applied sciences

MDPI

Article

Skin Lesion Segmentation Based on Vision Transformers and Convolutional Neural Networks—A Comparative Study

Yonis Gulzar [1,*] and Sumeer Ahmad Khan [2,*]

[1] Department of Management Information Systems, College of Business Administration, King Faisal University, Hofuf 31982, Saudi Arabia
[2] Biological and Environmental Sciences and Engineering, King Abdullah University of Science and Technology, Jeddah 23955, Saudi Arabia
* Correspondence: ygulzar@kfu.edu.sa (Y.G.); sumeer.khan@kaust.edu.sa (S.A.K.); Tel.: +966-545-719-118 (Y.G.)

Abstract: Melanoma skin cancer is considered as one of the most common diseases in the world. Detecting such diseases at early stage is important to saving lives. During medical examinations, it is not an easy task to visually inspect such lesions, as there are similarities between lesions. Technological advances in the form of deep learning methods have been used for diagnosing skin lesions. Over the last decade, deep learning, especially CNN (convolutional neural networks), has been found one of the promising methods to achieve state-of-art results in a variety of medical imaging applications. However, ConvNets' capabilities are considered limited due to the lack of understanding of long-range spatial relations in images. The recently proposed Vision Transformer (ViT) for image classification employs a purely self-attention-based model that learns long-range spatial relations to focus on the image's relevant parts. To achieve better performance, existing transformer-based network architectures require large-scale datasets. However, because medical imaging datasets are small, applying pure transformers to medical image analysis is difficult. ViT emphasizes the low-resolution features, claiming that the successive downsampling results in a lack of detailed localization information, rendering it unsuitable for skin lesion image classification. To improve the recovery of detailed localization information, several ViT-based image segmentation methods have recently been combined with ConvNets in the natural image domain. This study provides a comprehensive comparative study of U-Net and attention-based methods for skin lesion image segmentation, which will assist in the diagnosis of skin lesions. The results show that the hybrid TransUNet, with an accuracy of 92.11% and dice coefficient of 89.84%, outperforms other benchmarking methods.

Keywords: melanoma; lesion; segmentation; transformers; convolutional neural networks

Citation: Gulzar, Y.; Khan, S.A. Skin Lesion Segmentation Based on Vision Transformers and Convolutional Neural Networks—A Comparative Study. *Appl. Sci.* **2022**, *12*, 5990. https://doi.org/10.3390/app12125990

Academic Editors: Hyuntae Park, Do-Young Kang and Sangjin Kim

Received: 17 May 2022
Accepted: 10 June 2022
Published: 12 June 2022

1. Introduction

Cancer is one of the deadliest diseases in the world. According to WHO, in 2020, nearly 10 million people have died due to cancer. According to 2020 data, among the new cancer cases found, the most common cases were breast cancer (2.27 M), lung cancer (2.21 M), colon and rectum (1.93 M), skin cancer (1.2 M) and stomach (1.09 M) [1]. Skin cancers are distinguished into two categories, namely non-melanoma and melanoma. Non-melanoma is a more common type of skin cancer whereas melanoma is a less common skin cancer. However, melanomas are the most dangerous ones and have a high spread rate. Even though the amount is only 5% of skin malignancy, the mortality rate is over 75% [2]. Skin cancers can spread to other parts of the body and are often not curable; detection at a very early stage can help them from spreading as they may be curable. Melanoma skin cancers are a highly aggressive type of cancer, and their incidence has dramatically increased over the past three decades [3]. To stop their spread and treat them, it is important to detect such cancers at a very early stage. For melanoma skin

cancer, the five-year survival rate is 98% if detected at an early stage; however, it drops by 14% if detected in later stages. Due to the nature of melanoma, it is very important to detect it in a timely and accurate manner. The best way to detect skin lesions is through dermatoscopy. This technique is used as a primary examination to detect skin lesions. Due to the high resolution and quality, and enhanced visualization competence of dermoscopic images, these images help dermatologists to examine the skin lesions with naked eyes. Nevertheless, good expertise and deep knowledge are required to make the right diagnosis and such diagnoses/interpretations may differ from one dermatologist to another.

It is evident from the literature that skin lesion (melanoma) detection based on CNN (Convolutional Neural Networks) has obtained better performance than dermatologists' [4]. Incorporating AI in smartphones has become a trend and is proven productive in terms of accuracy. Advancement in image processing and incorporating the deep learning techniques, especially CNN, has given a different edge to use in medical science. Processing of images using CNN and its automatic analysis methods can be proven a powerful tool and can provide a user-friendly intelligent system to scan the lesion images to detect melanoma outside the clinic [5,6]. As a result, automatic analysis of skin lesions has become an important step in computer-aided diagnosis [7].

Using dermoscopy images, different approaches have been proposed to detect melanoma skin lesions based on the appearance of colour and texture patterns of skin for instant classical pattern analysis [8], ABCDE rules [9] and a seven-point checklist [10,11]. In ABCDE rules, an easy and general framework has been provided in order to identify melanoma. The defined rules are border irregularity, un-uniform colour, 6 mm diameter or above, and growing lesions in colour, shape or size. Usually, it is the borders of the melanoma which are uneven and may contain rough edges, which are imprecisely defined. Hence, skin segmentation is performed at an early stage to get the border information or regions of interest (ROI). Such approaches have been proven to be beneficial for the subsequent classification or detection task [11,12].

It is a challenging task to perform automatic skin lesion segmentation on some skin lesions where lesions are not that clear or have a light pigment. It gets complicated when colour or visual patterns are similar, and the boundaries are alike for different melanomas. In such cases, skin segmentation becomes enormously difficult. It gets more complicated for computers to read images directly which are of high resolution, and resource intensive. Having such properties make the process slow. To expedite the process, the image size is reduced using down-sampling, which in turn has a negative effect on the textures and subtlety. Due to that, it gets more complicated to differentiate the boundaries of these skin lesions. Moreover, the colour and texture distribution gets affected by some elements present in skin lesions such as blood vessels, colour illumination and hair artefacts, which have a negative impact on learning.

Figure 1 shows some sample images of the ISIC dataset [13]. ISIC contains the digital skin images of melanoma. The aim of keeping such images publicly available is to educate the public/professionals about melanoma, examine them and provide some diagnoses through teledermatology, clinical decision support and automated diagnosis. Applying skin lesion segmentation on such a dataset is a challenging task due to the characteristics of these images. As it can be seen from the figure some images are covered with hair, some lesions have blood vessels around them and some lesion images have fuzzy boundaries, which makes skin lesion segmentation difficult. Different approaches have been proposed for skin lesion segmentation by incorporating different CNN architectures with multi-scale information [12,14,15] and multi-task learning frameworks [16,17] with auxiliary information [18]. The aim of these proposed approaches is to utilize as much data available to detect skin lesions and make the right diagnosis. Nevertheless, such methods either require extra labelling or use extensive parameters which are impractical in real situations.

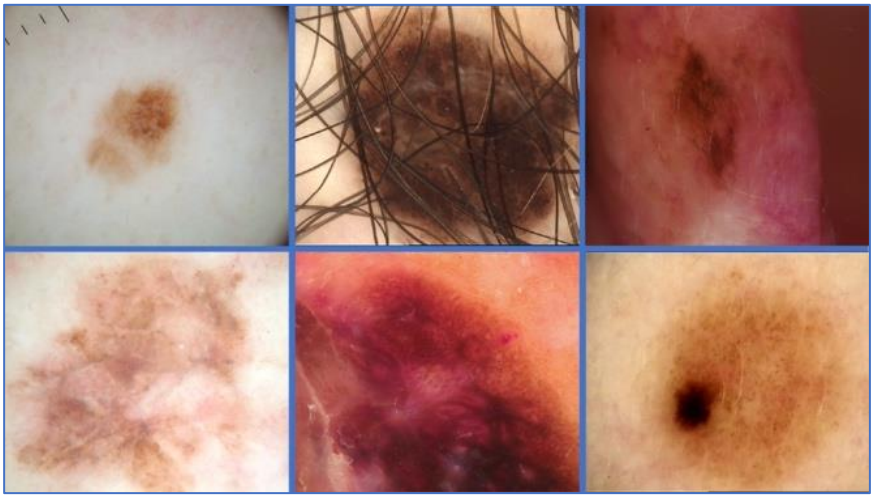

Figure 1. Sample of Skin Lesion Images.

