DLCE-MIL: Depth-aware Local Context Enhanced Multiple Instance Learning for Patient-Level IVCM Fungal Keratitis Classification

Xiaoyan Men[1], Ling-Xiao Zhang[2], Fang-Lue Zhang[3], Jie Zhou[2], Mengying Qi[4], Fen Chen[4], Qingyan Zeng[4,*], Lin Gao[2,5,*], and Yiqiang Chen[2]

[1] Shandong Normal University, China
13145342161@163.com
[2] Institute of Computing Technology, Chinese Academy of Sciences, China
{zhanglingxiao, zhoujie, gaolin, yqchen}@ict.ac.cn
[3] The University of New South Wales, Sydney, Australia
fanglue.zhang@unsw.edu.au
[4] Wuhan Aier Hankou Eye Hospital, Wuhan, China
{qimengying2014@163.com, 1402796997@qq.com, zengqingyan@aierchina.com}
[5] University of Chinese Academy of Sciences, Beijing, China

Abstract. Fungal keratitis (FK) is a vision-threatening corneal infectious disease. Among its various types, Fusarium and Aspergillus are the most common fungal genera, requiring accurate identification for treatment. In vivo confocal microscopy (IVCM) captures corneal images sequentially along the depth axis, generating structured slice sequences with depth-dependent pathological features. However, existing multiple instance learning (MIL) methods treat slices as independent and identically distributed (IID) instances, neglecting both the local inter-slice correlations produced by continuous scanning and the critical diagnostic indicator of image depth. To address this, we propose a Depth-aware Local Context Enhanced Multiple Instance Learning (DLCE-MIL) framework. This framework incorporates physical imaging depth by integrating learnable depth embeddings directly with discriminative features extracted by DINOv2. The depth information then conditionally modulates instance representations within the latent space, providing crucial contextual information. Secondly, motivated by the local dependencies among clustered slices observed during clinical acquisition, we design a Local Instance Context Enhancement (LICE) module. This module effectively captures short-range contextual relationships among neighboring slices in clinician-guided, irregularly acquired IVCM sequences, complementing global MIL aggregation with explicit local context modeling.

Keywords: Fungal keratitis · In vivo confocal microscopy · Multiple instance learning.

* Corresponding authors.

1 Introduction

Fungal keratitis (FK) is a rapidly progressing and sight-threatening infectious corneal disease, with a rising incidence in developing countries [2]. In clinical practice, accurately distinguishing between fungal subtypes, such as *Fusarium* and *Aspergillus*, is crucial for determining antifungal treatment regimens, as different genera exhibit significant variations in drug sensitivity and prognosis [23, 12]. While deep learning research based on slit-lamp photographs has shown value in the preliminary screening of infectious keratitis (IK) [24, 8, 10], macro-level imaging struggles to capture definitive micro-pathological evidence.

In Vivo Confocal Microscopy (IVCM), an important adjunctive tool, enables the observation of fungal hyphae morphology at cellular resolution [1]. However, fungal genera often exhibit highly similar filamentous branching structures, frequently forming complex intertwined networks, which makes expert interpretation highly subjective and dependent on clinical experience [1, 4]. While some studies have applied deep learning to fungal keratitis and the classification of pathogenic genus in IVCM images [18, 13, 22, 6], most remain limited to predictions at the image-level, neglecting the spatial and structural information inherent in IVCM sequences [13]. Recent work on IVCM decision-support using corneal layer information further highlights the importance of structural cues [15], but targets Acanthamoeba keratitis rather than the classification of fungal genus.

The imaging mechanism of IVCM naturally aligns with Multiple Instance Learning (MIL) for patient-level diagnosis [13]. Nevertheless, FK subclassification from IVCM presents a practical structural challenge: clinician-guided and non-uniform axial acquisition. During clinical acquisition, clinicians dynamically adjust the focal plane along the corneal z-axis to track lesions [11], producing image sequences where the input order does not strictly correspond to the actual physical depth. This non-monotonic sampling introduces a complex structure, characterized by local tissue continuity between adjacent slices and occasional cross-layer jumps. Many MIL frameworks formulate a bag as an unordered set and aggregate instance features through permutation-invariant pooling or attention-based bag aggregation, as in ABMIL and CLAM [9, 17]. Although more advanced variants such as TransMIL further model inter-instance relations [21], they still do not explicitly encode the physical acquisition structure or absolute depth in IVCM sequences.

