WSNet: Towards An Effective Method for Wound Image Segmentation

Subba Reddy Oota[1] , Vijay Rowtula[1,2], Shahid Mohammed[1], Minghsun Liu[1], Manish Gupta[2*]
[1]Woundtech Innovative Healthcare Solutions, [2]IIIT-Hyderabad, India
{soota, vrowtula, shmohammed, mliu}@woundtech.net, manish.gupta@iiit.ac.in

Abstract

Medical image segmentation is critical for effective computer-aided diagnosis and localization of ailments. Automated segmentation of wound regions from patient images can aid clinicians in measuring and managing chronic wounds and monitoring the wound healing trajectory. While there exists a plethora of work on general medical image segmentation, there is hardly any work on wound image analysis and segmentation. Existing methods are limited to segmenting a smaller subset of ulcers, such as foot ulcers, with no special processing for wound images. In this paper, we build segmentation models for eight different types of wound images. Wound image analysis is a challenging problem due to the lack of availability of extensive data (labeled or unlabeled), and annotation is also challenging due to the shortage of well-trained wound care clinicians. To handle these challenges, we contribute WOUNDSEG[1]*, a large and diverse dataset of segmented wound images. Generic wound image segmentation is complex due to the heterogeneous appearance of wound area across images of similar wound types. We propose a novel image segmentation framework,* WSNET*, which leverages (a) wound-domain adaptive pretraining on a large unlabeled wound image collection and (b) a global-local architecture that utilizes full image and its patches to learn fine-grained details of heterogeneous wounds. On* WOUNDSEG*, we achieve a decent Dice score of 0.847. On existing AZH Woundcare and Medetec datasets, we establish a new state-of-the-art. Further, we show the impact of using segmentation for improving the accuracy of downstream tasks like wound area and volume prediction.*

1. Introduction

The magnitude of wound care costs is on a rise in the health care industry. In 2018, around 8 million people suffered from wounds, with the medicare cost estimates ranging from $28.1 billion to $96.8 billion [28]. Time is crucial since the longer a wound is allowed to fester, the harder it becomes to treat, and a quick response is the best way to reduce wound care costs. Wound image segmentation is critical for quick marking the region of interest (the wound area), which can play a significant role in effective computer-aided diagnosis and localization of ailments. Accurate wound image segmentation can be helpful in automatically identifying wound attributes like wound area, wound volume, and wound stage. Such attributes can, in turn, help evaluate and manage chronic wounds, monitor the wound healing trajectory, determine future interventions, predict the patient's hospitalization risk, or predict the time to heal [25], significantly reducing hospital admissions and amputations.

Despite this critical need for automated wound image segmentation methods, previous work has primarily focused on broader medical image segmentation [6, 20] with relatively sparse literature on *wound image segmentation* [33, 35]. Broadly, medical image segmentation aims to make anatomical or pathological structure changes more clear in images. Popular medical image segmentation tasks include liver and liver-tumor segmentation [18], brain and brain-tumor segmentation [23], optic disc segmentation [2], cell segmentation [27], lung segmentation, and pulmonary nodules [37]. Lei et al. [17] and Hesamian et al. [9] provide a good overview of popular methods in the area. In this work, we focus on the segmentation of *wound images*.

Our goal is to build a model highlighting the wound region from a given image. Developing such a system can help clinicians reduce the time spent to diagnose the wound by automatically measuring its area and volume, reducing healing time and wound care costs. The problem is challenging because wound segmentation can depend on multiple factors like wound/ulcer type, wound location, wound size, variations across images in terms of lighting conditions, and skin color variations. The lack of large amounts of publicly available labeled (or even unlabeled) data is yet another challenge. Existing datasets like AZH Woundcare [35] and Medetec [33] are small or contain only foot ulcer images.

[*] The author is also a Principal Applied Scientist at Microsoft.

[1] Link to code and data: https://github.com/subbareddy248/WSNET

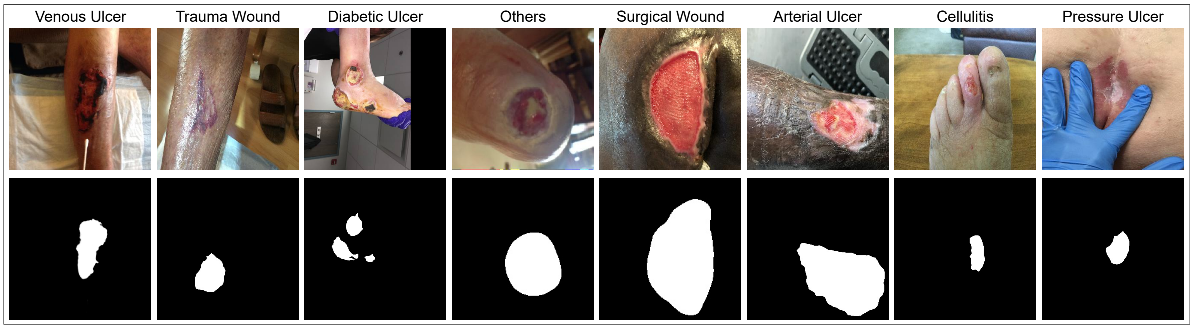

Figure 1. Examples of eight different wound types from our WoundSeg dataset. The top row shows the actual wound images, while the bottom row shows segmentation masks.

