IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 November 2022, accepted 17 November 2022, date of publication 21 November 2022,
date of current version 28 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3224005

RESEARCH ARTICLE

Semi-Supervised Skin Lesion Segmentation With Coupling CNN and Transformer Features

MOHAMMAD D. ALAHMADI[1], (Member, IEEE), AND WAJDI ALGHAMDI[2]
[1]Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah 23890, Saudi Arabia
[2]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Mohammad D. Alahmadi (mdalahmadi@uj.edu.sa)

ABSTRACT An automatic skin lesion segmentation algorithm not only facilitates the dermatologist's workload on skin cancer analysis but also provides a platform for early cancer prediction. Over the years, several deep learning methods have been proposed to addDepartment of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabiaress the skin lesion segmentation problem. However, training deep models usually requires a large-scale annotated dataset, which is not feasible in the medical domain due to the annotation burden. In addition, the low data regime highly increases the overfitting potential for the neural network. To address these limitations in an end-to-end manner, we propose to incorporate unlabelled samples during the training process. Our network offers a semi-supervised training schema, wherein the first stage performs a supervised training strategy to learn semantic segmentation map while the second step focuses on the unsupervised technique to enrich the encoder module. Specifically, unlike the literature work on skin lesion segmentation, we design a surrogate task on top of the convolutional and Transformer representations to learn data-driven features from the image itself to alleviate the requirement of the large annotated dataset. The effectiveness of the proposed method is demonstrated using three different skin lesion segmentation datasets, namely ISIC 2018 (dice score 0.905), ISIC 2017 (dice score 0.898) and PH2 (dice score 0.940). Particularly we observed that including the unsupervised samples can increase the dice score by 2%.

INDEX TERMS Skin lesion, CNN, transformer, semi-supervised, segmentation.

I. INTRODUCTION

Computer-Aided Diagnosis (CAD) is a severe counterpart for medical experts to assist them in their daily treatment diagnosis by interpreting medical images [1]. Deep Learning (DL) brought a solid foundation for computer vision tasks, and CAD systems are no exception [2], [3]. Among many medical image analysis tasks, image segmentation is a de facto step in which its presence is not negligible. Medical image segmentation is embedded in various medical applications, including skin lesion segmentation. Human skin tissue consists of three types, i.e., dermis, epidermis, and hypodermis. The epidermis is a susceptible tissue, which under severe solar radiation, could trigger the embedded melanocytes to produce melanin at a significant level. Fatal skin cancer is a result of melanocyte growth, which is known as melanoma. The *American Cancer Society* anticipated the approximate melanoma skin cancer cases around 99,780, with death cases of 7,650, 7.66% of all cases [4] for 2022. Early disease recognition plays a crucial role in medical diagnosis, as it has been reported that detection of melanoma in early phases could increase the relative survival rate to 90% [5]. Although dermatologists could detect malignant melanoma in medical images from dermoscopy, it could be a tiresome task and needs the proficiency of a dermatologist [6]. To this end, skin lesion segmentation is highly desired and could assist the dermatologist with appropriate treatment.

Automatic segmentation plans to cut out desired regions from irrelevant counterparts by pixel-wise classification. Hence, for skin lesions, the segmentation task is a binarization most of the time, separating the malignant region from its neighbor. Explicitly, automated skin lesion segmentation is interfered with by occasional intraclass factors, i.e., skin colors, textures, tissue size, the geometrical shape of a lesion,

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyu Zhou.

illumination and contrast due to the various dermoscopic imaging tools, and the interclass factors such as the presence of hair, blood vessels, ruler marks, and occlusion. Conventional automatic skin lesion segmentation techniques are typically based on classical computer vision and machine learning approaches such as adaptive thresholding [7], active contours [8], region growing [9], and unsupervised clustering [10], [11]. The methods, as mentioned earlier, heavily depend on reliable engineered handcraft features to determine lesion boundaries from the background. Therefore, DL methods revolutionized this domain through their end-to-end automatic feature extraction and classification baseline.

In the last decade, Ciresun et al. [12] made the first attempt to use the Convolutional layers in the medical image segmentation task. Afterward, several architectures were proposed to enhance the segmentation performance, not particularly in the medical domain, such as Fully Convolutional Network (FCN) [13], FC-DenseNet [14], and U-Net [15] for medical image segmentation. These architectures advanced the image segmentation such as the images obtained from medical domain. U-Net, an encoder-decoder alongside skip connections network, has demonstrated tremendous State of the Arts (SOTA) performance in medical image segmentation since 2015. To this end, various modifications have been introduced for various medical applications with different image modalities, e.g., U-Net++ [16], U-Net3+ [17], ResU-Net [18], DenseU-Net [19], 3D U-Net [20], V-Net [21], S3D U-Net [22]. Ramani et al. [23] used seminal U-Net for melanoma lesion segmentation in the skin lesion segmentation task. Bi et al. [24] employed a cascade multi-stage FCN ensemble model to produce a segmentation map. MS-UNet [25] is a multi-stage U-Net-based model that utilizes a deep supervision loss schema to learn intermediate features better which in turns increases the segmentation performance. These methods suffer from a common problem as they cannot capture long-range context information for the accurate localization of semantic features to produce monotonous segmentation results. This drawback is caused by the Convolutional Neural Network (CNN) deficiency due to the convolution layers' limited receptive field. Loss of abstract localization features through the layers is not the desired result for semantic segmentation especially in the medical domain that requires an accurate extraction of boundary regions of organs and tissues. Thus, supplementing long-range dependencies and learning conceptualized features from the image is required.

The strength of U-Net is based on the symmetrical design of the encoder-decoder and the intersection of the encoder path to the decoder path with skip connections. Feature representation in CNN layers loses its localization due to the successive convolution and downsampling operations. In addition, the successive upsampling operation makes the model lose more detailed spatial features. Although the U-Net structure tries to alter this loss of global and contextual information with skip connections, these shortcuts are still insufficient. As a result, this outline inspires that segmentation representation improves drastically if the model hinders the loss of spatial information besides capturing long-range dependencies and integrating them into the decoder path. Moreover, a mechanism to include un-labeled dataset in the training stage is not presented in the U-Net model to enrich the feature representation capacity.

