Skin Lesion Semantic Segmentation based on Transformer

Siyi Du, Ben Hers
Department of Electrical and Computer Engineering, University of British Columbia
Vancouver, BC, Canada
{siyi, bhers}@ece.ubc.ca

Abstract

Skin cancer is one of the most serious diseases worldwide, causing millions of deaths every year. Automatic segmentation can facilitate diagnosis and treatment planning, and improve the patients' survival rate. Although Convolutional Neural Network (CNN) based models performed well but most existing architectures are still inefficient due to inductive bias caused by the lack of global context. The newly released Transformer with attention mechanisms was considered a promising tool to encode global dependencies, but it also faces some shortcomings. First, Transformers can't extract sufficient local details to distinguish ambiguous boundaries. Second, their complex architecture require large datasets to train. We build several Transformer based models to tackle these challenges and devote ourselves to finding the best architecture to segment skin lesions. We replace the linear layers with convolutional layers before the Transformer encoder to get a small feature map and integrate a boundary-aware attention gate to enable the network to attend to ambiguous boundaries. After that we imitate Unet architecture that uses skip connection and upsampling to get the final predictive map. We implement detailed experiments on our networks and also compare them against state-of-the-art CNN based models.

1. Author Contribution

Two authors contributed equally in this project.
Idea Design: Ben and Siyi.
Experiments: Ben coded and ran UNET, UNET++, and BAT, wrote metrics and visualization code. Siyi preproceeded the data, coded and ran BAT and BATU.
Report: Ben wrote Related Works and Evaluation. Siyi wrote Abstract, Introduction and Method.

2. Introduction

Cancer is the most serious disease worldwide, causing nearly 10 million deaths in 2022, or about one in six deaths, and skin is among the 6 most common body parts for cancer [19], while skin lesions are easily to be noticed and early diagnosis and treatment can greatly improve the patients' survival rate [3]. Diagnosis and detection of skin cancer diseases have been traditionally carried out by dermatologists via manual screening and visual inspection, which is time consuming, complex and error-prone. Thanks to the development of computer vision, automatic skin lesion diagnosis can flag which ones it thinks are cancerous, freeing physicians from tedious inspection tasks.

Lesion segmenting, the key step of the whole automatic diagnosis, receives much research attention, while still a quite challenging task due to the reasons outlined below. First, skin lesion images may include hair, blood vessels and other noises to mutilate or cover the lesion boundaries. Also, low contrast in-between the surrounding skin and the lesion area as well as diverse illumination may cause ambiguous boundaries and increase segmenting difficulties. Besides, skin lesions vary in shapes, sizes, and colors, which occludes the methods to achieve high accuracy. Classical hand-crafted feature based methods like threshold-based [9] or edge and region-based methods [1] [8] are powerless when facing these issues. Convolutional Neural Network (CNN) methods that can be optimized to generate more useful features were introduced in this area and achieved great results, such as U-Net based [23] and Fully Convolutional Network (FCN) based [2] methods. However, CNN based architectures are not at a point to completely take over dermatologists' work due to the inductive bias caused by the lack of global context [7]. In every CNN layer, some 3×3 or 5×5 filters (very small size compared to the whole picture size like 224×224) slides over the whole image, making each pixel in the convoluted feature map only attends to its close neighbors.

The Transformer architecture was first introduced in 2017 [22] to solve Recurrent Neural Network (RNN)'s slow processing speed, then quickly became dominant in natural language processing. Its Multi-head attention mechanism enables each word to have a global view of the whole sentence and then was introduced into the computer vision [7] [16]. These Transformer based models overcame the inductive bias issue in CNNs, thus realizing state-of-the-art results in many visual tasks. Nevertheless, Transformers are difficult to deploy for skin lesion segmentation. First, Transformers cannot extract sufficient local details to distinguish ambiguous boundaries. Also, its complex architecture requires huge datasets to train while the general skin lesion dataset only contains thousands of images. To conquer these obstacles, some papers built hybrid networks via combining Transformer and CNN together [25] [26], which shows great performance on small skin datasets. To step further to let the model pay special attention to the lesion ambiguous boundaries, Wang *et al*. [24] incorporated the Transformer with boundary prior so that the network predicts boundary key-points maps to guide the attention module. We follow the insight of the boundary prior research [24] to propose a Boundary-Aware Transformer-Unet (BATU) framework that works well with small datasets and pays attention to the global information as well as local details. This framework consists of two stages: Encoder and Decoder. In the Encoder stage, we implement a convolutional extractor to extract a compact feature map and get the insight of [24] to introduce a Boundary-Aware Transformer Encoder (BATE) that captures the global context and notices key boundaries. In the Decoder stage, we imitate the Unet [18] architecture that is suitable for medical imaging segmentation, including convolution, upsampling, and skip-connection. Our main contributions can be summarized as follows:

(1) We propose a framework combining the advantages of CNN and Transformer and utilizing the prior boundary information.

