U-Net-based DFU Tissue Segmentation and Registration on Uncontrolled Dermoscopic Images

Yanexis Toledo[1][a], Leandro A. F. Fernandes[1][b], Silena Herold-Garcia[2][c] and Alexis P. Quesada[3]

[1] *Instituto de Computação, Universidade Federal Fluminense, Niterói, Brazil*
[2] *Facultad de Matemática y Computación, Universidad de Oriente, Santiago de Cuba, Cuba*
[3] *Hospital General Dr. Juan Bruno Zayas Alfonso, Santiago de Cuba, Cuba*

Keywords: Semantic Segmentation, Uncontrolled Viewpoint, Epipolar Constraints, Image Registration.

Abstract: Diabetic Foot Ulcers (DFUs) are aggressive wounds with high morbimortality due to their slow healing capacity and rapid tissue degeneration, which cause complications such as infection, gangrene, and amputation. The automatic analysis of the evolution of tissues associated with DFU allows the quick identification and treatment of possible complications. In this paper, our contribution is twofold. First, we present a new DFU dataset composed of 222 images labeled by specialists. The images followed the healing process of patients of an experimental treatment and were captured under uncontrolled viewpoint and illumination conditions. To the best of our knowledge, this is the first DFU dataset whose images include the identification of background and six different classes of tissues. The second contribution is an U-Net-based segmentation and registration procedure that uses features computed by hidden layers of the network and epipolar constraints to identify pixelwise correspondences between images of the same patient at different healing stages.

1 INTRODUCTION

During the treatment of Diabetic Foot Ulcers (DFUs), the patient's foot undergoes significant transformations. The different tissues alert the specialists about the clinical evolution of the patient's condition. Typically, the analysis of these ulcers is carried out visually by specialists, which is a process prone to errors. The automatic analysis of the evolution of DFUs can mitigate the problems arising from manual inspection. Two processes are required as part of the automatic analysis: the segmentation of the image regions corresponding to different tissues and the registration of the images of each patient throughout the treatment. In this paper, we address both processes. The semantic segmentation of DFU images requires robust computer vision techniques in the face of low contrast and variations on the image acquisition conditions. Deep learning models are state-of-the-art in semantic segmentation (Hao et al., 2020). However, the success of these models depends on the quality of training datasets. Unfortunately, large sets of DFU images are not available, and the existing ones label image pixels as *wound* and *non-wound* only. Our first challenge in this work was to handle these issues. In our approach, we have used images from the Foot Ulcer Segmentation Challenge 2021 (FUSeg) and the Medetec Wound Database to train an initial version of our U-Net-based segmentation model considering the two usual classes. Segmentation considering six different classes of tissue, plus background, was obtained by transfer learning from the 2-class to the 7-class model trained with a new DFU image dataset built for this work (Figure 1). Our dataset includes images captured under uncontrolled conditions from patients undergoing an experimental treatment in the General Dr. Juan Bruno Zayas Alfonso Hospital in Cuba. Each image in our dataset has a mask that labels the pixels as *background*, *epithelialization*, *healthy skin*, *granulation tissue*, *slough tissue*, *necrotic tissue*, and *exposed tendon*.

The second challenge in this work was to develop an image registration technique to keep track of wound evolution along with the treatment. In contrast to DFU analysis techniques described in the literature (Solís-Sánchez et al., 2016), which consider highly controlled environments during image acquisition, in this work, we assume that hard capturing

[a] https://orcid.org/0000-0003-4103-0225
[b] https://orcid.org/0000-0001-8491-793X
[c] https://orcid.org/0000-0001-9238-3472

Toledo, Y., Fernandes, L., Herold-Garcia, S. and Quesada, A.
U-Net-based DFU Tissue Segmentation and Registration on Uncontrolled Dermoscopic Images.
DOI: 10.5220/0010868600003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 510-517
ISBN: 978-989-758-555-5; ISSN: 2184-4321

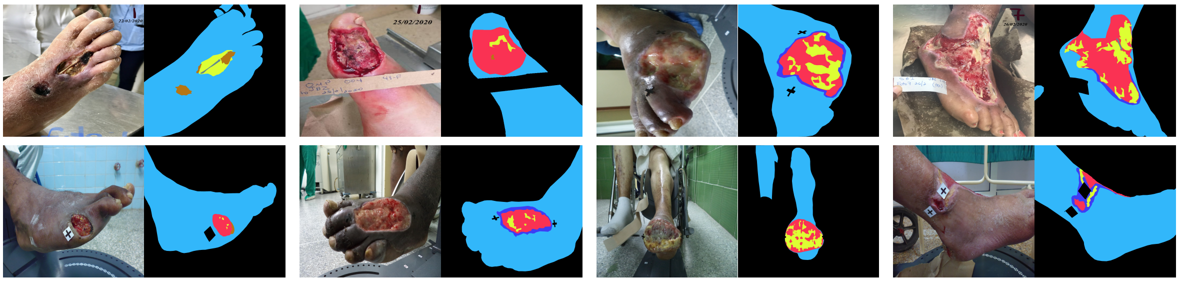

Figure 1: Images from our DFU dataset and their respective segmentation masks. The mask colors correspond to background ■, epithelialization 🟦, healthy skin 🟦, granulation tissue 🟥, slough tissue 🟨, necrotic tissue 🟧, and exposed tendon ■.

