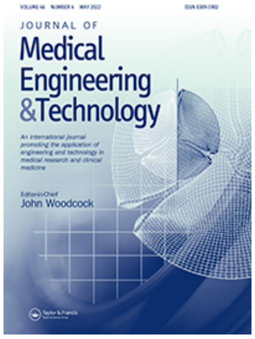


Taylor & Francis
Taylor & Francis Group

Journal of Medical Engineering & Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ijmt20

Diabetic foot thermal image segmentation using Double Encoder-ResUnet (DE-ResUnet)

Doha Bouallal, Hassan Douzi & Rachid Harba

To cite this article: Doha Bouallal, Hassan Douzi & Rachid Harba (2022): Diabetic foot thermal image segmentation using Double Encoder-ResUnet (DE-ResUnet), Journal of Medical Engineering & Technology, DOI: 10.1080/03091902.2022.2077997

To link to this article: https://doi.org/10.1080/03091902.2022.2077997

Published online: 31 May 2022.

JOURNAL OF MEDICAL ENGINEERING & TECHNOLOGY
https://doi.org/10.1080/03091902.2022.2077997

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Diabetic foot thermal image segmentation using Double Encoder-ResUnet (DE-ResUnet)

Doha Bouallal[a], Hassan Douzi[a] and Rachid Harba[b]

[a]IRF-SIC Laboratory, Ibn Zohr University, Agadir, Morocco; [b]Prisme Laboratory, Polytech Orléans, Orléans, France

ABSTRACT

The use of thermography in the early diagnosis of Diabetic Foot (DF) has proven its effectiveness in identifying areas of the plantar foot that are susceptible to ulcer development. Segmentation of the foot sole is one of the most pertinent technical issues that must be performed with great precision. However, because of the inherent difficulties of foot thermal images, such as unclarity and the existence of ambiguities, segmentation approaches have not demonstrated sufficiently accurate and reliable results for clinical use. In this study, we aim to develop a fully automated, robust and accurate segmentation of the diabetic foot. To this end, we propose a deep neural network architecture adopting the encoder-decoder concept called Double Encoder-ResUnet (DE-ResUnet). This network combines the strengths of residual network and U-Net architecture. Moreover, it takes advantage of RGB (Red, Green, Blue) colour images and fuses thermal and colour information to improve segmentation accuracy. Our database consists of 398 pairs of thermal and RGB images. The population includes two groups. The first group of 54 healthy subjects. And a second group of 145 diabetic patients from the National Hospital Dos de Mayo in Peru. The dataset is splitted into 50% for training, 25% for validation and the last 25% is used for testing. This proposed model provided robust and accurate automatic segmentations of the DF and outperformed other state of the art methods with an average intersection over union (IoU) of 97%. In addition, it is able to accurately delineate the part of toes and heels which are high risk regions for ulceration.

ARTICLE HISTORY
Received 3 September 2021
Revised 7 April 2022
Accepted 10 May 2022

KEYWORDS
Diabetic foot; thermal imaging; mobile health; deep learning; image segmentation

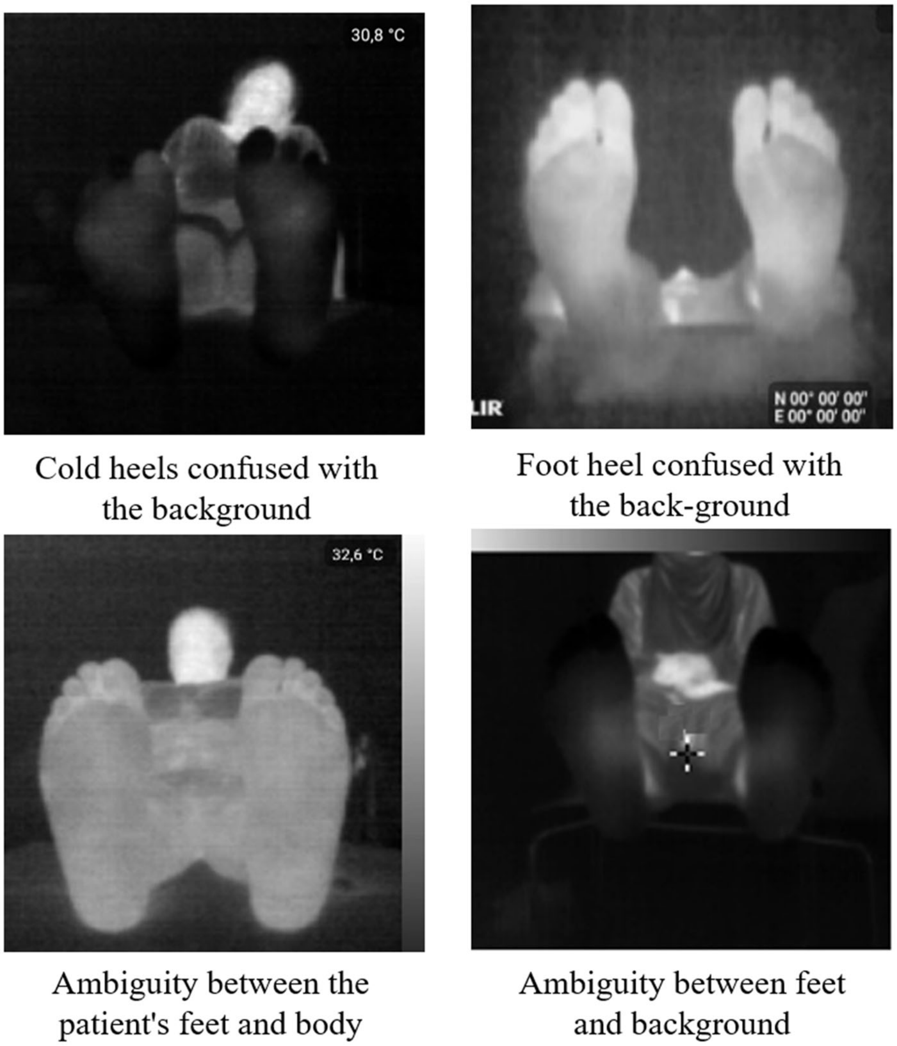
1. Introduction

Diabetes-related diseases mainly affect the feet, eyes, heart, kidneys, nervous system and blood vessels [1, 2]. This work solely concerns Diabetic Foot disease (DF), which is defined as infection, ulceration or destruction of the deep tissues of the foot [3]. It also includes peripheral arterial disease and neuropathy [4]. The severity of DF can lead to hospitalisation or even to lower limb amputation, which imposes a major burden to society and great loss in health-related quality of life for patients. In many cases, early detection of DF and preventive care have proven their efficiency limiting the development of foot ulcers and related amputation. Once diabetic foot disease is detected, the patient can be treated with specific education, regular foot care and therapeutic shoe inserts. Consequently, the incidence of serious complications, i.e., ulceration and amputations, could be further reduced, according to diabetes experts.

Abnormal temperature variation in the patient's foot can be an early indicator of diabetic foot disorders [5, 6]. Accordingly, skin temperature is an important factor in the assessment of foot health. Nevertheless, this data is currently not well exploited for diabetes-related disease monitoring and detection. The most common and clinically effective monitoring protocol for DF ulcers is the daily temperature comparison of six contralaterally matched plantar zones as described in [5]. This self-monitoring procedure is time consuming and there is often a lack of adherence to monitoring protocols, there is also the assessment of foot temperature by manual palpation, which is neither reliable nor efficient method and depends on the expert's level of competence.

Thermography is a non-invasive, safe, accessible, non-contact and easily reproducible technique that has been used in several fields, such as military [7], space [8], civil applications [9] and medicine [5, 10, 11]. In the medical field, thermography has been used for the diagnosis and detection of soft tissue

CONTACT Doha Bouallal doha.bouallal@uiz.ac.ma Ibn Zohr University Faculty of Science Agadir, Computer Science and Mathematics, B.P 8106 Cité Dakhla Agadir, AGADIR, 80000 Morocco



Cold heels confused with the background

Foot heel confused with the back-ground

Ambiguity between the patient's feet and body

Ambiguity between feet and background

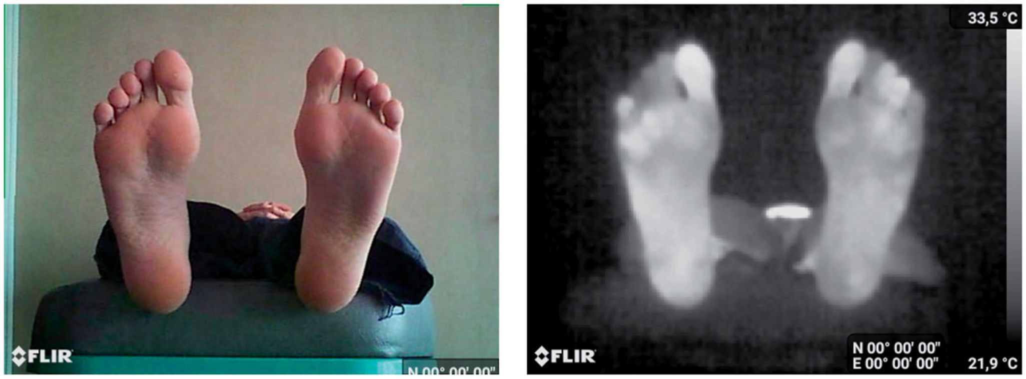
Figure 1. Examples of thermal images difficult to segment.

pathologies based on temperature measurement [12–14] and has proven its effectiveness in identifying possible ulcerous regions in the plantar foot.

