DMSNet : Dual multi scale networks for diabetic foot ulcer segmentation using contrastive learning

Akshat Keshav Dhamale
Dept. of Computer Science (AI&ML)
Vellore Institute of technology (VIT)
Chennai, India
akshatkeshav.dhamale2020@vitstudent.ac.in

***Abstract*—This paper presents DMSNet, an advanced model for the segmentation of diabetic foot ulcers. DMSNet employs a dual multi-scale architecture, combining the computational efficiency of EfficientNet B5 with the contextual understanding of the Pyramid Vision Transformer (PVT). Integration of a multi-scale module in both encoders enhances the model's capacity to capture intricate details across various resolutions, enabling precise delineation of complex ulcer boundaries. Notably, DMSNet incorporates contrastive learning with a novel pixel-wise contrastive loss function during training, contributing to heightened segmentation accuracy and improved generalization capabilities. The model's performance is demonstrated through experimental evaluation on the MICCAI 2021 Diabetic Foot Ulcer Segmentation Challenge dataset, where it achieves a Dice score of 83.51 and an IoU score of 75.65. These metrics underscore DMSNet's proficiency in accurately delineating diabetic foot ulcers, positioning it as a valuable tool in clinical settings for enhanced diagnosis and treatment planning. By advancing segmentation accuracy through innovative architectural design, multi-scale modules, and contrastive learning techniques, DMSNet represents a significant stride in the field, with potential implications for improved patient care and outcomes. The code for following work can be found in https://github.com/AkshatDhamale/Diabetic-Foot-Ulcer-segmentation**

***Index Terms*—Diabetic foot ulcer, Contrastive learning, multi-scale architecture, encoder-decoder based model, medical image segmentation**

I. Introduction

Diabetic foot ulcers, prevalent complications of diabetes mellitus, stand as a formidable threat to global public health, exacting a toll on individuals' well-being and contributing to escalating healthcare costs. A staggering 15% of individuals [1] [2] with diabetes are projected to encounter a foot ulcer during their lifetime, amplifying the gravity of this complication on a global scale. Diabetic foot ulcers are a primary precursor to lower limb amputations, with a substantial majority of amputations attributed to complications stemming from these ulcers. This grim reality accentuates the urgent need for effective preventive measures, precise diagnostics, and timely intervention strategies.

Automated solutions in medical image analysis have surfaced as pivotal instruments in enhancing the accuracy and efficiency of diagnosing diabetic foot ulcers. These automated solutions expedite the diagnostic process, offering the potential for early detection and facilitating proactive medical intervention. Significantly, improved segmentation of diabetic foot ulcers plays a crucial role by providing clinicians with intricate spatial information for informed decision-making. By automating the segmentation process, these solutions contribute not only to a more efficient allocation of healthcare resources but also to improved patient outcomes.

Amid the imperative for automated segmentation solutions, the existing landscape of diabetic foot ulcer segmentation models is both extensive and diverse. Various methodologies have been explored, ranging from traditional image processing techniques to sophisticated deep learning approaches. Prominent models in this domain include U-Net [11], FCN [12], DeepLabV3 [13], PSPNet [14], and others. The diversity of these models reflects the ongoing exploration of effective methodologies to tackle the nuanced challenges in diabetic foot ulcer segmentation. Furthermore, studies such as those by Wang et al. (2020), Liu et al. (2019), and Zhang et al. (2018) have contributed valuable insights, pushing the boundaries of segmentation accuracy and clinical utility.Highlights of this work are mentioned below :

- We propose a novel architecture - DMSNet which is constructed considering to efficiently combine convolution and attention features in a multi-scale fashion.
- First contrastive learning based approach applied on contrastive learning as well as on a dual encoder-decoder based model.
- Boundary learning is integrated through convolution based network with attention based mask learning. Appropriate loss functions to account for both of them are used.
- The model demonstrates potential application in medical image segmentation and discusses effectiveness of contrastive learning in a supervised approach.