Although CNNs have achieved the best results on various vision tasks, they have still shown limited performance on skin lesion segmentation tasks due to a lack of understanding of long-range spatial relations in skin lesion images. In order to overcome this limitation, the recently proposed Vision Transformer (ViT) [19] for image classification employs a purely self-attention-based model that learns long-range spatial relations to focus on the image's relevant parts. To achieve better performance, existing transformer-based network architectures require large-scale datasets. However, because medical imaging datasets are small, applying pure transformers to medical image analysis is difficult. ViT emphasizes low-resolution features, claiming that the successive downsamplings result in a lack of detailed localization information, rendering it unsuitable for skin lesion image classification. To improve the recovery of detailed localization information, several ViT-based image segmentation methods have recently been combined with ConvNets in the natural image domain. Furthermore, different architecture called UNet [20], based on CNN with few changes in its architecture, has been developed and incorporated into biomedical images not only to just classify the disease but also to identify the area of infection. Due to its success rate on medical images, many researchers [21–24] have adopted this model and use it for skin segmentation. This study provides a combined architecture that combines ViT and ConvNet to provide an efficient skin lesion image segmentation, which will assist in the diagnosis of skin lesions.

Recently, some of the feature extraction studies have shown some improvement in terms of accuracy [25–27]. Such studies mainly follow parallel-based and serial-based approaches. The aim of these studies is to enrich the information related to the subject gathered from different sources. Nevertheless, incorporating this technique gives the rise to the number of predictors, which in turn has a negative impact on computational time. Due to this reason, many researchers only focus on introducing techniques that select only the best features. Feature-based techniques are of two types: heuristic-based and meta-heuristic technique. Meta-heuristic techniques are more valuable for the selection process and provide a fewer number of predictors.

In this paper, a detailed comparative analysis based on the properties of the Transformers and CNNs is carried out. The following points summarize the contribution of the paper:

- We discuss the challenges in the skin lesion segmentation.
- We provide a detailed comparative analysis of the state-of-the-art methods for the task of skin lesion segmentation.
- We evaluate the efficacy of the state-of-the-art methods, and the experiments demonstrate the effectiveness of the U-Net and Transformer-based methods for skin lesion segmentation.

The rest of the paper is organized as follows: related works to this research are reported in Section 2. In Section 3, the details about the comparative methods are presented. Section 4 shows the experimental details and performance evaluation of the comparative methods and the benchmarking models. Finally, Section 5 concludes the paper.

2. Related Work

In different computer vision tasks, the convolutional neural network models have made exceptional progress [28–32], and have been proven to be the best in solving segmentation problems. Skin lesion classification accuracy directly depends on the lesion segmentation. The more accurate the lesion segmentation is, the better the result of lesion classification. Lately, a lot of research has been done to develop such models, which can automatically conduct the lesion segmentation in such a way that differentiates the normal area of skin from the lesion. Different techniques have been used, such as region growing, threshold, edge detection, etc. to develop such models.

Segmentation problems have been treated as classification problems by some researchers, such as Jafari et al. [33]. They proposed an automatic skin segmentation model for non-dermoscopic images. For better performance, this model uses the guided filter technique [34] as a pre-processing technique to reduce the noise artefacts on the input images. A window around the pixel is placed to identify whether it belongs to a lesion area or not and then is fed to a CNN and output labels pixel in the centre of the patch. A segmentation map is generated with the label as 1 (lesion) and 0 (normal) for each pixel of the pre-processed images and then the segmentation mask is generated by selecting the largest connected component of an image. The output of this model is the predicted labels of the pixels. Nonetheless, such a model requires a dense prediction because it is based on pixel-level prediction. U-NET is a well-known model proposed by Ronneberger et al. [20]. It has a good success rate for medical images when it comes to segmentation problems. Due to its popularity, many models have been proposed [21–24] adopting U-NET architecture. They have used it for melanoma skin segmentation and classification. In [23], the authors have modified the original U-NET by widening the kernel at each convolutional block using a dilated convolution technique in order to expand the receptive field of the proposed model; whereas in [24], the authors have utilized colour transformation to select different colour bands in order to improve the performance of the proposed approach. Another model proposed by Yuan et al. [35] incorporated the deconvolution method and replaced the regular cross-entropy function with a loss function based on Jaccard distance; whereas a full resolution CNN model is proposed by Al-masni et al. [36], in which models directly learn the full resolution features of every single pixel from the input images without reducing the size of input images. However, there is an issue of using full resolution CNN as an overfitting of the model, which is likely to happen on the features of non-melanoma skin lesions. In that case, such models perform poorly when it comes to complex features of melanoma skin lesions due to their inconsistent boundary appearance and different textures. Bi et al. [37] proposed a new model to overcome the limitation of full-resolution CNN. Their model learns the necessary skin lesion characteristics for each class (be it melanoma or non-melanoma) individually. They incorporated a new probability-based, stepwise integration to combine the segmentation output generated from each class. In another study [38] of skin lesion segmentation, a dilated residual network has been used along with pyramid pooling networks. To attain sharp boundaries, they have used a negative log-likelihood and endpoint error loss together. In a recent study, a bootstrapping CNN method was proposed by Xie et al. [17] for both skin lesion segmentation and skin classification problem. In this method, one task aids another in a bootstrapping way where a coarse segmentation network is trained, and then a predicted coarse mask is utilized to guide the classification network. Simultaneously, localization maps are generated to increase the prediction of masks outperforming the coarse mask. Recently, another model of CNN architecture using auxiliary information has been proposed [18]. The model does not need any pre-post-processing of data. This model predicts edges and does the classification

of skin lesions simultaneously by using two parallel branches. Predicting edges helps the neural network to pay attention to segmentation mask boundaries.

In brief, the aforementioned methods focused mainly on feature extraction. They used pre-trained CNN models in order to achieve it. Furthermore, they incorporated CNN segmentation models for lesion detection. Their primary aim is to improve the model's accuracy in terms of performing skin segmentation. However, hardly anyone has focused on reducing the systems' time in terms of prediction.

3. Methods

In this work, we selected convolution-deconvolution based, (U-Net [20], V-Net [39]) and Attention-based (Attention U-Net [40], TransUNet [41], Swin-UNet [42]) medical image segmentation models based on their documented performance on the existing medical image segmentation datasets.

Since the task of skin lesion segmentation is quite different from multiorgan segmentation, as in skin lesion segmentation, the task is to predict the binary mask rather than the multiclass pixel-wise segmentation. Therefore, we modified the implementation of these models in order to apply these models for this specific task of skin lesion segmentation. The following subsections outline a brief overview of these methods and the architectural changes that were carried out for this specific task of skin lesion segmentation.

3.1. U-Net

As stated in [20], U-Net consists of contracting and expansive path (downsampling and upsampling) as shown in Figure 2. For this particular task, we used four downsampling and upsampling levels with one bottom level. Two convolutional layers per downsampling level and one convolutional layer after concatenation per up sampling level. Gaussian Error Linear Unit (GELU function) was used as an activation function with batch normalization and Softmax as an output function. GELU multiplies its input by the cumulative density function of the normal distribution. The reason for using GELU is that it provides a well-defined negative gradient to prevent neurons from dying while also limiting how far into the negative regime activations can have an effect, allowing for improved feature mixing between layers. In addition to this, max pooling was used for downsampling and reflective padding for upsampling.

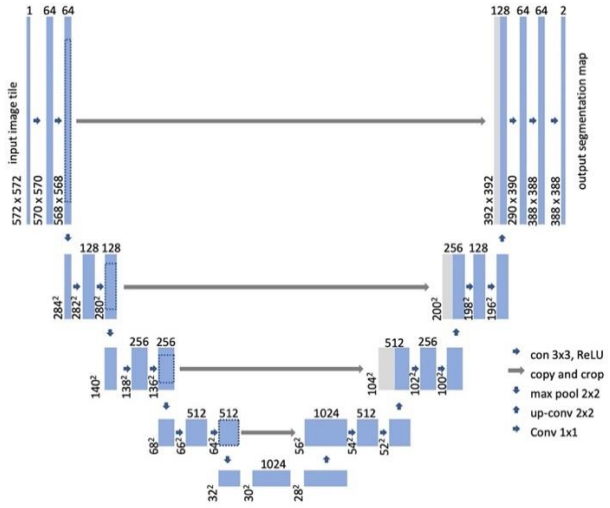

Figure 2. U-Net architecture adapted from [20].

