npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01706-y

Advanced and interpretable corneal staining assessment through fine grained knowledge distillation

Check for updates

Yuqing Deng[1,2,3,13], Pujin Cheng[4,5,6,13], Ruiwen Xu[1,2,3], Lirong Ling[1,2,3], Hongliang Xue[7,8], Shiyou Zhou[1,2,3], Yansong Huang[4], Junyan Lyu[4,9], Zhonghua Wang[4,5], Kenneth K. Y. Wong[6], Yimin Zhang[1,2,3], Kang Yu[1,2,3], Tingting Zhang[1,2,3], Xiaoqing Hu[1,2,3], Xiaoyi Li[10], Xiaoying Tang[4,5] ✉, Yan Lou[11] ✉ & Jin Yuan[12] ✉

The assessment of corneal fluorescein staining is essential, yet current AI models for Corneal Staining Score (CSS) assessments inadequately identify punctate lesions due to annotation challenges and noise, risk misrepresenting treatment responses through "plateau" effects, and highlight the necessity for real-world evaluations to enhance disease severity assessments. To address these limitations, we developed the Fine-grained Knowledge Distillation Corneal Staining Score (FKD-CSS) model. FKD-CSS integrates fine-grained features into CSS grading, providing continuous and nuanced scores with interpretability. Trained on corneal staining images collected from dry eye (DE) patients across 14 hospitals, FKD-CSS achieved robust accuracy, with a Pearson's r of 0.898 and an AUC of 0.881 in internal validation, matching senior ophthalmologists' performance. External tests on 2376 images from 23 hospitals across China further validated its efficacy (r: 0.844–0.899, AUC: 0.804-0.883). Additionally, FKD-CSS demonstrated generalizability in multi-ocular-surface-disease testing, underscoring its potential in handling different staining patterns.

Cornea is the transparent, dome-shaped structure at the front of the eye that plays a crucial role in focusing light onto the retina[1,2]. Corneal epithelial staining is a significant indicator refer to epithelial defects of various ocular surface diseases[3–5] (e.g., dry eye (DE), neurotrophic keratopathy, various types of keratitis) and certain systemic conditions (e.g., diabetes-related corneal damage)[6,7]. These conditions significantly increase the risks of visual impairment and can negatively impact overall quality of life[7,8]. Fluorescein dye staining allows for the visualization of corneal epithelium lesions[9] associated with epithelial erosions or defects in the epithelial cell barrier[10], making it an important diagnostic and assessment tool for evaluating the severity of ocular surface diseases[11]. However, the current assessment for staining relies on manual grading, which is time-consuming and labor-intensive. Moreover, subjective variation caused by the reliance upon the experiences and individual perspectives of physicians makes it challenging for large-scale data analysis and reduces data comparability between different centers. While computer-assisted quantification systems have emerged to enhance objective evaluation[12–18], their diagnostic accuracy, sensitivity, and clinical utility warrant further validation.

Prior approaches have introduced image processing and traditional machine learning techniques into corneal staining evaluation, including thresholding[16], edge detection[15,17] and random forest[18]. However, those methodologies rely on pre-selected and typically fixed thresholds and hyperparameters, challenged, e.g., by confluent staining[19] (the stained area comprises multiple punctate lesions with indistinct borders) in counting punctate staining, which had limited performance on images with large individual variability. Compared to feature-based machine learning, deep

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China. [2]Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou, China. [3]Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China. [4]Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China. [5]Jiaxing Research Institute, Southern University of Science and Technology, Jiaxing, China. [6]Department of Electrical and Electronic Engineering, the University of Hong Kong, Hong Kong, China. [7]The Key Laboratory of Advanced Interdisciplinary Studies, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China. [8]Department of Nutrition, School of Public Health, Guangzhou Medical University, Guangzhou, China. [9]Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia. [10]Zhaoke (Guangzhou) Ophthalmology Pharmaceutical Ltd, Guangzhou, China. [11]Department of Computer, School of Intelligent Medicine, China Medical University, Shenyang, China. [12]Beijing Institute of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing Key Laboratory of Ophthalmology & Visual Sciences, Beijing, China. [13]These authors contributed equally: Yuqing Deng, Pujin Cheng.
✉e-mail: tangxy@sustech.edu.cn; ylou04@cmu.edu.cn; yuanjincornea@126.com

learning (DL) enables end-to-end prediction[20] with superior feature discrimination and state-of-the-art performance. Currently, several existing digital methods have introduced DL into corneal staining evaluation. Deng et al. have introduced a dataset for the classification of corneal ulcers and the segmentation of flaky ulcers[12]. Building upon this work, Wang et al. have proposed a framework for the segmentation of patchy ulcers, utilizing adjacent scale fusion and position embedding[21]. Wang et al. combining patchy ulcer segmentation and staining grading, have employed a Transformer-based network for corneal staining evaluation, incorporating additional patchy ulcer segmentation branch[13]. Qu et al. have designed a DL grading system for epithelial erosions using ResNet34[14], but failed to provide interpretability.

However, all of the aforementioned AI-based corneal staining methods encounter several limitations in their clinical applications. Firstly, while DL excels in detecting well-defined patchy lesions[13], it underperforms in identifying punctate lesions[13]. Punctate lesions appear as scattered dots or confluent areas confined to the epithelium upon staining, which is often associated with DE and frequently found surrounding the primary site of most corneal ulcers, related to inflammatory attacks and epithelial repair, which indicate disease progression and prognosis[22], occurred commonly in most ocular surface disease[6,8,23,24]. However, annotating these lesions is costly due to their small and discrete nature[25], and often accompany with “noise”, e.g., dye residue, which poses a significant obstacle to the training of existing AI models[13,26]. Numerous approaches have been developed to tackle the issue of learning from noisy labels, including adversarial training[27], Bayesian learning[28], and knowledge distillation[29]. However, there have been no such study applying noisy learning to quantitative analyses of corneal staining. Secondly, the current discrete classification-based AI-assist staining evaluation systems risk plateau effects, where the results may remain within the same clinical grading range even when a significant number of stained points decreases[16]. This limitation could misrepresent treatment response in clinical practices. Lastly, there is an urgent need for real-world evaluation on corneal staining AI systems to facilitate precise evaluation of disease severity in different medical institutions and provide accurate endpoints for large-scale, multi-center clinical trials in ocular surface disease medication.

Our study aims to address the challenges in the evaluation of corneal punctate staining. Firstly, we proposed a deep learning framework for general corneal punctate staining grading, then we introduced a fine-grained corneal staining scoring model using knowledge distillation (FKD-CSS), which employs knowledge distillation to feature dual decoder branches for CSS grading and staining segmentation. This innovative framework not only optimizes grading performance but also provides interpretability through staining area visualization to enhance the clinical applicability and accessibility. Secondly, this study will focus on DE due to its high prevalence (5% to 50%[22,30] globally) and its characteristic epithelial punctate staining[31]. Additionally, corneal staining is a recognized standard for diagnosis and severity in this population[32,33] and is also an FDA-approved endpoint for efficacy[11]. Therefore, the DE datasets prospectively recruited from a multi-center real-word clinical condition will be used as the primary training and validation dataset for the FKD-CSS model. This dataset contributes to the robustness and representativeness of our study. Furthermore, an independent external test of multiple ocular surface abnormalities was also utilized to validate the model’s generalizability in handling different staining patterns.

Results

Datasets characteristics

The images of developing, internal validation, and primary external test datasets were mainly collected prospectively from a 3-month follow-up dataset of DE patients from 37 tertiary hospitals in seven regions of China (Fig. 1) where the model is most likely to be adopted. After an automatic quality control step of the model, we retained 1471 out of 1477 corneal fluorescein staining images from Eastern China of DE dataset for both training and 5-fold cross-validation. The images were from 14 tertiary hospitals in seven provinces and municipalities (Supplementary Table 2-4, 12). The remaining images from 6 regions of China (Central, South, North, Northeast, Northwest and Southwest) except for Eastern were used for multi-center external testing of the model in the scenario of DE, yielding

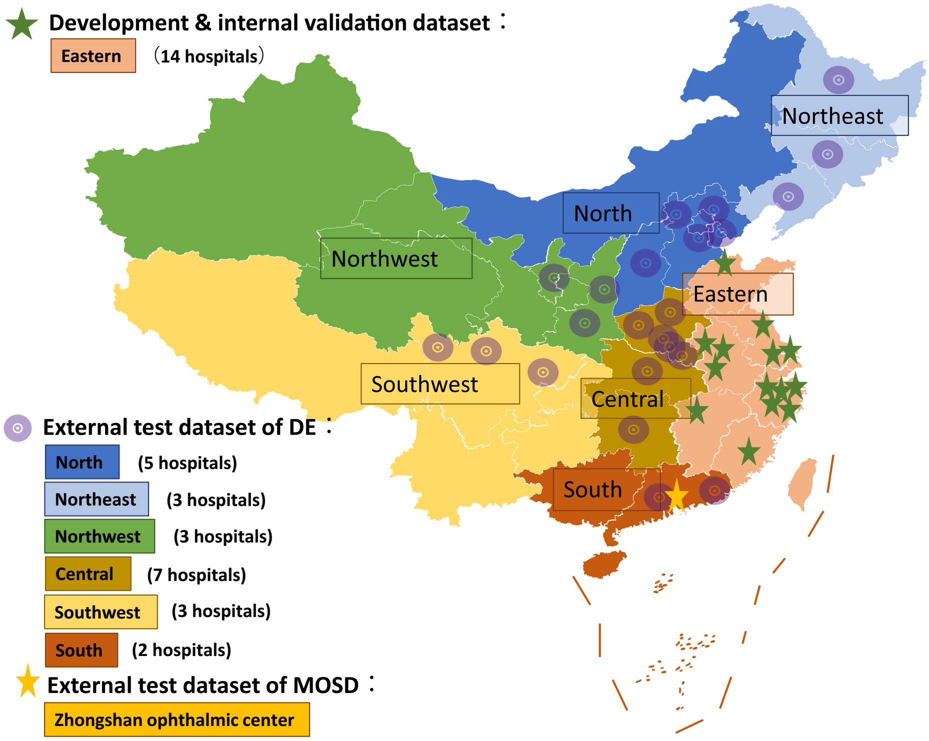

Fig. 1 | The distribution and collection of the datasets encompasses the entirety of China. DE dry eye, MOSD multiple ocular surface diseases.

