An Improvement of the Performance of Polar-Space-Trained DuAT Models for the Segmentation of Dermoscopic Images

Parham Izadi Ghahfarokhi*

Abstract

Malignant melanoma is a type of skin cancer, which has a very high fatality rate if not diagnosed and treated in its early stages. This has created an incentive for the design of systems for its automated diagnosis, using dermoscopic images of skin lesions. One of the steps in such systems would be the segmentation of the dermoscopic image, so that the lesion is separated from the healthy skin tissue surrounding it. In the recent years, with the prolific usage and research on neural networks and deep learning methods, many new deep-learning-based methods have emerged that aim to improve the performance and the validity of automatically generated segmentation masks. This work explores the possible areas of improvement upon earlier methods, such as the Dual Aggregate Transformer method and analysis in polar space, by using spatial encodings and novel loss functions. Furthermore, the discrepancies between the results in polar and cartesian spaces are examined. It is shown in this work that the dice score for the generated segmentation mask can increase by 0.0059 points by using polar coordinates, and by a further 0.0017 points by using the mixed loss function. Furthermore, we can see an increase of 0.0152 points by comparing the results with what the previous studies have achieved in the polar space.

Keywords: Medical Image Segmentation, Deep Learning, Neural Networks

1 Introduction

Melanoma, also known as malignant melanoma, is a type of skin cancer caused by the out-of-control growth and multiplication of the melanocyte cells that are responsible for the pigmentation of the skin. Such abnormal growth causes lesions to appear on the skin, which will first grow on the two upper layers of the skin. Since the tumor, at this stage, has not yet reached any blood vessels, the possibility of metastasis is very low, and the tumor can be removed with a surgery.[2]

In the recent years, this has created an incentive for researchers in the fields of image processing and machine learning for the development of methods which are able to detect the skin lesions caused by melanoma automatically, providing a method for alerting potential patients to possible melanoma tumors.[3]

While the conventional diagnosis of melanoma relies on a variety of visual and physical features [4, 5], not all of these features can be used by a computer model that decides the risk of melanoma based on a stationary image, therefore making the system less reliable than a trained healthcare worker, but still useful as an early alert mechanism, which can be used to reduce the capital and human cost of diagnosing melanoma.

The first step in most of these automated methods for the detection of melanoma is creating a segmentation mask, which separates the lesion tissue from the surrounding skin. As the shape and color of a lesion are important clues in its classification, a segmentation mask can be used as a clue to its shape and relative pigmentation compared to the surrounding skin tissue [6].

With the recent advances in the field of deep learning, a number of new methods have been introduced which utilize deep neural networks for the segmentation of dermoscopic images. In this work, the Dual Aggregate Transformer (DuAT) architecture [7] is used in combination with the method used by Bencevic et al. [1] to examine the possible improvements in its performance by analyzing the images in polar space. While the analysis in polar space is shown to greatly increase the dice score in the work by Bencevic et al. [1], the relation between the changes of dice score in the polar and cartesian spaces is explored in this work in more detail. Having examined these methods, the effects of adding spatial encodings to the feature maps is explored, alongside a number of loss functions, which cause improvements in the network's performance. Finally, a segmentation method is designed that shows higher performance than the previously proposed methods.

2 Previous Work

As stated above, many deep-learning-based methods are proposed for the segmentation of dermoscopic images, which mostly rely on a U-Net [8] encoder-decoder structure. A great number of improvements have been made on the original U-Net structure, such as the usage of vision-transformer-based encoders and decoders [7, 9],

*Department of Electrical and Computer Engineering, Isfahan University of Technology, parham.izadi@ec.iut.ac.ir

which are shown to have superior performance compared to classical convolutional neural networks in this application. Another method of improving upon the conventional U-Net design, is by transforming images into polar coordinates, which is shown to greatly increase the dice score of segmentation masks generated for biomedical images, namely dermoscopic images that are the focus of this work. [1]

2.1 U-Net-Based Architectures for Image Segmentation

The U-Net architecture, is a class of neural networks which are widely used for image segmentation. [8] The general structure of these networks consists of an encoder and a decoder part. In the encoder, a feature map is extracted from the image in each level, and the dimensions of the image are decreased. The decoder uses up-convolution layers to extract more high-level features and combines them with low-level features from the encoder via concatenation.