(2) We implement detailed experiments comparing the proposed model with CNN based networks and baseline to show the model's good performance.

(3) We also conduct an ablation study to investigate the influence of some modules.

3. Related Work

Skin Lesions Segmentation and classification have attracted attention because skin cancer is one of the most common forms of cancer [14] [13]. Until recently traditional CNNs have been dominating research into skin lesion segmentation. Yu et al. [29] used deep residual networks to first predict a segmentation mask for the skin lesion and then passed the cropped skin lesion through another network to predict whether or not it was dangerous or malignant. The way they were able to get incredible performance was by being the first to train very deep networks for this difficult task. One downside to their approach was they did a segmentation task to segment out the lesion in the image then used a label whether it was cancerous or not. This two stage pipeline further complicates the process and difficulty of collecting data by requiring not only someone to segment out the image but also label it. One fix to this is training each section separately which could cause issues. Xu et al. [28] leveraged recent advances in CNN architectures and created a UNET based network that implemented depthwise convolutions and split attention. One new development they made was their CSA block which split the feature maps channel wise into two streams and then did a series of skip connections and convolutions that allows more capacity and has more parameters than a typical residual block. However, in their decoder or DC block they used 7x7 convolutions which if used sequentially either required immense zero-padding or would limit the ability to accurately segment around the edges of the image. Although CNN based architectures achieve great results in the skin lesion segmentation task, the inductive bias problem prevents them from further development.

Vision Transformers was introduced in 2020 by Dosovitskiy et al.[7]. They leveraged the Transformer architecture from natural language processing [22] and adapted it to computer vision, and then lots of Transformer based models [16][30] were proposed and flourished from there. The biggest benefit of these Transformers is their multi-head attention blocks have global attention so they can compare features across the entire image. CNNs have a limited receptive field so when predicting the class of a pixel it can't grasp features across the entire image. This can be an issue in the case that one corner of a skin lesion image has a feature that is important to another corner but it sits outside the receptive field of that pixel. However, Transformers are not perfect. One limitation of them is that they lack the translation equivariance and locality [7] that CNNs have, so they are more difficult to train and often need more data to perform as well as or better than CNN based architectures.

Transformers in Medical Imaging has gained much attention in recent years, but medical imaging tasks generally have scarce data resources with millions of pixels in one image, making the complex Transformer models hard to train. Researchers have gone 2 directions to tackle this issue: building hybrid models that combine convolutional and multi-head attention layers and improving purely transformer based models in architecture or loss function. Hatamizadeh et al. [10] created a hybrid model where they had Transformer based encoders but used convolutional based decoders to get the best performance for MRI segmentation. One impressive feat of this research was that they were able to create a Transformer encoder with a convolutional decoder that might have had more parameters but overall FLOPS for a forward pass was smaller than other state-of-the-art methods. On the flip side of this is that since the model has a lot of parameters, it could limit the ability of researchers to deploy and train or fine-tune these models if they don't have access to the same resources.