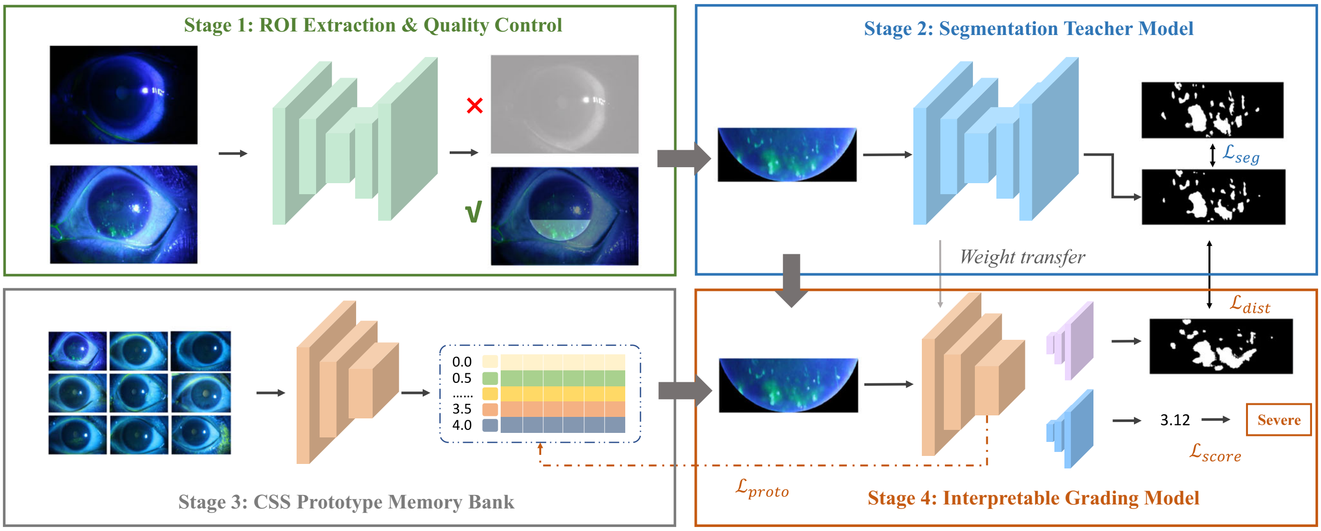
Stage 1: ROI Extraction & Quality Control
×
√
Stage 2: Segmentation Teacher Model
$\mathcal{L}_{seg}$
Weight transfer
$\mathcal{L}_{dist}$
0.0
0.5
......
3.5
4.0
Stage 3: CSS Prototype Memory Bank
3.12
Severe
$\mathcal{L}_{score}$
$\mathcal{L}_{proto}$
Stage 4: Interpretable Grading Model

Fig. 2 | The proposed FKD-CSS framework consists of four components. (1) a corneal segmentation network to extract the region of interest and employ a quality control, (2) a weakly-supervised teacher segmentation network to guide the grade model towards fine-grained lesions, and (3) a novel prototype memory bank to cluster similar semantic information across different CSS grades, (4) an interpretable grading model to predict the CSS and disease severity. ROI region of interest. FKD fine-grained knowledge distillation. CSS corneal staining score.

2376 out of 2386 images from 23 hospitals in 15 provinces and municipalities (Supplementary Table 4-7, 12). Additionally, 231 images from 18 kinds of common ocular surface disease without DE (e.g., Acanthamoeba keratitis, viral keratitis, neuroparalytic keratitis) with punctate and patchy lesions from Zhongshan ophthalmic center were annotated for external test of generalization capability of our FKD-CSS in disease staging (Supplementary Table 9). The assignment of datasets is shown in Supplementary Fig. 1. And the detailed distribution of the datasets is described in “datasets” of the “Methods” section.

The images were captured using various types of clinical commonly used cameras (Supplementary Table 10). Corneal staining image were captured followed the standard practice (Supplementary Table 1). The images were labeled with CSS[34] (0 - 4), which had been utilized as primary efficacy assessment in various clinical trials[34–36]. By a triple read and arbitration method (Supplementary Fig. 2), the label was required to be accurate to 0.5 points, resulting in 9 classes, ranging from 0 to 4 in 0.5 intervals, as the ground-truth of CSS. To facilitate model performance comparison with the commonly used clinical classification of corneal staining, the CSS (score x) was also converted into discrete categories to classifies ocular epithelium damage[34,35,37,38]: Staining Negative or Without Clinical Significance (SNWC): x<0.5. Mild: 0.5 ≤ x<1.5, indicating mild damage. Moderate:1.5 ≤ x<2.5, indicating mild inflammation & worsening defects. Severe:2.5 ≤ x<3.5, indicating chronic inflammation & extensive damage. Critical:x ≥ 3.5, indicating severe inflammation or significant epithelial damage. Additionally, senior ophthalmologists provided their best clinical judgment of CSS that took into account all the morphological features of the lesions present in the Region of Interest (ROI) of the images by coarsely annotating 100 images randomly selected from the development datasets, allow the FKD-CSS grading framework to effectively transfer prior knowledge about the correlation between staining grades and lesion locations to a larger and more generalized dataset using knowledge distillation, 80 annotated images from each of the two external test sets were used for staining detection validation for external test of the model in DE and multi-disease scenarios. Notably, the assessment for DE dataset focused primarily on the lower third of the cornea instead of complete corneal, which has been demonstrated to present the highest degree of corneal staining in symptomatic DE patients most commonly[39,40], while evaluation was focused in the entire corneal area in dataset of multiple ocular surface diseases (MOSD). For details, see “corneal staining test” section in Methods.

System architecture

We developed an advanced automatic CSS grading system that utilizes fine-grained staining information to output continuous and consistent scores. As illustrated in Fig. 2, the framework’s training pipeline encompasses three key stages: ROI cropping, segmentation teacher network training, and CSS grading network training. In the inference phase, the system initially detects the ROI areas in the cornea before directly outputting the CSS score. Notably, due to the incorporation of fine-grained knowledge distillation in the grading network’s training, the system is adept at identifying crucial staining areas that significantly influence CSS grading. A comprehensive explanation of this process is detailed in the Methods: “Algorithm Construction of FKD-CSS” section.

Performance of baseline model between regression and categorization

Many existing methods treat the cornea grading as a classification problem[13,14], However, the suitability of these methods for finer-grained lesions is questionable. Classification focuses solely on the accuracy of individual grades, overlooking the overall consistency and correlation among different grades. To address this, we conducted experiments on ResNet50, the fundamental deep learning backbone of FKD-CSS, under both classification and regression settings. The primary difference between these two models lies in their final fully-connected layer: the classification model generates a probability distribution across 9 discrete output values, corresponding to each CSS grade increment (ranging from 0 to 4 in 0.5-step intervals), whereas the deep regression model directly predicts a continuous CSS value. Our results indicate the classification model’s limitations in CSS tasks, achieving only a Pearson Correlation of 0.727 and an area under the curve of receiver operating characteristic (AUC-ROC) of 0.438. In contrast, the regression model scores 0.828 and 0.803, respectively. It is evident that the regression model is more suitable for providing consistent and robust CSS grading, aligning closely with the nuanced demands of CSS assessment.

Performance and interpretability of FKD-CSS compared with other deep learning algorithms

To assess the performance of the proposed FKD-CSS model, we compared it with established medical image feature extractors that are widely used for ocular staining grading or segmentation: ResNext50[41], DenseNet121[42], and ResNest50[43]. We adapt all of these models as regression models.

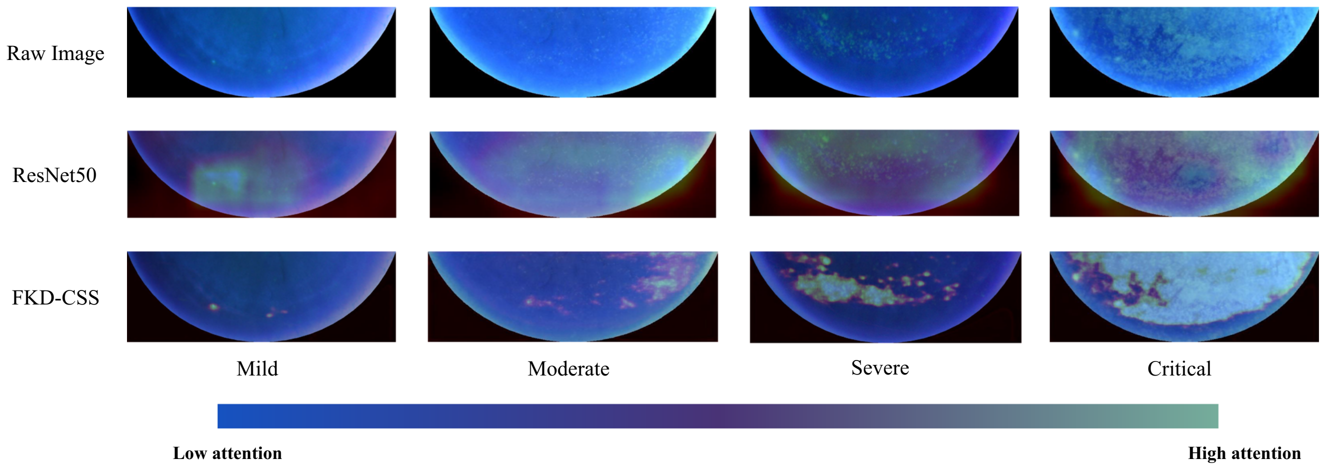

Fig. 3 | Feature visualization of the proposed FKD-CSS model and ResNet50 under randomly chosen samples with different levels of dry eye. The first row is the original raw image. The second row is the class activation mapping plus plus (CAM++) of the ResNet50. The third row is the fined-grained lesion detection visualization from the proposed FKD-CSS model. FKD fine-grained knowledge distillation. CSS corneal staining score.