To address these challenges, we propose **DLCE-MIL** (Depth-aware Local Context Enhanced Multiple Instance Learning) for IVCM fungal keratitis classification. First, we employ DINOv2 [19] as a feature extraction backbone to capture micro-hyphae textures. Next, we introduce the Local Context Enhancement module, which leverages 1D convolutions to explicitly model local continuity between adjacent slices. Finally, we implement a depth-aware fusion mechanism that incorporates physical imaging depth as a conditional modulation signal, enabling the model to perceive absolute axial positions and better exploit the structural information of the sequences.

Our main contributions are as follows:

1. We propose DLCE-MIL (Depth-aware Local Context Enhanced Multiple Instance Learning), a novel framework for fungal keratitis classification in IVCM images that enables accurate patient-level diagnosis.
2. We design a depth-aware fusion mechanism that incorporates physical imaging depth as a conditional modulation signal. Additionally, we introduce a Local Instance Context Enhancement module based on lightweight 1D convolution, which operates on sequential instance features to explicitly capture short-range inter-slice continuity induced by continuous IVCM scanning and enhance local contextual modeling in MIL.
3. We construct a large-scale IVCM dataset, and extensive experiments demonstrate that our method achieves state-of-the-art performance with 79.32% accuracy and 0.8192 AUC.

2 Method

2.1 Overall Framework

The overall architecture of the proposed **DLCE-MIL** framework is illustrated in Fig. 1, showing the complete data flow from raw IVCM instance sequences to patient-level classification. We propose the **DLCE-MIL** framework for patient-level fungal keratitis subtype classification. Traditional Multiple Instance Learning (MIL) typically operates under the IID assumption, treating instances within a bag as a permutation-invariant set [9]. However, in clinical IVCM acquisition, images are not captured in a strictly monotonic depth order; instead, clinicians dynamically adjust the focal plane across different corneal layers [11]. As a result, patient samples form irregularly acquired axial sequences that exhibit local tissue continuity between adjacent slices as well as occasional cross-layer transitions across distinct depths [4].

This acquisition characteristic presents two challenges: (1) adjacent captured images usually possess local structural continuity; (2) the acquisition order does not necessarily equal the physical depth order [11, 20, 3]. To address this, we construct instance sequences following the original acquisition order:

$$\mathcal{B}_p = \{(I_{p,i}, d_{p,i})\}_{i=1}^{N_p} \tag{1}$$

where $I_{p,i}$ and $d_{p,i}$ denote the i-th frame and its corresponding depth. Visual features $\mathbf{X}_p \in \mathbb{R}^{N_p \times 384}$ are extracted using a frozen DINOv2 encoder [19]. We first introduce the Local Instance Context Enhancement (LICE) module, which applies 1D convolutions along the acquisition sequence to establish explicit information propagation paths between adjacent observations.

Furthermore, to resolve the discrepancy between acquisition order and physical depth, we introduce a Depth-aware Fusion mechanism. By embedding physical depth as an independent axial coordinate:

$$\mathbf{H}^{(d)} = \mathbf{H} + \alpha \cdot \mathbf{E}(d) \tag{2}$$

the model explicitly incorporates absolute axial position while preserving sequence modeling under irregular clinical acquisition.

