
TPNet: Transformer Feature Pyramid Network for Diabetic Retinopathy Multi-lesion Segmentation

Hongkun Gong Yitian Huang

Abstract

1. Introduction

Diabetic Retinopathy (DR) is the leading cause of blindness among working-age adults, a prevalent eye condition experienced by diabetic patients (CDC, 2022). Given that the incidence of diabetes is projected to surge by over 54% from 2015 to 2030 (Rowley et al., 2017), and with no cure for DR available, there's an increasing need for enhanced early detection methods, which enable timely interventions, potentially preventing or mitigating vision loss. Although a dilated eye exam serves as a diagnostic tool for DR, its interpretation can be challenging and prone to human error. To address this, leveraging Semantic Image Segmentation for eye lesion detection in exam images can enable physicians to more precisely discern the presence and progression stages of DR.

Deep Convolutional Neural Networks (CNNs) have emerged as fundamental tools in machine learning, particularly in the domains of computer vision and medical image processing. Over the years, various iterations of CNN architectures have been introduced. One seminal model in this domain is the U-Net, which served as a foundational framework for medical image segmentation (Ronneberger et al., 2015). Although the U-Net architecture pioneered image segmentation, newer models have since outperformed it. Recent research, including models like TransUnet, SwinUnet, and others, has sought to build upon the U-Net structure, aiming to enhance segmentation capabilities by integrating Transformers (Chen et al., 2021; Cao et al., 2022). However, as pointed out by Huang et al., many of these endeavors have primarily focused on network design, potentially overlooking critical pathological connections in medical image segmentation. (2022). Inspired by their work, we propose the TransPyramid model aiming to integrate pathologically related features with image features from a TransUnet backbone through Transformer's attention mechanism and Feature Pyramid Network's semantically rich structure (Vaswani et al., 2017; Lin et al., 2017).

2. Related Work

2.1. Diabetic Retinopathy Pathological Analysis

Segmenting DR (Diabetic Retinopathy) lesions is challenging due to the intricate nature of the disease. Different visual markers typically correspond to various stages of DR. These stages can be categorized as mild nonproliferative, moderate nonproliferative, severe nonproliferative, and proliferative diabetic retinopathy. (Alghadyan, 2011)

In the earliest stage of DR, microaneurysms (MAs) emerge in the retina's blood vessels. MAs manifest as small circular dots with a width smaller than that of a blood vessel. When MAs rupture or when capillaries leak, Retinal Hemorrhages (H) develop. These hemorrhages are typically larger than MAs and exhibit irregular edges and coloration. The leakage from MAs also leads to the accumulation of lipoproteins and other proteins in the eye, forming Hard Exudates (HE). Vascular occlusion, on the other hand, results in the appearance of white lesions known as soft exudates, which often surround hard exudates (Huang et al., 2022).

In the study by Huang et al. (Huang et al., 2022), it was observed that MAs are typically situated among the capillaries, while SEs predominantly appear near the upper and lower arteries' trunks. Although fundus tissue can often complicate the segmentation of lesions, the research team discovered that recognizing patterns in fundus tissue can aid in differentiation.

2.2. Feature Pyramid Network

Fully Convolutional Networks (FCNs) have been extensively employed for semantic image segmentation. However, they frequently suffer from reduced spatial resolution in their deeper layers, leading to suboptimal predictions. A team led by Yuan Yuan tackled this issue with FCNs by utilizing feature maps generated by the multi-scale and pyramidal hierarchy of the Feature Pyramid Network. This approach facilitates a more comprehensive and pixel-level classification of the input image, resulting in a detailed label map for semantic image segmentation (Yuan et al., 2019).

2.3. TransUnet

Unet performs well in image segmentation, especially due to its precise localization in the expansive path(Ronneberger et al., 2015). However, Unet lacks the ability to capture the relation between distant features. TransUnet solves this by integrating Transformer’s self-attention mechanism in the encoder path of Unet(Chen et al., 2021). Instead of only performing convolution operations in the encoder, TransUnet first uses CNN to extract features from the image. Feature map from CNN is then fed to a Transformer to capture the global context through Transformer’s global self-attention mechanism. By doing so, TransUnet effectively addresses the limitation of U-Net in capturing distant feature relations, making it a powerful model for medical image segmentation.

2.4. RTNet

Ever since the introduction of the Vision Transformer, researchers have been trying to combine the model with other image processing models to leverage the benefits of different models. Researchers led by Shiqi Huang built Relation Transformer Net (RTnet) with the inspiration of Transformer’s performance in medical images and Deep Neural Network’s success in DR Lesion Segmentation (Huang et al., 2022). Their model precompute the preprocessed medical image with UNet and feed them into a global Transformer block and relation Transformer block for further process. They achieved significant results, outperforming other state-of-the-art works.

3. Dataset and Evaluation

3.1. Datasets

To maintain the consistency of the experiments, we adopted the same datasets Huang, et al. used in their research(2022).

The first dataset is IDRiD Dataset(Porwal et al., 2018). It consists of three parts and each part has a certain number of images all with a size of 4288*2848. Part A contains 81 original fundus and ground truth lesion label images separated into 54 train and 27 test images. Part B contains 516 original fundus images and ground truth labels for the severity grade of Diabetic Retinopathy and Diabetic Macular Edema. Part C contains 516 original fundus images and ground truth labels for Optic Disc Center Location and Fovea Center Location.

