Medical Image Analysis 77 (2022) 102357

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Fully transformer network for skin lesion analysis

Xinzi He[a], Ee-Leng Tan[b], Hanwen Bi[a], Xuzhe Zhang[a], Shijie Zhao[c,*], Baiying Lei[d,*]

[a] *Department of Biomedical Engineering, Columbia University, New York, NY, USA*
[b] *Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore*
[c] *Northwestern Polytechnical University, Xian, China*
[d] *National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Marshall Laboratory of Biomedical Engineering, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China*

ARTICLE INFO

Article history:
Received 25 April 2021
Revised 26 December 2021
Accepted 6 January 2022
Available online 18 January 2022

MSC:
41A05
41A10
65D05
65D17
Skin lesion segmentation
Disease classification
Attention
Transformer
Skin lesion,

ABSTRACT

Automatic skin lesion analysis in terms of skin lesion segmentation and disease classification is of great importance. However, these two tasks are challenging as skin lesion images of multi-ethnic population are collected using various scanners in multiple international medical institutes. To address them, most recent works adopt convolutional neural networks (CNNs) for skin lesion analysis. However, due to the intrinsic locality of the convolution operator, CNNs lack the ability to capture contextual information and long-range dependency. To improve the baseline performance established by CNNs, we propose a Fully Transformer Network (FTN) to learn long-range contextual information for skin lesion analysis. FTN is a hierarchical Transformer computing features using Spatial Pyramid Transformer (SPT). SPT has linear computational complexity as it introduces a spatial pyramid pooling (SPP) module into multi-head attention (MHA)to largely reduce the computation and memory usage. We conduct extensive skin lesion analysis experiments to verify the effectiveness and efficiency of FTN using ISIC 2018 dataset. Our experimental results show that FTN consistently outperforms other state-of-the-art CNNs in terms of computational efficiency and the number of tunable parameters due to our efficient SPT and hierarchical network structure. The code and models will be public available at: https://github.com/Novestars/Fully-Transformer-Network.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

A report from the American cancer society shows that melanoma accounts for a 92% incidence and a 60% death rate among skin cancers (Siegel et al., 2021). A recent report reveals that the 5 years survival rate of melanoma can reach 99% with early diagnosis, and delayed diagnosis can lead to a dramatic decrease in survival rate from 99% to 23% (Balch et al., 2009). Dermoscopy is a noninvasive tool to distinguish melanomas from nevi. It is an in-vivo microscopy that acquires images of skin lesions with enhanced sub-skin structures via eliminating interfacial reflection (Pehamberger, 1993). However, dermoscopic image analysis is heavily dependent on skilled dermatologists and the visual inspection is both time- consuming and subjective (Vestergaard et al., 2008). Thus, automatic skin lesion analysis, which comprises skin lesion segmentation and disease classification, is highly desired as they can assist doctors in delivering diagnosis efficiently, accurately and objectively.

An early step in any automatic computer-aided diagnosis system of melanoma is to segment out the skin lesion automatically (Celebi et al., 2009). In fact, accurate segmentation plays an important role in extracting representative features for classifying melanoma. However, automatic skin lesion segmentation to separate the lesion from the surrounding healthy skin is a complicated and challenging task. As shown in the first two rows of Fig. 1, many cases contain the large variations such as irregular fuzzy boundaries, heavy hairs, circular field of view, ruler marks and dyed areas. The low contrast between lesions and health tissues further enhance the challenges. Moreover, subject-specific properties such as illumination, scale, shape and location contribute additional obstacles to the segmentation task. As for disease classification, the difficulties can be summarized as low inter- and high intra-class variations between melanoma and non-melanoma (benign). The last two rows in Fig. 1 shows image examples for disease classification.

In previous studies, many segmentation and classification methods have been proposed for skin lesion analysis based on typ-

[*] Corresponding authors.
E-mail addresses: shijiezhao666@gmail.com (S. Zhao), leiby@szu.edu.cn (B. Lei).

https://doi.org/10.1016/j.media.2022.102357
1361-8415/© 2022 Elsevier B.V. All rights reserved.

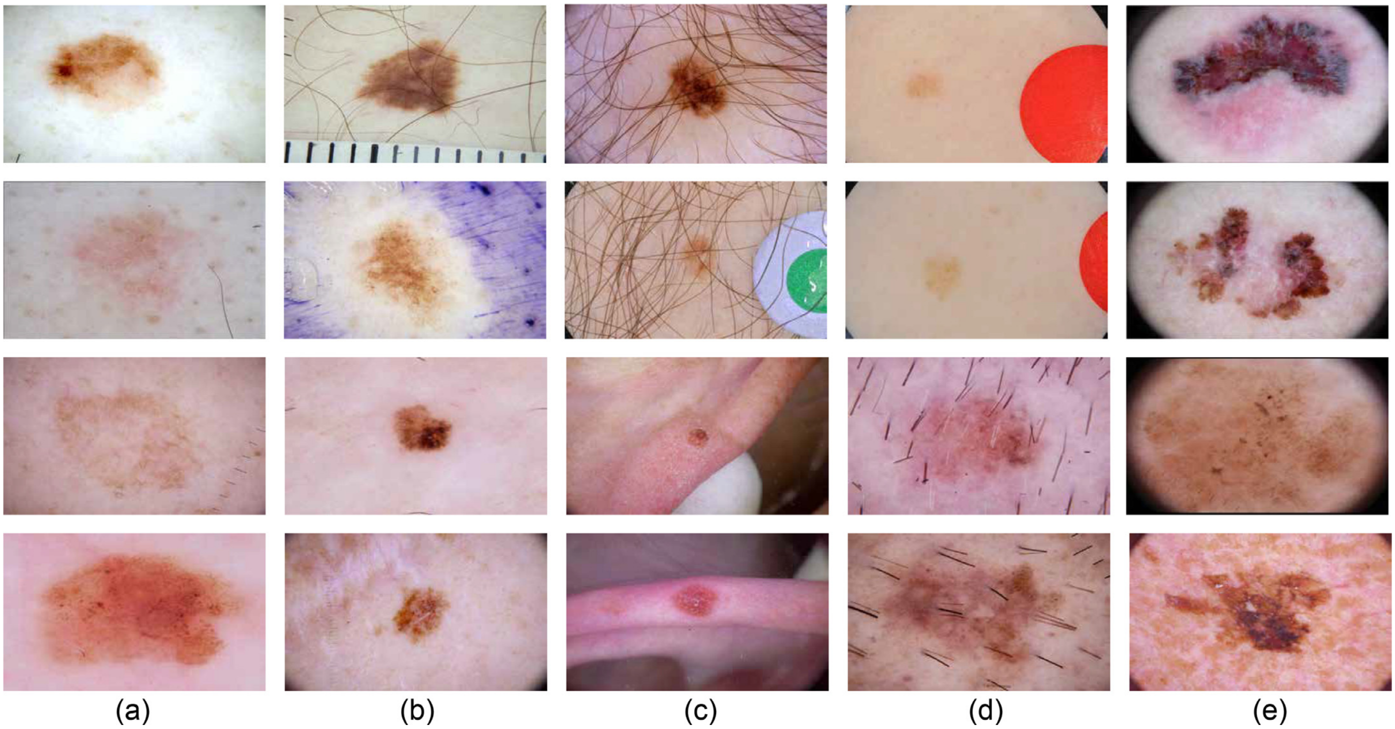


Fig. 1. Examples of dermoscopic images of skin lesions. Images of the first two rows are from ISIC 2018 skin lesion segmentation dataset, while the last two rows are images from ISIC 2018 skin lesion disease classification dataset. Main challenges of segmentation of skin lesions are: (a) low contrast between the lesions and background; (b) dyed background area; (c) lesions with thick hairs; (d) the appearance of diagnostic marks in the background; (e) circular field of view. Main challenges of classification of skin lesions are: low inter-class variations (benign (the third row) vs. melanoma (the fourth row)) and high intra-class variations (a-e).

ical machine learning approaches. Histogram thresholding methods adapt one or more appropriate thresholds to separate the skin lesions from the neighboring tissues (Silveira et al., 2009; Emre Celebi et al., 2013; Peruch et al., 2014; Yuksel and Borlu, 2009). Unsupervised clustering methods use handcrafted features such as the color space feature, lesion shapes and texture to acquire the target objects (Silveira et al., 2009). Active contour methods utilize the evolution curve or region to segment out skin lesions (Zhou et al., 2011). As for classification, (Schaefer et al., 2014) ensembled multiple classifiers which employ color, shape and texture features extracted from skin lesions to find malignant melanoma. (Yu et al., 2019) usesd support vector machines with encoded local descriptors to differentiate melanoma and non-melanoma skin lesions. However, most of these methods heavily rely on human intervention or hand-crafted features, which is unable to extract features discriminative enough to address the aforementioned challenges.