A temperature difference between the two feet of more than 2.2 °C is considered abnormal and is called hyperthermia. This hyperthermia can be present up to two weeks before the development of a foot ulcer. The Early detection of hyperthermia on the sole of the foot in at-risk patients decreases the incidence of foot ulcers by 3 times. This major result clearly indicates the potential of such system. Therefore, the development of new effective diagnostic tools using thermal cameras has become an attractive need. Various technologies have been developed to measure foot temperature in order to detect hyperthermia [15]. Unfortunately, none of these systems are effective or easy to use.

In recent years, the use of thermal cameras has started to gain interest in the medical field, essentially since the price of infra-red cameras has fallen sharply while their technical capabilities have increased considerably. Therefore, these technologies are strong candidates for detecting thermal changes in diabetic foot disorders. Nonetheless, the implementation of a thermal camera-based monitoring system requires the resolution of several technical issues before being integrated into clinical care protocols. These technical issues include the choice of camera, the acquisition protocol, automating the processing of the acquired images and extracting the maximal amount of thermal information from these data.

Among the most challenging tasks in the development of such a system, there is the automatic segmentation of the plantar soles. It is a crucial and indispensable step in this early diagnosis system. This step must be automatic, fully unsupervised and without any interaction with the user. Manual segmentation of the plantar sole depends on the observer and consumes a lot of time, hence the essence of automating this essential step.



(a)

(b)

Figure 2. Acquisition example: (a) the RGB image and (b) the corresponding thermal image.

To deal with the segmentation task, most of the existing works have defined restrictive acquisition protocols, which consist in homogenising the background of the images and masking all thermal sources except those coming from the plantar foot [16–18] which facilitates the separation of the foot from the rest of the background. The majority of studies are based only on thermal images in the segmentation [11, 19–21]. These thermal images consisting of a single channel suffer from important limitations; they are unclear, and some regions cannot be detected, for example the toes or cold heels (Figure 1). These areas may go unnoticed or confused with other parts in the background. Similarly, some heat sources in the body or in the background may be considered as part of the foot as they have similar statistical characteristics. In addition, the thermal sensors used in previous works are not equipped with visible light cameras to provide spatially registered RGB images on the thermal images, which requires an additional camera.

Lately, multi-modal fusion strategies have gained attention due to the decreasing price of sensors. They are usually based on existing modality-specific methods that, once combined, enrich the representation of the scene in such a way that the strengths of one modality offset the weaknesses of another. That's why we thought of integrating the colour information in our segmentation process [22] to overcome the limitations of classical methods. In this context, several works adopted the fusion of thermal and colour image to improve the segmentation performance. In the field of autonomous vehicles, [23] and [24] have developed two neural networks RTFNet and MFNet, respectively that take as inputs both thermal and RGB images. Their objective is to be able to detect pedestrians and urban scenes in different light conditions. Especially during the night. And this fusion strategy was very efficient in this case. These two architectures are based on the encoder decoder concept. Similar to FuseNet [25], which fuses depths images and RGB images to indoors scenes segmentation. Visual images RGB provide detailed morphological information and clear delineation of the feet unlike IR images. Moreover, we used the Flir one pro camera which contains two sensors, thermal and RGB respectively. This allows it to capture spatially registered RGB and IR images.

The object of this paper is to improve the accuracy and robustness of diabetic foot segmentation. Based on two main ideas; the first one is to create a novel network that combine the architectures U-Net and ResNet which has proven their efficiency in several studies [26, 27] and the second one is to take advantage of thermal cameras, by fusing RGB and thermal information to reach a better performance.

The rest of the paper is organised as follows. Section 2 describes the acquisition protocol and the proposed DE-ResUnet method. Section 3 is devoted to experiments and results of the tested methods, and finally, discussion and conclusion are presented in the last sections.

2. Materials and methods

2.1. Data acquisition

The RGB and thermal images were acquired with a FlirOne Pro camera [28]. The chosen camera is designed to be plugged into a smartphone. We used a Samsung Galaxy S8 smartphone. This camera consists of two sensors. A thermal sensor that measures heat through infra-red emission, characterised by a thermal image resolution of 160 × 120, and the spectral range of the thermal sensor is 8–14 μm. The other camera is a conventional 1440 × 1080 pixel visual camera, designed to work in parallel with the thermal core

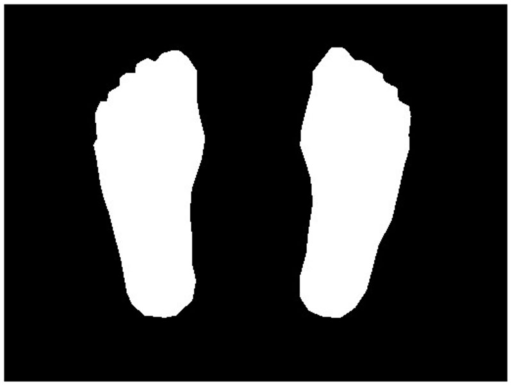


Figure 3. Illustrative example showing ground truth of image in Figure 2.

to produce images and video with a much higher resolution than 160 × 120.

The camera software includes an alignment control technology called Multi-Spectral Dynamic Imaging (MSX) [29], allowing it to provide spatially registered RGB and thermal images. MSX adds visible light details to thermal images in real time for greater clarity, embedding edge and outline detail onto thermal readings. Unlike image fusing (merging of a visible light and thermal image), MSX does not dilute the thermal image or decrease thermal transparency. Every time you take a picture, both a thermal image and a visible image are captured simultaneously (Figure 2). Another feature that was very important in our study is that Flir One Pro camera can detect temperature differences of 0.1 °C which is sufficient to detect possible hyperthermia variations that may appear in the sole. Combined with the ability to measure higher temperatures than the third-generation Flir One or the Flir One Pro LT, the Flir One Pro is a powerful model.

Our database consists of 398 pairs of thermal and RGB images. The population included two groups. The first group of 54 healthy subjects, 23 women and 31 men with an age range of 24 to 70 years, who were part of the staff members of the two laboratories IRF_SIC of Ibn Zohr University and PRISME laboratory of the University of Orleans. And a second group of 145 diabetic patients participated in an acquisition campaign conducted from 14 January 2019 to 9 March 2019, within the diabetic service of National Hospital Dos de Mayo (HNDM), Lima, Peru. The Ethic committee of HNDM has approved this study on 10 January 2019. For type II diabetes, exclusion criteria were defined, such as patients suffering from ulcers, neurodegenerative diseases, or amputations. These criteria are comparable to those used in other similar studies [30–32]. 145 type II diabetic patients accepted to participate in our study and signed the informed consent form. These patients were taken in charge by qualified nurses and medical doctors. General data include age, time of diagnosis (TOD), and body mass index (BMI). Clinical data concern the evaluation of foot deformity, the neurological assessment and the vascular assessment.

Two images were acquired for each subject and stored in Portable Network Graphic (PNG) format. The first was taken at the beginning of the examination (T0), as soon as the person sits, he/she is asked to remove shoes and socks. The person lays down on a stretcher and places the feet at the end of the stretcher in an upright position. The second acquisition (T10) was taken 10 min later, in order to allow the feet to return to their normal temperature. Meanwhile the subject was at the same resting position. Note that no subject participated in the acquisition more than once. The acquisition of the images is done freehandedly, without the use of any object to homogenise the background as has been done in other works [16, 17].

Our images were manually segmented in order to extract the soles of the feet from the background. No expert was needed to accomplish this essential task (Figure 3). Each image was segmented using the Image labeller app provided by MATLAB software (version R2018a) [33]. There are two classes in our images; namely the plantar sole and the background which groups the rest of the image and any object different from the foot. The resolution of the images is 480 × 640.

2.2. The proposed network

In the field of medical imaging, the samples available to train deep neural networks are often limited, difficult to collect or inaccessible. This is the case in our study. And to solve this problem of lack of data, we have two solutions, either to use a pre-trained network and then refine it on the target data set, as it has been done in several works [34], or the second solution is to use data augmentation. U-Net [35] is one of the famous Fully Convolutional Networks (FCN) [36], which have been highly successful at biomedical image segmentation especially in small datasets. U-Net's strength consists of copying low-level features to the corresponding high levels which creates a propagation of information, facilitates the backward propagation during training and compensates the finer details of the low-level to the high-level semantic features.