The second dataset is DDR Dataset(Li et al., 2019). It contains 13673 fundus images that are already preprocessed to remove the black background. These images are from 147 hospitals spread across 23 provinces in China and each image is classified by the severity grade of Diabetic Retinopathy.

3.2. Evaluation

We will use area under precision-recall curve(AUPRC) and area under receiver operating characteristic curve(AUROC) as the evaluation matrices. As previous research widely adopted these two matrices for medical image segmentation evaluation, using them helps compare our work with models others proposed. The focus will lean mostly toward AUPRC. In the context of pixel-wise segmentation for medical images, the positive labels, the lesions, are much fewer than the negative labels. Since AUPRC is generally a better matrix than AUROC if the class labels are unbalanced(Davis & Goadrich, 2006), focusing on AUPRC better reflects the actual performance and effectiveness in finding the actual lesions of our model.

4. Methods

4.1. Baseline

The baseline of our task is a Unet architecture(Ronneberger et al., 2015). It has been shown by many previous researches that Unet does a great job in medical image segmentation. In Unet, an image is fed as the original input to the model. It then goes through a U-shaped encoder and decoder structure. The encoder has multiple layers of 3*3 convolution followed by ReLU and 2*2 max pooling. The decoder has the same number of layers as the encoder and each layer of decoder performs 3*3 convolution and ReLU followed by a 2*2 up convolution to restore the size of the image. Feature maps produced in each layer of encoder are cropped and then concatenated with the feature maps in each decoder layer. The final output is a pixel-wise feature map with channel number equal to the number of classes.

4.2. Architecture Overview

Our model tries to exploit the accuracy of TransUnet in image segmentation and the pathological connection of lesions and retinal blood vessels. The model consists of two modules and an additional input vessel feature from a pre-trained model.

The pre-trained model we pick is a TransUnet neural network(Chen et al., 2021). It will be trained on a different dataset to make pixel-wise label predictions for retinal blood vessels. The model is, then, applied to our training data to generate pseudo-vessel labels as input features to our model.

The first module is also a TransUnet for lesion features(Chen et al., 2021). The input of this module is the preprocessed training image and the output is a lesion feature map that will be fed to the next module.

The second module is the TransPyramid that performs feature pyramid resizes on both lesion features and retinal blood vessel features(Lin et al., 2017). Transformer self-

attention and cross-attention are, then, applied to each layer of the lesion feature pyramid and retinal blood vessel feature pyramid. The outputs from each layer's Transformer are then upsampled to the original spacial feature size and normalized to get the final lesion mask.

4.3. TransPyramid

The TransPyramid we propose integrates Feature Pyramid Network (FPN) with a Multi-Head Transformer for lesion features. The method begins with adapting pre-computed lesion features into a spatial format suitable for FPN processing. These reshaped features are then used to construct an FPN, creating multi-scale feature maps either by upsampling or downsampling. Both vessel and lesion features computed previously go through this process. For each level of the FPN, the spatial dimensions are flattened and encoded with positional information, transforming them into sequences can be processed by a multi-head attention Transformer.

There are two primary heads within the multi-head attention Transformer: the self-attention head and the cross-attention head. The self-attention head focuses on the lesion features, computing relationships and dependencies within each level of the FPN. On the other hand, the cross-attention head is designed to contextualize each FPN level in relation to the vessel features, offering a more holistic understanding of the pathological environment.

After the Transformer's processing, the outputs are reshaped back into spatial feature maps, which are subsequently upsampled to target resolution. Once all features map are upsampled, feature maps from cross-attention head and self-attention head can undergo feature fusion by concatenation and go through final downstream processing of FPN to obtain Lesion masks.

5. Experiments

6. Discussion

7. Conclusion

8. Style Guide

References

- Alghadyan, A. A. Diabetic retinopathy – an update. *Saudi Journal of Ophthalmology*, 25(2):99–111, 2011. doi: <https://doi.org/10.1016/j.sjopt.2011.01.009>.
- Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.
- CDC. Diabetes and vision loss, 2022. URL <https://www.cdc.gov/diabetes/managing/diabetes-vision-loss.html>.
- Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 233–240. Association for Computing Machinery, 2006. doi: 10.1145/1143844.1143874.
- Huang, S., Li, J., Xiao, Y., Shen, N., and Xu, T. Rtnet: relation transformer network for diabetic retinopathy multi-lesion segmentation. *IEEE Transactions on Medical Imaging*, 41(6):1596–1607, 2022.
- Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., and Kang, H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019. doi: <https://doi.org/10.1016/j.ins.2019.06.011>.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.
- Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabudhe, V., and Meriaudeau, F. Indian diabetic retinopathy image dataset (idrid), 2018. URL <https://dx.doi.org/10.21227/H25W98>.
- Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.
- Rowley, W. R., Bezold, C., Arikian, Y., Byrne, E., and Krohe, S. Diabetes 2030: Insights from yesterday, today, and future trends. *Population Health Management*, 20(1):6–12, 2017. doi: 10.1089/pop.2015.0181. PMID: 27124621.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Yuan, Y., Fang, J., Lu, X., and Feng, Y. Spatial structure preserving feature pyramid network for semantic image segmentation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(3), 2019. ISSN 1551-6857. doi: 10.1145/3321512.

A. Appendix