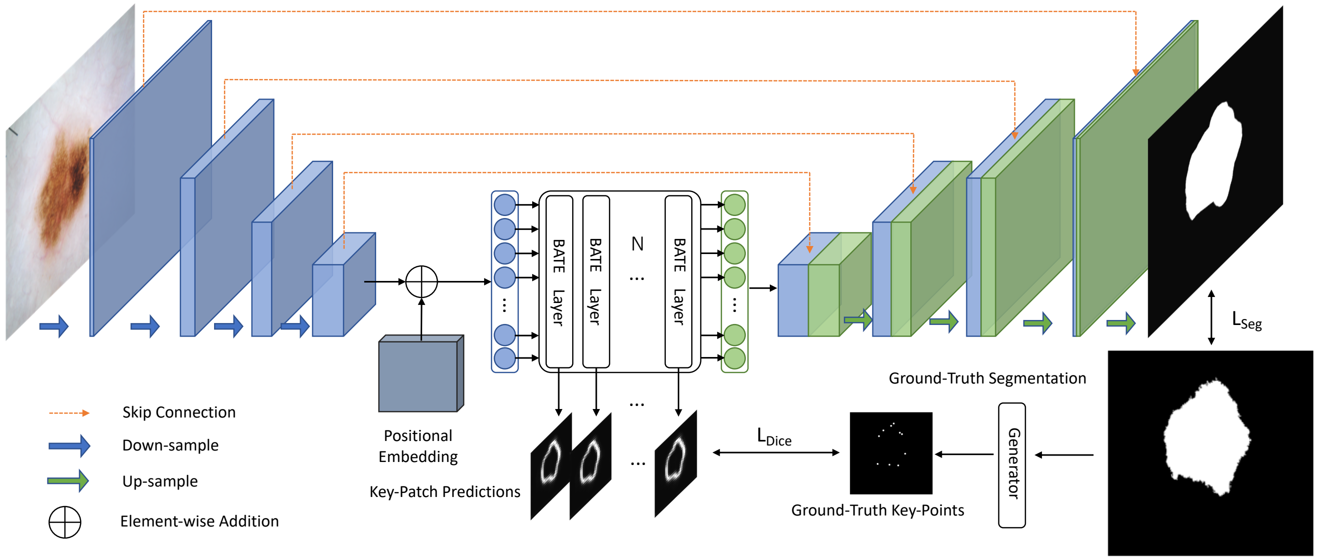

Figure 1. Overall structure of the proposed BATU model

These researchers had access to a state-of-the-art DGX-1 cluster to train on. Xie et al. [27] leveraged deformable attention [32] blocks in the bottleneck of the UNET to beat standard UNET based architectures but kept convolutional based encoders and decoders, which reduces the computational complexity of using Transformers for researchers with constrained resources. One tradeoff though in this paper is by using deformable attention they sampled subsets of the attention to reduce complexity. Other research that has full or normal attention have more parameters but tend to outperform this method. This is a useful tool to know if researchers have limited access to GPU's.

Valanarasu et al. [21] created a purely transformer based model but leveraged more efficient versions of attention to speed up computation and improve performance. This method helped tackle the problem of limited data in medical imaging by being creative and incorporating axial attention [12] to help factor attention into two 1D attention matrices. One downside of this paper was that they have two branches that mimic the UNET style of doing segmentation yet this paper is trying to say it's a medical transformer paper. Cao et al. [4] created the Swin Transformer which was a UNET type network except they replaced every convolutional aspect with multi-head attention but kept the UNET structure with encoder, bottleneck, decoder and skip connections from encoder to decoder. They were able to leverage the structure of UNET which is proven to be effective but also incorporate attention which let them achieve impressive performance. One downside of this research is that they use pre-trained weights from ImageNet which may be suboptimal when transferring over to medical imaging.

Transformers for skin lesions is starting to gain traction, wherein a recent conference Wang et al. [24] developed a Transformer based framework that swapped the initial linear layer in the Transformer for a small number of convolutional layers and helped the training process by not only getting the network to segment out the skin lesion but also predict boundary. Lots of research has gone into integrating structural boundary information [24] into CNN's so they seek to bring this into the transformer model. However, they used DeepLabV3 [5] as the backbone that is suitable for semantic segmentation in general computer vision tasks. UNET [18] architecture is more common in medical imaging segmentation area for its skip connection facilitates the information transmission from encoder to decoder. In this paper, we proposed Boundary-aware Transformer-Unet (BATU) framework that combine unet backbone and transformer encoder together and utilize boundary prior information, which beats CNN based models and has higher generalizing ability compared to Boundary-aware Transformer [25].

4. Method

In the skin lesion segmentation task, the model is required to output a binary segmentation map $\mathbf{S} \in \mathbb{R}^{H \times W \times 1}$ based on a RGB skin image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$. H and W are the height and width of the image. As shown in Figure 1, our *Boundary-Aware Transformer-UNet* (BATU) model consists of two stages: encoder and decoder. We first implement a convolutional neural network (CNN) to get a feature map $\mathbf{I}_f$ of the input image, and then utilize a Boundary-Aware Transformer Encoder (BATE) module to grasp global context as well as enable the model to attend on ambiguous boundaries by predicting key-points maps. In the second stage, We got inspiration from UNet [18].

The encoded representation is sent into a decoder that skip-connects encoder convolutional layers and repeatedly conducts upsampling to get the final segmentation map. We will introduce the details of each module in this section.