In this paper, we propose to couple CNN and Transformer encoders to capture both local and global representation. Nevertheless, training CNN/Transformer models usually require a large labelled dataset, which is not always available in the medical domain. Besides that, although integrating large encoder modules (e.g., CNN/Transformer) increases the model freedom (high number of parameters) to learn underlying data distribution, a lack of labelled dataset results in an unstable and overfitted model. To overcome this limitation, we propose to incorporate the unlabelled samples during the training process. Particularly, we offer a semi-supervised training technique, where the first step takes the advantage of the supervised training strategy to learn semantic segmentation map whereas the second step focuses on leveraging the unsupervised data during the training process. Specifically, we design a surrogate task to learn data-driven features from the image itself to alleviate the requirement of the large annotated dataset.

Our contributions can be summarized as follows:

- Coupling CNN and Transformer modules to model local and global representation
- Semi-supervised technique to utilize unlabelled samples during the training process
- State-of-the-art resutls on three challenging skin lesion segmentation benchmarks

II. RELATED WORKS

A. SKIN LESION SEGMENTATION

In contradiction with conventional feature engineering methods, DL does not need further hand-crafted feature extraction, and can be used effectively in skin lesion segmentation [26], [27]. Broadly speaking, Yuan et al. [28] proposed a CNN with deep layers with small convolutional kernels to generalize their model with various image acquisition qualities. Alahmadi et al. [29] proposed a network that captures both local and global representation of medical images using a supervised learning technique. MSU-Net [25] has been proposed as a multi-stage U-Net-based network that simultaneously captured low-level features with fused context information in two successive stages of U-Net with a recursive perspective. Taghanaki et al. [30] proposed a modification for the U-Net skip connection to capture the most informative channel in feature map channels in each stage and transfer it to the corresponding stage in the decoder path. This transformation minimized the parameters, which led to light weighing of the network and better feature aggregation. DSM [31] utilized a multi-scale connection block within skip connection to handle the tissue variation size and aggregate the multi-stage output in the decoder path in a deep supervision strategy. DPFCN [32] employed a dense pooling schema

with overlapping windows to acquire densely feature maps. Xie et al. [33] proposed the MB-DCNN model with two segmentation networks, i.e., coarse-SN and enhanced-SN, alongside a mask-CN classification network. The first localization information extracted with coarse-SN was transfered to the classification network, and the resultant class activation map was fed into enhanced-SN to obtain accurate lesion segmentation. Pourya et. al. [34] addressed the automatic skin lesion segmentation challenge from a differnet perspective. Their design offers a multi-scale representation with a scale-wise fusion mechanism to alleviate the effect of overlapped background with the object of interest (skin lesion). More precisely, their approach utilizes the dilated pyramid convolution to capture multi-scale representation, by proposing a scale-wise fusion module they model the interaction among scales to enrich feature representation in the boundary area.

All the reviewed methods have a mutual bottleneck of ignoring global context information, which is a crucial factor in the medical image segmentation task. Hence, a parallel module to compensate for the loss of global contextual representation seems necessary.

B. TRANSFORMER

Li et al. [35] employed dense deconvolutional layers with cascade pooling to extract features hierarchically to capture long-range dependencies. SSP [36] developed an FCN with a shape prior information to preserve the global context of the segmentation region by penalizing non-star segmentation results. SegAN [37] leaned the segmentation map by an adversarial learning strategy with a multi-scale loss to enhance the long-range spatial dependencies. Wang et al. [38] leveraged simultaneously spatial and channel attention to recalibrate the feature representation by updating each feature value by a weighted sum of all other features. FCA-Net [39] proposed a factorized channel attention block to determine relevant channel patterns from feature maps. Abraham et al. [40] inspired by Attention U-Net [41] integrated spatial attention gate in skip shortcuts of encoder-decoder for the better interweaving of localization feature maps and coarse feature maps alongside focal tversky loss. CPFNet [42] applied a pyramid module on feature maps to capture global context. Attention Deeplabv3+ [43] applied a two-stage attention mechanism to capture informative channels and scale relevant from atrous convolution layers.

Unlike the mentioned attention mechanism, Transformer emerged by Vaswani et al. [44] proposed self-attention mechanism in Natural Language Processing (NLP) domain for machine translation tasks where it was a pure encoder-decoder network. Its success over traditional recursive leveraged modules and layers made it out to a vision domain. The first pioneering Vision Transformer (ViT) by Dosovitskiy et al. [45] was a simple stacked encoder built by Transformer blocks. After the Vision Transformer (ViT) success in major vision tasks and prior knowledge of the importance of attention mechanism in segmentation, ViT is broadly used either as a complement to CNNs or a standalone backbone design in these tasks. TransU-Net [46] was one of the earliest impressions of ViT in medical image segmentation tasks, where it embraced the Transformer as a complement to CNNs in the encoder path to capture long-range dependencies. However, due to the quadratic computational complexity of Transformers, they were not offered as a single standalone backbone until the Swin Transformers [47] for their linear computational complexity versus being the solitarily Transformer. Swin U-Net [48] is a solely Swin Transformer network based on U-Net design. It captures long-range dependencies for better medical segmentation results due to the deformable nature of body organs and tissues.

C. SEMI-SUPERVISED SEGMENTATION

The semi-supervised technique can be categorized into traditional hand-crafted features and novel deep learning-based approaches. The former uses prior knowledge (e.g., clustering) to perform feature matching whereas the deep learning-based methods utilize representational learning to learn data-driven features. An iterative procedure developed by Bai et al. [49], where pseudo labels for mask-free images are predicted by the network and distilled by Conditional Random Forest (CRF), and these labels are used to fed to the network again. Zhang et al. [50] proposed a new Deep Adversarial Network (DAN) to utilize unlabelled data in a semi-supervised way for predicting unannotated images. Yu et al. [51] developed the mean teacher model with uncertainty map guidance for semi-supervised left atrium segmentation. Zhang et al. [52] utilized shape-aware prior information to leverage the unlabelled data and impose a geometric shape constraint on the segmentation output. What all these methods have in common, is their prior knowledge assumption, which might not be feasible for any task. Differently, our unsupervised technique learns a mapping function which is consistent over different augmentations. More precisely, we create two augmented versions of the input image and then fed each augmented image into the encoder module then using an auxiliary decoder module, we minimize the cross-entropy loss between the two generated feature maps. Hence, using cross-entropy loss, our encoder architecture learns the mapping function which is robust to slight variation (e.g., augmentation) and consequently can learn more generic representation from unlabelled samples.

