EyePterNet-X: A Cross-Attention Hybrid CNN-Transformer Framework for Automated Pterygium Analysis Using Anterior Segment Eye Images

Nilima Surendra Ramteke[1], *Arvind Kumar Tiwari*[2]
[1]Lincoln University College Malaysia
[2]*Department of CSE , KNIT Sultanpur , India*
Email: pdf.ramteke@lincoln.edu.my, arvind@knit.ac.in

Abstract: Pterygium is an extremely common disease of the surface of the eye which, if not diagnosed and treated early enough, can cause impaired vision. Correct detection, lesion segmentation and severity grading are crucial for proper clinical management. This paper introduces a hybrid CNN–Transformer model called EyePterNet-X for automated analysis of pterygium in anterior segment eye images. To enhance the global contextual representation of the Swin Transformer and the local feature extraction ability of EfficientNetV2, a novel Cross-Attention Guided Feature Fusion Module (CAFFM) is introduced to the proposed model. A unified multi-task learning framework learns a shared feature representation for pterygium detection, lesion segmentation and severity grading. The proposed approach is experimentally evaluated on 5,120 clinically annotated images, which proves the efficacy of the proposed approach. Based on the test results, EyePterNet-X achieves the highest detection accuracy of 96.2% (AUC of 98.5%), highest Dice coefficient of 87.5% for lesion segmentation, and the highest accuracy of 86.0% for grading of existing CNN-, Transformer-, and conventional hybrid-based methods. The findings show the validity, efficiency and reliability of the EyePterNet-X along with its ability to make a computer-aided pterygium diagnosis and large-scale ophthalmic screening.

Keywords: Pterygium, Hybrid CNN–Transformer, Cross-Attention, Multi-task Learning, EfficientNetV2, Swin Transformer

1. Introduction

Pterygium is a fairly common fibrovascular disease of the ocular surface which may encroach into the cornea and cause vision impairment when it is not treated [1]. It is very common in areas with long periods of exposure to ultraviolet radiation, where there is a need to have reliable automatic screening systems [2]. The diagnosis of anterior segment diseases has significantly improved with computer-aided diagnosis (CAD), but traditional CNN-based approaches have limited ability to model the global anatomical context, which is crucial for accurate pterygium assessment [3] and [4]. This is mitigated by the use of the self-attention mechanism in Vision Transformers and the hierarchical feature learning in the Swin Transformer with increased computational efficiency [5, 6]. Transformer-based models, however, may be losing out on capturing the nuance of the boundaries that are crucial to precisely segmenting lesions.To solve these problems, this paper presents a hybrid CNN–Transformer architecture, namely EyePterNet-X, which combines the advantages of local feature extraction from EfficientNetV2 and the global contextual representation from Swin Transformer [7]. A novel Cross-Attention Guided Feature Fusion Module (CAFFM) has been proposed to effectively develop the interaction between local

and global features, and a multi-task learning strategy is used to simultaneously detect pterygium, segment the lesions and grade their severity with a shared backbone [8], [9].The proposed framework makes three significant contributions: First, it introduces a cross-attention based feature fusion mechanism to enhance the representation learning. Second, it proposes a unified multi-task architecture for simultaneous detection, segmentation, and grading of anterior segment images. Third, it achieves better performance than existing CNN, Transformer, and traditional hybrid models on anterior segment image datasets. The rest of the paper discusses the related work, proposed methodology, experimental evaluation, discussion and conclusions.