4.1. Encoder Stage

Convolutional Feature Extractor The Transformer architecture requires a sequence as an input. Generally, a Transformer like ViT [7] splits the image into 2D patches and then directly flattens and applies a linear projection to these patches, which has millions of learnable parameters. To overcome the lack of skin data and utilize the inductive bias in CNN, we follow [25] to replace the linear projection as a convolutional feature extractor before the Transformer module, acquiring a denser feature map $\mathbf{I}_f \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C}$, i.e. the patch size is 16×16. C is the number of channels. The feature map is then passed through a convolutional layer to adjust the number of channels to K to reduce the computational cost of BATE. The output is flattened into 1D sequence embedding and added by a learnable positional embedding that is randomly initialized to compensate for 2D spatial information destroyed by sequentialization. Finally, the BATE input embedding is $\mathbf{I}_p \in \mathbb{R}^{L\times K}$. $L = \frac{H\times W}{256}$ is the sequence length and K is the embedding dimension.

Boundary-Aware Transformer Encoder (BATE)

Boundary-Aware Transformer was first introduced by Wang *et al.* [25] as an upgraded basic Transformer module with a boundary-wise attention gate. We imitate their architecture and propose the Boundary-Aware Transformer Encoder (BATE) where we include a general Transformer encoder and boundary-wise attention gate. As shown in Figure 1 and 2, we stack N encoder layers, each of which consists of multi-head self-attention (MSA), multi-layer perceptron (MLP), and boundary-wise attention gate (BAG).

MSA and MLP follows the typical design [22]. The former one gets three branches (query, key, and value) that are linearly transformed from the layer input, and then calculates the attended result. The latter module, a fully connected feed-forward network, is applied to each position in the sequence separately and identically. It contains two linear transformations with a ReLU activation function in between. Assuming the input of the i-th encoder layer is $\mathbf{Z}_{i-1}$ (specially, $\mathbf{Z}_0 \leftarrow \mathbf{I}_p$), we can write the calculation as follows:

$$\tilde{\mathbf{Z}}_i = MSA(LN(\mathbf{Z}_{i-1})) + \mathbf{Z}_{i-1} \quad (1)$$

$$\hat{\mathbf{Z}}_i = MLP(LN(\tilde{\mathbf{Z}}_i)) + \tilde{\mathbf{Z}}_i \quad (2)$$

BAG utilizes boundary information from the ground-truth segmentation, which helps the whole model pay more attention to ambiguous boundaries and increases segmentation performance. It first predicts a binary key-patch sequence $\mathbf{M}_i = \sigma(conv(\hat{\mathbf{Z}}_i)) \in \mathbb{R}^{L\times 1}$, which will be used to compute the distance with ground-truth key point map M_{gt} and optimized to reduce the difference. Next, we utilize this boundary-aware binary map as a gate to filter important information related to boundaries, which will be added by a residual connection to get the final enhanced representation.

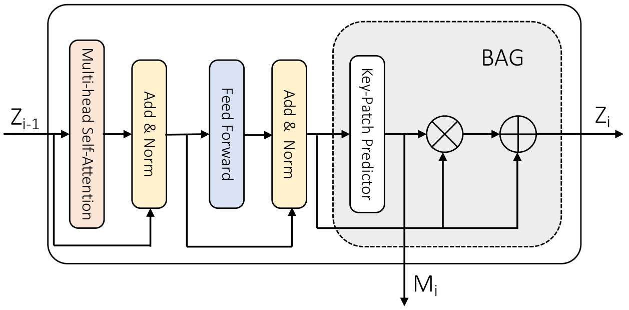

Figure 2. The illustration of i-th BATE layers

$$\mathbf{Z}_i = \hat{\mathbf{Z}}_i + (\hat{\mathbf{Z}}_i \times \mathbf{M}_i) \quad (3)$$

To get the supervised ground-truth key-point map, we conduct a conventional edge detection algorithm as Wang *et al.* [24] did on the segmentation mask to get a set of boundary points. For each point in the set, we draw a circle of radius r and calculate the proportion p of lesion area in this circle, scoring them with $|p - 0.5|$. A higher score indicates the boundary is not smooth in the circle. Those points with higher scores are pitched up and filtered by the classical Non-maximum suppression algorithm in computer vision to get the final key points. The ground-truth key-point map $\mathbf{M}_{gt} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1}$, in where the patches having key points are1 and others are 0. Notice we will flatten the ground-truth map as $\mathbb{R}^{L\times 1}$ to be the same size as predictions.