Wound Type	Diabetic	Pressure	Trauma	Venous	Surgical	Arterial	Cellulitis	Other	All
# Images	441	636	368	690	268	99	113	71	2686
Wound Area (%)	9.0	12.5	10.7	13.3	12.5	8.7	13.0	18.9	12.3

Table 1. Basic statistics of WoundSeg Dataset.

Dataset	Wound Types	Size
WoundSeg (ours)	8 types (diabetic, pressure, trauma, venous, surgical, arterial, cellulitis, and others)	2686
AZH	1 type (foot ulcer)	1109
Medetec	1 type (foot ulcer)	607

Table 2. Proposed WoundSeg dataset is more diverse and large compared to existing datasets.

Our first goal was to create a wound image segmentation dataset with good diversity. Three qualified wound specialists provided annotations for 3000 images from Jun 2015–Mar 2019 using Django-labeller[2] with an intra-annotator agreement of ∼0.95. As part of the annotation, we requested experts to include the peri-wound skin area, i.e., the tissue surrounding a wound. After two sanity checks, we obtained 2686 masked images in size 384x384. Our dataset is highly diverse, with images for eight types of wound ulcers (diabetic, pressure, trauma, venous, surgical, arterial, cellulitis, and others). We provide statistics of our WoundSeg dataset in Table 1. Fig. 1 shows a few samples of wound images for each wound type from our dataset, along with their segmentation masks. We also capture the comparison of WoundSeg with existing datasets in detail in Table 2. We make our WoundSeg dataset available publicly[1].

To build accurate wound image segmentation models, we start by performing extensive experiments with four popular image segmentation architectures and 17 different backbones. To improve upon these baseline models, we propose a novel framework called WSNet. The framework is based on convolutional neural networks (CNNs).

WSNet has the following main features:

- Usage of popular segmentation models with well-known CNN backbones,
- specialized wound image specific pretraining on a large unlabeled set,
- data augmentation strategies, and
- a novel global-local architecture to leverage multi-level, heterogeneous information effectively.

Overall, in this paper, we make the following main contributions.

- To the best of our knowledge, WoundSeg is the largest wound segmentation dataset covering eight types of wound ulcers (diabetic, pressure, trauma, venous, surgical, arterial, cellulitis, and others) unlike previous work, which focused on only diabetic foot ulcer images.
- We experiment with four different image segmentation architectures and 17 different backbones and a multi-level heterogeneous ensemble model. We improve upon baseline models via wound-domain adaptive pretraining, dataset augmentation, and a novel global-local segmentation architecture.
- Extensive experiments show that methods in our proposed framework, WSNet, outperform baselines on existing AZH Woundcare and Medetec datasets, as well on our contributed WoundSeg dataset. Our highly accurate wound image segmentation model also helps us achieve state-of-the-art results in wound area and volume prediction downstream tasks.

2. Related Work

In this section, we discuss related work in the areas of wound image analysis and wound image segmentation.

[2] https://github.com/Britefury/django-labeller

2.1. Wound Image Analysis

Chronic wound diagnosis, monitoring, and measuring the area and volume of a wound is an ongoing research area in medical image analysis. While several methods have been proposed in the literature to classify the wound tissue or segmentation of related skin lesions, these experiments fail to provide a robust tool for process automation [36, 39, 40]. Classical image processing techniques such as color descriptors and texture detectors have been used to extract features from wound images, classify the skin patches as normal or abnormal and to monitor the healing process automatically in [15, 16, 22, 26, 34, 38]. Recently,Oota et al. [24] and [25] leverage wound attributes like area and volume for tasks like re-admission risk prediction and hospitalization probability, respectively. While Oota et al. [24] obtain wound area and volume manually from clinicians, [25] predict them directly from wound images. We leverage wound segmentation models to obtain a better estimate of wound area and volume prediction. In this work, we rigorously study the wound segmentation problem by contributing a large dataset with eight wound types and proposed a novel deep learning-based segmentation method.

2.2. Wound Image Segmentation

Previous work on wound image segmentation was based on feature engineering-based traditional machine learning and image processing methods. Song et al. [30] described 49 features extracted from a wound image using K-means clustering, edge detection, thresholding, and region growing in both grayscale and RGB. Other such methods include (1) generating a Red-Yellow-Black-White probability map followed by optimal thresholding or region growing [4], (2) flood fill on contour obtained using energy minimizing discrete dynamic contour algorithm [10], (3) K-means clustering on the Independent Component Analysis (ICA) output of pre-processed RGB images [7, 41]. Such feature engineering methods require manual effort in designing features, cannot handle heterogeneity in images, and lead to relatively lower accuracies compared to recently proposed deep learning methods. Additionally, the performance is evaluated on a small biased dataset.

Several researchers have used deep learning models for segmenting and classifying various ulcer images in [3, 5, 14]. However, across all such works, the corpus size is minimal (e.g., Medetec wound images dataset [33] has only 607 images), leading to relatively brittle systems. Liu et al. [21] proposed a fully convolutional neural networks (FCN) architecture for the task on a semi-automatically annotated dataset. Recently, Wang et al. [35] proposed a new wound image dataset called AZH Woundcare consisting of 1109 foot ulcer images. They propose a convolutional framework based on MobileNetV2 to segment wound regions from natural images. However, such methods work with the entire image and fail to capture subtle local signals. We compare these approaches and show improved results in Section 5. In [25], the authors proposed an automated wound image heal classification to predict the risk of hospitalization of the wound along with wound area and volume. However, the efficacy of area and volume prediction is comparatively low. In this paper, we improve upon their results, leading to the new state-of-the-art for wound area and volume prediction tasks.