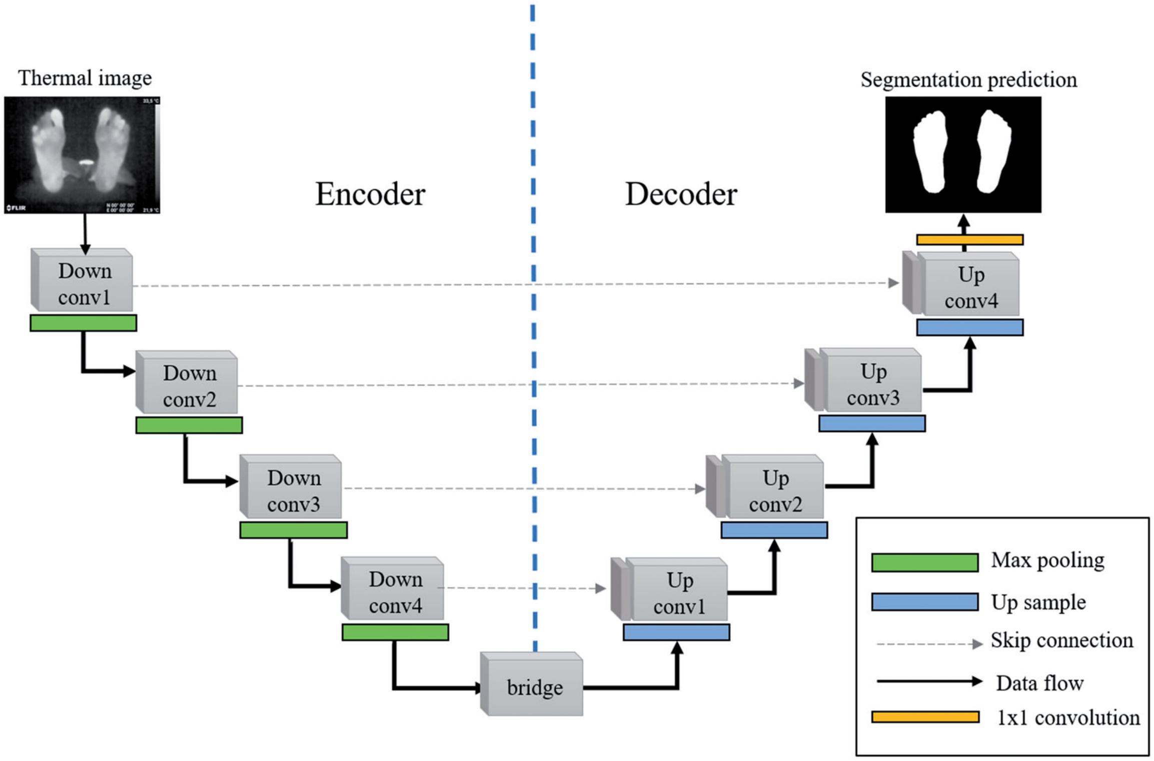


Figure 4. The U-Net Architecture. Comprises an encoder and a decoder pathway, with skip connections between the corresponding layers.

2.2.1. Unet

As shown in the figure below (Figure 4), UNet [35] has a "U" shape, which justifies its name. This architecture, similar to FCN and SegNet [37], uses fully convolutional layers to perform the semantic segmentation task. It consists of an encoder (also called contraction path) that extracts the image's spatial features and a decoder (expansion path) which builds the segmentation map from the encoded features. The encoder is composed of four contraction blocks. Each block consists of a sequence of two 3×3 convolution operations, followed by a max pooling operation of size 2×2 and stride of 2. After each down-sampling, the number of filters in the convolutional layers is doubled. A block of two 3×3 convolutional operations followed by a 2×2 up sampling layer acts as a bridge between the encoder and the decoder. Symmetrically, the decoder also consists of a set of expansion blocks. Each block passes the input to two 3×3 convolutional layers followed by a 2×2 up sampling operation which halves the feature channels. Similarly, to the encoder this sequence of up-sampling and two convolution operations is repeated four times. Finally, a 1×1 convolution operation is performed to generate the final segmentation map.

UNet is a network and training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization, and most importantly, it can be trained end-to-end from very few images. That is why we adopted it in our architecture, since we have a limited database.

2.2.2. Residual blocks

The ResNet deep network [38] has arguably been the most ground-breaking work in the computer vision and deep learning community in recent years, after blowing people away in 2015 with its famous victory in the ILSVRC classification competition. It outperformed humans with 3.6 classification errors and expanded the network to a depth of 1202 layers. ResNet *a*llows training up to hundreds or even thousands of layers, while achieving convincing performance. The key to ResNet's success is the adoption of the central idea of identity shortcut connection that allows skipping one or more layers, as shown in Figure 5(b).

Many computer vision applications other than image classification have been improved by taking advantage of the power of ResNet, such as object detection, face recognition and semantic segmentation. The residual units make the deep network easy

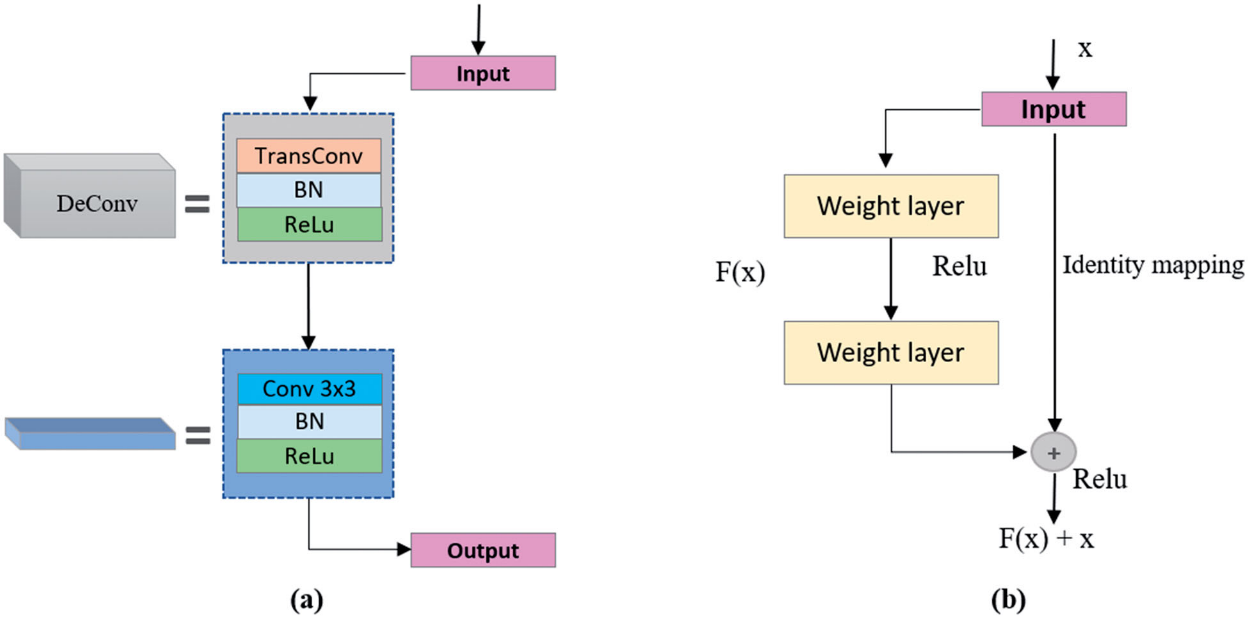


Figure 5. (a) DE-ResUnet decoder up sampling block with one convolution operation (b) illustrative scheme of a residual block [38].

to train and the skip connection in the network helps to propagate the information without degradation, improving the neural network design by decreasing the parameters with comparable or increased performance on the semantic segmentation task. Another advantage of ResNet is that it abandons Fully Connected and instead adopts Average pool, which avoids overfitting and significantly improves the accuracy. For all these reasons we used a pre-trained ResNet as the backbone of the architecture. The mathematical formula behind each residual block can be illustrated as follows:

$$\begin{aligned} y_l &= h(x_l) + F(x_l,\ w_l) \\ x_{l+1} &= f(y_l) \end{aligned} \tag{1}$$

Where x_l and x_{l+1} are the input and output of the l-th residual unit, $F(.)$ is the residual function, $f(y_l)$ is activation function and $h(x_l)$ is the identity mapping function.

2.2.3. Double Encoder residual UNet (DE-ResUnet)

In this section we present the novel neural architecture called Double Encoder Residual UNet (DE-ResUnet) for semantic segmentation of diabetic foot images by fusing thermal and RGB information. The overall architecture of DE_ResUnet is shown in Figure 6. Our network also adopts the encoder-decoder structure that proved its effectiveness in several semantic segmentation architectures [39]. DE-ResUnet is composed of three parts; a thermal encoder, an RGB encoder, which extract features from the thermal and RGB images respectively, and the decoder, which recover the representations to a pixel-wise categorisation.