4.2. Decoder Stage

Following UNet [18] decoder architecture, we implement a CNN decoder to perform feature upsampling and pixel segmentation.

Feature Mapping layer reshapes the encoded representation $\mathbf{Z}_{N+1} \in \mathbb{R}^{L\times K}$ to $\frac{H}{16} \times \frac{W}{16} \times K$, then recovers it to C channels though a convolutional layer. The final feature map $\mathbf{Z} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C}$ is the same size as the Convolutional feature extractor output.

Progressive Feature Upsampling contains cascaded upsampling operations and convolutional blocks to get the final segmentation map $\mathbf{S} \in \mathbb{R}^{H\times W\times 3}$. We also use skip-connections from the feature extractor side to enrich the feature's information.

4.3. Objective Function

We optimize the whole architecture through two losses. The first loss function we seek to minimize is the Dice loss which measures the difference between the predicted and ground truth segmentation map. The second is Cross-Entropy loss between ground-truth key-point map and predicted key-patch maps (notice each BATE layer will output a key-patch map, so there are N maps). The final objective function can be written as follows:

$$L_{total} = L_{Seg} + \sum_{i=1}^{N+1} L^{i}_{Point} \tag{4}$$

$$L_{Seg} = \Phi_{Dice}(\mathbf{S}_{gt}, \mathbf{S}) \tag{5}$$

$$L^{i}_{Point} = \Phi_{CE}(\mathbf{M}_{i}, \mathbf{M}_{gt}) \tag{6}$$

5. Experiments

Datasets: We implement extensive experiments on widely used skin lesion segmentation datasets: ISIC2018 [6] [20] and PH2 [17]. International Skin Imaging Collaboration (ISIC) archived high-quality skin lesion images from International Symposium on Biomedical Imaging (ISBI) and hosted an annual lesion detection and analysis challenge from 2016 to boost the research in this area. ISIC 2018 is the dataset for the challenge in 2018, which contains 2594 samples with corresponding ground-truth segmentation masks. Mendonça *et al.* first introduced PH2 database consisting of 200 dermoscopic images with manual segmentation labels and clinical diagnosis. We split the ISIC 2018 into train and test sets in 7:3 proportion and treat PH2 as an extra test set.

Metrics: The metrics we used to quantitatively analyze the results is the Dice score and Intersection over Union (IoU). We use the Dice score to measure how close the predicted segmentation map is to the ground truth segmentation map. Our network outputs probabilities between 0 and 1 and we use a threshold of 0.5 to predict whether or not a pixel is apart of the skin lesion. Using this we can calculate how many true positives (TP) (where prediction is 1 and ground truth is 1), false positives (FP) (where prediction is 1 but ground truth is 0), and false negatives (FN) (where prediction is 0 but ground truth is 1) to calculate the Dice score.

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{7}$$

IoU is another good metric and measures how well our predicted segmentation map overlaps with the ground truth maps. To calculate IoU we do the same thresholding at 0.5 and then compare the values. Using a logical AND as well as OR we can calculate the IoU of the maps.

$$IoU = \frac{Predicted \cap Actual}{Predicted \cup Actual} \tag{8}$$

Implementation Details: We conduct 8 convolutional layers and one max-pooling layer in the feature extractor to reduce the height and width of the image to 1/16. The BATE's hidden dimension is 128 and attention head is 8. We stack 6 BATE layers and each of them will output a predicted key-patch map. Images are augmented through adding Gaussian noise, horizontal flipping, vertical flipping and randomly shifting/scaling/rotating to boost data diversity, then resized to $512 \times 512 \times 3$. We use Adam [15] optimizer to train the model with initial learning rate 1×10^{-4}. The learning rate changes through *CosineAnnealingWarmRestarts* scheduler in the PyTorch. We deploy the model on a single TITAN V GPU and train it with batch size 4 and 200 epochs.

5.1. Overall Results

Using the parameters described in the implementation details we got the following results. We compared our BATU model against two classical convolutional based methods UNET, and UNET++. Both UNET and UNET++ are standard encoder-decoder architectures where we encode our image into a higher dimensional feature map which we then decode through a series of convolutions and interpolations. UNET++[31] improves on the classical UNET by using dense and nested skip pathways which helps the optimizer have an easier task to learn. We utilize skip connections across the *U* shaped architecture to help the network look at features from both the lower dimensional encoder as well the higher dimensional features coming out of the decoder. We also compared our BATU model against the implementation in the boundary-aware transformer (BAT) [24] which we based our method off of. BAT utilized DeepLabV3 [5] as the down-sample and up-sample backbone and also inserted the Transformer decoder, which is different from BATU architecture.

