] modified ResNet in a way to control the flow of information through a connection. Lin et al. adopted one step fusion without filtering the channels.

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. Hu et al. [51] proposed 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. Wang et al. [146] added an attention module to the deep residual network (ResNet) for image classification. Their proposed attention module consists of several encoding-decoding layers. Fu et al. [31] proposed dual attention networks that apply both spatial and channel-based attention operations.

Li et al. [76], proposed a pyramid attention based network, for semantic segmentation. They combined an attention mechanism and a spatial pyramid to extract precise dense features for pixel labeling instead of complicated dilated convolution and artificially designed decoder networks. Chen et al.[22] applied attention to DeepLab which takes multi-scale inputs.

2.5. Adversarial Semantic Image Segmentation

Goodfellow et al. [37] proposed an adversarial approach to learn deep generative models. Their generative adversarial networks (GANs) take samples z from a fixed (e.g., standard Gaussian) distribution $p_z(z)$, and transform them using a deterministic differentiable deep network $p(.)$ to approximate the distribution of training samples x. 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

$$\begin{aligned}\ell\left(\boldsymbol{\theta}_s, \boldsymbol{\theta}_a\right) = & \sum_{n=1}^{N} \ell_{\mathrm{mce}}\left(s\left(\boldsymbol{x}_n\right), \boldsymbol{y}_n\right) \\ & -\lambda\left[\ell_{\mathrm{bce}}\left(a\left(\boldsymbol{x}_n, \boldsymbol{y}_n\right), 1\right)\right. \\ & \left.+\ell_{\mathrm{bce}}\left(a\left(\boldsymbol{x}_n, s\left(\boldsymbol{x}_n\right)\right), 0\right)\right],\end{aligned} \tag{1}$$

where $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_a$ denote the parameters of the segmentation and adversarial model, respectively. l_{bce} and l_{mce} are binary and multi-class cross-entropy losses, respectively. In this setup, the segmentor tries to produce segmentation maps that are close to the ground truth, i.e., which look more realistic.

The main models being used for image segmentation mostly follow encoder-decoder architectures as U-Net. Recent approaches have shown that dilated convolutions and feature pyramid pooling can improve the U-Net style networks. In Section 3, we summarize how these methods and their modified counterparts have been applied to medical images.

3. Architectural Improvements Applied to Medical Images

In this section, we review the different architectural based improvements for deep learning-based 2D and volumetric medical image segmentation.

3.1. Model Compression based Image Segmentation

To perform image segmentation in real-time and be able to process larger images/(sub)volumes in case of processing volumetric and high-resolution 2D images such as CT, MRI, and histopathology images, several methods have attempted to compress deep models. Weng et al. [153] applied a neural architecture search method to U-Net to obtain a smaller network with a better organ/tumor segmentation performance on CT, MR, and ultrasound images. Brugger et al. [14] by leveraging group normalization [157] and leaky ReLU function, redesigned the U-Net architecture in order to make the network more memory efficient for 3D medical image segmentation. Perone et al. [110] and Bonta et al. [13] designed a dilated convolution neural network with fewer parameters as compared to the original convolution-based one. Some other works [106, 160] have focused on weight quantization of deep networks for making segmentation networks smaller.

3.2. Encoder Decoder based Image Segmentation

Drozdal et al. [28] proposed to normalize input images before segmentation by applying a simple CNN prior to pushing the images to the main segmentation network. They showed improved results on electron microscopy segmentation, liver segmentation from CT, and prostate segmentation from MRI scans. Gu et al. [40] proposed using a dilated convolution block close to the network's bottleneck to preserve contextual information.

Vorontsov et al. [143] (using a dataset defined in [25], proposed an image-to-image based framework to transform an input image with object of interest (presence domain) like a tumor to an image without the tumor (absence domain) i.e. translate diseased image to healthy; next, their model learns to add the removed tumor to the new healthy image. This results in capturing detailed structure from the object, which improves the segmentation of the object. Zhou et al. [180] proposed a rewiring method for

the long skip connections used in U-Net and testes their method on nodule segmentation in the low-dose CT scans of the chest, nuclei segmentation in the microscopy images, liver segmentation in abdominal CT scans, and polyp segmentation in colonoscopy videos. Goyal et al. [38] applied DeeplabV3 [23] to color dermoscopic images to segment skin lesions.

3.3. Attention based Image Segmentation

Nie et al. designed an attention model to segment prostate from MRI images with higher accuracy compared to baseline models e.g. V-Net [91] and FCN [86]. Sinha et al. [128], proposed a multi-level attention based architecture for abdominal organ segmentation from MRI images. Similarly, Qin et al. [115] proposed a dilated convolution base block to preserve more detailed attention in 3D medical image segmentation. Similarly, other papers [80, 56, 77, 101, 102, 105, 124] have leveraged the attention concept into medical image segmentation as well.

3.4. Adversarial Training based Image Segmentation

Khosravan et al. [66], proposed an adversarial training framework for pancreas segmentation from CT scans. Son et al. [129] applied generative adversarial networks for retinal image segmentation. Xue et al. [161], used a fully convolutional network as a segmenter in the generative adversarial framework to segment brain tumors from MRI images. Other papers [26, 27, 59, 94, 99, 118, 162, 168] have also successfully applied adversarial learning to medical image segmentation.

3.5. Recurrent Neural Network-based Models

The Recurrent Neural Network (RNN) was designed for handling sequences. The long short-term memory (LSTM) network is a type of RNN that introduces self-loops to enable the gradient flow for long durations [48]. In the medical image analysis domain, RNNs have been used to model the temporal dependency in image sequences. Bai et al., [8] proposed an image sequence segmentation algorithm by combining a fully convolutional network with a recurrent neural network, which incorporates both spatial and temporal information into the segmentation task. Similarly, Gao et al. [34], applied LSTM and CNN to model temporal relationship in brian MRI slices to improve segmentation performance in 4D volumes. Li et al., [75] applied U-Net to obtain initial segmentation probability maps and further improve them using LSTM for pancreas segmentation from 3D CT scans. Similarly, other works have also applied RNNs (LSTMs) [4, 18, 164, 173, 174] to medical image segmentation.

4. Optimization Function based Improvements

In addition to improved segmentation speed/accuracy using architectural modifications as those mentioned in section 2, designing new loss functions also resulted in improvements in subsequent inference-time segmentation accuracy.

4.1. Cross Entropy

The most commonly used loss function for the task of image segmentation is a pixel-wise cross entropy loss (Eq. 2). This loss examines each pixel individually, comparing the class predictions vector to the one-hot encoded target (or ground truth) vector. For the case of binary segmentation, let $P(Y=0)=p$ and $P(Y=1)=1-p$. The predictions are given by the logistic/sigmoid function $P(\hat{Y}=0)=\frac{1}{1+e^{-x}}=\hat{p}$ and $P(\hat{Y}=1)=1-\frac{1}{1+e^{-x}}=1-\hat{p}$, where x is output of network. Then cross entropy (CE) can be defined as:

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

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

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

4.2. Weighted Cross Entropy

The cross-entropy loss evaluates the class predictions for each pixel vector individually and then averages over all pixels, which implies equal learning to each pixel in the image. This 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. WCE was defined as:

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

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].

