HHS Public Access

Author manuscript

Med Image Anal. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

Med Image Anal. 2021 January ; 67: 101841. doi:10.1016/j.media.2020.101841.

Segmentation of Cellular Patterns in Confocal Images of Melanocytic Lesions in vivo via a Multiscale Encoder-Decoder Network (MED-Net)

Kivanc Kose[a,1,*], **Alican Bozkurt**[b,2], **Christi Alessi-Fox**[d], **Melissa Gill**[e,f], **Caterina Longo**[g], **Giovanni Pellacani**[g], **Jennifer Dy**[c], **Dana H. Brooks**[c], **Milind Rajadhyaksha**[a]

[a]Dermatology Service, Memorial Sloan Kettering Cancer Center, New York, 11377, NY, USA

[b]Paige, New York, 10036, NY, USA

[c]Electrical and Computer Engineering Department, Northeastern University, Boston, 02115, MA, USA

[d]Caliber Imaging and Diagnostics, Rochester, Rochester, 14623, NY, USA

[e]Department of Pathology at SUNY Downstate Medical Center, New York, 11203, NY, USA

[f]Skin Medical Research Diagnostics, P.L.L.C., Dobbs Ferry, 10522, NY, USA

[g]University of Modena and Reggio Emilia, Reggio Emilia, Italy

Abstract

In-vivo optical microscopy is advancing into routine clinical practice for non-invasively guiding diagnosis and treatment of cancer and other diseases, and thus beginning to reduce the need for traditional biopsy. However, reading and analysis of the optical microscopic images are generally

[*]Corresponding author: Tel.: +1-646-888-6241; kosek@mskcc.org;.

[1]Shared first authorship

[2]Work done while author was affiliated with Northeastern University

Credit Author Statement

Kivanc Kose : Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing the Original Draft, Reviewing & Editing the final manuscript.

Alican Bozkurt : Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing the Original Draft, Reviewing & Editing the final manuscript.

Christ Alessi-Fox : Conceptualization, Data Curation, Formal Analysis, Investigation, Validation, Reviewing & Editing the final manuscript.

Melissa Gill : Conceptualization, Data Curation, Formal Analysis,Investigation, Validation, Reviewing & Editing the final manuscript.

Caterina Longo : Data Curation, Validation, Reviewing & Editing the final manuscript.

Giovanni Pellecani : Data Curation, Validation, Reviewing & Editing the final manuscript.

Jennifer Dy : Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project administration, Validation, Writing the Original Draft, Reviewing & Editing the final manuscript.

Dana H. Brooks : Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project administration, Validation, Writing the Original Draft, Reviewing & Editing the final manuscript.

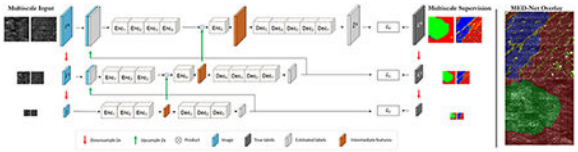
Milind Rajadhyaksha : Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project administration, Validation, Writing the Original Draft, Reviewing & Editing the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

still qualitative, relying mainly on visual examination. Here we present an automated semantic segmentation method called Multiscale Encoder-Decoder Network (MED-Net) that provides pixel-wise labeling into classes of patterns in a quantitative manner. The novelty in our approach is the modeling of textural patterns at multiple scales (magnifications, resolutions). This mimics the traditional procedure for examining pathology images, which routinely starts with low magnification (low resolution, large field of view) followed by closer inspection of suspicious areas with higher magnification (higher resolution, smaller fields of view). We trained and tested our model on non-overlapping partitions of 117 reflectance confocal microscopy (RCM) mosaics of melanocytic lesions, an extensive dataset for this application, collected at four clinics in the US, and two in Italy. With patient-wise cross-validation, we achieved pixel-wise mean sensitivity and specificity of 70 ± 11% and 95 ± 2%, respectively, with 0.71 ± 0.09 Dice coefficient over six classes. In the scenario, we partitioned the data clinic-wise and tested the generalizability of the model over multiple clinics. In this setting, we achieved pixel-wise mean sensitivity and specificity of 74% and 95%, respectively, with 0.75 Dice coefficient. We compared MED-Net against the state-of-the-art semantic segmentation models and achieved better quantitative segmentation performance. Our results also suggest that, due to its nested multiscale architecture, the MED-Net model annotated RCM mosaics more coherently, avoiding unrealistic-fragmented annotations.

Graphical Abstract



Keywords

Reflectance confocal microscopy; melanocytic lesion; semantic segmentation; dermatology; *in vivo* segmentation

1. Introduction

Many areas of medical and biological imaging have seen a recent upsurge in automated diagnosis systems using deep neural nets (DNNs). This trend is pretty much similar in many areas of traditional pathology (Litjens et al., 2017; Campanella et al., 2019; Chen et al., 2018). However, the clinical application of medical imaging often involves extraction and analysis of information specific to needs of the imaging modality which can not be effectively assessed using methods designed for natural images. Typical challenges in these settings include large intrinsic variability, weak or inconsistent contrast, the presence of key structures in the images at distinct scales, significant class imbalance, the laborious and involved data labeling process, and the need for interpretability in terms of clinically relevant physiological features. These challenges prevent standard DNNs, even those designed for analyzing standard microscopy-based histopathological images, from achieving clinical utilization. In this work, we address one edge case of this type, analysis of morphological patterns of cellular structures in reflectance confocal microscopy (RCM) images of pigmented skin lesions.

As we explain below, RCM has been shown to have the potential for a high impact on the assessment of such lesions and can significantly improve clinicians' ability to make accurate and reliable screening decisions on which lesions to biopsy. However, a wider adoption of RCM is hindered significantly because the images are very different visually from standard histopathology, thus making them an edge case in that context. For that same reason, automated analysis tools require solutions that go beyond standard DNN approaches and that address the challenges listed in the previous paragraph. We report here on the motivation, structure, and evaluation of a DNN architecture, which we call Multiscale Encoder-Decoder Network (MED-Net), that was explicitly designed to overcome these edge case challenges.

Analysis of pigmented skin lesions is critical, with skin cancer being a serious medical problem worldwide. About 5.4 million new cases detected in the USA and another million in other regions (primarily parts of Europe, Canada, UK, Australia, New Zealand) (Nikolaou and Stratigos, 2014). Diagnosis costs are about $3 billion, and treatment costs another $8 billion per year in the USA (Guy Jr et al., 2015). RCM is an emerging non-invasive optical diagnostic tool based on examination of living tissue morphology directly on patients, on the fly, and at the bedside or in the clinic. After more than two decades of development and translation, *in vivo* RCM is advancing into clinical practice for non-invasively guiding diagnosis and treatment of cancer (Rajadhyaksha et al., 2017). RCM imaging, combined with the current clinical standard for visual examination, known as dermoscopy, reduces the benign-to-malignant biopsy ratio by about a factor of two compared to dermoscopy alone (Alarcon et al., 2014; Pellacani et al., 2014, 2016; Borsari et al., 2016).

Although RCM images have a μm-level resolution like standard histopathology, their appearance is quite different because they are collected *in vivo*. One difference is that the images are acquired in an en face orientation, as opposed to the "vertical" (i.e. normal to the skin surface) sections typically used in the pathology of excised specimens. Another is that, due to lack of *in vivo* contrast agents, images have only one source of contrast, reflectance, and therefore are displayed in grayscale, whereas standard H&E pathology is in color contrast (the purple and pink appearance). Instead of color contrast, skin and cellular structures are differentiated by intricate multiscale textural patterns in RCM images.

Diagnosis of melanocytic lesions using RCM is primarily based on the identification of four cellular morphological patterns in RCM mosaics acquired at the dermal-epidermal junction (DEJ). These mosaics typically span rectangular-shaped areas with 4-6 mm at one side (Scope et al., 2017). The patterns in the mosaics are composed of heterogeneous cellular formations, appear at highly varying scales with highly varying shapes, and with diffused transition boundaries in between. Moreover, the images are contaminated by intrinsic speckle noise. All these aspects are characteristic of high-resolution optical microscopy *in vivo*.

These characteristics present challenges for human readers who are trained extensively to interpret H&E pathology. Learning to read and perform a qualitative examination of RCM images demands significant effort and time for novices, and results tend to be highly subjective, with high levels of inter-reader variability even among experts. The steep

learning-curve and large inter-reader variability have become a significant impediment to broader RCM adoption by clinicians, which strongly motivates the development of automated computational tools for both clinical guidance and clinical training.

Existing medical image segmentation applications are developed for identifying target structures that typically have

1. predefined shapes with noticeable boundaries (e.g. organs (Nie et al., 2016; Yu et al., 2017), cells (Ronneberger et al., 2015; Falk et al., 2019)),
2. distinct contrast compared to the background (e.g. cells, retinal vessels (Fu et al., 2016)),
3. predefined spatial location within the view (e.g. organs, retinal layers (Gu et al., 2019), lesions (Marchetti et al., 2018; Codella et al., 2018)).

On the other hand, the morphological structures encountered in RCM images are complex in shape, have ambiguous boundaries, vary in size, change appearance under inherent speckle noise, and appear at arbitrary spatial locations within the field of view. Therein our experience has convinced us that neither the existing semantic segmentation approaches developed for other medical imaging modalities (Ronneberger et al., 2015; Falk et al., 2019; Nie et al., 2016; Yu et al., 2017; Marchetti et al., 2018; Codella et al., 2018) nor the existing very deep neural network architectures (Badrinarayanan et al., 2017; Chen et al., 2016) can be effectively used for RCM mosaics. These models contain very large numbers of parameters to optimize, making them prone to overfitting with the type of limited and class-imbalanced training data available for RCM. Moreover, in deep network architectures with limited training data, the training of the layers which are farther away from the output is challenging as the partial derivatives that define the coefficient updates tend to get smaller as the error propagates from the output towards the input layers.