Recently, skin lesion segmentation and classification based on convolutional neural networks (CNN) has attracted tremendous attentions (Yu et al., 2019; 2017; Xie et al., 2020; Hasan et al., 2020; Lei et al., 2020; Gessert et al., 2020a). Deep-learning methods are powerful tools to extract prominent features from images rather than hand-craft features. In skin lesion segmentation, encoder-decoder convolutional neural networks (ED-CNNs) have achieved tremendous successes, especially fully convolutional network (FCN)(Shelhamer et al., 2016) and UNet (Ronneberger et al., 2015). ED-CNNs are characterized by progressively reducing the spatial resolution using an encoder, increasing the spatial resolution using a decoder and utilizing skip-connections from the encoder to the decoder to enhance local details. ED-CNNs have become a paradigm and the top choice for designing models for various medical image segmentation tasks, such as organ segmentation from computer tomography, tumor segmentation from magnetic resonance imaging and skin lesion segmentation from dermatoscopy. Recently, ED-CNNs have been introduced in skin lesion segmentation. For example, (Yu et al., 2017) proposed a two-stage model for skin lesion segmentation and classification. Their fully convolutional residual network ranked the second place in ISIC 2016 skin lesion analysis challenge (Gutman et al., 2016). In the same year, (Yuan et al., 2017) employed FCN for skin lesion segmentation, which introduces Jaccard distance as a loss function to solve the lesion-tissue imbalance. (Esteva et al., 2017) utilized a pre-trained deep network for automatic skin cancer classification. This CNN-based approach used ImageNet pre-trained GoogleNet Inception v3 to extract representative features to distinguish melanoma and non-melanoma (benign) (Szegedy et al., 2016). In 2018, (Yu et al., 2019) used features extracted from CNNs to construct local deep descriptor to build a global image representation. This representation is then fed into classifiers to enhance the performance of directly using CNNs features.

Though CNNs proved their effectiveness in skin lesion analysis, including skin lesion segmentation and disease classification, they fail to capture contextual information effectively due to the intrinsic locality of the convolution operator (Dosovitskiy et al., 2020; Vaswani et al., 2017). Long-range dependencies are prerequisite for forming contextual information and such information is crucial for both segmentation and classification tasks. So far, many improvements have been proposed to improve CNNs (Ibtehaz and Rahman, 2020; Schlemper et al., 2019; He et al., 2019). DeepLab utilizes dilated convolution to increase the receptive fields without increasing the kernel size of filters (Chen et al., 2018; Wang et al., 2020). NonLocal network uses a non-local module for capturing long-range dependencies by directly computing the interaction between two positions (He et al., 2019). Both dilated/atrous convolution operator and the non-local module sit at the top of CNNs or to be used as a sub-module, hence features from other layers still suffer from a paucity of contextual information (Schlemper et al., 2019).

Lately, Transformer, a fully self-attention model designed for language processing, has attracted wide attention and has been recently applied to vision tasks such as image classification, semantic segmentation and object detection (Vaswani et al., 2017; Zheng et al., 2021). Among these works, ViT is the first fully Transformer model proposed to recognize images (Dosovitskiy et al., 2020). ViT splits an image into 16x16 non-overlapping patches (a.k.a., token), then the concatenation of flattened tokens and a class token are fed into the network following the same procedure as in Transformer for language processing. The Transformer in ViT models the global relation in a simple and efficient manner. ViT pre-trained on large datasets like JFT 300 or ImageNet 21k significantly benefits more challenging downstream applications such as semantic segmentation (Dosovitskiy et al., 2020) or object detection(Beal et al., 2020). However, when Transformers are trained from scratch for skin lesion analysis (Touvron et al., 2021; Han et al., 2021), they are ineffective for this application.

We hypothesize these downsides stem from the limitation of ViT, specifically, the straightforward tokenization of the input image using the hard split operation (Yuan et al., 2021). To overcome the issue that Transformer lacks local information such as edges, lines and corners in each token, we proposed an FTN for skin lesion segmentation and disease classification in this paper. FTN presents a self-attention paradigm for image segmentation and other downstream tasks without pretraining on large dataset like JFT-300M or ImageNet-21k (Dosovitskiy et al., 2020). Our contributions are as follows:

- We propose an FTN for skin lesion segmentation and disease classification on ISIC 2018 dataset, which is fully relied on Transformer, and outperforms other competing CNN-based methods.
- We leverage Sliding Window Tokenization (SWT) instead of the native non-overlapping tokenization to construct hierarchical features. The hierarchical Transformer design not only extracts discriminative features but also preserves more fine-grained texture information.
- Spatial Pyramid Transformer (SPT) is proposed to improve efficiency, since it only requires linear computational complexity with respect to the feature dimensions. Low level high-resolution features can be extracted without increasing too much computation and memory usage.
- A Transformer decoder is proposed to aggregate hierarchical features extracted by our hierarchical Transformer. The Transformer decoder module contains skip-connections from the encoder to the decoder to compensate the loss from the reduction of spatial resolution using SWT and utilizes the proposed SPT for better performance and efficiency.

2. Related work

2.1. CNN-Based methods

The convolutional layer is first introduced to recognize digital hand-writing numbers. A convolutional layer contains kernels, whose weights are shared over the whole image to capture translation-invariant local patterns. The development of computational resource, especially GPUs, enables the training very deep neural networks possible. For the CNNs, VGG demonstrates that the deep-narrow structure is a good balance between computational efficiency and accuracy (Simonyan and Zisserman, 2015; Gessert et al., 2020a; Tang et al., 2020; Xie et al., 2020). More recently, ResNet with skip-connections in CNNs achieves state-of-the-art performance in various tasks including classification, segmentation and detection(He et al., 2016; Xie et al., 2021).

Motivated by the success of CNNs in computer vision, CNNs have been widely adopted for ISIC skin lesion analysis challenges(Tschandl et al., 2018; Gutman et al., 2016; Codella et al., 2018; 2019; Combalia et al., 2019; Rotemberg et al., 2021). (Yuan et al., 2017) introduced a fully automatic method using 19 layers CNN based on the Jaccard Index loss function. (Xie et al., 2020) used a CNN to improve disease classification accuracy by filtering out useless background with attention mechanisms. (Jahanifar et al., 2019) used ensemble CNNs to detect skin lesion attributes by formulating detection tasks into segmentation tasks.

2.2. Self-attention in CNN

Self-attention mechanism has been widely applied to CNN in ISIC skin lesion analysis challenges. In (Hu et al., 2019), authors focus on the channel relationship and propose a novel channel attention unit, termed squeeze and excitation block, which calibrates feature channel-wisely. Followed this idea, (Gessert et al., 2020b) utilized the patch-based attention to aggregate context information. Compared with most of the works using the channel attention, some works used the spatial attention to explore long-range spatial dependency. (Zhang et al., 2019) proposed a spatial attention-based network called attention residual learning convolution neural network for skin lesion classification. However, all these works partially rely on attention mechanisms which is based on the top of the network or local attention (Schlemper et al., 2019). Different from previous works, we construct FTNs which purely rely on self-attention and extend them in skin lesion segmentation skin disease classification.

2.3. Transformer

Transformer is the first method that fully depends on the self-attention mechanism to explore long-range dependency in natural language processing. The Transformer layer is comprised of a multi-head attention (MHA) layer and a multi-layer perceptron (MLP). Layer normalization and skip-connection are applied in both MHA and MLP layers. ViT is purely composed of Transformer layers, which has been proposed for image recognition and it has been demonstrated that only transformer can achieve state-of-the-art performance. However, authors found that ViT relies on large image dataset such as ImageNet-21K and JFT-300M for model pretraining (Dosovitskiy et al., 2020).

3. Methodology

We propose an FTN for skin lesion segmentation and disease classification. In skin lesion segmentation, FTN has an encoder-decoder structure similar to CNNs. Therefore, FTN, which is illustrated in Fig. 2, comprises 1) SPT for efficient and effective self-attention; 2) a Transformer encoder which utilizes SWT modules for extracting hierarchical representation; 3) a Transformer decoder that refines high-level features using hierarchical representation generated by the encoder. These components are introduced in the following sections.

3.1. Spatial pyramid transformer

The standard ViT conducts global self-attention and the relationships between a token and all other tokens are computed. This operation leads to quadratic complexity regarding the length of tokens. Such high computation is unsuitable for skin lesion segmentation and disease classification as these tasks need numerous tokens for dense prediction to represent a high-resolution image. To improve efficiency and alleviate the difficulty of capturing all-pair relations, we propose to compute self-attention on downsampled features. As we use the pooling operation to downsample the

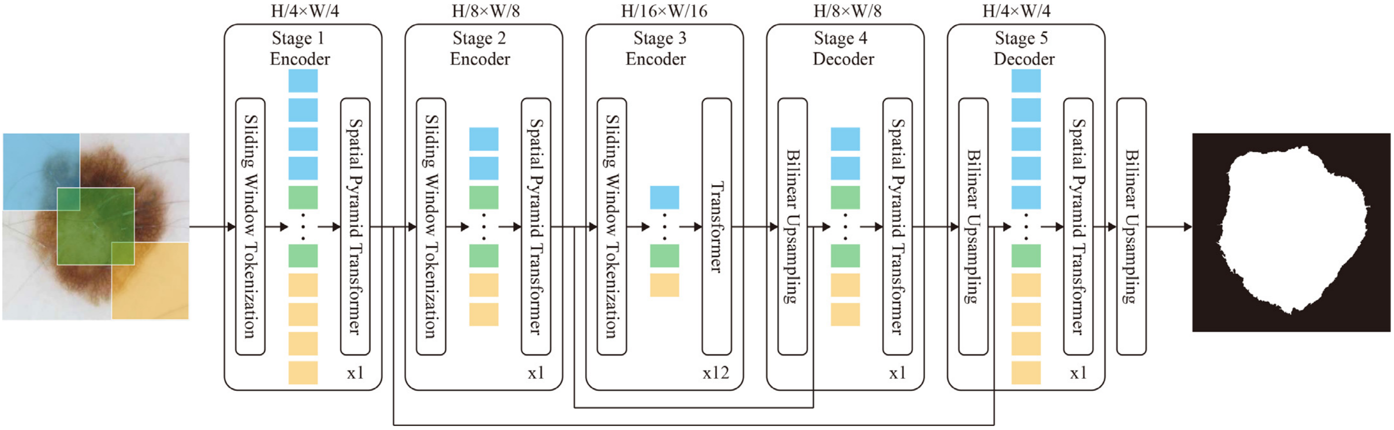


Fig. 2. The overall framework of FTN. The entire model is divided into five stages and each stage contains 1) the SWT module; 2) the corresponding Transformer layers. The output resolution of each SWT is displayed above each stage. The first three stages are the encoder of FTN and the last two stages are the decoder of FTN. Skip-connections from the encoder to the decoder introduce low-level information for compensating for the high-frequency information loss due to progressive spatial reduction in encoder stages.

key-value pair, we refer the transformer as SPT, which is shown in Fig. 4.