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

Ronnenberger et al. [119], added a distance function to the cross-entropy function to enforce learning distance between the components to enforce better segmentation in case of having very close objects to each other as follows:

$$\mathrm{BCE}(p,\hat{p}) + w_0 \cdot \exp\left(-\frac{(d_1(x)+d_2(x))^2}{2\sigma^2}\right) \quad (6)$$

where $d_1(x)$ and $d_2(x)$ are two functions that calculate the distance to the border of nearest and second cells in their cell segmentation problem.

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 $(1-\hat{p})^\gamma$ to the cross entropy loss as:

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

setting $\gamma = 0$ the equation will be equivalent to BCE.

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:

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

Similarly, the Jaccard metric (intersection over union: IoU) is computed as:

$$\mathrm{IoU} = \frac{TP}{TP+FP+FN} = \frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|} \quad (9)$$

where X and Y are the predicted and ground truth segmentation, respectively. TP is the true positives, FP false positives and FN false negatives. We can see that $\mathrm{DC} \geq \mathrm{IoU}$.

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

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

where $p \in \{0,1\}^n$ and $0 \leq \hat{p} \leq 1$. p and $\hat{p}$ are the ground truth and predicted segmentation and $\langle\cdot,\cdot\rangle$ denotes dot product.

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:

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

setting $\beta = 0.5$ simplifies the equation to Eq. 10.

4.4.3 Exponential Logarithmic Loss

Wong et al. [156] proposed using a weighted sum of the exponential logarithmic Dice loss ($\mathcal{L}_{\mathrm{eld}}$) and the weighted exponential cross-entropy loss ($\mathcal{L}_{\mathrm{wece}}$) 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.

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

where

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

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

$\mathbf{x}$, i, and l denote the pixel position, the predicted label, and the ground truth label. D_i denotes the smoothed Dice loss (obtained by adding an $\epsilon = 1$ term to the numerator and denominator in Eq. 10 in order to handle missing labels while training, and γ_D and γ_{CE} are used to control the non-linearities of the respective loss functions.

4.4.4 Lovász-Softmax loss

Since it has been shown that the Jaccard loss (IoU loss) is submodular [10], Berman et al. [11] proposed using the Lovász hinge with the Jaccard loss for binary segmentation, and proposed a surrogate of the Jaccard loss, called the Lovász-Softmax loss, which can be applied for the multi-class segmentation task. The Lovász-Softmax loss is, therefore, a smooth extension of the discrete Jaccard loss, and is defined as

$$\mathcal{L}_{\mathrm{LovaszSoftmax}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\overline{\Delta_{J_c}}\left(\boldsymbol{m}(c)\right), \quad (15)$$

where $\Delta_{J_c}(\cdot)$ denotes the convex closure of the submodular Jaccard loss, $\bar{\cdot}$ denotes that it is a tight convex closure and polynomial time computable, $\mathcal{C}$ denotes all the classes, and J_c and $\boldsymbol{m}(c)$ denote the Jaccard index and the vector of errors for class c respectively.

4.4.5 Boundary Loss

Kervadec et al. [64], proposed to calculate boundary loss $\mathcal{L}_B$ along with the generalized Dice loss $\mathcal{L}_{GD}$ function as

$$\alpha \mathcal{L}_{GD}(\theta) + (1-\alpha)\mathcal{L}_B(\theta), \tag{16}$$

where the two terms in the loss function are defined as

$$\begin{aligned}\mathcal{L}_{GD}(\theta) = 1 - \\ 2\left[\left(w_G \sum_{p\in\Omega} g(p)s_\theta(p) + \right.\right. \\ \left.\left. w_B \sum_{p\in\Omega} (1-g(p))\,(1-s_\theta(p))\right) / \right. \\ \left(w_G \sum_{p\in\Omega} [s_\theta(p) + g(p)] + \right. \\ \left. \left. w_B \sum_{p\in\Omega} [2 - s_\theta(p) - g(p)]\right)\right], \text{ and}\end{aligned} \tag{17}$$

$$\mathcal{L}_B(\theta) = p \in \Omega \phi_G(p) s_\theta(p), \tag{18}$$

where $\phi_G(p) = -\|p - z_{\partial G}(p)\|$ if $p \in G$ and $\phi_G(p) = \|p - z_{\partial G}(p)\|$, otherwise. The general form integral $\sum_\Omega g(p) f\left(s_\theta(p)\right)$ is for foreground and $\sum_\Omega (1-g(p)) f\left(1-s_\theta(p)\right)$ for background. $w_G = 1/\left(\sum_{p\in\Omega} g(p)\right)^2$ and $w_B = 1/(\sum_\Omega (1-g(p)))^2$. Ω shows the spatial domain.

4.4.6 Conservative Loss

Zhu et al. [181] proposed the Conservative Loss for in order to achieve a good generalization ability in domain adaptation tasks by penalizing the extreme cases and encouraging the moderate cases. The Conservative Loss is defined as

$$CL(p_t) = \lambda(1 + \log_a(p_t))^2 * \log_a(-\log_a(p_t)), \tag{19}$$

where p_t is the probability of the prediction towards the ground truth and a is the base of the logarithm. a and λ are empirically chosen to be e (Euler's number) and 5 respectively.

Other works also include approaches to optimize the segmentation metrics [104], weighting the loss function [120], and adding regularizers to loss functions to encode geometrical and topological shape priors [9, 92].

A significant problem in image segmentation (particularly in medical images) is to overcome class imbalance for which overlap measure based methods have shown reasonably good performance in overcoming the imbalance. In Section 5, we summarize the approaches which use new loss functions, particularly for medical image segmentation or use the (modified) loss functions mentioned above.

In Figure 11, we visualize the behavior of different loss functions for segmenting large and small objects. For the parameters of the loss functions, we use the same parameters as reported by the authors in their respective papers. Therefore, we use $\beta = 0.3$ in Eqn. 11, $\alpha = 0.25$ and $\gamma = 2$ in Eqn. 7, and $\gamma_D = \gamma_{CE} = 1$, $w_{\text{eld}} = 0.8$, and $w_{\text{wece}} = 0.2$ in Eqn. 12. Moving from the left to the right for each plot, the overlap of the predictions and ground truth mask becomes progressively smaller, i.e., producing more false positives and false negatives. Ideally, the loss value should monotonically increase as more false positives, and negatives are predicted. For large objects, almost all the functions follow this assumption; however, for the small objects (right plot), only combo loss and focal loss penalize *monotonically* more for larger errors. In other words, the overlap-based functions highly fluctuate while segmenting small and large objects (also see Fig. 12), which results in unstable optimization. The loss functions which use cross-entropy as the base and the overlap measure functions as a weighted regularizer show more *stability* during training.

5. Optimization Function based Improvements Applied to Medical Images

The standard CE loss function and its weighted versions, as discussed in section 4, have been applied to numerous medical image segmentation problems [56, 77, 80, 101, 102, 105, 124]. However, Miletari et al. [91] found that optimizing convolutional neural network for DL (Eq. 10) in some cases, e.g. in the case of having very small foreground objects in a large background, works better than the original cross-entropy.

Li et al. [79] proposed adding the following regularization term to the cross entropy loss function to encourage smooth segmentation outputs.

$$R = \sum_{i=1}^{N} \mathbb{E}_{\xi',\xi} \left\| f\left(x_i;\theta,\xi'\right) - f\left(x_i;\theta,\xi\right)\right\|^2 \tag{20}$$

where ξ' and ξ are different perturbation (e.g., Gaussian noise, network dropout, and randomized data transformation) applied to the input image x_i.