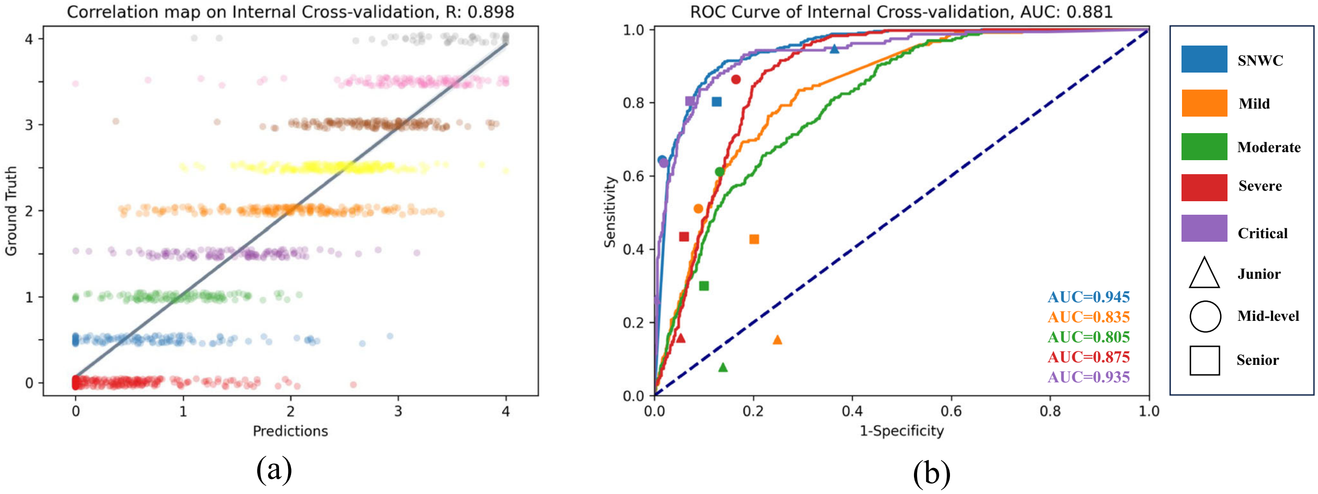

Fig. 4 | Performance of FKD-CSS in internal cross-validation set. a the correlation map between the FKD-CSS and the ground-truth CSS grading. **b** the AUC-ROC curves of the FKD-CSS model and ophthalmologists with different seniority. FKD fine-grained knowledge distillation. CSS corneal staining score. AUC area under the curve, ROC receiver operating characteristic. SNWC staining negative or without clinical significance.

However, despite their ability to extract high-level visual features due to pre-training on ImageNet, these methods exhibit deficiencies in the extraction of staining features, as illustrated in Table 2. The performance of ResNeSt ($r = 0.834$, AUC = 0.802), DenseNet121 ($r = 0.842$, AUC = 0.818), and ResNext50 ($r = 0.858$, AUC = 0.818) still falls short when compared to the proposed FKD-CSS model ($r = 0.898$, AUC = 0.881). Furthermore, we compared the performance of FKD-CSS with other methods using samples from SNWC to evaluate the model's ability to distinguish normal populations. FKD-CSS exhibits the best performance (AUC = 0.881) compared to other comparative methods (AUC:0.825–0.847). (Supplementary Table 11).

We provide a feature visualization comparison between the FKD-CSS model and the baseline backbone, ResNet50. The visualization for ResNet50 is achieved using gradient class activation mapping plus plus (GradCAM + +)[44]. As depicted in Fig. 3, the FKD-CSS model successfully captures fine-grained staining areas in corneal staining images of different CSS gradings, whereas ResNet50 only captures coarse-grained staining areas. These visualization results serve as visual evidence of the proposed FKD-CSS model's ability to generate meaningful and informative interpretability.

Achieving expert performance in clinical internal validation

In comparison to six ophthalmologists (Pearson's $r \leq 0.858$), the FKD-CSS model achieved the highest Pearson's r of 0.898, indicating a strong positive correlation with the ground-truth CSS grading and affirming its accuracy. Additionally, it attained the highest AUC of 0.881, which is surpasses that of the senior ophthalmologists (AUCs ≤ 0.761), exemplifying its superior ability to distinguish between different CSS grades (Fig. 4, Table 1).

Achieving expert performance in 6 external datasets of dry eye

We conducted a comprehensive performance test on six external test datasets of DE, sourced from 6 out of 7 regions across China. The Pearson's r between the FKD-CSS model's predictions and the actual clinical assessments ranges from 0.844 to 0.899, and the AUCs shown the performance of classification of the FKD-CSS model in comparison with ground-truth CSS grading ranges from 0.804 to 0.883 (Fig. 5), demonstrate performance on par with senior ophthalmologists (AUC: 0.691–0.806) (Table 2). Notably, the model's best performance in classification is most frequently observed in SNWC (AUC: 0.921–0.966) and Critical (AUC: 0.927–0.990), followed by Mild (AUC: 0.753–0.868) and Severe (AUC: 0.782–0.871), while its worst performance is most commonly seen in distinguishing Moderate severity

Table 1 | Comparison of the performance of the proposed FKD-CSS model and different AI models and ophthalmologists on internal cross-validation

Comparison	Pearson	AUC
AI Methods		
ResNeSt50	0.834	0.827
ResNeXt50	0.858	0.847
DenseNet121	0.842	0.825
Ophthalmologists		
Junior A	0.755	0.631
Junior B	0.665	0.600
Mid-level A	0.821	0.714
Mid-level B	0.861	0.749
Senior A	0.837	0.739
Senior B	0.858	0.761
Proposed		
FKD-CSS	**0.898**	**0.881**

The best results are highlighted in bold a and the second-best results are underlined.
All *p*-value of Pearson test are statistically significant (*p*-value < 0.001).
FKD fine-grained knowledge distillation, *CSS* corneal staining score, *AUC* area under the curve.

(AUC: 0.705–0.818) across the test sets from six different regions. In contrast, senior ophthalmologists show lower performance in these stages, with AUC: 0.780–0.911 for SNWC, 0.587–0.814 for mild severity, 0.518–0.749 for moderate severity, 0.609–0.789 for severe severity, and 0.620–0.934 for critical severity. These findings highlight the FKD-CSS model's robustness, outperforming senior ophthalmologists in these critical assessments.

To assess the explainability of the FKD-CSS model, we randomly selected 80 images from the external test dataset of DE (20 images for each category from 1 to 4) for expert annotation, which was used for lesion detection analysis. The FKD-CSS model successfully detected 95.3% of these annotated lesion areas. This high detection rate (Recall) demonstrates the model's capability in providing insights that align with the expertise of ophthalmologists in identifying corneal staining lesions.

Model generalizes to multiple ocular surface diseases

To further assess the generalization ability of the proposed FKD-CSS model, we conducted a clinical test on a MOSD dataset encompassing 18 common ocular surface diseases except for DE. The FKD-CSS model achieved a Pearson's r of 0.816 and an AUC of 0.807 (0.716–0.968 in the specific classification), demonstrating its robust generalization capabilities across different ocular surface diseases (Fig. 6).

In evaluating the FKD-CSS model's fine-grained detection ability, senior ophthalmologists annotated the fine-grained staining areas of 80 images randomly selected from the MOSD dataset. Remarkably, the FKD-CSS model detected 78.5% of these areas. We also present visualization examples of the FKD-CSS model applied to various diseases. These visualizations highlight the model's effectiveness in capturing the nuanced staining areas characteristic of these diseases, further illustrating its capability to accurately assess the severity of fine-grained staining areas in a range of ocular surface conditions (Fig. 7).

Interactive visualization tool

Given the rich and varied nature of our grading data, which encompasses inputs from six different ophthalmologists across seven regions and multiple diseases, we have developed an interactive visualization software for more effective data analysis (Supplementary Software 1). This tool is designed to present comprehensive data distributions and facilitate various comparisons. Users can interactively select different options to compare the performance of various levels of ophthalmologists and the AI system across different regions and disease categories. Such functionality allows for in-depth exploration and a better understanding of the data, enhancing the overall analysis process.

Discussion

The contribution of this study is three-fold: (1) Rather than segmenting only patchy lesions or just generating a CSS grading, the FKD-CSS model is able to give the CSS grades and the coarse staining segmentation simultaneously with the dual-decoder architecture, which addresses the gap in identifying the punctate lesion of corneal staining with interpretability by using fine-grained knowledge distillation. (2) Unlike traditional methods that only offer categories of grading, FKD-CSS offers rich, fine-grained pathological features, producing continuous CSS score to show the severity level of pathological epithelium damage, which are valuable for detailed assessments in clinical trials; delivering fine-grained lesion detection with visualizations of the shape, size, and location of lesions, which contain etiological information[39,40]. As such, the AI system thereby assisting clinicians in evaluating etiology, severity, and treatment efficacy. (3) The FKD-CSS showed satisfactory performance for evaluating CSS in real-world settings of DE and external set of multiple ocular surface diseases using prospectively collected corneal staining images, and so could allow the system to be implemented and adopted for clinical care.

Due to limitations in both data size and quality, traditional DL methods are unable to extract rich features of punctate lesion for CSS grading training. While transfer learning can help by pre-training on large-scale natural image datasets, biases introduced may not optimally represent CSS-specific features. The complexity of staining patterns—irregular morphologies, heterogeneous intensities, and sparse distributions—requires task-specific inductive biases to align feature learning with clinical criteria, reducing reliance on extensive high-quality datasets[45,46]. Incorporating fine-grained information through multi-task learning and segmentation supervision presents a solution, yet the discrete nature of punctate lesions complicates precise or even coarse labeling on a large scale. To overcome this issue, we propose a fine-grained knowledge distillation approach that leverages annotations from a small-sized and coarse-labeled dataset focusing on punctate lesions in a weakly supervised manner. In fact, coarse annotations are valuable in weakly supervised learning in medical image analysis[47], enabling practical, scalable, and effective training of models while alleviating challenges in obtaining high-quality, detailed annotations[48], thus enhancing a model's robustness and generalizability[44]. This allows the model to make educated guesses about finer details, effectively learning to localize or segment the target structures even without accurate pixel-level annotations' guidance. The coarse-labeled dataset consists of diverse corneal stained images across various severity levels, allowing the AI model to learn punctate patterns effectively (details are presented in Sec.2.1). A teacher segmentation network, trained on this dataset, serves as a foundation to initialize and provide fine-grained supervision for the grading network, especially for large-scale data without manual staining mask annotation. By utilizing this pre-trained segmentation teacher network, the grading model initializes robust encoder parameters for subsequent CSS grading training. Furthermore, the knowledge distillation approach minimizes the Kullback–Leibler divergence (KL-divergence) between the distributions of staining segmentation by the teacher and grading network. This alignment in fine-grained information extraction across segmentation and grading tasks substantially enhances the model's robustness in complex real-world scenarios[49–52]. It is important to note that knowledge distillation is applied only during the training phase of our study. Once the training dataset has been quality-controlled to ensure high-quality, knowledge distillation can effectively transfer fine-grained features from the lesion detection task to the CSS grading task. However, given the high cost and imprecision of annotations, direct quantitative analysis poses challenges. Instead, we assess the impact through the improvement in grading performance. As shown in Table 3, the integration of segmentation knowledge distillation significantly enhances the FKD-CSS model's grading performance, allowing it to share latent features between segmentation and grading components and utilize