To respond to these particular challenges of automated analysis of RCM images, we developed a multiscale neural network called MED-Net for semantic segmentation of textural patterns in segmented lesions, based on the morphological patterns that have been defined by expert RCM readers. The architecture of MED-Net was driven by two key observations about clinical practice. First, our multiscale structure was inspired by the typical procedure for examining pathology in RCM mosaics clinically, which routinely starts with low magnification and low resolution in a large field of view (2X-4X, ~1-5 μm/px, over 5-10 mm) followed by closer inspection of suspicious areas with higher magnification and higher resolution in smaller fields of view (10X-40X, 0.2-1.0 μm/px, over 0.5-2 mm), and then often returns to lower magnification to integrate features found at higher resolution into a broader semantic setting. MED-Net models textural patterns at multiple scales (magnifications, resolutions), starting from a coarse scale and proceeding to finer scales. Semantic segmentation at each scale is handled by subnetworks, which are fully convolutional encoder-decoder neural networks capable of generating label maps at the same scale as their input. The capacity (number of layers and coefficients) of the subnetworks depends on the complexity of the segmentation task at the given scale (e.g. coarser scales use smaller subnetworks as there is less detail at those scales). Consecutive subnetworks in the multiscale hierarchy explicitly cooperate, leveraging the correlation across scales. Each

subnetwork utilizes the encoded feature representation (called the bottleneck representation) from the immediate predecessor subnetwork by integrating it into its feature representation at the equivalent level.

Similarly, the semantic segmentation estimation of each subnetwork is used as a prior in the subnetwork at the finer scale, so that each subnetwork only refines the coarser-scale estimates rather than solving the whole segmentation problem from scratch. However, using several subnetworks in a cascaded fashion makes the model rather deep and can make training difficult. To solve this problem, we employ a method called “deep supervision” (Zhu et al., 2017). We compare the output of the subnetwork at every scale against ground truth segmentation downsampled to the same scale. This supervision gives us direct access to deeper layers (early subnetworks) and allows efficient updates to avoid vanishing gradients during training.

Second, we use a set of four cell-morphological patterns (textural structures) that have been identified by clinicians (Scope et al., 2017) along with two “extra” classes for artifacts and non-lesion background. Rather than designing a binary classifier to simply classify lesions as suspicious or non-suspicious, we aim to respond to clinicians’ need for reasoning in diagnostics by providing them a scheme that reports more finely grained results in this setting. Indeed, given this critical need and its advantage for both rapid reader throughput and education, it is of critical importance to generate segmentation masks for each pattern class rather than just binary classifications. Similarly, we chose pixel-wise instead of image-wise classification, because in the latter, the clinician only has access to the final diagnostic prediction, while pixel-wise segmentation reports the spatial location of the diagnostic findings, making the diagnostic process more interpretable.

The precursor to MED-Net, named MUNet, was developed as a feasibility study (Bozkurt et al., 2018). Here we significantly extend MUNet in the following ways:

1. MUNet only provides feedback between consecutive layers via output label maps, whereas MED-Net also shares feature representations between consecutive subnetworks (Fig. 3, Section 2.1).
2. We trained MED-Net using a novel loss function that incorporates a total variation constraint to regularize the smoothness of the output label maps (Section 2.2).
3. We greatly expanded the dataset used to train and test MED-Net compared to MUNet, using what is, in the RCM context, an unprecedentedly rich set of labeled data, 117 mosaics, collected at six different clinics in the US (4) and Italy (2). In addition to only having more data available, here we were able to carry out cross-validation with data stratified by clinic-of-origin, providing a more realistic prediction of future performance. We note that while in the context of DNNs, this is a rather small dataset, it is large for RCM due to the difficulty of labeling, an aspect of the “edge case” nature of this problem.

Labeling datasets is laborious and challenging, even for experts. Indeed, only 58% of the pixels in the dataset were labeled by our experts due to these difficulties. Thus, another

feature of MED-Net is the ability to train on "partially-labeled" data, where only arbitrarily-shaped parts of training images are labeled, but be capable of classifying full images. In our quantitative evaluations, we can only compare to the labeled pixels as we only have ground-truth there, but we show our visual segmentation results on the full images (Fig. 4). We evaluated the segmentation performance of MED-Net using the Dice coefficient, as well as the sensitivity and the specificity of the model in identifying the patterns. We compared MED-Net results against 4 well-known DNN models (FCN (Long et al., 2015), SegNet (Badrinarayanan et al., 2017), DeepLab (Chen et al., 2016) and UNet (Ronneberger et al., 2015)).

In the following sections, we discuss the design of MED-Net in detail, explain the algorithmic choices we made to overcome unique issues encountered in semantic segmentation of *in vivo* microscopy images, and present the results of our tests on mosaics of melanocytic skin lesions.

2. Materials and Methods

Our study set is composed of 117 RCM mosaics of melanocytic skin lesions collected at the DEJ level. 31 of these mosaics were acquired at 4 different clinics in the US (Memorial Sloan Kettering Cancer Center (New York, NY), University of Rochester (Rochester, NY), Loma Linda University Health (Loma Linda, CA), and Skin Cancer Associates (Plantation, FL)) and the other 86 at clinics at the University of Modena and Reggio Emilia (Italy). All mosaics were collected under the required IRB (USA) and Ethics Committee (EU) approvals and de-identified (patient metadata was removed). The study set was chosen to reflect the data diversity encountered in daily clinical practice. At each clinic, the imaging was carried out with a commercial confocal microscope (Vivascope 1500, Caliber I.D.) with a spatial resolution of 0.5 μm/px. Mosaic sizes varied from 7000 × 8000 pixels up to 12000 × 12000 pixels, corresponding to an area between 14 and 36 mm^2. The size of the mosaics was determined by the clinical need to be able to evaluate the cellular morphological patterns that characterize melanocytic lesions accurately.

We set as our goal the segmentation of these mosaics into six clinically important classes. Four of them are cellular morphological patterns, i.e. ring, meshwork, nested, and aspecific. These patterns are routinely observed in RCM mosaics of melanocytic neoplasm collected at the DEJ (Scope et al., 2017). We added two additional classes for non-lesion areas and areas dominated by imaging artifacts (Gill et al., 2019), leading to six total classes in our segmentation task.[3] Exemplars of these six classes are shown in Fig. 1.

Ground truth maps for these six classes came from labels determined by the consensus of 2 expert readers (co-authors MG and CAF), labeled using the open-source software package Seg3D (University of Utah, (CIBC, 2016)). Labeling was conducted in a non-exhaustive manner, meaning that pixels not labeled as any of the six classes were given a distinct

[3]Regions that were beyond the lesion borders or within the lesion but not at the DEJ were classified as non-lesion. Artifact regions and their imaging characteristics include saturation (too bright), under-illumination (too dark). Air bubbles in the index-matching oil (appear as bright highly reflecting blobs or amorphous structures) or in the immersion media (appear as large dark round/oval areas) can also obscure the underlying tissue morphology, resulting in artifacts in the images (Gill et al., 2019).

"ignore" label. Pixels were not labeled either because the distinction between the labels was not clear due to the existence of mixed patterns or because they would have required excessive time and effort to label, in the readers' judgement. Overall, 58% of the pixels were labeled (Table 1). We show a sample labeled mosaic in Fig. 2. The unlabeled portions of the mosaics were omitted during both training and quantitative testing. However, the readers qualitatively assessed the algorithm's segmentations even for these unlabeled regions. The distribution balance of the six labels over the whole dataset is given in Table 1.

2.1. Semantic Segmentation Network Architecture

MED-Net is composed of multiple encoder-decoder subnetworks nested together (Fig. 3). Each subnetwork processes the input image starting at a specific scale and outputs a segmentation map at the same scale. To the best of our knowledge, MED-Net is different from existing networks in the following aspects. In similar existing approaches (Lin et al., 2017; Jiang et al., 2018; Amirul Islam et al., 2017; Chen et al., 2016; Zhao et al., 2017; Fu et al., 2018; Zhou et al., 2018; Gu et al., 2018; Li et al., 2017; Zhang et al., 2019), the subnetworks are cascaded so that they share only features (ultimate output of encoder blocks) across networks, or else they independently solve the same segmentation problem and then, only at the end, fuse the results. More similar to MED-Net, Eigen and Fergus (2015) use three separate networks to process the input images at different scales in a cascaded manner resembling our approach. They feed the output of subnetworks into the input of the following subnetworks, so the individual models provide feedback to each other. However, in their approach, due to lack of feedback at the individual subnetwork level (e.g. deep supervision (Zhu et al., 2017)), the output of each subnetwork is not final output (e.g. in their case, a depth map) at respective scale, but a feature representation. Auto-Context is another similar approach that was first introduced for MRF/CRF based brain segmentation (Tu and Bai, 2009). Similar to our approach, auto-context based methods solve the segmentation problem in a recursive manner. In Mirikharaji et al. (2018), the authors adapted this recursive structure to skin lesion segmentation in dermoscopy images but they do not use it for multi-resolution analysis and require partitioning the dataset into distinct subsets of the images to train networks at different levels. Our model can be trained end-to-end at all levels, without dataset partitioning. In Mohseni Salehi et al. (2017), rather than using a multi-scale approach, the authors used multiple fields of view, which, while effective for brain MRI, is different from our approach.

Unlike all these approaches, MED-Net shares intermediate results in two ways. It shares the segmentation outputs across subnetworks (Fig. 3) by using them as a prior that becomes part of the input for subsequent subnetworks. Through the use of deep supervision (Zhu et al., 2017), the output of each subnetwork is compared against a ground truth segmentation and forced to be an intermediate label prediction at the given scale it operates.

Moreover, MED-Net also shares feature representations between matching levels of consecutive subnetworks (orange colored boxes in Figure 3). These subnetwork interconnections are not present in previous approaches (Lin et al., 2017; Jiang et al., 2018; Amirul Islam et al., 2017; Chen et al., 2016). Backpropagating the final loss through the network can lead to inefficient training of the layers that are farther from the output.