Xu et al. [24], proposed leveraging traditional active contour energy minimization into convolutional neural networks via the following loss function.

$$\text{Loss} = \text{Length} + \lambda \cdot \text{Region} \tag{21}$$

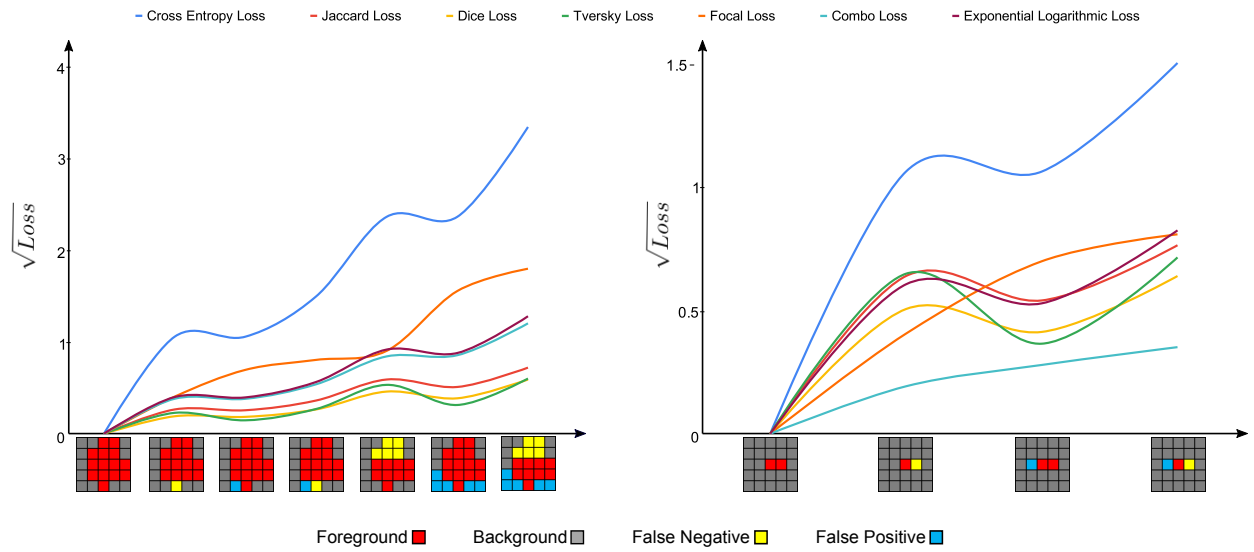

Figure 11: A comparison of seven loss functions for different extends of overlaps for a large (left) and a small (right) object.

$$\text{Length} = \sum_{\Omega}^{i=1,j=1} \sqrt{\left|\left(\nabla u_{x_{i,j}}\right)^2 + \left(\nabla u_{y_{i,j}}\right)^2\right| + \epsilon} \quad (22)$$

where x and y from $u_{x_{i,j}}$ and $u_{y_{i,j}}$ are horizontal and vertical directions, respectively.

$$\begin{aligned}\text{Region} = &\left|\sum_{\Omega}^{i=1,j=1} u_{i,j}\left(c_1 - v_{i,j}\right)^2\right| \\ &+ \left|\sum_{\Omega}^{i=1,j=1} \left(1 - u_{i,j}\right)\left(c_2 - v_{i,j}\right)^2\right|\end{aligned} \quad (23)$$

where u and v are represented as prediction and a given image, respectively. c1 is set to 1 and c2 to 0. Similar to Li et al., [79], Zhou et al. [178], proposed adding a contour regression term to the weighted cross entropy loss function.

Karimi et al. [62], optimized Hausdorff distance based function between a predicted and ground truth segmentation as follows.

$$f_{\text{HD}}(p,q) = \text{Loss}(p,q) + \lambda\left(1 - \frac{2\sum_{\Omega}(p \circ q)}{\sum_{\Omega}\left(p^2 + q^2\right)}\right) \quad (24)$$

where the second term is the Dice loss function and the first term can be replaced with three different versions of the Hausdorff distance for p and q i.e. ground truth and predicted segmentations respectively, as follows;

$$\text{Loss}(q,p) = \frac{1}{|\Omega|}\sum_{\Omega}\left((p-q)^2 \circ \left(d_p^{\alpha} + d_q^{\alpha}\right)\right) \quad (25)$$

The parameter α determines the level of penalty for larger errors. d_p is the distance map of the ground-truth segmentation as the unsigned distance to the boundary δp. Similarly, d_q is defined as the distance to δq. The $\circ$ is hadamard operation.

$$\text{Loss}(q,p) = \frac{1}{|\Omega|}\sum_{k=1}^{K}\sum_{\Omega}\left((p-q)^2 \ominus_k B\right)k^{\alpha} \quad (26)$$

where $\ominus_k$ denotes k successive erosions. where

$$B = \begin{pmatrix} 0 & 1/5 & 0 \\ 1/5 & 1/5 & 1/5 \\ 0 & 1/5 & 0 \end{pmatrix} \quad (27)$$

$$\begin{aligned}\text{Loss}(q,p) = \frac{1}{|\Omega|}\sum_{r\in R} r^{\alpha}\sum_{\Omega}\big[&f_s\left(B_r * \overline{p}^C\right) \circ f_{\overline{q}\setminus\overline{p}} \\ &+ f_s\left(B_r * \overline{p}\right) \circ f_{\overline{p}\setminus\overline{q}} \\ &+ f_s\left(B_r * \overline{q}^C\right) \circ f_{\overline{p}\setminus\overline{q}} \\ &+ f_s\left(B_r * \overline{q}\right) \circ f_{\overline{q}\setminus\overline{p}}\big]\end{aligned} \quad (28)$$

where $f_{\overline{q}\setminus\overline{p}} = (p-q)^2 q$. f_s indicates soft thresholding. B_r denotes a circular-shaped convolutional kernel of radius r. Elements of B_r are normalized such that they sum to one. $\overline{p}^C = 1 - \overline{p}$. Ground-truth and predicted segmentations, denoted with $\overline{p}$ and $\overline{q}$,

Caliva et al. [15] proposed to measure distance of each voxel to the boundaries of the objects and use the weight matrices to penalize a model for error on the boundaries. Kim et al., [67] proposed using level-set energy minimization as a regularizer summed with standard multi-class cross

entropy loss function for semi-supervised brain MRI segmentation as:

$$\mathcal{L}_{\text{level}}(\Theta; x) = \sum_{n=1}^{N} \int_{\Omega} \left| x(r) - c_n^{\Theta} \right|^2 y_n^{\Theta}(r) dr + \lambda \sum_{n=1}^{N} \int_{\Omega} \left| \nabla y_n^{\Theta}(r) \right| dr \tag{29}$$

with

$$c_n^{\Theta} = \frac{\int_{\Omega} x(r) y_n^{\Theta}(r) dr}{\int_{\Omega} y_n^{\Theta}(r) dr} \tag{30}$$

where $x(r)$ is the input, $y_n^{\Theta}(r)$ is the output of softmax layer, Θ refers to learnable parameters.