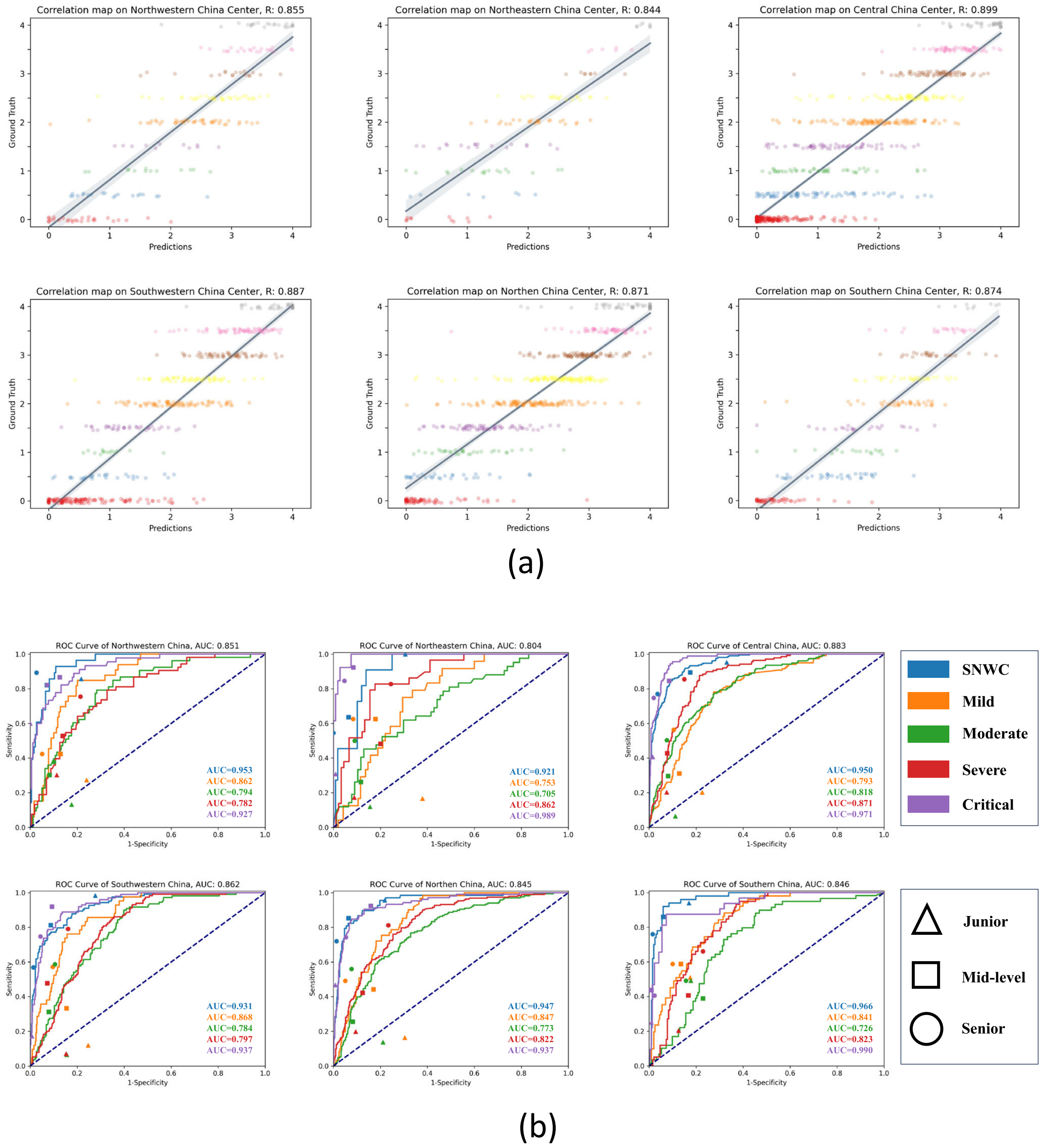

Fig. 5 | Performance of FKD-CSS in 6 external test datasets of dry eye. **a** Correlation map between the FKD-CSS model and clinical assessments (ground-truth) on six external test datasets. **b** AUC-ROC of the FKD-CSS model and different levels of ophthalmologists on six external validation datasets. FKD fine-grained knowledge distillation. CSS corneal staining score. AUC area under the curve. ROC receiver operating characteristic. SNWC staining negative or without clinical significance.

spatial information effectively. Additionally, we evaluated the segmentation performance on coarsely annotated images from external DE and MOSD datasets (descriptions are shown in Results section). The coarse nature of these annotations, often broader than actual punctate staining areas, introduces potential for false positives, leading us to choose Recall as our evaluation metric. Such visualizations do not merely delineate crucial diagnostic areas but also allow clinicians to visualize and comprehend the model's decision-making process, thereby augmenting both the transparency and interpretability. In contrast to many existing AI-assisted image reading methods, which are still often criticized for their 'black box' nature[44,53]. Although knowledge distillation has been explored in medical image analysis, this study introduces a method that opens new avenues for intelligent evaluation of multi-disciplinary medical conditions using pretext assessments like grading—not limited to ocular surface diseases—particularly addressing conditions requiring early diagnosis through subtle lesion detection.

Table 2 | Comparison of the performance of the proposed FKD-CSS model and different ophthalmologists on six external validation datasets

Ophthalmologist	Northern		Southern		Central		Northwestern		Southwestern		Northeastern		All	
	Pearson	AUC	Pearson	AUC	Pearson	AUC	Pearson	AUC	Pearson	AUC	Pearson	AUC	Pearson	AUC
Junior A	0.771	0.622	0.809	0.676	0.821	0.624	0.840	0.678	0.783	0.568	0.795	0.563	0.800	0.655
Junior B	0.752	0.578	0.801	0.714	0.733	0.682	0.820	0.644	0.661	0.650	0.784	0.511	0.729	0.608
Mid-level A	0.830	0.705	0.801	0.712	0.858	0.731	0.854	0.749	0.891	0.756	0.842	0.675	0.855	0.729
Mid-level B	0.835	0.725	0.824	0.680	0.876	0.746	0.849	0.723	0.900	0.756	0.797	0.671	0.864	0.733
Senior A	0.884	0.740	0.834	0.714	**0.914**	0.772	0.883	0.749	**0.915**	0.806	**0.881**	0.740	**0.898**	0.762
Senior B	**0.890**	0.730	0.840	0.691	0.867	0.749	**0.885**	0.753	0.859	0.742	0.802	0.620	0.865	0.734
FKD-CSS	0.871	**0.845**	**0.874**	**0.846**	0.899	**0.883**	0.855	**0.851**	0.887	**0.862**	0.844	**0.804**	0.882	**0.860**

The best results are highlighted in bold a and the second-best results are underlined. All *p*-value of Pearson test are statistically significant (*p*-value < 0.001).
FKD fine-grained knowledge distillation, *CSS* corneal staining score, *AUC* area under the curve.

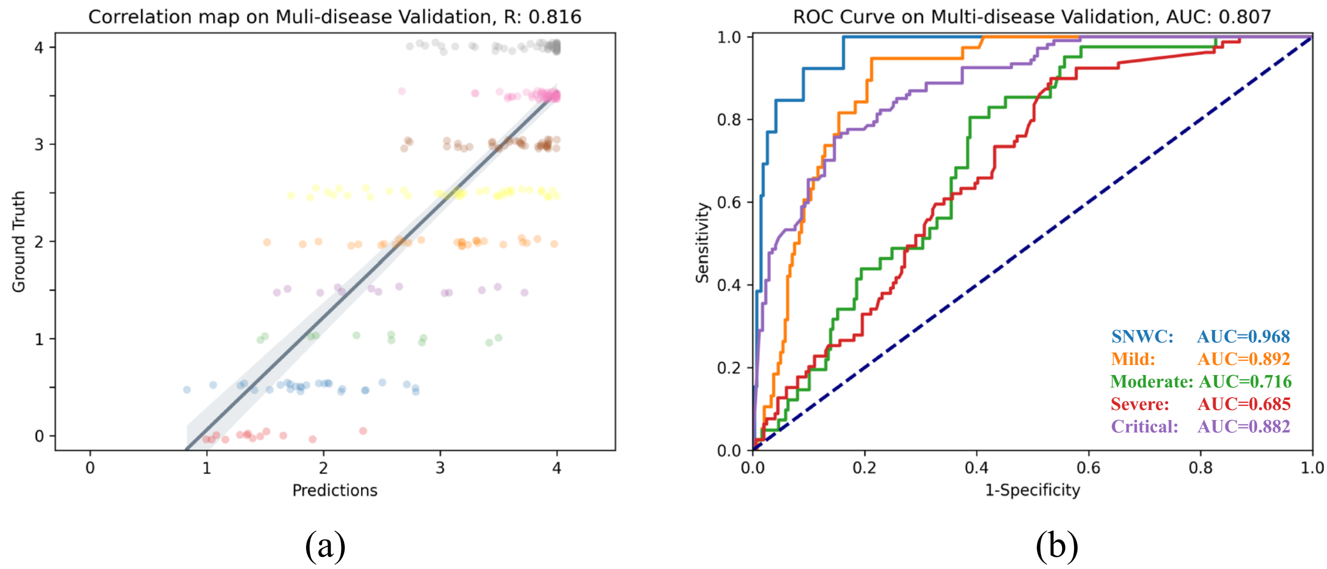

Fig. 6 | Performance of FKD-CSS in the external test dataset of MOSD. a The correlation map between the model and clinical assessments (ground-truth) (**b**) the ROC curve of the model compared with clinical assessments (ground-truth). FKD fine-grained knowledge distillation. CSS corneal staining score. AUC area under the curve, ROC receiver operating characteristic. MOSD multiple ocular surface disease. SNWC staining negative or without clinical significance.