Model	ISIC 2018		PH2	
	Dice ↑	IoU↑	Dice↑	IoU↑
UNET[18]	0.898	0.821	0.861	0.789
UNET++[31]	0.875	0.786	0.867	0.776
BAT[24]	**0.910**	**0.844**	0.894	0.818
BATU (our)	0.906	0.841	**0.902**	**0.828**

Table 1. Performance on ISIC 2018 and PH2 Datasets

From Table 1 we find it beats state-of-the-art CNN models (outperforms UNET by 0.008 on Dice score and 0.02 for IoU for ISIC2018). We found that the original BAT model outperforms our model for the ISIC2018 dataset but our BATU model beats the BAT baseline for the PH2 dataset. We believe the structure of our model generalizes better to unseen data(PH2) and has fewer parameter resulting in the best performance. We can see that both transformer models outperform CNN based models and we believe this is because the global attention mechanism in the BAT/BATU enable the models to have a better understanding of the whole image. Besides, BAT/BATU introduces the ambiguous boundary information from the label side, which enforces the model to attend on these important positions.

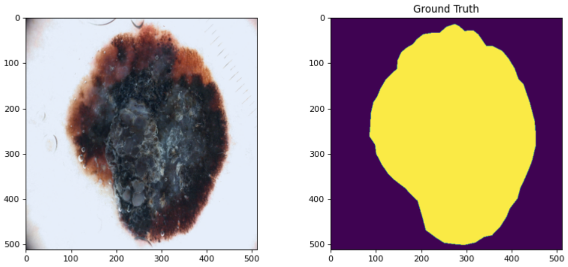

Figure 3. Ground truth for predictions

5.2. Qualitative Analysis

We found that our BATU model didn't have the best Dice score for ISIC2018 but we can see that our model outputs a more consistent prediction than conventional convolutional based models. We can highlight the biggest difference being the center of the skin lesions where in the third example in Figure 4 the convolutional models in the center of the skin lesion inconsistently predict the skin lesion. If we look at our model we can see the effect of forcing the model to look at the boundary prior information because it predicts the skin lesion consistently. Since the convolutional models have a fixed receptive field they cannot attend to the global information of the skin lesion.

We show in the appendix failure modes for skin lesion segmentation where our BATU outperforms the UNET model as well as a failure mode where our BATU model does worse than the UNET based model. We can see that when a skin lesion has an unusual shape or appearance our BATU model is model is more capable at predicting a consistent border because it is able to look globally and was forced to focus on ambigious boundary information. One limitation of our model is strange looking scar tissue around the lesions seem to confuse our BATU model more so then the UNET based model. In this example in the appendix the skin lesion is similar in appearance to the surrounding skin and there is scar tissue in the shape of a semi-circle. We believe this throws off the BATU model because it is unsure if everything inside the scar tissue is apart of the skin lesion or just the inner skin lesion itself.

In the appendix we have included examples where an abundance of hair causes the UNET based models to light up where the hair is close to the lesion but our BATU network is able to identify the hair and not include it in our segmentation maps. This illustrates how the addition of global attention in our BATU network enables us to get state-of-the-art performance that is much more consistent than classical UNET based architectures.

5.3. Ablation Study

Since we could see a much more consistent prediction in our model we wanted to study the effect of the MSA layers into the model as well as the effect of adding the boundary prior information. We performed three experiments to study these effects. The first experiment we performed was taking our convolutional encoder and decoder and instead of adding in MSA layers we simply passed the output of the encoder directly into the input of the decoder. This mimics an architecture similar to UNET and gives us a baseline of no attention or boundary prior information in the model. In the second experiment we took our BATU model but left out the boundary prior information. This architecture resembles a UNET but with our MSA layers in the bottleneck of the model. We can see in the second line in Table 2 that the Dice score stays approximately the same but the addition of the MSA layers shows a noticeable increase in the IoU score. This shows us there is a small but noticeable gains from adding global attention into the bottleneck of an UNET. The third and final experiment we performed was adding both the MSA layers and the boundary prior information. We can see this gives us a gain of approximately 1% for our Dice score and 2% for IoU. Moreover, we could find general DeepLabV3 model performs worse than Unet about 0.019 in Dice score and 0.011 in IoU, thus proving Unet architecture is better than DeepLabV3 in medical imaging area.