Taghanaki et al. [139]], discussed the risks of using solo overlap based loss functions and proposed to use them as regularizes along with a weighted cross entropy to explicitly handle input and output imbalance as follows;

$$\begin{aligned} Combo\ Loss = \alpha \Bigg(-\frac{1}{N} \sum_{i=1}^{N} \beta \left(t_i - \ln p_i \right) + \\ (1-\beta) \left[(1-t_i) \ln (1-p_i) \right] \Bigg) \\ - (1-\alpha) \sum_{i=1}^{K} \left(\frac{2 \sum_{i=1}^{N} p_i t_i + S}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} t_i + S} \right) \end{aligned} \tag{31}$$

where α controls the amount of Dice term contribution in the loss function L, and $\beta \in [0, 1]$ controls the level of model penalization for false positives/negatives: when β is set to a value smaller than 0.5, FP are penalized more than FN as the term $(1-t_i) \ln (1-p_i)$ is weighted more heavily, and vice versa. In our implementation, to prevent division by zero, we perform add-one smoothing (a specific instance of the additive/Laplace/Lidstone smoothing) [121], i.e., we add unity constant S to both the denominator and numerator of the Dice term. The majority of the methods discussed in Section 5 have attempted to handle the class imbalance issue in the input images i.e., small foreground versus large background with providing weights/penalty terms in the loss function. Other approaches consist of first identifying the object of interest, cropping around this object, and then performing the task (e.g., segmentation) with better-balanced classes. This type of cascade approach has been applied for the segmentation of multiple sclerosis lesions in the spinal cord [39].

6. Image Synthesis based Methods Applied to Medical Image Segmentation

Deep convolutional neural networks are heavily reliant on big data to avoid overfitting and class imbalance issues, and therefore this section focuses on data augmentation, a data-space solution to the problem of limited data. Apart from standard online image augmentation methods such as geometric transformations, color space augmentations, in this section, we discuss image synthesis methods, the output of which are novel images rather than modify existing images. Since GANs based augmentation techniques for segmentation, tasks have been used for a wide variety of problems - from remote sensing imagery [95] to filamentary anatomical structures [171], this section covers GAN based augmentation works in the field of medical image analysis.

Chartsias et al. [19] used a conditional GAN to generate cardiac MR images from CT images. They showed that utilizing the synthetic data increased the segmentation accuracy and that using only the synthetic data led to only a marginal decrease in the segmentation accuracy. Similarly, Zhang et al. [169] proposed a GAN based volume-to-volume translation for generating MR volumes from corresponding CT volumes and vice versa. They showed that synthetic data improve segmentation performance on cardiovascular MRI volumes. Huo et al. [54] proposed an end-to-end synthesis and segmentation network called SynSeg-Net to simultaneously synthesize CT images from unpaired MR images and to segment CT splenomegaly on unlabeled CT images and showed that their approach yielded better segmentation performance than even segmentation obtained using models trained using the manual CT labels. Abhishek et al. [2] trained a conditional GAN to generate skin lesion images from and confined to binary masks, and showed that using the synthesized images led to a higher skin lesion segmentation accuracy. Zhang et al. [167] trained a GAN for translating between digitally reconstructed radiographs and X-ray images and achieved similar accuracy as supervised training in multi-organ segmentation. Shin et al. [125], proposed a method to generate synthetic abnormal MRI images with brain tumors by training a generative adversarial network using two publicly available data sets of brain MRI. Similarly, other works [42, 163, 165] have leveraged GANs to synthesize brain MR images.

7. Weakly Supervised Methods

Collecting large-scale accurate pixel-level annotation is time-consuming and financially expensive. However, unlabeled and weakly-labeled images can be collected in large amounts in a relatively fast and cheap manner. Therefore a promising direction for semantic image segmentation is to develop unsupervised and weakly supervised models.

Kim et al. [68] proposed a weakly supervised seman-

tic segmentation network using unpooling and deconvolution operations, and used feature maps from the deconvolutions layers to learn scale-invariant features, and evaluated their model on the PASCAL VOC and chest X-ray image datasets. Lee et al. [73] used dropout [131] to choose features at random during training and inference and combine the many different localization maps to generate a single localization map, effectively discovering relationships between locations in an image, and evaluated their proposed approach on the PASCAL VOC dataset.

7.1. Weakly Supervised Models Applied to Medical Images

The scarcity of richly annotated medical images is limiting supervised deep learning-based solutions to medical image analysis tasks [112], such as localizing discriminatory radiomic disease signatures. Therefore, it is desirable to leverage unsupervised and weakly supervised models. Kervadec et al. [65] introduced a differentiable term in the loss function for datasets with weakly supervised labels, which reduced the computational demand for training while also achieving almost similar performance to full supervision for segmentation of cardiac images. Afshari et al. [3] used a fully convolutional architecture along with a Mumford-Shah functional [98] inspired loss function to segment lesions from PET scans using only bounding box annotations as supervision. Mirikharaji et al. [93] proposed to learn spatially adaptive weight maps to account for spatial variations in pixel-level annotations and used noisy annotations to train a segmentation model for skin lesions. Taghanaki et al. [138] proposed to learn spatial masks using only image-level labels with minimizing mutual information between the input and masks, and at the same time maximizing the mutual information between the masks and image labels. Peng et al. [108] proposed an approach to train a CNN with discrete constraints and regularization priors based on the alternating direction method of multipliers (ADMM). Perone et al. [111] expanded the semi-supervised Mean Teacher [141] approach to segmentation tasks on MRI data, and show that it can bring important improvements in a realistic small data regime. In another work, Perone et al. [109] extended the method of unsupervised domain adaptation using self-ensembling for the semantic segmentation task. They showed how this approach could improve the generalization of the models even when using a small amount of unlabeled data.

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.

Bischke et al. [12] proposed a cascaded multi-task loss to preserve boundary information from segmentation masks for segmenting building footprints and achieved state-of-the-art performance on an aerial image labeling task. Chaichulee et al. [17] extended the VGG16 architecture [126] to include a global average pooling layer for patient detection and a fully convolutional network for skin segmentation. The proposed model was evaluated on images from a clinical study conducted at a neonatal intensive care unit, and was robust to changes in lighting, skin tone, and pose. He et al. [46] trained a U-Net [119]-like encoder-decoder architecture to simultaneously segment thoracic organs from CT scans and perform global slice classification. Ke et al. [63] trained a multi-task U-Net architecture to solve three tasks - separating wrongly connected objects, detecting class instances, and pixelwise labeling for each object, and evaluated it on a food microscopy image dataset. Other multi-task models have also been proposed for segmentation and classification for detecting manipulated faces in images and video [100] and diagnosis of breast biopsy images [89] and mammograms [72]

He et al. extended Faster R-CNN [117] by adding a new branch to predict the object mask along with a class label and a bounding box, and the proposed model was called Mask R-CNN [44]. Mask R-CNN has been used extensively for multi-task segmentation models for a wide range of application areas [1], such as adding sports fields to OpenStreetMap [61], detection and segmentation for surgery robots [133], understanding climate change patterns from aerial imagery of the Arctic [166], converting satellite imagery to maps [96], detecting image forgeries [150], and segmenting tree canopy [175].

8.1. Multi-Task Models Applied to Medical Images

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]. Mask R-CNN has also been extended to work with 3D volumes and has been evaluated on lung nodule detection and segmentation from CT scans and breast lesion detection and categorization on diffusion MR images [57, 70].