This architecture is built based on the U-Net and ResNet networks, and inspired by multispectral networks such as FuseNet [25], MFNet [24] and RTFNet [23] which use both types of information; thermal and RGB images. The differences between our DE-ResUnet and the original U-Net consist of three main characteristics. First, the network contains two encoders instead of one, in order to extract features from two different sources, i.e., the thermal image and the RGB image at the same time. Second, the pre-trained ResNet [38] was employed as the feature extractor in each encoder. Finally, a small decoder with a single convolutional layer in each block was designed in order to reduce the parameters of the architecture as well as to speed up the inference.

DE-ResUnet contains two encoders that are used for the extraction of thermal and RGB features. These two identical encoders are built with ResNet units. Starting with an initial block that contains a convolutional layer, a BN layer (batch normalisation layer) and a ReLU activation layer (rectified linear unit activation layer). The two encoders are identical to each other except for the number of input channels in the first layer. Since ResNet is designed to use 3-channel RGB images, we changed the number of input channels of the convolutional layer in the initial block of the thermal encoder to 1. Just after the convolutional block, a max pooling layer of size 3 × 3 and stride of 2

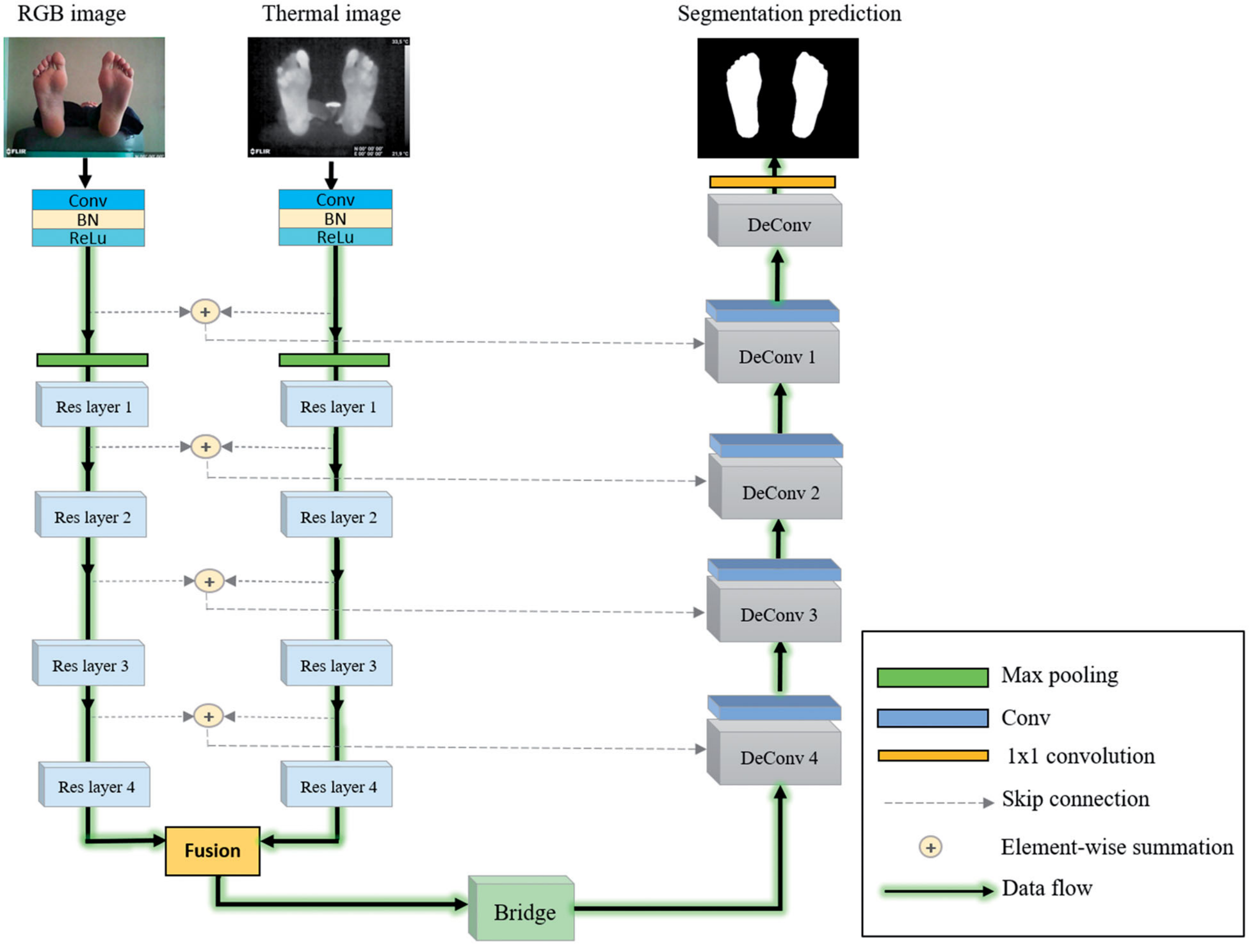


Figure 6. The architecture of the proposed DE-ResUnet.

followed by four residual layers are employed sequentially to progressively reduce the resolution and increase the number of channels of the feature maps.

In the later stage of the encoders, we merge the RGB and thermal information by applying a concatenation operation of the corresponding RGB and thermal feature maps. The shape of the feature map is not changed after the fusion operation. The output of the last fusion layer is taken as input to the decoder. A block of 3 × 3 convolutional operation acts as a bridge between the encoder and the decoder.

In the decoder part, the resolution of the feature map is gradually restored to that of the input images. DE-ResUnet decoder is designed symmetrically to the two encoders, which is illustrated in the architecture diagram (Figure 6). This allows the network to keep the U shape inspired by the original UNet network.

Just like in UNet, our decoder consists of up-sampling and concatenation followed by regular convolution operations. So, in our network each unit of the decoder consists of an up-sampling block followed by a convolution operation to produce dense feature maps as detailed in Figure 5(a). Since up-sampling is a sparse operation, we need a good prior from the previous stages to represent the localisation better. This is why we concatenate the higher resolution feature maps from the encoders with the up-sampled features. Illustrated in Figure 6 by skip connections, this operation preserves the shallow information and recovers fine details in the prediction. After the last level of the decoding path, a 1 × 1 convolution is used to project the multi-channel feature maps into the desired segmentation.

3. Results

In this section we evaluate our proposed DE-ResUnet architecture by comparing it with five state-of- the-art neural networks. We've tested all the networks on our database of diabetic foot images.

3.1. Implementation details

All experiments in this work were run on google colaboratory, often abbreviated to "Colab", a hosted service of jupyter notebooks pre-configured with

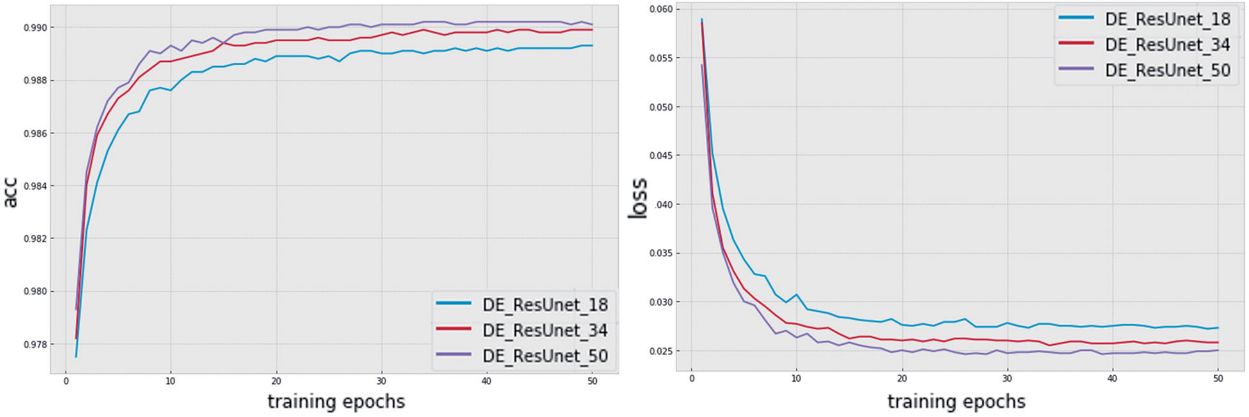


Figure 7. Comparison between 3 different ResNet as our Network Encoder. The graph on the left illustrates the validation accuracy of the 3 models, while the graph on the right represents their validation loss.

essential machine learning and artificial intelligence libraries, such as TensorFlow, PyTorch and Keras. It is suitable for machine learning, requires no configuration and allows access to computing resources, including GPUs. We implemented and executed our experiments in Python, using the PyTorch library and the used GPU was NVIDIA Tesla T4. We train the proposed architecture DE-ResUnet using stochastic gradient descent (SGD) optimiser with a momentum of 0.9 and weight decay of 0.0005. The initial learning rate is set to 0.01 and at each epoch it was multiplied by a decay rate of 0.94, in order to decrease gradually. The dataset is splitted into 50% for training, 25% for validation and the last 25% is used for testing. Moreover, the training set has been augmented to obtain more data by making modifications such as horizontal flips, rotations, blur filters and contrast changes in order to avoid overfitting. The total number of images in the training set is now 1393 images. Before each epoch the training images are randomly shuffled. Each network was trained until convergence, at which no further decrease in the loss is observed. The size of a mini-batch was set to 4 and each image was used only once in an epoch. We employed the cross-entropy loss as the objective function for backpropagation.