Therefore, to effectively train the individual subnetworks, we provide direct feedback to them $\mathscr{L}_i$ at the end of each network, a method known as deep supervision (Zhu et al., 2017). Overall, sharing intermediate feature representation, using intermediate label predictions as priors, and deep supervision to individual subnetworks are the three main innovations in the MED-Net architecture.

The elementary units of subnetworks in MED-Net consist of residual blocks (He et al., 2016), which are generally concatenations of convolutions, non-linearities (e.g. leaky ReLU), and batch normalizations. The sequence of downsampling processing blocks (encoder) is followed by a sequence of upsampling processing blocks (decoder). Downsampling is carried out via the non-unity stride of the first convolution operation in the residual blocks, and bilinear upsampling is applied to processing block outputs. If we had a single scale, the architecture would be very similar to a Fully Convolutional Network (FCN32) (Long et al., 2015) with encoder-decoder topology. However, here we have M subnetworks that solve the segmentation problem starting from a different scale of the input image. Subnetworks in this cross-scale hierarchy share information (feature representations) directly through skip connections from bottleneck representations of their predecessor scale subnetwork – the orange colored boxes and green arrow connections in Figure 3. This information exchange is done via multiplication of tensor representations at comparable scales to act like attention mechanisms (Roy et al., 2018). Also, the output segmentation probability map (a vector of six probabilities per pixel) at each scale (except the finest) is upsampled and then concatenated with the original or directly downsampled image at the next finer scale and used as the input for the subnetwork at that next scale. More precisely, let I^0 and L^0 be the original image and corresponding ground truth labeled image, and I^m and L^m be those images after 2^m times downsampling in both spatial dimensions ($m = 0, \ldots, M-1$). The subnetwork at the coarsest scale takes only I^{M-1} as input and produces a probability map $\hat{L}^{M-1}$, which represents the likelihood of each pixel belonging to a particular class. For all other subnetworks (i.e. $m \in [0, M-2]$), we fuse the segmentation coming from subnetwork $m+1(\hat{L}^{m+1})$ with the level m version of the input (I^m) via concatenation as illustrated by blue-white colored inputs to the second and third level networks. The final segmentation probability map is $\hat{L}^0$, which is at the same resolution as the input image of the overall model.

The subnetwork depth parameter M is a design choice, and one can also vary the scale factor between subnetworks, which we set to 2, leading to a 3-level version of MED-Net. Likewise, the scale difference of the input between consecutive levels is another design choice and can be determined according to needs and computational capabilities. In addition, the overall architecture is modular in the sense that one can replace our subnetwork architecture (including a different design of the processing blocks) with any other relevant subnetwork architecture and then assemble a MED-Net version of that network.

Each MED-Net subnetwork for $m > 0$ has the same architecture as the subnetwork at scale $m-1$ but with two additional blocks: One encoder block before the bottleneck feature representation and one deconvolution block at the input of the decoder. Note that the weights in each corresponding block differ across subnetworks; weights are not shared between

layers. Information is shared between subnetworks only through the skip connections described above.

2.2. Loss function

We modified the soft-Dice loss function (Milletari et al., 2016) to take three distinct factors into account:

1. Appropriateness of segmentation (e.g. generating labels that change smoothly across the image).
2. Ability to handle imbalances in label distribution of the training data.
3. Applicability to multiclass labeling. Thus we used a modified version of the soft-Dice loss calculated between $\hat{L}^m$ and L^m (see Fig. 3) at each level.

The standard Dice Similarity Coefficient DSC(A, B) = $2|A \cap B|/(|A| + |B|)$ (Dice, 1945) is commonly used for binary segmentation and is known to be robust against label imbalance in the data. In its original binary formulation, DSC explicitly represents only true-positive samples, while true-negative cases are automatically optimized simultaneously. However, similar to Salehi et al. (2017), we found that directly extending this formulation to the multilabel case by treating each label as a binary classification task did not put enough emphasis on true-negatives samples. Our modified soft-Dice coefficient combines two components. The first component measures the soft Dice similarity only considering true positive samples, whereas the second component takes the true negative samples into account during the loss calculation, as described next.

Suppose we have $W \times H \times K$ sized tensors L^m and $\hat{L}^m$, where L^m is one-hot encoded ground truth at the subnetwork level m. The entries $\hat{L}^m_{ij} = \mathbf{e}_k$ if pixel (i, j) is labeled as class k, where $\mathbf{e}_k$ is a one-hot vector of length K with 1 in its k^{th} entry and 0 everywhere else. $\hat{L}^m$ is the neural network output, such that at each (i, j) pixel, $\hat{L}^m_{ijk} \in [0, 1]$ and $\sum_k \hat{L}^m_{ijk} = 1$. Our modified loss function is:

$$\mathrm{MDSC}\left(L^m, \hat{L}^m\right) = \sum_{k=0}^{K-1}\left(1 - \frac{2\sum_{i,j} L^m_{ijk}\hat{L}^m_{ijk}}{\sum_{i,j}\left(L^m_{ijk}\right)^2 + \left(\hat{L}^m_{ijk}\right)^2 + \epsilon}\right) + \sum_{k=0}^{K-1}\left(1 - \frac{2\sum_{i,j}\left(1 - L^m_{ijk}\right)\left(1 - \hat{L}^m_{ijk}\right)}{\sum_{i,j}\left(1 - L^m_{ijk}\right)^2 + \left(1 - \hat{L}^m_{ijk}\right)^2 + \epsilon}\right)$$

where ϵ is a small value in order to avoid division by zero. The first part of the equation is the standard soft-Dice loss, which encourages agreement between true positive labels, while the second part of the equation also encourages agreement between true negative predictions. To ensure smoothness of the prediction label map and avoid small isolated segmentation labels, we regularize the loss function using the total variation (TV) of the output label map.

$$\mathrm{TV}\left(\hat{L}^m\right) = \sum_{i,jk}\left|\hat{L}^m_{i+1,j,k} - \hat{L}^m_{i,j,k}\right| + \left|\hat{L}^m_{i,j+1,k} - \hat{L}^m_{i,j,k}\right|.$$

Combining MDSC and TV losses, the loss applied at each subnetwork level is

$$\mathscr{L}_m = \mathrm{MDSC}\left(L^m, \hat{L}^m\right) + \gamma TV\left(\hat{L}^m\right).$$

We set the regularization parameter empirically, $\gamma = 10^{-6}$, which kept the total variation cost to [0.1, 0.01] of the soft-Dice loss. In our experiments, we observed that keeping the total variation cost within this range of the soft-Dice loss provided a good balance between smoothness and the accuracy of produced label maps.

As shown in Fig. 3, we calculate $\mathscr{L}^m$ between outputs of each subnetwork and the label map at the respective scale for each scale m, and the overall loss as the sum of losses across all subnetworks/scales $\mathscr{L} = \sum_{m=0}^{M-1} \mathscr{L}^m$. Doing so, we effectively gain direct access to the deeper layers of the network, as is done with deep supervision (Zhu et al., 2017). However, the subnetworks are not trained disjointly as they are connected via skip connections, resulting in joint optimization of all subnetwork parameters.

2.3. Implementation Details

In this section, we discuss specific parameter choices in our implementation of MED-Net on RCM mosaics. These choices were made to fit available hardware resources (e.g. GPU memory, number of GPUs) and problem characteristics (e.g. data sampling and augmentation scheme). We report them so that readers can replicate our work, and we also anticipate that they will provide a guideline towards applying this structure to other segmentation problems.

Before training the MED-Net model, we needed to make two important choices regarding; (i) the resolution of the mosaics to be processed and (ii) the size of the input images to the network. Although the network architecture can segment arbitrarily sized images, we processed the RCM mosaics in patches (portions of the mosaic) due to memory limitations of the GPU we used. Note that the patches needed to be larger than 2^4 pixels per dimension because we used 2-strides (effectively downsampling by 2) at least at 4 levels of encoder blocks. To determine useful patch-sizes, we consulted our expert readers, who reported that in their experience, the morphological patterns of interest could still be reliably identified at 2 μm/px resolution, 4-times lower than that of the RCM acquisition system. Thus before feeding the mosaics to MED-Net, we downsampled them by 4. The readers also reported that a 0.5 mm × 0.5 mm field of view is typically large enough to identify these same patterns reliably. Thus we processed the mosaics in patches of 256 × 256 pixels after downsampling.

All models are trained using the same training parameters. We trained each model for 200 epochs, using a base learning rate of 0.01, batch size of 48, and weight decay of 10^{-8}. We exponentially decayed the learning rate to one-tenth of the base value throughout the training. For a fair comparison, we kept the number of trainable parameters for all networks at 6 million. All the convolutional layers are initialized with He Normal initialization (He et al., 2016).

We also implemented data augmentation through spatial sampling. In order to cover all possible patches that could be extracted from the mosaic, we devised the following patch extraction procedure. Before each epoch, we extract 512 × 512 pixels patches in a sliding window fashion with a 50% overlap. Then, at each epoch of training, we extracted 256 × 256 pixel patches at random locations within the larger patches.

In order to account for inevitable variations during RCM image acquisition, such as changes in laser power (illumination intensity), distortion in tissue, speckle noise, and the orientation of the microscope, we applied data augmentation on the extracted patches. At each epoch, we

1. rotated each patch at a random angle up to 180 degrees
2. randomly flipped the patch horizontally and vertically,
3. added a random intensity value within [−20, 20][4]
4. zoomed in/out randomly up to 10%,
5. randomly sheared the patches ($\theta = 0.2$),
6. added signal-dependent Gaussian-distributed pseudospeckle noise (with uniform random multiplication parameter of 0.2).