II. Literature survey

In the realm of diabetic foot ulcer segmentation, numerous models have been developed, each bringing unique strengths to the challenge. U-Net, a pioneering architecture introduced by [11], employs a U-shaped design with skip connections for effective global and local feature capture, making it highly relevant in discerning intricate ulcer boundaries. Fully Convolutional Networks (FCN), proposed by [12], revolutionized semantic segmentation by introducing fully convolutional layers, enabling end-to-end pixel-wise predictions, proving beneficial in accurately delineating ulcer regions. DeepLabV3 [13] incorporates atrous convolution and a dilated spatial pyramid pooling module, facilitating effective feature extraction at multiple scales and handling varied lesion sizes and shapes. The Pyramid Scene Parsing Network (PSPNet) by [14]. (2017) utilizes a pyramid pooling module to capture multi-scale contextual information, providing a holistic view of ulcers for precise delineation and comprehensive analysis. Meanwhile, Mask R-CNN [16], renowned for instance segmentation, integrates region-based convolutional neural networks for predicting both class labels and segmentation masks, proving valuable in analyzing multiple ulcers within a single image.

Attention U-Net [6] builds upon the success of U-Net by incorporating attention gates, selectively highlighting relevant features during segmentation and enhancing accuracy in diabetic foot ulcer analysis. SegNet [4], designed for semantic segmentation, employs a fully convolutional network with an encoder-decoder structure, offering accurate segmentation particularly in resource-constrained environments. Capsule Networks [15], introducing capsule layers, aim to overcome limitations in traditional neural networks and offer improved hierarchical feature representation, potentially capturing nuanced ulcer characteristics more effectively. UNet++ [7] extends the U-Net architecture by incorporating dense skip pathways for richer feature propagation, crucial for handling complex anatomical structures and varying ulcer appearances. LinkNet [8], with its symmetric encoder-decoder structure, focuses on maintaining information flow across different scales, contributing to robust segmentation performance in diabetic foot ulcer analysis. Lastly, RefineNet [17] addresses challenges in capturing fine details by employing chained residual pooling modules, progressively refining segmentation predictions and enhancing the model's ability to delineate intricate ulcer boundaries. These diverse models collectively contribute to the ongoing advancements in diabetic foot ulcer segmentation, pushing the boundaries of precision and clinical utility.

Recent trends also showcase the integration of both attention and dual encoder-decoder features in hybrid architectures. Models like Dual Attention Network [5] combine dual pathways for comprehensive feature extraction with attention mechanisms for selective information utilization. This amalgamation aims to exploit the benefits of both approaches, ensuring not only the capture of global and local features but also their strategic emphasis during segmentation. In the context of diabetic foot ulcers, hybrid architectures represent a promising avenue, offering a nuanced understanding of ulcer characteristics and spatial relationships crucial for accurate segmentation.

III. Proposed method

The foundation of DMSNet integrates two key components: EfficientNet B5 [9] as a convolutional encoder and the Pyramid Vision Transformer (PVT) [10] as an attention-based encoder. This strategic fusion aims to harness the strengths of both computational efficiency and contextual understanding in capturing intricate details within diabetic foot ulcer regions.

A. *Dual Multi-scale architecture*

DMSNet incorporates a multi-scale module within both the EfficientNet B5 and PVT encoders. This integration allows the model to capture features at various resolutions, addressing the diverse spatial intricacies present in diabetic foot ulcer images. The dual multi-scale architecture enhances the model's capacity to discern fine details, contributing to improved segmentation accuracy and the delineation of complex ulcer boundaries.