3.2. V-Net

V-Net [39] originally proposed for 3d MRI images was modified to 2d inputs of skin lesion images as shown in Figure 3. The same number of downsampling and upsampling levels were used as in U-Net implementation, i.e., four with one bottom level. The number

of stacked convolutional layers in the residual path increases from one to three downsampling levels (symmetrically decreasing with upsampling levels). Unlike U-Net, the Parametric ReLU (PReLU) activation function was used with batch normalization and Softmax as an output activation function. PReLU is a type of LeakyReLU that instead of having a predetermined slope, makes it a parameter for the neural network to figure out itself, which helps in the better adaption to other parameters such as weights and biases. Downsampling was carried through stridden convolutional layers and upsampling through transpose convolutional layers.

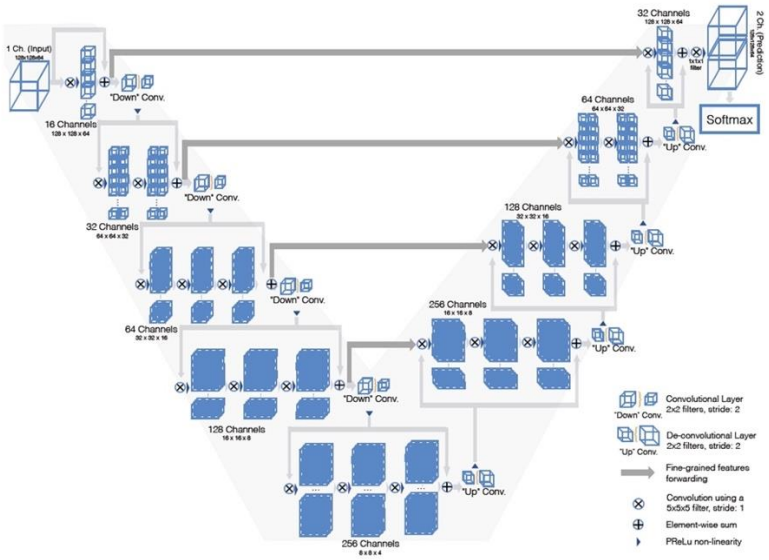

Figure 3. V-Net architecture adapted from [39].

3.3. Attention U-Net

Attention U-Net [40] was proposed for multi-class CT abdominal image segmentation as shown in Figure 4. We modified this model for skin lesion image segmentation. We used four downsampling and upsampling levels with two convolutional layers per downsampling level and two convolutional layers after concatenation per upsampling level. We used the ReLU activation function with batch normalization. In addition to this, we used additive attention and ReLU attention activation. We used stridden convolution layers for downsampling and bilinear interpolation for upsampling.

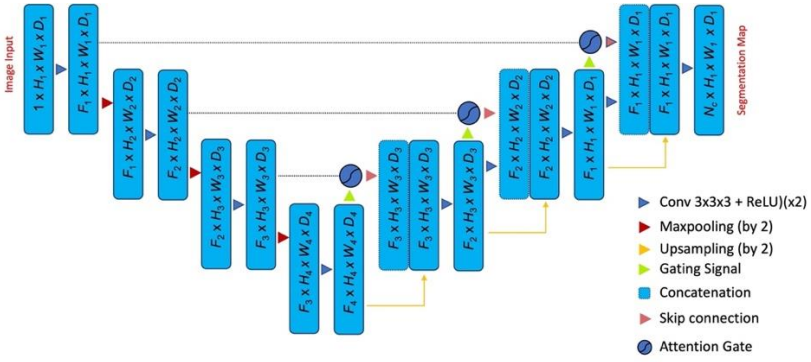

Figure 4. Attention U-Net architecture adapted from [40].

3.4. TransUNet

The TransUNet [41] architecture has been used for the medical image segmentation on the Synapse multi-organ segmentation dataset for segmenting the abdominal CT images into eight abdominal organ classes as shown in Figure 5. Since the task of skin lesion segmentation is quite different from multiorgan segmentation, as in skin lesion segmentation,

the task is to predict the binary mask rather than the multiclass pixel-wise segmentation. Therefore, we modified the basic architecture of the TransUNet, which is a combination of the Transformer and U-Net architectures in order to address the challenges in skin lesion image segmentation. For an image, instance $i \in \mathbb{R}^{H \times W \times C}$, with a spatial resolution of $H \times W$ and a channel count of C is determined. The goal is to forecast the label map with size $H \times W$ that corresponds to the related ground truth label map. The most common and conventional method is to train a CNN (such as U-Net) to encode images into high-level feature representations, which are subsequently decoded back to full spatial resolution. Unlike other approaches, this method uses Transformers to incorporate self-attention mechanisms into the encoder architecture.

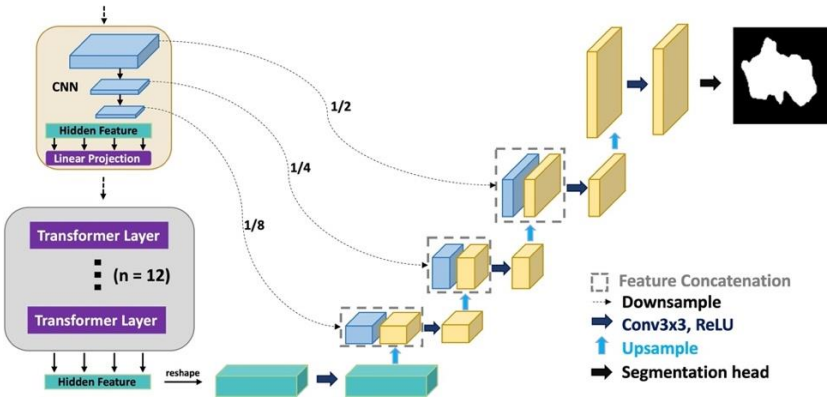

Figure 5. TransUNet architecture adapted from [41].

The input to the model is (512, 512, 3) images. The model has four down and upsampling levels-two convolutional layers per downsampling and upsampling layers, respectively. We used 12 transformers blocks and the number of attention heads was set to 12. The other changes that have been made to the basic architecture of the TransUNet are the number of nodes in the multi-layer perceptron (MLP) and the embedding dimensions. In the proposed model, each transformer has 1536 multi-layer perceptron (MLP) nodes that embed the images into 384 dimensions. The reason for modifying these was to make the network focus on more precise localization and produce accurate binary masks as compared to the multiclass segmentation in the base TransUNet. For the transformer, an MLPs Gaussian Error Linear Unit (GELU) was used as an activation function and upsampling through bilinear interpolation was carried out. Sigmoid was used as an output function.

3.5. Swin-UNet

Swin-UNet [42] is transformer-based U-shaped architecture with an encoder, bottleneck, decoder and skip connections as shown in Figure 6. The encoder, bottleneck and decoder are all built using the Swin-Transformer block. The inputs are divided into non-overlapping image patches. Each patch is treated as a token and fed into the Transformer-based encoder to learn deep feature representations. The extracted context features are then up-sampled by the decoder with a patch expanding layer and fused with the multi-scale features from the encoder via skip connections to restore the spatial resolution of the feature maps and further perform segmentation prediction. Three downsamplings and upsampling with one bottom level were used. Two Swin transformers per downsampling level and two Swin transformers per upsampling level were used. Two by two patches were extracted from the input and then embedded into 64 dimensions. In addition to this, 512 nodes per Swin transformer, shift attention windows and Softmax output function were used.