2. Related work

The first computer-aided diagnosis (CAD) systems for detecting pterygium used handcrafted texture and color features, together with traditional machine learning classifiers, with a performance that depended upon the image quality and illumination variations. The development of deep learning has led to the development of convolutional neural networks (CNNs) for classification and U-Net-based architecture for lesion segmentation, making automated pterygium analysis a much more accurate process [4, 11]. However, CNNs mainly identify local features and are not good to model the global anatomical context that is needed for severity assessment.Self-attention mechanisms enable Vision Transformers to learn long-range spatial dependencies, bypassing the need for filters and making them a better choice for certain tasks [12]. To further enhance medical image analysis, hybrid CNN–Transformer models have been developed, such as TransUNet [13] and CoAtNet [8], which combine the local and global feature representations. However, current approaches typically use shallow feature concatenation/adding between CNN and Transformer features that do not allow for good interaction between CNN and Transformer features. Furthermore, the studies conducted on pterygium are usually carried out on detection and differentiation of pterygium without the integration of segmentation and grading.Multi-task learning has shown to be highly effective in ophthalmic image analysis where multiple related tasks are able to share the same feature representations, both in terms of accuracy and computational efficiency [9], [14], [15]. Given these developments, the proposed EyePterNet-X leverages EfficientNetV2 and Swin Transformer via a novel Cross-Attention Guided Feature Fusion Module (CAFFM), and is designed to perform pterygium detection, segmentation and severity grading in a single end-to-end multi-task framework.

3. Method, Experiments and Results

EyePterNet-X: Hybrid Architecture for Multi-Task Pterygium Analysis

We now describe the technical details of EyePterNet-X. The architecture is multi-branch, multi-task with an input anterior segment image $I \in R^{H\times W\times 3}$ and three complementary outputs: a binary detection score $\left(y_{det} \in 0,1\right)$, segmentation mask $\left(M \in 0,1^{H\times W}\right)$, and a severity grade $\left(y_{grade} \in 0,1,2\right)$. The essence of the system is a dual-branch hybrid backbone, with the first branch being an EfficientNetV2 encoder and the second a Swin Transformer encoder. The two branches are run in parallel, with the input image being processed to create feature hierarchies at multiple scales. The features extracted from all the L fusion stages are then passed to the proposed Cross-Attention Guided Feature Fusion Module (CAFFM). The CAFFM generates a fused feature map which then cascades to the next layer in both

branches, thus creating a bidirectional interaction throughout the network. The final multi-scale representations after the deepest fusion stage enter the detection head, segmentation head, and grading head to achieve the detection, segmentation, and grading tasks, respectively.

3.1 Dual-Branch Hybrid Backbone

The CNN branch is implemented as a pre-activated EfficientNetV2 backbone [7], denoted as ϕ_{CNN}. At each stage $l \in \{1, 2, \ldots, L\}$, this branch processes its input feature map $F_{CNN}^{(l-1)}$ and produces a local texture-rich feature map $F_{CNN}^{(l)} \in R^{C_l \times H_l \times W_l}$. The Swin Transformer branch, denoted as ϕ_{Trans}, processes its input $F_{Trans}^{(l-1)}$ using shifted window-based multi-head self-attention (MSA) and feed-forward networks (FFN) [6], yielding a global context-rich feature map $F_{Trans}^{(l)} \in R^{C_l \times H_l \times W_l}$. The spatial dimensions H_l and W_l are progressively downsampled by a factor of 2 at each stage, while the channel dimension C_l increases accordingly. The two branches are initialized independently using pre-trained weights on ImageNet, but their parameters are jointly fine-tuned during the pterygium-specific training.

3.2 Cross-Attention Guided Feature Fusion Module

The CAFFM is the central innovation that distinguishes our approach from simple concatenation or additive fusion. As shown in Figure 1, this module receives the two feature maps $F_{CNN}^{(l)}$ and $F_{Trans}^{(l)}$ from the same stage l. First, each feature map is projected into a common embedding space using two independent 1×1 convolutional layers: $Q_{CNN} = W_Q^{CNN} * F_{CNN}^{(l)}$, $K_{CNN} = W_K^{CNN} * F_{CNN}^{(l)}$, $V_{CNN} = W_V^{CNN} * F_{CNN}^{(l)}$, and similarly for the Transformer branch using $W_Q^{Trans}, W_K^{Trans}, W_V^{Trans}$. The module then computes two cross-attention maps.