9. Summary of the Main Models Tested on Natural Images

Table 1 lists a summary of selected papers from this review, the nature of their proposed contributions, and the datasets that they were evaluated on. For the papers that evaluated their models on the PASCAL VOC

2012 dataset [30], a popular image semantic segmentation dataset, we also list their reported mean IoU (intersection over union) scores. As can be seen in Table 1, the focus has been mostly on architectural improvements. Comparing the first deep learning-based model (i.e. FCN) to the state-of-the-art model (i.e. DeepLabV3+) there is a large improvement (i.e. $\sim 27\%$; 62.2% to 89.0%) in terms of mean IoU. The latter model leverages a more sophisticated decoder, dilated convolutions, and feature pyramid pooling.

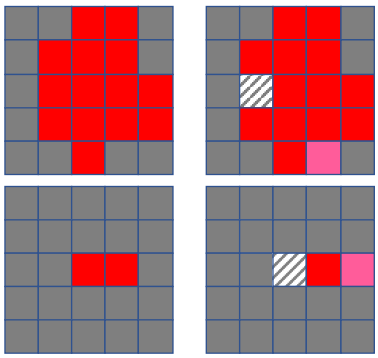

Figure 12: Comparison of cross entropy and Dice losses for segmenting small and large objects. The red pixels show the ground truth and the predicted foregrounds in the left and right columns respectively. The striped and the pink pixels indicate false negative and false positive, respectively. For the top row (i.e., large foreground), the Dice loss returns 0.96 for one false negative and for the bottom row (i.e., small object) returns 0.66 for one false negative, whereas the cross entropy loss function outputs 0.83 for both the cases. By considering a false negative and false positive, the output value drops even more in case of using Dice but the cross entropy stays smooth (i.e., Dice value of 0.93 and 0.50 for large and small object versus cross entropy loss value of 1.66 for both.)

10. Discussion and Future Directions

By reviewing the literature of both natural and medical image segmentation, we observe potential difficulties for medical image segmentation. For example, medical images can be in high dimensions (both 2D and volumetric), which does not fit into current GPUs; thus, they need to be processed as sub-volumes/images. This prevents models to capture spatial information/relationships properly. Sometimes medical devices generate unique and hard-to-detect noise patterns (biases), which makes it hard to generalize in inference time. Another potential difficulty regarding the medical domain is the lack of annotated data, which encourages training semi and unsupervised models. Encoding prior knowledge in medical image analysis models is generally more possible as compared to natural images.

In the following sections, we discuss in detail the potential future research direction for semantic image segmentation.

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. They equip encoder-decoder-like networks with richer feature representations, but at the cost of higher memory usage, computation, and possibly resulting in transferring non-discriminative feature maps. Similar to Taghanaki et al.'s method [136], one future work direction is to optimize the amount of data is being transferred through skip connections. As for the cell level architectural design, our study shows that Atrous convolutions with feature pyramid pooling modules are highly being used in the recent models. These approaches are somehow modifications of the classical convolution blocks. Similar to radial basis function layers in [90, 134], a future work focus can be designing new layers that capture a new aspect of data as opposed to convolutions or transform the convolution features into a new manifold.

10.2. Sequenced Models

For image segmentation, sequenced models can be used to segment temporal data such as videos. These models have also been applied to 3D medical datasets, however the advantage of processing volumetric data using 3D convolutions versus the processing the volume slice by slice using 2D sequenced models. 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. Therefore a future direction in this area can be through analysis of sequenced models versus volumetric convolution-based approaches.

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. Incorporating domain/prior knowledge (such as coding the location of different organs explicitly in a deep model) is more sensible in the medical datasets. As shown in [139], when only a distance-based or overlap-based loss function is used in a network, and the final layer applies Sigmoid function, the risk of gradient vanishing increases. Although overlap based loss function are used in case of a class imbalance (small foregrounds), in Figure 12, we show how using (*only*) overlap based loss functions as the main

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%

term can be problematic for smooth optimization where they highly penalize a model under/over-segmenting a small foreground. However, the cross-entropy loss returns a reasonable score for the same cases. Besides using integrated cross-entropy based loss functions, future work can be exploring a single loss function that follows the behavior of the cross-entropy and at the same time, offers more features such capturing contour distance. This can be achieved by revisiting the current distance and overlap based loss functions. Another future path can be exploring auto loss function (or regularization term) search similar to the neural architecture search mentioned above.

10.4. Other Potential Directions

- Going beyond pixel intensity-based scene understanding via incorporating prior knowledge. Currently, deep models receive matrices of intensity values, and usually, they are not forced to learn prior information. Without explicit reinforcement, the models might still learn object relations to some extent. However, it is difficult to interpret a learned strategy.

- Because of the large number of imaging modalities, the significant signal noise present in imaging modalities such as PET and ultrasound, and the limited amount of medical imaging data mainly because of high acquisition cost compounded by legal, ethical, and privacy issues, it is difficult to develop universal solutions that yield acceptable performances across various imaging modalities. Therefore, a proper research direction would be along Raghu et al.'s work [116] on image classification models, studying the risks of using non-medical pre-trained models for medical image segmentation.

- 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.

- Exploring reinforcement learning approaches similar to [130, 151] for semantic (medical) image segmenta-

tion to mimic the way human does delineation. Deep CNNs are successful in extracting features of different classes of objects, but they loss the local spatial information of where the borders of an object should be. Some researchers seek to traditional computer vision method to overcome this problem, like conditional random field (CRF), however, CRF adds more computation time to models.

- Studying the causes for some models/datasets being prone to false positive in the image segmentation context and some other to false-negative as found by the authors in [11, 139].

- Exploring segmentation-free [55, 97, 114, 137, 176] approaches, i.e., bypassing the segmentation step according to the target problem.

- Weakly supervised segmentation using image-level labels versus a few images with segmentation annotations. Most new weakly supervised localization methods apply attention maps or region proposals in a multiple instance learning formulations. While attention maps can be noisy, leading to erroneously highlighted regions, it is not simple to decide on an optimal window/bag size for multiple instance learning approaches.

- Modifying input instead of the model, loss function, and adding more train data. Drozdal et al. [28], showed that attaching a pre-processing module at the beginning of a segmentation network improves the network performance. Taghanaki et al. [135] leveraged the gradients of a trained segmentation network with respect to the input to transfer it to space where the segmentation accuracy improves.

References

[1] W. Abdulla. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. https://github.com/matterport/Mask_RCNN, 2017.

[2] K. Abhishek and G. Hamarneh. Mask2Lesion: Mask-constrained adversarial skin lesion image synthesis. *Medical Image Computing and Computer-Assisted Intervention Workshop on Simulation and Synthesis in Medical Imaging*, pages 71–80, 2019.

[3] S. Afshari, A. BenTaieb, Z. Mirikharaji, and G. Hamarneh. Weakly supervised fully convolutional network for PET lesion segmentation. In *Medical Imaging 2019: Image Processing*, volume 10949, page 109491K. International Society for Optics and Photonics, 2019.

[4] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari. Recurrent residual U-Net for medical image segmentation. *Journal of Medical Imaging*, 6(1):14006, 2019.

[5] M. Amirul Islam, M. Rochan, N. D. Bruce, and Y. Wang. Gated feedback refinement network for dense image labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3751–3759, 2017.

[6] R. Anantharaman, M. Velazquez, and Y. Lee. Utilizing Mask R-CNN for detection and segmentation of oral diseases. *2018 IEEE International Conference on Bioinformatics and Biomedicine*, pages 2197–2204, 2018.

[7] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv:1511.00561*, 2015.

[8] W. Bai, H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. M. Matthews, and D. Rueckert. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 586–594. Springer, 2018.