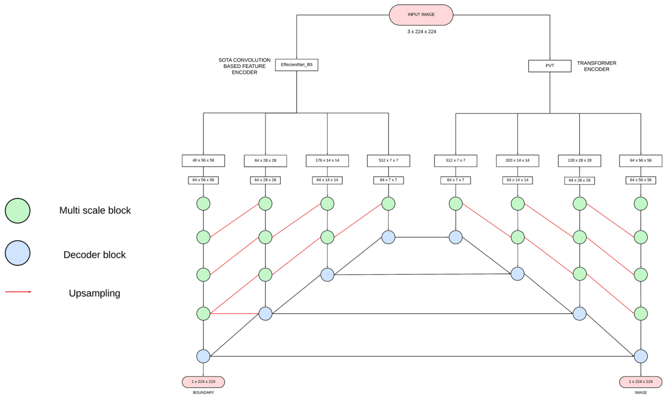


Fig. 1. DMSNet architecture

This model is largely divided into 3 sections - dual encoder, multi-scale block and dual decoder. Dual encoder consists of 4 set of features from effecientNet and pyramid vision transformer (PVT). The deepest features of both encoder are scaled and added to the next densest feature. This structure keeps repeating to the shallowest levels. All the encoder features are convoluted to 64 channels so that additional convolutions need not be incorporated.

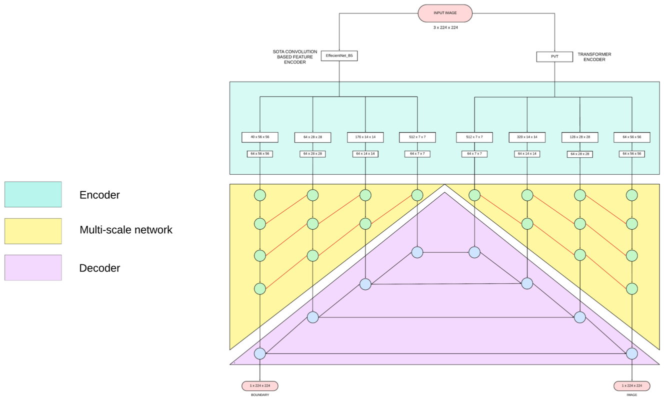


Fig. 2. Labeled architecture of DMSNet

B. *Dual Decoder output*

A distinctive feature of DMSNet is its dual decoder output, which comprises both boundary and segmentation masks. The dual output provides a comprehensive representation of the diabetic foot ulcer region, offering valuable insights into both the extent of the ulcer and its boundaries. This dual information output enhances the clinical interpretability of the segmentation results, supporting healthcare professionals in making informed decisions.

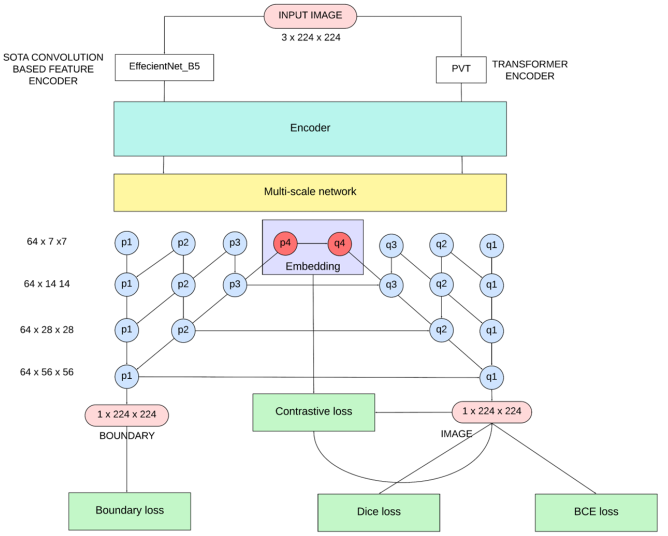


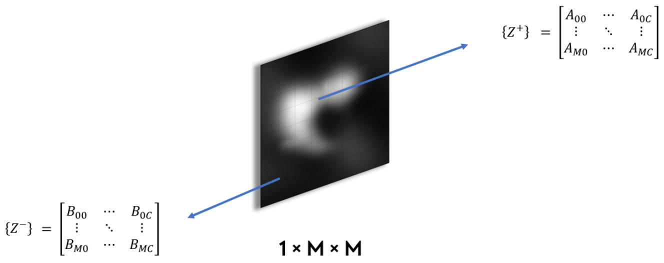
Fig. 3. Labeled architecture of DMSNet

Dual decoder outputs are mapped in such a way that attention based network PVT is trained on groud truth masks with dice and cross entropy loss and convolution based network EfficientNet is trained on boundary segments using boundary loss. Boundaries of ground truths are obtained using mophological operations of dilation and erosion. Ground truths are first eroded over a factor of δ pixels and then subtracted from a dilation of δ pixels. Here we keep δ as 8 therefore getting a boundary 16 pixels wide.