3.1.1. Traditional transformer

A typical Transformer layer comprises two sub-layers: a MHA layer and a MLP layer. For each of two sub-layers, a residual connection is applied around it and followed by a layer normalization. Namely, the output of a traditional Transformer can be written as:

$$TF(t_{in}) = MLP(LayerNorm(MHA(LayerNorm(t_{in})))) \tag{1}$$

$$MHA(Q,K,V) = Concat(head_0, \ldots head_{N_i})W^o \tag{2}$$

$$head(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

$$Q = t_{in}W^Q; K = t_{in}W^K; V = t_{in}W^V \tag{4}$$

where TF is a Transformer layer; t_{in} is the input tokens; Q, K, V are the query, key and value tokens, respectively and W^O, W^Q, W^K, W^V denote the parameter matrices in MHA.

3.1.2. Multi-head spatial pyramid attention

MHA layer dominates the computation and memory occupation. Specifically, the large matrix multiplication in the MHA layer between K and V mainly leads to the inefficiency.

As shown in Fig. 3, the multi-head spatial pyramid attention (MSPA) uses the same scaled dot-product attention as the traditional Transformer, where the query Q and the key-value pair K, V are mapped to an output Zhao et al. (2017); Zhu et al. (2019). SPT is constructed by replacing the key-value pair by their downsampled counterparts, with other layer kept the same. As illustrated in Fig. 3, we use adaptive pooling layers to generate a feature pyramid of four levels. In the pyramid, the levels are of size 1, 2, 4 and 16. The difference between them is that the key-value pair our MSPA received are generated by the spatial pyramid pooling(SPP) module. Mathematically, MSPA is defined as:

$$head(Q,K,V) = softmax(\frac{QK'^T}{\sqrt{d_k}})V' \tag{5}$$

$$K' = Concat(Pooling_0(K), \ldots, Pooling_i(K)W^K \tag{6}$$

$$V' = Concat(Pooling_0(V), \ldots, Pooling_i(V))W^V \tag{7}$$

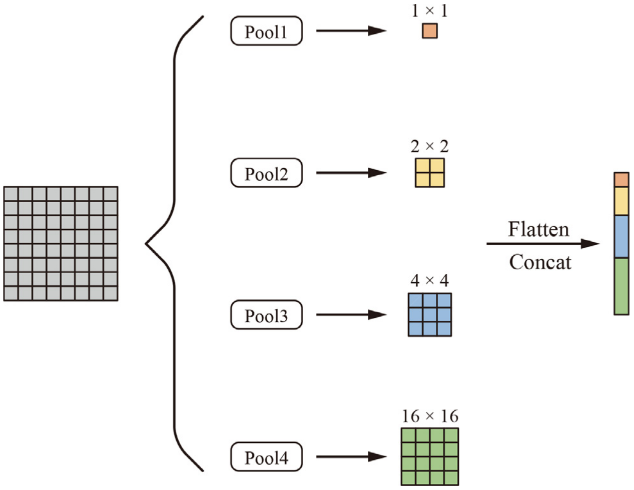


Fig. 3. The pipeline of the spatial pyramid pooling module. Pooling operations are applied to the input features to generate hierarchical features, followed by flattening and concatenation to form the output of the spatial pyramid pooling module.

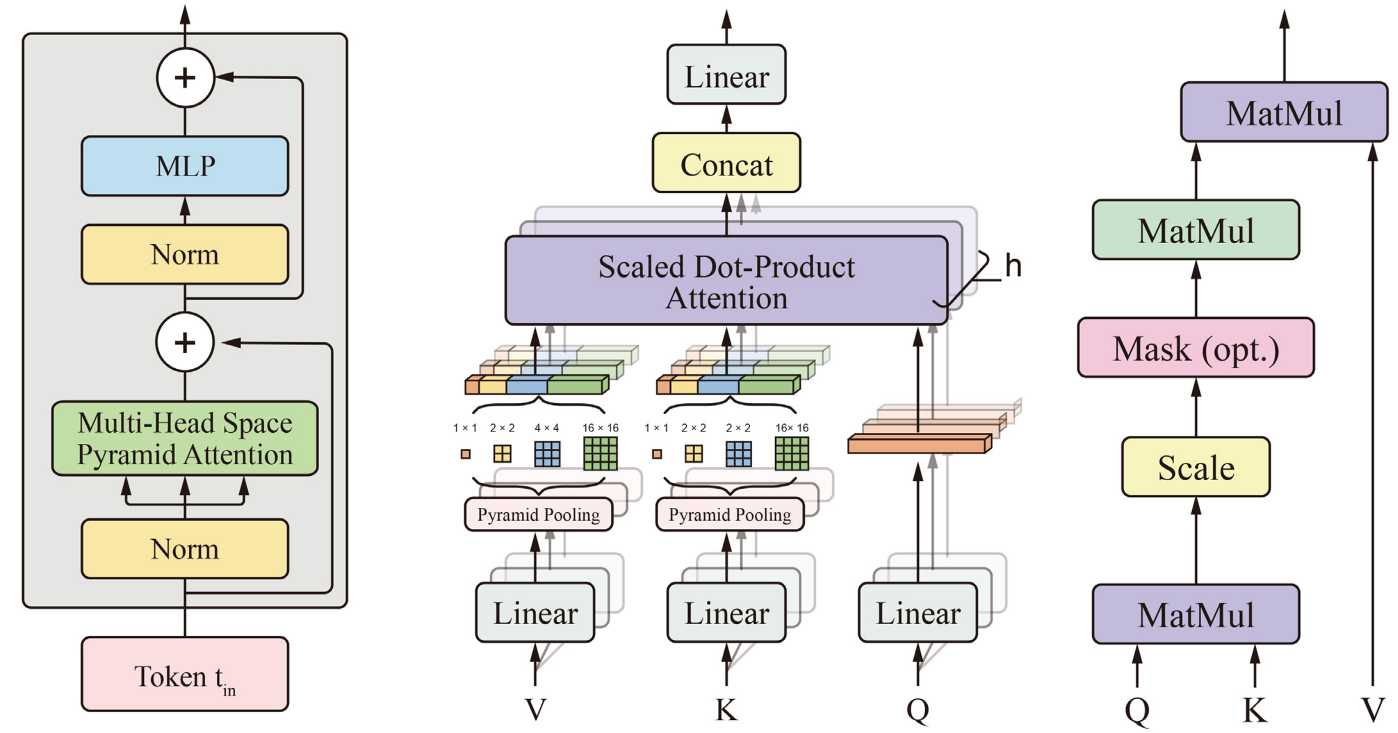
where $Pooling_i$ is the i th pooling layer and i is set as 4 in all fully transformer networks. The output size of each pooling layer is given in Table 1.

3.2. Sliding window tokenization

To achieve pixel-level classification for segmentation, models should consist of three essential parts. An encoder which is responsible for generating deep features by reducing spatial dimensions of features. Taking inspiration from Tokens-to-Token module (Yuan et al., 2021), we introduce a SWT to progressively reduce the resolution of features to replace the straightforward simple tokenization. The repeated use of SWT leads to a hierarchical Transformer design and the hierarchical architecture constructs a feature pyramid. By constructing the feature pyramid, we can capture both local structure and contextual information for segmentation and classification. As shown in Fig. 5, SWT consists of two main steps to generate tokens: 1) sliding window sampling (SWS); 2) feature reconstruction.

3.2.1. Sliding window sampling

To construct a feature pyramid, we apply sliding window sampling to generate new tokens and reduce the length of tokens si-



(a) Spatial Pyramid Trasnformer (b) Multi-head Spatial-Pyramid Attention (c) Scaled Dot-Product Attention

Fig. 4. The architecture of Spatial Pyramid Transformer.

Table 1
Detailed structures of the proposed FTNs for segmentation. Input image size is 512 × 512 by default. SWT: Sliding Window Tokenization. MSPA: Multi-head Spatial Pyramid Attention. MHA: Multi-head Attention. MLP: Multi-Layer Perceptron. Upsample: Bilinear Upsampling. P_i represents the output size of pyramid pooling. H_i and D_i is the number of heads and the embedding feature dimension in the ith MSPA module. R_i is the feature dimension expansion ratio in the ith MLP layer.