3.2. Evaluation metrics

To quantitatively evaluate the precision of our architecture, we adopted two evaluation metrics that are most commonly used in the field of semantic segmentation. The first one is the accuracy per class, and the second one is the IoU (Intersection over Union). The average values across all the classes for these two metrics are denoted as mAcc and mIoU, respectively.

$$mAcc = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \tag{2}$$

mAcc is the average value of accuracy on each class and is calculated using the Equation (2). While IoU is the intersection of inferred segmentation and ground truth, divided by the union, and mIoU is the average value of IoU of each class (Equation 3).

$$mIoU = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{3}$$

3.3. Comparative results

In this section we compare the DE-ResUnet architecture to RTFNet [23], MFNet [24], UNet [35], SegNet [37] and DeepLabv3 [40]. UNet, SegNet and DeepLabv3 architectures are designed for 3-channel RGB images. Therefore, to compare them to our architecture we train them with the 4-channel RGB-Thermal data obtained by stacking the 3-channel RGB data with the 1-channel thermal data. We modify the input layers of these three networks to adapt them to the 4 channels RGB-Thermal data. As described in section II of the paper, we opted to use pre-trained ResNet as the backbone of our architecture in the two encoders in order to take advantage of the strength of transfer learning, which reduces the training time of the model and improves its performance. Initially, it was necessary to choose which ResNet to use for our task. So, we tested our model using 3 different types, namely ResNet-18, 34 and 50. We settled for these three architectures to avoid the problem of overfitting. As more layers (101 or 152) can cause overfitting especially in our case where the database is limited. Figure 7 shows the superiority of ResNet-50 used as an encoder in our network compared to ResNet-18 and 34. In the

Table 1. Comparative results (%) on the test dataset. The results for the background class are less informative. (4c) denotes the use of stacked four channels (RGB and Thermal).

	Background		Foot		mAcc	mIoU
Methods	Acc	IoU	Acc	IoU		
SegNet (4c)	99,10	98,12	95,97	92,63	97,54	95,37
UNet (4c)	99,03	98,03	95,90	92,28	97,46	95,15
DeepLabv3(4c)	**99,55**	98,61	96,21	94,55	97,78	96,58
MFNet	99,13	98,25	96,40	93,12	97,76	95,69
RTFNet	99,11	98,67	96,41	94,60	97,76	96,63
DE-ResUnet (ours)	99,43	**98,72**	**97,39**	**95,20**	**98,41**	**97**

following, all the results represented in the tables and figures correspond to our network with ResNet-50 as a pre-trained backbone.

Table 1 displays the quantitative comparative results for the networks. As most of the pixels in our images correspond to the unlabelled background, the evaluation results for the Background class are similar across the different networks (~99%). They are less informative in our study because our region of interest is the foot, which should be segmented with more precision.

Figures 8 and 9, display some prediction examples of the data-fusion networks. In general, we can see that DE-ResUnet can robustly and accurately segment the feet under various acquisition difficulties and challenging images.

Several medical studies [41] have shown that certain areas of the diabetic foot are more prone to ulceration than others. These areas are called the high-risk areas for ulceration. Among these areas are the toes and the heels. Mainly in these zones there is more pressure while walking. It is therefore necessary that these areas with a high risk of ulceration should be very well segmented. Returning to the results, we notice that our network is more efficient than other architectures in detecting fine details and delimits with more precision the toes and heels of the sole.

Based on the quantitative results we can see that the accuracy values are close between all architectures. Despite the fact that the accuracy is close, we noticed that the strong point of DE-ResUnet is the detection of the toes and heels of the feet which represent the regions at risk of ulceration, and these details of the feet are not detected by other architectures, despite their high accuracy value. This can be explained by the use of skip connections that preserve the finest details in addition to the use of residual blocks.

Another example is shown in Figure 10 in which we see DE-ResUnet outperforming RTFNet in the segmentation of the big toe that was not detected by RTFNet, despite their close IoU values.

We measure the inference speed of the networks with an NVIDIA Tesla T4 graphics card. From Table 2 we can see that the average time cost of our network is lower than RTFNet. While UNet remains the most rapid with an average time cost of 3.35 ms.

3.4. Ablation study

The main objective of this work is to demonstrate that the use of both thermal and colour information will increase the semantic segmentation accuracy of diabetic feet. For this purpose, we carried out an ablation study, where we compare our multimodal network DE-ResUnet with two other variants using only thermal encoder or RGB encoder. Therefore, we test DE-ResUnet by removing the RGB encoder to see the benefits brought by using the colour information. We term the variant T-ResUnet (Thermal-ResUnet). Similarly, we remove the thermal encoder from DE-ResUnet to see how the network performs when only given the RGB information. This variant has no Thermal Encoder, so we call it R-ResUnet (RGB-ResUnet).

Figure 11 illustrates the ablation study results. By comparing the results of R-ResUnet and T-ResUnet, we find that R-ResUnet generally gives better performance, but they are both inferior to our DE-ResUnet. This proves that data fusion is an effective approach to increase performance. And the RGB information contributes remarkably to the data fusion. We could find that only using the RGB information gives better results in comparison with using only thermal information. This is expected because RGB images are more informative than thermal ones.

4. Discussion

Semantic segmentation of medical images is a challenging problem and a rapidly growing research topic. This field in particular requires very high exactitude and precision. Thermal images are characterised by the difficulty of detecting tissue boundaries and, therefore, manual segmentation of these images is strongly dependent on the observer and is prone to errors. However, sometimes we deal with thermal images that are difficult to segment even by an expert, as shown in the examples in Figure 1. In these cases, a better strategy is needed. This is the main reason why we thought of integrating the colour image in our segmentation process. Certainly, RGB visual images provide detailed morphological information and a clear delineation of the feet, unlike IR images.

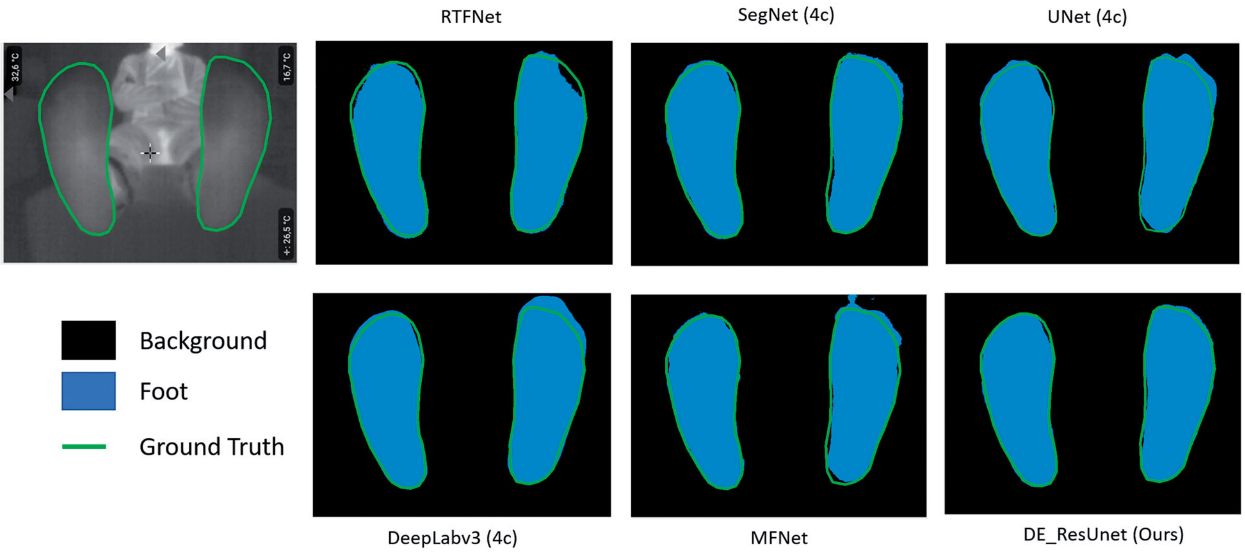


Figure 8. Illustrative example showing the input thermal image and the predictions achieved by all the networks. Ground truth mask is represented by the green outline.