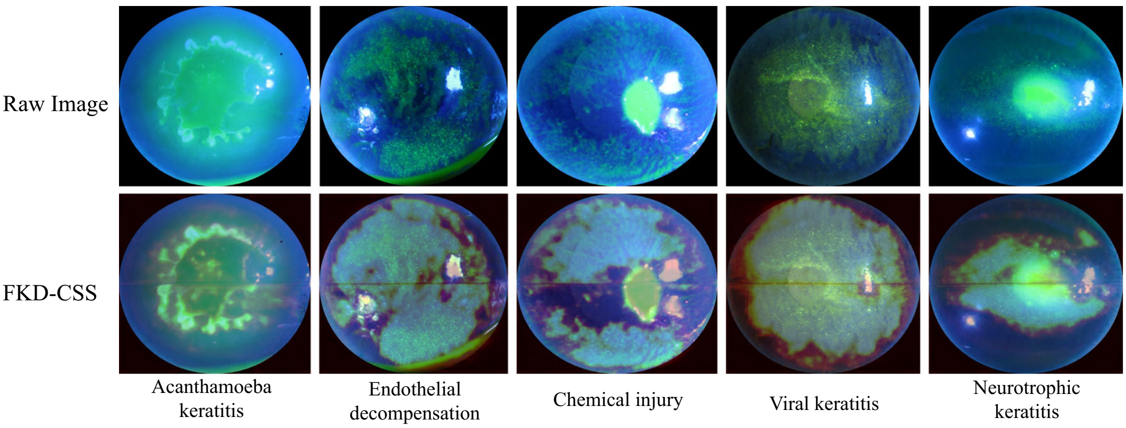

Fig. 7 | The visualization of fine-grained features of FKD-CSS model of randomly chosen samples in MOSD dataset. The first row is the original raw image, and the second row is the fined-grained lesion detection visualization from the proposed FKD-CSS. FKD fine-grained knowledge distillation. CSS corneal staining score. MOSD multiple ocular surface disease.

Table 3 | Ablation study of the proposed FKD-CSS model on the internal dataset cross-validation

ROI extraction	CPMB	Knowledge distillation	Pearson	AUC
			0.828	0.803
✓			0.865	0.848
✓	✓		0.889	0.871
✓		✓	0.883	0.861
✓	✓	✓	0.898	0.881

The first row represents the baseline model (regression ResNet50). All *p*-value of Pearson test are statistically significant (*p*-value < 0.001).
FKD fine-grained knowledge distillation, *CSS* corneal staining score, *ROI* region of interest, *AUC* area under the curve, *CPMB* CSS prototype memory bank.

Given that the existing DL-assisted models for corneal staining focus solely on discrete classification tasks, assigning fixed categorical grades (e.g., mild, moderate, or severe) to corneal staining samples[14,54], which are relatively less sensitive in detecting clinically significant subtle changes, especially in treatment response[55]. The continuous regression grading approach of the FKD-CSS model aligns more effectively with clinical assessments, offering a nuanced measurement of the relative distances between different scores. For instance, in a regression context, the error between scores of 3.5 and 1.0 is considered larger than that between 2.0 and 1.0, whereas a categorization approach would treat these differences as equivalent. This nuanced understanding is crucial in clinical settings where subtle variations in severity can have significant implications. However, current deep learning methods often face challenges in regression tasks, primarily due to the imbalanced distribution of data[56]. To counter this, the FKD-CSS introduces a memory bank that stores key features for each discrete category, effectively bridging the gap between regression and classification tasks. This memory bank allows for a balanced approach, harnessing the strengths of both methodologies to enhance overall performance and accuracy in grading. As a result, the continuous CSS output of FKD-CSS presented considerable consistency with expert team, which will assist in clinical monitoring of disease progression and therapeutic response, and serve as a sensitive endpoint for treatment trials. On the other hand, the continuous score of FKD-CSS can transfer to categories by employing reasonable threshold, which can correspond to various existing scoring methods of corneal staining (e.g., Oxford scale[24], Baylor scale[57], and NEI/Industry scale[58]), and can flexibly adapt to the clinical needs of severity assessment of disease.

The ultimate goal of medical AI is to serve clinical needs, necessitating thorough testing and validation in real-world settings before deployment[59]. Previous studies have showed only moderate intra- and interrater agreement in manually grading punctate staining[26], with correlation coefficients ranging from 0.65 to 0.961[16,54,60] and kappa values between 0.417 and 0.463[26,61]. This variability stems from difficulties in counting fine lesions and ambiguities in defining adjacent CSS grades, inducing subjectivity in clinician evaluations. Our study also found relatively low correlation values for senior doctors' independent scoring compared to a consensus ground truth. Therefore, AI-assisted CSS offers significant advantages in corneal damage assessment due to its stability and repeatability. To date, only two studies have utilized deep learning to evaluate corneal staining[14,54]. One focuses only on punctate erosions with limited interpretability, validating on a single internal test set (153 images)[14], while the other employed external testing from a single center (93 image)[54]. Such single-center, small-scale datasets pose risks of overfitting and limited generalizability[62,63]. In contrast, our model's primary testing datasets were sourced from a multicenter clinical trial across China, encompassing 37 hospitals in 22 provinces from all seven major regions, offering an in-depth comparison across different camera types, slit-lamp platform platforms and regions with different population distribution, ensuring a realistic evaluation of performance across diverse settings. It is important to recognize that there are no absolutely precise annotations for CSS, which complicates performance evaluation of all existing CSS approaches[24,34,58,64]. This study aims to establish a reference close to the true ground truth based on consensus among senior doctors. If our automated approach matches or surpasses the consistency and accuracy of experienced ophthalmologists, it is commendable. The FKD-CSS model exhibited comparable performance (*r*:0.844–0.899, AUC:0.804–0.883) to experienced ocular surface experts (*r*:0.802-0.915, AUC:0.620–0.806), which is also excellent when compared to currently reported correlation of existing DL-assisted methods with human grading in single-center (*r*:0.863)[54]. This automation alleviates the workload on doctors and eases the burden on remote hospitals lacking specialists. The stable and consistent scoring facilitates long-term patient tracking, enabling better assessment of recovery processes and the evaluation of clinical drugs in both dry eye and various ocular surface diseases. In addition, differences in model performance across various centers exist, and future research should explore how demographic characteristics and other confounding variables influence model performance to optimize the model's real-world application. Furthermore, this study designs an interactive data visualization tool to present raw data from large-scale, multicenter research (Supplementary Software 1), illustrating variations in data distribution and grading consistency across physicians, facilitating a comprehensive evaluation of the model's performance in complex environments.

Unlike human grading, AI-based evaluation is not bound to a specific scale or disease, but simply evaluates predefined affected areas[26]. The focus on the inferior one-third of the cornea aligns with existing dry eye staining assessment methods, specifically the inferior cornea staining score (ICSS), which has been a primary efficacy measure in various clinical trials. Previous studies indicate that staining in the inferior cornea is the most severe and typical sign of dry eye[34–36] and shows significant improvement after treatment[39,40]. Moreover, limited zone assessment is highly efficient and avoids the plateau effect caused by weighted calculations of staining scores in other regions, making it ideal for evaluating severity and treatment effectiveness in DE. Nevertheless, the vulnerable areas of the corneal vary across different diseases[6,7,39,65], and evaluating the entire cornea may provide more comprehensive pathological information. The ability to switch between the ROI within the partial cornea area and the entire cornea surface is crucial for digital CSS model's generalizability across different ocular surface abnormalities. We tested the FKD-CSS model externally on a MOSD dataset, which includes multiple ocular surface diseases with punctate and patchy lesions. The model performed well, consistently categorizing and scoring corneal staining, correctly recognizing 78.5% of staining areas across the different staining patterns. These results These results indicate that the techniques can be applied beyond dry eye and punctate lesions to various corneal staining images across multiple diseases. Despite these advantages, the robustness of AI algorithms in complex clinical scenarios is still a challenge, as misidentifying key staining areas or subtle corneal changes could lead to incorrect diagnoses. Traditional clinical grading methods often struggle with providing reliable confidence measures, as deep regression approaches lack uncertainty quantification, and classification methods suffer from overconfidence due to reliance on the cross-entropy loss[54,66,67]. Weakly supervised methods with coarse annotations can indeed introduce risks due to the absence of fine-grained labels. Our approach calculates confidence scores based on the distance between the image of interest and prototypes (Eq. 3), reducing the risks of false negatives and positives through knowledge distillation and prototype-based confidence estimation. This allows for reliable anomaly alerts for manual review when necessary. Such uncertainty analysis has already been widely used in fields like autonomous driving[68,69]. Additionally, the AI model offers practical benefits, built on the widely-used ResNet50 framework, known for easy implementation across various platforms and low computational resource requirements. It operates without an internet connection, making it suitable for deployment in diverse medical and research settings, even in areas with limited connectivity, thus facilitating its real-world application.

This study has the following limitations. Firstly, the model's ability in specific grading of the severities performed the worst in the moderate category in most of the external test sets, with subpar performance in the adjacent mild and severe categories as well. This may be attributed to the semantic ambiguity in the image reading criteria used as a reference for expert decision-making, which the model learned from. Specifically, the distinction between the moderate and the mild category defined by image reading criteria is whether the lesions are countable, while the distinction from the severe category is whether it is merged[34]. These definitions are prone to subjective variability in image reading, which is an inherent limitation of all the existing corneal staining scoring criteria[24,34,58,65]. Consequently, the accuracy of model performance evaluation is constrained by this limitation. Secondly, it remains a significant challenge to simultaneously consider the model's generalization performance and accuracy in specific. The real-world settings of this study mainly focus on DE while the model preliminary demonstrates promising generalization in adapting to multiple ocular surface diseases, it is recommended to expand the training and testing samples as much as possible to enhance the model's capability before applying it to specific ocular surface diseases other than DE. Thirdly, this study only explored the Asian population, subsequent research efforts could aim to expand the range of ethnic diversity scenarios to improve the model's generalization capacity. Moreover, the database included in this study only from tertiary hospitals. The main challenge is the lack of imaging resources in primary or community hospitals and the difficulty in data quality control. Lastly, the proposed AI model still heavily relies on the training distribution and is unable to handle data from out-of-control healthcare systems, which has the potential to produce unforeseen and inexplicable outcomes. Moving forward, our plan is to incorporate robust abnormal analysis and uncertainty estimation methods into the FKD-CSS framework. It is worth mentioning that unconventional corneal staining such as adhesion of secretions or filaments on the surface of the cornea, and noticeable ocular surface irregularities lead to dye accumulation, such staining was not in the application scenario of this study since it is uncommon and lacks clear clinical implications. For future work, we plan to extend our analysis to include these out-of-distribution (OOD) data. By incorporating a broader range of corneal staining types, we aim to enhance the model's robustness and generalizability. This will allow us to better understand and address the variability in clinical presentations and improve the model's applicability to a wider array of ocular surface conditions.