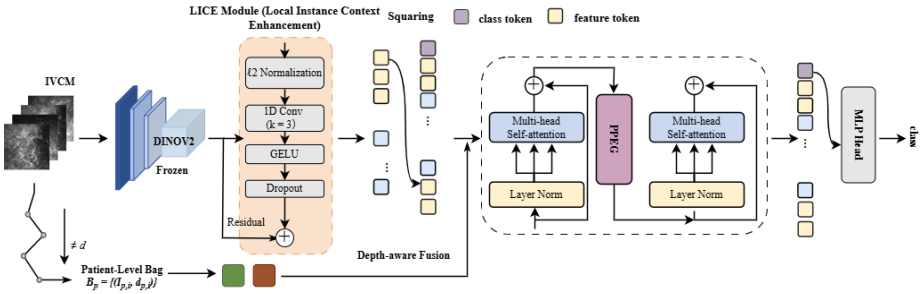

Fig. 1. Overview of the DLCE-MIL framework. The raw IVCM sequence is first processed by a frozen DINOv2 to extract instance features. The LICE module enhances local contextual information, followed by a Depth-aware Fusion mechanism to integrate absolute axial coordinates. The final patient-level prediction is generated via a Transformer-based MIL backbone with PPEG.

2.2 Local Instance Context Enhancement (LICE)

In IVCM-based patient-level MIL, positive instances often originate from adjacent corneal regions or similar depths, leading to local correlations in the sequence. While Transformer-based backbones model global dependencies, they lack explicit local inductive biases. We propose LICE to facilitate lightweight local information propagation before global aggregation.

Let $\mathbf{X} \in \mathbb{R}^{B\times N\times C}$ be the instance features. To reduce feature scale variance across patients and stabilize cross-validation, we first apply instance-wise ℓ_2 normalization:

$$\tilde{\mathbf{X}}_{b,n,:} = \frac{\mathbf{X}_{b,n,:}}{\|\mathbf{X}_{b,n,:}\|_2 + \epsilon} \tag{3}$$

where ϵ is a small constant for numerical stability. This step prevents high-amplitude features from dominating the local aggregation. We then apply a 1D convolution along the instance dimension:

$$\mathbf{U} = \mathrm{Conv1D}\big(\tilde{\mathbf{X}}^{\top}; k = 3\big) \in \mathbb{R}^{B\times C\times N} \tag{4}$$

where $(\cdot)^{\top}$ denotes the transpose between $\mathbb{R}^{B\times N\times C}$ and $\mathbb{R}^{B\times C\times N}$. With a kernel size $k = 3$, each instance aggregates context from its immediate neighbors. The output is processed via a GELU activation $\sigma(\cdot)$, Dropout, and Group Normalization (GN):

$$\mathbf{V} = \mathrm{GN}\big(\mathrm{Dropout}_{0.1}(\sigma(\mathbf{U}))\big) \tag{5}$$

LICE utilizes a residual connection to inject the local response into the original features:

$$\mathrm{LICE}(\mathbf{X}) = \mathbf{X} + \mathbf{V}^{\top} \tag{6}$$

This design preserves the original visual representation while enhancing local context. The complexity is $\mathcal{O}(BNCk)$, which is significantly lower than the $\mathcal{O}(BN^2C)$ of self-attention.

2.3 Depth-aware Fusion

Unlike standard 2D images, IVCM frames contain explicit axial depth information [11]. Since fungal invasion depth and morphology may vary across corneal layers, depth provides a complementary structural cue for subtype discrimination [4]. Given visual features $\mathbf{H} \in \mathbb{R}^{N\times 384}$ extracted by the frozen DINOv2 encoder [19], we first project them into the latent dimension $C_h = 512$ of the MIL backbone:

$$\mathbf{H}_{\text{proj}} = \text{Linear}(\mathbf{H}) \in \mathbb{R}^{N\times C_h}. \tag{7}$$

For each instance, its physical depth d_i is mapped to a learnable embedding $\mathbf{e}(d_i) \in \mathbb{R}^{C_h}$, and the depth-enhanced representation is obtained by gated additive fusion:

$$\mathbf{H}_i^{(d)} = \mathbf{H}_{\text{proj},i} + \alpha \cdot \mathbf{e}(d_i), \tag{8}$$

where α is a learnable scalar initialized to 0. This design preserves the original feature dimensionality while allowing depth to act as an axial modulation signal in the visual semantic space. Initializing $\alpha = 0$ makes the model start from a purely visual representation and allows depth information to be gradually incorporated during optimization.