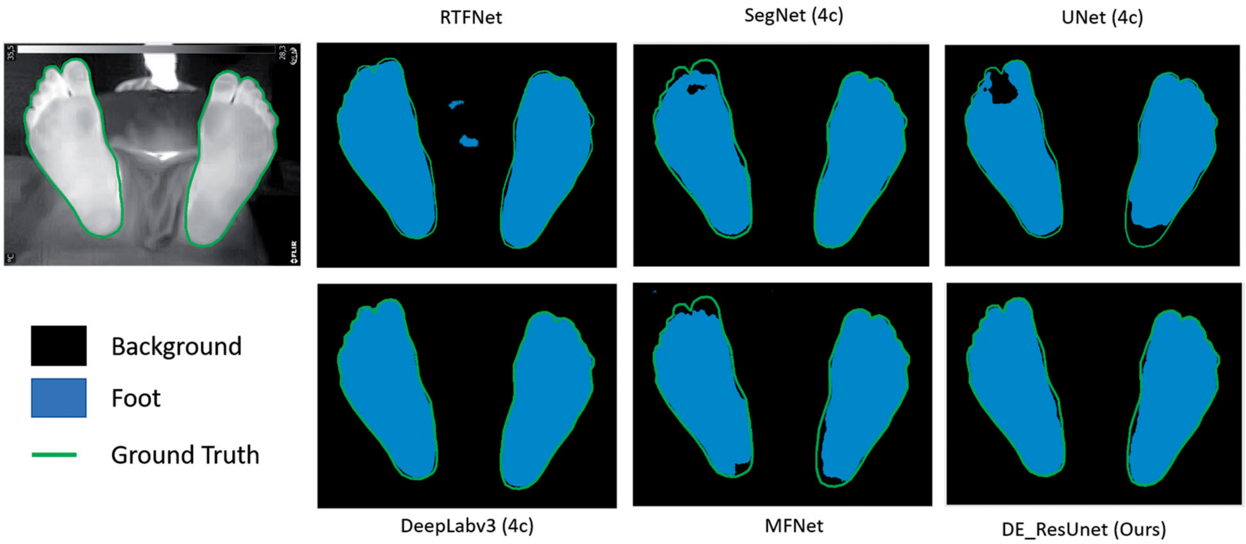


Figure 9. Illustrative example showing the robustness of DE-ResUnet and DeepLabv3 in the precise delineation of fine details of toes and heels of the feet.

But this does not prevent that in some situations even the colour images are affected by the light conditions in the image acquisition room as well as by individual characteristics such as skin tone, age, gender and body parts studied. Other factors affect background colours, shadows and motion blur etc. This creates a certain complicity between the thermal information which is not affected by light or skin colour changes and the RGB information which gives more precise foot contours.

In this study the data set is characterised by a variety in the subjects' skin tones, as well as the lighting conditions and the colours and objects in the background. This increased the difficulty of our task and makes our solution robust and adaptable to different situations. From the beginning, our goal has been to propose a segmentation approach that does not require a constrained and well-defined acquisition protocol such as homogenisation of the background by polyurethane foam or cold towel as it was the case in some works [16, 17]. All our images are acquired freehandedly with a smartphone and its linked thermal camera. Moreover, this approach did not require any preprocessing of the training data set. Unlike [16, 20, 21] who prepared their images by partitioning them to use a single foot and a single orientation instead of processing the whole image with both right and left feet. In this paper we proposed a

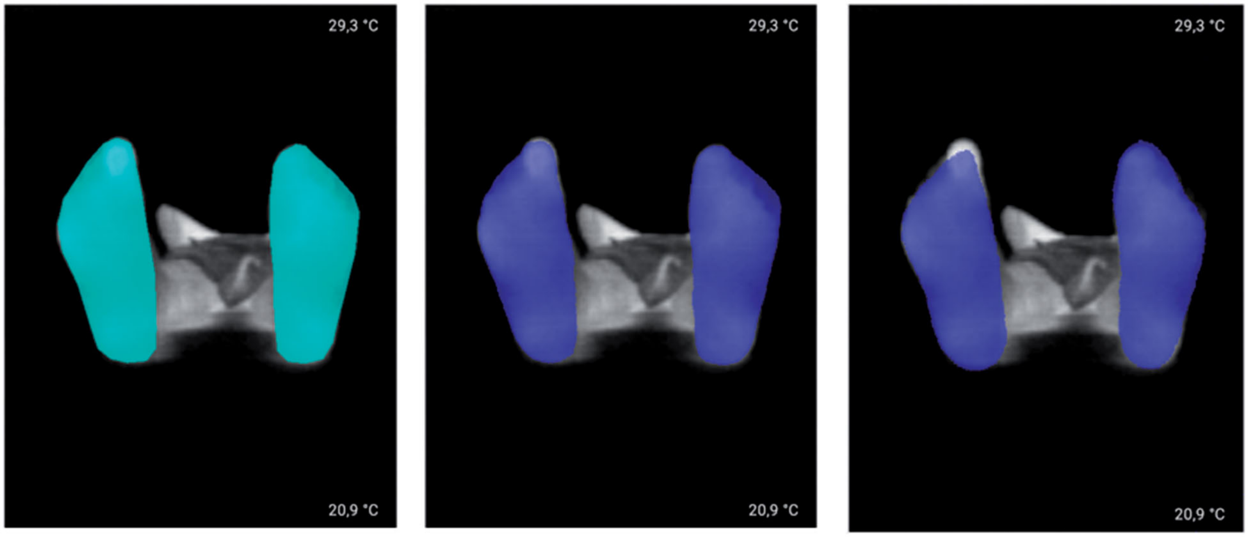


Figure 10. Example in which DE-ResUnet is able to segment the foot with the toes in comparison with RTFNet. Left image represents thermal image with ground truth mask overlaid. The image in the middle represents DE-ResUnet prediction and right image represents RTFNet [24] segmentation.

Table 2. The Inference speed for each network. ms represents the time cost in millisecond and the FPS represents Frame-Per-Second.

	Tesla T4 GPU	
Methods	Ms	FPS
SegNet (4c)	4.75 ms	210.53
UNet(4c)	**3.35 ms**	**298.91**
DeepLabv3(4c)	11.6 ms	86.08
MFNet	6.76 ms	147.89
RTFNet	15.78 ms	63.36
DE-ResUnet	11.89 ms	84.08

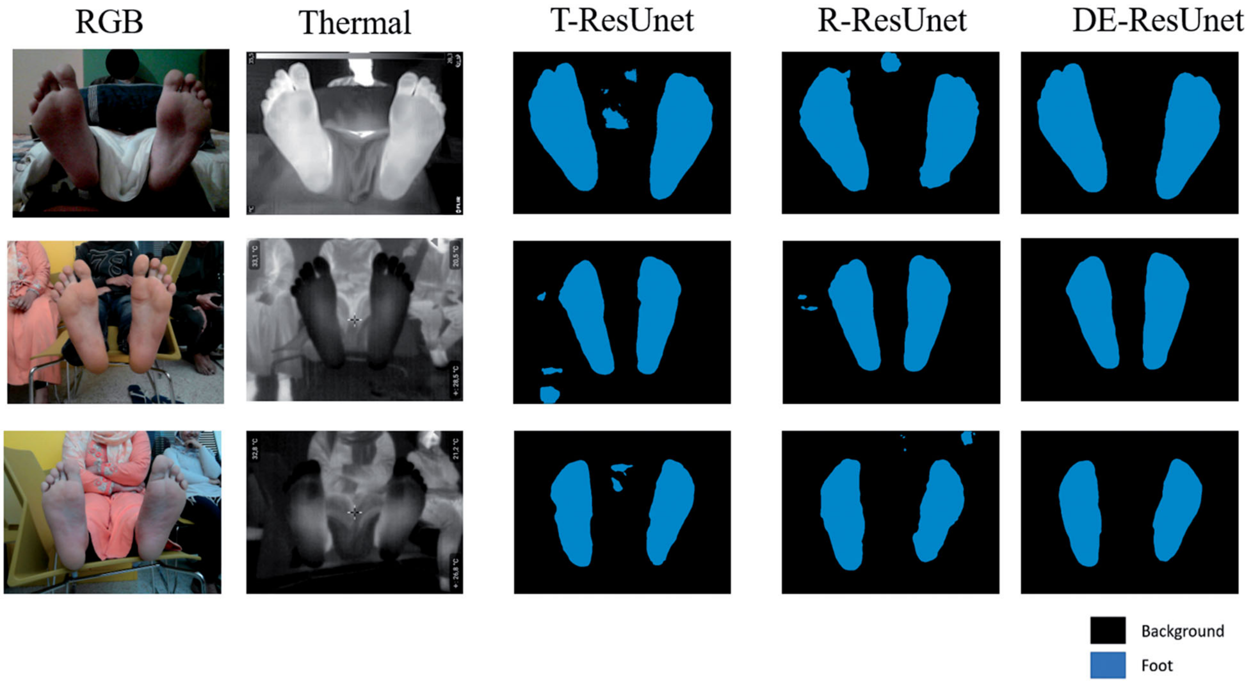


Figure 11. Illustrative example showing the input images (RGB and thermal) and the final prediction of DE-ResUnet and its two variants T-ResUnet and R-ResUnet. T-ResUnet is the network with thermal encoder only and R-ResUnet contain RGB encoder only. Whereas DE-ResUnet is our Double encoder proposed approach.