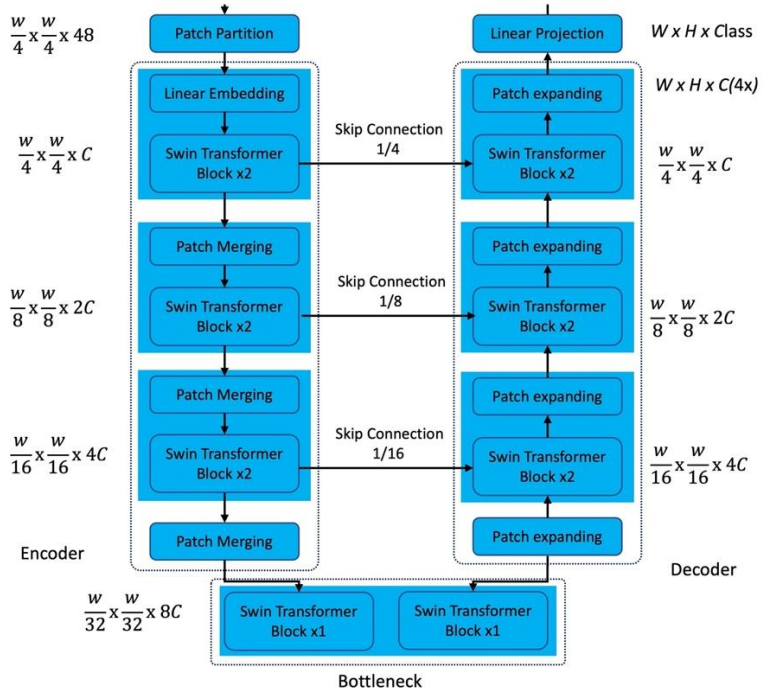

Figure 6. Swin-UNet architecture adapted from [42].

4. Experimental Results

The implementation details of the comparative methods are presented in this section. After that, the dataset and evaluation metrics are described. Finally, the performance of the different methods is assessed and compared to state-of-the-art approaches.

4.1. Implementation Details

All the models are trained using ISIC 2018 [18] training data, and their performance is assessed using testing data. A dermoscopic image (input), its matching ground-truth segmentation mask and an edge (contour) image (outputs) are required for training the proposed model. By using a contour detection technique, the ground truth of the edge picture may be automatically extracted from the ground truth of the segmentation mask. We trained all the models with stochastic gradient descent (SGD) with a learning rate of 0.001, momentum = 0.9 and decay = 0.0005. The batch size is set to 1 and the models were trained for 100 to 200 epochs with early stopping criteria. The models were trained with binary cross-entropy loss.

The methods were implemented in Keras with the TensorFlow backend. All the experiments were carried out on Nvidia Tesla V100 GPUS with 32 GB memory.

4.2. Dataset

ISIC 2018 Dataset: ISIC dataset [13] contains of three parts—skin lesion segmentation, attribute detection and classification of the skin lesion into type. For segmentation, the dataset contains 2594 and 1000 number of images for training and testing, respectively. The training set is provided along with ground truth images. The images were resized to a uniform dimension of 512 by 512 as a pre-processing step in order to have a uniform evaluation of all the methods. The images present in dataset have illumination variations as well as different artefacts such as hair, colour-marks, rulers and glue. Figure 7 shows some of the image samples of dataset ISIC 2018.

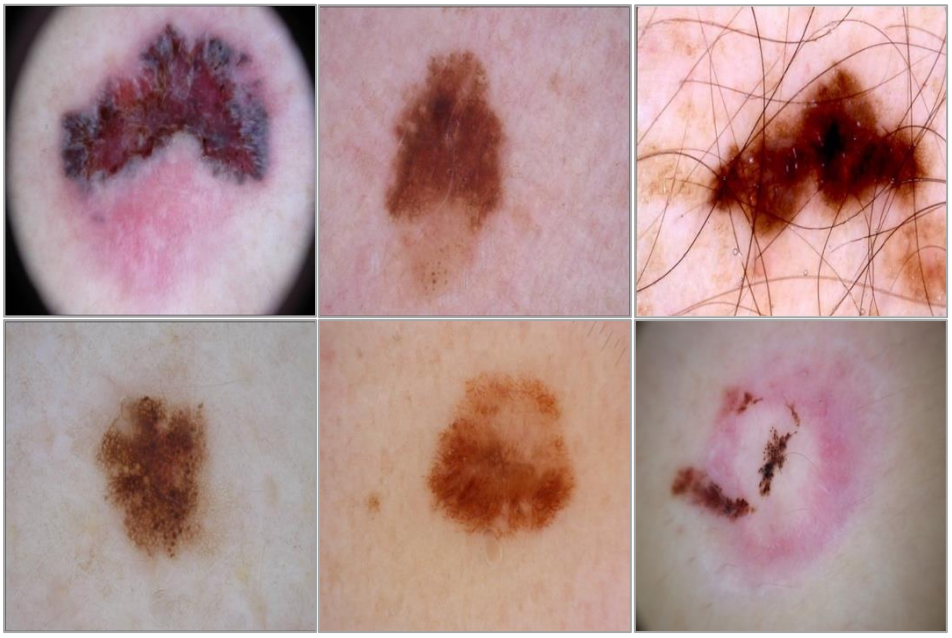

Figure 7. Example images from ISIC 2018 dataset.

4.3. Evaluation Metrics

For the evaluation, different matrices have been used in this research work, such as Intersection over Union (*IoU*), Precision (*P*), Recall (*R*) and Accuracy (*ACC*). These requirements are as defined as follows:

$$IoU = \frac{TP}{TP + FP + FN}$$

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

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

wherein *TP* (True Positive) denotes the number of correctly recognized foreground pixels (interest region), the number of background pixels accurately recognized as a background is referred to as *TN* (True Negative) (skin region), the number of background pixels labelled as the foreground is known as *FP* (False Positive) and the number of foreground pixels incorrectly labelled as background is known as *FN* (False Negative).

The ratio of overlapping and union areas between the expected segmentation mask and the ground truth mask is represented by *IoU*. The *IoU* metric measures how similar the prediction mask is to the ground truth mask. The percentage of correctly identified pixels in the total number of pixels is represented by *ACC*. The fraction of successfully segmented foreground pixels versus the total number of foreground pixels is represented by recall (*R*).

4.4. Qualitative Results

A qualitative examination of the different method's performance is undertaken in this section. The segmentation and edge prediction branches' final results are shown in Figures 8–12. In most cases, as illustrated in Figures 8, 9 and 11, the attention-based methods combined with the properties of U-Net can appropriately segregate the pigment regions as compared to the U-Net based methods when used separately. The first row shows the output predictions given an input image, which is a simple scenario because the colour contrast between the foreground and background regions of the input image is

high. As a result, the attention-based approaches can accurately detect the pigment region. The input images in the bottom rows have hairs and low contrast and deformed shapes, which can distort the textures of the skin lesions and make learning difficult. Despite this, the attention-based methods (TransUNet and Attention U-Net) succeed in segmenting the pigment regions. To put it another way, the attention-based methods show robust performance to the distortions and exhibit good overall performance.