2.4 Transformer-based MIL Backbone

To perform patient-level diagnosis from IVCM image collections, we formulate each patient as a bag and each image frame as an instance. Specifically, all positive IVCM images from the same patient are organized as an instance sequence and jointly processed within a multiple instance learning (MIL) framework, where the bag label corresponds to the patient-level fungal subtype. Based on the depth-enhanced instance features $\mathbf{H}^{(d)}$, we employ TransMIL [21] to model both global inter-instance correlations and bag-level aggregation for patient-level prediction. Following the original TransMIL terminology, we first perform "squaring" by padding the N instance tokens to $\hat{N} = S^2$ tokens, where $S = \lceil\sqrt{N}\rceil$. The padded tokens can then be reshaped into a pseudo-spatial $S \times S$ grid before the PPEG module. Here, "squaring" only refers to padding the sequence length to a square number, rather than element-wise squaring of token features. We then prepend a learnable CLS token:

$$\mathbf{Z}_0 = [\mathbf{h}_{cls}; \text{Pad}(\mathbf{H}^{(d)})] \in \mathbb{R}^{B\times(1+\hat{N})\times C_h} \tag{9}$$

The resulting sequence is processed by stacked Transformer blocks to capture long-range dependencies among instances, and the final CLS token is used for patient-level classification.

3 Experiments

3.1 Datasets

The dataset utilized in this study consists of in vivo confocal microscopy (IVCM) images acquired via Heidelberg HRT-III devices. The cohort includes 10,643 JPG

images (384×384 resolution) from 127 patients. The distribution comprises 75 patients with *Fusarium* keratitis (5,693 images) and 52 patients with *Aspergillus* keratitis (4,950 images). All labels were confirmed through clinical culture results and expert review. This retrospective study used anonymized clinical IVCM images and was conducted in accordance with institutional ethical requirements.

To ensure robust evaluation and prevent data leakage, we adopted patient-level 5-fold cross-validation. Specifically, the 127 patients were partitioned into five mutually exclusive folds using a stratified sampling strategy, so as to preserve, as much as possible, the overall class distribution of pathogen categories (Fusarium and Aspergillus) in each fold. During data splitting, the patient was treated as the minimum unit, ensuring that all images from the same patient were strictly assigned to either the training set or the test set within a given fold. This design prevents data leakage and avoids overestimating model performance. Finally, the mean performance across the five folds was reported as the overall evaluation metric.

3.2 Implementation Details

We employ a frozen DINOv2-Small model as the backbone for feature extraction. Images are resized to 336×336 and normalized using ImageNet statistics. For patient-level modeling, each patient is treated as a single bag containing all extracted instance features. Acquisition depth values are aligned with their corresponding frames; missing values (none encountered in this study) are handled with a -1 placeholder.

The MIL framework is optimized using the Lookahead-RAdam optimizer with a learning rate of 2×10^{-4} and weight decay of 1×10^{-3} for 100 epochs. Training is conducted with a batch size of 1 and gradient accumulation of 2 on an NVIDIA RTX 3090 GPU. We evaluate model performance using AUC, Accuracy (Acc), and F1-score (F1). The model was trained using the cross-entropy loss for binary fungal subtype classification. For each fold, the checkpoint with the best validation accuracy was selected.

3.3 Comparative Experiments

To comprehensively evaluate the performance of the proposed depth-aware MIL model, we designed comparative experiments across three dimensions:

Backbone Comparison We selected three representative deep vision models as feature extractors (backbones) and compared them under the same MIL classification framework to evaluate differences in representation capabilities: (1) **ResNet18**: A classic convolutional neural network baseline [7]; (2) **ConvNeXt-v2**: A modern pure convolutional architecture [25]; (3) **Swin Transformer**: A vision Transformer based on shifted window attention [16]. To ensure fairness, all backbones were used to extract image features for the same TransMIL [21] classifier, with consistent training strategies, epochs, and data partitioning.