restrictions may prevent the application of an analysis technique in realistic scenarios. For such, the registration process must be robust to the temporal evolution of the wound's appearance, variations on distance, point of view, and lighting conditions. We use feature maps computed by the semantic segmentation model and its segmentation results to extract robust descriptors for tiles in the wounds and health skin regions. In turn, our registration approach uses epipolar constraints (Hartley and Zissermann, 2004) to establish correspondences and reduce the occurrence of matching outliers between pairs of images of the same patient acquired on different dates.

Our main contributions can be summarized as: (i) a new DFU dataset composed of 222 images whose pixels were labeled by specialists considering six types of tissues and background; and (ii) a new U-Net-based segmentation and registration procedure for DFU images. To the best of our knowledge, our DFU dataset is the one including the highest number of labels in the annotation masks. Also, our registration approach is the first to use features extracted from an U-Net as image descriptors for image registration.

2 RELATED WORK

DFU Image Segmentation. Convolutional Neural Networks (CNNs) are cutting edge in the image segmentation task of complex wounds such as DFUs. Although their performance varies according to the training dataset, they all show promising results for the 2-class segmentation problem (*wound* and *non-wound*) (Goyal et al., 2017; Wang et al., 2020; Wagh et al., 2020), and few address 4 classes (*external skin*, *necrosis*, *granulation*, and *slough*) (Kaswan et al., 2020). The main advantages of the U-Net architecture are the reduced times necessary for training and the good performance even with few training samples. The higher convergence speed of U-Nets during the training phase is related to the jump connections between the encoder and the decoder blocks, which contribute to smooth the descent path of the gradient towards the global minimum.

Image Registration. Klein et al. (2009) perform elastic alignment of intensity medical image via the estimation of displacement vectors. Klein's et al. approach is capable of representing complex local distortions but is prone to compatibility issues and requires excessive execution times. By combining non-elastic and elastic alignments, Zhang et al. (2019) are capable of performing natural (non-medical) image registration with the robustness of the parametric alignment. Experimental results show that Zhang's et al. method is accurate and surpasses a selection of last-generation image alignment approaches. Unfortunately, the implementation is not available, which prevents its use and continuity by other researchers. Long et al. (2014) proposed the applicability of feature vectors extracted from CNNs in tasks such as semantic segmentation and robust correspondence estimation. Recent works have attempted to analyze and explain this overwhelming success (Dara and Tumma, 2018). Our work extends these studies to DFU images using features computed in a segmentation model.

Current techniques that analyze the evolution of wounds consider controlled environments (Solís-Sánchez et al., 2016). The absence of restrictions in the image capture process shows that our wound registration technique goes beyond state of the art approaches.

3 MATERIALS AND METHODS

Our approach uses an U-Net CNN architecture (Ronneberger et al., 2015) to segment DFU images considering six different types of tissues, plus the background (Figure 2, top), and to extract robust descriptors for efficient region matching and subsequent image registration (Figure 2, bottom).

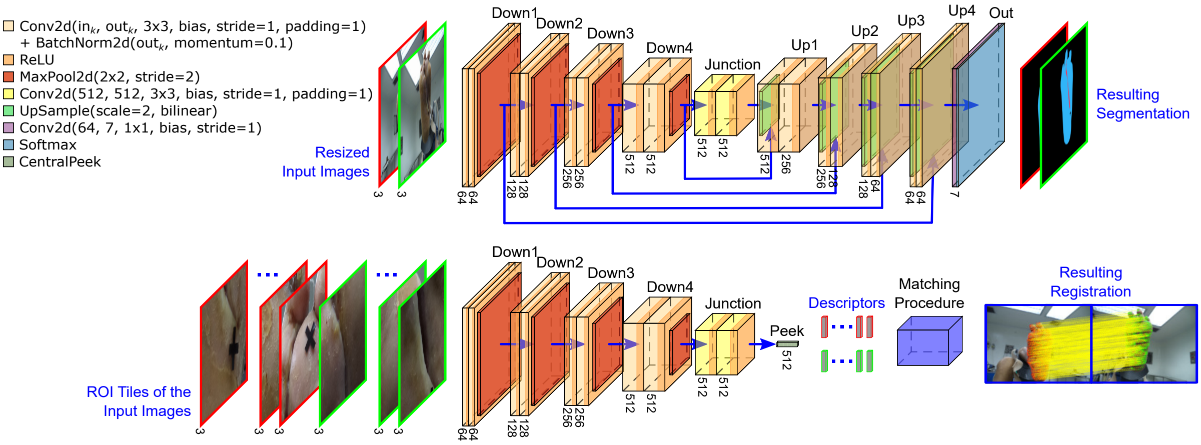
□ Conv2d(in$_k$, out$_k$, 3x3, bias, stride=1, padding=1) + BatchNorm2d(out$_k$, momentum=0.1)
□ ReLU
□ MaxPool2d(2x2, stride=2)
□ Conv2d(512, 512, 3x3, bias, stride=1, padding=1)
□ UpSample(scale=2, bilinear)
□ Conv2d(64, 7, 1x1, bias, stride=1)
□ Softmax
□ CentralPeek
Resized Input Images
Down1
Down2
Down3
Down4
Junction
Up1
Up2
Up3
Up4
Out
Resulting Segmentation
ROI Tiles of the Input Images
Peek
Descriptors
Matching Procedure
Resulting Registration
3
64
128
256
512
7