C. Contrastive learning

DMSNet adopts contrastive learning as the training methodology, introducing a novel pixel-wise contrastive loss function. This innovative approach involves training the model to distinguish between positive and negative pairs of image patches, facilitating improved feature learning and embedding. By incorporating contrastive learning, DMSNet aims to enhance segmentation accuracy and generalization capabilities, particularly in scenarios where traditional supervised learning approaches may face limitations.

For an image z, Contrastive loss is defined in equation (1) where z_i represents i'th pixel. z^+ represents similar region of interest whereas z_- depicts similar backgrond region. τ acts as a constant for determing contrast levels. Here it is set to 1. It is a logarithmic function which finds positive (mask) and negative (background) regions using contrast between these areas. z is generally taken to be the densest feature of encoder and is compared with ground truth mask. This pixel wise contrastive loss approach is adopted primarily from [3].



$\{Z^+\} = \begin{bmatrix} A_{00} & \cdots & A_{0C} \\ \vdots & \ddots & \vdots \\ A_{M0} & \cdots & A_{MC} \end{bmatrix}$
$\{Z^-\} = \begin{bmatrix} B_{00} & \cdots & B_{0C} \\ \vdots & \ddots & \vdots \\ B_{M0} & \cdots & B_{MC} \end{bmatrix}$
1 × M × M

Fig. 4. Contrastive learning for a binary image

$$\mathcal{L}_{con}(z_i, \left\{z^+\right\}, \left\{z^-\right\}) = -\frac{1}{|\{z^+\}|} log \frac{\sum_{z_j \in \{z^+\}} exp(sim(z_i, zj)/\tau)}{\sum_{z_k \in \{z^+\} \cup \{z^-\}} exp(sim(z_i, zk)/\tau)} \tag{1}$$

By encouraging the model to discern subtle differences between pixels in positive and negative pairs, it becomes adept at understanding the contextual significance of each pixel within the medical image. This approach not only improves segmentation accuracy but also enables the model to adapt and generalize well to variations in medical imaging data, a crucial aspect given the inherent diversity in medical image datasets. Pixel-wise contrastive learning aligns with the broader trend in the field of leveraging self-supervised learning techniques for medical image analysis. The ability to learn from the intrinsic structure of the data without the reliance on extensive annotations is particularly advantageous in medical imaging scenarios where obtaining labeled datasets can be challenging and time-consuming.

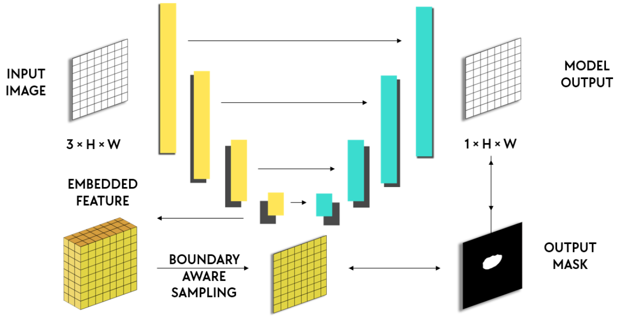


Fig. 5. Contrastive learning for an encoder-decoder based network

D. Loss function

Total loss function (equation 2) consists of weighted dice loss and weighted binary cross entropy loss with boundary and pixel wise contrastive loss. The combination of weighted Dice Loss and weighted Binary Cross Entropy (BCE) Loss in segmentation models provides advantages in handling class imbalances, offering customized importance for different classes based on their significance in the task. By assigning higher weights to underrepresented classes, the model becomes