III. PROPOSED METHOD

The overall structure of our proposed network is depicted in Figure 1. To incorporate the unlabelled samples during the training process, our method utilizes an auxiliary decoder module to learn consistency over the augmentation map. In addition, our design offers a combination of CNN and Transformer encoder for robust local to global representation. In the next subsections, each module will be presented comprehensively.

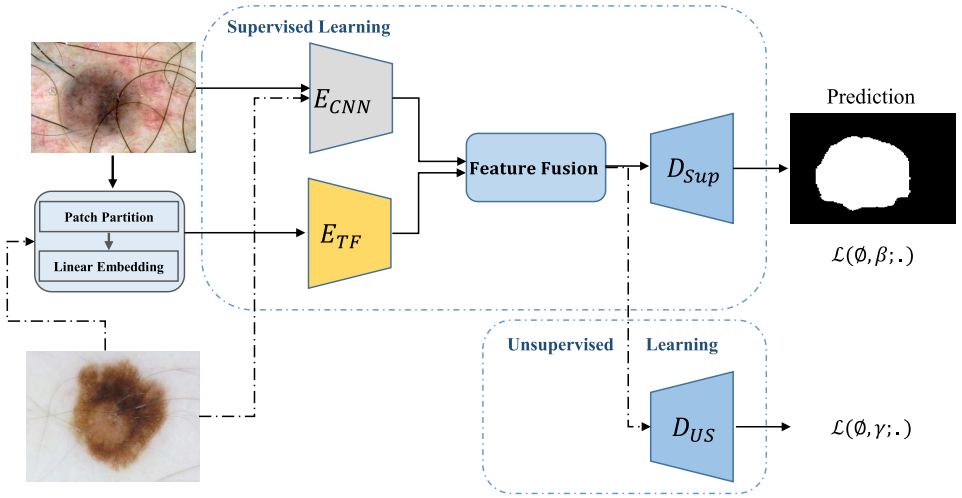


FIGURE 1. The proposed semi-supervised skin lesion segmentation network. In our model, the top stream applies a combination of CNN/Transformer model followed by the decoder module to learn supervised segmentation while the bottom stream utilizes an unsupervised technique to enrich the encoder block by learning an auxiliary task.

A. CNN REPRESENTATION

Figure 1 shows that our proposed method is built with two encoder flow branches that complement each other. We use a seminal U-Net [15] network in the first branch to model CNN representation E_{CNN}. This architecture E_{CNN} parameterized with θ_1, applies successive convolutional layers on a given image $x \in \mathbb{R}^{H\times W\times C}$ (H, W, C are spatial height, width and channels dimension, respectively.) to extract pixel-level contextual information. More precisely, in our design, we follow the original structure of the U-Net model [15] and deploy a four-block encoder architecture, wherein in each block we use two convolutional layers followed by the Relu action and max pooling operations to produce the feature map. The resulting feature map contains local semantic information, however, due to the locality nature of the convolutional operation, it is ineffective in capturing object-level (e.g., global) representation. Therefore, to alleviate this limitation, we utilize the Transformer module as a complementary feature extractor.

B. LONG-RANGE CONTEXTUAL REPRESENTATION

The second branch (E_{TF} parameterized with θ_2) objective is compensating convolutions deficiency in capturing long-range dependencies by utilizing Transformer. Similarly to [45], we feed the input image $x \in \mathbb{R}^{H\times W\times C}$ with respect to the first branch to the Transformer module by dividing it into the $N = [\frac{HW}{p^2}]$ non-overlapping patches where $p \times p$ is the dimension of each patch. Later a patch encoder $E(x_p; \omega)$ applies on serialized patches to project from $p^2 \cdot c$ space to K embedding space. A 1-D learnable positional embedding $I_{\text{pos}} \in \mathbb{R}^{N\times K}$ adds to the projected sequence of each patch to preserve each patch's spatial information:

$$t_0 = \left[x_p^1 I; x_p^2 I; \cdots; x_p^N I\right] + I_{\text{pos}} \tag{1}$$

where $I \in \mathbb{R}^{(p^2\cdot C)\times K}$ denotes the projected patch embedding. We then stack the multiple Transformer blocks to learn long-range dependencies. Each Transfomer block composed with Multi-head Self Attention (MSA) where consists of M parallel self-attention heads to scale different patch interaction learning's:

$$t_i' = \text{MSA}\left(\text{Norm}\left(t_{i-1}\right)\right) + t_{i-1}, \quad i = 1, \cdots, L \tag{2}$$

and Multi Layer Perceptron (MLP) modules to learn long-range contextual dependencies by:

$$t_i = \text{MLP}\left(\text{Norm}\left(t_i'\right)\right) + t_i', \quad i = 1, \cdots, L. \tag{3}$$

Norm() depicts layer normalization [53] and $t_i \in \mathbb{R}^{\frac{HW}{p^2}\times d}$ represents encoded semantic representation. In our design we used the public implementation of the vision transformer [45] with three self attention head to encode the global representation.

C. FEATURE FUSION

As presented in the previous two subsections, our encoder module applies both CNN and Transformer encoders to extract local and global representation. To combine these two feature sets, we first reshape the Transformer representation into the same spatial dimension as the CNN feature set, then we simply concatenate these two feature sets to create the final encoder representation.

D. SEGMENTATION DECODER

Our decoder module utilizes the same structure as seminal U-Net model to produce the segmentation map. More precisely, in our supervised section we utilize four block CNN decoder (similar to the CNN encoder but with replaced upsampling instead of pooling operation) D_{SUP} with parameters ϕ to progressively increase the spatial dimension while reducing the feature map to predict the skin lesion area. We apply dice loss $\mathcal{L}(\theta, \phi; .)$ between the predicted segmentation map and the ground truth mask to learn the segmentation task in a supervised manner, where $\theta = \theta_1 \cup \theta_2$ indicate the CNN and

Transformer encoder's parameters and ϕ represents network parameters related to the segmentation task.