Figure 2: Overview of the proposed segmentation (top) and registration (bottom) methods.

3.1 Our Diabetic Foot Ulcer Dataset

Our dataset was collected at a hospital center over one year, between 2019 and 2020, by specialists in vascular wounds who carried out an experimental study on DFUs. The dataset is composed of 222 annotated DFU images of 28 patients. The image history of each patient consists of weekly image captures from the first time the patient attended the treatment until the wound is completely closed, with an average of 9 captures per subject. The images were taken by four different digital camera models under uncontrolled illumination conditions, with various viewpoints and backgrounds. Image resolution ranges from 522 × 692 to 4608 × 3456 pixels. The characteristics of the patients are pretty diverse, moreover, there is an evident imbalance among the types of tissues observed in the images. Necrotic tissue and exposed tendon are rare, while most images contain granulation tissue, slough tissue, and epithelialization (Figure 1). The segmentation masks in our dataset were created by an specialist using the free online Computer Vision Annotation Tool. All skin information within the image, whether or not it was from the patient, was marked as healthy skin tissue, and interfering objects such as paper markers, among others, were annotated as background.

3.2 Segmentation Network

The network expects as input a batch of RGB images having 224 × 224 pixels and produces one tensor with 224 × 224 pixels and seven channels for each image in the batch. Images with a resolution different than expected must be resized to fit the input shape, and the result is resized back to the original image resolution. The channels of the output encode the probabilities of a given pixel be related to the classes considered in our dataset. The segmentation network has a 4-block encoder-decoder structure. From the first to the last encoder block `Down`K, the number of output channels in the convolutions is 64, 128, 256, and 512, respectively. The encoder blocks are followed by the `Junction` block. These five blocks are in charge of extracting the essential characteristics of the input images, which will be used here for segmentation and in Section 3.3 as descriptors for image registration. The decoder blocks `Up`K in Figure 2 recover the resolution of the input image. The result of each up-sample in these blocks is concatenated to the feature map produced by the encoder layer of its same level. Thus, the number of input channels of the first convolution in each `Up`K block is, respectively, 1024, 512, 256, and 128. The `Out` block returns per-pixel probability distributions of the seven classes.

3.3 Image Registration Procedure

We restrict the registration process to the portions of the input image classified as non-background. According to our experience, classic algorithms for automatic detection of features such as SIFT and SURF fail on describing health skin and wound textures for three main reasons: (i) the keypoints are usually attracted by the boundaries of the foot; (ii) health skin regions are poor in texture; and (iii) the appearance of the wounded regions changes over time. In contrast, it has been shown that features computed by the image classification networks such as VGG and ResNet are successful as a source of image descriptors (Long et al., 2014; Dara and Tumma, 2018). We use feature vectors computed by the `Junction` block of the U-Net as source of region descriptors, even though the amount of DFU images available for training is much smaller than the number of images in datasets for image classification, like CIFAR-10 and COCO.

The steps of our registration procedure are *feature extraction*, *feature matching*, and *epipolar pruning with dense registration*.

Feature Extraction. We split the (original) input image into overlapping tiles having the size expected by the first layer of the U-Net (*i.e.*, 224×224 pixels) and stride s. In our experiments, s was empirically set to 10 pixels after analyzing its impact on processing time. The tiles whose central pixel were classified as one of the six types of tissue during the segmentation process (Figure 2, top) are submitted to the encoder blocks of the U-Net (Figure 2, bottom), and the feature vector at the center of the $14 \times 14 \times 512$ map produced by the `Junction` block is taken as the descriptor for a given tile (`Peek` operation). Thus, for each tile we have a feature vector d_{p} associated to the image location p at tile's center.