	Output Size	Layer Name	FTN-4	FTN-8	FTN-12
Stage1	128 × 128	SWT	7 × 7, stride 4, padding 2		
	128 × 128	MSPA MLP	$\begin{bmatrix} P_1 = \{1, 2, 4, 16\} \\ H_1 = 1, D_1 = 64 \\ R_1 = 1 \end{bmatrix} \times 1$	$\begin{bmatrix} P_1 = \{1, 2, 4, 16\} \\ H_1 = 1, D_1 = 64 \\ R_1 = 1 \end{bmatrix} \times 1$	$\begin{bmatrix} P_1 = \{1, 2, 4, 16\} \\ H_1 = 1, D_1 = 64 \\ R_1 = 1 \end{bmatrix} \times 1$
Stage2	64 × 64	SWT	3 × 3, stride 2, padding 1		
	64 × 64	MSPA MLP	$\begin{bmatrix} P_2 = \{1, 2, 4, 16\} \\ H_2 = 1, D_2 = 64 \\ R_2 = 1 \end{bmatrix} \times 1$	$\begin{bmatrix} P_2 = \{1, 2, 4, 16\} \\ H_2 = 1, D_2 = 64 \\ R_2 = 1 \end{bmatrix} \times 1$	$\begin{bmatrix} P_2 = \{1, 2, 4, 16\} \\ H_2 = 1, D_2 = 64 \\ R_2 = 1 \end{bmatrix} \times 1$
Stage3	32 × 32	SWT	3 × 3, stride 2, padding 1		
	32 × 32	MSPA MLP	$\begin{bmatrix} P_3 = \{1, 2, 4, 16\} \\ H_3 = 2, D_2 = 256 \\ R_3 = 2 \end{bmatrix} \times 4$	$\begin{bmatrix} P_3 = \{1, 2, 4, 16\} \\ H_3 = 3, D_2 = 384 \\ R_3 = 2 \end{bmatrix} \times 8$	$\begin{bmatrix} P_3 = \{1, 2, 4, 16\} \\ H_3 = 4, D_2 = 512 \\ R_3 = 2 \end{bmatrix} \times 12$
Stage4	64 × 64	Upsample	2× bilinear upsample(concat(stage2, stage3))		
	64 × 64	MSPA MLP	$\begin{bmatrix} P_4 = \{1, 2, 4, 16\} \\ H_3 = 1, D_2 = 64 \\ R_3 = 1 \end{bmatrix} \times 1$	$\begin{bmatrix} P_4 = \{1, 2, 4, 16\} \\ H_3 = 1, D_2 = 64 \\ R_3 = 1 \end{bmatrix} \times 1$	$\begin{bmatrix} P_4 = \{1, 2, 4, 16\} \\ H_3 = 1, D_2 = 64 \\ R_3 = 1 \end{bmatrix} \times 1$
Stage5	128 × 128	Upsample	2× bilinear upsample(concat(stage1, stage4))		
	128 × 128	MSPA MLP	$\begin{bmatrix} P_4 = \{1, 2, 4, 16\} \\ H_3 = 1, D_2 = 64 \\ R_3 = 1 \end{bmatrix} \times 1$	$\begin{bmatrix} P_4 = \{1, 2, 4, 16\} \\ H_3 = 1, D_2 = 64 \\ R_3 = 1 \end{bmatrix} \times 1$	$\begin{bmatrix} P_4 = \{1, 2, 4, 16\} \\ H_3 = 1, D_2 = 64 \\ R_3 = 1 \end{bmatrix} \times 1$
	Params		8 M	14 M	19 M

multaneously. In sliding window sampling, tokens are produced by sliding a kernel of size k with stride s over an input image or features. We set the kernel size larger than the stride size to avoid information loss. Each patch can establish a prior that there should be a stronger correlation with the surrounding patches. Given an image $b \times c \times h \times w$, sliding window sampling generates patches whose dimensions are:

$$p = \left\lfloor \frac{h + 2p - k}{k - s} + 1 \right\rfloor \times \left\lfloor \frac{w + 2p - k}{k - s} + 1 \right\rfloor \tag{8}$$

where b is the batch size, c is the channel numbers, h and w are the height and width of the input image and k, s and p are the kernel size, the stride and the number of zero padding.

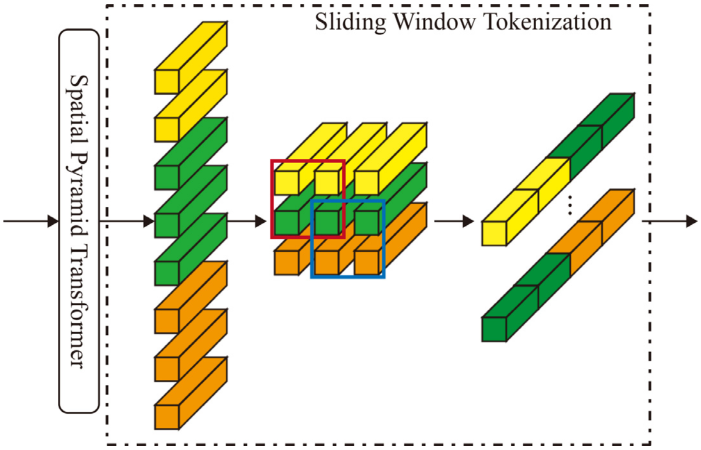


Fig. 5. The pipeline of SWT. The tokens are first mapped to token embedding by a SPT. Then reconstruction module reshapes tokens back to image features, followed by sliding window sampling to generate tokens as the input for the next SWT. Sliding window sampling: a kernel of size 2 slides over image features along spatial dimensions to form new tokens.

After sliding a kernel over a whole image, 2D patches are flattened into 1D tokens. As the tokens will serve as the input of an SPT layer in stages 1 and 2 or a classic Transform layer in stage 3. Note that we adopt a linear projection to adjust the output channel number.

3.2.2. Reconstruction

As shown in Fig. 5, the reconstruction module reconstructs one dimensional 'words' tokens back to two dimensional image features, which is defined as:

$$I = Reshape(t_{in}) \tag{9}$$

This module can be implemented by *reshape* and *transpose* functions in PyTorch.

3.3. Fully transformer network encoder

To produce hierarchical features, the number of tokens is reduced gradually by the SWT module. The encoder of FTN consists of three stages: In stage 1, the image is directly fed to the sliding window sampling layer to generate tokens. SPT is applied afterwards to transform generated tokens. In stages 2 and 3, the procedure is kept the same as stage 1, except the input tokens are reconstructed to the image features at the beginning. The model sizes, throughput and hyperparameters of FTNs are listed in Table1.

3.4. Transformer decoder

To recover object segmentation details, we propose an efficient and effective Transformer decoder. As shown in Fig. 2, the Transformer decoder is comprised of stage 4 and 5, which progressively restores the resolution from $\frac{H}{16} \times \frac{W}{16}$ to $\frac{H}{4} \times \frac{W}{4}$. In stage 4, the embedding dimension and the resolutions of the tokens from stage 3 are reduced and upsampled to the same as the tokens from stage 2 using a linear projection and bilinear upsampling layers. The resized high-level features are then concatenated into the reshaped low-level features. After the concatenation, we utilize an SPT module to refine features. This process repeats again in stage 5. We output the final prediction by bilinear upsampling the final feature by a factor of 4.

3.5. Fully transformer network architecture

FTN for skin lesion segmentation contains five stages and each stage contains: SWT and Transformer layers. Thus FTNs for segmentation have 3 sliding window sampling and 3 reconstruction. The kernel sizes for the five soft window sampling are respectively [7,3,3], the overlapping are [3,1,1], strides are [4,2,2] and padding are [2,1,1]. Note that there is no reconstruction module in stage 1, but a reconstruction module transform tokens into image feature for final output. Given the image of 512 × 512, the input feature resolution of Transformer backbone is 32 × 32. Compared to the traditional ViT, FTN adopts deep-narrow structure with smaller hidden dimension (256–512) and MLP size (512–1536) than ViT. We instantiate FTNs with different paramters by varying the number of transformer blocks and hiden dimensions used, as shown in Table 2.

Table 2
Detailed structures of the proposed FTN for classification. We use input image size 224 × 224 as an example. SWT: Sliding Window Tokenization. MSPA: Multi-head Spatial Pyramid Attention. MLP: Multi-Layer Perceptron. P_i represents the output size of pyramid pooling. H_i and D_i is the number of heads and the embedding feature dimension in the ith MSPA module. R_i is the feature dimension expansion ratio in the ith MLP layer.

	Output Size	Layer Name	FTN-12-Cla
Stage1	56 × 56	SWT	7 × 7, stride 4, padding 2
	56 × 56	MSPA MLP	$\begin{bmatrix} P_1 = \{1,2,4,12\} \\ H_1 = 1, D_1 = 64 \\ R_1 = 1 \end{bmatrix} \times 1$
Stage2	28 × 28	SWT	3 × 3, stride 2, padding 1
	28 × 28	MSPA MLP	$\begin{bmatrix} P_2 = \{1,2,4,12\} \\ H_2 = 1, D_2 = 64 \\ R_2 = 1 \end{bmatrix} \times 1$
Stage3	14 × 14	SWT	3 × 3, stride 2, padding 1
	14 × 14	SWT MSPA MLP	$\begin{bmatrix} 3 \times 3, \text{stride1, padding1} \\ P_3 = \{1,2,4,12\} \\ H_3 = 2, D_2 = 256 \\ R_3 = 2 \end{bmatrix} \times 4$
		Global Average Pooling Linear Projection	
	Params		8 M

In order to apply FTN to skin disease classification, we remove the Transformer decoder and add an adaptive average pooling layer followed by a layer normalization layer and linear projection layer at the end of the network.

4. Experiments

To evaluate the performance of FTN in the skin lesion segmentation and classification tasks, we conduct the extensive experiments on the ISIC 2018 dataset to compare performance of FTN with other CNN-based baseline frameworks(Alom et al., 2019; He et al., 2016; Zhang et al., 2018). Note that we use FTN to represent FTN-12 in the following section without explicit notation as FTN-12 performs the best in the ablation study with regard to the depth of FTN. In Section 4.1, we demonstrate that the proposed FTN achieves state-of-the-art performance in skin lesion segmentation; In Section 4.2, we conduct experiments on FTN and other CNN-based classification models (including widely-used methods: VGG, ResNet and DesneNet), which shows that FTN is a powerful alternative for CNNs on skin disease classification; In Section 4.3, comprehensive ablation studies are done to validate the effectiveness of the proposed key components.

