Vision Transformer for Fine-Grained Wound Stage Classification

Siddharth Indora[1]
Computer Science and Engineering
University of California Santa Cruz
sindora@ucsc.edu

Michael Briden[2]
Computer Science and Engineering
University of California Santa Cruz
mbriden@ucsc.edu

Narges Norouzi[3]
Computer Science and Engineering
University of California Santa Cruz
nanorouz@ucsc.edu

***Abstract*—The classification of wound severity is a critical step in wound diagnosis. An effective classifier can help medical professionals categorize wound conditions more quickly and affordably, allowing them to choose the best treatment option. This study uses wound images to train a neural network-based wound classifier that classifies the images into one of four stages in the wound healing process. The dataset contains single wound images from different mice pictured daily till the wound maturates. Multi-Head attention is added using Vision Transformer to create a model which analyses the dataset. Attention maps of the images are analyzed to see where the model pays attention. The attention maps of two models are compared . The first one has a pre-training task to detect temporal validity followed by wound stage classification. The second model is trained from scratch to classify images into wound stages. Both model achieves good results in wound stage estimation.**

I. Introduction and Related Work

According to a 2018 retrospective investigation [1], more than 8 million people suffered open wounds, While the cost of their treatment was an estimated 28.1 billion. This notable figure indicates the number of people affected by wounds and the number of resources directed towards their treatment. Wound stage estimation is an essential task in the wound assessment process. Identifying, tracking, and predicting wound heal-stage progression is a fundamental task toward proper diagnosis, effective treatment, facilitating healing, and reducing pain. Wound healing routinely falls into four stages: hemostasis, inflammation, proliferation, and maturation.

Much work has been devoted to classifying wound stages and segmenting wounds. Wang et al. [2]. propose a lightweight and less compute-intensive convolutional framework based on MobileNetV2 and connected component labeling to segment wound regions from natural images. To train and test the deep learning models, they built an image dataset of 1109 foot ulcer images from 889 patients.D. M. Anisuzzaman et al. [3]l use transfer learning and stacked deep learning models. They chose nine classification models from VGG16, DenseNet, ResNet pre-trained on ImageNet, and stacked models, a mixture of the previous nine pre-trained individual models. The dataset contains 420 wound images of different wounds, including diabetic, pressure, and venous ulcers. Marcos V. Conde [4] proposes an interpretable multi-stage model based on Vision Transformers (ViT) [5] and detection-based FGVC methods that allow localizing and recognizing informative regions in the image using the inherent multi-head attention mechanism on CUB-200-2011 dataset [6]. Also, the authors explore the potential of Visual Transformers for fine-grained classification.

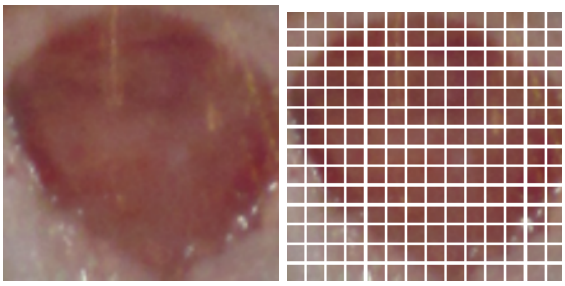

Fig. 1. Splitting of an image into patches of size 16 x 16 by Vision Transformer

Transformers have already revolutionized the way we think about attention. The paper "Attention is all you need" [7] paved the way for different types of architectures in Deep Learning which can be used in various applications as well. ViT have achieved remarkable performance recently on a variety of supervised computer vision tasks [8] [9].

ViT have replaced Convolutional Neural Networks due to inherent attention mechanisms and provide a better region of interest [10]. A region of interest (ROI) is a portion of an image that you want to filter or operate on somehow. ViT models are more robust when pointing to arbitrary ROI than CNN networks. Transformers use the self-attention mechanism as the core module to extract long-range dependencies between sequences from low-level feature sequences to achieve good performance.