3. WSNET Methodology

This section discusses our experiments with the large set of image segmentation models with CNN backbones. Further, we discuss wound data-specific pre-training, our data augmentation strategy, and our global-local segmentation architecture.

3.1. Wound Segmentation Models

First, we experiment with the following four popular segmentation architectures: U-Net [27], LinkNet [1], PSPNet [43], and FPN [19]. Further, we experiment with 17 backbones to explore the accuracy versus model size trade-off for each segmentation model. These backbones belong to eight popular architecture groups: VGGNet [29], ResNet [8], ResNeXt [42], Inception [31], DenseNet [13], SEResNet [12], EfficientNet [32], and MobileNet [11]. For each of these, we use ImageNet pretrained weights. Out of the 17 backbones, we select three backbones (DenseNet121, DenseNet169, and MobileNet) for further experiments. Results are reported in Table 4. We chose the two DenseNet backbones since they perform the best on average across all four architectures. MobileNet is chosen since it is the smallest of backbones and is commonly used across multiple baseline systems.

3.2. Wound-Domain Adaptive Pretraining (WDAP)

The results of the 17 backbones with four segmentation models are reported in Table 4. To further improve the performance of wound image segmentation models, we create pretrained models specifically on the wound image dataset instead of using Imagenet pretrained weights.

We select three backbones (DenseNet121, DenseNet169, and MobileNet) for such special pretraining. To create pretrained models on the wound image dataset, we classify wound types into five different ulcer types: diabetic, pressure, surgical, trauma, and venous, using the three backbones. The input to each model is a wound image, and the five wound-type classes are given as our target output. We use Adam optimizer with a learning rate of 0.001. We use the categorical-cross-entropy loss with a batch size of 32, and the model is trained for 20 epochs over a dataset of

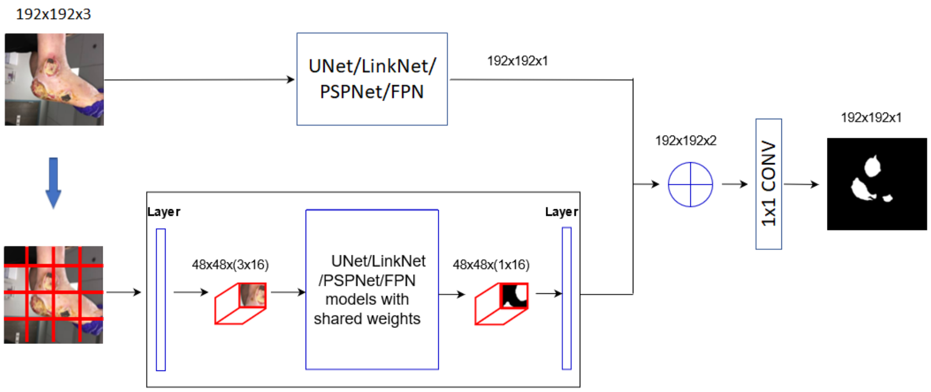

Figure 2. Architecture of our two-stream global-local image segmentation method.

125711 images from [25]. The class distribution is as follows: diabetic ulcer (19773), pressure ulcer (47541), surgical wound (12238), trauma wound (13667), venous ulcer (32492). We use Adam optimizer, a learning rate of 0.001, batch size of 32, 20 epochs. To overcome the over-fitting problem, we checked for early stopping if the validation loss did not decrease for five consecutive epochs.

3.3. Fine-tuning and Data Augmentation

Pretrained models are fine-tuned on labeled image segmentation data. Note that this data is disjoint from the data used for pretraining. While our pretraining is done on data labeled for wound type classification, our fine-tuning is done on separate data labeled for wound segmentation task. The encoder model weights corresponding to each backbone are frozen, and the decoder weights are fine-tuned over the wound image dataset for the four segmentation models.

We validated several transformation methods and finally chose horizontal flip, random rotation, optical distortion, grid distortion, blur, random brightness contrast, and transpose to perform the data augmentation. We applied the same augmentation to both image and its corresponding mask. Augmentation transformations are performed on the training set only and not on the test set.

3.4. Global-Local Architecture

Wound images have less semantically distinctive information. They usually contain just skin and wound; hence, more granular analysis is necessary to segment the wound from normal skin accurately. Thus, for effective segmentation, it is essential to obtain (global) signals from the entire image and (local) signals from individual patches extracted to capture the intricate details in wound images. Local context information is essential for semantic segmentation of wound images that contain different patches with similar features. Convolution operations on local patches process one local neighborhood at a time, which may cause incomplete segmentation of large wound patches. As shown in Fig. 2, the Global-Local architecture can help us extract high-level multi-scale semantic information. It consists of two sub-networks – the top sub-network is the global segmentation network to process the complete image, and the bottom one, i.e., the local segmentation network, processes a patch of the image.