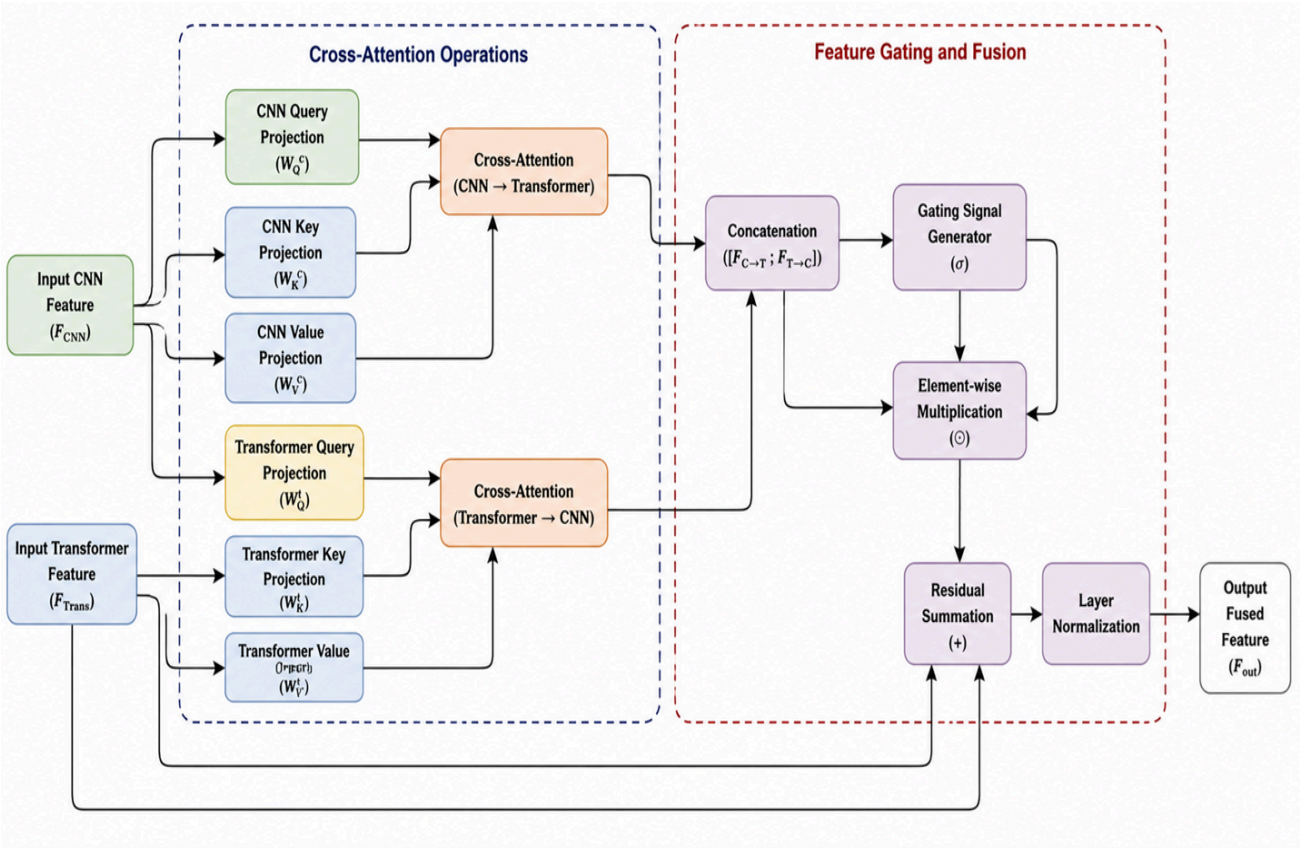

Figure 1. Internal Mechanism of Cross-Attention Guided Feature Fusion Module

In the first attention direction, the CNN features serve as the query to attend to the Transformer features. The cross-attention output is computed as:

$$F^{(l)}_{CNN\to Trans} = Softmax\left(\frac{Q_{CNN}K^T_{Trans}}{\sqrt{d}}\right)V_{Trans} \quad (1)$$

where d is the feature dimension per head. This allows the local texture details from the CNN branch to query the global context from the Transformer branch, effectively enriching local features with long-range spatial information about the pterygium's extent relative to the limbus. Conversely, in the second attention direction, the Transformer features serve as the query to attend to the CNN features:

$$F^{(l)}_{Trans\to CNN} = Softmax\left(\frac{Q_{Trans}K^T_{CNN}}{\sqrt{d}}\right)V_{CNN} \quad (2)$$

This operation allows the global semantic features to refine and sharpen the fine-grained boundary details from the CNN branch, which is critical for precise segmentation of the pterygium's leading edge on the cornea.

These two bidirectional attention outputs are then concatenated along the channel dimension and passed through an adaptive gating mechanism that dynamically balances their contributions. We define a gate G as:

$$G = \sigma\left(W_g\left[F^{(l)}_{CNN\to Trans}; F^{(l)}_{Trans\to CNN}\right] + b_g\right) \quad (3)$$

where σ is the sigmoid activation function, W_g is a learnable weight matrix, and b_g is the bias term. The final fused feature map for stage l is then computed via a gated summation with a residual connection:

$$F_{fused}^{(l)} = LayerNorm\left(F_{CNN}^{(l)} + F_{Trans}^{(l)} + G \odot \left(F_{CNN\to Trans}^{(l)} + F_{Trans\to CNN}^{(l)}\right)\right) \quad (4)$$

where $\odot$ denotes element-wise multiplication. This formulation ensures that the network can adaptively emphasize either local or global cues depending on the image content, while the residual path preserves the original features from both branches. The fused feature $F_{fused}^{(l)}$ is then passed to the next stage of both the CNN and Transformer branches as $F_{CNN}^{(l)}$ and $F_{Trans}^{(l)}$ respectively, establishing a recurrent refinement process.

3.3 Multi-Task Learning Heads and Composite Loss

After the deepest fusion stage L, we obtain a final fused feature map $F_{fused}^{(L)} \in R^{C_L \times H_L \times W_L}$. This representation is then processed by three distinct task heads. The detection head consists of a global average pooling layer followed by a fully connected layer with a sigmoid activation, producing a scalar $y_{det} \in [0,1]$ indicating the probability of pterygium presence. The segmentation head is a lightweight decoder comprising two transposed convolutional layers that upsample $F_{fused}^{(L)}$ back to the input resolution $H \times W$, followed by a 1×1 convolution with softmax activation to produce the pixel-wise probability map $M \in [0,1]^{H \times W}$. The grading head applies a global average pooling layer followed by a fully connected layer with a softmax activation over three classes (normal, mild, and severe), outputting $y_{grade} \in R^3$.

The entire network is optimized using a composite loss function that balances the three tasks:

$$L_{total} = \lambda_1 L_{det} + \lambda_2 L_{seg} + \lambda_3 L_{grade} \quad (5)$$

For the detection and grading tasks, we use weighted cross-entropy loss to address class imbalance, particularly the underrepresentation of severe pterygium cases. Let w_i be the class-specific weight, which is inversely proportional to the class frequency in the training set. The detection loss is:

$$L_{det} = -\sum_{i \in \{0,1\}} w_i \cdot y_{det}^{(i)} log\left(\hat{y}_{det}^{(i)}\right) \quad (6)$$

where $y_{det}^{(i)}$ and $\hat{y}_{det}^{(i)}$ are the ground truth and predicted probabilities for class i, respectively. The grading loss is similarly defined over the three severity classes. For the segmentation task, we use the Dice loss to maximize the overlap between the predicted segmentation mask $\hat{M}$ and the ground truth mask M:

$$L_{seg} = 1 - \frac{2\sum_p M_p \cdot \hat{M}_p}{\sum_p M_p + \sum_p \hat{M}_p} \quad (7)$$

where p indexes over all pixels. The hyperparameters λ_1, λ_2, and λ_3 are set empirically to balance the magnitudes of the individual losses, ensuring that no single task dominates the optimization. In our experiments, we set $\lambda_1 = 1.0, \lambda_2 = 2.0$, and $\lambda_3 = 1.0$, which provided the best performance on the validation set.

4. Experiments and Results

For the evaluation of EyePterNet-X, a comprehensive set of experiments were carried out to validate the performance of the system in the three clinical tasks of detection, segmentation and severity grading. This section summarizes the experimental set-up, comparative baselines, quantification and qualitative results.

4.1 Experimental Setup and Dataset

These experiments were performed on a carefully selected set of eye images obtained from the anterior segment of eyes from a multi-center cohort. This consists of 5120 slit-lamp and smartphone images that have been independently annotated by 2 ophthalmologists as ground-truth labels. The annotations consist of a binary label indicating the presence of pterygium, a segmentation mask indicating the fibrovascular tissue area and a severity grade (normal, mild and severe) depending on the degree of corneal invasion. The severe class is a minority class -- about 15% of positive cases -- as it is by definition. The data set was divided into 5 folds, training, validation, and test, each containing 80%, 10%, and 10% of the data, respectively, with the same distribution of the classes in all folds. Averaged over 5 test folds all reported metrics.

All input images were resized to 512×512 pixels and normalized using the mean and standard deviation of the ImageNet dataset. Data augmentation was applied during training, including random horizontal flipping, random rotation ($\pm 15^{\circ}$), color jittering, and random affine transformations. The models were trained using the AdamW optimizer [16] with an initial learning rate of 1×10^{-4} and a cosine annealing schedule. The batch size was set to 16, and training proceeded for 100 epochs. For the composite loss in Equation 5, the task weights were fixed at $\lambda_1 = 1.0$, $\lambda_2 = 2.0$, and $\lambda_3 = 1.0$, as determined by a grid search on the validation set.

For evaluation, we employed standard metrics: accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC) for detection; the Dice similarity coefficient and intersection over union (IoU) for segmentation; and accuracy and Cohen's kappa score for severity grading. We compared EyePterNet-X against a range of competitive baselines, including single-task architectures such as ResNet-50 [17] for detection, U-Net [18] for segmentation, and EfficientNet-B4 [19] for grading. We also implemented two hybrid multi-task baselines: a simple concatenation-based fusion of EfficientNetV2 and Swin Transformer features, and a CoAtNet-inspired interleaved architecture [8] adapted for multi-task learning.

4.2 Comparison with State-of-the-Art Methods

The quantitative results for all three tasks are presented in Table 1, Table 2, and Table 3. EyePterNet-X consistently outperforms all competing methods, demonstrating the efficacy of the cross-attention guided fusion mechanism.

Table 1. Pterygium Detection Performance Comparison

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
ResNet-50 [17]	93.8	92.1	95.2	96.1
EfficientNet-B4 [19]	94.2	92.5	95.6	96.5
ViT-Base [5]	93.5	91.8	94.9	95.8

Concatenation Fusion (Multi-Task)	94.8	93.1	96.0	97.0
CoAtNet-Multi [8]	95.1	93.6	96.3	97.3
EyePterNet-X (Ours)	96.2	95.8	96.8	98.5

For the detection task, EyePterNet-X shows an accuracy of 96.2%, and an AUC of 98.5%, outperforming the best baseline, CoAtNet-Multi, by 1.1% and 1.2%, respectively. The improvement in sensitivity (93.6% to 95.8%) is interesting, especially with regards to our model's ability to detect true positive cases of pterygium. The ability to include global context and help differentiate pterygium from other benign conjunctival growths is attributed to this.