4.1. Evaluation metric

For the performance evaluation of lesion segmentation, we adopt metrics recommended by the ISIC, including: Accuracy (Acc), Sensitivity (Sens), Specificity (Spec), Jaccard Index (JI) and Dice coefficient (DSC). As for the performance of disease classification, we apply four widely used performance metrics: the area under receive operation curve (AUC), Sens, Spec and Precision (Pre). Note that we use AUC as the main criteria in the classification task, as it is recommended by the ISIC. Among the segmentation metrics, JI is the main metrics for verifying the performance of a segmentation model. JI is a metric that measures the similarity between the prediction and the groundtruth.

4.2. Segmentation results

4.2.1. Dataset and preprocessing

We utilize the public ISIC 2018 skin lesion segmentation dataset to evaluate the segmentation performance of FTN. The training dataset contains 2594 RGB images and corresponding groundtruth. We randomly split them into 2076 images for training and 518 images for testing.

As the ISIC 2018 skin lesion segmentation dataset is a multi-source dataset and images are of different resolutions. Thus, we uniformly resize all images to 512x512 for both training and testing. Since this work focuses on methodology innovation, to eliminate the effect of confounders of data augmentation in comparison, we only utilize simple geometrical data augmentation strategies, including random horizontal and vertical flip, as well as random rotation (i.e., from 0° to 360°). Note, random variables are taken from uniform distributions.

4.2.2. Experimental setup

All experiments are implemented on PyTorch frame with an NVIDIA RTX TITAN GPU. We employ an AdamW, a variant of Adam optimizer for 100 epochs using batch size = 20, weight decay = 0.0001, and momentum = 0.9. Initial learning rate of 0.0002 and a cosine decay learning rate scheduler are used.

4.2.3. Results

To verify the performance of our method, we compare FTN with other state-of-the-art CNNs in medical image segmentation using the ISIC 2018 skin lesion segmentation dataset. The CNNs used for comparisons are: FCN(Shelhamer et al., 2016), UNet(Ronneberger et al., 2015), ResUNet(Ibtehaz and Rahman, 2020), R2UNet(Alom et al., 2019) and AttUNet(Schlemper et al., 2019). We introduce these networks briefly and details can be found in the reference. FCN uses VGG as the encoder and adopts skip-connections to integrate global information from deep, coarse layers with texture information from shallow, fine layers. UNet consists of an encoder and a decoder. The encoder uses multiple encoder blocks which is made up of two convolutional layers, each is followed by a rectified linear unit layer and a 2x2 max pooling layer. Every step in the decoder consists of a upsampling layer followed by a convolutional layer, which reduces the channel dimensions to half, and a concatenation with features from the corresponding encoder block followed by two convolutional layers. ResUNet is the variants of UNet using Residual blocks in both encoder and decoder blocks. R2UNet, a UNet family member, introduces the proposed recurrent convolution layers to replace the convolutional layer in UNet which used to halve the channels of features. AttUNet integrates the proposed addictive attention layer into UNet.

Table 4 shows the quantitative result of the comparisons. From the table, we can see our method ranked at the first place using five important metrics: ACC, Sen, Spec, Ji and DSC. The segmentation results presented demonstrate that the superiority of FTN over well-established CNNs. More importantly, it proves that FTN with carefully designed tokenization strategy and efficient and effective transformer layers can be trained from scratch without pretraining and transfer learning in skin lesion segmentation. In sum, FTN is the first Fully Transformer Network in medical image segmentation and the success in skin lesion segmentation can stimulate more research on other medical image segmentation tasks using Transformer.

In addition to the remarkable performance, FTN is efficient in terms of parameters number (millions) and throughput. To verify it, we also calculate the Frames Per Second (FPS) of all methods on a NVIDIA RTX TITAN GPU card, given the input image of size 512x512. As shown in Table 4, we can see FTN is more efficient than state-of-the-art methods. Specifically, FTN is 2.18, 5.36, 6.38, 12.5 and 5 times more efficient than UNet, FCN, ResUNet, R2UNet and AttUNet, respectively. Moreover, in the table, FTN shows that it contains far less parameters compared to UNet, R2UNet and AttUNet, as FTN only contains linear projections. Though FCN contains fewer parameters, the difference is fairly small and our method delivers better performance and throughput.

4.2.4. Visual results

The visualization of segmentation performance of FTN and other state-of-the-art networks are shown in Fig. 6. It is observed that FTN achieves better segmentation accuracy compared to other networks despite the blur boundaries. Compared with other convolution-based methods, it shows that contextual information is crucial for the segmentation of challenging cases, and FTN can capture more contextual information using the self-attention mechanism. In column 2,3,5,9, it is observed that FTN provides better predictions of lesions when there are hairs, rulers and marks. By contrast, other methods fail to capture the shape of lesions accurately as they are unable to identify these artifacts with limited receptive fields. Moreover, the results demonstrate the applicability of Transformer-based method on complicated medical image segmentation tasks.

4.2.5. Comparison with state-of-the-art methods

To further verify the effectiveness of our method, we also compare our methods with several state-of-the-art methods that appeared on the public leaderboard on the official ISIC 2018 evaluation set. The official evaluation dataset contains 100 RGB skin lesion images. As the source code of these methods are not released by authors, we just report these methods based on information from their publications. For comparison, we listed top three methods in Table 3, including 1) MAsk-Rcnn2+Segmentation; 2) ensemble with CRF; 3) Automatic Skin Lesion Segmentation by DCNN. Here, we briefly introduce these three methods, and readers can refer to their publications for more details. The top method adopts a two-stage process where the first stage is the detection model Mask-Rcnn and the second stage is the skin lesion segmentation using an encoder-decoder model (Qian et al., 2018). In addition to

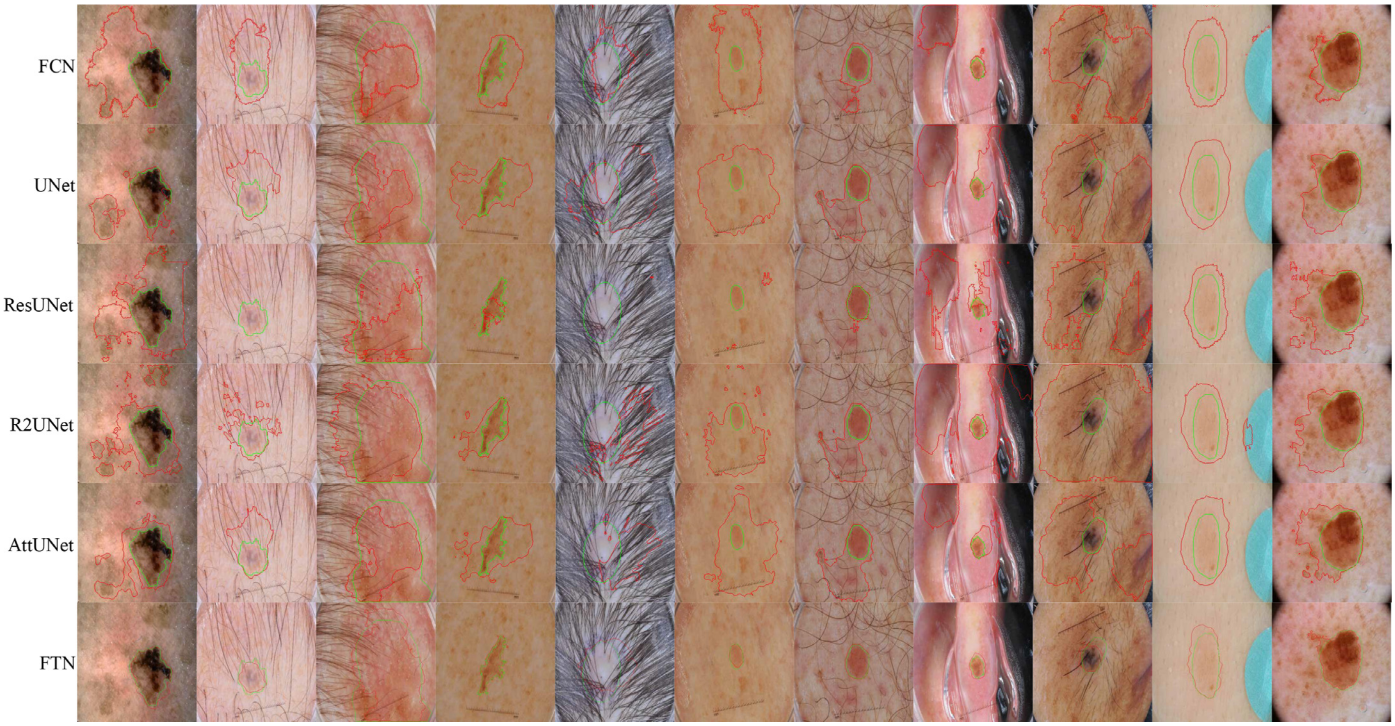


Fig. 6. Segmentation performance comparison among FTN and other state-of-the-art methods on ISIC 2018 dataset. From the first row to the last row, they represent FCN, UNet, ResUNet, R2UNet, AttUNet and FTN, respectively. Each column stands for different case. The green and red contours are the boundaries of the groundtruth and the predictions, respectively.

Table 3
Results of skin lesion segmentation.

Rank	User	Method Description	JI	Acc	Sens	Spec	DSC
1	Chengyao Qian	MAsk-Rcnn2+Segmentation	81.6	/	/	/	/
2	Jeon Young Seok	ensemble with CRF_v3	81.3	/	/	/	/
3	Yuan Xue	Automatic Skin Lesion Segmentation by DCNN	81.0	/	/	/	/
Ours	Xinzi He	FTN	82.9	94.6	93.9	94.4	90.0

Table 4
Results of skin lesion segmentation.