E. SURROGATE TASK

One of the challenges in medical image segmentation is to provide large annotated dataset to train the segmentation network. To tackle this issue, we propose integrating a supplementary decoding head D_{US} to alleviate the lack of labelled data during the training process. To this end, we propose to include an auxiliary loss function to reduce the dissimilarity between the representation of two augmented versions (x_1 and x_2) of the same image, where $x_i = Aug(x)$ and $Aug()$ indicates a random augmentation function. We denote this loss as $\mathcal{L}(\theta, \gamma; .)$, where γ represents network parameters related to the surrogate task. To model a surrogate task, several methods have been proposed in the literature, including predicting rotation [54], solving jigsaw puzzles [55], and filling removed parts of an image [56]. Note that skin tissue in dermatology concept is direction variant. In contrast, consider a rotated car image with wheels above the car roof; in this example, it is obvious to predict the rotation [57]. Therefore, it is evident that due to the nature of the application, we should consider an appropriate surrogate task. To accomplish this, we utilize an auxiliary dataset, $D_{unlabelled} = \{(X_i^{\mathcal{U}})\}_{i=1}^{N}$, with N unlabelled samples. We apply data augmentation technique two times to each images, resulting in a peer-to-peer mapping of augmented images $(Y_{i,1}^{\mathcal{U}}, Y_{i,2}^{\mathcal{U}})$ for each image in the dataset $(X_i^{\mathcal{U}})$. Using MSE loss, we force the encoder module to learn the feature representation space which is robust to slight variation (e.g., reducing the feature dissimilarity for two augmented version of the image). Choosing MSE loss over other losses was empirical, and it is evidence of using MSE loss as a reconstruction loss in the Auto encoder-decoder concept. Equation 4 is formulated the used MSE loss as follows:

$$\mathcal{L}_{sur}(\theta, \gamma; Y_{i,1}^{\mathcal{U}}, Y_{i,2}^{\mathcal{U}}) = -\frac{1}{N} \sum_{i}^{N_{\mathcal{U}}} \sum_{j}^{H \times W} (Y_{i,1}^{\mathcal{U}} - Y_{i,2}^{\mathcal{U}})^2, \quad (4)$$

F. JOINT OBJECTIVE

The final objective function during the training is a weighted sum of two counterparts dedicated to the semantic segmentation task and surrogate task, respectively. The first term, $\mathcal{L}(\theta, \phi; .)$, is a function of the parameters θ and ϕ of semantic segmentation encoder-decoder term. Also, $\mathcal{L}(\theta, \gamma; .)$ is a function of encoder parameters θ and surrogate network parameters of γ. Equation 5 represents the joint loss functions of network as follows:

$$\min_{\theta, \phi, \gamma} \mathcal{L}_{segmentation}(\theta, \phi; D_{train}) + \lambda \mathcal{L}_{sur}(\theta, \gamma; Y^{\mathcal{U}}), \quad (5)$$

where λ is a regularized term to control the weight of surrogate task.

IV. EXPERIMENTS

A. DATASETS

We applied our proposed method to three publicly available dermoscopic datasets to demonstrate the efficacy of our module. The first two datasets belong to *International Skin Imaging Collaboration (ISIC)*, i.e., *ISIC 2017* [1] [58] and *ISIC 2018* [2] [59]. The last dataset is PH^2 [3] [60], published by the dermatology service of *Pedro Hispano Hospital*, Matosinhos, Portugal. Dataset specifications are as follows:

- *ISIC 2017* - This dataset contains 2,750 RGB images, where 2,000 images are for training, 150 images are for validation, and 600 images belong to the test phase. Melanoma-positive cases form 18.7%, 20%, and 19.5% of each set portion, respectively. Samples resolution varies from 540×722 to $4,499 \times 6,748$ pixels.
- *ISIC 2018* - This dataset contains 3,694 RGB images, where originally 2,594 images are for training, 100 images are for validation, and 1,000 images belong to the test phase. Melanoma-positive cases form 20% of the training set portion. As the available testing dataset was unlabelled, we randomly split the training dataset to 1,815 images for training, 259 for validation and 520 for testing. Samples resolution varies from 540×576 to $4,499 \times 6,748$ pixels.
- PH^2 - This dataset contains 200 RGB images, where we randomly split it into 140 samples for training, 20 samples for validation, and 40 samples for the testing set. Melanoma-positive cases form 20% of the dataset. Samples resolution varies from 553×763 to 577×769 pixels.

B. IMPLEMENTATION DETAILS

We implemented our proposed method using PyTorch framework on a single NVIDIA RTX 3090 GPU. All the samples from the datasets were resized to 224×224 resolution. In all of the settings of our experiment, the networks weights initialized by ImageNet pre-trained weights. We used a polynomial learning rate decay with initial learning rate of 1×10^{-3} for better convergence, where i denotes the i-th epoch of training as follows:

$$lr_i = lr_{i-1} \times \left(1 - \frac{i}{\text{Total No. of Epochs}}\right)^{0.9} \quad (6)$$

We set the batch size and the total number of epochs to 4 and 100, respectively. SGD optimizer with momentum 0.9 and weight decay 0 is employed. To alleviate the low samples of datasets and generalize our proposed network, we utilized unlabelled samples during the training process to benefit from unsupervised techniques to enrich encoder representation. Note that during the training using the ISIC 2018 dataset, we used ISIC 2017 samples as an unsupervised dataset. Similarly, unlabelled samples from ISIC 2018 are utilized during the training of the model on the ISIC 2017 and the PH2 datasets.

C. EVALUATION METRICS

For the performance evaluation and present comparison results with other methods, we used four metrics, i.e., Sen-

[1] https://challenge.isic-archive.com/data
[2] https://challenge.isic-archive.com/data
[3] https://www.fc.up.pt/addi/ph2%20database.html

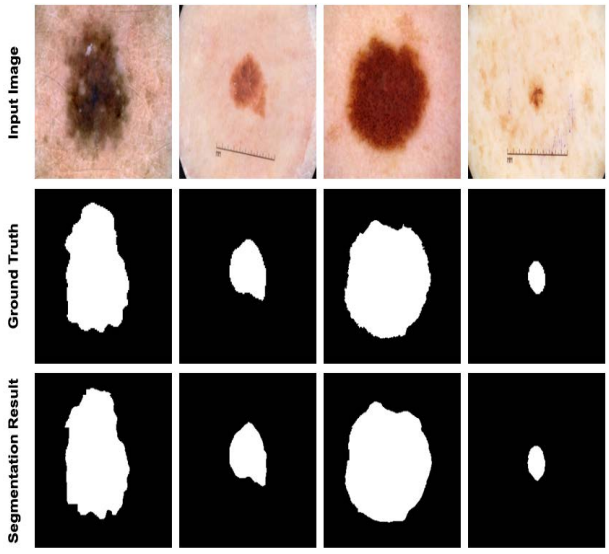


FIGURE 2. Segmentation maps obtained by the suggested network on the ISIC 2017 dataset. It is obvious that the method precisely produces a smooth segmentation results on the object boundary.