Overall, this study proposes a knowledge distillation strategy for fine-grained corneal staining score assessment, addressing the gap in the identifying fined-grained lesions with interpretability. The model also employs flexible output for both continuous scores and severity grading for treatment efficacy assessment. The performance of the FKD-CSS model has been validated in real-world settings of DE, which demonstrated the potential in facilitating standardized and precise medical care and supporting the conduct of multicenter clinical trials of DE, and it also exhibits promising generalizability across diverse ocular surface diseases. This work represents a foundational step, to further support analysis of correlations between feature matrices, pathological information, and final scores by linking automatic scoring with fine-grained features in specific condition to provide a new direction for quantifying clinical evaluations.

Methods

Ethics approvals

This study was approved by the medical ethics committee of Zhongshan Ophthalmic Centre, Sun Yat-sen University, China (protocol number: 2020YWPJ001). All procedures were conducted following the Declaration of Helsinki (1983). Informed consent was obtained from each of the participants prior to data collection.

Datasets

The images of developing, internal validation and primary external test datasets were mainly collected prospectively from a DE dataset of a multicenter phase III study for cyclosporine-A eye drops (ClinicalTrials.gov; identifier: NCT04541888), involving 37 tertiary hospitals from seven regions of China (Fig. 1). The dataset consists of one corneal staining image per eye of each follow-up of DE patients who tested positive for epithelial staining at baseline between November 2020 and October 2021. The follow-up visits including baseline, improvement or exacerbation after treatment, and also from subjects who have recovered to corneal staining negativity after treatment, encompasses corneal abundant punctate, discrete, and coalescent patchy epithelial lesions. For eyes that completely returned to normal after treatment, we included only one staining-negative image to avoid duplication. Images from different eyes of the same patient, taken at various time points, can be deemed independent since the sampling intervals (≥2 weeks) exceed the epithelial cell turnover time (≤1 week), which is sufficient for significant morphological changes in corneal epithelial lesions[70]. The multicenter operators had all received standard training before capturing corneal fluorescein staining images. Additionally, the generalizability of FKD-CSS was tested on an independent annotated corneal staining image set of multiple ocular surface diseases (MOSD) without DE from Zhongshan ophthalmic center. There was no overlap between patients in the development and all test sets.

Development and internal validation datasets

All images from Eastern China of the phase III study dataset were enrolled for development and internal validation of FKD-CSS model, resulting in 1,477 corneal fluorescein staining images of 299 DE patients from 14 tertiary hospitals in seven provinces and municipalities. The internal validation was conducted via 5-fold cross-validation, with 30% of the training data in each fold used for hyper-parameter selection. The prediction for the external validation is ensemble averaging across five models trained during the cross-validation process.

Specifically, to establish a correlation between CSS grading and staining lesions, we randomly selected 100 DE images with CSS grades 1-4 from the development dataset (25 images per grade). These images were coarsely annotated for staining lesions by senior ophthalmologists. Integrating these fine-grained but imprecise annotations allows the FKD-CSS framework to effectively transfer prior knowledge about the correlation between staining grades and lesion locations to a larger and more generalized dataset using knowledge distillation. Then we conducted extensive ablation studies to analyze the contribution of key architectural components in FKD-CSS model and compared its performance and interpretability with representative deep learning methods in internal validations.

External test datasets for Dry eye

The proposed model was tested in six external independent sets from 6 regions of China (Central, South, North, Northeast, Northwest, and Southwest China). Each external set involved 2–7 tertiary hospitals, yielding a total of 2386 photographs of 476 DE patients from 23 hospitals in 15 provinces and municipalities. Each dataset contains all ranges of CSS.

External test for multiple ocular surface diseases

To assess the generalizability of FKD-CSS, 231 corneal staining images of 231 patients with multiple ocular surface diseases from Zhongshan ophthalmic center were collected retrospectively (between October 2019 and November 2023). The inclusion criteria focused on patients with corneal epithelial injuries from any cause except DE. The exclusion criteria included factors which could lead to false-positive sodium fluorescein staining, such as adhesion of secretions or filaments on the surface of the cornea, noticeable ocular surface irregularities that can lead to significant dye accumulation, and a condition that the investigator felt may have confounded the study results. Notably, the FKD-CSS model was initially trained on DE cases, focusing primarily on the lower third of the cornea. However, multiple ocular surface diseases affect the entire corneal area. To address this discrepancy, we divided each image into upper and lower sub-images and independently predicted their CSS. The final CSS for each image of MOSD dataset was then determined by selecting the higher value from either the upper or lower corneal area.

Reading process of corneal staining images

All images were assigned to three senior ophthalmologists who graded each image three times. If there was a dispute, they discussed and reached consensus. If the reading group encountered difficult cases, they could apply for expert arbitration done by two ocular surface experts. The principle of CSS was scored by evaluation of distribution and density of the punctate staining[34] as follows: no staining = 0, few/rare punctate lesions = 1, discrete and countable lesions = 2, lesions too numerous to count, but not coalescent = 3, coalescent = 4. The manual scoring label was required to be accurate to 0.5 points according to a precise evaluation in practice. 228 corneal staining images were randomly re-read to check the grading consistency. The total eligibility rate of spot-check was 84.21%. In order to establish a detailed correlation between CSS and staining lesions, we randomly selected 100 DE images with CSS ranging from 1 to 4 from the development dataset, 25 images for each grade. The staining lesions of these images were then coarsely annotated by the team of 3 experienced doctors using the PAIR® annotation software package (https://www.aipair.com.cn/en/, Version 2.7, RayShape, Shenzhen, China). It is worth noting that the annotation process for punctate lesions is both expensive and prone to inaccuracies. Consequently, there is a notable limitation in the quantity and quality of these lesion annotations. However, the integration of these fine-grained yet imprecise annotations enables the proposed FKD-CSS framework to effectively transmit prior knowledge regarding the correlation between staining grades and lesion locations to a more expansive and generalized dataset by using knowledge distillation. Since all of those patients are DE, corneal staining was annotated and scored from the 1/3 inferior area of cornea, which has been demonstrated to present the highest degree of corneal staining in symptomatic DE patients most commonly[39,40].

Algorithm construction of FKD-CSS

The proposed FKD-CSS framework consists of three stages (Fig. 2). Initially, a corneal segmentation network is utilized to extract the inferior corneal region, followed by a quality control step to exclude low-quality images with indistinct corneal regions to ensure that each image has clearly visible corneal boundaries and consistent, uniform illumination. Subsequently, a weakly-supervised teacher segmentation network is introduced in the second stage to guide the CSS grading model towards the detection of detailed but coarsely-grained lesions. Additionally, a prototype-based stratified screening mechanism was incorporated to enable self-correction and uncertainty estimation for misclassified samples, ensuring the model's reliability and suitability for real-world clinical applications.

We trained a U-Net on the SUSTech-SYSU dataset[12] utilizing a pretrained ResNet50 backbone on ImageNet for automatic corneal segmentation. To maintain topological accuracy, we applied ellipse fitting as a post-processing technique. However, it is essential to highlight that the ellipse fitting process may encounter challenges with low-quality images that have unclear corneal boundaries or uneven illumination. To address this, we implemented an automatic exclusion strategy to ensure quality control. As a result of this quality control step, we retained 1471 out of 1477 images for both training and 5-fold cross-validation, and 2376 out of 2386 images for external test of DE, 231 images for external test of MOSD purposes. The training-validation ratio in cross-validation is set at 7:3 in each fold.

CSS grading depends heavily on the accurate identification of punctate staining areas, yet existing DL methods often fall short due to their inability to extract such fine-grained information during training. To address this difficulty, we introduced a weakly-supervised segmentation teacher network utilizing a limited number of coarse lesion segmentation annotations. Specifically, we manually annotated 100 images from the developing dataset with coarse annotations, as mentioned in the previous section. We then utilized a U-Net[71] with a ResNet50 backbone, pre-trained from our previous corneal segmentation task. The teacher network's primary objective is to infuse spatial staining information into the subsequent student grading model, thereby enriching the automated grading process with robust explanations, as will be detailed in the following paragraphs. By leveraging weak segmentation supervision alongside spatial context, our proposed approach aims to enhance the accuracy and reliability of automated CSS grading by while being cost-effective and efficient, requiring only a few coarse annotations.

The grading network approaches the CSS grading task as a regression problem, generating continuous predictions that can be easily converted into discrete grades by applying a specified threshold. To optimize this regression task, the network aims to minimize the L1 loss function, which is defined in Eq. (1):

$$\mathcal{L}_{\text{score}} = \frac{1}{B}\sum_{i=1}^{B} || \hat{y}_i - y_i ||, \tag{1}$$

where B is the batch size, $\hat{y}_i$ denotes the predicted grade, and y_i is the ground-truth CSS grade. As discussed earlier, most of the existing methods lack explainability and do not yield interpretable results and fine-grained highlights of staining areas. In response to this issue, we adopt a novel approach of utilizing intermediate features for pretext tasks, specifically referred to as interpretable segmentation distillation. Our method employs a straightforward knowledge distillation technique, drawing insights from the pretrained weakly supervised segmentation network, as shown in Eq. (2):

$$\mathcal{L}_{\text{dist}} = \frac{1}{B}\sum_{i=1}^{B} \text{KL}\left(\text{s}_\text{i}^\text{s}, \text{s}_\text{i}^\text{t}\right), \tag{2}$$

where s_i^t is the pseudo soft label generated by the teacher segmentation network, s_i^s denotes the segmentation probabilities based on the features from the CSS grading encoder, employing a U-Net segmentation head. The KL-divergence is used to measure the difference between these two probability distributions. This segmentation knowledge distillation task effectively aligns the CSS grading features with staining locations. Furthermore, the segmentation branch also provides interpretable results, making it valuable for clinical applications. In accordance with the CSS grading standard, annotations are divided into nine discrete classes, ranging from 0 to 4 in increments of 0.5. We operate under the assumption that subjects in each category should display similar levels of clinical symptoms. To capitalize on this assumption, we propose the integration of a CSS Prototype Memory Bank (CPMB). The CPMB is specifically designed to group similar semantic information from predicted successive grades, which is able to cluster facilitates more effective knowledge representation and retrieval, enabling the system to draw upon a rich database of categorized symptoms and corresponding grades. To employ it, we consider the InfoNCE loss[20] to align the similar samples with the same category in Eq. (3):

$$\mathcal{L}_{\text{proto}} = -\frac{1}{\text{B}}\sum\nolimits_{\text{i}=1}^{\text{B}} \log\frac{\exp\left(\text{f}_\text{i}\cdot \text{p}_{\text{y}_\text{i}}/\tau\right)}{\sum\nolimits_{\text{j}=1}^{\text{C}}\exp\left(\text{f}_\text{i}\cdot \text{p}_\text{j}/\tau\right)}, \tag{3}$$

where f_i represents the feature vector of i-th sample, and p_j denotes the j-th prototype in CPMB, C is the number of classes in CPMB (which is set as 9 in this study) and τ is the temperature parameter. We employ a moving average (MA) strategy to update the clustering center in CPMB. Given the i-th sample labeled with y_i, the MA process is formulated in Eq. (4):

$$p_{y_i} = \alpha p_{y_i} + (1-\alpha) f_i, \tag{4}$$

where α represents the momentum coefficient. The overall objective function is given by Eq. (5):

$$\mathcal{L} = \mathcal{L}_{\text{score}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist.}} \tag{5}$$