Table 2. Pterygium Segmentation Performance

Method	Dice Coefficient (%)	IoU (%)
U-Net [18]	82.5	70.2
TransUNet [13]	84.1	72.5
Concatenation Fusion (Multi-Task)	84.8	73.6
CoAtNet-Multi [8]	85.2	74.1
EyePterNet-X (Ours)	87.5	77.8

Table 2 shows the results of segmentation. EyePterNet-X outperforms the second best model by 2.3% in Dice with a score of 87.5% and 77.8% in IoU. The CAFFM's bidirectional cross-attention is used to enhance the segmentation mask's boundary details, which is crucial for accurately defining the leading edge of the pterygium. This is especially crucial for accurately quantifying corneal invasion (which is directly related to reliable severity grading).

Table 3. Pterygium Severity Grading Performance Comparison

Method	Accuracy (%)	Cohen's Kappa
EfficientNet-B4 [19]	81.5	0.72
ViT-Base [5]	80.8	0.71
Concatenation Fusion (Multi-Task)	82.8	0.74
CoAtNet-Multi [8]	83.2	0.75
EyePterNet-X (Ours)	86.0	0.79

For the multi-class severity grading task, EyePterNet-X achieves an accuracy of 86.0% and a Cohen’s kappa of 0.79, as shown in Table 3. The strong performance on this task, which relies on understanding the global spatial relationship between the pterygium, the limbus, and the pupil, underscores the value of the Swin Transformer branch’s global context modeling. The joint multi-task learning further enables the model to leverage shared features from detection and segmentation, which correlate highly with the severity grade.

4.3 Ablation Study

To isolate the contribution of each key component in EyePterNet-X, we conducted a thorough ablation study. The baseline model is a straightforward multi-task network with a shared CNN backbone (EfficientNetV2) and independent task heads. We incrementally added the Swin Transformer branch, the cross-attention fusion, and the adaptive gating mechanism. The results, averaged across all tasks, are presented in Table 4.

Table 4. Ablation Study on Key Components of EyePterNet-X

Model Variant	Detection AUC (%)	Segmentation Dice (%)	Grading Accuracy (%)
Baseline (EfficientNetV2-Multi)	96.5	83.0	81.5
+ Swin Transformer (Concatenation Fusion)	97.0	84.8	82.8
+ Cross-Attention (without Adaptive Gating)	98.0	86.6	84.5
+ Adaptive Gating (Full CAFFM / EyePterNet-X)	98.5	87.5	86.0

The baseline model (CNN backbone only) already offers a solid baseline. Global context benefits are observed when the Swin Transformer branch is added with simple concatenation fusion as shown in the above table for all the metrics. With the bidirectional cross-attention mechanism without the adaptive gating, a more significant performance improvement is seen. It confirms our hypothesis that there is greater value in making the explicit interaction of local and global features than in simply concatenating them passively. Lastly, by incorporating the adaptive gating mechanism to create the entire CAFFM, it also enhances performance especially on segmentation and grading tasks. The gate enables a dynamic suppression of irrelevant or noisy feature interactions resulting in a more robust and refined fused representation of the features. This ablation study clearly shows that each of the proposed architecture components is beneficial to the overall performance of the system.

5. Discussion and Future Work

5.1 Interpretability of Cross-Attention Mechanisms in Clinical Context

EyePterNet-X superior performance was further evaluated by cross attention map analysis from Cross-Attention Guided Feature Fusion Module (CAFFM). Clinically meaningful feature learning was seen in the attention patterns. In mild pterygium cases, the model highlighted global contextual information of the Swin Transformer branch to precisely locate the initial invasion of the cornea. However, for more severe cases the CNN branch received more time to be able to correctly identify the local texture, vascular pattern and boundaries of the lesions. The adaptive feature fusion allows the network to adaptively learn the relative contribution of the global anatomical context and the local structural details according to the gravity of the disease in order to improve the prediction accuracy and interpretability of the model.