During inference, the output of the networks is six probability maps, one for each label (represented as a 256 × 256 × 6 tensor) over a 0.5 mm^2 field of view. Due to the use of padded convolutions, the network produces less reliable segmentation results at the borders of the patches. To compensate, we extracted and processed patches in an overlapping fashion, resulting in multiple soft decisions for each pixel. Specifically, we extracted patches at a stride of 32 pixels, leading to up to 8 different decisions per pixel. We then weighted each patch's probability map for each label with a spatial Gaussian mask whose variance was half of the patch size before summing the overlapping probability maps. Finally, we chose the class with the highest resulting probability for each pixel.

3. Results

We report the results of testing on two distinct training scenarios. In Scenario 1, we pooled data across all sites, then stratified by the patient for training, validation, and testing (5-fold stratified cross-validation). In Scenario 2, we first stratified by clinics, only used the data from clinics in Europe for training and validation, and then tested only on data from the US. The validation set was used to probe the performance of the model throughout training, and the test set was used to evaluate the performance of the trained models quantitatively. We chose to train on the European data and test on the US data, and not vice-versa, both due to the limited size of the US data set and also because the US data came from a larger number of clinics, thus better mimicking a more realistic application scenario. Results from the first scenario are described in Section 3.1 and results from the second scenario in Section 3.2. Each fold used in Scenario 1 is also stratified by the class label in the training/test split to ensure a representative sampling of training data in the face of the class imbalance in our

[4]At the augmentation stage, the pixels values are in the range [0,255] so we clipped the added intensity for the brightest pixels.

data. Specifics of the data distribution over the training, validation, and test sets for both scenarios are given in Table 1.

In addition to MED-Net, we also tested 4 other widely used deep segmentation networks; FCN (Long et al., 2015), SegNet (Badrinarayanan et al., 2017), DeepLab (Chen et al., 2016), and UNet (Ronneberger et al., 2015) for comparison purposes. To try to ensure fair comparisons, we used a similar number of trainable parameters in each network ($\sim 6 \times 10^6$). All of the networks were trained using similar training parameters (e.g. learning rate, weight decay, batch size) for 200 epochs using the MDSC+TV loss described above. The performance of the models are presented in terms of sensitivity, specificity and Dice coefficient metrics. As our task is multiclass segmentation, we use the one vs. rest approach to determine binary outputs and binary ground truth for each class, which we then use to calculate sensitivity, specificity, and the Dice coefficient.

3.1. Scenario-1: Patient-Wise Cross-Validation Experiment

As described above, in this scenario, we “patient-wise partitioned” the dataset into 5 stratified folds, meaning that each fold contained similar proportions of class labels. Training, validation, and test sets approximately corresponded to 70, 10, and 20 percent of the data in each fold, respectively.

In Table 2, we present the segmentation performance of all four networks for Scenario 1 in terms of sensitivity, specificity, and the Dice coefficient. On average, MED-Net modestly outperforms the other networks in terms of sensitivity (by 0.02 to 0.12), although the comparison differs across classes. On specificity, all four networks perform similarly both on average and by class. The Dice coefficient values are consistently better for MED-Net than the compared methods except for FCN on the Nest class. In general, FCN was the closest to MED-Net.

A closer comparison of the model output with ground truth labels revealed that in general, all models confused the meshwork class with the ring and aspecific classes. This result is interesting, because anecdotally we have been told that novice clinicians also suffer from the same problem due to the wide range of variations in the appearance of the meshwork pattern. Moreover, visual examination of the results by our experts confirmed that most of the falsely classified meshwork pattern samples contain “deformed” variations of the pattern, which they reported are typically also misclassified by novice readers.

To obtain a qualitative assessment of MED-Net outputs, we presented the segmentation maps produced by MED-Net to our experts. In particular, we asked them to review the automated annotation of the algorithm over the “unlabeled areas”. Their qualitative assessment of the results was very positive and confirmed that the model performed well in annotating most of the unlabeled areas in the mosaics. Beyond this qualitative assessment, we do not claim any success measure in those regions. We show an example in Fig. 4. The gray-colored areas in the figure represent the unlabeled areas. MED-Net typically extended the labels of the neighboring labeled areas over the unlabeled sections, providing smoother label maps than the other methods.

3.2. Scenario-2: Clinic-wise Cross-Validation

To assess how the models generalize across clinical settings, we trained them over the data collected in Italy (86 mosaics) and tested on data collected at 4 US clinics (31 mosaics). In this case, we were not able to keep the incidences of the labels in the training and test sets at similar levels (Table 1). In the training set, [18,20,21,6,23,12] percent of the labeled pixels were, [background, artifact, meshwork, nested, ring and aspecific] patterns respectively; whereas in the test set the ratios were [8,23,23,5,36,5] percent. We used the same network model architectures and training parameters that we used in Scenario 1 for both MED-Net and the other networks.

In Table 3, we summarize the segmentation performance of these networks in terms of sensitivity, specificity, and Dice coefficient. In general, performances of all the networks were close to what we observed on the patient-wise stratification, with only modest decreases in the performance metrics. Overall, MED-Net outperformed all the other networks in terms of averages across classes, particular with regards to sensitivity and Dice coefficient. Specificity values were generally very high for all networks on all classes, and for some classes, other networks had sightly higher specificity than MED-Net.

3.3. Ablation Studies

We conducted 2 ablation studies to investigate how multiscale analysis and the proposed loss function each affect performance. We compared ablation results to our baseline model (the 3-level MED-Net trained using MDSC+TV loss, see Section 3.2). We followed the same training and testing procedures in Section 3.2.

To test the effect of the multiscale approach, we trained 1-level and 2-level MED-Net models and compared them to the 3-level MED-Net. For a fair comparison, the number of trainable parameters for all the models is kept at 6 million. The results in Table 4 show that using the multiscale analysis improves the segmentation performance. We stopped at 3 levels because a fourth level would necessarily decrease the resolution below the size of the most of the relevant features in the images.

To test the effect of the loss function, we trained the same baseline MED-Net model using cross-entropy, Dice loss functions, and compare the results against our MDSC+TV loss defined in Section 2.2. The results in Table 5 show that using MSDC+TV as the loss function results in the best segmentation performance in terms of average Dice coefficient over all classes.

4. Discussion and Conclusions

In this article, we present a deep-learning based semantic segmentation algorithm developed specifically for a class of *in vivo* microscopy applications. Textural patterns of cellular morphology in in-vivo optical microscopy images are unique and different from those in natural images. Hence, the features developed for natural images do not perform well on such microscopy images. Deep learning-based semantic segmentation models provide the possibility of learning both the feature representation and the classification model in an integrated fashion, allowing greater flexibility in capturing the relationships between pixels

that encode complex morphological patterns like those present in RCM images. Semantic segmentation also addresses the need for interpretable machine-learning-based image analysis by providing more granular information (e.g. pixel wise label prediction rather than just a single prediction score) to the user. This reasoning and interpretability can facilitate acceptance and adoption of machine learning-based approaches among a clinical community (Goodman and Flaxman, 2017).

We report several promising results in this study. Although the average sensitivity is moderate, the specificity is very high; MED-Net performed very well at detecting the absence of particular patterns and reported a small number of false positives. Hence, a clinician could be highly confident about the accuracy of true negative results reported by the model. Moreover, Dice coefficients of 0.73-0.75 show that the model is not only good at detecting the existence of a pattern but also successfully finds the location and the extent of the pattern. On the other hand, due to its modest sensitivity, clinicians should be aware that the model may miss patterns that are present (true positives) in the data. We highlight that the transition between the distinct morphological patterns in RCM images is smooth (e.g. it is common to see ring pattern smoothly evolving into meshwork or vice versa) without clear borders, unlike other medical applications such as organ segmentation or brain tumor segmentation, where the borders are sharp at the resolution of the imaging. This limits the precision of the ground truth labels and therefore also limits the upper-bound for algorithm performance. Feedback from the clinicians has been positive, especially with regards to the algorithms segmentation of background, artifact and ring patterns showing that a dice level of 0.8 is a plausible performance level to be achieved. Nonetheless, clinicians do not make diagnostic decisions such as biopsy based solely on individual pixels. Rather they make decisions more comprehensively, based on a gestalt that includes clinical visual and dermoscopic examination, analysis of cellular and structural patterns, and supplementary clinical (non-imaging) factors including age and gender, physical location on the body, the appearance of other lesions on the same patient, patient history, genetic and other risk factors, etc. In this respect, the metric of pixel-wise accuracy cannot be directly translated into clinical utility. A further reader study, which takes into account these other factors as well as inter-reader variability, would be required to truly judge clinical utility. In this regard, the current version of MED-Net has already been integrated in two RCM devices at our institute, working closely with the manufacturer, and is soon to be tested real-time at the clinic for performance analysis. We believe that, with the availability of larger datasets that will be collected through this clinical testing, we will be able to improve the model performance on under-performing classes and be able to bring all the classes to the same performance level.

Compared to the other network models that we tested, MED-Net achieved consistently higher quantitative metrics. Among other approaches, FCN performed best and had average sensitivity, specificity, and Dice coefficient similar to MED-Net. The qualitative results provided in Fig. 2 suggest that MED-Net avoided inaccurately fragmented annotations. Note that both networks used the same loss function, which included an overfragmentation penalty. Thus we conclude that this result was achieved via the multiresolution feedback mechanism introduced in the network, which provides the output of the coarser network as a

prior estimate to the finer level (Fig. 3). In this way, the model was observed to provide more coherent segmentations compared to FCN.

Previous work has suggested that a double review of cases is preferable for remote interpretation (Witkowski et al., 2017), but this can be logistically difficult due to the limited availability of experts. Having an integrated machine learning based segmentation analysis serve as a second review may be a reasonable alternative to ensure accurate and consistent care. Moreover, beyond diagnostic analysis, MED-Net could serve as a quality assurance layer for experts; it has already been shown to provide a quantitative measure of artifacts in the collected mosaics (Kose et al., 2019) and could this guide the clinician to acquire mosaics in which diagnostic content is not obscured by artifacts.