The hyper-parameters λ_{proto} and λ_{dist} are configured as 0.5 and 0.05, respectively. We set the initial learning rate to 0.0001 and utilize the cosine decay scheduler for learning rate adjustment. The student grade network is initialized using the parameters from the teacher network encoder and is trained for a total of 50 epochs. A batch size of 64 is employed during training. For distillation, the temperature parameter τ is set to 0.1 and the momentum coefficient α is set to 0.999. The proposed framework is implemented using PyTorch and is trained on a single NVIDIA Titan RTX GPU. We plan to make the code publicly available once the paper is accepted.

Ablation study

To evaluate the specific impact of each component within the FKD-CSS model, we undertook an ablation study utilizing our internal validation set. The quantitative outcomes of this study, detailing various combinations of three critical components, are presented in Table 3. The "ROI extraction" component refers to the preprocessing step that crops the ROI areas, focusing on the 1/3 inferior area of the cornea for Dry Eye cases and the entire cornea for other conditions. The "CPMB" stands for the CSS Prototype Memory Bank training strategy, mathematically denoted as $\mathcal{L}_{\text{proto}}$. Lastly, "Knowledge distillation" indicates the fine-grained segmentation training strategy, represented as $\mathcal{L}_{\text{dist}}$.

Compared to the performance of the baseline ($r = 0.828$, AUC = 0.803), the successive inclusion of ROI extraction, Knowledge distillation, and CPMB can elevate the algorithm's r to 0.865, 0.883, and 0.889, and increased the AUCs to 0.848, 0.861, and 0.871, respectively. When combine all three components, the performance of the model was the best ($r = 0.898$, AUC = 0.881) (Table 3). The findings from this study clearly demonstrate the critical role played by each component in ensuring accurate CSS grading. Notably, the removal of any single component from the model led to a significant decline in its performance.

Statistical analysis

All statistical analyses were conducted in the SPSS version 24.0 software (IBM, Armonk, New York). Since the output from FKD-CSS initially consisted of continuous scores that could be converted into these discrete categories, we employed two statistical metrics to evaluate the regression consistency and categorization performance, namely correlation coefficient and the AUC-ROC. Normality of the continuous CSS scores was confirmed via D'Agostino-Pearson test ($p > 0.05$). We evaluated the correlation between clinical grading scores and automated grading scores by Pearson test, r was utilized to determine the correlation coefficients. AUC-ROCs were computed by using the absolute error between the predicted grades and the ground truth as the threshold. ROC curves were generated by sweeping error thresholds from 0 (perfect match) to maximum observed error, calculating true/false positive rates at each grade. The overall AUC was derived from a clinical prevalence-weighted average of all five grade-specific AUCs. All tests were two-tailed, $p < 0.001$ was considered statistically significant.

Comparison with ophthalmologists

Six clinicians at an academic ophthalmology center, with varying levels of clinical experience, were assigned to grade the severity of corneal staining in the validation and test images. Their experience ranged as follows: two had less than 5 years of experience, two had between 5 and 10 years, and two had more than 10 years. All of these clinicians were instructed to make a referral grading decision on each image in development & internal validation datasets, and six external test datasets.

Staining detection test

To assess the fine-grained capability of our model, we performed a staining detection testing by randomly selected 80 images from the external DE test dataset and 80 images from the MOSD test dataset. Notably, the assessment for DE dataset was focusing primarily on the lower third of the cornea, which has been demonstrated to be the most predictive key zone for DE severity, while evaluation was focused on the entire corneal area in dataset of multiple ocular surface diseases.

We used the segmentation branch integral to the knowledge distillation process for feature visualization. This branch shares latent features with the grading network, meaning that its segmentation outputs are indicative of the area most relevant to CSS grading. We harness these outputs to create a probability map, which is then employed as a heatmap for detecting staining areas. This approach allows for a visual representation of the model's focus areas, thereby highlighting its precision in identifying key regions relevant to CSS grading. It is worth noting the inherent challenge of annotating point staining, where the ground truth area is typically larger than discrete staining points. Therefore, we use the recall of the segmentation results as a metric to evaluate the effectiveness of our model in detecting staining.

Uncertainty estimation

To further mitigate potential risks of misdiagnosis of the FKD-CSS, our algorithm incorporated a novel confidence estimation mechanism. By leveraging prototype distances (Eq. 3), it calculated confidence scores to gauge uncertainty. Testing on an external validation set showed a statistically significant confidence gap ($p < 1e^{-10}$) between correctly graded (error < 0.5) and misgraded samples (error ≥ 0.5). With a confidence threshold of 0.27, the model detected 97.77% of misgraded samples, flagging low-confidence cases for manual review.

This confidence estimation mechanism enhances reliability in clinical AI applications, ensuring safe and effective large-scale screenings. By achieving performance comparable to senior clinicians while integrating safeguards for patient safety, our FKD-CSS model fulfills the demands of clinical use with reduced risk.

Data availability

The datasets utilized in this study are not publicly available due to strict privacy and security concerns. The data, derived from corneal staining assessments across multiple medical centers, are governed by confidentiality agreements and cannot be shared with individuals outside of the Institutional Review Board-approved research collaborations. However, information on the datasets and potential access for legitimate research inquiries can be provided upon reasonable request, subject to appropriate review and approval processes.

Code availability

The code utilized for dataset preparation, statistical modeling, and evaluation of the findings in this study is available in our GitHub repository (https://github.com/QtacierP/FKD-CSS).

Received: 16 April 2024; Accepted: 5 May 2025;
Published online: 23 May 2025

References

1. DelMonte, D. W. & Kim, T. Anatomy and physiology of the cornea. *J. Cataract Refract Surg.* **37**, 588–598 (2011).
2. Massoudi, D., Malecaze, F. & Galiacy, S. D. Collagens and proteoglycans of the cornea: importance in transparency and visual disorders. *Cell Tissue Res.* **363**, 337–349 (2016).
3. Kim, M., Chun, Y. S. & Kim, K. W. Different perception of dry eye symptoms between patients with and without primary Sjogren's syndrome. *Sci. Rep.* **12**, 2172 (2022).
4. Lee, D., Lee, G. W. & Yoon, S. H. Relationship between ocular surface temperature and 0.1% cyclosporine a in dry eye syndrome with meibomian gland dysfunction. *PLoS One* **18**, e0293472 (2023).
5. Stapleton, F. et al. CLEAR - Contact lens complications. *Cont. Lens Anterior Eye* **44**, 330–367 (2021).
6. Dohlman, C. H. The function of the corneal epithelium in health and disease. The Jonas S. Friedenwald Memorial Lecture. *Investig. Ophthalmol.* **10**, 383–407 (1971).
7. Han, S. B., Yang, H. K. & Hyon, J. Y. Influence of diabetes mellitus on anterior segment of the eye. *Clin. Inter. Aging* **14**, 53–63 (2019).
8. Yang, A. Y., Chow, J. & Liu, J. Corneal innervation and sensation: the eye and beyond. *Yale J. Biol. Med.* **91**, 13–21 (2018).
9. Kim, J. The use of vital dyes in corneal disease. *Curr. Opin. Ophthalmol.* **11**, 241–247 (2000).