Trans.	BAG	Backbone	Dice Score ↑	IoU ↑
		DeepLabV3	0.879	0.810
		Unet	0.898	0.821
√		Unet	0.897	0.830
√	√	Unet	**0.906**	**0.841**

Table 2. Effects of Trans. and BAG on ISIC 2018

Overall these experiments prove that combining both convolutional and MSA based components and that the boundary prior information helps the model learn to look at the shape and size of the skin lesion. This is especially important due to the fact that we were able to generate the boundary keypoints automatically using non-maximal suppression.

6. Discussion

After seeing the performance difference between our BATU model and the original BAT model on the PH2 benchmarking dataset, we wanted to study the differences between them. First of all their architectures differ slightly in terms of encoder/decoder size but we think the number of training parameters played a significant role in the discrepancy. Since our model has 9 million less parameters or 57% of the parameters, it generalized better to the unseen PH2

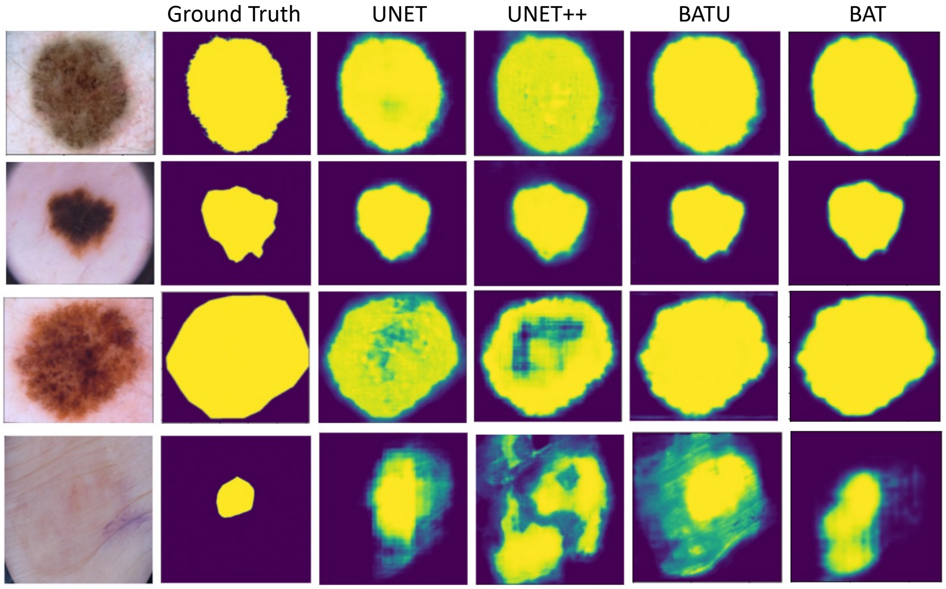

Figure 4. Difference between ground truth and different UNET/BATU models

dataset. Since our encoder and decoder differ, it changed our maximum model size for our GPU memory. We believe that the lower amount of parameters plus different encoder architecture made our model generalize better. Due to the time limitations of this project, we didn't have time to run more ablation tests where we limit the BAT parameters to approximately the same amount of parameters.

Model	Training Parameters
BAT	22.11 million
BATU	12.73 million

Table 3. Comparing number of training parameters

7. Future Work

One area of improvement we saw for our model was in the encoder and decoder architecture. We used a simple combination of 2D convolutions, ReLU layers, and BatchNorm2d layers to create the encoders and decoders but the model performance may increase by using a series of residual blocks as described by He et al.[11]. The skip connections inside these blocks as well as across the UNET architecture could improve performance. Another limitation of this project due to the time limit was that we only compared our model against the original BAT paper and UNET/UNET++. There are different state of the art methods such as TransBTS [25] that have very similar architectures to our BATU method without the boundary prior information. This would be an interesting baseline to compare our model against if we turned off the boundary section off in our BATU model as well.

8. Conclusion

In this novel work, we built upon the work of Wang et al. [24] where they created the BAT architecture. By improving the encoder and decoder of the model we were able to create our BATU model with less parameters that generalized better than the original BAT baseline. Since BAT [24] is one of the few works in skin lesion segmentation using transformers, we also base-lined our model against popular CNN based architectures and showed our model outperforms classical CNN based models as well. Since we wanted to study the additional benefit adding in MSA layers added, we performed ablation where we studied the effects of each of the new blocks, both transformer and boundary blocks gave to the model. We found that just the addition of the MSA layers gave us a small increase in performance but the boundary prior information gave our model a big increase in performance.