Feature Matching. After extracting sets of descriptors from two images acquired from the same patient, the feature matching step normalizes the feature vectors to unit vectors using the L2 norm and exhaustively computes the Sum of Squared Distances (SSD) between the normalized features from each set. For each normalized entry of the first set, the procedure returns the closest one in the second set if the SSD distance is less than the match threshold of 1% from a perfect match. Due to normalization, the SSD values range is $[0,4]$. Multiple features in the first set can match one feature in the second one. The feature matching step returns pairs of correspondent sparse image locations $\mathrm{p} \leftrightarrow \mathrm{p}'$, where p belongs to a non-background region of image I and p' is in a non-background region of image I'.

Epipolar Pruning with Final Dense Registration. We use the correspondences $\mathrm{p} \leftrightarrow \mathrm{p}'$ to estimate the epipolar geometry of the cameras used to capture the DFU images. For that, we apply RANSAC, which in addition removes most ambiguities, as it keeps track of the set of inliers related to the estimated fundamental matrix F, but that nevertheless also usually contains false-positive matches. RANSAC alone cannot perform elastic (non-linear) dense image registration from a sparse set of keypoints. Algorithm 1 describes how we use the RANSAC's consensus set, epipolar constraints, and angular pondering to perform dense elastic registration between DFU images I and I'. As a result of the image registration, one must expect coherence on displacement vectors computed from the location of a point $\mathrm{x} \in I$ to its corresponding point $\mathrm{x}' \in I'$ (see Figure 4, right). The same applies to the keypoints in the subset of inliers returned by RANSAC. Thus, after determining the fundamental matrix F (Algorithm 1, line 1), we use the consensus set to compute the cumulative distribution functions (CDF) of the angles between the x-axis and the displacement vectors $\vec{\mathrm{v}} = \mathrm{p}' - \mathrm{p}$ of the $\mathrm{p} \leftrightarrow \mathrm{p}'$ pairs in the subset of inliers (lines 2 and 3). We have assumed a discrete set of 360 angular values to compute the CDF. Next, for each pixel location $\mathrm{x} = (x, y, 1)^{\mathsf{T}}$ within a non-background region of I, we look for its correspondent pixel location $\mathrm{x}' = (x', y', 1)^{\mathsf{T}}$ in image I' (lines 4 to 10). We use the epipolar constraint given by $\mathrm{l}' = F\,\mathrm{x}$ to compute the epipolar line $\mathrm{l}' = (A', B', C')^{\mathsf{T}}$ (line 5), and the Bresenham algorithm to traverse the portions of l' that intersect with a non-background region in I' (lines 6 to 8). A, B, and C are the coefficients of the general form of the equation of a straight line, *i.e.*, $Ax + By + Cw = 0$. Recall that we extracted feature vectors for tiles in both images. Here, we assume that pixels x and x' are related to the feature vectors d_{p} and $d_{\mathrm{p}'}$ computed at the tile center closest to x and x', respec-

Algorithm 1: Epipolar Pruning with Final Dense Registration.

```
   Data: Sparse set of correspondences $\mathrm{p} \leftrightarrow \mathrm{p}'$
         of keypoints from images $I$ and $I'$ of
         the same patient
   Result: Dense elastic registration of the
           non-background regions of the input
 1 Use RANSAC to estimate the matrix $F$ that
     best fit the input set of correspondences and
     the respective subset of inliers
 2 Compute the angles between the $x$-axis and
     displacement vectors $\vec{\mathrm{v}} = \mathrm{p}' - \mathrm{p}$ using
     $\mathrm{p} \leftrightarrow \mathrm{p}'$ pairs from the subset of inliers
 3 Compute the normalized cumulative
     histogram of angles and call it the CDF of
     angular values
 4 for each pixel location x segmented as
     non-background in image $I$ do
 5     Use the epipolar constraint to compute
         the epipolar line $\mathrm{l}' = F\,\mathrm{x}$ in image $I'$
 6     for each pixel $\mathrm{x}' \in \mathrm{l}'$ and in a
         non-background region of $I'$ do
 7         Compute and store the weighted
             distance $\frac{1}{p_{\mathrm{x},\mathrm{x}'}} \|d_{\mathrm{x}} - d_{\mathrm{x}'}\|_{L2}^{2}$ within x
             (use the CDF to compute $p_{\mathrm{x},\mathrm{x}'}$)
 8     end
 9     Take the smallest distance and set the
         respective point $\mathrm{x}'$ as correspondent to x
10 end
11 The set of $\mathrm{x} \leftrightarrow \mathrm{x}'$ pairs define to the dense
     elastic registration of $I$ and $I'$
```

tively. Rather than simply comparing the features associated with x and x′, we use the inverse angle probabilities as a weighting factor of the squared distance of the descriptors while traversing the epipolar lines (line 7). Formally, for a given pixel location x, we retrieve $\mathrm{x}' = \arg\min_{\mathrm{x}' \in \mathrm{l}'} \frac{1}{p_{\mathrm{x},\mathrm{x}'}} \|d_{\mathrm{x}} - d_{\mathrm{x}'}\|^2_{L2}$ as its corresponding pixel location, where $p_{\mathrm{x},\mathrm{x}'}$ is the probability estimated for the angle between the *x*-axis and the displacement vector from x to x′. The use of epipolar constraints enhanced with angular restrictions is an original idea of our work. Our experience shows that the presence of outliers is mitigated when $\frac{1}{p_{\mathrm{x},\mathrm{x}'}}$ is used to weight the distance between features.