5.2 Limitations Regarding Domain Generalization and Data Diversity

Although EyePterNet-X is successful, there are a number of restrictions associated with it. The data primarily have been collected in the highest exposure ultraviolet areas which may limit the generalizability of the data to other populations and geographic areas [1]. Furthermore, the system relies on color photographs of the anterior segment only and fails to incorporate the complementary imaging modalities such as anterior segment optical coherence tomography (AS-OCT) which may include additional useful information about the depth to be included in severity assessment [20].

Another constraint is that of the acquisition devices for images. The model did not perform quite as well with the images taken using a smartphone, indicating sensitivity to variations in image quality and illumination, as the images were mostly slit-lamp images. Future work will involve building a more extensive multi-center database, multimodal imaging, and including domain adaptation techniques to enhance robustness and generalization to other clinical settings and imaging devices [21, 22].

6. Conclusion

In this paper, authors presented EyePterNet-X, a novel hybrid CNN-Transformer framework for automated analysis of pterygium, from anterior segment eye images. The proposed architecture effectively leverages the fine-grained textural details and the long-range spatial dependencies by synergistically integrating an EfficientNetV2 backbone with a Swin Transformer branch. The Cross-Attention Guided Feature Fusion Module (CAFFM) is the main technical contribution, which uses bidirectional cross-attention and adaptive gating mechanism to dynamically adjust the local and global features. This enables network to focus on local details for accurate segmentation while also tapping into the global context for accurate grading of severity. Moreover, unified detection, segmentation and grading within a single end-to-end model is provided by multi-task learning heads that are optimized with a composite loss function.

It was validated on an extensive dataset on a large clinical study, with all three clinical tasks, with the experts' annotations, and it was shown that EyePterNet-X always outperforms the state-of-the-art single-task baseline and naive hybrid baseline. The attention maps show a dynamic gate behavior, switching between global geometry and local texture depending on the level of severity of the disease, with clinically interpretable insights. The model is well performing and has some limitations in terms of the ability of domain generalization for a variety of imaging devices and populations. Moving forward, clinical validation with prospective data sets, incorporation of temporal analysis for progression monitoring, and investigation of privacy-preserving federated learning paradigms to explore and facilitate real-world deployment in telemedicine settings will be explored.

References

[1] T. Shahraki, A. Arabi, and S. Feizi, “Pterygium: An update on pathophysiology, clinical features, and management,” *Therapeutic Advances in Ophthalmology*, vol. 13, pp. 1–21, 2021, doi: 10.1177/25158414211020152.

[2] R. Asokan, R. S. Venkatasubbu, L. Velumuri, V. Lingam, and R. George, “Prevalence and associated factors for pterygium and pinguecula in a South Indian population,” *Ophthalmic and Physiological Optics*, vol. 32, no. 1, pp. 39–44, 2012, doi: 10.1111/j.1475-1313.2011.00882.x.

[3] M. K. S. V. and R. G., “Computer-aided diagnosis of anterior segment eye abnormalities using visible wavelength image analysis based machine learning,” *Journal of Medical Systems*, vol. 42, Art. no. 128, 2018, doi: 10.1007/s10916-018-0980-z.

[4] X. Fang, M. Deshmukh, M. L. Chee, Z. D. Soh, *et al*., “Deep learning algorithms for automatic detection of pterygium using anterior segment photographs from slit-lamp and hand-held cameras,” *British Journal of Ophthalmology*, vol. 106, no. 8, pp. 1107–1113, 2022.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[6] M. Tan and Q. V. Le, “EfficientNetV2: Smaller models and faster training,” in *Proc. 38th International Conference on Machine Learning (ICML)*, *Proc. Machine Learning Research*, vol. 139, pp. 10096–10106, 2021.