Model	Acc	Sens	Spec	JI	DSC	Param	FPS
FCN	94.0	92.1	95.1	73.2	82.3	18.6	1.37
UNet	93.9	90.1	96.3	73.8	83.0	34.5	0.56
ResUNet	93.3	83.9	96.2	71.1	80.2	13.0	0.47
R2UNet	94.0	91.5	95.6	74.1	83.6	39.1	0.24
AttUNet	94.6	86.2	98.5	77.2	85.7	34.9	0.60
Ours	96.2	96.2	97.5	82.8	89.8	19.9	3.0

the operation, the authors also adopt multi-angle inference to further improve the model performance. The second method is an ensemble method which aggregates multiple deeplab v2 models whose backbones are Inception v4 (Szegedy et al., 2016), ResNet (He et al., 2016) and DenseNet (Huang et al., 2017), respectively. As for the third model, the authors took advantage of both adversarial learning and ensemble strategy to boost the performance of UNet-like structure (Lei et al., 2020). From Table 3, we can see our method has better performance compared to the ones in leaderboard by a factor of 1.5%. As compared to the methods listed in the leaderboard, our method does not require ensemble of methods as it generates significantly more compact feature as compared to CNN-based methods as shown in Fig. 7.

4.2.6. Discussion and limitations

As shown in Fig. 6, FTN achieves significant improvement and ranks the first place over five methods on our ISIC 2018 evaluation dataset and outperforms the top three methods listed in the leaderboard on the public evaluation dataset. It is shown that FTN gets the best performance using various metrics apart from JI including: Acc, Sen and DSC. The reasons are that SWT builds the feature pyramid like encoders in traditional encoder-decoder CNNs; the introduction of SPT can simultaneously extract global information efficiently; and the novel transformer decoder overcomes the drawback of the locality of CNNs.

Theoretically, researchers proved that ViT splitting image into 16x16 patches can aggregate information through all-pair global attention. However, recent researches showed that training such networks needs millions of annotated images and this is impractical for medical image segmentation. We introduce the well-established principle, the hierarchical design, back into the Transformer field to complement tokenization by proposing the module referred to SWT. Through this design, we are able to mitigate the requirement of millions of data with labels with less computational complexity, as all-pair self-attention is of quadratic computation complexity in terms of the length of tokens. To reduce the computation and memory usage, we then propose the SPT module. SPT consists of SPP, which extracts global information and reduces the computational complexity from quadratic level to polynomial level. The usage of SPP leads to the efficient feature encoder and decoder. Moreover, it has been proved in the ablation study section that this design does not affect performance. Therefore, FTN outperforms other state-of-the-art methods, and can be expected to be applied to most skin lesions segmentation tasks.

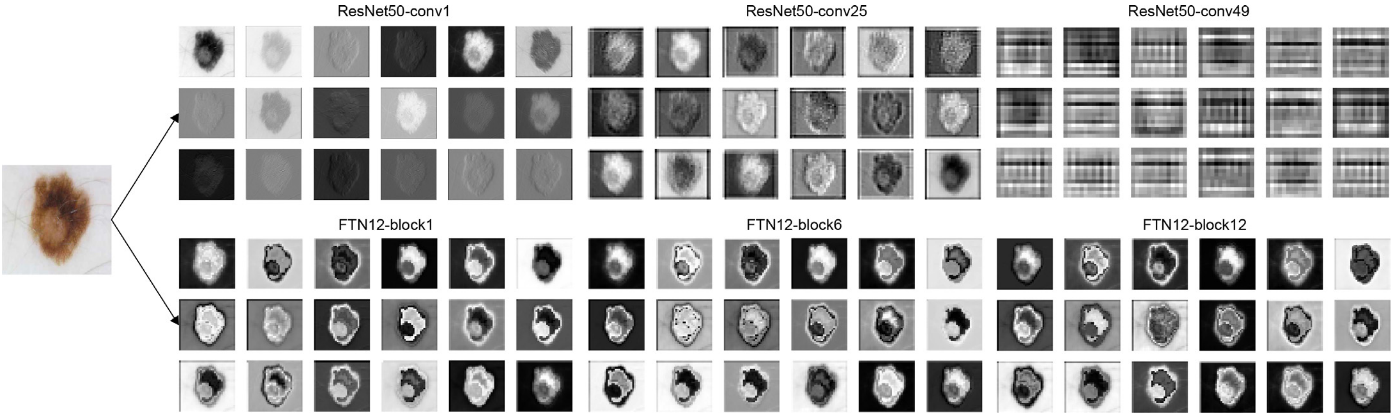


Fig. 7. Feature visualization of ResNet-50 and FTN-12. Features of ResNet-50 are extracted from convolution layer 1, 25 and 49, respectively, whereas FTN's features shown above is reshaped from the tokens of block 1, 6 and 12, respectively.

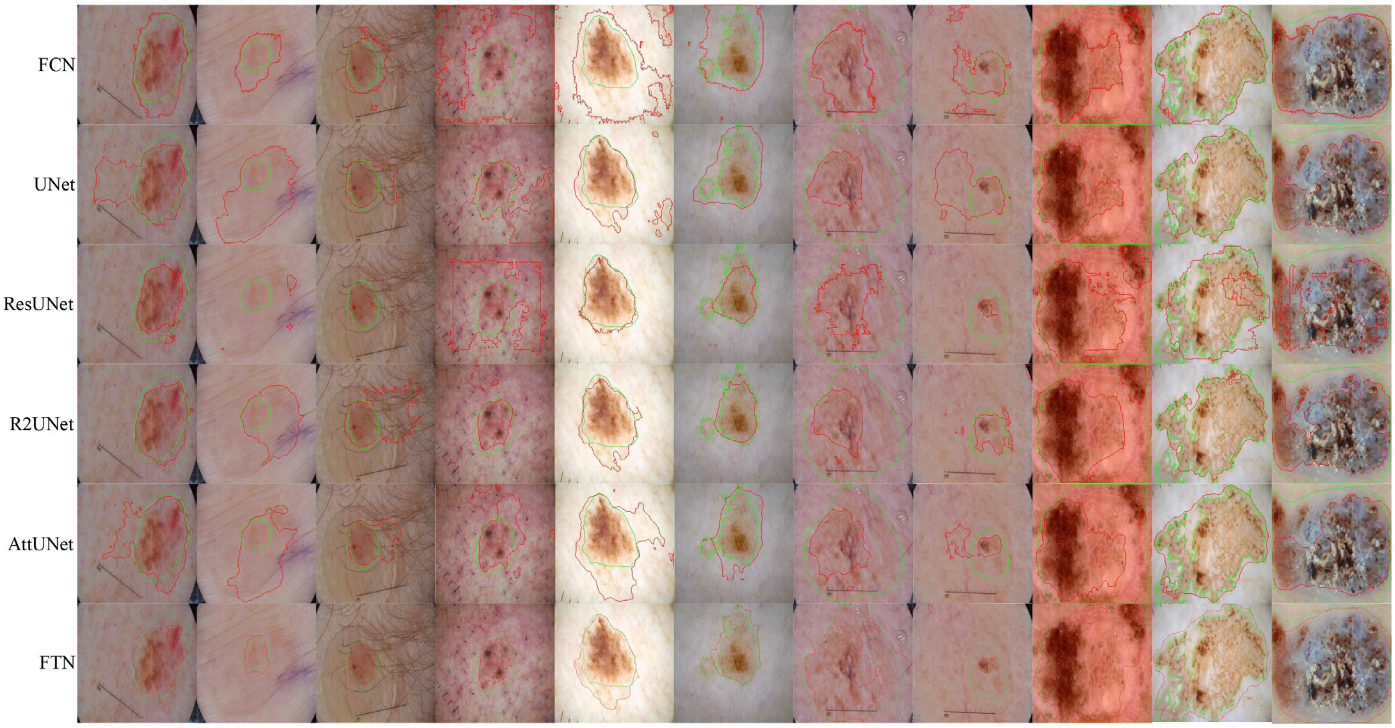


Fig. 8. Visual comparison of failure cases among FTN and other state-of-the-art methods on ISIC 2018 dataset. From the first row to the last row, they represent FCN, UNet, ResUNet, R2UNet, AttUNet and FTN, respectively. The green and red contours are the boundaries of the groundtruth and the predictions, respectively.

Although FTN is able to predict correctly for most cases, it fails to deal with some extreme cases. As shown in Fig. 8, when the contrast between the background and foreground is extremely low, our method is unable to identify lesions as well. Nevertheless, comparing to other methods, our method performs better in terms of boundaries.

4.3. Classification results

In order to apply FTN to disease classification problems, we remove the Transformer decoder and add an adaptive average pooling layer followed by a layer normalization layer and linear projection at the end of the network.

4.3.1. Dataset and preprocessing

For the skin lesion classification task, ISIC 2018 challenge provides 10,025 images and corresponding labels. We randomly choose 8200 images as the training dataset and the remaining 2005 images are added into the testing dataset.

We apply the same geometrical transformation as mentioned in Section 4.1.1 for data augmentation. The only difference is that we resize all images to 224x224.

4.3.2. Experiment setup

All our experiments are implemented on PyTorch frame with an NVIDIA RTX TITAN GPU. We employ an AdamW, a variant of Adam optimizer for 100 epochs using batch size = 100, weight decay = 0.05, and momentum = 0.9. Initial learning rate of 0.0005 and a cosine decay learning rate scheduler are used.

4.3.3. Results

Table 5 shows that our method achieves remarkable results in the skin lesions classification problem. AUC represents the area under the curve, while MAUC denotes MAUC represents the average AUC (the aera under the curve) over all categories (Melanoma,

Table 5
Results of skin lesion classification.