4 EXPERIMENTS AND RESULTS

We have implemented the network (Section 3.2) using PyTorch 1.4 and the image registration procedure (Section 3.3) using MATLAB R2020a[1]. We ran our experiments in an Intel Xeon E5-2698 v4 CPU with 2.2Ghz, 512Gb of RAM, and 8 GPUs NVIDIA Tesla P100-SXM2 with 16Gb of memory each. Section 4.1 presents experiments tailored to compare DFU segmentation results produced by our U-Net against ENet (Paszke et al., 2016), Deeplab V3 with ResNet-50 backbone (Chen et al., 2017), and SegNet (Badrinarayanan et al., 2017). Since we don't have enough images to train models that consider 7 classes from scratch, we first initialized the parameters of the networks with random values and trained the models for the *wound* and *non-wound* classes. Then, we applied transfer learning and performed the second phase of the training process using the 7-class dataset. Our experiments demonstrate the influence of data augmentation and transfer learning in the segmentation model's quality. Section 4.2 presents the results of the automatic registration procedure for pairs of DFU images from the same patient captured at different weeks. We take advantage of the segmentation model obtained in Section 4.1 to extract sparse feature descriptors from DFU images regions.

4.1 Segmentation Results

Data Preparation. We built the dataset used to train binary segmentation models by joining the FUSeg Dataset (Wang et al., 2021) and the Medetec Wound Database (Medetec, 2021). Altogether, this 2-class dataset includes 1,919 images. For the 7-class segmentation problem, the dataset used in our experiments includes the 222 images described in Section 3.1, plus 35 DFU images from the Medetec Wound Database not included in the binary segmentation dataset. These extra images were also annotated by a specialist. We took each set of images and randomly split them into training (60%), validation (20%), and test (20%) subsets. Then, the training subsets were augmented with images produced by the Albumentations library considering nine random transformations: Gaussian blur, motion blur, optical distortion, brightness contrast, scale, translation, rotation, and horizontal and vertical flip. In the end, the training subsets of the 2- and 7-class datasets included 5,765 and 1,550 images, respectively.

Hyperparameter Tuning. We performed hyperparameter tuning in both phases of the training process of the networks via a Bayesian approach (Bergstra et al., 2013), computing the F1-Score and the Cross Entropy (CE) as metrics for validation in, respectively, the 2- and 7-class segmentation tasks. Also, we used Hyperband (Li et al., 2017) as stopping criteria, with $\texttt{min\_iter} = 20$ and $\eta = 3$, and a limit of 30 runs per sweep due to time constraints. The implementations of Bergstra et al. (2013) and Li et al. (2017) procedures are available in the Weights & Biases toolset. The hyperparameters considered were: batch size (from 2 to 20), learning rate (from 10^{-7} to 10^{-1}), and optimizer type (Adam, SGD, and RMSprop). Tables 1 and 2 show the hyperparameter values of the best models found in each scenario. The final scores were calculated on the test subset.

Discussion. DeepLab models have better average performance in 2-class (Table 1) and 7-class segmentation tasks (Table 2), closely followed by U-Net models. However, it is important to comment that we have observed advantages in using U-Net to meet the purpose of our research (*i.e.*, segmentation *and registration* of DFU images). The first advantage is that they require less training time and significantly less GPU memory than DeepLab models. The second advantage is that the features extracted by U-Net, which are also used to register pairs of images, are vectors with only 512 components, while the latent features produced by DeepLab have 2,048 components. We have observed that using descriptors with four times more elements does not improve the quality of the registration process of DFU images but dramatically impacts its computational cost. In general, the performance of ENet and SegNet in DFU image segmentation tasks is much lower than U-Net and DeepLab, specially on the 7-classes case. In Table 2, the difference between mean F1-Score and mean Accuracy (Acc) of SegNet to U-Net reaches 0.15, and the dif-

[1] https://github.com/Prograf-UFF/DFU

Table 1: Hyperparameters, training values, and scores of 2-class segmentation models trained on different scenarios of data augmentation (DA). Bold and underlined values highlight the best and second-best results among the compared models.