MIL Frameworks Comparison We compared our proposed model with mainstream Multiple Instance Learning (MIL) frameworks. All models utilized DINOv2 [19] as the input to verify the effectiveness of our improved modules:

- **ABMIL**: A classic pooling method based on attention mechanisms [9].
- **CLAM**: An MIL model incorporating clustering constraints and attention weights [17].
- **TransMIL**: The original Transformer-based MIL framework [21].
- **Proposed (Ours)**: Our model introduces the Local Instance Context Enhancement (LICE) module and a Depth fusion mechanism on top of TransMIL.

Keratitis-specific SOTA To objectively evaluate the clinical advancement of this study, we reproduced and compared two representative keratitis classification algorithms: Erukulla et al. [5] and Tang et al. [22]. As the original datasets for these studies are not public, we retrained and tested these architectures on our own clinical dataset based on the descriptions in their respective papers. Since these methods were originally designed for instance-level classification, whereas our task targets patient-level diagnosis, we adopted a simple adaptation strategy: the final patient-level prediction P_{patient} is obtained by averaging the predicted probabilities over all n images from the same patient [14]:

$$P_{\text{patient}} = \frac{1}{n} \sum_{i=1}^{n} p_i \tag{10}$$

Results and Discussion The comparative results are summarized in Table 1. DLCE-MIL achieves the best performance across AUC, Accuracy, and F1-score.

Notably, under the same TransMIL aggregation head, ResNet18 yields a higher mean performance with lower cross-fold variance on our clinical IVCM dataset. This suggests that, in our IVCM setting with limited sample size, noisy slices, and substantial acquisition heterogeneity, the inductive bias of lightweight CNNs may be advantageous for learning robust local textures and pathological morphologies, whereas more complex backbones may be more sensitive to data scale and regularization. Accordingly, we use the backbone study to analyze representation differences, while fixing DINOv2 features as a unified input in the MIL framework comparison to control confounding factors, ensure a fair evaluation, and attribute performance gains more clearly to our core modules (LICE and depth fusion).

3.4 Ablation Studies

We conducted ablation experiments under the same patient-level 5-fold cross-validation protocol, and the results are summarized in Table 2.

Table 1. Comparative results of different MIL frameworks and keratitis-specific methods on our clinical IVCM dataset. Performance is reported as average ± standard deviation across 5-fold cross-validation.

Method	AUC	Accuracy	F1-Score
ConvNeXt-v2	0.7236 ± 0.0813	0.6989 ± 0.0520	0.5804 ± 0.0535
Swin Transformer	0.7198 ± 0.0750	0.7156 ± 0.0381	0.5515 ± 0.0139
ResNet18	0.7669 ± 0.0942	0.7534 ± 0.0478	0.5758 ± 0.0888
ABMIL	0.6996 ± 0.1194	0.7306 ± 0.0940	0.6339 ± 0.1558
CLAM	0.7762 ± 0.0657	0.7641 ± 0.0692	0.6518 ± 0.0034
TransMIL	0.7856 ± 0.0858	0.7312 ± 0.0879	0.6351 ± 0.1513
Erukulla et al.	0.7253 ± 0.0418	0.6638 ± 0.0830	0.5857 ± 0.1401
Tang et al.	0.7091 ± 0.0663	0.6394 ± 0.0813	0.3011 ± 0.1629
Ours (DLCE-MIL)	$\mathbf{0.8192 \pm 0.0697}$	$\mathbf{0.7932 \pm 0.0809}$	$\mathbf{0.7494 \pm 0.0968}$

Table 2. Ablation study of the proposed modules on our clinical IVCM dataset.

Num.	Baseline	LICE	Depth Fusion	AUC	Accuracy	F1-Score
1	✓			0.7856 ± 0.0858	0.7312 ± 0.0879	0.6351 ± 0.1513
2	✓	✓		0.7890 ± 0.0751	0.7762 ± 0.0840	0.7182 ± 0.1211
3	✓		✓	0.7975 ± 0.0732	0.7666 ± 0.0512	0.7248 ± 0.0545
4 (Ours)	✓	✓	✓	$\mathbf{0.8192 \pm 0.0697}$	$\mathbf{0.7932 \pm 0.0809}$	$\mathbf{0.7494 \pm 0.0968}$

Impact of LICE Module. Integrating the LICE module (Num. 2) improved Accuracy from 0.7312 to 0.7762. Unlike global attention alone, LICE captures neighborhood-dependent morphological patterns (e.g., hyphal branching) through 1D convolutions. This suggests that local structural priors play an important role in distinguishing fine-grained fungal features in IVCM sequences.