References

[1] Qaisar Abbas, M Emre Celebi, Irene Fondón García, and Muhammad Rashid. Lesion border detection in dermoscopy images using dynamic programming. *Skin Research and Technology*, 17(1):91–100, 2011.

[2] Mohammed A Al-Masni, Mugahed A Al-Antari, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Computer methods and programs in biomedicine*, 162:221–231, 2018.

[3] Charles M Balch, Jeffrey E Gershenwald, Seng-jaw Soong, John F Thompson, Michael B Atkins, David R Byrd, Antonio C Buzaid, Alistair J Cochran, Daniel G Coit, Shouluan Ding, et al. Final version of 2009 ajcc melanoma staging

and classification. *Journal of clinical oncology*, 27(36):6199, 2009.

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] M Emre Celebi, Hassan A Kingravi, Hitoshi Iyatomi, Y Alp Aslandogan, William V Stoecker, Randy H Moss, Joseph M Malters, James M Grichnik, Ashfaq A Marghoob, Harold S Rabinovitz, et al. Border detection in dermoscopy images using statistical region merging. *Skin Research and Technology*, 14(3):347–353, 2008.

[9] M Emre Celebi, Quan Wen, Sae Hwang, Hitoshi Iyatomi, and Gerald Schaefer. Lesion border detection in dermoscopy images using ensembles of thresholding methods. *Skin Research and Technology*, 19(1):e252–e258, 2013.

[10] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[13] Imran Iqbal, Muhammad Younus, Khuram Walayat, Mohib Ullah Kakar, and Jinwen Ma. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized Medical Imaging and Graphics*, 88:101843, 2021.

[14] Muhammad Attique Khan, Muhammad Sharif, Tallha Akram, Robertas Damaševičius, and Rytis Maskeliūnas. Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. *Diagnostics*, 11(5):811, 2021.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021.

[17] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[19] WHO staff. Cancer. `https://www.who.int/news-room/fact-sheets/detail/cancer`, 2022.

[20] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[21] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2021.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[23] Sulaiman Vesal, Nishant Ravikumar, and Andreas Maier. Skinnet: A deep learning framework for skin lesion segmentation. In *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, pages 1–3. IEEE, 2018.

[24] Jiacheng Wang, Lan Wei, Liansheng Wang, Qichao Zhou, Lei Zhu, and Jing Qin. Boundary-aware transformers for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 206–216. Springer, 2021.

[25] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021.

[26] Huisi Wu, Shihuai Chen, Guilian Chen, Wei Wang, Baiying Lei, and Zhenkun Wen. Fat-net: Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis*, 76:102327, 2022.

[27] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on med-*

ical image computing and computer-assisted intervention, pages 171–180. Springer, 2021.

[28] Qing Xu, Wenting Duan, and Na He. Dcsau-net: A deeper and more compact split-attention u-net for medical image segmentation. *arXiv preprint arXiv:2202.00972*, 2022.

[29] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE transactions on medical imaging*, 36(4):994–1004, 2016.

[30] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021.

[31] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.

[32] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A. Appendix

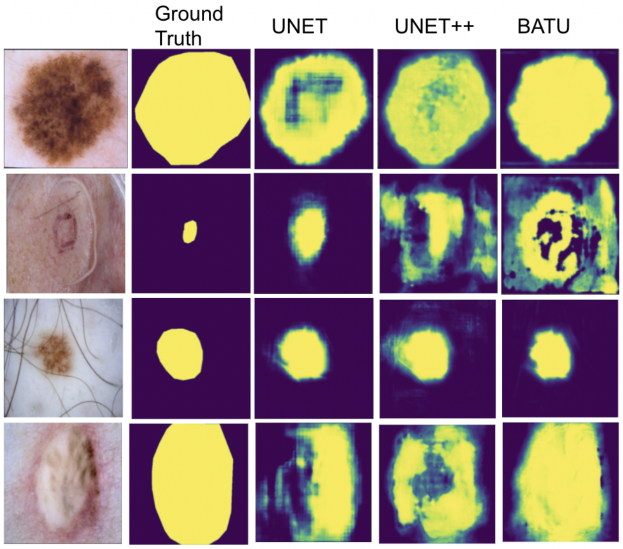

Figure 5. Results showing good cases and failure modes for different models