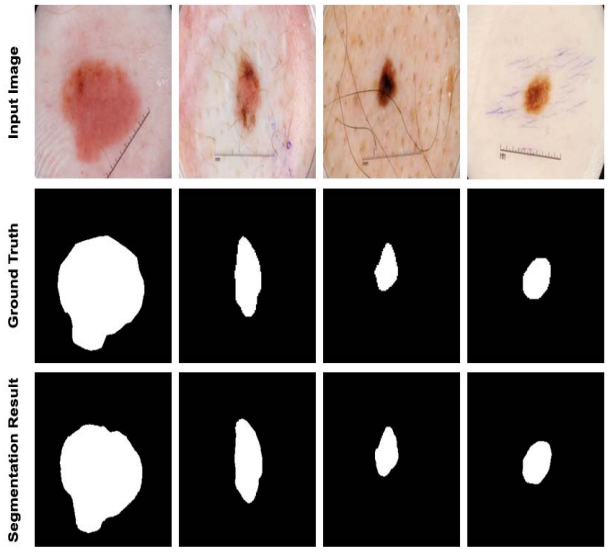


FIGURE 3. Segmentation maps obtained by the suggested network on the ISIC 2018 dataset. Our method precisely produces a smooth segmentation results on the object boundary.

sitivity (SE), Specificity (SP), Accuracy (ACC), and Dice coefficient (Dice) as follows:

$$SE = \frac{TP}{TP + FN} \tag{7}$$

$$SP = \frac{TN}{TN + FP} \tag{8}$$

$$ACC = \frac{TP + FN}{TP + TN + FP + FN} \tag{9}$$

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{10}$$

where TP and TN represent the correct number of skin lesion pixels and background pixels, respectively. FP is a number of background pixels that are miss-labelled with the skin lesion label, and FN denotes the number of skin lesion pixels that are incorrectly predicted as background pixels.

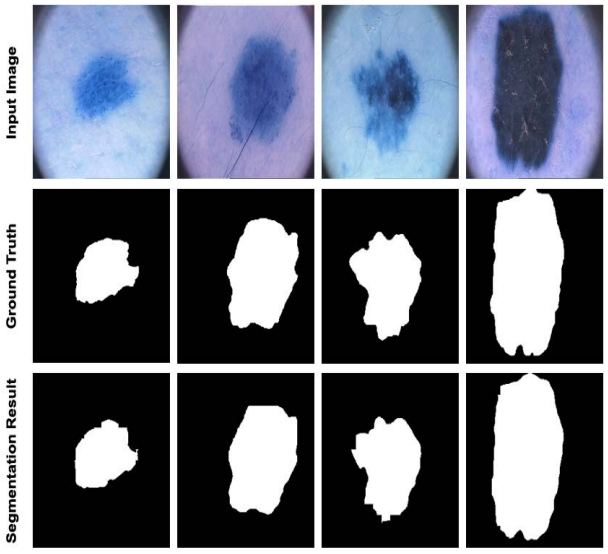


FIGURE 4. Segmentation maps obtained by the suggested network on the PH2 dataset.

D. RESULTS ON THE ISIC 2017 DATASET

We compared our method under the same circumstances with the SOTA approaches. In Table 1, the comparison results of the proposed network comparing to the seminal U-Net [15], Att U-Net [61], DAGAN [62], TransUNet [46], MCGU-Net [63], MedT [64], FAT-Net [65], and MSA-UNet [66] is provided. Our network improved the DSC and accuracy metrics of the seminal U-Net model by 8.99% and 4.27%, respectively. Furthermore, comparing to the CNN based approaches [15], [63] our network produces better results in all metrics, which indicate the effectiveness of both Transformer module incorporated in our structure and the semi-supervised technique used in our strategy. Besides that, comparing to the recently proposed MSA-UNet [66], our method exhibits a better performance due to the strength of the unsupervised technique utilized in our method. We also displayed a visual comparison of the obtained results in Figure 2. As can be seen from Figure 2, our proposed method produces a soft and precise segmentation results on the object boundary and effectively separates the skin lesion with irregular shapes and scales from the overlapped background.

E. RESULTS ON THE ISIC 2018 DATASET

We dissected our method with SOTA methods in the literature, including seminal U-Net [15], Att U-Net [61], DAGAN [62], TransUNet [46], MCGU-Net [63], MedT [64], FAT-Net [65], and MSA-UNet [66]. The evaluation settings are the same for all methods for a fair comparison. The statistical comparison is illustrated in the Table 2. As it is clear from Table 2 the MCGU-Net [63] with an attention mechanism and pretrained VGG backbone produces better performance than other CNN based approaches. MSA-UNet [66] outperformed both CNN and Transformer based approaches

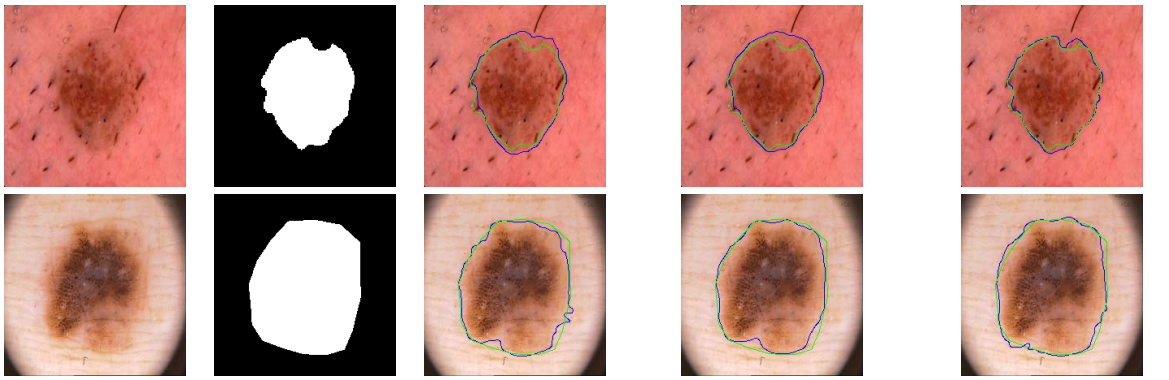


FIGURE 5. Visual comparisons of different setting of the proposed method on the *ISIC2018* skin lesion segmentation dataset. The green boundaries indicate the Ground truth while the blue color shows the predicted boundary.