The performance of MED-Net on RCM images depends on several domain-specific design choices that we made throughout the design of the network topology and training procedure, in order to utilize the model and the available data to their full extent. All these choices were carefully made by observing how clinician examine these images. Moreover, we increased the model's invariance against variations in the signal (e.g. speckle noise inherent in optical imaging of scattering tissue, tissue deformations under microscope pressure) by designing augmentation techniques that simulate these modality-specific variations. Our experience is that careful design of data augmentation helped to ameliorate these problems and increased both sensitivity and specificity. We believe this is a manifestation of the widely-held understanding that, even if deep learning methods provide powerful solutions to represent the data of interest and carry out classification tasks, without the proper domain-specific choices, one may not achieve good results.

Another way to potentially increase the performance would be to increase the amount of available training data. For example, as mentioned in Section 3, "deformed" variants of the meshwork pattern were misclassified by MED-Net, decreasing the segmentation performance. We believe that it is possible to overcome this problem by using more meshwork patterns that includes such deformations for training. Similar strategies could be followed to cover variations of all the patterns and increase the segmentation performance.

However, preparing data to train semantic segmentation models is logistically challenging. Unlike widely used classification models, where collecting image-wise labels are sufficient for training, data labeling for semantic segmentation is laborious and time-consuming, as it requires identifying precise and exhaustive boundaries in each test image. As mentioned earlier, unlike labeling natural scenes where the object borders are well defined, subjectivity is a common issue in labeling microscopic images. For example, even if meshwork and ring patterns are considered two different morphological patterns in their canonical form, it was not at all uncommon in our data for one of the patterns to slowly morph into the other, leading to a region with a blend of both patterns. One way to ease the experts' labeling workload, which we adopted here, was to ask experts to label only relatively clear and distinct single-pattern regions, rather than exhaustively labeling all pixels. Specifically, we asked the experts to label only the areas that they thought represented clear examples of the six given patterns. The result was that they labeled 57% of the training data pixels across the 117 mosaics. Once trained, MED-Net was able to predict labels for the entire mosaic,

although we were not able to calculate quantitative metrics on the unlabeled regions due to lack of ground truth. To allow this level of flexibility for the labelers, we designed our training procedure to be capable of handling partially labeled data by calculating and backpropagating the error over only the labeled pixels.

However, based on our experience, we believe that even this “partial labeling” scheme will not be sustainable in the long run if we want to significantly increase the size and variety of data available for further training and development. We are currently investigating ways of utilizing “weakly-labeled” data for semantic segmentation purposes. In such a scheme, the expert would provide only mosaic-wise labels (or maybe quadrant-wise, or smaller portions of the mosaics), similar to what is done for classification problems. These labels would then be extended by the network to full semantic segmentation maps. These regions could be singly or multiply labeled according to both the ML scheme and the nature of the data. Campanella et al. (2019) investigate a multiple instance learning based approach for the segmentation of histopathology slides. In histopathology, large amounts of weakly-labeled data are available through pathology slides and the respective pathology reports (e.g. Campanella et al. used 12 thousand pathology slides). RCM imaging, on the other hand, is likely to remain in the realm of small data, at least, in the foreseeable future. We hope that this work, and specifically the availability of MED-Net, will help to accelerate the adoption of RCM imaging, in turn leading to larger data availability in the coming years. With the increase in data availability, it will become possible to explore semi-supervised and weakly-supervised learning approaches, where data labeling needs will be decreased substantially. Indeed, we are currently working on a weakly supervised method that requires image-wise, rather than pixel-wise labeling (D’Alonzo (2020)).

Our study has several limitations we wish to mention here. The algorithm’s analysis capability is limited to mosaics collected at the DEJ level of the melanocytic lesions. While DEJ-level mosaics are the key components in RCM-based diagnostic analysis of melanocytic lesions, images collected in the relatively deeper dermis and also relatively superficial epidermis may also be used for a definitive diagnosis. Unfortunately, cellular morphological patterns do look different at these other layers. Therefore MED-Net will require further training with mosaics from those layers to be able to provide a more complete diagnostic analysis of the overall lesion. Similarly, melanoma is only one particular type of skin cancer among the several different skin conditions, and each has characteristic appearance features in tissue. In order to provide a more complete analysis, MED-Net models should be trained for skin conditions beyond melanoma. Finally, the sample size is not at the level of many other imaging modalities (e.g. dermoscopy) limiting current level of performance, Nonetheless, as a novel optical imaging modality that has recently entered to the wider clinical practice, the sample size is unprecedentedly big. The challenges inherent in adopting a new optical imaging technology into clinical practice makes the throughput of image collection low, resulting in slow accumulation of high quality images for studies like this. With the recent availability of reimbursement codes (Rajadhyaksha et al. (2017)), clinicians have started to collect data in a more standardized fashion and thus increase the imaging throughput, resulting in the expectation of availability of larger datasets in the future. Availability of such larger datasets will eventually help us to improve model performance as well as to expand applicability to other diseases.

Finally, MED-Net was explicitly designed as a segmentation tool that can be used for other imaging modalities. The multiscale cellular and morphological textural patterns seen in RCM images of melanocytic skin lesions have fundamental underlying similarities to patterns seen in other conditions (e.g. non-melanocytic skin lesions, skin pre-cancers, oral pre-cancers and cancers, benign and inflammatory conditions in skin (Flores et al., 2019; Peterson et al., 2019; Longo et al., 2012)) and with other emerging optical microscopic imaging approaches (optical coherence tomography (OCT), multimodal OCT-and-RCM, and optical coherence microscopy (OCM)) (Schneider et al., 2019; Boone et al., 2015). Thus we also hope that MED-Net will eventually help to drive wider applicability, acceptance and adoption of in vivo optical microscopy in clinical practice. To this end, the model and the corresponding code will be made available at https://github.com/kkose/MED-Net upon publication.

Acknowledgment

This project was supported by NIH grant R01CA199673 from NCI and in part by MSKCCs Cancer Center core support NIH grant P30CA008748 from NCI. The authors would like to thank NVIDIA Corporation for the Titan V GPU donation through their GPU Grant Program.

Alican Bozkurt is a current employee at Paige, New York This company played no role in the sponsorship, design, data collection and analysis, decision to publish, or preparation of the manuscript. Christi Alessi-Fox is a current employee and shareholder at CaliberID. Milind Rajadhyaksha is a former employee of and holds equity in CaliberID, manufacturer of a confocal microscope. Prof. Giovanni Pellacani received honoraria for courses on confocal microscopy from Mavig GmbH, and served as advisory board member for CaliberID.

References

Alarcon I, Carrera C, Palou J, Alos L, Malvehy J, Puig S, 2014 Impact of in vivo reflectance confocal microscopy on the number needed to treat melanoma in doubtful lesions. British journal of Dermatology 170, 802–808.

Amirul Islam M, Rochan M, Bruce ND, Wang Y, 2017 Gated feedback refinement network for dense image labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3751–3759.

Badrinarayanan V, Kendall A, Cipolla R, 2017 Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39, 2481–2495. [PubMed: 28060704]

Boone M, Marneffe A, Suppa M, Miyamoto M, Alarcon I, Hofmann-Wellenhof R, Malvehy J, Pellacani G, Del Marmol V, 2015 High-definition optical coherence tomography algorithm for the discrimination of actinic keratosis from normal skin and from squamous cell carcinoma. Journal of the European Academy of Dermatology and Venereology 29, 1606–1615. [PubMed: 25656269]

Borsari S, Pampena R, Lallas A, Kyrgidis A, Moscarella E, Benati E, Raucci M, Pellacani G, Zalaudek I, Argenziano G, et al., 2016 Clinical indications for use of reflectance confocal microscopy for skin cancer diagnosis. JAMA dermatology 152, 1093–1098. [PubMed: 27580185]

Bozkurt A, Kose K, Alessi-Fox C, Gill M, Dy J, Brooks D, Rajadhyaksha M, 2018 A multiresolution convolutional neural network with partial label training for annotating reflectance confocal microscopy images of skin, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer pp. 292–299.

Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ, 2019 Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine 25, 1301–1309.

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL, 2016 Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915.

Chen PHC, Gadepalli K, MacDonald R, Liu Y, Nagpal K, Kohlberger T, Dean J, Corrado GS, Hipp JD, Stumpe MC, 2018 Microscope 2.0: An augmented reality microscope with real-time artificial intelligence integration. arXiv preprint arXiv:1812.00825.

CIBC, 2016 Seg3D: Volumetric Image Segmentation and Visualization. Scientific Computing and Imaging Institute (SCI), Download from: http://www.seg3d.org.

Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, Kalloo A, Liopyris K, Mishra N, Kittler H, et al., 2018 Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC), in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE pp. 168–172.

D'Alonzo M, 2020 Semantic Segmentation of Reflectance Confocal Microscopy Mosaics of Pigmented Lesions using Weak Labels. Master's thesis Northeastern University Boston, MA, USA.

Dice LR, 1945 Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Eigen D, Fergus R, 2015 Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Proceedings of the IEEE international conference on computer vision, pp. 2650–2658.

Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, Böhm A, Deubner J, Jäckel Z, Seiwald K, et al., 2019 U-net: deep learning for cell counting, detection, and morphometry. Nature methods 16, 67. [PubMed: 30559429]

Flores E, Yélamos O, Cordova M, Kose K, Phillips W, Lee E, Rossi A, Nehal K, Rajadhyaksha M, 2019 Peri-operative delineation of nonmelanoma skin cancer margins in vivo with handheld reflectance confocal microscopy and video-mosaicking. Journal of the European Academy of Dermatology and Venereology 33, 1084–1091. [PubMed: 30811707]

Fu H, Cheng J, Xu Y, Wong DWK, Liu J, Cao X, 2018 Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE transactions on medical imaging 37, 1597–1605. [PubMed: 29969410]

Fu H, Xu Y, Wong DWK, Liu J, 2016 Retinal vessel segmentation via deep learning network and fully-connected conditional random fields, in: 2016 IEEE 13th international symposium on biomedical imaging (ISBI), IEEE pp. 698–701.