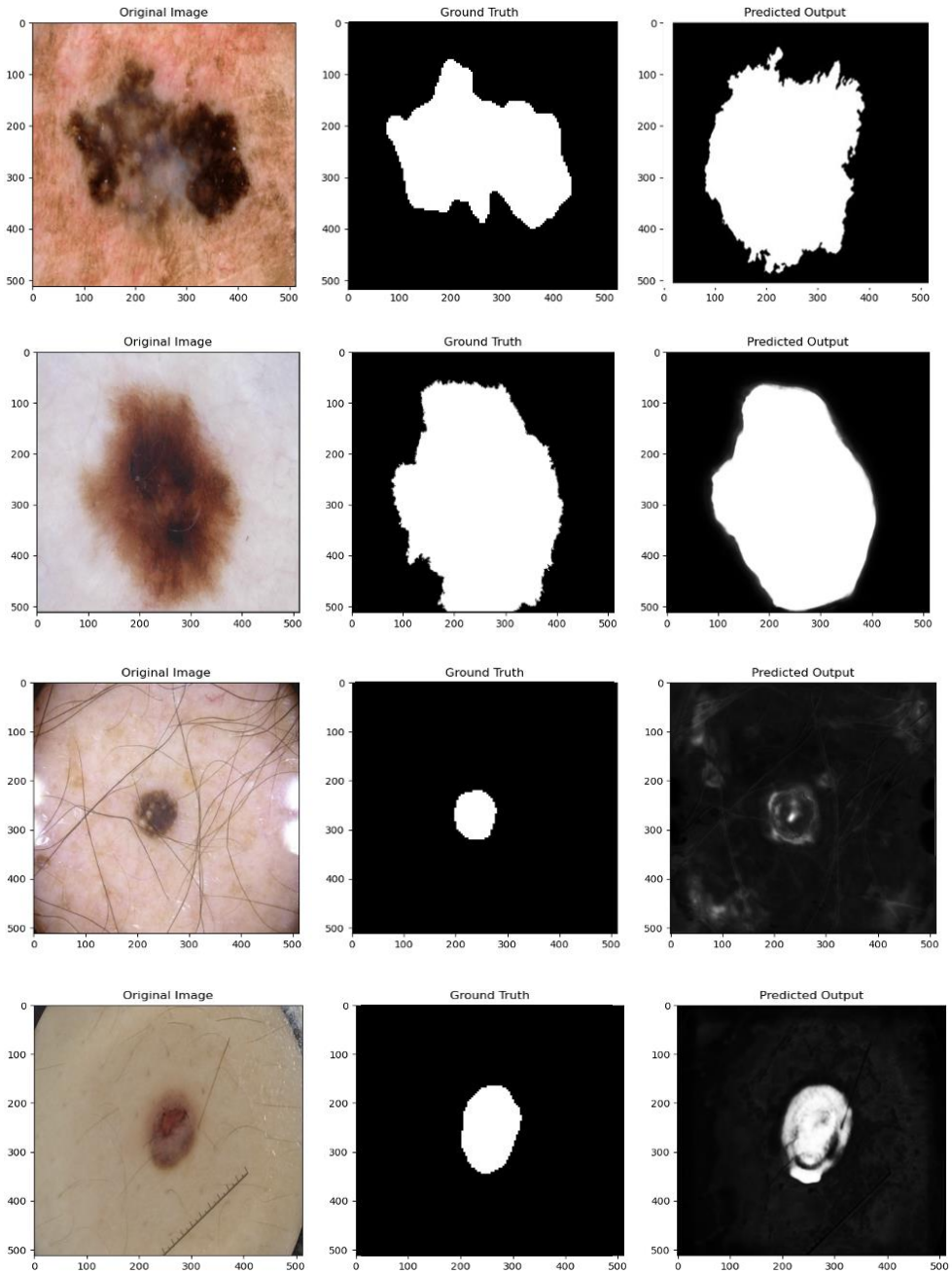

Figure 8. The TransUNet method output visualizations: (**a**) input test image; (**b**) the corresponding ground truth segmentation mask and (**c**) the output probability map of the segmentation prediction.

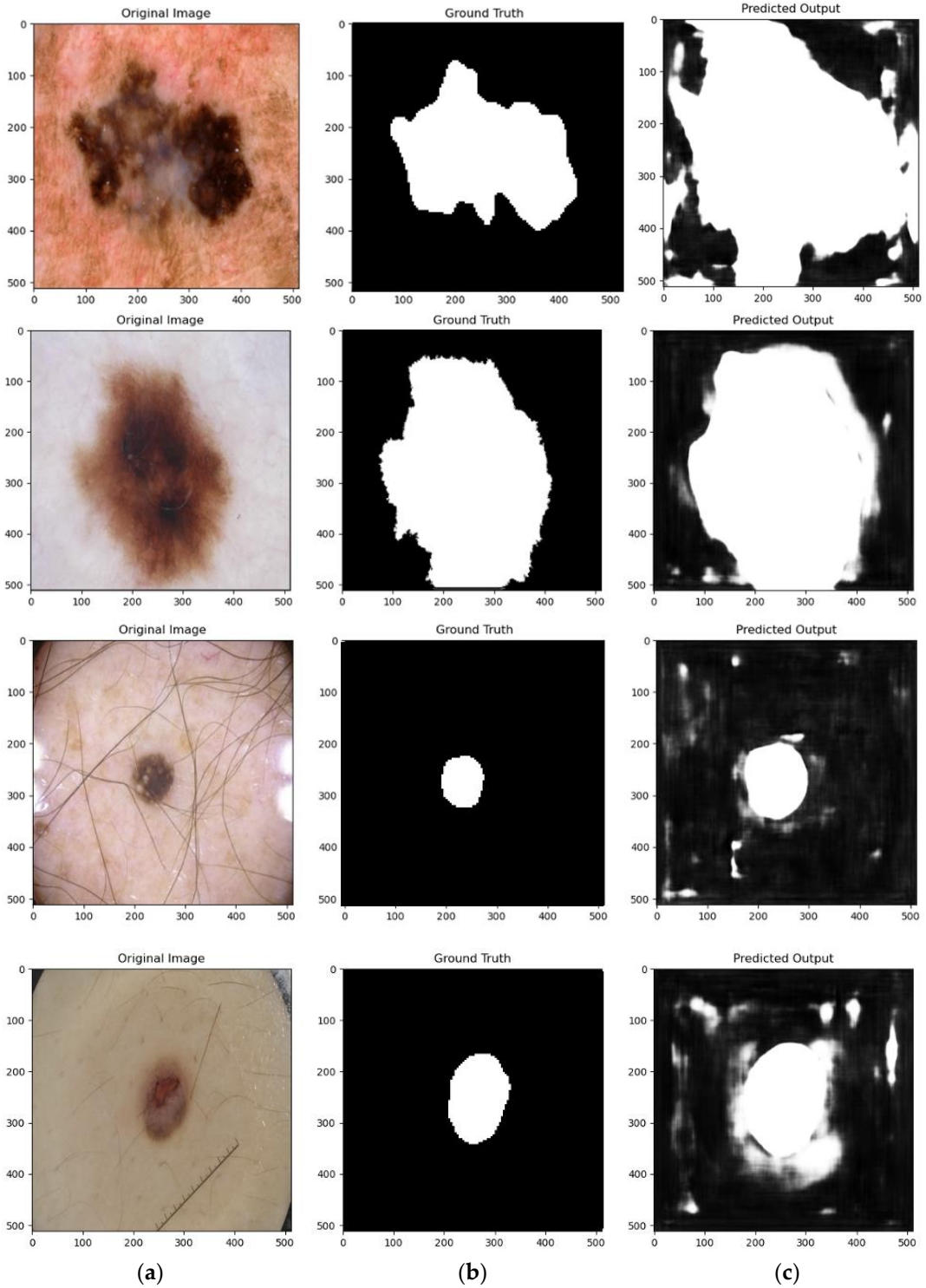

Figure 9. The U-Net output visualizations: (**a**) input test image; (**b**) the corresponding ground truth segmentation mask and (**c**) the output probability map of the segmentation prediction.