Model	Acc	Sens	Pre	Spec	AUC	MAUC
VGG-16	94.3	76.2	72.7	96.5	86.4	96.8
ResNet-18	95.3	71.4	83.3	98.3	84.8	95.5
ResNet-50	93.3	61.9	72.2	97.1	79.5	90.6
ResNet-100	90.7	47.6	58.8	95.9	71.8	83.1
DenseNet-121	94.3	61.9	81.2	98.3	80.1	93.4
DenseNet-169	94.3	76.2	72,7	96.5	86.4	96.1
Ours	89.6	85.7	51,4	90.1	87.9	94.2
Ours+Seg	92.7	85.7	62.1	93.6	89.7	97.3

Table 6
Results of skin lesion classification.

Model	Score
All Data Are Ext	94.90
Ours	94.85
aloe-18	94.85
Deloitte Analytics Spain-50	94.84

Melanocytic nevus, Basal cell carcinoma Actinic keratosis / Bowens disease (intraepithelial carcinoma), Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis), Dermatofibroma, Vascular lesion). FTN ranks top over three metrics: Sens, Spec and AUC. In comparison with ResNet-50, AUC and Sens increase 9.5% and 27.7%. In addition, as the training sample is extremely limited in terms of training a robust disease classification, we adopt transfer learning to further improve the classification performance. We use FTN pre-trained on the segmentation dataset to stabilize the training process in disease classification. It can be seen that the model pre-trained on the segmentation dataset can further boost the performance of classification model.

4.3.4. Comparison with state-of-the-art methods

To demonstrate that FTNs are able to achieve state-of-the-art, we compare our methods with state-of-the-art methods on ISIC 2020 challenge dataset. The training set of ISIC 2020 challenge dataset provides 33,126 images and corresponding labels. However, there are only 584 melanomas.Given the small percentage of melanomas in ISIC 2020 challenge dataset, we use ISIC 2019 challenge data as external data which has more melanomas to alleviate class imbalance. The testing set has 10,982 images. We evaluate our method’s performance by submitting classification results on the testing set to the official server. The final submission of our method is the ensemble of 4 FTN-12 trained with different image sizes (224, 384, 512, 768). On ISIC 2020 challenge dataset, the only metric (score) they used to sort results is AUC. As shown in Table 6, our method achieved 94.85 in terms of AUC and ranked the second on the private leaderboard. Here, we briefly introduce other listed methods. The best result was achieved by team ’All Data Are Ext’ which ensembles 18 models trained on both ISIC 2019 and 2020 challenge datasets. ’aloe-18’ adapted the similar strategy, but the total number of models they used to ensemble are far fewer than the first team. In addition, they retrained their models using pseudolabels. As for ’Deloitte Analytics Spain-50’, their final submission is an ensemble of 8 models using ISIC 2019 and 2020 challenge dataset. According to the result, we found that our method can achieve state-of-the-art without ensembling a large number of models.

5. Discussion and limitations

Table 7
Ablation results of SWT.

Model	Acc	Sens	Spec	JI	DSC
ViT	94.0	81.2	98.9	74.2	83.6
FTN	96.2	90.2	97.5	82.8	89.8

Table 8
Ablation results of skip connection.

Model	Acc	Sens	Spec	JI	DSC
$\text{FTN}_{wo\ skip}$	94.4	89.3	85.7	76.1	84.6
FTN_{fcn}	96.2	90.0	97.5	82.6	89.6
FTN	96.2	90.2	97.5	82.8	89.8

As shown in Table 5, our methods which trained from scratch or fine-tuning from the segmentation ranks at the first place in terms of AUC. We believe that the inferior performance in terms of other metrics are contributed by the insufficiency of annotated data. Unlike the segmentation task, the classification is an image-level task where each model is trained with image level label. In theory, though the receptive field of CNN cover the whole image by using downsampling and overlapping kernels, the inductive bias of CNN force itself to focus on nearby image elements which leads to great sample efficiency. However, FTN, which fully relies on the self-attention mechanism, is prone to over-fit on small datasets. It can be seen that the Pre and Spec is improved if we used the FTN pre-trained on the segmentation dataset. Though FTN didn’t achieve superiority over all metrics, it performs better in terms of Sens and AUC.

5.1. Ablation studies

To verify the effectiveness of newly-introduced components (namely SWT, Transformer decoder and SPT), we conduct ablation studies on our proposed method on skin lesion segmentation using the same setting illustrated in Section 4.2.

5.1.1. Sliding window tokenization

To identify the effect of SWT, we conduct the following ablation studies using two models: ViT and FTN, where ViT directly uses tokens generated by flattening 16x16 patches. The results are shown in Table 7. Compared with ViT, FTN with SWT yields 8.6% improvement in terms of JI.

5.1.2. Skip connection

Similar to other medical image segmentation tasks using deep hierarchical features, skin lesion segmentation also needs to consider between semantics and fine boundary. As low level features in feature pyramid encodes more location information and high level feature contains more semantics, many encoder-decoder CNNs use skip-connections to boost the feature richness in the decoder. Among encoder-decoder CNNs, the design of skip-connections falls into two categories: FCN-style and UNet style. Here, we design an ablation study to explore whether skip-connections are also helpful for FTN and which style (FCN style vs. UNet style) is more effective. In this ablation study, we compare the experimental results of three different models: $\text{FTN}_{wo\ skip}$, FTN_{fcn}, FTN. $\text{FTN}_{wo\ skip}$ represents the model with the same architecture as FTN but without skip connections. The FTN with FCN style decoder is named as FTN_{fcn}, while the FTN with the default setting UNet style decoder is named as FTN. From Table 8, we can see UNet-style decoder is the most effective decoder for skin lesion segmentation.

5.1.3. Spatial pyramid pooling

The selected pooling method has an impact on the performance of SPT. The widely adopted pooling methods include: average pooling and max pooling. There are two options for the combination with SPP: pyramid average pooling and pyramid max pooling. To

Table 9
Ablation results of SPP.

Model	Acc	Sens	Spec	JI	DSC
Max	95.9	88.9	97.7	82.1	89.2
Avg	96.2	90.5	97.4	82.6	89.7
Pyramid Max	96.1	90.0	97.3	82.4	89.5
Pyramid Avg	96.2	90.3	97.5	82.9	89.8

Table 10
Ablation results of SPP.

Model	Acc	Sens	Spec	JI	DSC
FTN-4	95.6	89.5	96.8	80.9	88.0
FTN-8	96.1	89.9	97.1	81.8	89.1
FTN-12	96.2	90.2	97.5	82.8	89.8

Table 11
Ablation results of strides of the last Bilinear upsampling.

Model	Acc	Sens	Spec	JI	DSC
Stride 8	95.6	89.8	96.2	79.4	87.6
Stride 4	96.2	90.2	97.5	82.8	89.8
Stride 2	96.0	90.9	96.2	81.5	88.5
Stride 1	95.7	90.4	94.1	80.8	88.1

study the performance of different pooling methods, we conduct an ablation study on four models: Max, Avg, Pyramid max, Pyramid Avg. Here, Max represents max pooling, while Avg is average pooling. As for Pyramid, it represents whether using multiple pooling layers to generate pooling features of different resolutions (Mahbod et al., 2020). It shows that average pyramid pooling performs better than other three approaches in Table 9.

5.1.4. Depth

As the depth of CNNs is of great importance, we build three FTNs with backbones of different depths to verify the effect. As shown in Table 1, the three FTNs we constructed are FTN-3, FTN7 and FTN-12, where the number represents the amount of SPT in stage 3. From Table 10, we find that the depth of FTN increases the performance accordingly. Thus, we use FTN-12 in both segmentation and classification experiments.

5.1.5. Upsampling stride

Different strides of the final upsampling have different performance. Bilinear upsampling with high strides for generating final outputs may be unable to recover boundaries of skin lesions, while bilinear upsampling with low strides may be over-fitting on irregular boundaries and is computationally inefficient.

In order to find which upsampling stride leads to the best performance, we conduct the experiments on bilinear upsampling with strides 8, 4, 2 and 1. The results of strides of upsampling strides are shown in 11. It turns out that the best performance is achieved by stride 4. Thus, we stick to use stride 4 in the final bilinear upsampling.

5.2. Feature visualization

To better understand where the success of FTN stems from, we need to understand what features our model has learned. We plot the visualization results of the features in Fig. 7, which shows the activated features by feeding an image into our FTN and well-established CNN model ResNet-50. From left to right, features from the early, middle and later layers are shown and 18 activation maps are presented for each layer by random selection. As we can see, ResNet forms more discriminative features as the depth increases and constructs semantic abstracts at the last layer. However, FTN can successfully search the representative features even at the first block. We also observe that the feature of the lesion become more and more distinct and compact as the depth increases which reveals the effectiveness of FTN. Note that FTN can even discover some lesion attributes such as pigment networks, streaks and abrupt border cutoff to assist the segmentation and classification without explicit supervision.

6. Conclusion

In this paper, we propose an FTN consisting of SWT and SPT for skin lesion segmentation and classification. SWT extracts feature pyramids instead of using convolutional layer. SPT implements self-attention in an efficient and effective way, as we perform self-attention on downsampled features. Compared to the traditional CNN, FTN is purely composed of Transformer layer, which significantly increases global information essential for skin lesion segmentation. In addition, we propose a Transformer decoder which can efficiently aggregate low and high level features. With these novel modules, FTN can be trained from scratch and achieve better segmentation and classification results than other state-of-the-art methods. We conduct extensive experiments using the public ISIC 2018 skin lesion segmentation and classification datasets. Our comparisons with other methods show that FTN is not only more efficient but also able to achieve better performance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Xinzi He: Formal analysis, Data curation, Writing – review & editing. **Ee-Leng Tan:** Formal analysis, Data curation, Writing – review & editing. **Hanwen Bi:** Formal analysis, Data curation, Writing – review & editing. **Xuzhe Zhang:** Formal analysis, Data curation, Writing – review & editing. **Shijie Zhao:** Formal analysis, Data curation, Writing – review & editing. **Baiying Lei:** Formal analysis, Data curation, Writing – review & editing.