Recently, several data-efficient ViT's have been proposed to alleviate the dependence on large-size datasets. For example, DeiT [11] improved the efficiency of ViT by employing data augmentations and regularizations and realized knowledge

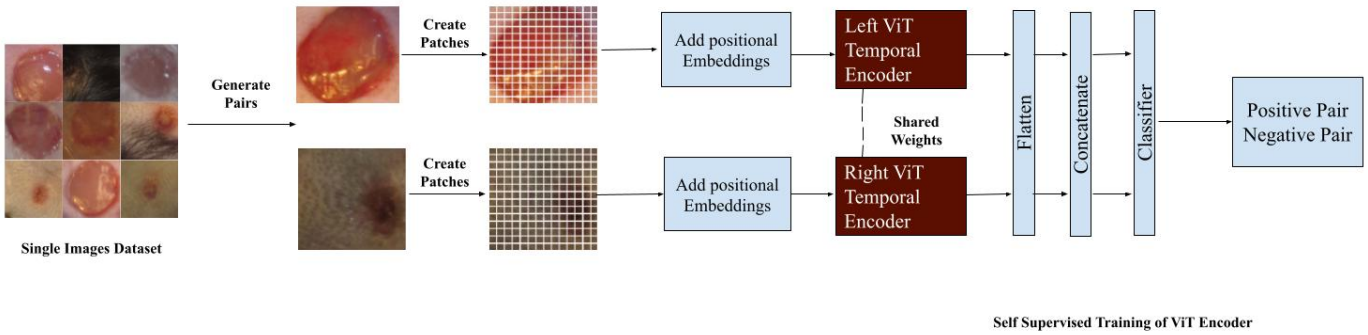

Fig. 2. Positive and Negative pairs of wound images of size (224x224) are generated, and passed through ViT encoder with shared weights forming a Siamese network. The Siamese network outputs 16-dimensional representations, which are concatenated together and passed through a binary classifier, which checks the temporal effectiveness of the encoder. The learned embedding in the encoder is used to train the downstream model to predict the wound stage in the healing process.

distillation by introducing the distillation token concept. T2T [12] used a tokenization method that flattened overlapping patches and applied a transformer. This makes it possible to learn local structure information around a token. PiT [13] produced various receptive fields through spatial dimension reduction based on the pooling structure of a convolutional layer.

ViT leverage the fact that image patches are synonymous with sequence tokens (like words) in NLP tasks. There are six steps in ViT : i) Split an image into patches ii) Flatten the patches iii) Produce lower-dimensional linear embedding from the flattened patches iv) Add positional embedding vi) If needed,pre-train the model with image labels (fully supervised on a vast data set) vii) Fine-tune on the downstream data set for image classification.

Our contributions are as follows: We transfer the original ViT model implementation to do temporal classification by making the following changes: We remove the final token layer and instead attach a lightweight, fully connected dense layer for temporal validity. Temporal validity here refers to the ability of a model to differentiate between positive pair of images and negative pair of images. The pair generation technique is defined in the Dataset section. The trained encoder, which outputs 16-dimensional embedding, is then used to create clusters and these embeddings create pseudo labels(based on clusters) that are used to assign each image a healing stage class. Thus this model follows a self-supervised learning approach. In turn, we also analyze the attention maps at each stage of the healing process.

II. Dataset

The wound images [14] are collected from eight mice: four young and four old. A wound is present on each side of the mice and is photographed daily over 16 days. The resulting dataset is a set of 256 images (16 days x 2 wounds x 8 mice). The wound images from one young, and one old mouse is reserved for validation and test set to understand any changes in the healing process between different ages of mice. The original image consists of a splint and a wound. Therefore, we perform data pre-processing. In data pre-processing, we separate the splint from the wound. Data augmentation techniques used are random flipping, random zoom, random rotation, and random contrast. These augmentations are performed on the whole image and not on the patches created. After data pre-processing, we perform data augmentations techniques, including resizing each image to 224 x 224.

For the downstream classification, we use two datasets. The first one uses pseudo labels generated by K means clustering to perform wound stage classification. The second one uses true labels of the images to perform the same task. Mouse wound true labels are created by a group of 10 non-experts for supervised classification. Each annotator labeled the images with one of the four stages: hemostasis, inflammation, proliferation, and maturation.