		Hyperparameters			Training		Scores				
							F1				
DA	Network	Batch Size	Learning Rate	Optimizer	Epochs	Runtime	Val.	Test	IoU	Acc	Recall
–	U-Net	6	0.0486	SGD	48	0:35:03	0.82	0.89	0.90	**1.00**	0.88
	ENet	20	0.0016	Adam	56	**0:12:27**	0.79	0.86	0.87	0.99	0.84
	DeepLab	5	0.0586	SGD	36	2:02:14	**0.87**	**0.90**	**0.91**	**1.00**	**0.92**
	SegNet	3	0.0554	Adam	36	0:33:20	0.50	0.72	0.78	0.99	0.59
✓	U-Net	6	0.0853	SGD	66	**2:43:54**	**0.86**	**0.91**	0.91	**1.00**	0.89
	ENet	5	0.0323	SGD	87	3:04:01	0.85	0.89	0.90	**1.00**	0.88
	DeepLab	8	0.0643	SGD	20	5:49:51	0.85	**0.91**	**0.92**	**1.00**	**0.93**
	SegNet	10	0.0116	SGD	67	3:48:44	0.76	0.86	0.87	0.99	0.79

Table 2: Hyperparameters, training values, and scores of 7-class segmentation models on different scenarios of DA and TL.

			Hyperparameters			Training		Scores					
								CE					
TL	DA	Network	Batch Size	Learning Rate	Optimizer	Epochs	Runtime	Val.	Test	Mean F1	Mean IoU	Mean Acc	Mean Recall
–	–	U-Net	12	0.0228	SGD	41	0:12:26	0.28	0.31	0.91	0.40	0.91	0.46
		ENet	4	0.0041	RMSprop	56	0:18:08	0.40	0.38	0.88	0.29	0.88	0.33
		DeepLab	3	0.0142	SGD	26	**0:09:25**	**0.18**	**0.20**	**0.94**	**0.45**	**0.94**	**0.53**
		SegNet	3	0.0051	RMSprop	58	0:09:45	0.50	0.48	0.82	0.22	0.82	0.27
–	✓	U-Net	11	0.0105	SGD	41	0:31:17	0.18	0.21	0.94	0.46	0.94	0.55
		ENet	20	0.0286	Adam	57	**0:24:01**	0.20	0.23	0.93	0.41	0.93	0.47
		DeepLab	7	0.0534	SGD	24	1:48:58	**0.14**	**0.15**	**0.95**	**0.49**	**0.95**	**0.57**
		SegNet	11	0.0223	SGD	54	0:36:25	0.34	0.34	0.89	0.34	0.89	0.42
✓	–	U-Net	7	0.0853	SGD	44	0:09:05	0.23	0.24	0.92	**0.38**	0.92	**0.44**
		ENet	13	0.0521	SGD	52	0:04:11	0.27	0.31	0.91	0.32	0.91	0.38
		DeepLab	4	0.0484	SGD	24	0:09:46	**0.21**	**0.21**	**0.94**	**0.38**	**0.94**	**0.44**
		SegNet	7	0.0941	SGD	26	**0:03:15**	0.59	0.68	0.77	0.17	0.77	0.21
✓	✓	U-Net	9	0.0890	SGD	40	**0:38:30**	0.17	0.19	0.94	**0.48**	0.94	0.55
		ENet	6	0.0871	SGD	51	0:54:00	0.20	0.23	0.93	0.39	0.93	0.45
		DeepLab	7	0.0549	SGD	30	1:27:50	**0.14**	**0.17**	**0.95**	**0.48**	**0.95**	**0.56**
		SegNet	7	0.0445	SGD	39	0:48:03	0.47	0.44	0.85	0.32	0.85	0.39

Table 3: Per class metric scores of U-Net models trained for 7 classes on different scenarios of transfer learning (TL) and data augmentation (DA). Color/Class legend in Figure 1.

		F1							IoU						
TL	DA	⬛	🟦	🟦	🟥	🟨	🟫	🟩	⬛	🟦	🟦	🟥	🟨	🟫	🟩
–	–	0.96	0.00	0.88	0.71	0.46	0.40	**0.00**	0.92	0.00	0.78	0.55	0.30	0.25	**0.00**
–	✓	**0.98**	0.21	0.92	0.77	**0.57**	0.46	**0.00**	0.95	0.11	0.85	0.63	**0.39**	0.30	**0.00**
✓	–	0.96	0.03	0.89	0.75	0.48	0.05	**0.00**	0.93	0.02	0.80	0.60	0.31	0.03	**0.00**
✓	✓	**0.98**	**0.26**	**0.93**	**0.81**	0.54	**0.47**	**0.00**	**0.96**	**0.15**	**0.87**	**0.68**	0.37	**0.31**	**0.00**

ference between mean Intersection over Union (IoU) and mean Recall of the same architectures are up to 0.21 and 0.23, respectively. For ENet models, the higher differences in the mean scores concerning U-Net range from 0.03 (F1 and Acc) to 0.13 (Recall). According to our experiments on binary segmentation, data augmentation improved the F1, IoU, and Recall scores of U-Net by 2.25%, 1.11%, and 1.14%, respectively. In recent work, Wang et al. (2020) achieved F1-Scores of 0.90 using a MobileNetV2 with Connected-Component Labeling (CCL) as post-processing stage, and a dataset with 1,109 DFU images. We managed to achieve an F1-Score of 0.91 on the test data without post-processing. On the segmentation of 7-classes, data augmentation alone improved the scores by 3.3%, 15.0%, and 19.57%, respectively.