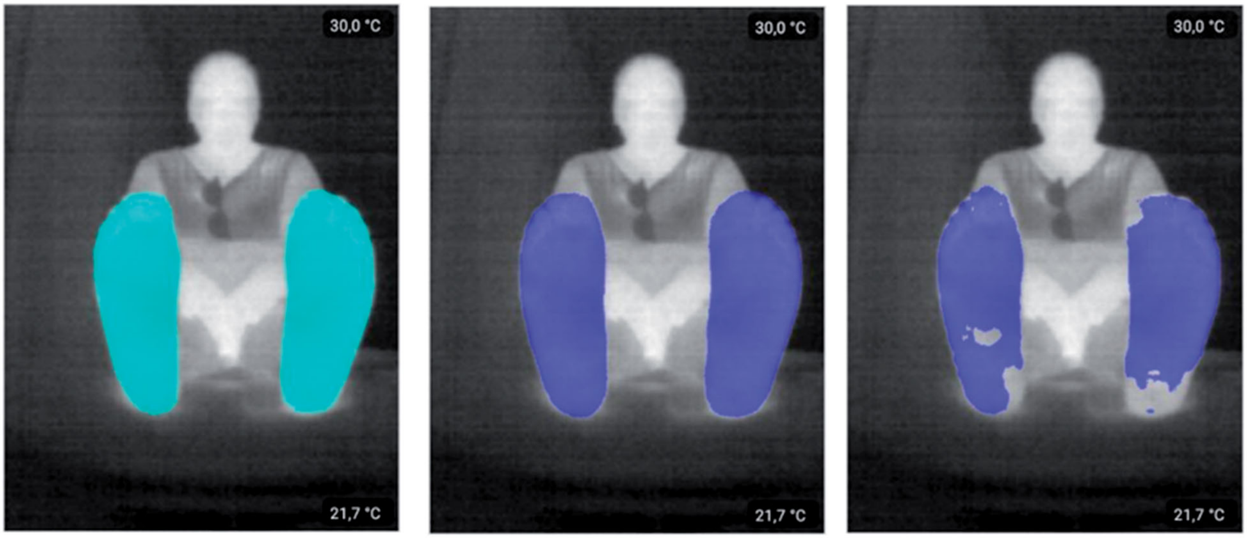


Figure 12. Illustrative example in which DE-ResUnet accurately segments the heels of the feet in comparison with a state-of-the-art method "MFNet". left image is thermal image with ground truth. our network prediction is presented in the middle image and right image is the result of MFNet [24].

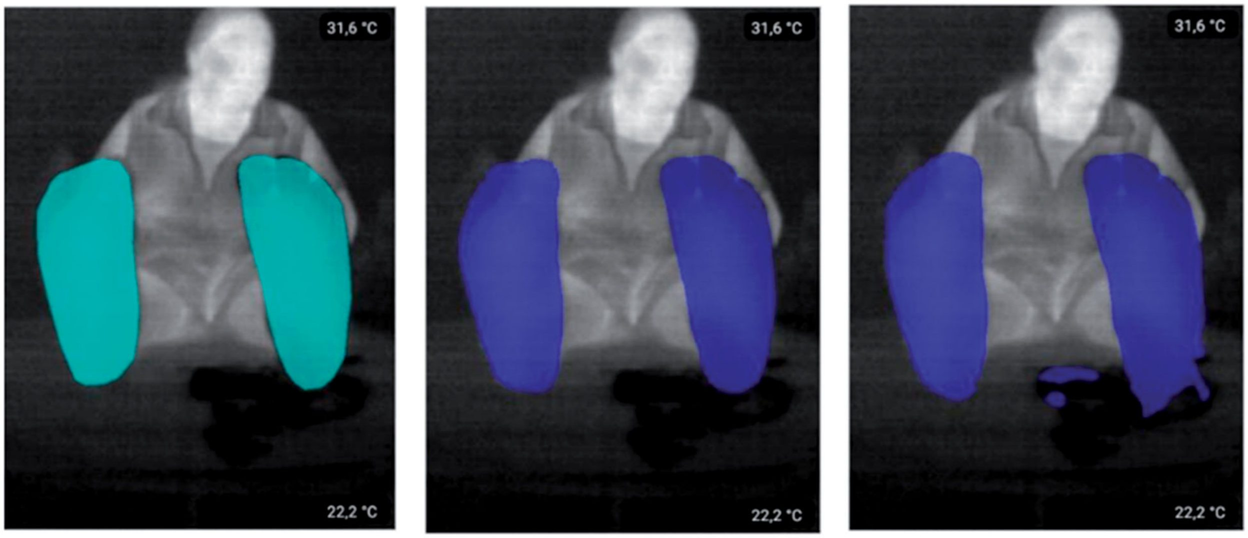


Figure 13. Example of a difficult image to segment due to the ambiguity between the cold heel and the background. DE-ResUnet performs better than the original UNet [36] and can segment the heel correctly. Left image corresponds to thermal image with overlaid mask, in the middle DE-ResUnet and the right one corresponds to UNet (4c).

segmentation approach that combines the strength of both UNet and ResNet architectures on one hand and on the other hand merges both thermal and RGB image types. The goal is to improve the accuracy and performance of segmentation of the foot. We compared our architecture to other state-of-the-art approaches, such as SegNet, UNet, DeepLabv3, RTFNet and MFNet to evaluate the performance of this architecture. And according to the quantitative evaluation, we deduced that DE-ResUnet showed the best performance.

Although the accuracy values of all the approaches are close, we noticed a very important point about DE-ResUnet, in several test images we can see that our network is able to detect parts of the foot such as toes and heels with a very high precision compared to other architectures, Figures (10,12–14). This is a strong point of DE-ResUnet that could be an added value especially since toes and heels are the most susceptible regions to be ulcerated. This can be explained by the use of skip connections that preserve the finest details in addition to the use of residual blocks that make the deep network easy to train and helps to propagate the information without degradation, improving the neural network design by decreasing the parameters with comparable or increased

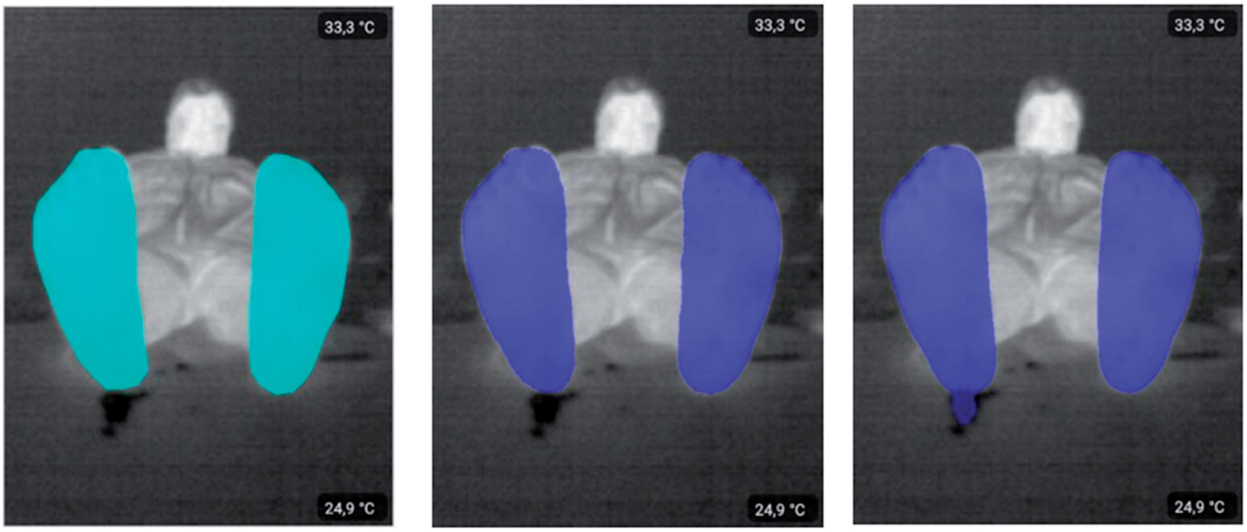


Figure 14. Example where DE_ResUnet outperforms DeepLabv3 [41]. Left image corresponds to thermal image with overlaid mask, in the middle DE-ResUnet and the right one corresponds to DeepLabv3 (4c).

performance on the semantic segmentation task. We also performed an ablation study, in which our objective was to prove that the use of both thermal and RGB images probably improves the accuracy of semantic segmentation of the foot. For this purpose, we compared our multimodal network DE-ResUnet with its two other variants using only a thermal encoder or RGB encoder.