III. Our Contribution

A. Pretext Task

Wound pairs were created for the pre-text learning task to capture the temporal dynamics. Figure 1 shows the pre-text task process. The process of generating wound pairs follows i)Positive pairs: the image passed to the left ViT encoder is captured before the one passed to the right ViT encoder in the wound healing process. ii)Negative pairs: the image passed to the right ViT encoder is captured before the one passed to the right ViT encoder in the wound healing process.

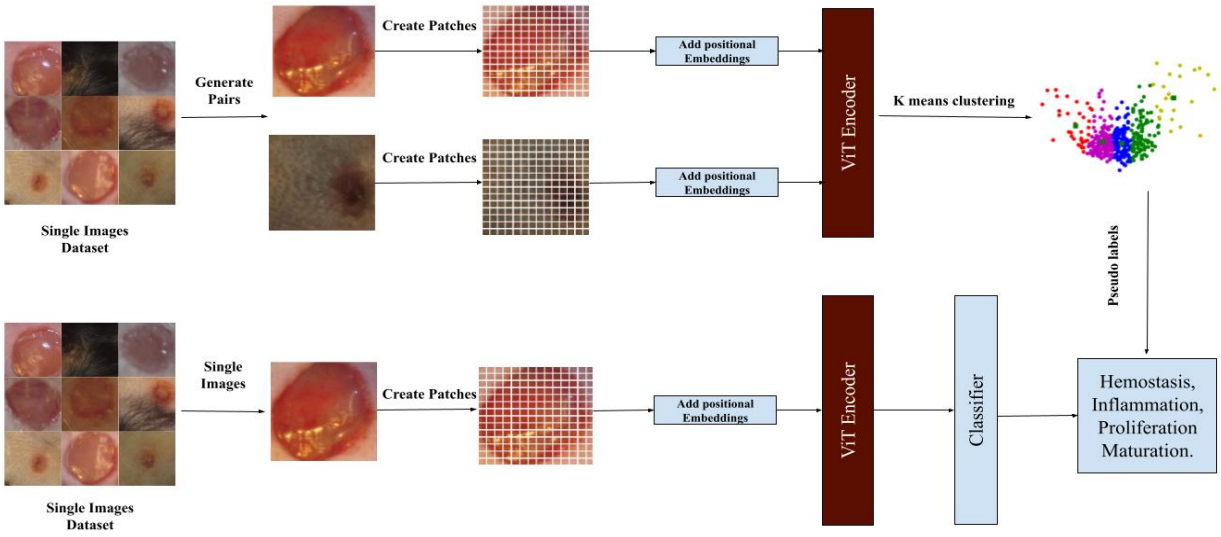

Fig. 3. For the pretext task, Positive and Negative pairs of wound images of size (224x224) passed through ViT encoder that outputs 16-dimensional representations. The output is used to perform K means clustering, an unsupervised machine learning algorithm.Using K means clustering, we create pseudo labels which act as labels for all data points inside that cluster. The pseudo labels help train downstream classifier to predict the wound healing stage.

B. Unsupervised Wound Stage Clustering

In order to identify and cluster wound healing stages, we use unsupervised K-means clustering. The ViT temporal encoder is used to create clusters using K-Means(k=4) so that images with similar characteristics are grouped into one cluster. We set k=4 to mimic stages of wound healing. We use these cluster mappings to assign pseudo-labels to our dataset.

C. Downstream Classification

The final step of the pipeline is to re-task and fine-tune the temporal encoder towards predicting the healing stage of input images based on cluster assigned pseudo-labels. The downstream classification includes replacing the previous head with a fully connected layer resembling a more traditional image classifier, as shown in Figure 2. The attention maps are generated and compared that show where the downstream model is paying attention inside each image.

IV. Model

In this work, we build upon original ViT implementation. An important point to note here is that the dataset used was not meant to train deep learning models and is relatively small, with only 256 images. The advantage of building upon original ViT implementation is to analyze the model's performance and show the benefits of transformers as a fine-grained medical image recognition framework. In addition, we can use the extensible body of knowledge surrounding ViT.