We feed a 192×192×3 resized image as input. The global sub-network uses one of these four popular segmentation architectures: U-Net [27], LinkNet [1], PSPNet [43], and FPN [19]. For the local sub-network, the image is split into 16 different non-overlapping 48×48×3 patches, which are stacked to obtain a 48×48×(3×16) volume. We experimented with different patch sizes; 48×48×3 gave the best results. The patches are dispatched in parallel to 16 local models with shared weights for better throughput and efficiency, as seen in Fig. 2. Another layer is used at the end of the local network to stitch back the patches into a full-size mask image (192×192×1). Outputs from global and local networks are stacked to get a 192×192×2 output. A 1×1 convolution is then applied to get a 192×192×1 predicted mask. The predicted mask is compared with the ground-truth mask, and the loss is back-propagated to train the network end-to-end.

4. Experiments

4.1. Baselines

Besides the models discussed in Section 3.1, we experiment with the following traditional machine learning models: linear (Ridge Classifier) and tree-based models (Extra Trees Classifier). Since the target mask contains multiple outputs, we followed the multi-output classification during model training[3]. By tuning on validation data, we chose the following hyper-parameters: Ridge Classifier (regularization strength= 1.0, CV as leave-one-out cross-validation), Extra Trees Classifier (number of estimators=500, crite-

[3] https://scikit-learn.org/stable/modules/multiclass.html

rion=*gini*, and maximum depth=3). We also tried other methods like Gradient Boosting Decision Trees, Random Forests, and KNN-based multi-output regression methods. However, they did not lead to better results, so we do not present them here.

As another baseline, we consider a patch-based approach as follows. We split the wound image into different non-overlapping patches as input, and the corresponding mask patch is used as a target to train the model. We build a simple CNN model (patch-based CNN) consisting of two regular convolutional layers with 32 and 64 filters, followed by separable convolutional layers to predict the mask patch.

4.2. Evaluation Metrics

In order to validate the image segmentation performance of our models, we use the following popular evaluation metrics: IoU score and Dice score.

IoU Score: Intersection over Union (IoU) is a standard metric that allows us to evaluate how similar our predicted segmented mask is to the ground truth mask. To measure the IoU score, we compare the ratio of the area where the two masks overlap to the total combined area of the two masks. Given set A of predicted wound pixels and set B of ground truth wound pixels, IoU is calculated as $\frac{|A \cap B|}{|A \cup B|}$.

Dice Score: The Dice score captures the similarity between the segmented and the ground truth mask. More specifically, the Dice score ranges from 0 to 1 where a Dice score of 1 denotes perfect and complete overlap. Given set A of predicted wound pixels and set B of ground truth wound pixels, the Dice coefficient is calculated as $\frac{2|A \cap B|}{|A|+|B|}$.

4.3. Experimental Settings

We perform 5-fold cross-validation to assess our models' accuracy and report 5-fold average results. 4-folds were used for training, 1-fold is for the test. In order to train/fine-tune all the segmentation models, we use Dice loss as a loss function to measure the error between ground-truth and predicted masks. Note that Dice loss=1-(Dice score). We set the number of epochs as 100, batch size of 32, and optimizer as Adam with a learning rate of 0.001. We set the number of augmented samples to 8000. The experiments were performed on a single V100 16GB RAM GPU machine.

5. Results

5.1. Baseline Results on WoundSeg

Wound segmentation results on WoundSeg using baseline models are shown in Table 3. Baseline IoU and Dice scores are poor. We observe that the Patch-based CNN performs better than traditional machine learning methods.

Table 4 compares the performance of four popular segmentation models with 17 backbones where each backbone is loaded with ImageNet pretrained weights. We observe that both *DenseNet121* and *DenseNet169* backbones display better IoU and Dice Scores across three segmentation models: LinkNet, PSPNet, and FPN. On the other hand, EfficientNetB1 backbone displays a higher Dice score in the U-Net segmentation model. Overall, out of the 17 backbones, both *DenseNet121* and *DenseNet169* backbones yield higher Dice scores across the four segmentation models; hence, we use these backbones for further experiments in this paper. We also consider the *MobileNet* backbone due to its minimal training parameters and better portability of models in hand-held devices. We also show results for DenseNet121, DenseNet169 and MobileNet in Table 5 (Part A).

Model	IoU	Dice Score
RidgeClassifierCV	0.250	0.500
ExtraTreesClassifier	0.278	0.530
Patch-based CNN	0.280	0.540

Table 3. Comparison for baseline methods. We observe that both full image and patches-based baseline methods give similar results.

5.2. WSNet Results on WoundSeg

WSNet combines the global-local model, WDAP, and data augmentation. Table 5 shows results using WSNet-FF with backbone weights frozen during fine-tuning (Part F) and WSNet with end-to-end fine-tuning (Part G). End-to-end fine-tuning helped the global-local model achieve the best results. Overall, our best combination (Part G) leads to IoU=0.713 and Dice=0.847. Table 9 lists the performance of individual models on different wound types in our dataset. We observed that the best model performance is uniform across all wound types.

5.3. Ablation Results on WoundSeg

We experiment with different variations: pretraining (ImageNet/WDAP), data augmentation (yes/no), architecture (local/ global/ global-local). Table 5 (Part B) reports the results of four segmentation models by using specialized wound image pretrained weights corresponding to each backbone. Compared to ImageNet pretraining (Part A), WDAP (Part B) shows an improvement in both IoU and Dice scores for all the models. We achieve the highest Dice score of 0.8 using LinkNet with DenseNet121 in Part B. Further, Part C reports the data augmentation results along with WDAP and global-only models. We observe that data augmentation improves each metric performance for each segmentation model by approximately 2-3% (Part B vs. Part C). We also wanted to check the effectiveness of local-only models. We observe from Part D that the local model only results are worse compared to even Part A, implying that it is insufficient to capture local signals only. Lastly, does the pretraining type matter if we use global-local models?