Gill M, Alessi-Fox C, Kose K, 2019 Artifacts and landmarks: pearls and pitfalls for in vivo reflectance confocal microscopy of the skin using the tissue-coupled device. Dermatology online journal 25.

Goodman B, Flaxman S, 2017 European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine 38, 50–57.

Gu F, Burlutskiy N, Andersson M, Wilén LK, 2018 Multi-resolution networks for semantic segmentation in whole slide images, in: Computational Pathology and Ophthalmic Medical Image Analysis Springer, pp. 11–18.

Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, Zhang T, Gao S, Liu J, 2019 Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging.

Guy GP Jr, Machlin SR, Ekwueme DU, Yabroff KR, 2015 Prevalence and costs of skin cancer treatment in the us, 2002-2006 and 2007-2011. American journal of preventive medicine 48, 183–187. [PubMed: 25442229]

He K, Zhang X, Ren S, Sun J, 2016 Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Jiang J, Hu YC, Liu CJ, Halpenny D, Hellmann MD, Deasy JO, Mageras G, Veeraraghavan H, 2018 Multiple resolution residually connected feature streams for automatic lung tumor segmentation from ct images. IEEE transactions on medical imaging 38, 134–144.

Kose K, Bozkurt A, Alessi-Fox C, Brooks DH, Dy JG, Rajadhyaksha M, Gill M, 2019 Utilizing machine learning for image quality assessment for reflectance confocal microscopy. Journal of Investigative Dermatology doi:10.1016/j.jid.2019.10.018.

Li J, Sarma KV, Ho KC, Gertych A, Knudsen BS, Arnold CW, 2017 A multi-scale u-net for semantic segmentation of histological images from radical prostatectomies, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association p. 1140.

Lin G, Milan A, Shen C, Reid I, 2017 Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1925–1934.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI, 2017 A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88. [PubMed: 28778026]

Long J, Shelhamer E, Darrell T, 2015 Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Longo C, Zalaudek I, Argenziano G, Pellacani G, 2012 New directions in dermatopathology: in vivo confocal microscopy in clinical practice. Dermatologic clinics 30, 799–814. [PubMed: 23021059]

Marchetti MA, Codella NC, Dusza SW, Gutman DA, Helba B, Kalloo A, Mishra N, Carrera C, Celebi ME, DeFazio JL, et al., 2018 Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. Journal of the American Academy of Dermatology 78, 270–277. [PubMed: 28969863]

Milletari F, Navab N, Ahmadi S, 2016 V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571.

Mirikharaji Z, Izadi S, Kawahara J, Hamarneh G, 2018 Deep auto-context fully convolutional neural network for skin lesion segmentation, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 877–880.

Mohseni Salehi SS, Erdogmus D, Gholipour A, 2017 Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. IEEE Transactions on Medical Imaging 36, 2319–2330.

Nie D, Wang L, Gao Y, Shen D, 2016 Fully convolutional networks for multi-modality isointense infant brain image segmentation, in: 2016 IEEE 13Th international symposium on biomedical imaging (ISBI), IEEE. pp. 1342–1345.

Nikolaou V, Stratigos A, 2014 Emerging trends in the epidemiology of melanoma. British journal of dermatology 170, 11–19.

Pellacani G, Pepe P, Casari A, Longo C, 2014 Reflectance confocal microscopy as a second-level examination in skin oncology improves diagnostic accuracy and saves unnecessary excisions: a longitudinal prospective study. British Journal of Dermatology 171, 1044–1051.

Pellacani G, Witkowski A, Cesinaro A, Losi A, Colombo G, Campagna A, Longo C, Piana S, De Carvalho N, Giusti F, et al., 2016 Cost–benefit of reflectance confocal microscopy in the diagnostic performance of melanoma. Journal of the European Academy of Dermatology and Venereology 30, 413–419. [PubMed: 26446299]

Peterson G, Zanoni DK, Ardigo M, Migliacci JC, Patel SG, Rajadhyaksha M, 2019 Feasibility of a video-mosaicking approach to extend the field-of-view for reflectance confocal microscopy in the oral cavity in vivo. Lasers in Surgery and Medicine 51, 439–451.

Rajadhyaksha M, Marghoob A, Rossi A, Halpern AC, Nehal KS,2017 Reflectance confocal microscopy of skin in vivo: From bench to bedside. Lasers in surgery and medicine 49, 7–19. [PubMed: 27785781]

Ronneberger O, Fischer P, Brox T, 2015 U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer pp. 234–241.

Roy AG, Navab N, Wachinger C, 2018 Recalibrating fully convolutional networks with spatial and channel squeeze and excitation blocks. IEEE transactions on medical imaging 38, 540–549.

Salehi SSM, Erdogmus D, Gholipour A, 2017 Tversky loss function for image segmentation using 3d fully convolutional deep networks, in: International Workshop on Machine Learning in Medical Imaging, Springer pp. 379–387.

Schneider SL, Kohli I, Hamzavi IH, Council ML, Rossi AM, Ozog DM, 2019 Emerging imaging technologies in dermatology: Part ii: Applications and limitations. Journal of the American Academy of Dermatology 80, 1121–1131. [PubMed: 30528310]

Scope A, Guitera P, Pellacani G, 2017 Rcm diagnosis of melanocytic neoplasms: Terminology, algorithms and their accuracy and clinical integration., in: González S, Rajadhyaksha M, Ardigo M, Longo C, Carrera C, Ulrich M, Moscarella E (Eds.), Reflectance Confocal Microscopy of Cutaneous Tumors, 2nd Ed. Boca Raton, CRC Press, pp. 168–186.

Tu Z, Bai X, 2009 Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE transactions on pattern analysis and machine intelligence 32, 1744–1757.

Witkowski A, Łudzik J, Arginelli F, Bassoli S, Benati E, Casari A, De Carvalho N, De Pace B, Farnetani F, Losi A, et al., 2017 Improving diagnostic sensitivity of combined dermoscopy and reflectance confocal microscopy imaging through double reader concordance evaluation in telemedicine settings: A retrospective study of 1000 equivocal cases. PloS one 12, e0187748. [PubMed: 29121636]

Yu L, Yang X, Chen H, Qin J, Heng PA, 2017 Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images, in: AAAI, pp. 66–72.

Zhang S, Fu H, Yan Y, Zhang Y, Wu Q, Yang M, Tan M, Xu Y, 2019 Attention guided network for retinal image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer pp. 797–805.

Zhao H, Shi J, Qi X, Wang X, Jia J, 2017 Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890.

Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J, 2018 Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support Springer, pp. 3–11.

Zhu Q, Du B, Turkbey B, Choyke PL, Yan P, 2017 Deeply-supervised cnn for prostate segmentation, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE pp. 178–184.

Highlights

- In-vivo imaging based non-invasive diagnosis is advancing into clinical practice
- MED-Net mimicks clinicians' examination of in vivo RCM images by multiscale analysis
- Trained and tested on 117 RCM mosaics collected at 6 clinics in 2 countries
- An efficient training strategy via multi-scale deep supervision
- Multi-scale analysis capability of MED-Net increase in vivo segmentation performance

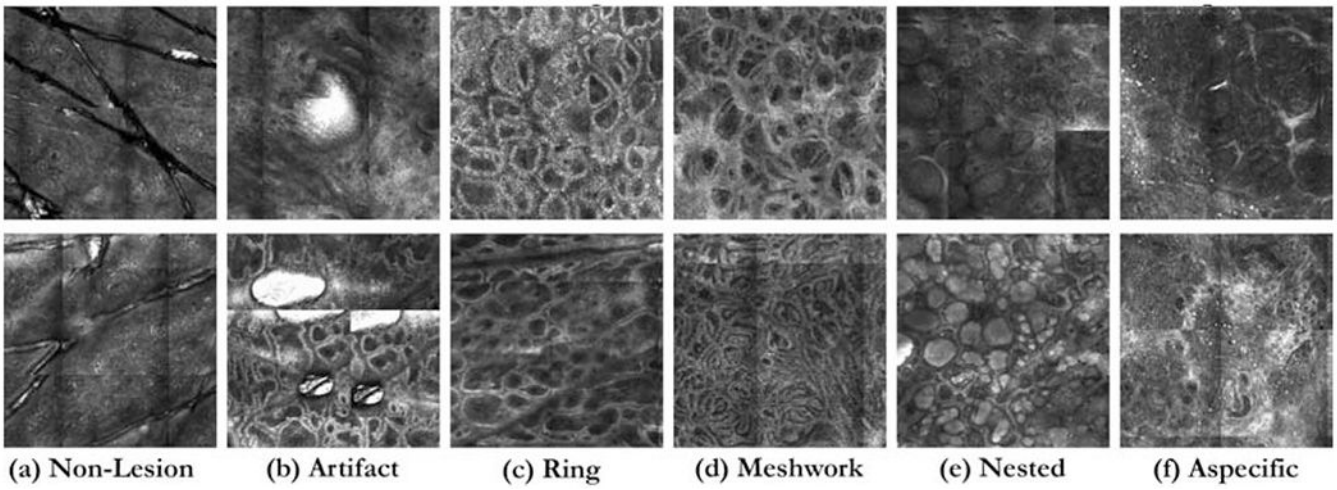


Fig. 1.

Two examples for each of the six distinct patterns (four cellular morphological and two other patterns). as seen in reflectance confocal mosaics at the dermal-epidermal junction in melanocytic skin lesions.