TABLE I
COMPARITIVE RESULTS OF DMSNET WITH DIFFERENT MODELS

Model	**mDice**	**mIOU**	**Accuracy**	**MAE**	**WFM**	**SM**	**EM**
UNET + PCL	0.8072	0.7158	0.9961	0.0039	0.8681	0.8244	0.9145
UNET + RES2NET	0.7905	0.6986	0.9966	0.0034	0.8432	0.8157	0.9083
Unet++	0.7547	0.6552	0.9947	0.0053	0.8097	0.8062	0.8993
MSNet	0.8058	0.7149	0.9968	0.0032	0.8736	0.8346	0.9387
PVT	0.7896	0. 6893	0.9945	0.0045	0.8408	0.7968	0.9001
DMSNet	0.8171	0.7321	0.9969	0.0031	0.8595	0.8422	0.9426
DMSNet + PCL	0.8335	0.7565	0.9979	0.0021	0.8684	0.8569	0.9483

TABLE II
ABLATION STUDY RESULTS OF DUAL ENCODER SYSTEM

Model	Dice	IOU	Accuracy	MAE	AvgF	WFM	SM	EM
PVT + Resnet	0.7896	0. 6893	0.9945	0.0045	0.8012	0.8408	0.7968	0.9001
PVT + Res2Net	0.7967	0. 7036	0.9956	0.0040	0.8125	0.8498	0.8034	0.9120
Swin + EfficientNet	0.7836	0. 6783	0.9942	0.0047	0.7986	0.8386	0.7864	0.8867
PVT + EfficientNet	0.8335	0.7565	0.9979	0.0021	0.8374	0.8684	0.8569	0.9483

more sensitive to their spatial details, as emphasized by the Dice Loss, and pixel-wise information, as emphasized by the BCE Loss. This dual approach enhances the robustness of the model, allowing it to learn complex patterns and converge more effectively during training. The combined use of these losses improves the model's generalization ability and ensures its adaptability to scenarios with imbalanced datasets, contributing to more accurate and context-aware segmentation in medical imaging or similar applications.

$$\mathcal{L}_{total} = \alpha_1 \cdot \mathcal{L}_{BCE} + \alpha_2 \cdot \mathcal{L}_{Dice} + \mathcal{L}_{bound} + \mathcal{L}_{con} \qquad (2)$$

Here we keep the constant values α_1 and α_2 as 0.5 for both. Here the boundary loss is a modified version of focal-tversky loss function.

IV. RESULTS

In this section, we test the performance of our method to Diabetic Foot ulcer segmentation dataset proposed in MICCAI 2021. To provide deeper insight into the model performance, we further introduce four other metrics which are widely used in the field of object detection. The weighted f-measure (WFM) is used to amend the "Equal-importance flaw" in Dice. The MAE metric is utilized to evaluate the pixel-level accuracy. To evaluate pixel-level and global-level similarity, we adopt the recently released enhanced-alignment metric EM [6]. Since WFM and MAE are based on a pixel-wise evaluation system and ignore structural similarities, SM [5] is adopted to assess the similarity between predictions and ground-truths.

The comparison results are shown table I, II and III. We compare our performance with Unet, Unet++, MSNet and PVT. The mDice score exceeds by 2.7% the second best model which is MSNet.