Models→	U-Net		LinkNet		PSPNet		FPN	
Backbone	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
VGG16	0.574	0.726	0.568	0.722	0.522	0.684	0.541	0.701
VGG19	0.581	0.733	0.597	0.746	0.544	0.702	0.575	0.729
ResNet18	0.588	0.738	0.603	0.751	0.559	0.714	0.601	0.750
ResNet34	0.596	0.744	0.569	0.724	0.542	0.702	0.577	0.731
ResNext50	0.599	0.746	0.588	0.739	0.567	0.722	0.610	0.755
DN121	0.617	0.761	0.617	0.762	0.585	0.736	**0.623**	**0.766**
DN169	0.613	0.758	**0.624**	**0.768**	**0.596**	**0.745**	0.614	0.760
SERN18	0.581	0.732	0.596	0.745	0.546	0.704	0.568	0.722
SERN34	0.600	0.747	0.571	0.724	0.569	0.725	0.614	0.760
ENB0	0.610	0.755	0.588	0.738	0.551	0.708	0.607	0.754
ENB1	**0.621**	**0.764**	0.594	0.743	0.561	0.716	0.605	0.750
ENB2	0.606	0.753	0.597	0.747	0.588	0.739	0.588	0.739
ENB3	0.597	0.744	0.585	0.736	0.541	0.700	0.605	0.752
ENB4	0.602	0.750	0.602	0.750	0.567	0.721	0.596	0.745
MN	0.593	0.742	0.571	0.724	0.561	0.717	0.594	0.743
MNV2	0.506	0.667	0.488	0.651	0.391	0.556	0.525	0.687
IV2	0.606	0.753	0.533	0.692	0.554	0.711	0.613	0.759

Table 4. Results of performance of four image segmentation models on WoundSeg dataset using 17 different Imagenet pretrained backbone models. EN=EfficientNet, DN=DenseNet, MN=MobileNet, SERN=SEResNet, IV2=InceptionV2

		U-Net		LinkNet		PSPNet		FPN	
		IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
(A) Models with ImageNet pretraining	DenseNet121	0.617	0.761	0.617	0.762	0.585	0.736	0.623	0.766
	DenseNet169	0.613	0.758	0.624	0.768	0.596	0.745	0.614	0.760
	MobileNet	0.593	0.742	0.571	0.724	0.561	0.717	0.594	0.743
(B) Models with wound domain adaptive pretraining (WDAP)	DenseNet121	0.648	0.783	0.657	0.800	0.625	0.765	0.652	0.793
	DenseNet169	0.647	0.781	0.651	0.788	0.636	0.773	0.637	0.773
	MobileNet	0.615	0.760	0.611	0.755	0.563	0.718	0.616	0.758
(C) Models with WDAP and data augmentation	DenseNet121	0.680	0.818	0.687	0.820	0.653	0.797	0.680	0.817
	DenseNet169	0.672	0.810	0.675	0.812	0.656	0.801	0.664	0.807
	MobileNet	0.636	0.778	0.647	0.780	0.598	0.744	0.634	0.775
(D) Local (patch-based) models with WDAP	DenseNet121	0.527	0.689	0.537	0.698	0.520	0.682	0.532	0.694
	DenseNet169	0.534	0.696	0.530	0.691	0.519	0.681	0.533	0.696
	MobileNet	0.512	0.673	0.514	0.677	0.493	0.660	0.510	0.670
(E) Global-local models with ImageNet pretraining and data augmentation	DenseNet121	0.648	0.784	0.649	0.786	0.621	0.763	0.651	0.792
	DenseNet169	0.649	0.787	0.650	0.790	0.624	0.767	0.648	0.785
	MobileNet	0.620	0.761	0.621	0.763	0.565	0.722	0.618	0.760
(F) WSNet-FF: Global-local models with WDAP and data augmentation	DenseNet121	0.685	0.823	0.706	0.840	0.663	0.805	0.700	0.834
	DenseNet169	0.684	0.821	0.694	0.830	0.675	0.815	0.680	0.818
	MobileNet	0.650	0.790	0.651	0.792	0.590	0.740	0.651	0.792
(G) WSNet: Global-local models with WDAP, data augmentation, end-to-end fine-tuning	DenseNet121	0.695	0.831	**0.713**	**0.847**	0.683	0.820	**0.707**	**0.840**
	DenseNet169	**0.701**	**0.834**	0.707	0.841	**0.686**	**0.823**	0.697	0.832
	MobileNet	0.661	0.800	0.662	0.800	0.601	0.748	0.661	0.798

Table 5. Performance results of image segmentation models on WoundSeg dataset.

Hence, we trained global-local models with ImageNet pretraining (Part E) and found them significantly worse than WDAP-trained ones (Parts F and G). Dice score improvements for WSNet (Parts F and G) indicate that the WDAP and data augmentation has helped the model improve due to training on a more extensive set of in-domain examples.