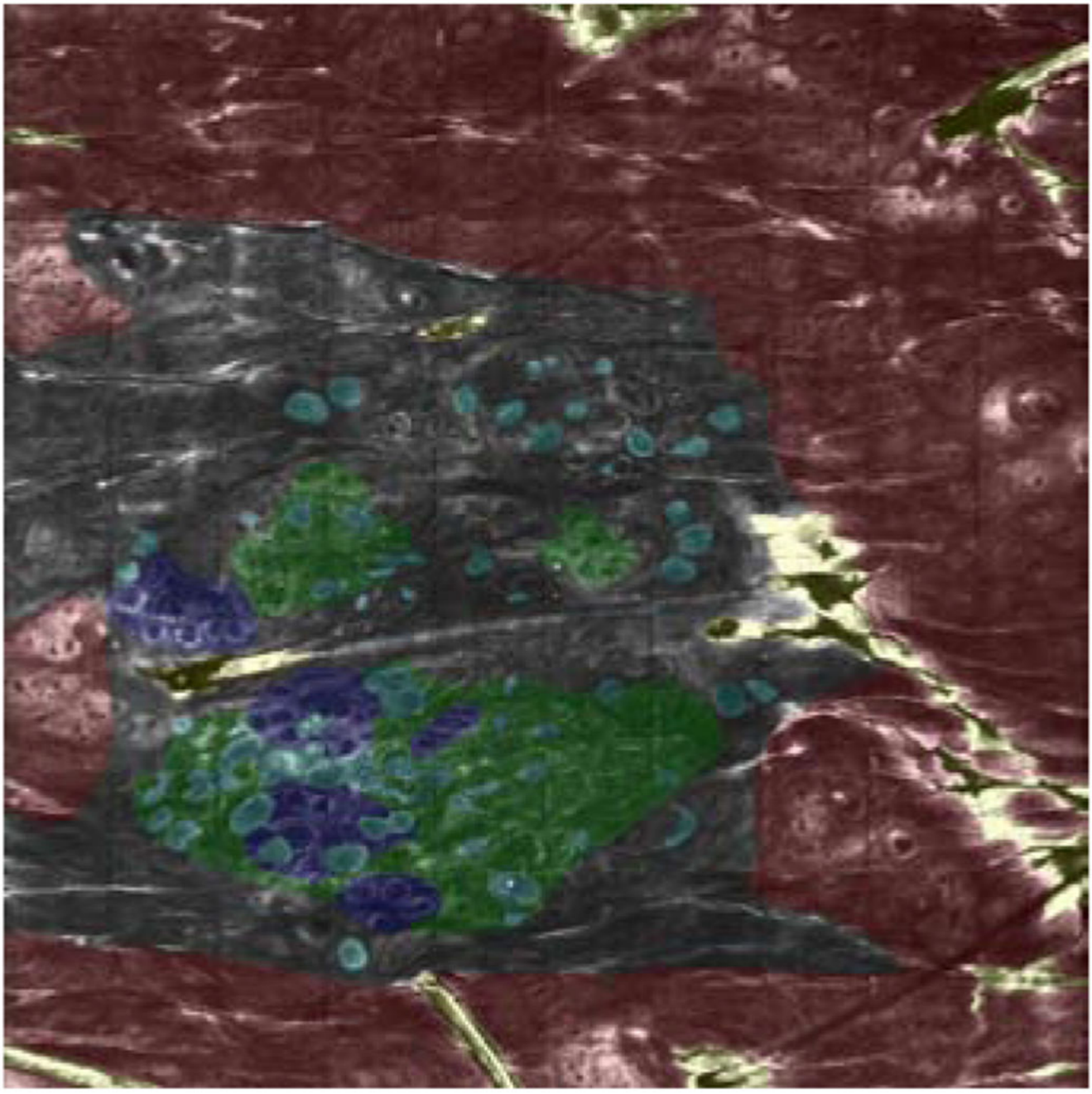


Fig. 2.

An example mosaic and its corresponding expert labeling. Colors indicate the labels; Red: Non-Lesion, Yellow: Artifact, Green: Meshwork, Blue: Ring, Cyan: Nested. Grey colored areas are not labeled, and are ignored in training and quantitative evaluation.

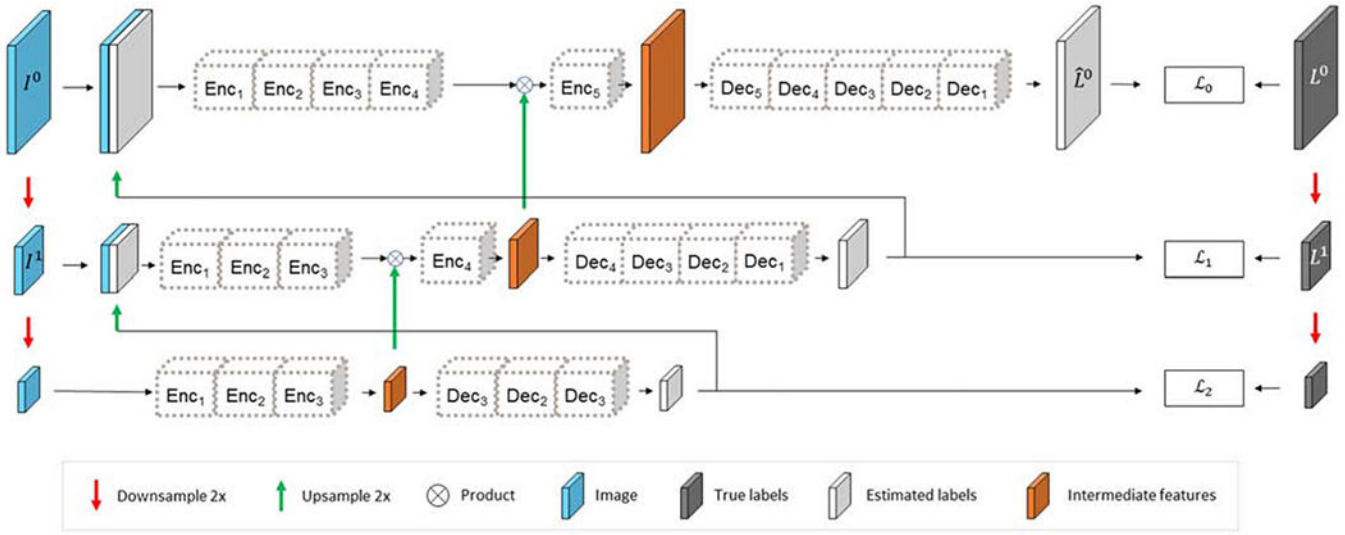


Fig. 3.

Our architecture is composed of 3 nested fully convolutional networks that generate semantic segmentation at different scales. Red arrows denote 2x downsampling, and green arrows denote 2x upsampling. Output segmentations at lower magnifications are fed into the next level via concatenation. The loss at each level (scale) is calculated and backpropagated for deep supervision of the subnetworks.

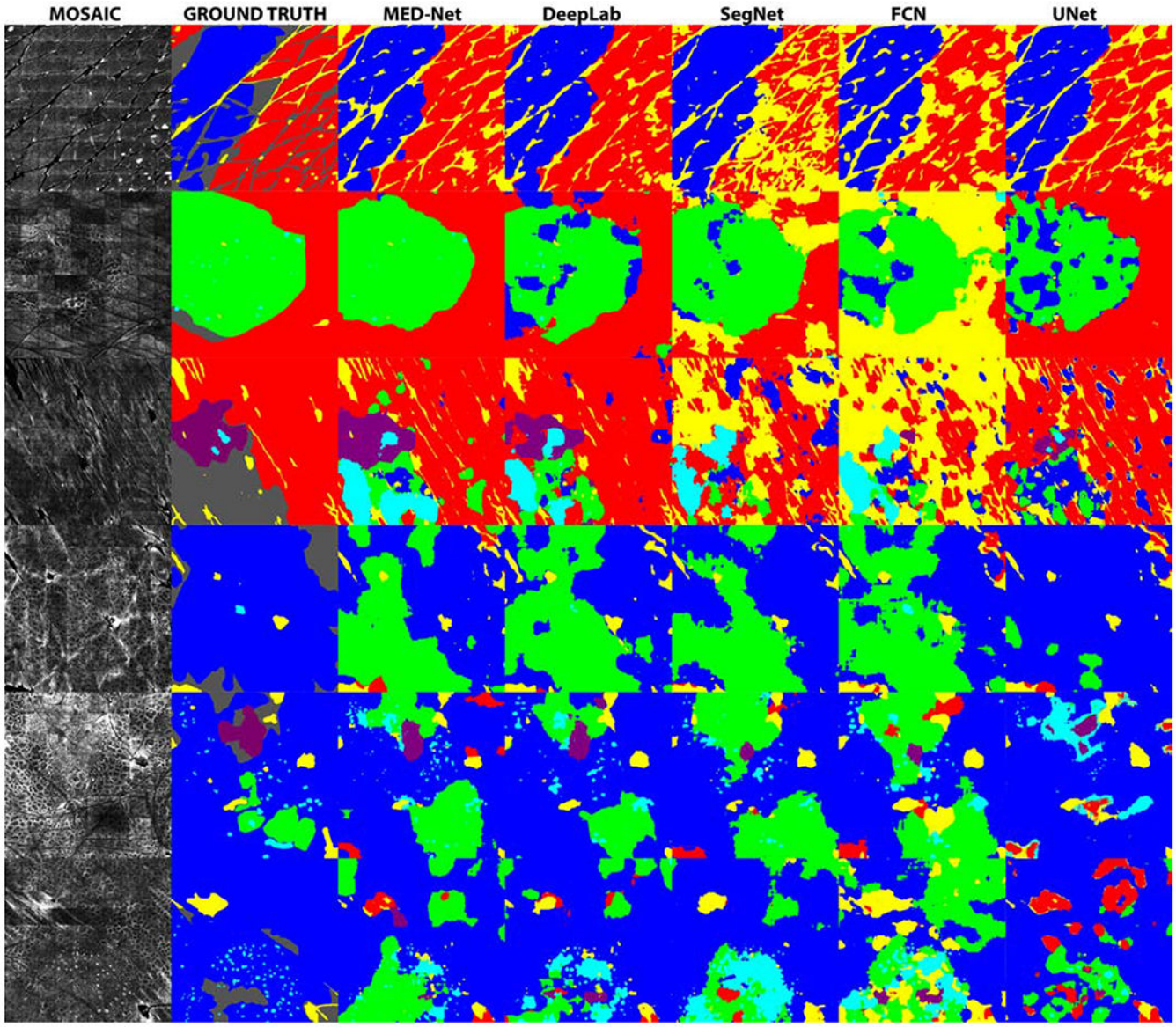


Fig. 4.