[9] A. BenTaieb and G. Hamarneh. Topology aware fully convolutional networks for histology gland segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 460–468. Springer, 2016.

[10] M. Berman, M. B. Blaschko, A. R. Triki, and J. Yu. Yes, IoU loss is submodular-as a function of the mispredictions. *arXiv preprint arXiv:1809.01845*, 2018.

[11] M. Berman, A. Rannen Triki, and M. B. Blaschko. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. *arXiv:1705.08790*, 2017.

[12] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel. Multi-task learning for segmentation of building footprints with deep neural networks. In *2019 IEEE International Conference on Image Processing*, pages 1480–1484. IEEE, 2019.

[13] L. R. Bonta and N. U. Kiran. Efficient segmentation of medical images using dilated residual networks. In *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*, pages 39–47. Springer, 2019.

[14] R. Brügger, C. F. Baumgartner, and E. Konukoglu. A partially reversible U-Net for memory-efficient volumetric image segmentation. *arXiv preprint arXiv:1906.06148*, 2019.

[15] F. Caliva, C. Iriondo, A. M. Martinez, S. Majumdar, and V. Pedoia. Distance map loss penalty term for semantic segmentation. *International Conference on Medical Imaging with Deep Learning*, 2019.

[16] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[17] S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, G. Green, K. McCormick, A. Zisserman, and L. Tarassenko. Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 266–272. IEEE, 2017.

[18] A. Chakravarty and J. Sivaswamy. Race-net: a recurrent neural network for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 23(3):1151–1162, 2018.

[19] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 3–13. Springer, 2017.

[20] L.-C. Chen, M. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens. Searching for efficient multi-scale architectures for dense image prediction. In *Advances in Neural Information Processing Systems*, pages 8699–8710, 2018.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[22] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2016.

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818, 2018.

[24] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams, and Y. Zheng. Learning active contour models for medical image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11632–11640, 2019.

[25] J. P. Cohen, M. Luck, and S. Honari. Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. *Medical Image Computing & Computer Assisted Intervention*, 2018.

[26] P. Costa, A. Galdran, M. I. Meyer, M. D. Abràmoff, M. Niemeijer, A. M. Mendonça, and A. Campilho. Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*, 2017.

[27] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, and E. P. Xing. Scan: Structure correcting adversarial network for organ segmentation in chest x-rays. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 263–273. Springer, 2018.

[28] M. Drozdzal, G. Chartrand, E. Vorontsov, M. Shakeri, L. Di Jorio, A. Tang, A. Romero, Y. Bengio, C. Pal, and S. Kadoury. Learning normalized inputs for iterative estimation in medical image segmentation. *Medical Image Analysis*, 44:1–13, 2018.

[29] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[30] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[31] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[32] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu. Stacked deconvolutional network for semantic segmentation. *IEEE Transactions on Image Processing*, 2019.

[33] H. Gamage, W. Wijesinghe, and I. Perera. Instance-based segmentation for boundary detection of neuropathic ulcers through Mask-RCNN. In *International Conference on Artificial Neural Networks*, pages 511–522. Springer, 2019.

[34] Y. Gao, J. M. Phillips, Y. Zheng, R. Min, P. T. Fletcher, and G. Gerig. Fully convolutional structured LSTM networks for joint 4D medical image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pages 1104–1108. IEEE, 2018.

[35] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[36] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[38] M. Goyal and M. H. Yap. Automatic lesion boundary segmentation in dermoscopic images with ensemble deep learning methods. *arXiv preprint arXiv:1902.00809*, 2019.

[39] C. Gros, B. De Leener, A. Badji, J. Maranzano, D. Eden, S. M. Dupont, J. Talbott, R. Zhuoquiong, Y. Liu, T. Granberg, et al. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. *Neuroimage*, 184:901–915, 2019.

[40] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu. CE-Net: Context encoder network for 2D medical image segmentation. *IEEE Transactions on Medical Imaging*, 2019.

[41] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7(2):87–93, 2018.

[42] C. Han, K. Murao, S. Satoh, and H. Nakayama. Learning more with less: Gan-based medical image augmentation. *Medical Imaging Technology*, 37(3):137–142, 2019.

[43] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally. EIE: efficient inference engine on compressed deep neural network. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture*, pages 243–254. IEEE, 2016.

[44] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[45] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[46] T. He, J. Guo, J. Wang, X. Xu, and Z. Yi. Multi-task learning for the segmentation of thoracic organs at risk in CT images. In *SegTHOR@ISBI*, 2019.

[47] M. H. Hesamian, W. Jia, X. He, and P. Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging*, pages 1–15, 2019.

[48] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[49] S. Honari, J. Yosinski, P. Vincent, and C. Pal. Recombinator networks: Learning coarse-to-fine feature aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5743–5752, 2016.

[50] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[51] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[52] Y. Hu, Z. Chen, and W. Lin. RGB-D semantic segmentation: A review. In *2018 IEEE International Conference on Multimedia & Expo Workshops*, pages 1–6. IEEE, 2018.

[53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[54] Y. Huo, Z. Xu, S. Bao, A. Assad, R. G. Abramson, and B. A. Landman. Adversarial synthesis learning enables segmentation without target modality ground truth. In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pages 1217–1220. IEEE, 2018.

[55] M. A. Hussain, A. Amir-Khalili, G. Hamarneh, and R. Abugharbieh. Segmentation-free kidney localization and volume estimation using aggregated orthogonal decision CNNs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 612–620. Springer, 2017.

[56] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, et al. nnU-Net: Self-adapting framework for U-Net-based medical image segmentation. In *Bildverarbeitung für die Medizin 2019*, pages 22–22. Springer, 2019.

[57] P. F. Jaeger, S. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H.-P. Schlemmer, and K. H. Maier-Hein. Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. *arXiv preprint arXiv:1811.08661*, 2018.

[58] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.

[59] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura. CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 732–740. Springer, 2018.

[60] J. W. Johnson. Adapting Mask R-CNN for automatic nucleus segmentation. *arXiv preprint arXiv:1805.00500*, 2018.

[61] jremillard. Images to OSM. https://github.com/jremillard/images-to-osm, 2018.

[62] D. Karimi and S. E. Salcudean. Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. *arXiv preprint arXiv:1904.10030*, 2019.

[63] R. Ke, A. Bugeau, N. Papadakis, P. Schütz, and C.-B. Schönlieb. A multi-task U-Net for segmentation with lazy labels. *ArXiv*, abs/1906.12177, 2019.

[64] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed. Boundary loss for highly unbalanced segmentation. In *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 285–296, London, United Kingdom, 08–10 Jul 2019. PMLR.

[65] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed. Constrained-CNN losses for weakly supervised segmentation. *Medical Image Analysis*, 54:88–99, 2019.

[66] N. Khosravan, A. Mortazi, M. Wallace, and U. Bagci. PAN: Projective adversarial network for medical image segmentation. *arXiv preprint arXiv:1906.04378*, 2019.

[67] B. Kim and J. C. Ye. Multiphase level-set loss for semi-supervised and unsupervised segmentation with deep learning. *arXiv preprint arXiv:1904.02872*, 2019.

[68] H.-E. Kim and S. Hwang. Scale-invariant feature learning using deconvolutional neural networks for weakly-supervised semantic segmentation. *ArXiv*, abs/1602.04984, 2016.

[69] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530*, 2015.