Methods	AZH	Medetec
WSeg [21]	NA	0.919
VGG16 [35]	0.810	0.790
SegNet [35]	0.851	0.730
M-RCNN [35]	0.902	0.932
MNV2 [35]	0.903	0.938
MNV2+CCL [35]	0.905	0.940

Table 6. Dice-score comparison on the AZH and Medetec Datasets for baseline methods. WSeg results for AZH are not available.

5.4. Model size, Latency and Accuracy Tradeoff

For our best combination (global-local with WDAP, data augmentation, and end-to-end finetuning), we analyze the model size, latency (sec for a batch of 50 images), and Dice score tradeoff in Fig. 3 across three different backbones and four different architectures. Out of U-Net, LinkNet, PSPNet, and FPN, typically, PSPNet models have fewer parameters. Also, as expected, the Dice score is higher for larger-sized models. We obtain the best Dice score of 0.847 using a 16.7M parameter DenseNet121-LinkNet model with a latency of 0.146s. However, for resource-constrained deployments, we recommend the DenseNet169-PSPNet model, which has a small size of 6.3M parameters, 0.112s latency, and a decent Dice score of 0.823.

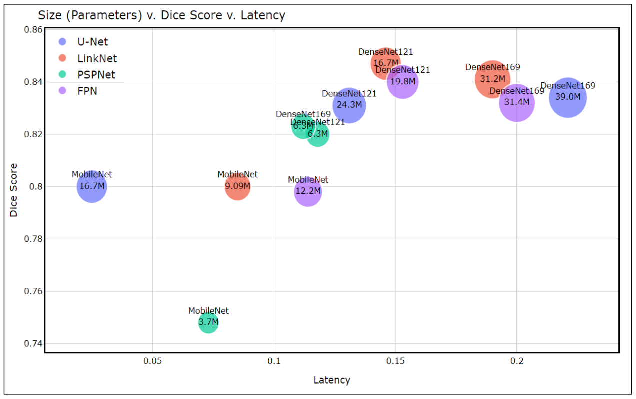

Figure 3. Comparison of model size (million), Dice score and latency (sec for a batch of 50 images) on WoundSeg for global-local models with data augmentation, specialized pretraining and end-to-end fine-tuning.

5.5. WoundSeg Case Studies

We show a few difficult examples in Fig. 4. The top illustration shows multiple wounds, the middle example shows two (one tiny and another bigger), while the bottom example shows an image with low lighting conditions. We observe that in all cases, at least one of our models obtains a finer boundary for the wound image than the ground truth mask. In general, FPN performs better than the other models in these cases.

Models	Methods	WDAP		WDAP+GloLocal	
		AZH	Medetec	AZH	Medetec
U-Net	DN121	0.920	0.948	0.923	0.951
	DN169	0.923	**0.956**	0.924	0.956
	MN	0.915	0.946	0.920	0.948
LinkNet	DN121	**0.927**	0.932	0.927	0.940
	DN169	0.913	0.945	0.921	0.948
	MN	0.913	0.933	0.915	0.938
PSPNet	DN121	0.908	0.930	0.912	0.932
	DN169	0.910	0.932	0.911	0.935
	MN	0.890	0.920	0.897	0.925
FPN	DN121	0.910	0.942	0.916	0.947
	DN169	0.912	0.953	0.920	0.954
	MN	0.911	0.941	0.918	0.944

Table 7. Dice-score comparison on the AZH and Medetec Datasets for our proposed WSTech methods. DN=DenseNet, MN=MobileNet.

Method	Area MAE	Volume MAE
HealTech [25]	1.14	1.28
WSNet with U-Net	0.66	**0.78**
WSNet with LinkNet	**0.65**	**0.78**
WSNet with PSPNet	0.71	0.82
WSNet with FPN	0.66	**0.78**

Table 8. Wound Area and Volume Prediction Results. Lower MAE is better. Wound area and volume are in cm^2 and cm^3. MAE is computed on quantities after taking log.

5.6. Results on Existing Benchmark Datasets

We also experiment with two existing benchmark datasets: AZH [35] and Medetec [33], as described in Section 2.

Table 6 and 7 show Dice score comparison results for previously proposed methods and our WSNet methods respectively. WSNet outperforms existing baselines by a significant margin across both the benchmarks. The improvements are statistically significant at 95% confidence. For our WSNet methodology, we show results with just WDAP and WDAP combined with global-local architecture. The results show that WDAP+global local is better than just WDAP, and both outperform existing baselines by a significant margin across both the benchmarks. Using patches in the local network helped detect wound borders with better precision.

5.7. Wound Area and Volume Prediction Results

We analyze the effect of pretrained wound image segmentation models on downstream tasks such as area and volume prediction by fine-tuning task-specific labeled data. The wound area data has 122417 images, while the wound volume dataset has 99719 images. We divide the data into 70:10:20 for train, validation, and test splits. We use Mean