Acknowledgement

This work was supported partly by the Key-Area Research and Development Program of Guangdong Province (2019B010110001), National Natural Science Foundation of China (Nos. 61806167, 82060336, 61936007, U1801265, 61871274 and 61801305), Guangdong Basic and Applied Basic Research Foundation (Nos. 2019B1515120029 and 2019A1515111205), Science and Technology Support Program of Guizhou Province (No. [2021]432) and (Nos. SGDX20201103095802007, KCXFZ20201221173213036, JCYJ2018 0507184647636, JCYJ20190808155618806, GJHZ20190822095414576, and JCYJ20190808145011259).

References

Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K., 2019. Recurrent residual u-net for medical image segmentation. J. Med. Imaging 6 (1), 014006. doi:10.1117/1.JMI.6.1.014006.

Balch, C.M., Gershenwald, J.E., Soong, S.-j., Thompson, J.F., Atkins, M.B., Byrd, D.R., Buzaid, A.C., Cochran, A.J., Coit, D.G., Ding, S., Eggermont, A.M., Flaherty, K.T., Gimotty, P.A., Kirkwood, J.M., McMasters, K.M., Mihm, M.C., Morton, D.L., Ross, M.I., Sober, A.J., Sondak, V.K., 2009. Final version of 2009 AJCC melanoma staging and classification. Journal of Clinical Oncology 27 (36), 6199–6206. doi:10.1200/JCO.2009.23.4799.

Beal, J., Kim, E., Tzeng, E., Park, D.H., Zhai, A., Kislyuk, D., 2020. Toward trans-former-based object detection. arXiv:2012.09958 [cs].

Celebi, M., Iyatomi, H., Schaefer, G., Stoecker, W.V., 2009. Lesion border detection in dermoscopy images. Computerized Medical Imaging and Graphics 33 (2), 148–153. doi:10.1016/j.compmedimag.2008.11.002.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40 (4), 834–848. doi:10.1109/TPAMI.2017.2699184.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A., 2019. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). arXiv:1902.03368 [cs].

Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A., 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, Washington, DC, pp. 168–172. doi:10.1109/ISBI.2018.8363547.

Combalia, M., Codella, N.C.F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., Malvehy, J., 2019. BCN20000: Dermoscopic lesions in the wild. arXiv:1908.02288 [cs, eess].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929 [cs].

Emre Celebi, M., Wen, Q., Hwang, S., Iyatomi, H., Schaefer, G., 2013. Lesion border detection in dermoscopy images using ensembles of thresholding methods. Skin Research and Technology 19 (1), e252–e258. doi:10.1111/j.1600-0846.2012.00636.x.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115–118. doi:10.1038/nature21056.

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A., 2020. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. MethodsX 7, 100864. doi:10.1016/j.mex.2020.100864.

Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., Schlaefer, A., 2020. Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. IEEE Trans. Biomed. Eng. 67 (2), 495–503. doi:10.1109/TBME.2019.2915839.

Gutman, D., Codella, N.C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A., 2016. Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv:1605.01397 [cs].

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y., 2021. Transformer in transformer. arXiv:2103.00112 [cs].

Hasan, M.K., Dahal, L., Samarakoon, P.N., Tushar, F.I., Martí, R., 2020. DSNet: Automatic dermoscopic skin lesion segmentation. Comput. Biol. Med. 120, 103738. doi:10.1016/j.compbiomed.2020.103738.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA, pp. 770–778. doi:10.1109/CVPR.2016.90.

He, X., Yang, S., Li, G., Li, H., Chang, H., Yu, Y., 2019. Non-Local Context Encoder: Robust Biomedical Image Segmentation against Adversarial Attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, pp. 8417–8424. doi:10.1609/aaai.v33i01.33018417.

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2019. Squeeze-and-excitation networks. arXiv:1709.01507 [cs].

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

Ibtehaz, N., Rahman, M.S., 2020. Multiresunet : rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Networks 121, 74–87. doi:10.1016/j.neunet.2019.08.025.

Jahanifar, M., Tajeddin, N.Z., Koohbanani, N.A., Gooya, A., Rajpoot, N., 2019. Segmentation of skin lesions and their attributes using multi-scale convolutional neural networks and domain specific augmentations. arXiv:1809.10243 [cs, stat].

Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., Qin, J., Chen, S., Wang, T., Wang, S., 2020. Skin lesion segmentation via generative adversarial networks with dual discriminators. Med Image Anal 64, 101716. doi:10.1016/j.media.2020.101716.

Mahbod, A., Schaefer, G., Wang, C., Ecker, R., Dorffner, G., Ellinger, I., 2020. Investigating and exploiting image resolution for transfer learning-based skin lesion classification. arXiv:2006.14715 [cs].

Pehamberger, H., 1993. In vivo epiluminescence microscopy: improvement of early diagnosis of melanoma. J. Invest. Dermatol. 100 (3), 7.

Peruch, F., Bogo, F., Bonazza, M., Cappelleri, V.-M., Peserico, E., 2014. Simpler, faster, more accurate melanocytic lesion segmentation through MEDS. IEEE Trans. Biomed. Eng. 61 (2), 557–565. doi:10.1109/TBME.2013.2283803.

Qian, C., Liu, T., Jiang, H., Wang, Z., Wang, P., Guan, M., Sun, B., 2018. A detection and segmentation architecture for skin lesion segmentation on dermoscopy images. arXiv:1809.03917 [cs].

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Vol. 9351. Springer International Publishing, Cham, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J., Soyer, H.P., 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data 8 (1), 34. doi:10.1038/s41597-021-00815-z.

Schaefer, G., Krawczyk, B., Celebi, M.E., Iyatomi, H., 2014. An ensemble classification approach for melanoma diagnosis. Memetic Computing 6 (4), 233–240. doi:10.1007/s12293-014-0144-8.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal 53, 197–207. doi:10.1016/j.media.2019.01.012.

Shelhamer, E., Long, J., Darrell, T., 2016. Fully convolutional networks for semantic segmentation. arXiv:1605.06211 [cs].

Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A., 2021. Cancer statistics, 2021. CA Cancer J Clin 71 (1), 7–33. doi:10.3322/caac.21654.

Silveira, M., Nascimento, J.C., Marques, J.S., Marcal, A.R.S., Mendonca, T., Yamauchi, S., Maeda, J., Rozeira, J., 2009. Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. IEEE J Sel Top Signal Process 3 (1), 35–45. doi:10.1109/JSTSP.2008.2011119.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: International Conference on Learning Representations.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Tang, P., Liang, Q., Yan, X., Xiang, S., Zhang, D., 2020. GP-CNN-DTEL: Global-part CNN model with data-transformed ensemble learning for skin lesion classification. IEEE J Biomed Health Inform 24 (10), 2870–2882. doi:10.1109/JBHI.2020.2977013.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning. PMLR, pp. 10347–10357.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5 (1), 180161. doi:10.1038/sdata.2018.161.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 5998–6008.

Vestergaard, M.E., Macaskill, P., Holt, P.E., Menzies, S.W., 2008. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. British Journal of Dermatology 159 (3), 669–676. doi:10.1111/j.1365-2133.2008.08713.x.

Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.-C., 2020. MaX-Deeplab: end-to-end panoptic segmentation with mask. transformers.

Xie, Y., Zhang, J., Lu, H., Shen, C., Xia, Y., 2021. SESV: Accurate medical image segmentation by predicting and correcting errors. IEEE Trans Med Imaging 40 (1), 286–296. doi:10.1109/TMI.2020.3025308.

Xie, Y., Zhang, J., Xia, Y., Shen, C., 2020. A mutual bootstrapping model for automated skin lesion segmentation and classification. arXiv:1903.03313 [cs].

Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.-A., 2017. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging 36 (4), 994–1004. doi:10.1109/TMI.2016.2642839.

Yu, Z., Jiang, X., Zhou, F., Qin, J., Ni, D., Chen, S., Lei, B., Wang, T., 2019. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. IEEE Trans. Biomed. Eng. 66 (4), 1006–1016. doi:10.1109/TBME.2018.2866166.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F.E., Feng, J., Yan, S., 2021. Tokens-to-token ViT: training vision transformers from scratch on imagenet. arXiv:2101.11986 [cs].

Yuan, Y., Chao, M., Lo, Y.-C., 2017. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. IEEE Trans Med Imaging 36 (9), 1876–1886. doi:10.1109/TMI.2017.2695227.

Yuksel, M., Borlu, M., 2009. Accurate segmentation of dermoscopic images by image thresholding based on type-2 fuzzy logic. IEEE Trans. Fuzzy Syst. 17 (4), 976–982. doi:10.1109/TFUZZ.2009.2018300.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2019. Attention residual learning for skin lesion classification. IEEE Trans Med Imaging 38 (9), 2092–2103. doi:10.1109/TMI.2019.2893944.

Zhang, Z., Liu, Q., Wang, Y., 2018. Road extraction by deep residual u-net. IEEE Geosci. Remote Sens. Lett. 15 (5), 749–753. doi:10.1109/LGRS.2018.2802944.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, pp. 6230–6239. doi:10.1109/CVPR.2017.660.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L., 2021. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers. In: 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6881–6890.

Zhou, H., Schaefer, G., Celebi, M.E., Lin, F., Liu, T., 2011. Gradient vector flow with mean shift for skin lesion segmentation. Computerized Medical Imaging and Graphics 35 (2), 121–127. doi:10.1016/j.compmedimag.2010.08.002.

Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X., 2019. Asymmetric Non-Local Neural Networks for Semantic Segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Seoul, Korea (South), pp. 593–602. doi:10.1109/ICCV.2019.00068.