TABLE III
COMPARATIVE RESULTS OF MODELS WITH AND WITHOUT PCL LOSS

Model	Dice	IOU	Model	Dice	IOU
Unet	0.7698	0.6895	Unet	0.8072	0.7158
Unet++	0.7547	0.6552	Unet++	0.7834	0.6813
MSNet	0.8058	0.7149	MSNet	0.8135	0.7243
PVT	0.7896	0. 6893	PVT	0.7965	0.6945
DMSNet	0.8171	0.7321	DMSNet	0.8335	0.7565

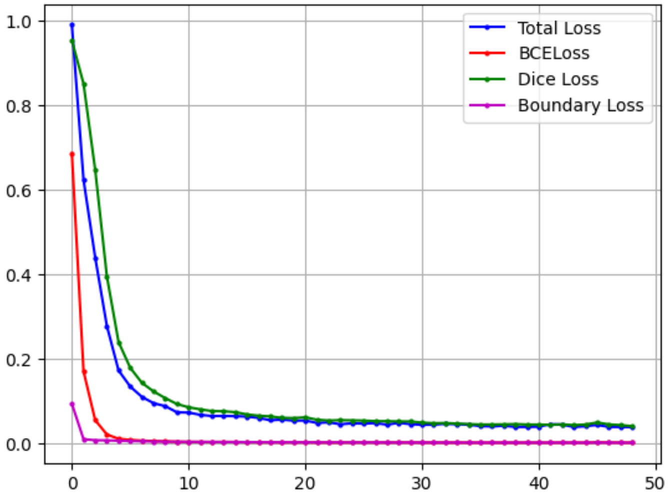


Fig. 6. Loss curve of DMSNet trained on Diabetic foot ulcer segmentation challenge for 50 epochs.

V. CONCLUSION

We have presented a novel architecture, DMSNet, for automatically segmenting Diabetic foot ulcer from RGB images. Extensive experiments demonstrated that DMSNet consistently outperforms all state-of-the-art approaches by a good margin (¿2%). Another advantage is that DMSNet is universal and flexible, meaning that more effective modules can be added to further improve the accuracy. The current encoders can be swapped with any other encoder and can be improved in future with better feature encoders. We hope this study will offer the community an opportunity to explore more powerful models on other medical image segmentation tasks such as Lung segmentations, abdominal organs and brain tumor.

References

[1] Sun, Hong, et al. ”IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045.” Diabetes research and clinical practice 183 (2022): 109119.

[2] Williams, Rhys, et al. ”Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas.” Diabetes research and clinical practice 162 (2020): 108072.

[3] Chaitanya, Krishna, et al. ”Contrastive learning of global and local features for medical image segmentation with limited annotations.” Advances in neural information processing systems 33 (2020): 12546-12558.

[4] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. ”Segnet: A deep convolutional encoder-decoder architecture for image segmentation.” IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.

[5] Fu, Jun, et al. ”Dual attention network for scene segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[6] Oktay, Ozan, et al. ”Attention u-net: Learning where to look for the pancreas.” arXiv preprint arXiv:1804.03999 (2018).

[7] Zhou, Zongwei, et al. ”Unet++: A nested u-net architecture for medical image segmentation.” Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer International Publishing, 2018.

[8] Chaurasia, Abhishek, and Eugenio Culurciello. ”Linknet: Exploiting encoder representations for efficient semantic segmentation.” 2017 IEEE visual communications and image processing (VCIP). IEEE, 2017.

[9] Koonce, Brett, and Brett Koonce. ”EfficientNet.” Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization (2021): 109-123.

[10] Dong, Bo, et al. ”Polyp-pvt: Polyp segmentation with pyramid vision transformers.” arXiv preprint arXiv:2108.06932 (2021).

[11] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. ”U-net: Convolutional networks for biomedical image segmentation.” Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[12] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. ”Fully convolutional networks for semantic segmentation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[13] Chen, Liang-Chieh, et al. ”Encoder-decoder with atrous separable convolution for semantic image segmentation.” Proceedings of the European conference on computer vision (ECCV). 2018.

[14] Zhao, Hengshuang, et al. ”Pyramid scene parsing network.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[15] Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. ”Dynamic routing between capsules.” Advances in neural information processing systems 30 (2017).

[16] He, Kaiming, et al. ”Mask r-cnn.” Proceedings of the IEEE international conference on computer vision. 2017.

[17] Lin, Guosheng, et al. ”Refinenet: Multi-path refinement networks for high-resolution semantic segmentation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.