The multitude of U-Net architectures used for segmentation utilize a variety of encoders and decoders, such as the ResNet backbone [10], which can improve the process of feature extraction.

2.2 Polar Image Transformation in Biomedical Segmentation

While dermoscopic images are captured in a Cartesian coordinate system, it is possible to transform such images to other coordinate systems, including the polar coordinate system. In [1] it is shown that doing so in the pre-processing step can significantly improve the performance of the network.

This method begins by finding the mass center of the skin lesion, which is then used as the origin of the polar coordinate system. The image is then transformed into a polar representation of itself, where the i and j axes represent the radius and the angle of each pixel relative to the mass center.

This transform is shown to cause improvements in accuracy, dice score, and IoU (intersect over union) of several U-Net-based convolutional neural networks, including Res-U-Net++ and DeepLabV3+.

This approach creates the secondary problem of detecting the position of the mass center without having the ground truth mask. In [1], multiple approaches to this problem are evaluated, such as using a separate cartesian network to estimate the mass center, and using a stacked hourglass network [11] to create a mass center heatmap. The stacked hourglass approach achieved the best results in [1], and is used in this work to estimate the mass center of lesions.

2.3 Dual Aggregate Transformer Architecture

The basis for the neural network architecture used in this work is the DuAT (Dual Aggregate Transformer) architecture proposed by Tang et al. [7], which uses a transformer-based architecture for both the encoder and the decoder blocks of the network. Transformers have been the preferred architecture in language-processing and other 1-D sequence processing applications in the recent years [12], as well as a multitude of applications related to image processing, computer vision, and image generation [13].

The DuAT architecture uses the PVT (Pyramid Vision Transformer) architecture as its backbone, which contains no convolutional layers [9]. The PVT architecture operates by extracting patches from the image, which are then mixed with positional embeddings, and finally fed into a transformer encoder, which uses the patch data to generate a multi-channel output, similar to that of a convolutional neural network.

To decode the extracted features and generate the segmentation mask, the DuAT architecture uses several GLSA (Global Local Spatial Attention) blocks, and an SBA (Selective Boundary Aggregation) block. These GLSA blocks work by splitting the input features into two local and global categories, which are processed differently, and then mixed together. The SBA block mixes the local and global features with a calibration mechanism.

This architecture has shown a vast improvement over the previously proposed segmentation models, and therefore, it serves as the foundation of the neural network architecture used in this work.

2.4 Combined Loss Function

Another feature of the previously mentioned DuAT architecture is its use of a combined loss function, which uses a weighted sum the binary cross-entropy and the IoU loss of the output during the training phase. The weights assigned to each pixel are calculated according to their difference with their neighboring area. The use of combined loss functions has been shown to improve the performance and the convergence speed of the network in previous works [14, 15].

3 Method

The proposed method builds on top of the previously mentioned methods by utilizing both the polar transform and the DuAT model, while making several other changes in the form of spatial encoding, and post-processing steps.

3.1 The Segmentation Process

3.1.1 Pre-processing

The pre-processing step consists of reading the image and resizing it into a 256 × 256 image. Firstly, an estimate of the mass center is made by feeding the image to a previously trained stacked hourglass network. The polar transform is then applied to the image using the center as the origin point. The radius dimension is scaled so that the resulting polar image would fit in the same 256 × 256 dimensions.

3.1.2 Encoder

The polar image is then fed to a PVT backbone, which extracts 4 feature maps in 4 successively decreasing dimensions. This encoder is trained alongside the decoder of the network.

3.1.3 Spatial Encoding

A point of difference between the used architecture and the original DuAT architecture is that the extracted features are adjusted using a spatial encoder, which calculates a certain vector for every pixel, according to its radius value, and then adds it to the feature vector extracted by the encoder.

The positional encodings were implemented by two different methods, each of which were trained and implemented differently.

The first method was by using per-radius positional encodings, meaning that each pixel in the image would have an embedding vector added to its feature vector. The optimal values of the positional embedding vector are found during the training phase.

An alternative to this method is using polynomial functions to approximate positional encodings for each of the features in the feature vector. An n-th order polynomial encoding is shown in (1), where r is the distance of the point from the origin, a is the polynomial coefficient, and p vectors are the trainable parameters of the function.

$$\vec{a}_{(r)} = \sum_{i=0}^{n} r^i \vec{p}_i \tag{1}$$

While this approach simplifies the positional encoding vector, by doing so, it might make the model easier to train, and reduce the issue of overfitting. The second model was trained with a 5-th order polynomial positional embedding.