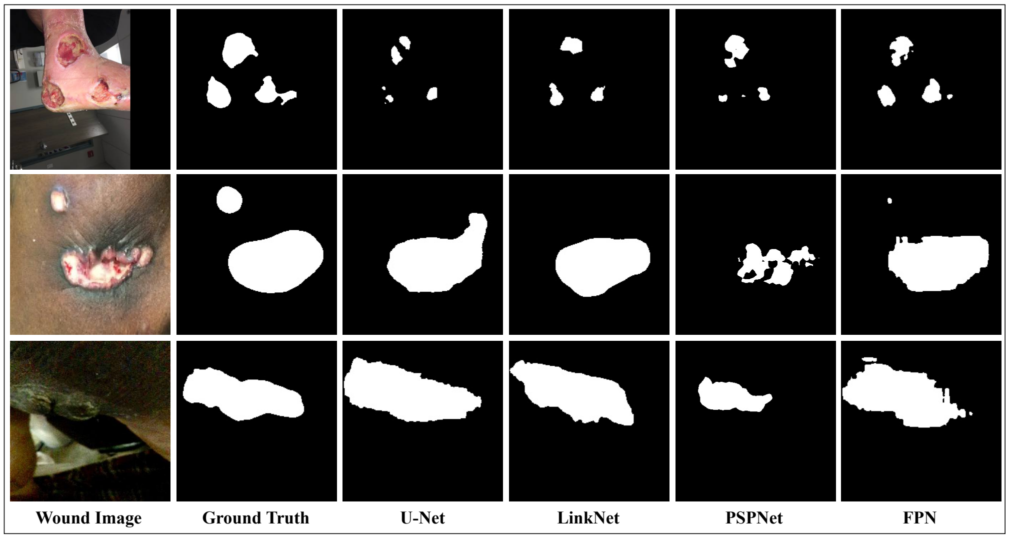

Figure 4. WSNet Predictions using the four global-local architectures (i) U-Net, (ii) LinkNet, (iii) PSPNet and (iv) FPN.

Models	Methods	Wound Type							
		Diabetic	Pressure	Surgical	Venous	Trauma	Arterial	Cellulitis	Other
U-Net	DN121	0.744	0.792	0.786	0.761	0.749	0.747	0.745	0.825
	DN169	0.742	0.789	0.771	0.761	0.757	0.752	0.752	0.826
	MN	0.719	0.749	0.745	0.755	0.737	**0.768**	0.736	0.786
LinkNet	DN121	0.733	0.774	0.767	0.748	0.748	0.745	0.761	0.815
	DN169	**0.763**	**0.803**	**0.800**	**0.774**	**0.769**	0.760	**0.794**	0.811
	MN	0.719	0.744	0.740	0.738	0.720	0.729	0.734	0.772
PSPNet	DN121	0.630	0.662	0.654	0.674	0.640	0.643	0.662	0.642
	DN169	0.616	0.640	0.643	0.660	0.633	0.621	0.627	0.653
	MN	0.580	0.591	0.587	0.584	0.594	0.590	0.572	0.613
FPN	DN121	0.747	0.779	0.768	0.764	0.760	0.742	0.783	0.806
	DN169	0.747	0.794	0.796	0.770	**0.769**	0.756	0.787	**0.839**
	MN	0.722	0.771	0.760	0.770	0.753	0.751	0.782	0.803

Table 9. Dice-score comparison on the WoundSeg Dataset for our proposed WSNet methods. DN=DenseNet, MN=MobileNet.

Absolute Error (MAE) as the metric, with the Adam optimizer and a learning rate of 0.001. Since the variation in area and volume values is large, we applied the log scaling to the target values. The average area is 7.74, and the max area of the wound is 14.25 on a log scale. Similarly, the average volume is 8.95, and the max wound volume is 16.86 on a log scale.

Oota et al. [25] directly used wound images for wound area and volume prediction. We used our models pretrained on wound segmentation for fine-tuning to predict the wound area and volume. From Table 8, we observe that our LinkNet models lead to the lower area and volume MAE compared to the HealTech baseline [25]. The results indicate that the segmentation training helped the model focus on core wound regions, unlike earlier models. We believe that the pretraining on a larger dataset and fine-tuning the wound segmentation model helped develop the depth sense needed for volume prediction.

6. Conclusion

We contribute a diverse dataset, WoundSeg, of 2686 images across eight wound types for the wound image segmentation task. We experimented extensively with four CNN model architectures and 17 backbones. We propose a novel WSNet framework that consists of wound-domain adaptive pretraining, data augmentation, global-local architecture, and end-to-end fine-tuning. The proposed methods outperform baselines on existing benchmark datasets, show beneficial results on the WoundSeg dataset, and even establish a new state-of-the-art on wound area and volume prediction tasks. We comprehensively studied the current best image segmentation models and improved the segmentation performance by introducing a novel method. We also contribute a better and larger wound image dataset, which can help the research community to advance wound image analysis further.

References

[1] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *VCIP*. IEEE, 2017.

[2] Jun Cheng, Jiang Liu, Yanwu Xu, Fengshou Yin, Damon Wing Kee Wong, Ngan-Meng Tan, Dacheng Tao, Ching-Yu Cheng, Tin Aung, and Tien Yin Wong. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. *IEEE T-MI*, 2013.

[3] Mohammed Elmogy, Begoña García-Zapirain, Connor Burns, Adel Elmaghraby, and Ayman Ei-Baz. Tissues classification for pressure ulcer images based on 3d CNN. In *ICIP*. IEEE, 2018.

[4] Mohammad Faizal Ahmad Fauzi, Ibrahim Khansa, Karen Catignani, Gayle Gordillo, Chandan K Sen, and Metin N Gurcan. Computerized segmentation and measurement of chronic wound images. *Computers in Biology and Medicine*, 2015.

[5] Begoña García-Zapirain, Mohammed Elmogy, Ayman El-Baz, and Adel S Elmaghraby. Classification of pressure ulcer tissues with 3d CNN. *Medical & Biological Engineering & Computing*, 2018.