10. Bron, A., Argüeso, P., Irkec, M. & Bright, F. Clinical staining of the ocular surface: mechanisms and interpretations. *Prog. Retin. Eye Res.* **44**, 36–61 (2015).
11. Begley, C. et al. Review and analysis of grading scales for ocular surface staining. *Ocul. Surf.* **17**, 208–220 (2019).
12. Deng, L. et al. The SUSTech-SYSU dataset for automatically segmenting and classifying corneal ulcers. *Sci. Data* **7**, 23 (2020).
13. Wang, S. et al. AES-CSFS: an automatic evaluation system for corneal sodium fluorescein staining based on deep learning. *Ther. Adv. Chronic Dis.* **14**, 20406223221148266 (2023).
14. Qu, J. H. et al. Fully automated grading system for the evaluation of punctate epithelial erosions using deep neural networks. *Br. J. Ophthalmol.* **107**, 453–460 (2023).
15. Peterson, R. C. & Wolffsohn, J. S. Objective grading of the anterior eye. *Optom. Vis. Sci.* **86**, 273–278 (2009).
16. Rodriguez, J. D. et al. Automated grading system for evaluation of superficial punctate keratitis associated with dry eye. *Investig. Ophthalmol. Vis. Sci.* **56**, 2340–2347 (2015).
17. Chun, Y. S., Yoon, W. B., Kim, K. G. & Park, I. K. Objective assessment of corneal staining using digital image analysis. *Investig. Ophthalmol. Vis. Sci.* **55**, 7896–7903 (2014).
18. Bunya, V. Y. et al. Development and evaluation of semiautomated quantification of Lissamine green staining of the bulbar conjunctiva from digital images. *JAMA Ophthalmol.* **135**, 1078–1085 (2017).
19. Wolffsohn, J. S. et al. TFOS DEWS II diagnostic methodology report. *Ocul. Surf.* **15**, 539–574 (2017).
20. Oord, A. V. D., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. https://doi.org/10.48550/arXiv.1807.03748 (2018).
21. Wang, Z., Lyu, J., Luo, W. & Tang, X. Adjacent scale fusion and corneal position embedding for corneal ulcer segmentation. In *Ophthalmic Medical Image Analysis: 8th International Workshop, OMIA 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France*, *September 27, 2021, Proceedings*, 1–10 (Springer, 2021).
22. Stapleton, F. et al. TFOS DEWS II epidemiology report. *Ocul. Surf.* **15**, 334–365 (2017).
23. Hill, G. M., Ku, E. S. & Dwarakanathan, S. Herpes simplex keratitis. *Dis. Mon.* **60**, 239–246 (2014).
24. Bron, A. J., Evans, V. E. & Smith, J. A. Grading of corneal and conjunctival staining in the context of other dry eye tests. *Cornea* **22**, 640–650 (2003).
25. Rasmussen, A. et al. Reproducibility of ocular surface staining in the assessment of sjogren syndrome-related keratoconjunctivitis sicca: implications on disease classification. *ACR Open Rheumatol.* **1**, 292–302 (2019).
26. Kourukmas, R., Roth, M. & Geerling, G. Automated vs. human evaluation of corneal staining. *Graefes Arch. Clin. Exp. Ophthalmol.* **260**, 2605–2612 (2022).
27. Dong, C., Liu, L. & Shang, J. Label noise in adversarial training: a novel perspective to study robust overfitting. *Adv. Neural Inf. Process. Syst.* **35**, 17556–17567 (2022).
28. Tran, T., Vu, H., Carneiro, G. & Bui, H. Bayesian metric learning for robust training of deep models under noisy labels. *Preprint at* https://openreview.net/forum?id=uRuGNovS11 (2020).
29. Li Y. et al. Learning from noisy labels with distillation. In *Proc. of the IEEE International Conference on Computer Vision*, 1910–1918 (IEEE, 2017).
30. Craig, J. P. et al. TFOS DEWS II report executive summary. *Ocul. Surf.* **15**, 802–812 (2017).
31. Narayanan, S. et al. The diagnosis and characteristics of moderate dry eye in non-contact lens wearers. *Eye Contact Lens* **31**, 96–104 (2005).
32. Baudouin, C. et al. Diagnosing the severity of dry eye: a clear and practical algorithm. *Br. J. Ophthalmol.* **98**, 1168–1176 (2014).
33. The definition and classification of dry eye disease: report of the definition and classification subcommittee of the International Dry Eye Workshop (2007). *Ocul. Surf.* **5**, 75-92 (2007).
34. Holland, E. J. et al. Lifitegrast clinical efficacy for treatment of signs and symptoms of dry eye disease across three randomized controlled trials. *Curr. Med. Res. Opin.* **32**, 1759–1765 (2016).
35. Sheppard, J. D. et al. Lifitegrast ophthalmic solution 5.0% for treatment of dry eye disease: results of the OPUS-1 phase 3 study. *Ophthalmology* **121**, 475–483 (2014).
36. Peng, W. et al. Cyclosporine A (0.05%) ophthalmic gel in the treatment of dry eye disease: a multicenter, randomized, double-masked, phase III, COSMO trial. *Drug Des. Dev. Ther.* **16**, 3183–3194 (2022).
37. Tauber, J. et al. Lifitegrast ophthalmic solution 5.0% versus placebo for treatment of dry eye disease: results of the randomized phase III OPUS-2 study. *Ophthalmology* **122**, 2423–2431 (2015).
38. Semba, C. P. et al. A phase 2 randomized, double-masked, placebo-controlled study of a novel integrin antagonist (SAR 1118) for the treatment of dry eye. *Am. J. Ophthalmol.* **153**, 1050–1060.e1051 (2012).
39. Fenner, B. J. & Tong, L. Corneal staining characteristics in limited zones compared with whole cornea documentation for the detection of dry eye subtypes. *Investig. Ophthalmol. Vis. Sci.* **54**, 8013–8019 (2013).
40. Woods, J., Hutchings, N., Srinivasan, S. & Jones, L. Geographic distribution of corneal staining in symptomatic dry eye. *Ocul. Surf.* **18**, 258–266 (2020).
41. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 1492–1500 (IEEE, 2017).
42. Tabrizchi, H., Mosavi, A., Vamossy, Z. & Varkonyi-Koczy, A. R. Densely connected convolutional networks (DenseNet) for diagnosing coronavirus disease (COVID-19) from chest X-ray imaging. In *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 1–5 (IEEE, 2021).
43. Zhang, H. et al. Resnest: Split-attention networks. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition.* 2736–2746 (IEEE, 2021).
44. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision*, 839–847 (IEEE, 2018).
45. Goyal, A. & Bengio, Y. Inductive biases for deep learning of higher-level cognition. *Proc. R. Soc. A* **478**, 20210068 (2022).
46. Tay, Y. et al. Scaling laws vs model architectures: How does inductive bias influence scaling? In *The 2023 Conference on Empirical Methods in Natural Language Processing.* https://doi.org/10.48550/arXiv.2207.10551 (2022).
47. Chen, Z., Tian, Z., Zhu, J., Li, C. & Du, S. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11676-11685 (IEEE, 2022).
48. Fries, J. A. et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nat. Commun.* **10**, 3111 (2019).
49. Klingner, M. et al. X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13343-13353 (IEEE, 2023).
50. Li, Y. et al. Proceedings of the IEEE/CVF international conference on computer vision. in *Proc. of the IEEE/CVF International Conference on Computer Vision*. 6715-6724.
51. Yuan, J., Phan, M. H., Liu, L. & Liu, Y. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 595–605.
52. Liu, Y. et al. Generic perceptual loss for modeling structured output dependencies. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*. 2604-2613 (IEEE, 2021).
53. Chan, K. H. R. et al. ReduNet: A white-box deep network from the principle of maximizing rate reduction. *J. Mach. Lear Res*. **23**, 4907–5009 (2022).

54. Kim, S. et al. Deep learning-based fully automated grading system for dry eye disease severity. *Plos one* **19**, e0299776 (2024).
55. Novack, G. D. et al. TFOS DEWS II clinical trial design report. *Ocul. Surf.* **15**, 629–649 (2017).
56. Zhang, S., Yang, L., Mi, M. B., Zheng, X. & Yao, A. Improving Deep Regression with Ordinal Entropy. In *The Eleventh International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.2301.08915 (2023).
57. De Paiva, C. S. & Pflugfelder, S. C. Corneal epitheliopathy of dry eye induces hyperesthesia to mechanical air jet stimulation. *Am. J. Ophthalmol.* **137**, 109–115 (2004).
58. Lemp, M. A. Report of the National Eye Institute/Industry workshop on Clinical Trials in Dry Eyes. *CLAO J.* **21**, 221–232 (1995).
59. Lin, D. et al. Application of Comprehensive Artificial Intelligence Retinal Expert (CARE) system: a national real-world evidence study. *Lancet Digit Health* **3**, e486–e495 (2021).
60. Amparo, F., Wang, H., Yin, J., Marmalidou, A. & Dana, R. Evaluating corneal fluorescein staining using a novel automated method. *Investig. Ophthalmol. Vis. Sci.* **58**, BIO168–BIO173 (2017).
61. Danis, R. P. et al. Methods and reproducibility of grading optimized digital color fundus photographs in the age-related eye disease study 2 (AREDS2 report number 2). *Invest Ophthalmol. Vis. Sci.* **54**, 4548–4554 (2013).
62. Sorbara, L., Peterson, R., Schneider, S. & Woods, C. Comparison between live and photographed slit lamp grading of corneal staining. *Optom. Vis. Sci.* **92**, 312–317 (2015).
63. Liu, Q., Chen, C., Qin, J., Dou, Q. & Heng, P.-A.Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*. 1013-1023. (IEEE, 2021).
64. Foulks, G. N. Treatment of dry eye disease by the non-ophthalmologist. *Rheum. Dis. Clin. North Am.* **34**, 987–1000 (2008). x.
65. Winkler, S. Punctate Epithelial Defects/Erosions. In: Schmidt-Erfurth, U., Kohnen, T. (eds) *Encyclopedia of Ophthalmology*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-69000-9_929 (2018).
66. Yang, J., Shi, S., Ding, R., Wang, Z. & Qi, X. Towards efficient 3d object detection with knowledge distillation. *Adv. Neu Info Pro Sys* **35**, 21300–21313 (2022).
67. Ren, S., Wu, Z. & Zhu, K. Q. EMO: Earth Mover Distance Optimization for Auto-Regressive Language Modeling. In *The Twelfth International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.2310.04691 (2023).
68. Suk, H., Lee, Y., Kim, T. & Kim, S. Addressing uncertainty challenges for autonomous driving in real-world environments. In *Advances in Computer*. Vol. 134, 317–361 (Elsevier, 2024).
69. Doula, A., Mühlhäuser, M. & Guinea, A. S. AR-CP: Uncertainty-Aware Perception in Adverse Conditions with Conformal Prediction and Augmented Reality For Assisted Driving. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 216–226 (IEEE, 2024).
70. Wang, Y., Li, D., Su, W. & Dai, Y. Clinical features, risk factors, and therapy of epithelial keratitis after cataract surgery. *J. Ophthalmol.* **2021**, 6636228 (2021).
71. Ronneberger, O., Fischer, P. & 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*, 234–241 (Springer, 2015).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (82230033 and 82201143), the fellowship of China Postdoctoral Science Foundation (2021M703703), the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (22qntd3901), the National Key Research and Development Program of China (2023YFC2415400), the National Natural Science Foundation of China (62071210); the Shenzhen Science and Technology Program (RCYX20210609103056042), the Shenzhen Science and Technology Innovation Committee (KCXFZ2020122117340001), the Guangdong Basic and Applied Basic Research (2021A1515220131).

Author contributions

All authors have read and approved the manuscript. Conception or design of the work: Y.D., P.C., Y.L., X.T. and Y.J. Acquisition, images reading, analysis, or interpretation of data: Y.D., P.C., R.X., L.L., H.X., S.Z., Y.Z., K.Y., T.Z., X.H. and X.L. Code writing, figures plotting: P.C., Y.H., J.L. and Z.W. Drafting the work or revising it critically: Y.D., P.C., R.X., K.Y.W., Y.Z. and K.Y. X.T., Y.L. and Y.J. take full responsibility for the work as a whole, including the study design, access to data, and the decision to submit and publish the manuscript.

Competing interests

Xiaoyi Li is affiliated with Zhaoke Ophthalmology Ltd, which is the sponsor of the phase III study for cyclosporine-A eye drops (ClinicalTrials.gov; identifier: NCT04541888). The images of developing, internal validation, and primary external test datasets of the FKD-CSS model were mainly collected from this phase III study. The authors declare no other competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01706-y.

Correspondence and requests for materials should be addressed to Xiaoying Tang, Yan Lou or Jin Yuan.