3.1.4 Decoder

The features extracted by the encoder are then fed to the DuAT decoder, which uses deep-supervision to generate two feature maps from the aforementioned features. The loss function is the sum of losses for both of the outputs, but the reported dice scores are calculated using only the final output.

3.1.5 Post-Processing

In the post-processing step, an erosion operation is first applied to the segmentation mask, to detach the remote parts of the segmentation mask. Then, the leftmost connected object, i.e. the object on the center of the mask, is selected, and all other objects are removed. Finally, a dilation operation is applied to the mask to undo the erosion that was applied earlier.

After the removal of the disconnected parts, the image is transformed from polar coordinates back to Cartesian coordinates. This is necessary for the calculation of the dice scores, since the polar coordinate system tends to increase the pixel density in the regions close to the origin, which generally results in a higher-than-usual dice score in this scenario.

In the last step, a closing operation is applied to the Cartesian segmentation map to close any gaps inside of it, and the result is compared with the ground-truth mask provided in the dataset.

3.2 Loss Functions

Three different loss functions were used in this work. Firstly, the combined loss function used by the DuAT model. This loss function, as implemented by the DuAT model, is shown to increase the performance of segmentation models.

Secondly, the focal loss function [16] is used, with the weighing factor of each pixel being equal to the absolute value of error. For a segmentation output h and a ground truth value y, the focal loss with a γ parameter of 1 can be calculated as seen in (2).

$$L_{focal}(h, y) = -|h - y|(y \log(1 - y) + (1 - y) \log y) \tag{2}$$

Finally, mixed loss, an alteration of the combined loss function is used, which multiplies the combined loss function by the weighing factor used above, as can be seen in (3).

$$L_{mixed}(h, y) = |h - y| L_{combined}(h, y) \tag{3}$$

3.3 Configurations

Four different models were tested, firstly the original DuAT model, secondly the DuAT model trained on polar space images, thirdly the DuAT model with positional embeddings, and finally the DuAT model with polynomial embeddings. Each of these models were trained and evaluated with the 3 different loss functions mentioned above, and the average results of the runs were reported.

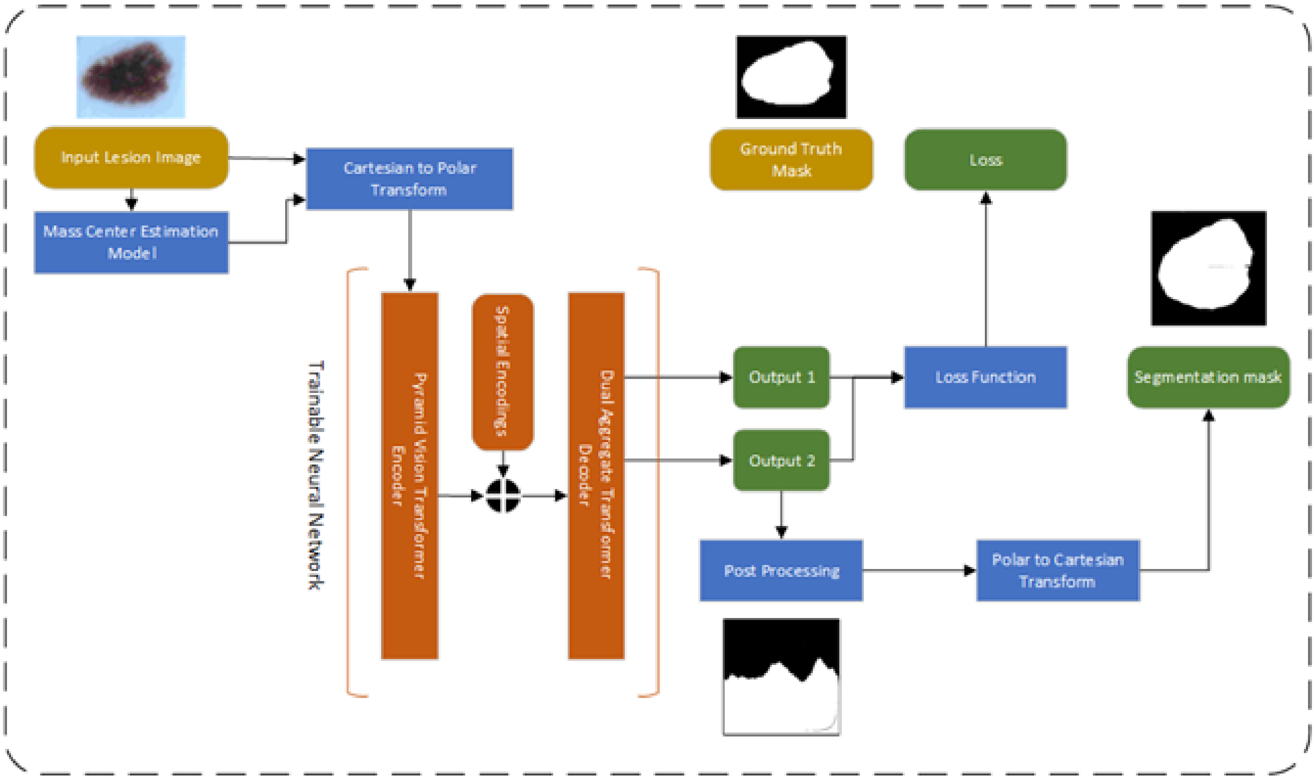
Input Lesion Image
Mass Center Estimation Model
Cartesian to Polar Transform
Trainable Neural Network
Pyramid Vision Transformer Encoder
Spatial Encodings
Dual Aggregate Transformer Decoder
Ground Truth Mask
Loss
Output 1
Output 2
Loss Function
Segmentation mask
Post Processing
Polar to Cartesian Transform