Example segmentation results of 6 mosaics for Scenario 1. Color scheme is the same as used in Fig. 2. The ground truth segmentations are compared to the outputs of MED-Net and other state-of-the-art-methods. Images are not exhaustively annotated by the readers. Pixels that are not annotated (dark grey label) are ignored during training. During the testing phase, these pixels are discarded from sensitivity and specificity calculations.

Table 1.

Class distribution statistics: The top portion reports the distribution of labels for both scenarios. In Scenario 1, we were able to balance distribution across training and test sets to within 1% (stratified cross-validation). Class distributions in training and test sets are explicitly given for Scenario 2. In the bottom portion, we report on the size of the datasets in terms of both images and labeled pixels, as well as on the overall fraction of pixels that were labeled.

	Scenario 1	Scenario 2	
	Whole Dataset	**Europe (Train)**	**US (Test)**
Background	17% (83.7M)	19% (65.3M)	13% (18.4M)
Artifact	19% (92.6M)	20% (67.7M)	18% (24.5M)
Mesh	20% (97.5M)	21% (73.7M)	17% (23.8M)
Nest	5% (25.6M)	6% (19.0M)	5% (6.6M)
Ring	28% (136.6M)	23% (78.8M)	40% (57.8M)
Aspecific	10% (50.5M)	12% (40.3M)	7% (10.1M)
labeled Pixels	57% (486.6M)	51% (344.7M)	60% (141.8M)
# labeled Images	117	86	31

Table 2.

Results for Scenario 1 Patient-Wise. The best results for each metric and label are highlighted in bold.

	Sensitivity				
	MED-Net	**DeepLab**	**SegNet**	**FCN**	**UNet**
Background	0.83 ± 0.05	0.91 ± 0.05	**0.94 ± 0.02**	0.86 ± 0.04	0.85 ± 0.11
Artifact	**0.83 ± 0.10**	0.58 ± 0.18	0.57 ± 0.18	0.81 ± 0.08	0.78 ± 0.11
Meshwork	**0.59 ± 0.09**	0.47 ± 0.10	0.45 ± 0.14	**0.59 ± 0.08**	0.22 ± 0.15
Nest	0.59 ± 0.13	0.49 ± 0.22	0.53 ± 0.16	**0.60 ± 0.14**	0.34 ± 0.20
Ring	0.85 ± 0.07	0.82 ± 0.10	0.74 ± 0.12	0.82 ± 0.14	**0.87 ± 0.08**
Aspecific	**0.53 ± 0.22**	0.21 ± 0.32	0.38 ± 0.031	0.40 ± 0.28	0.48 ± 0.28
Average	**0.74**	0.64	0.63	0.72	0.65
	Specificity				
	MED-Net	**DeepLab**	**SegNet**	**FCN**	**UNet**
Background	**0.99 ± 0.01**	0.84 ± 0.09	0.81 ± 0.14	0.97 ± 0.01	0.96 ± 0.01
Artifact	0.92 ± 0.02	**0.96 ± 0.02**	0.94 ± 0.03	0.94 ± 0.01	0.89 ± 0.04
Meshwork	0.91 ± 0.02	0.92 ± 0.04	0.92 ± 0.06	0.89 ± 0.05	**0.96 ± 0.03**
Nest	**0.99 ± 0.01**	**0.99 ± 0.01**	**0.99 ± 0.00**	**0.99 ± 0.01**	**0.99 ± 0.02**
Ring	0.88 ± 0.06	0.087 ± 0.04	**0.92 ± 0.02**	0.86 ± 0.05	0.81 ± 0.04
Aspecific	0.98 ± 0.01	**0.99 ± 0.00**	0.97 ± 0.02	**0.99 ± 0.01**	0.95 ± 0.05
Average	**0.92**	0.90	0.91	0.91	0.90
	Dice Coefficient				
	MED-Net	**DeepLab**	**SegNet**	**FCN**	**UNet**
Background	**0.87 ± 0.03**	0.65 ± 0.17	0.65 ± 0.20	0.86 ± 0.02	0.83 ± 0.06
Artifact	**0.78 ± 0.04**	0.65 ± 0.13	0.62 ± 0.10	**0.78 ± 0.04**	0.70 ± 0.03
Meshwork	**0.58 ± 0.10**	0.53 ± 0.13	0.48 ± 0.11	0.56 ± 0.12	0.30 ± 0.18
Nest	0.66 ± 0.11	0.55 ± 0.16	0.61 ± 0.13	**0.67 ± 0.11**	0.43 ± 0.11
Ring	**0.82 ± 0.08**	0.79 ± 0.09	0.78 ± 0.10	0.79 ± 0.11	0.78 ± 0.07
Aspecific	**0.60 ± 0.19**	0.25 ± 0.34	0.39 ± 0.23	0.48 ± 0.26	0.42 ± 0.19
Average	**0.74**	0.61	0.61	0.71	0.61

Table 3.

Results for Scenario 2 Clinic-Wise Cross-Validation Experiments. The best results for each metric and label are highlighted in bold.

	Sensitivity				
	MED-Net	**DeepLab**	**SegNet**	**FCN**	**UNet**
Background	0.89	0.94	**0.95**	0.90	0.77
Artifact	0.81	0.69	0.67	0.78	**0.92**
Meshwork	**0.67**	0.57	**0.67**	0.66	0.17
Nest	**0.50**	0.27	0.19	0.37	0.23
Ring	**0.79**	0.75	0.74	0.71	0.70
Aspecific	0.77	0.72	0.47	0.65	**0.87**
Average	**0.77**	0.71	0.69	0.72	0.68
	Specificity				
	MED-Net	**DeepLab**	**SegNet**	**FCN**	**UNet**
Background	**0.99**	0.93	0.90	0.95	0.91
Artifact	0.92	0.93	0.95	0.91	**0.98**
Meshwork	0.91	0.90	0.85	0.89	**0.95**
Nest	0.99	**1.00**	**1.00**	**1.00**	**1.00**
Ring	**0.94**	0.91	0.94	0.93	**0.94**
Aspecific	0.95	0.95	**0.98**	**0.98**	0.79
Average	**0.94**	0.87	0.88	0.88	0.89
	Dice Coefficient				
	MED-Net	**DeepLab**	**SegNet**	**FCN**	**UNet**
Background	**0.92**	0.85	0.80	0.86	**0.92**
Artifact	**0.76**	0.71	0.72	0.73	0.72
Meshwork	**0.67**	0.59	0.60	0.63	0.25
Nest	**0.62**	0.41	0.32	0.51	0.37
Ring	**0.79**	0.73	0.75	0.73	0.73
Aspecific	**0.72**	0.70	0.57	0.71	0.50
Average	**0.77**	0.70	0.69	0.72	0.64

Table 4.

Ablation study results for training versions of MED-Net with 3 different levels.

	Dice Coefficient						
	Background	**Artifact**	**Meshwork**	**Nest**	**Ring**	**Aspecific**	***Average***
1 Level	0.90	0.69	0.62	0.48	0.71	0.63	*0.67*
2 Level	0.90	0.70	0.65	0.52	0.77	0.63	*0.69*
3 Level	**0.92**	**0.76**	**0.67**	**0.62**	**0.79**	**0.72**	***0.75***
	Sensitivity						
	Background	**Artifact**	**Meshwork**	**Nest**	**Ring**	**Aspecific**	***Average***
1 Level	0.87	0.66	**0.72**	0.34	0.67	0.69	*0.66*
2 Level	0.85	0.68	0.71	0.39	0.74	0.72	*0.68*
3 Level	**0.89**	**0.81**	0.67	**0.50**	**0.79**	**0.77**	***0.74***
	Specificity						
	Background	**Artifact**	**Meshwork**	**Nest**	**Ring**	**Aspecific**	***Average***
1 Level	**0.99**	**0.94**	0.84	**1.00**	**0.94**	0.93	*0.94*
2 Level	**0.99**	0.93	0.87	0.99	**0.94**	0.92	*0.94*
3 Level	**0.99**	0.92	**0.91**	0.99	**0.94**	**0.95**	***0.95***

Table 5.

Ablation study results for training MED-Net with different loss functions.

	Dice Coefficient		
	Cross Entropy	**Dice Loss**	**MSDC+TV**
Background	0.91	**0.92**	**0.92**
Artifact	0.78	**0.78**	0.76
Meshwork	0.57	0.64	**0.67**
Nest	0.56	0.57	**0.62**
Ring	0.76	0.75	**0.79**
Aspecific	0.68	0.71	**0.72**
Average	*0.74*	*0.75*	***0.77***
	Sensitivity		
	Cross Entropy	**Dice Loss**	**MSDC+TV**
Background	0.89	**0.93**	0.89
Artifact	0.78	0.74	**0.81**
Meshwork	0.50	**0.71**	0.67
Nest	**0.51**	0.44	0.50
Ring	0.76	0.71	**0.79**
Aspecific	**0.88**	0.76	0.77
Average	*0.73*	*0.73*	***0.77***
	Specificity		
	Cross Entropy	**Dice Loss**	**MSDC+TV**
Background	0.99	0.98	**0.99**
Artifact	0.95	**0.96**	0.92
Meshwork	**0.93**	0.86	0.91
Nest	0.98	0.99	**0.99**
Ring	0.93	**0.95**	0.94
Aspecific	0.90	**0.95**	**0.95**
Average	***0.94***	***0.94***	***0.94***