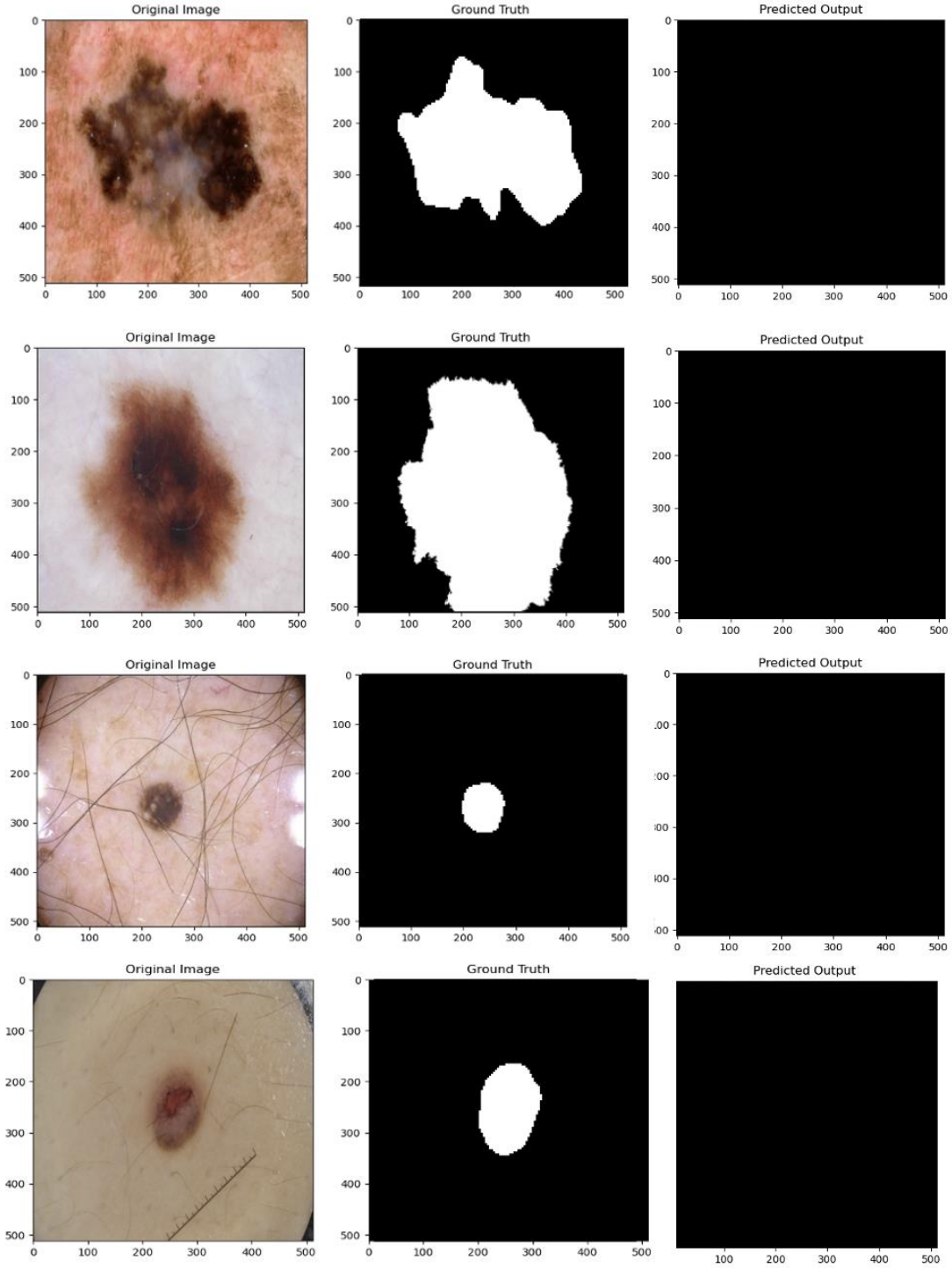

Figure 10. The V-Net output visualizations: (**a**) input test image; (**b**) the corresponding ground truth segmentation mask and (**c**) the output probability map of the segmentation prediction.

It can be clearly seen in the output results from U-Net and Attention U-Net that these two methods generate larger segmentation maps with unclear boundaries as compared to the TransUNet method, whereas the Swin-UNet completely fails to segment the lesions, thus depicting the efficacy of the hybrid method for skin lesion segmentation.

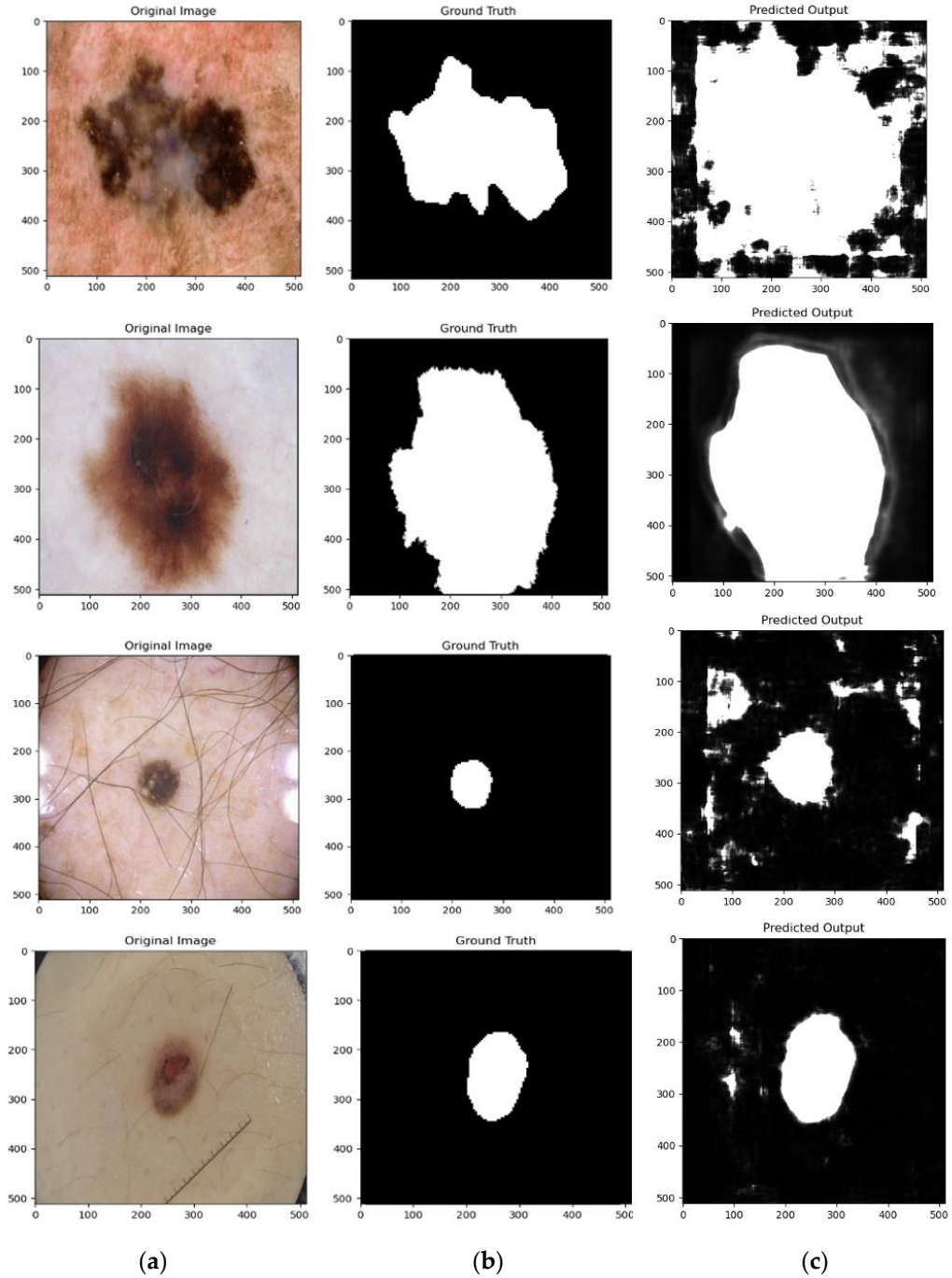

Figure 11. The Attention U-Net output visualizations: (**a**) input test image; (**b**) the corresponding ground truth segmentation mask and (**c**) the output probability map of the segmentation prediction.

To quantitatively evaluate the prediction results, we have used Intersection over Union (*IoU*), Precision (*P*), Recall (*R*) and Accuracy (*ACC*) metrics. The training and testing results are shown in Tables 1 and 2, respectively. The accuracy rate of the TransUNet model reached 50–55% within the 10–12 iterations and the accuracy rates constantly increased up to 120 iterations. From 120 iterations, the accuracy remains almost consistent and reaches 92%. It is due to the fact that the hybrid CNN-transformer, as the hybrid CNN-transformer encoder with long range dependencies, retains the high-level features and passes directly to each stage of decoder and thus results in better output predictions and precise localizations, whereas when compared with other models, it can be seen that there is not much consistency found in terms of accuracy while training the other models (U-Net and Attention U-Net). It can be seen that the accuracy reached 89% and 87% in Attention U-Net and U-Net models,

respectively. The results in Tables 1 and 2 summarize the performance of the different benchmarking architectures in terms of (IoU), Precision (P), Recall (R) and Accuracy (ACC) metrics and validates the efficacy of the TransUNet as compared to the other benchmarking architectures. In addition to the evaluation metrics, we also compared the performance of these models on the basis of training time, inference time and dataset size as shown in Table 3. It can be seen in the results that the hybrid architectures take more time to train than the U-Net and V-Net architectures and take more time for the inference as well.