Figure 1: Block diagram of the proposed method. Note the difference in the output used for training the network and the one used as the overall segmentation mask generated by the model.

4 results

The neural network model was trained for 30 epochs on a GTX1050-Ti GPU with 4 GB of VRAM. The Adam optimizer was used with a batch size of 4, learning rate of , and an l2 regularization constant of . The ISIC-2018 dataset was used, which provided 2594 sample images of varying resolutions, paired with their ground truth segmentation masks. These images were resized to 256×256 pixels during the training and evaluation phases. While reporting the results on the test set, however, the output of the neural network were resized to 512×384 pixels and compared to the ground truth mask, in order to match the resolution used by [7]. The weights of the mass center detection network trained by Bencevic et al. were published online, and as such, they were used for the Stacked Hourglass model which estimated the mass center of the image, in order to transform it into polar space. [17]

In the beginning of each training run, the dataset was randomly separated into training, validation, and test subsets, each consisting of 80, 10, and 10 percent of the original set respectively. The model was then trained for 30 epochs on the training set, and evaluated on the validation set in the end of each epoch. The best-performing model on the validation set was chosen as the optimal output of its respective training run.

In the end of each run, the optimal model of that run was evaluated on the test dataset, and the resulting dice score reported. Each configuration of the model was evaluated for 5 runs, and the average results were reported.

For models that used polar coordinates, dice scores were measured on both the polar ground truth (similar to the method used by [1]) and the original cartesian ground truth.

4.1 Usage of Polar Coordinates

By comparing the results between the cartesian and polar models, it can be seen that the usage of polar coordinates has a positive effect on the model's performance, as seen in Table 1. However, by comparing the cartesian-space dice scores with polar-space dice scores in Table 2, we observe that evaluating the dice score in polar space causes a large increase in dice score, which is not necessarily retained after the mask is transformed back into cartesian coordinates. This is especially evident for the model trained with polynomial encodings and mixed loss function, which is revealed to perform worse than the cartesian network after its generated masks are transformed into cartesian space.

	Composite loss mean DSC	**Focal loss mean DSC**	**Mixed loss mean DSC**
Cartesian DuAT	0.8799 ± 0.0008	0.8777 ± 0.0005	0.8798 ± 0.0010
Polar DuAT	0.8858 ± 0.0004	0.8834 ± 0.0022	0.8875 ± 0.0009
Polar DuAT + Spatial Embedding	0.8858 ± 0.0015	0.8772 ± 0.0040	0.8802 ± 0.0012
Polar DuAT + Polynomial Embedding	0.8789 ± 0.0013	0.8739 ± 0.0031	0.8792 ± 0.0016

Table 1: Mean dice scores in cartesian space for every tested configuration

4.2 Positional Encoding

When the models are trained with the combined loss function, a noticeable improvement can be seen in the polar coordinate dice scores, as seen in Table 2. This trend is however reversed when examining the models with focal and mixed loss functions. In these models, adding positional encodings seems to have an adverse effect on the resulting dice score.