By comparing the results of R-ResUnet and T-ResUnet, we find that R-ResUnet generally gives better performance, but they are both inferior to our DE-ResUnet. This proves that the data fusion is an effective approach to increase the performance. And the RGB information contributes remarkably to data fusion.

One of the difficulties we encountered during our study is that we could not make more image acquisitions in hospitals for a larger number of patients, due to the health situation related to covid 19 worldwide. Therefore, we were initially limited to the data set that we collected before the pandemic. Our database consists of images corresponding to diabetic patients and also healthy people.

In addition, our work is ongoing, on one hand acquiring more images and increasing the size of our multimodal database that will be in the future shared in public for scientific research. On the other hand, we are working on a medical study in which we'll require healthy subjects as well as pathological subjects who are affected by diabetic disorders. The images of the healthy subjects will be used to represent the normality of the temperature patterns. In contrast, the images of the pathological subjects will aim to establish a relationship between the temperature patterns and the underlying diabetic condition.

5. Conclusion

This paper proposed a new multispectral segmentation architecture (DE-ResUnet) compared with five other Deep Learning methods for the segmentation of diabetic foot thermal images. The comparison performed on our test database showed the superiority of DE-ResUnet with a mIoU of 97%. This proposed method is robust, provides good results and has demonstrated its effectiveness in segmenting both feet simultaneously with high precision. In addition, no constraining isolation system is required, the images are taken freehandedly with a smartphone equipped with a dedicated thermal camera, and the processing is fully automatic. As a perspective, a future study is planned to provide a more rigorous test of the system. This will comprise a larger study population and will also include participants with partial amputations. Furthermore, it would be interesting to show that this suggested friendly and automated segmentation method combined with other medical analyses could help doctors in hospitals or in medical centres for a better diagnosis of DF disorders.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Actions (grant agreement No #777661), aiming to develop smartphone applications for prevention and supervision of diabetic foot ulcers.

ORCID

Doha Bouallal http://orcid.org/0000-0001-6153-3570
Hassan Douzi http://orcid.org/0000-0002-2756-1399

References

[1] Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. Phys Ther. 2008;88(11):1254–1264.

[2] Harding JL, Pavkov ME, Magliano DJ, et al. Global trends in diabetes complications: a review of current evidence. Diabetologia. 2019;62(1):3–16.

[3] van Netten JJ, Bus AS, Apelqvist J, et al. Definitions and criteria for diabetic foot disease. Diabetes Metab Res Rev. 2020;36(S1):e3268.

[4] Mishra SC, Chhatbar KC, Kashikar A, et al. Diabetic foot. BMJ. 2017;359:j5064.

[5] Armstrong DG, Holtz-Neiderer K, Wendel C, et al. Skin temperature monitoring reduces the risk for diabetic foot ulceration in high-risk patients. Am J Med. 2007;120(12):1042–6.

[6] Adam M, Ng EYK, Tan JH, et al. Computer aided diagnosis of diabetic foot using infrared thermography: a review. Comput Biol Med. 2017;91:326–336.

[7] Muller AC, Narayanan S. Cognitively-engineered multisensor image fusion for military applications. Inf Fusion. 2009;10(2):137–149.

[8] Okada T. Thermography of asteroid and future applications in space missions. Appl Sci. 2020;10(6):2158.

[9] Younsi M, Diaf M, Siarry P. Automatic multiple moving humans detection and tracking in image sequences taken from a stationary thermal infrared camera. Expert Syst Appl. 2020;146:113171.

[10] Jiang LJ, Ng EYK, Yeo ACB, et al. A perspective on medical infrared imaging. J Med Eng Technol. 2005;29(6):257–267.

[11] Cajacuri LAV. Early diagnostic of diabetic foot using thermal images. 2013. p. 140.

[12] Saxena A, Ng EYK, Lim ST. Infrared (IR) thermography as a potential screening modality for carotid artery stenosis. Comput Biol Med. 2019;113:103419.

[13] Casas-Alvarado A, Mota-Rojas D, Hernández-Ávalos I, et al. Advances in infrared thermography: surgical aspects, vascular changes, and pain monitoring in veterinary medicine. J Therm Biol. 2020;92:102664.

[14] Singh D, Singh AK. Role of image thermography in early breast cancer detection- past, present and future. Comput Methods Progr Biomed. 2020;183:105074.

[15] Roback K. An overview of temperature monitoring devices for early detection of diabetic foot disorders. Expert Rev Med Dev. 2010;7(5):711–8.

[16] Vilcahuaman L, et al. Automatic analysis of plantar foot thermal images in at-risk type II diabetes by using an infrared camera. In: D. A. Jaffray, Éd. *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*, Cham: Springer International Publishing, 2015. p. 228–231.

[17] Fraiwan L, AlKhodari M, Ninan J, et al. Diabetic foot ulcer mobile detection system using smart phone thermal camera: a feasibility study. Biomed Eng OnLine. 2017;16(1):117.

[18] Liu C, van der Heijden F, Klein ME, et al. Infrared dermal thermography on diabetic feet soles to predict ulcerations: a case study, San Francisco, California, USA, mars 2013. p. 85720N.

[19] Kaabouch N, Chen Y, Hu W-C, et al. Early detection of foot ulcers through asymmetry analysis, Lake Buena Vista, FL, févr2009. p. 72621L.

[20] Bougrine A, Harba R, Canals R, et al. On the segmentation of plantar foot thermal images with deep learning. In: *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, 2019. p. 1–5.

[21] Bougrine A, Harba R, Canals R, et al. A joint snake and atlas-based segmentation of plantar foot thermal images. In: *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Montreal, QC, 2017. p. 1–6.

[22] Bouallal D, et al. Segmentation of plantar foot thermal images: application to diabetic foot diagnosis. In: *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Niterói, Brazil, 2020, p. 116–121.

[23] Sun Y, Zuo W, Liu M. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. IEEE Robot Autom Lett. 2019;4(3):2576–2583.

[24] Ha Q, Watanabe K, Karasawa T, et al. MFNet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, BC, 2017. p. 5108–5115.

[25] Hazirbas C, Ma L, Domokos C, et al. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In S.-H. Lai, V. Lepetit, K. Nishino, et Y. Sato, Éd. Computer vision – ACCV, Cham: Springer International Publishing, 2017. p. 213–228.

[26] Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-Net. IEEE Geosci Remote Sensing Lett. 2018;15(5):749–753.

[27] Diakogiannis FI, Waldner F, Caccetta P, et al. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J Photogramm Remote Sens. 2020;162:94–114.

[28] FLIR ONE Pro Thermal Imaging Camera for Smartphones | FLIR Systems. https://www.flir.com/products/flir-one-pro/. (consulté le mars 29, 2021.

[29] What is MSX®? https://www.flir.com/discover/professional-tools/what-is-msx/. (consulté le mars 29, 2021.

[30] Chan AW, MacFarlane IA, Bowsher DR. Contact thermography of painful diabetic neuropathic foot. Diabetes Care. 1991;14(10):918–922.

[31] Nagase T, et al. Variations of plantar thermographic patterns in normal controls and non-ulcer diabetic patients: novel classification using angiosome concept. J Plast Reconstr Aesthet Surg. 2011;64(7):860–6.

[32] Bastyr EJ, Price KL, Bril V. Development and validity testing of the neuropathy total symptom score-6: questionnaire for the study of sensory symptoms of diabetic peripheral neuropathy. Clin Ther. 2005;27(8): 1278–1294.

[33] Label images for computer vision applications - MATLAB. https://www.mathworks.com/help/vision/ref/imagelabeler-app.html. (consulté le avr. 04, 2021.

[34] Tan C, Sun F, Kong T, et al. A survey on deep transfer learning. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, et I. Maglogiannis, Éd. Artificial neural networks and machine learning – ICANN. Cham: Springer International Publishing, 2018. p. 270–279.

[35] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation *ArXiv150504597 Cs*, mai 2015. Consulté le: janv. 10, 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/1505.04597.

[36] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation *ArXiv160506211 Cs*, mai 2016. Consulté le: nov. 20, 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/1605.06211.

[37] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation *ArXiv151100561 Cs*, oct2016. Consulté le: janv. 28, 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/1511.00561.

[38] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition *ArXiv151203385 Cs*, déc2015. Consulté le: janv. 28, 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/1512.03385.

[39] Garcia-Garcia A, Orts-Escolano S, Oprea S, et al. A Review on Deep Learning Techniques Applied to Semantic Segmentation *ArXiv170406857 Cs*, avr2017. Consulté le: mars 29, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1704.06857.

[40] Chen L-C, Papandreou G, Schroff F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation *ArXiv170605587 Cs*, déc2017. Consulté le: juin 16, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1706.05587.

[41] Hernandez-Contreras D, Peregrina-Barreto H, Rangel-Magdaleno J, et al. Narrative review: diabetic foot and infrared thermography. Infrared Phys Technol. 2016; 78:105–117.