Impact of Depth-aware Fusion. The individual introduction of physical depth information (Num. 3) improved AUC (from 0.7856 to 0.7975). As IVCM is a sequential scanning technique, the depth value of each frame reflects the physical corneal layer of fungal invasion. Our experiments demonstrate that fusing this physical spatial prior as a conditional modulation signal with visual features provides crucial contextual information for differentiating between Fusarium and Aspergillus.

Synergistic Effect (Final Model). The model achieved its optimal overall performance when simultaneously integrating both the LICE module and depth information. Compared with the baseline, the final model improved AUC from 0.7856 to 0.8192, Accuracy from 0.7312 to 0.7932, and F1-score from 0.6351 to 0.7494, demonstrating the complementary effect between local visual context enhancement and physical depth constraints.

4 Conclusion

We propose the DLCE-MIL, a depth-aware local context enhanced multiple instance learning framework for patient-level classification of fungal keratitis subtypes from IVCM images. By introducing the LICE to capture local inter-slice continuity and a learnable depth fusion mechanism to encode absolute axial position, our method effectively addresses the local continuity and depth irregularity inherent in clinical IVCM acquisition. Extensive experiments demonstrate that DLCE-MIL outperforms existing MIL approaches and keratitis-specific methods by a significant margin. In future work, we plan to extend our framework to multimodal inputs and explore its applicability to other ophthalmic imaging modalities.

Acknowledgments. This work was supported by the HNXJ Philanthropy Foundation (No. KY25008, KY24014), in part by the Natural Science Foundation of Hunan Province (No. 2024JJ9002) and in part by the National Natural Science Foundation of China (No. 62322210).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bakken, I.M., Jackson, C.J., Utheim, T.P., Villani, E., Hamrah, P., Kheirkhah, A., Nielsen, E., Hau, S., Lagali, N.S.: The use of in vivo confocal microscopy in fungal keratitis - progress and challenges. The Ocular Surface **24**, 103–118 (2022). https://doi.org/10.1016/j.jtos.2022.03.002
2. Brown, L., Leck, A.K., Gichangi, M., Burton, M.J., Denning, D.W.: The global incidence and diagnosis of fungal keratitis. The Lancet Infectious Diseases **21**(3), e49–e57 (2021). https://doi.org/10.1016/S1473-3099(20)30448-5
3. Chiang, J.C.B., Roy, M., Kim, J., Markoulli, M., Krishnan, A.V.: In-vivo corneal confocal microscopy: Imaging analysis, biological insights and future directions. Communications Biology **6**, 652 (2023)
4. Chidambaram, J.D., Prajna, N.V., Larke, N., et al.: In vivo confocal microscopy appearance of Fusarium and Aspergillus species in fungal keratitis. British Journal of Ophthalmology **101**(8), 1119–1123 (2017). https://doi.org/10.1136/bjophthalmol-2016-309656
5. Erukulla, R., Esmaili, K., Rahdar, A., Aminizade, M., Cheraqpour, K., Tabatabaei, S.A., Bibak-Bejandi, Z., Mohammadi, S.F., Yousefi, S., Soleimani, M.: Deep learning-based classification of fungal and acanthamoeba keratitis using confocal microscopy. The Ocular Surface **38**, 203–208 (2025). https://doi.org/10.1016/j.jtos.2025.07.012
6. Essalat, M., Abolhosseini, M., Le, T.H., Moshtaghion, S.M., Kanavi, M.R.: Interpretable deep learning for diagnosis of fungal and acanthamoeba keratitis using in vivo confocal microscopy images. Scientific Reports **13**(1), 8953 (2023). https://doi.org/10.1038/s41598-023-35085-9
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