[70] E. Kopelowitz and G. Engelhard. Lung nodules detection and segmentation using 3D Mask R-CNN, 2019.

[71] F. Lateef and Y. Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019.

[72] T.-L.-T. Le, N. Thome, S. Bernard, V. Bismuth, and F. Patoureaux. Multitask classification and segmentation for cancer diagnosis in mammography. *arXiv preprint arXiv:1909.05397*, 2019.

[73] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019.

[74] S. Leroux, P. Molchanov, P. Simoens, B. Dhoedt, T. Breuel, and J. Kautz. IamNN: Iterative and adaptive mobile neural network for efficient image classification. *arXiv preprint arXiv:1804.10123*, 2018.

[75] H. Li, J. Li, X. Lin, and X. Qian. Pancreas segmentation via spatial context based U-Net and bidirectional LSTM. *arXiv preprint arXiv:1903.00832*, 2019.

[76] H. Li, P. Xiong, J. An, and L. Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[77] S. Li, M. Dong, G. Du, and X. Mu. Attention dense-U-Net for automatic breast mass segmentation in digital mammogram. *IEEE Access*, 7:59037–59047, 2019.

[78] X. Li, Z. Liu, P. Luo, C. Change Loy, and X. Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 3193–3202, 2017.

[79] X. Li, L. Yu, H. Chen, C.-W. Fu, and P.-A. Heng. Transformation consistent self-ensembling model for semi-supervised medical image segmentation. *arXiv preprint arXiv:1903.00348*, 2019.

[80] S. Lian, Z. Luo, Z. Zhong, X. Lin, S. Su, and S. Li. Attention guided U-Net for accurate iris segmentation. *Journal of Visual Communication and Image Representation*, 56:296–304, 2018.

[81] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 603–619, 2018.

[82] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1925–1934, 2017.

[83] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[84] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[85] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 82–92, 2019.

[86] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[87] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[88] P. Luo, G. Wang, L. Lin, and X. Wang. Deep dual learning for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2718–2726, 2017.

[89] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro. Y-net: Joint segmentation and classification for diagnosis of breast biopsy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 893–901. Springer, 2018.

[90] B. J. Meyer, B. Harwood, and T. Drummond. Deep metric learning and image classification with nearest neighbour gaussian kernels. In *2018 25th IEEE International Conference on Image Processing*, pages 151–155. IEEE, 2018.

[91] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision*, pages 565–571. IEEE, 2016.

[92] Z. Mirikharaji and G. Hamarneh. Star shape prior in fully convolutional networks for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 737–745. Springer, 2018.

[93] Z. Mirikharaji, Y. Yan, and G. Hamarneh. Learning to segment skin lesions from noisy annotations. *International Workshop on Medical Image Learning with Less Labels and Imperfect Data*, 2019.

[94] P. Moeskops, M. Veta, M. W. Lafarge, K. A. Eppenhof, and J. P. Pluim. Adversarial training and dilated convolutions for brain MRI segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 56–64. Springer, 2017.

[95] S. Mohajerani, R. Asad, K. Abhishek, N. Sharma, A. van Duynhoven, and P. Saeedi. Cloudmaskgan: A content-aware unpaired image-to-image translation algorithm for remote sensing imagery. In *2019 IEEE International Conference on Image Processing*, pages 1965–1969. IEEE, 2019.

[96] S. P. Mohanty. Crowdai mapping challenge 2018 : Baseline with Mask RCNN. https://github.com/crowdai/crowdai-mapping-challenge-mask-rcnn, 2018.

[97] S. Mukherjee, I. Cheng, S. Miller, T. Guo, V. Chau, and A. Basu. A fast segmentation-free fully automated approach to white matter injury detection in preterm infants. *Medical & Biological Engineering & Computing*, 57(1):71–87, 2019.

[98] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42(5):577–685, 1989.

[99] T. Neff, C. Payer, D. Stern, and M. Urschler. Generative adversarial network based synthesis for supervised medical image segmentation. In *Proceedings of OAGM and ARW Joint Workshop*, 2017.

[100] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *ArXiv*, abs/1906.06876, 2019.

[101] Z.-L. Ni, G.-B. Bian, X.-L. Xie, Z.-G. Hou, X.-H. Zhou, and Y.-J. Zhou. RASNet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. *arXiv preprint arXiv:1905.08663*, 2019.

[102] D. Nie, Y. Gao, L. Wang, and D. Shen. ASDNet: Attention based semi-supervised deep networks for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 370–378. Springer, 2018.

[103] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[104] S. Nowozin. Optimal decisions from probabilistic models: the intersection-over-union case. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 548–555, 2014.

[105] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[106] M. Paschali, S. Gasperini, A. G. Roy, M. Y.-S. Fang, and N. Navab. 3dq: Compact quantized neural networks for volumetric whole brain segmentation. *arXiv preprint arXiv:1904.03110*, 2019.

[107] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2017.

[108] J. Peng, H. Kervadec, J. Dolz, I. B. Ayed, M. Pedersoli, and C. Desrosiers. Discretely-constrained deep network for weakly supervised segmentation. *arXiv preprint arXiv:1908.05770*, 2019.

[109] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *Neuroimage*, 194:1–11, Mar. 2019.

[110] C. S. Perone, E. Calabrese, and J. Cohen-Adad. Spinal cord gray matter segmentation using deep dilated convolutions. *Scientific reports*, 8(1):5966, 2018.

[111] C. S. Perone and J. Cohen-Adad. Deep semi-supervised segmentation with weight-averaged consistency targets. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 12–19, 2018.

[112] C. S. Perone and J. Cohen-Adad. Promises and limitations of deep learning for medical image segmentation. *Journal of Medical Artificial Intelligence*, 2, 2019.

[113] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4151–4160, 2017.

[114] H. Proenca and J. C. Neves. Segmentation-less and non-holistic deep-learning frameworks for iris recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[115] Y. Qin, K. Kamnitsas, S. Ancha, J. Nanavati, G. Cottrell, A. Criminisi, and A. Nori. Autofocus layer for semantic segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–611. Springer, 2018.

[116] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning with applications to medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

[117] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[118] M. Rezaei, K. Harmuth, W. Gierke, T. Kellermeier, M. Fischer, H. Yang, and C. Meinel. A conditional adversarial network for semantic segmentation of brain tumor. In *International Conference on Medical Image Computing and Computer Assisted Intervention, Brainlesion Workshop*, pages 241–252. Springer, 2017.

[119] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241. Springer, 2015.

[120] A. G. Roy, S. Conjeti, D. Sheet, A. Katouzian, N. Navab, and C. Wachinger. Error corrective boosting for learning fully convolutional networks with limited data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 231–239. Springer, 2017.

[121] S. J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[122] S. S. M. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 379–387. Springer, 2017.

[123] S. Saxena and J. Verbeek. Convolutional neural fabrics. In *Advances in Neural Information Processing Systems*, pages 4053–4061, 2016.

[124] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207, 2019.

[125] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 1–11. Springer, 2018.

[126] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[127] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[128] A. Sinha and J. Dolz. Multi-scale guided attention for medical image segmentation. *arXiv preprint arXiv:1906.02849*, 2019.

[129] J. Son, S. J. Park, and K.-H. Jung. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. *arXiv preprint arXiv:1706.09318*, 2017.

[130] G. Song, H. Myeong, and K. Mu Lee. Seednet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1760–1768, 2018.