[6] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE T-MI*, 2016.

[7] Ahmad Fadzil M Hani, Leena Arshad, Aamir Saeed Malik, Adawiyah Jamil, and Felix Yap Boon Bin. Haemoglobin distribution in ulcers for healing assessment. In *4th Intl. Conf. on Intelligent and Advanced Systems*, 2012.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *J. digital imaging*, 2019.

[10] NDJ Hettiarachchi, RBH Mahindaratne, GDC Mendis, HT Nanayakkara, and Nuwan D Nanayakkara. Mobile based wound measurement. In *IEEE PHT*, 2013.

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient CNNs for mobile vision applications. *arXiv:1704.04861*, 2017.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[14] Ashraf Khalil, Mohammed Elmogy, Mohammed Ghazal, Connor Burns, and Ayman El-Baz. Chronic wound healing assessment system based on different features modalities and non-negative matrix factorization (nmf) feature reduction. *IEEE Access*, 2019.

[15] Marina Kolesnik and Ales Fexa. Multi-dimensional color histograms for segmentation of wounds in images. In *Intl. conf. Image Analysis and Recognition*. Springer, 2005.

[16] Marina Kolesnik and Ales Fexa. How robust is the svm wound segmentation? In *NORSIG*, 2006.

[17] Tao Lei, Risheng Wang, Yong Wan, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: a survey. *arXiv:2009.13120*, 2020.

[18] Wen Li et al. Automatic segmentation of liver tumor in ct images with deep CNNs. *J. Computer and Communications*, 2015.

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[20] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017.

[21] Xiaohui Liu, Changjian Wang, Fangzhao Li, Xiang Zhao, En Zhu, and Yuxing Peng. A framework of wound segmentation based on deep convolutional networks. In *CISP-BMEI*. IEEE, 2017.

[22] Christos P Loizou, Takis Kasparis, Orthodoxia Mitsi, and Michalis Polyviou. Evaluation of wound healing process based on texture analysis. In *BIBE*, 2012.

[23] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE T-MI*, 2014.

[24] Subba Reddy Oota, Nafisur Rahman, Shahid Saleem Mohammed, Jeffrey Galitz, and Minghsun Liu. Wound and episode level readmission risk or weeks

to readmit: Why do patients get readmitted? how long does it take for a patient to get readmitted? In *8th ACM IKDD CODS and 26th COMAD*, 2021.

[25] Subba Reddy Oota, Vijay Rowtula, Shahid Mohammed, Jeffrey Galitz, Minghsun Liu, and Manish Gupta. Healtech-a system for predicting patient hospitalization risk and wound progression in old patients. In *WACV*, 2021.

[26] Elisabeth S Papazoglou, Leonid Zubkov, Xiang Mao, Michael Neidrauer, Nicolas Rannou, and Michael S Weingarten. Image analysis of chronic wounds for determining the surface area. *Wound Repair and Regeneration*, 2010.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015.

[28] CK Sen. Human wounds and its burden: An updated compendium of estimates. *Advances in Wound Care*, 8(2):39–48, 2019.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[30] Bo Song and Ahmet Sacan. Automated wound identification system based on image segmentation and artificial neural networks. In *BIBM*. IEEE, 2012.

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[32] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for CNNs. In *ICML*. PMLR, 2019.

[33] S Thomas. Medetec wound database, 2017.

[34] Francisco J Veredas, Héctor Mesa, and Laura Morente. Efficient detection of wound-bed and peripheral skin with statistical colour models. *Medical & Biological Engineering & Computing*, 2015.

[35] Chuanbo Wang, DM Anisuzzaman, Victor Williamson, Mrinal Kanti Dhar, Behrouz Rostami, Jeffrey Niezgoda, Sandeep Gopalakrishnan, and Zeyun Yu. Fully automatic wound segmentation with deep CNNs. *Scientific Reports*, 2020.

[36] Lei Wang, Peder C Pedersen, Diane Strong, Bengisu Tulu, and Emmanuel Agu. Wound image analysis system for diabetics. In *Medical Imaging 2013: Image Processing*, 2013.

[37] Shuo Wang, Mu Zhou, Zaiyi Liu, Zhenyu Liu, Dongsheng Gu, Yali Zang, Di Dong, Olivier Gevaert, and Jie Tian. Central focused CNNs: Developing a data-driven model for lung nodule segmentation. *Medical Image Analysis*, 2017.

[38] Hazem Wannous, Yves Lucas, and Sylvie Treuillet. Enhanced assessment of the wound-healing process by accurate multiview tissue classification. *IEEE T-MI*, 2010.

[39] Hazem Wannous, Sylvie Treuillet, and Yves Lucas. Supervised tissue classification from color images for a complete wound assessment tool. In *EMBC*. IEEE, 2007.

[40] Hazem Wannous, Sylvie Treuillet, and Yves Lucas. Robust tissue classification for reproducible wound assessment in telemedicine environments. *J. Electronic Imaging*, 2010.

[41] Kittichai Wantanajittikul, Sansanee Auephanwiriyakul, Nipon Theera-Umpon, and Taweethong Koanantakool. Automatic segmentation and degree identification in burn color images. In *The 4th 2011 Biomedical Engineering Intl. conf.*, 2012.

[42] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.