8. Hung, N.S., Shih, A.Y., Lin, C.Y., Kuo, M.T., Hwang, Y.S., Wu, W.T., Lin, Y.H., Hsieh, T.C., Hsu, B.W., Chen, K.J.: Using slit-lamp images for deep learning-based identification of bacterial and fungal keratitis: Model development and validation with different convolutional neural networks. Diagnostics **11**(7), 1246 (2021). https://doi.org/10.3390/diagnostics11071246
9. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning. pp. 2127–2136. PMLR (2018)
10. Jiang, J., Liu, W., Pei, M., et al.: Automatic diagnosis of keratitis using object localization combined with cost-sensitive deep attention convolutional neural network. Journal of Big Data **10**, 121 (2023). https://doi.org/10.1186/s40537-023-00800-w
11. Kumar, R.L., Cruzat, A., Hamrah, P.: Current state of in vivo confocal microscopy in management of microbial keratitis. Seminars in Ophthalmology **25**(5-6), 166–170 (2010). https://doi.org/10.3109/08820538.2010.518516
12. Lalitha, P., Shapiro, B.L., Srinivasan, M., et al.: Antimicrobial susceptibility of fusarium, aspergillus, and other filamentous fungi isolated from keratitis. Archives of Ophthalmology **125**(6), 789–793 (2007). https://doi.org/10.1001/archopht.125.6.789
13. Li, C.P., Dai, W., Xiao, Y.P., et al.: Two-stage deep neural network for diagnosing fungal keratitis via in vivo confocal microscopy images. Scientific Reports **14**(1), 18432 (2024). https://doi.org/10.1038/s41598-024-68768-y
14. Li, Z., et al.: Deep learning for multi-type infectious keratitis diagnosis: A nationwide, cross-sectional, multicenter study. npj Digital Medicine **7**, 181 (2024). https://doi.org/10.1038/s41746-024-01174-w
15. Lincke, A., Roth, J., Macedo, A.F., Bergman, P., Löwe, W., Lagali, N.S.: Ai-based decision-support system for diagnosing acanthamoeba keratitis using in vivo confocal microscopy images. Translational Vision Science & Technology **12**(11), 29 (2023). https://doi.org/10.1167/tvst.12.11.29
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
17. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021). https://doi.org/10.1038/s41551-020-00682-w
18. Lv, J., Zhang, K., Chen, Q., Chen, Q., Huang, W., Cui, L., Li, M., Li, J., Chen, L., Shen, C., Yang, Z., Bei, Y., Li, L., Wu, X., Zeng, S., Xu, F., Lin, H.: Deep learning-based automated diagnosis of fungal keratitis with in vivo confocal microscopy images. Annals of Translational Medicine **8**(11), 706 (2020). https://doi.org/10.21037/atm.2020.03.134
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt
20. Petroll, W.M., Robertson, D.M.: In vivo confocal microscopy of the cornea: New developments in image acquisition, reconstruction and analysis using the heidelberg

retina tomograph with rostock corneal module. The Ocular Surface **13**(3), 187–203 (2015)
21. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: Transformer-based correlated multiple instance learning for whole slide image classification. In: Advances in Neural Information Processing Systems. vol. 34, pp. 2136–2147 (2021), https://openreview.net/forum?id=LKUfuWxajHc
22. Tang, N., Jiang, L., Chen, Y., et al.: An artificial intelligence approach to classify pathogenic fungal genera of fungal keratitis using corneal confocal microscopy images. International Ophthalmology **43**(7), 2203–2214 (2023). https://doi.org/10.1007/s10792-022-02616-8
23. Thomas, P.A., Kaliamurthy, J.: Mycotic keratitis: epidemiology, diagnosis and management. Clinical Microbiology and Infection **19**(3), 210–220 (2013). https://doi.org/10.1111/1469-0691.12126
24. Won, Y.K., Lee, H., Kim, Y., Han, G., Chung, T.Y., Ro, Y.M., Lim, D.H.: Deep learning-based classification system of bacterial keratitis and fungal keratitis using anterior segment images. Frontiers in Medicine **10**, 1162124 (2023). https://doi.org/10.3389/fmed.2023.1162124
25. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16133–16142 (2023). https://doi.org/10.1109/CVPR52729.2023.01548