Additionally, by examining the dice scores shown in Table 1, we can observe that the previously observed increasing trend is not observed in cartesian space. On the contrary, a general decrease is seen in most of the models' performances. This shows that spatial embeddings implemented in this form fail to improve the models' performance in a meaningful way.

We can also infer from this observation that improvements in polar coordinate metrics do not necessarily correlate with improvements in cartesian space, and it is necessary to make comparisons in the cartesian space, if one is to determine if a certain change in the model has positive or negative effects.

4.3 Loss Functions

By examining the cartesian dice scores (Table 1), it can be seen that the focal loss function fails to improve the results achieved by the composite loss function in every configuration.

The mixed loss function, however, in the polar-trained model, and the model using polynomial spatial encodings, manages to increase the resulting dice score, giving the best result when paired with the polar-trained model without spatial encoding.

	Composite loss mean DSC	**Focal loss mean DSC**	**Mixed loss mean DSC**
Polar DuAT	0.9304 ± 0.0002	0.9380 ± 0.0013	0.9405 ± 0.0004
Polar DuAT + Spatial Embedding	0.9391 ± 0.0009	0.9334 ± 0.0040	0.9358 ± 0.0008
Polar DuAT + Polynomial Embedding	0.9352 ± 0.0009	0.9320 ± 0.0020	0.9351 ± 0.0011

Table 2: Mean dice scores in polar space for polar-trained networks

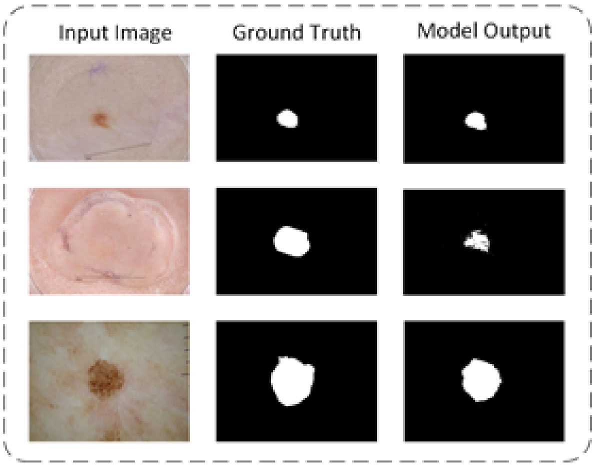

Figure 2: Images from the dataset, alongside the provided ground truth mask and the generated segmentation output

4.4 Computational Complexity and Parameters

The PVT backbone itself, has a lower number of parameters compared to the conventional convolutional backbones such as ResNet, as it can be seen in Table 3. While the addition of the spatial embeddings increases the number of parameters, this increase is negligible compared to the total number of parameters.

Similarly, the training and inference time for all models is within the same range, with the PVT models performing slightly better than the ResNet model. The extra processing necessary for transforming the image between the polar and cartesian spaces increases the training and inference times slightly for the polar models, but this increase remains negligible compared to the base value. The evaluation time was evaluated on the CPU as well, which while slower than the GPU, still can output an image within less than half a second.

	$N_{\text{parameters}}$	T_{training}	$T_{\text{evaluation}}$
DuAT (ResNet Backbone)	$27.69M$	$418ms$	$220ms$
DuAT	$25.07M$	$406ms$	$197ms$
DuAT (Cartesian Space)	-	$405ms$	$212ms$
DuAT (CPU)	-	-	$370ms$
DuAT + Spatial Embeddings	$25.43M$	$415ms$	$212ms$
DuAT + Polynomial Embeddings	$25.36M$	$418ms$	$204ms$

Table 3: A summary of the number of parameters and mean per-image training and evaluation time on one epoch

5 Conclusion

It can be concluded from the aforementioned results, that when training in polar spaces, positional encodings are unlikely to improve the results. While improvements are seen when examining the results in polar coordinates, these improvements are made by biasing the results towards the pixels whose importances are more emphasized by the polar coordinates, as the area close to the origin is more densely sampled in a cartesian-to-polar transform.