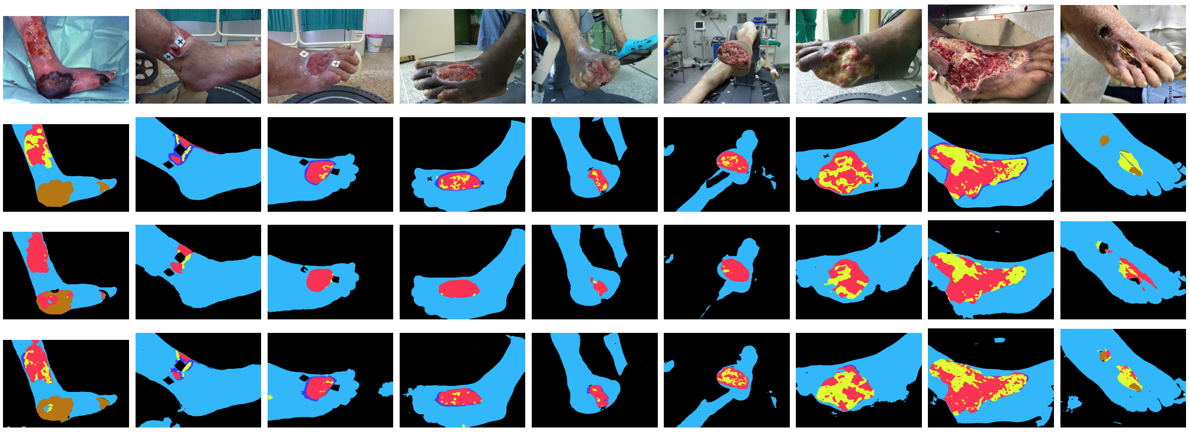

Figure 3: Results on the 7-class segmentation problem. The rows present, respectively, the input images, ground truth masks, segmentation by the model trained without TL nor DA, and by the model with TL and DA.

The use of transfer learning and data augmentation helped to increase the F1 and IoU scores of U-Net on almost all classes (Table 3). Unfortunately, the random split performed to compose the training, validation, and test subsets prevented the rarest case of tissue (*i.e.*, exposed tendon) from being identified by U-Net, ENet, DeepLab, and SegNet. We have only three images with an exposed tendon in our dataset, and the small area of this tissue tends to become unimpressive when the image is reduced to 224 × 224. However, this tissue is not as important in DFU analysis as are epithelization (🟦) and granulation (🟥), which indicate advances in ulcers healing, or slough (🟨) and necrotic (🟫) tissues, which represent barriers to recovery ulcers. By comparing the ground truth masks (Figure 3, second row) to segmentation result, its is clear that the 🟦, 🟥, 🟨, and 🟫 regions of the DFUs were better segmented using data augmentation combined with transfer learning (fourth row) than without the use of these techniques (third line).

4.2 Registration Results

The naive matching of descriptors computed by the U-Net leads to a large number of inconsistent correspondence pairs. For instance, note the crossed lines in the first column of examples shown in Figure 4. In all the cases analyzed we have noticed that the ratio between the amount of incorrect and correct matches changes favorably after RANSAC found the fundamental matrix that best fits the data. For example, in the second column in Figure 4, the remaining sets of matches are more consistent than the sets presented in the first column. The number of matching pairs dropped from 670 to 507 in the first row and 806 to 92 in the second row. Recall that sparse correspondence is not sufficient to carry out image registration. The last column of Figure 4 shows examples of dense registration achieved by the last step of our approach. The use of the angular weights combined with epipolar constraints was key in finding coherent dense correspondences because some descriptors are similar, especially on healthy skin which is poor in texture. The angular weighting showed good efficiency in the elimination of outliers.

5 CONCLUSIONS

This paper presented a DFU image dataset and a segmentation and registration procedure for such images. Our dataset includes 222 images and their respective segmentation masks considering six different types of tissues, plus background. We first trained the segmentation network assuming the wound and non-wound classes of different public datasets and then applied transfer learning to extend the classification to the 7-class semantic segmentation in our dataset. The registration process is based on the comparison of visual clues encoded by features computed by the encoder of the U-Net used for segmentation and geometrical constraints from the epipolar geometry of the cameras. The execution of both segmentation and registration methods performed well, especially considering that the wound of the same patient changes in appearance during treatment, making the dense registration more challenging. Our experiments proved that training U-Net models with small DFU datasets is sufficient to obtain feature descriptors representative enough to perform DFU image registration. This is an exciting result of our work, as other authors (*e.g.*, Long et al. (2014); Dara and Tumma (2018)) have pointed out that large datasets are needed to produce representative features for natural images.