TABLE 1. Experimental results of the proposed method against the SOTA approach on the ISIC 2017 dataset for skin lesion segmentation task.

Methods	DSC	SE	SP	ACC
U-Net [15]	0.8159	0.8172	0.9680	0.9164
Att U-Net [61]	0.8082	0.7998	0.9776	0.9145
DAGAN [62]	0.8425	0.8363	0.9716	0.9304
TransUNet [46]	0.8123	0.8263	0.9577	0.9207
MCGU-Net [63]	0.8927	0.8502	0.9855	0.9570
MedT [64]	0.8037	0.8064	0.9546	0.9090
FAT-Net [65]	0.8500	0.8392	0.9725	0.9326
MSA-UNet [66]	0.9032	0.8870	0.9714	0.9576
Proposed Method	**0.9058**	**0.9480**	**0.9766**	**0.9591**

TABLE 2. Experimental results of the proposed method against the SOTA approach on the ISIC 2018 dataset for skin lesion segmentation task.

Methods	DSC	SE	SP	ACC
U-Net [15]	0.8545	0.8800	0.9697	0.9404
Att U-Net [61]	0.8566	0.8674	0.9863	0.9376
DAGAN [62]	0.8807	0.9072	0.9588	0.9324
TransUNet [46]	0.8499	0.8578	0.9653	0.9452
MCGU-Net [63]	0.8950	0.8480	0.9860	0.9550
MedT [64]	0.8389	0.8252	0.9637	0.9358
FAT-Net [65]	0.8903	**0.9100**	0.9699	0.9578
MSA-UNet [66]	0.8960	0.8410	**0.9790**	0.9490
Proposed Method	**0.8988**	0.8820	0.9710	**0.9591**

due to the usage of the combination of CNN and Transformer modules. More precisely, the MSA-UNet utilizes a pyramidal feature representation underlying the network to compensate for the loss of global context in challenging samples, even though it can not achieve noticeable results regarding our work. in addition, comparing to both CNN and Transformer methods our semi-supervised training strategy not only achieved the highest score for most of the metrics but also outperformed all supervised learning strategies. Moreover, we depicted a visualization comparison in Figure 3. In some challenging samples, like the more minor contrast variance of lesion region with neighboring pixels, it is evident that our method still performs well.

F. RESULTS ON THE PH^2 DATASET

Finally, for further comparison studies, we investigated our method alongside some SOTA, including seminal U-Net [15], Att U-Net [61], DAGAN [62], TransUNet [46], MCGU-Net [63], MedT [64], FAT-Net [65], and MSA-UNet [66] on the PH^2 dataset. Like the previous experiments, the settings are the same for a fair comparison. The statistical comparison is depicted in the Table 3. Att U-Net [61], using an attention mechanism, achieved a better performance than U-Net. In addition, MSA-UNet [66] used a combination of CNN and Transformer rather than a conventional convolution, which became a facilitator to extract more discriminant features in an encoder, resulting in a better performance comparing to the other SOTA approaches. Our method utilizes the semi-supervised segmentation method

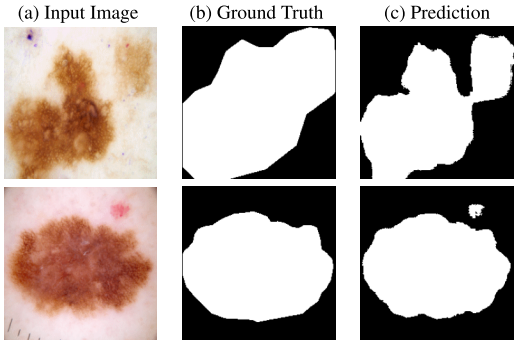
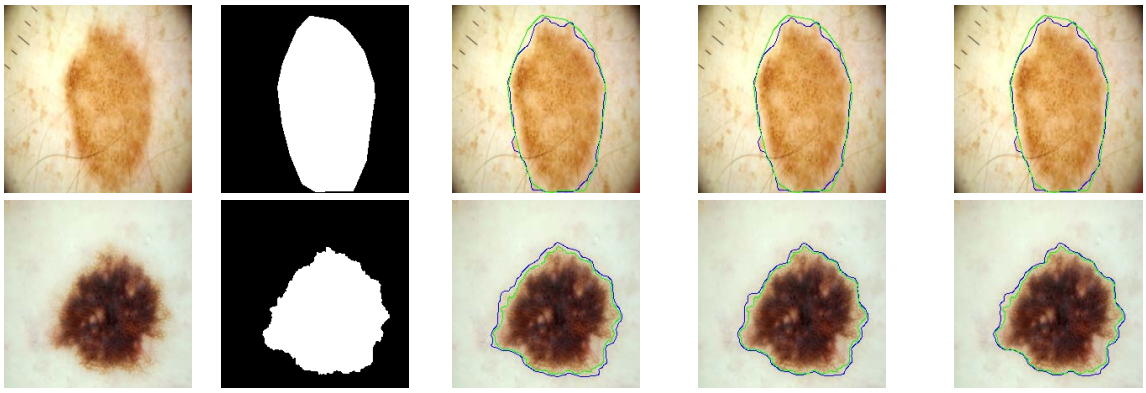


FIGURE 6. Some noisy annotation that exist in the ISIC 2018 dataset, where the model works quite well to predict the segmentation mask. The noisy annotations are a common scenario, so they can largely decrease model preference.

and outperformed the other SOTA approaches. In addition, we displayed a visual results in the Figure 4. It is evident that our method can easily handle complex brightness and contrast distributions which show the cogency and generalization of our network.

G. ABLATION STUDY

As part of this section, we conducted an ablation study to evaluate the impact of the proposed semi-supervised technique utilized in our pipeline and the Transformer module coupled with the CNN encoder to enrich the encoder representation. Different settings were used to analyze the contributions of each strategy. Our goal was to demonstrate how these techniques can be effectively incorporated into a skin lesion



(a) Input Image (b) Ground Truth (c) MCGU-Net [63] (d) MSA-UNet [66] (e) Proposed Method

FIGURE 7. Visual comparisons of segmentation prediction obtained by the proposed method against the SOTA approaches on the *ISIC2017* skin lesion segmentation dataset. The green boundaries indicate the Ground truth while the blue color shows the predicted boundary.