[131] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[132] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[133] SUYEgit. Mask R-CNN for Surgery Robot. https://github.com/SUYEgit/Surgery-Robot-Detection-Segmentation/, 2018.

[134] S. A. Taghanaki, K. Abhishek, S. Azizi, and G. Hamarneh. A kernelized manifold mapping to diminish the effect of adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11340–11349, 2019.

[135] S. A. Taghanaki, K. Abhishek, and G. Hamarneh. Improved inference via deep input transfer. *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2019.

[136] S. A. Taghanaki, A. Bentaieb, A. Sharma, S. K. Zhou, Y. Zheng, B. Georgescu, P. Sharma, S. Grbic, Z. Xu, D. Comaniciu, et al. Select, attend, and transfer: Light, learnable skip connections. *Medical Image Computing and Computer-Assisted Intervention Workshop on Machine Learning in Medical Imaging*, 2019.

[137] S. A. Taghanaki, N. Duggan, H. Ma, X. Hou, A. Celler, F. Benard, and G. Hamarneh. Segmentation-free direct tumor volume and metabolic activity estimation from pet scans. *Computerized Medical Imaging and Graphics*, 63:52–66, 2018.

[138] S. A. Taghanaki, M. Havaei, T. Berthier, F. Dutil, L. Di Jorio, G. Hamarneh, and Y. Bengio. InfoMask: Masked variational latent representation to localize chest disease. *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2019.

[139] S. A. Taghanaki, Y. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75:24–33, 2019.

[140] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. Chiang, Z. Wu, and X. Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *arXiv preprint arXiv:1908.10454*, 2019.

[141] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[142] H.-F. Tsai, J. Gajda, T. F. Sloan, A. Rares, and A. Q. Shen. Usiigaci: Instance-aware cell tracking in stain-free phase contrast microscopy enabled by machine learning. *SoftwareX*, 9:230–237, 2019.

[143] E. Vorontsov, P. Molchanov, W. Byeon, S. De Mello, V. Jampani, M.-Y. Liu, S. Kadoury, and J. Kautz. Boosting segmentation with weak supervision from image-to-image translation. *arXiv preprint arXiv:1904.01636*, 2019.

[144] A. O. Vuola, S. U. Akram, and J. Kannala. Mask R-CNN and U-net ensembled for nuclei segmentation. *arXiv preprint arXiv:1901.10170*, 2019.

[145] E. K. Wang, X. Zhang, L. Pan, C. Cheng, A. Dimitrakopoulou-Strauss, Y. Li, and N. Zhe. Multi-path dilated residual network for nuclei segmentation and detection. *Cells*, 8(5):499, 2019.

[146] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[147] G. Wang, P. Luo, L. Lin, and X. Wang. Learning object interactions and descriptions for semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5859–5867, 2017.

[148] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pages 1451–1460. IEEE, 2018.

[149] S. Wang, R. Rong, D. M. Yang, L. Cai, L. Yang, D. Luo, B. Yao, L. Xu, T. Wang, X. Zhan, et al. Computational staining of pathology images to study tumor microenvironment in lung cancer. *Available at SSRN 3391381*, 2019.

[150] X. Wang, H. Wang, S. Niu, and J. Zhang. Detection and localization of image forgeries using improved mask regional convolutional neural network. *Mathematical Bioscience and Engineering*, 2019.

[151] Z. Wang, S. Sarcar, J. Liu, Y. Zheng, and X. Ren. Outline objects using deep reinforcement learning. *arXiv preprint arXiv:1804.04603*, 2018.

[152] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2074–2082, 2016.

[153] Y. Weng, T. Zhou, Y. Li, and X. Qiu. NAS-Unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019.

[154] J. Wessel, M. P. Heinrich, J. von Berg, A. Franz, and A. Saalbach. Sequential rib labeling and segmentation in chest x-ray using Mask R-CNN. In *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, London, United Kingdom, 08–10 Jul 2019.

[155] Z. Wojna, V. Ferrari, S. Guadarrama, N. Silberman, L.-C. Chen, A. Fathi, and J. Uijlings. The devil is in the decoder. *arXiv preprint arXiv:1707.05847*, 2017.

[156] K. C. Wong, M. Moradi, H. Tang, and T. Syeda-Mahmood. 3D segmentation with exponential logarithmic loss for highly unbalanced object sizes. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 612–619. Springer, 2018.

[157] Y. Wu and K. He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018.

[158] Z. Wu, C. Shen, and A. Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.

[159] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1403, 2015.

[160] X. Xu, Q. Lu, L. Yang, S. Hu, D. Chen, Y. Hu, and Y. Shi. Quantization of fully convolutional networks for accurate biomedical image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8300–8308, 2018.

[161] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang. SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation. *Neuroinformatics*, 16(3-4):383–392, 2018.

[162] D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas, and D. Comaniciu. Automatic liver segmentation using an adversarial image-to-image network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 507–515. Springer, 2017.

[163] Q. Yang, N. Li, Z. Zhao, X. Fan, E. I. Chang, Y. Xu, et al. MRI cross-modality neuroimage-to-neuroimage translation. *arXiv preprint arXiv:1801.06940*, 2018.

[164] X. Yang, L. Yu, L. Wu, Y. Wang, D. Ni, J. Qin, and P.-A. Heng. Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[165] B. Yu, L. Zhou, L. Wang, J. Fripp, and P. Bourgeat. 3D cGAN based cross-modality MR image synthesis for brain tumor segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pages 626–630. IEEE, 2018.

[166] W. Zhang, C. Witharana, A. Liljedahl, and M. Kanevskiy. Deep convolutional neural networks for automated characterization of arctic ice-wedge polygons in very high spatial resolution aerial imagery. *Remote Sensing*, 10(9):1487, 2018.

[167] Y. Zhang, S. Miao, T. Mansi, and R. Liao. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 599–607. Springer, 2018.

[168] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 408–416. Springer, 2017.

[169] Z. Zhang, L. Yang, and Y. Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9242–9251, 2018.

[170] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 269–284, 2018.

[171] H. Zhao, H. Li, and L. Cheng. Synthesizing filamentary structured images with GANs. *arXiv preprint arXiv:1706.02185*, 2017.

[172] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[173] M. Zhao and G. Hamarneh. Retinal image classification via vasculature-guided sequential attention. *International Conference on Computer Vision workshop on Visual Recognition for Medical Images*, 2019.

[174] M. Zhao and G. Hamarneh. Tree-LSTM: Using LSTM to Encode Memory in Anatomical Tree Prediction from 3D Images. *Medical Image Computing and Computer-Assisted Intervention Workshop on Machine Learning in Medical Imaging*, 2019.

[175] T. Zhao, Y. Yang, H. Niu, D. Wang, and Y. Chen. Comparing U-Net convolutional network with Mask R-CNN in the performances of pomegranate tree canopy segmentation. In *Asia-Pacific Remote Sensing*, 2018.

[176] X. Zhen and S. Li. Towards direct medical image analysis without segmentation. *arXiv preprint arXiv:1510.06375*, 2015.

[177] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017.

[178] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen. High-resolution encoder-decoder networks for low-contrast medical image segmentation. *IEEE Transactions on Image Processing*, 2019.

[179] T. Zhou, S. Ruan, and S. Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, page 100004, 2019.

[180] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[181] X. Zhu, H. Zhou, C. Yang, J. Shi, and D. Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 568–583, 2018.