[7] Z. Dai, H. Liu, Q. V. Le, and M. Tan, “CoAtNet: Marrying convolution and attention for all data sizes,” in *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[8] R. Caruana, “Multitask learning,” *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997, doi: 10.1023/A:1007379606734.
[9] Q. Ji, W. Liu, Q. Ma, L. Qu, L. Zhang, and H. He, “A semantic segmentation-based automatic pterygium assessment and grading system,” *Frontiers in Medicine*, vol. 12, Art. no. 1507226, 2025, doi: 10.3389/fmed.2025.1507226.
[10] M. A. Zulkifley, “Automated segmentation of pterygium lesions using multiscale deep learning networks,” *Experimental Eye Research*, vol. 266, Art. no. 110928, 2026, doi: 10.1016/j.exer.2026.110928.
[11] B. Chen, Y. Liu, X. Zhang, *et al*., “Artificial intelligence assisted pterygium diagnosis: Current status and future perspectives,” *International Journal of Ophthalmology*, vol. 16, no. 9, pp. 1459–1468, 2023, doi: 10.18240/ijo.2023.09.04.
[12] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers,” *Medical Image Analysis*, vol. 97, Art. no. 103280, 2024, doi: 10.1016/j.media.2024.103280.
[13] K.-H. Hung, C. Lin, J. Roan, C.-F. Kuo, *et al*., “Application of a deep learning system in pterygium grading and further prediction of recurrence with slit lamp photographs,” *Diagnostics*, vol. 12, no. 4, Art. no. 888, 2022, doi: 10.3390/diagnostics12040888.
[14] H. He, L. Lin, Z. Cai, and X. Tang, “JOINED: Prior Guided Multitask Learning for Joint Optic Disc/Cup Segmentation and Fovea Detection,” in *Proceedings of Machine Learning Research, Medical Imaging with Deep Learning (MIDL)*, vol. 172, pp. 1–16, 2022.
[15] L. Pascal, O. J. Perdomo, X. Bost, *et al*., “Multi-task deep learning for glaucoma detection from color fundus images,” *Scientific Reports*, vol. 12, Art. no. 12985, 2022, doi: 10.1038/s41598-022-17126-6.
[16] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.
[17] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, *Proceedings of Machine Learning Research*, vol. 97, pp. 6105–6114, 2019.
[18] M. Shafiq and Z. Gu, “Deep Residual Learning for Image Recognition: A Survey,” *Applied Sciences*, vol. 12, no. 18, Art. no. 8972, 2022, doi: 10.3390/app12188972.
[19] W. Soliman and T. A. Mohamed, “Spectral-domain anterior segment optical coherence tomography assessment of pterygium and pinguecula,” *Acta Ophthalmologica*, vol. 90, no. 5, pp. e372–e378, 2012, doi: 10.1111/j.1755-3768.2011.02330.x.
[20] V. Kurilova, J. Goga, A. Thurzo, J. Pavlovicova, *et al*., “Review of Smartphone Deep Learning Applications Using Eye Imaging Diagnostic Techniques in Ophthalmology,” *IEEE Access*, vol. 13, pp. 187410–187441, 2025, doi: 10.1109/ACCESS.2025.3626696.
[21] Q. Xie, Y. Li, N. He, M. Ning, K. Ma, G. Wang, Y. Lian, and Y. Zheng, “Unsupervised Domain Adaptation for Medical Image Segmentation by Disentanglement Learning and Self-Training,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2795–2806, 2022.
[22] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, J. Wilcox, H. Ruamviboonsuk, A. Vardoulakis, Y. Choi, S. Schudel, L. Sayres, *et al*., “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12, doi: 10.1145/3313831.3376718.

[23] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[25] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in *International Conference on Learning Representations (ICLR)*, 2019.
[26] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, *Proceedings of Machine Learning Research*, vol. 54, pp. 1273–1282, 2017.