Our ViT model consists of a self-supervised learning procedure with MultiHead attention to predict wound stage classification. The temporal encoder in the pre-text task learns to differentiate between positive and negative pairs generated. These temporal encoder embeddings are then clustered into four groups to generate pseudo-labels. During the downstream classification of stages, the same temporal encoder trained earlier is used with a fully connected dense layer to predict the four stages of wound healing.

The input images in both pre-text and downstream tasks are broken into patch sizes of 16×16, allowing global attention. To encode the spatial and location information of a patch into tokens, we use position embeddings and patch embeddings [15] to preserve the information. A custom MultiHead attention was implemented for visualizing attention maps. Multi-Head attention is a core component of the transformer. The difference from Self Attention is that the multi-head mechanism splits the input into many small parts, calculates the scaled dot product for each input in parallel, and splices all attention to outputs to get the final result. On top of the MultiHead layer, we have a Multi-Layer Perception added. It is composed of Dropout and Dense layers separated by GeLU activation [16].

We used the Attention Rollout Technique [17] as done in the ViT paper to compute attention maps from the model to the input space. We averaged attention weights of ViT across all heads of one layer to get each layer's attention and all transformer layers to get aggregate maps. We then recursively multiplied the weight matrices of all layers to get attention maps.

V. Experimental Setting

We set AdamW optimizer with a learning rate of 0.001 and weight decay of 0.0001, and the maximum training epoch is 50. The code is implemented using Tensorflow.

To evaluate the model's performance, we use binary cross-entropy as a loss function and binary accuracy as a metric for the pretext task. We use categorical cross-entropy as a loss function and categorical accuracy as a metric for the downstream classifier.

For the ViT, the patch size is set to 16. The number of transformer layers is set to 4, with each layer having four heads. This configuration was determined experimentally.

VI. Results

Task	Training Accuracy	Validation Accuracy	Test Accuracy
Temporal Validity	93%	91%	87%
Downstream Classifier(Clustering)	74%	70%	69%
ViT using Supervised Learning	85%	68%	75%

Our initial attention maps generated with splint images tend to focus on the splint rather than the wound. Despite the dataset size, the ViT model achieves reasonable accuracy in detecting the temporal validity of the images. Attention maps showing the informative region detected by each layer in the ViT model are shown for each stage in the wound healing process. Figures 4 and 5 show the input images and attention maps for each stage in wound healing. An aggregate attention map is created by averaging the attention across all transformer layers. Each attention layer focuses on encoding different image features and visual receptive fields. The maps are unfazed to blur, occlusion, and illumination challenges in the dataset images.

A. ViT using Supervised Learning

We can see that aggregate attention maps for a ViT model trained from the scratch focus on both the wound as well as the edge of the wound. During the hemostasis stage, the model pays the most attention to the wound compared to the skin surrounding the wound. During the inflammation phase of wound healing, we see that most of the attention is still on the wound. The second layer attention map can completely separate the wound from the skin. During the proliferation phase, we see a shift of attention from the wound to the edge of the wound; this is in line with literature as the wound rebuilds itself and new skin grows over the wound area. In the maturation stage, when the wound completely closes and the skin is firm, the model looks for hair growth as seen in aggregate attention map to classify a wound into its maturation state.

B. ViT pre-trained to do temporal Validation

We can see that aggregate attention maps for a pre-trained ViT model focus mainly on the edge of the wound. During the hemostasis stage, the model pays all the attention to the wound compared to the skin surrounding the wound. During the inflammation phase of wound healing, we see a shift from the wound to the edge of the skin, although the attentions are not that strong yet. During the proliferation phase, the attention on the wound edges starts to become stronger and hence is depicted by bright yellow activations. In the maturation stage, when the wound completely closes and the skin is firm, the model looks for hair growth as seen in the aggregate attention map to classify a wound into its maturation state.