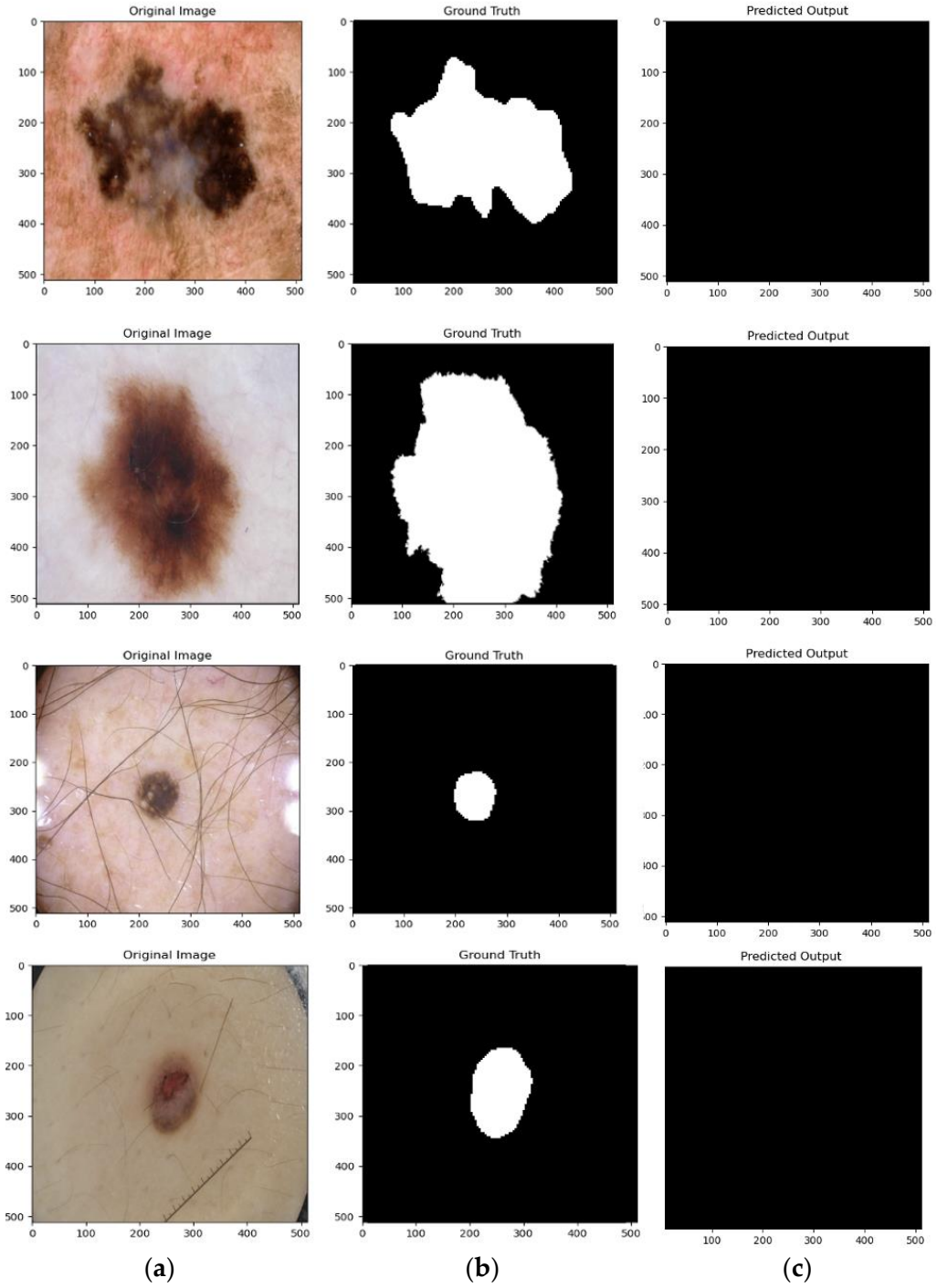

Figure 12. The Swin-UNet output visualizations (**a**) input test image; (**b**) the corresponding ground truth segmentation mask; (**c**) the output probability map of the segmentation prediction.

Table 1. Comparative results of the different methods on train set.

Method	IoU	Dice Coeff	Precision	Recall	Accuracy
U-Net [20]	**80.93**	**82.18**	**86.27**	**98.40**	**87.64**
V-Net [39]	15.95	17.31	21.52	52.04	28.02
Attention U-Net [40]	**81.21**	**83.27**	**86.75**	**98.71**	**88.74**
TransUNet [41]	**86.72**	**89.13**	**89.44**	**99.02**	**91.02**
Swin-UNet [42]	13.25	17.12	14.32	21.20	14.58

Table 2. Comparative results of the different methods on the test set.

Method	IoU	Dice Coeff	Precision	Recall	Accuracy
U-Net [20]	**83.77**	**84.12**	**86.31**	**98.61**	**87.93**
V-Net [39]	16.75	18.13	23.00	53.32	39.24
Attention U-Net [40]	**82.01**	**84.16**	**87.46**	**98.17**	**89.02**
TransUNet [41]	**87.96**	**89.84**	**89.93**	**99.38**	**92.11**
Swin-UNet [42]	13.23	16.98	14.35	21.22	12.21

Table 3. Comparative results in terms of training time, inference time and dataset limit.

Method	Training Time (s)	Inference Time (s)	Dataset Limit
U-Net [20]	24,588	12	2594
V-Net [39]	1172	6	2594
Attention U-Net [40]	34,380	15	2594
TransUNet [41]	77,976	38	2594
Swin-UNet [42]	62,568	27	2594

During the training, it can be depicted from the Table 1 that the TransUNet model outperforms U-Net, V-Net, Attention U-Net and SwinUNet in terms of IoU, Precision and Recall. The accuracy of TransUNet is significantly better than U-Net and Attention U-Net. During testing, again TransUNet performs better than U-Net and Attention U-Net. The accuracy rate increased to around 12% when compared to other models.

5. Conclusions

Skin cancer is considered as one of the deadliest cancers. It is important to detect skin lesions at a very early stage to control its spread as well as to cure it. In this paper, the different methods were evaluated on an ISIC 2018 skin lesion image dataset. Experimental results on the ISIC 2018 data have shown that the hybrid architecture of TransUNet, combining the properties of transformers and U-Net, outperforms other benchmarking methods both qualitatively and quantitatively in terms of various evaluation metrics. Furthermore, it has been shown that the TransUNet method is robust for various deformations, low contrast and noise. As a future avenue, the results of skin lesion segmentation can be combined with skin lesion classification for further diagnosis of lesion areas of skin in terms of identifying the possible progression of the lesion. Malignant skin lesions are considered fatal, and in order to identify such lesions at the right time, early diagnosis is highly recommended.

For that reason, incorporating the clinical knowledge into the lesion image identification system to automatically analyse and diagnose the current state of skin lesions with a high accuracy rate is crucial.

Author Contributions: Conceptualization, Y.G. and S.A.K.; Data curation, S.A.K.; Formal analysis, S.A.K.; Funding acquisition, Y.G.; Investigation, S.A.K.; Methodology, Y.G. and S.A.K.; Project administration, Y.G.; Resources, Y.G. and S.A.K.; Software, S.A.K.; Supervision, Y.G. and S.A.K.; Validation, Y.G. and S.A.K.; Visualization, Y.G. and S.A.K.; Writing—original draft, Y.G. and S.A.K.; Writing—review & editing, Y.G. and S.A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, Project No. GRANT382.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is a public dataset (ISIC 2018 Dataset [13]).

Conflicts of Interest: The authors declare no conflict of interest. The funder had no role in the design of the study, in the writing of the manuscript or in the decision to publish the results.