TABLE 3. Performance comparison betwen the proposed method and the SOTA approaches on the PH^2 dataset.

Methods	DSC	SE	SP	ACC
U-Net [15]	0.8936	0.9125	0.9588	0.9233
Att U-Net [61]	0.9003	0.9205	0.9640	0.9276
DAGAN [62]	0.9201	0.8320	0.9640	0.9425
TransUNet [46]	0.8840	0.9063	0.9427	0.9200
MCGU-Net [63]	0.9263	0.8322	0.9714	0.9537
MedT [64]	0.9122	0.8472	0.9657	0.9416
FAT-Net [65]	0.9440	**0.9441**	0.9741	0.9703
MSA-UNet [66]	0.9377	0.9430	0.9698	0.9617
Proposed Method	**0.9401**	0.9338	**0.9780**	**0.9711**

segmentation model to increase the model generalization performance by combining them. To demonstrate how the Transformer model can be used to encode a more robust representation and, thus, enhance the model performance, in one experiment we replaced the Transformer module with an Atrous convolution to increase the receptive field size and consequently capture the global representation (indicated with Atruous conv in table 4). In addition, we trained our model without using an auxiliary task to demonstrate the effect of the unsupervised technique incorporated in our strategy (denoted as Baseline + Transformer). According to our findings, each strategy contributes to the model performance and they provide a strong representation of the network features. Based on the experimental results shown in Table 4, using the Transformer module along with the hierarchical features of the seminal U-Net (baseline) helps the model to learn a multi-scale representation with rich and generic features, and significantly increases the model's performance. Moreover, the generalization performance is further enhanced by incorporating the auxiliary task. Our finding is in line with the semi-supervised literature [57] that the auxiliary task can enrich the segmentation encoder and consequently result in a better performance. Moreover, in terms of model selection, one should be noted that our method is not limited to a specific segmentation network, such as U-Net, and can be incorporated into any segmentation network for higher performance gain. To visually analyze the effect of suggested modules on the segmentation results,

TABLE 4. Performance comparison of different settings in our proposed method. Result reported using the ISIC 2017 dataset.

Methods	DSC	AC
Baseline (U-Net CNN)	0.8159	0.9164
Atrous Conv	0.8497	0.9045
Baseline + Transformer	0.8931	0.9562
Baseline + Transformer + Unsupervised	**0.9058**	**0.9591**

we provided sample comparison results in Figure 5. It is obvious that by incorporating each module the segmentation results become better. Specifically, comparing to the Atrous based and Baseline + Transformer methods the final setting (proposed method) works quite well on the boundary area without over and under estimation.

It should also be noted that in some of our experiences as can be seen in Figure 6, our method fails to segment the skin lesion area similar to the ground truth mask due to the noisy annotation provided by the dataset. In clinical applications, noisy annotations are a common scenario, so they can largely decrease model preference. This might explain why clean annotation is important in the training process.

To visualize the effectiveness of our suggested network compared to the SOTA approaches, we provided Figure 7. In our comparison, we provided the segmentation result achieved by the MCGU-Net [63] and MSA-UNet [66] approaches comparing to our suggested network. It can be observed that our method produces smooth segmentation results with precise boundary separation. We can also observe that comparing to the MCGU-Net [63], our method has better estimation of the skin lesion boundary and it is in line with MSA-UNet [66] approach. It is also worthwhile to mention that for the second sample, the MCGU-Net underestimates the segmentation map while the MSA-UNet slightly produces an overestimation. On the contrary, our method produces a better segmentation map for the second sample with slight FN predictions.

V. CONCLUSION

In this paper, we proposed a semi-supervised technique to enhance the semantic segmentation task. In our strategy, we proposed to incorporate the unlabelled samples during the

training process to encourage the feature learning paradigm. Our suggested network offers a semi-supervised training schema, wherein the first stage performs a supervised training strategy to learn semantic segmentation map while the second step focuses on the unsupervised technique to enrich the encoder module. Several experimental results on three public datasets demonstrated the effectiveness of our approach for the semantic segmentation task.

ACKNOWLEDGMENT

This research work was funded by Institutional Fund Projects under grant no. (IFPIP: 1332-611-1443). The authors gratefully acknowledge technical and financial support provided by the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

REFERENCES

[1] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2020.

[2] T. Falk, D. Mai, R. Bensch, and Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, and K. Seiwald, "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, 2019.

[3] R. Ali, R. C. Hardie, B. N. Narayanan, and T. M. Kebede, "IMNets: Deep learning using an incremental modular network synthesis approach for medical imaging applications," *Appl. Sci.*, vol. 12, no. 11, p. 5500, May 2022.

[4] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA Cancer J. Clin.*, vol. 72, no. 1, pp. 7–33, Jan. 2022.

[5] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi, "Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 250–258.

[6] H. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, and L. Uhlmann, "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018.

[7] M. Emre Celebi, Q. Wen, S. Hwang, H. Iyatomi, and G. Schaefer, "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Res. Technol.*, vol. 19, no. 1, pp. e252–e258, Feb. 2013.

[8] B. Erkol, R. H. Moss, R. J. Stanley, W. V. Stoecker, and E. Hvatum, "Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes," *Skin Res. Technol.*, vol. 11, no. 1, pp. 17–26, 2005.

[9] M. E. Celebi, Y. A. Aslandogan, W. V. Stoecker, H. Iyatomi, H. Oka, and X. Chen, "Unsupervised border detection in dermoscopy images," *Skin Res. Technol.*, vol. 13, no. 4, pp. 454–462, 2007.

[10] D. D. Gomez, C. Butakoff, B. K. Ersboll, and W. Stoecker, "Independent histogram pursuit for segmentation of skin lesions," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 157–161, Jan. 2007.

[11] R. Garnavi, M. Aldeen, M. E. Celebi, A. Bhuiyan, C. Dolianitis, and G. Varigos, "Automatic segmentation of dermoscopy images using histogram thresholding on optimal color channels," *Int. J. Med. Med. Sci.*, vol. 1, no. 2, pp. 126–134, 2010.