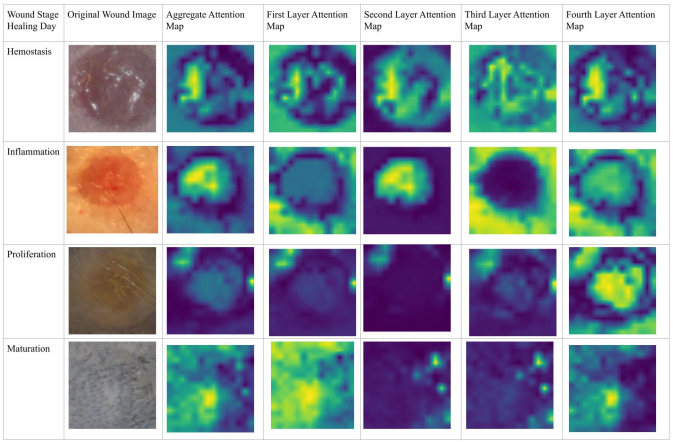

Fig. 4. Attention Maps for each stage in healing process of wound without pre-training temporal encoder

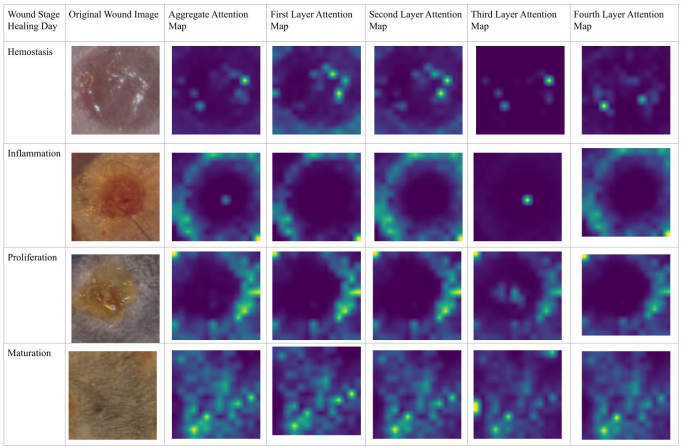

Fig. 5. Attention Maps for each stage in healing process of wound with pre-training temporal encoder

VII. Conclusion

This paper presents a self-supervised learning model for automatic wound stage classification. It compares the attention

maps generated from pre-training the model to detect temporal features, fine-tuning it in downstream classification to training the ViT from scratch. The architecture takes advantage of the Multi-Head attention mechanism in the ViT to show visual indicators in the healing process. This paper also shows the effectiveness of ViT over small datasets. Combining ViT with pretext learning tasks yields quality classification and visualization results for fine-grain classification.

REFERENCES

[1] Sen CK. Human wounds and its burden: An updated compendium of estimates. 2019.

[2] Anisuzzaman D.M. Williamson V. et al. Wang, C. Fully automatic wound segmentation with deep convolutional neural networks. 2020.

[3] D. M. Anisuzzaman, Yash Patel, Jeffrey Niezgoda, Sandeep Gopalakrishnan, and Zeyun Yu. Wound severity classification using deep neural network. 2022.

[4] Marcos V. Conde and Kerem Turgutlu. Exploring vision transformers for fine-grained classification. 2021.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale. 2020.

[6] P. Welinder P. Perona C. Wah, S. Branson and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[7] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Łukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. 2017.

[8] Dongyoon Han Sanghyuk Chun Junsuk Choe Byeongho Heo, Sangdoo Yun and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. 2021.

[9] Yue Cao Han Hu Yixuan Wei Zheng Zhang Stephen Lin Ze Liu, Yutong Lin and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. 2021.

[10] Raphael Couturier Stephane Cuenat. Convolutional neural network (cnn) vs vision transformer (vit) for digital holography. 2021.

[11] Matthijs Douze Francisco Massa Alexandre Sablayrolles Hugo Touvron, Matthieu Cord and Herve Jegou. Training ´ data-efficient image transformers distillation through attention. 2021.

[12] Tao Wang Weihao Yu Yujun Shi Zi-Hang Jiang Francis E.H. Tay Jiashi Feng Li Yuan, Yunpeng Chen and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. 2021.

[13] Dongyoon Han Sanghyuk Chun Junsuk Choe Byeongho Heo, Sangdoo Yun and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. 2021.

[14] Bagood Michelle Yang, Hsin-ya. Photographs of 15-day wound closure progress in c57bl/6j mice. 2021.

[15] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Łukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. pages 6,7, 2017.

[16] Kevin Gimpel Dan Hendrycks. Gaussian error linear units (gelus). 2016.

[17] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. 2020.