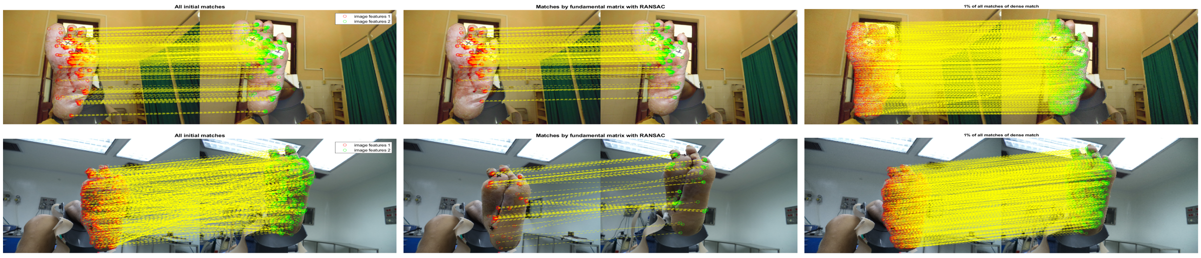
All initial matches
image features 1
image features 2
Matches by fundamental matrix with RANSAC
1% of all matches of dense match
All initial matches
image features 1
image features 2
Matches by fundamental matrix with RANSAC
1% of all matches of dense match

Figure 4: From the left to the right: sparse matching of features, consensus set returned by RANSAC, and dense registration. We present 1% of the correspondences to avoid clutter.

Future work include monitoring the evolution of wounds over time by calculating tissue area variations from the dense registration. Also, we are working on an attention mechanism to guide the model to classify pixels as DFU tissue, healthy skin, or background and then segment the five types of DFU tissues.

ACKNOWLEDGEMENTS

Y. Toledo and L. A. F. Fernandes are partially funded by Google LARA 2020. L. A. F. Fernandes is also funded by a FAPERJ grant (E-26/202.718/2018), and Y. Toledo is also funded by a CAPES fellowship.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *PAMI*, 39(12):2481–2495.

Bergstra, J., Yamins, D., and Cox, D. (2013). Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proc. ICML*, pages 115–123.

Chen, L. C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587.

Dara, S. and Tumma, P. (2018). Feature extraction by using deep learning: a survey. In *Proc. ICECA*, pages 1795–1801.

Goyal, M., Yap, M. H., Reeves, N. D., Rajbhandari, S., and Spragg, J. (2017). Fully convolutional networks for diabetic foot ulcer segmentation. In *Proc. SMC*, pages 618–623.

Hao, S., Zhou, Y., and Guo, Y. (2020). A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321.

Hartley, R. and Zissermann, A. (2004). *Multiple view geometry in computer vision*. Cambridge University Press, 2nd edition.

Kaswan, K. E., Jamil, N., and Roslan, R. (2020). Deep learning on wound segmentation and classification: a short review and evaluation of methods used. *Southeast Eur. J. Soft Comput.*, 9(2):6–10.

Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. W. (2009). Elastix: a toolbox for intensity-based medical image registration. *Trans Med Imaging*, 29(1):196–205.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2017). Hyperband: a novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.*, 18(1):6765–6816.

Long, J. L., Zhang, N., and Darrell, T. (2014). Do convnets learn correspondence? In *Proc. NeurIPS*, pages 1601–1609.

Medetec (2021). Medetec wound database. [Online]. Available: https://medetec.co.uk/slide\%20scans/foot-ulcers/index.html.

Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E. (2016). ENet: a deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, volume 9351, pages 234–241.

Solís-Sánchez, L. O., Ortiz-Rodriguez, J. M., Castañeda-Miranda, R., Martinez-Blanco, M. R., Ornelas-Vargas, G., Galván-Tejada, J. I., Galván-Tejada, C. E., Celaya-Padilla, J. M., and Castañeda-Miranda, C. L. (2016). Identification and evaluation on diabetic foot injury by computer vision. In *Proc. ICIT*, pages 758–762.

Wagh, A., Jain, S., Mukherjee, A., Agu, E., Pedersen, P. C., Strong, D., Tulu, B., Lindsay, C., and Liu, Z. (2020). Semantic segmentation of smartphone wound images: comparative analysis of AHRF and CNN-based approaches. *IEEE Access*, 8:181590–181604.

Wang, C., Anisuzzaman, D. M., Williamson, V., Dhar, M. K., Rostami, B., Niezgoda, J., Gopalakrishnan, S., and Yu, Z. (2020). Fully automatic wound segmentation with deep convolutional neural networks. *Scientific Reports*, 10(1):1–9.

Wang, C., Anisuzzaman, D. M., Williamson, V., Dhar, M. K., Rostami, B., Niezgoda, J., Gopalakrishnan, S., and Yu, Z. (2021). FUSeg dataset. [Online]. Available: https://github.com/uwm-bigdata/wound-segmentation/.

Zhang, X., Gilliam, C., and Blu, T. (2019). Parametric registration for mobile phone images. In *Proc. ICIP*, pages 1312–1316.