[12] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[14] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[17] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[18] G. Venkatesh, Y. Naresh, S. Little, and N. E. O'Connor, "A deep residual architecture for skin lesion segmentation," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Cham, Switzerland: Springer, 2018, pp. 277–284.

[19] M. K. Hasan, L. Dahal, P. N. Samarakoon, F. I. Tushar, and R. Martí, "DSNet: Automatic dermoscopic skin lesion segmentation," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103738.

[20] Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.

[21] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (DV)*, Oct. 2016, pp. 565–571.

[22] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3D-UNet: Separable 3D U-Net for brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2018, pp. 358–368.

[23] D. R. Ramani and S. S. Ranjani, "U-Net based segmentation and multiple feature extraction of dermascopic images for efficient diagnosis of melanoma," in *Computer Aided Intervention and Diagnostics (Don't Short) in Clinical and Medical Images*. Cham, Switzerland: Springer, 2019, pp. 81–101.

[24] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Biomed. Eng*, vol. 64, no. 9, pp. 2065–2074, Sep. 2017.

[25] Y. Tang, F. Yang, S. Yuan, and C. Zhan, "A multi-stage framework with context information fusion structure for skin lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1407–1410.

[26] R. Ali, R. C. Hardie, B. Narayanan Narayanan, and S. De Silva, "Deep learning ensemble methods for skin lesion analysis towards melanoma detection," in *Proc. IEEE Nat. Aerosp. Electron. Conf. (NAECON)*, Jul. 2019, pp. 311–316.

[27] E. Santos, R. Veras, H. Miguel, K. Aires, M. L. Claro, and G. B. Junior, "A skin lesion semi-supervised segmentation method," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Jul. 2020, pp. 33–38.

[28] Y. Yuan and Y.-C. Lo, "Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 519–526, Mar. 2019.

[29] M. D. Alahmadi, "Medical image segmentation with learning semantic and global contextual representation," *Diagnostics*, vol. 12, no. 7, p. 1548, Jun. 2022.

[30] S. A. Taghanaki, A. Bentaieb, A. Sharma, S. K. Zhou, Y. Zheng, B. Georgescu, P. Sharma, Z. Xu, D. Comaniciu, and G. Hamarneh, "Select, attend, and transfer: Light, learnable skip connections," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 417–425.

[31] G. Zhang, X. Shen, S. Chen, L. Liang, Y. Luo, J. Yu, and J. Lu, "DSM: A deep supervised multi-scale network learning for skin cancer segmentation," *IEEE Access*, vol. 7, pp. 140936–140945, 2019.

[32] E. Nasr-Esfahani, S. Rafiei, M. H. Jafari, N. Karimi, J. S. Wrobel, S. Samavi, and S. M. R. Soroushmehr, "Dense pooling layers in fully convolutional network for skin lesion segmentation," *Computerized Med. Imag. Graph.*, vol. 78, Dec. 2019, Art. no. 101658.

[33] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2482–2493, Dec. 2020.

[34] P. Shamsolmoali, M. Zareapoor, E. Granger, and H. Zhou, "Salient skin lesion segmentation via dilated scale-wise feature fusion network," 2022, *arXiv:2205.10272*.

[35] H. Li, X. He, F. Zhou, Z. Yu, D. Ni, S. Chen, T. Wang, and B. Lei, "Dense deconvolutional network for skin lesion segmentation," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 527–537, Mar. 2019.

[36] Z. Mirikharaji and G. Hamarneh, "Star shape prior in fully convolutional networks for skin lesion segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 737–745.

[37] Y. Xue, T. Xu, and X. Huang, "Adversarial learning with multi-scale loss for skin lesion segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 859–863.

[38] H. Wang, G. Wang, Z. Sheng, and S. Zhang, "Automated segmentation of skin lesion based on pyramid attention network," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 435–443.

[39] V. K. Singh, M. Abdel-Nasser, H. A. Rashwan, F. Akram, N. Pandey, A. Lalande, B. Presles, S. Romani, and D. Puig, "FCA-Net: Adversarial learning for skin lesion segmentation based on multi-scale features and factorized channel attention," *IEEE Access*, vol. 7, pp. 130552–130565, 2019.

[40] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 683–687.

[41] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[42] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[43] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 251–266.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

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

[46] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[48] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

[49] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, "Semi-supervised learning for network-based cardiac mr image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 253–260.

[50] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 408–416.

[51] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 605–613.

[52] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 552–561.

[53] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[54] N. Komodakis and S. Gidaris, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[55] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 69–84.

[56] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1058–1067.

[57] A. R Feyjie, R. Azad, M. Pedersoli, C. Kauffman, I. Ben Ayed, and J. Dolz, "Semi-supervised few-shot learning for medical image segmentation," 2020, *arXiv:2003.08462*.

[58] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "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 *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

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

[60] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "PH^2—A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jun. 2013, pp. 5437–5440.

[61] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[62] B. Lei, Z. Xia, F. Jiang, X. Jiang, Z. Ge, Y. Xu, J. Qin, S. Chen, T. Wang, and S. Wang, "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101716.

[63] M. Asadi-Aghbolaghi, R. Azad, M. Fathy, and S. Escalera, "Multi-level context gating of embedded collective knowledge for medical image segmentation," 2020, *arXiv:2003.05056*.

[64] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 36–46.

[65] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102327.

[66] M. D. Alahmadi, "Multiscale attention U-Net for skin lesion segmentation," *IEEE Access*, vol. 10, pp. 59145–59154, 2022.

MOHAMMAD D. ALAHMADI (Member, IEEE) received the M.Sc. and Ph.D. degrees from Florida State University, in 2018 and 2020, respectively. He is currently an Assistant Professor with the Software Engineering Department, University of Jeddah. His research interests include software engineering, computer vision, and machine learning. His research has been published in top journals and conferences in a wide variety of topics.

WAJDI ALGHAMDI received the Ph.D. degree specialized in computer science (data mining) from the Department of Computing, Goldsmiths College, University of London, London, U.K. His is currently an Assistant Professor at the Information Technology Department, Computing and Information Technology College, King Abdul-Aziz University, Jeddah, Saudi Arabia. He is mostly interested in knowledge discovery in databases, data mining and statistical computing. His research interests include applying machine learning and statistical learning methods to genotype, phenotype, and clinical data in-order to discover patterns of interest, including the identification of clinical and genetic predictors with respect to diseases.