References

1. WHO. Key Facts about Cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 5 April 2022).
2. Verma, R.; Anand, S.; Vaja, C.; Bade, R.; Shah, A.; Gaikwad, K. Metastatic Malignant Melanoma: A Case Study. *Int. J. Sci. Study* **2016**, *4*, 188–190.
3. Siegel, R.L.; Miller, K.D.; Goding Sauer, A.; Fedewa, S.A.; Butterly, L.F.; Anderson, J.C.; Cercek, A.; Smith, R.A.; Jemal, A. Colorectal Cancer Statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 145–164. [CrossRef] [PubMed]
4. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
5. Kroemer, S.; Frühauf, J.; Campbell, T.M.; Massone, C.; Schwantzer, G.; Soyer, H.P.; Hofmann-Wellenhof, R. Mobile Teledermatology for Skin Tumour Screening: Diagnostic Accuracy of Clinical and Dermoscopic Image Tele-evaluation Using Cellular Phones. *Br. J. Dermatol.* **2011**, *164*, 973–979. [CrossRef] [PubMed]
6. Alves, J.; Moreira, D.; Alves, P.; Rosado, L.; Vasconcelos, M.J.M. Automatic Focus Assessment on Dermoscopic Images Acquired with Smartphones. *Sensors* **2019**, *19*, 4957. [CrossRef] [PubMed]
7. Ngoo, A.; Finnane, A.; McMeniman, E.; Soyer, H.P.; Janda, M. Fighting Melanoma with Smartphones: A Snapshot of Where We Are a Decade after App Stores Opened Their Doors. *Int. J. Med. Inform.* **2018**, *118*, 99–112. [CrossRef] [PubMed]
8. Pehamberger, H.; Steiner, A.; Wolff, K. In Vivo Epiluminescence Microscopy of Pigmented Skin Lesions. I. Pattern Analysis of Pigmented Skin Lesions. *J. Am. Acad. Dermatol.* **1987**, *17*, 571–583. [CrossRef]
9. Stolz, W. ABCD Rule of Dermatoscopy: A New Practical Method for Early Recognition of Malignant Melanoma. *Eur. J. Dermatol.* **1994**, *4*, 521–527.
10. Argenziano, G.; Fabbrocini, G.; Carli, P.; De Giorgi, V.; Sammarco, E.; Delfino, M. Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions: Comparison of the ABCD Rule of Dermatoscopy and a New 7-Point Checklist Based on Pattern Analysis. *Arch. Dermatol.* **1998**, *134*, 1563–1570. [CrossRef]
11. Yu, L.; Chen, H.; Dou, Q.; Qin, J.; Heng, P.-A. Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Trans. Med. Imaging* **2016**, *36*, 994–1004. [CrossRef]
12. Liu, L.; Mou, L.; Zhu, X.X.; Mandal, M. Automatic Skin Lesion Classification Based on Mid-Level Feature Learning. *Comput. Med. Imaging Graph.* **2020**, *84*, 101765. [CrossRef]
13. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M. Skin Lesion Analysis toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (Isic). *arXiv* **2019**, arXiv:1902.03368.
14. Li, Y.; Shen, L. Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network. *Sensors* **2018**, *18*, 556. [CrossRef] [PubMed]
15. Singh, V.K.; Abdel-Nasser, M.; Rashwan, H.A.; Akram, F.; Pandey, N.; Lalande, A.; Presles, B.; Romani, S.; Puig, D. FCA-Net: Adversarial Learning for Skin Lesion Segmentation Based on Multi-Scale Features and Factorized Channel Attention. *IEEE Access* **2019**, *7*, 130552–130565. [CrossRef]
16. Yang, X.; Zeng, Z.; Yeo, S.Y.; Tan, C.; Tey, H.L.; Su, Y. A Novel Multi-Task Deep Learning Model for Skin Lesion Segmentation and Classification. *arXiv* **2017**, arXiv:1703.01025.

17. Xie, Y.; Zhang, J.; Xia, Y.; Shen, C. A Mutual Bootstrapping Model for Automated Skin Lesion Segmentation and Classification. *IEEE Trans. Med. Imaging* **2020**, *39*, 2482–2493. [CrossRef]
18. Liu, L.; Tsui, Y.Y.; Mandal, M. Skin Lesion Segmentation Using Deep Learning with Auxiliary Task. *J. Imaging* **2021**, *7*, 67. [CrossRef]
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
20. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
21. Berseth, M. ISIC 2017-Skin Lesion Analysis towards Melanoma Detection. *arXiv* **2017**, arXiv:1703.00523.
22. Chang, H. Skin Cancer Reorganization and Classification with Deep Neural Network. *arXiv* **2017**, arXiv:1703.00534.
23. Liu, L.; Mou, L.; Zhu, X.X.; Mandal, M. Skin Lesion Segmentation Based on Improved U-Net. In Proceedings of the 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), Edmonton, AB, Canada, 5–8 May 2019; pp. 1–4.
24. Abhishek, K.; Hamarneh, G.; Drew, M.S. Illumination-Based Transformations Improve Skin Lesion Segmentation in Dermoscopic Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 728–729.
25. Nasir, M.; Attique Khan, M.; Sharif, M.; Lali, I.U.; Saba, T.; Iqbal, T. An Improved Strategy for Skin Lesion Detection and Classification Using Uniform Segmentation and Feature Selection Based Approach. *Microsc. Res. Tech.* **2018**, *81*, 528–543. [CrossRef] [PubMed]
26. Afza, F.; Khan, M.A.; Sharif, M.; Rehman, A. Microscopic Skin Laceration Segmentation and Classification: A Framework of Statistical Normal Distribution and Optimal Feature Selection. *Microsc. Res. Tech.* **2019**, *82*, 1471–1488. [CrossRef] [PubMed]
27. Damian, F.A.; Moldovanu, S.; Dey, N.; Ashour, A.S.; Moraru, L. Feature Selection of Non-Dermoscopic Skin Lesion Images for Nevus and Melanoma Classification. *Computation* **2020**, *8*, 41. [CrossRef]
28. Khan, S.A.; Gulzar, Y.; Turaev, S.; Peng, Y.S. A Modified HSIFT Descriptor for Medical Image Classification of Anatomy Objects. *Symmetry* **2021**, *13*, 1987. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
31. Maninis, K.-K.; Caelles, S.; Pont-Tuset, J.; Van Gool, L. Deep Extreme Cut: From Extreme Points to Object Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 616–625.
32. Javed Awan, M.; Mohd Rahim, M.S.; Salim, N.; Mohammed, M.A.; Garcia-Zapirain, B.; Abdulkareem, K.H. Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach. *Diagnostics* **2021**, *11*, 105. [CrossRef]
33. Jafari, M.H.; Karimi, N.; Nasr-Esfahani, E.; Samavi, S.; Soroushmehr, S.M.R.; Ward, K.; Najarian, K. Skin Lesion Segmentation in Clinical Images Using Deep Learning. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, Cancun, Mexico, 4–8 December 2016; pp. 337–342.
34. He, K.; Sun, J.; Tang, X. Guided Image Filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1397–1409. [CrossRef]
35. Yuan, Y. Automatic Skin Lesion Segmentation with Fully Convolutional-Deconvolutional Networks. *arXiv* **2017**, arXiv:1703.05165.
36. Al-Masni, M.A.; Al-Antari, M.A.; Choi, M.-T.; Han, S.-M.; Kim, T.-S. Skin Lesion Segmentation in Dermoscopy Images via Deep Full Resolution Convolutional Networks. *Comput. Methods Programs Biomed.* **2018**, *162*, 221–231. [CrossRef]
37. Bi, L.; Kim, J.; Ahn, E.; Kumar, A.; Feng, D.; Fulham, M. Step-Wise Integration of Deep Class-Specific Learning for Dermoscopic Image Segmentation. *Pattern Recognit.* **2019**, *85*, 78–89.
38. Sarker, M.; Kamal, M.; Rashwan, H.A.; Akram, F.; Banu, S.F.; Saleh, A.; Singh, V.K.; Chowdhury, F.U.H.; Abdulwahab, S.; Romani, S. SLSDeep: Skin Lesion Segmentation Based on Dilated Residual and Pyramid Pooling Networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; Springer: Cham, Switzerland, 2018; pp. 21–29.
39. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 4th International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
40. Oktay, O.; Schlemper, J.; le Folgoc, L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
41. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
42. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv* **2021**, arXiv:2105.05537.