Similarly, the large gap seen between the dice scores of polar and cartesian spaces suggests that evaluation in polar coordinates can be misleading in many cases, and does not accurately describe improvement or deterioration in a method's performance.

The improvement seen by the usage of mixed loss function, that is consistent in both polar and cartesian spaces, suggests that this loss function can be more effective in training segmentation models than the composite loss used by [7], and as such, its performance in other similar tasks might be worth examination.

The final proposed method of this work, consists of a DuAT neural network, which is trained on polar-space images, uses a mixed loss function, and applies the post-processing steps described in the method section on its output. A comparison with the polar and cartesian-space methods can be seen in Table 4. While the results provided by [1] were evaluated in polar space, a very significant improvement can still be seen by comparing their provided metrics with that of the DuAT-based models.

	Dice score	**mIoU**
UNet	0.9234	0.8699
ResUNet++	0.9253	0.8743
DeepLabV3+	0.9235	0.8721
Our Method	**0.9405**	**0.8874**

Table 4: Comparison of polar-space metrics among our proposed method and methods investigated by [1]

References

[1] M. Benčević, I. Galić, M. Habijan, and D. Babin, "Training on polar image transformations improves biomedical image segmentation" *IEEE access*, vol. 9, pp. 133365-133375, 2021.

[2] Cancer.net, "Melanoma: statistics." *cancer.net/cancer-types/melanoma/statistics* (accessed 1/8/2023, 2023).

[3] A. Adegun and S. Viriri, ""Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art," *Artificial Intelligence Review,* vol. 54, no. 2, pp. 811-841, 2021/02/01 2021, doi: 10.1007/s10462-020-09865-y.

[4] F. Nachbar *et al.*, "The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions," (in eng) *J Am Acad Dermatol,* vol. 30, no. 4, pp. 551-9, Apr 1994, doi: 10.1016/s0190-9622(94)70061-3.

[5] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "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,* vol. 134, no. 12, pp. 1563-70, Dec 1998, doi:10.1001/archderm.134.12.1563.

[6] A. A. Adeyinka and S. Viriri, "Skin lesion images segmentation: A survey of the state-of-the-art," *Mining Intelligence and Knowledge Exploration: 6th International Conference,* MIKE 2018, Cluj-Napoca, Romania, December 20–22, 2018, Proceedings 6, 2018: Springer, pp. 321-330.

[7] F. Tang *et al.*, "DuAT: Dual-aggregation transformer network for medical image segmentation," *Chinese Conference on Pattern Recognition and Computer Vision (PRCV),* 2023: Springer, pp. 343-356.

[8] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access,* vol. 9, pp. 82031-82057, 2021.

[9] W. Wang *et al.,* "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *Proceedings of the IEEE/CVF international conference on computer vision,* 2021, pp. 568-578.

[10] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks* vol. 121, pp. 74-87, 2020.

[11] A. Newell, K. Yang, and J. Deng, “Stacked Hourglass Networks for Human Pose Estimation,“ Cham, 2016: Springer International Publishing, in Computer Vision – ECCV 2016, pp. 483-499.

[12] N. Patwardhan, S. Marrone, and C. Sansone, “Transformers in the real world: A survey on nlp applications,“ *Information* vol. 14, no. 4, p. 242, 2023.

[13] K. Han *et al.*, “A survey on visual transformer,“ *arXiv preprint arXiv:2012.12556,* 2020.

[14] J. Wei, S. Wang, and Q. Huang, “F3Net: fusion, feedback and focus for salient object detection,“ *Proceedings of the AAAI conference on artificial intelligence,* 2020, vol. 34, no. 07, pp. 12321-12328.

[15] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, “Basnet: Boundary-aware salient object detection,“ *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 2019, pp. 7479-7489.

[16] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, “Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation,“ *Computerized Medical Imaging and Graphics,* vol. 95, p. 102026, 2022/01/01/ 2022, doi: doi.org/10.1016/j.compmedimag.2021.102026.

[17] M. Benčević, “GitHub - marinbenc/medical-polar-training.“ https://github.com/marinbenc/medical-polar-training (accessed Jul 2024).