A Transformer-based Knowledge Distillation Network for Cortical Cataract Grading

Jinhong Wang, Zhe Xu, Wenhao Zheng, Haochao Ying, *Member, IEEE*, Tingting Chen, Zuozhu Liu, Danny Z. Chen, *Fellow, IEEE*, Ke Yao, and Jian Wu, *Member, IEEE*

***Abstract*—Cortical cataract, a common type of cataract, is particularly difficult to be diagnosed automatically due to the complex features of the lesions. Recently, many methods based on edge detection or deep learning were proposed for automatic cataract grading. However, these methods suffer a large performance drop in cortical cataract grading due to the more complex cortical opacities and uncertain data. In this paper, we propose a novel Transformer-based Knowledge Distillation Network, called TKD-Net, for cortical cataract grading. To tackle the complex opacity problem, we first devise a zone decomposition strategy to extract more refined features and introduce special sub-scores to consider critical factors of clinical cortical opacity assessment (location, area, density) for comprehensive quantification. Next, we develop a multi-modal mix-attention Transformer to efficiently fuse sub-scores and image modality for complex feature learning. However, obtaining the sub-score modality is a challenge in the clinic, which could cause the modality missing problem instead. To simultaneously alleviate the issues of modality missing and uncertain data, we further design a Transformer-based knowledge distillation method, which uses a teacher model with perfect data to guide a student model with modality-missing and uncertain data. We conduct extensive experiments on a dataset of commonly-used slit-lamp images annotated by the LOCS III grading system to demonstrate that our TKD-Net outperforms state-of-the-art methods, as well as the effectiveness of its key components. Codes are available at https://github.com/wjh892521292/Cataract_TKD-Net.**

***Index Terms*—Cataract Grading, Knowledge Distillation, Transformer, Medical Imaging Classification**

This research was partially supported by National Key R&D Program of China under grant No. 2018AAA0102102, National Natural Science Foundation of China under grants No. 62176231 and No. 62106218. *Jinhong Wang, Zhe Xu, and Wenhao Zheng contributed equally to this work. Corresponding authors: Haochao Ying and Ke Yao.*

Jinhong Wang, Wenhao Zheng, and Tingting Chen are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (email: wangjinhong@zju.edu.cn; zhengwenhao@zju.edu.cn; trista_chen0603@zju.edu.cn).

Zhe Xu and Ke Yao are with the Eye Center, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China (email: x_z_dahaizhe@zju.edu.cn; xlren@zju.edu.cn).

Haochao Ying is with the School of Public Health, Zhejiang University, Hangzhou 310058, China. (email: haochaoying@zju.edu.cn).

Zuozhu Liu is with ZJU-Angelalign Inc R&D Center for Intelligent Healthcare, ZJU-UIUC Institute, Zhejiang University, Haining, China (email: zuozhuliu@intl.zju.edu.cn).

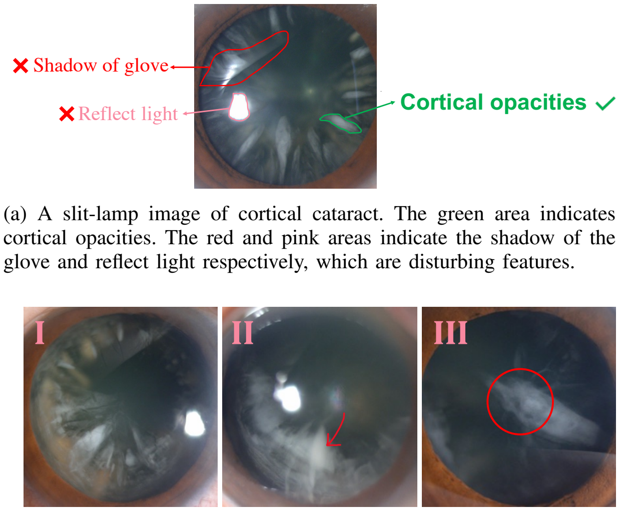
Danny Z. Chen is with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA (e-mail: dchen@nd.edu).

Jian Wu is with the Second Affiliated Hospital School of Medicine, School of Public Health, and the Institute of Wenzhou, Zhejiang University, Hangzhou 310058, China (e-mail: wujian2000@zju.edu.cn).

I. Introduction

CATARACT is the current leading cause of visual impairment and even blindness [1]. Studies have shown that about 314 million people worldwide suffer blindness or visual impairment caused by cataracts [2]. Cataracts can be categorized into three types: nuclear, cortical, and posterior subcapsular cataract [3]. In cortical cataracts, contrast sensitivity is significantly reduced at high spatial frequency in daylight and at low spatial frequency in night light. The prevalence of cortical and nuclear cataracts is much higher than posterior subcapsular cataracts. Unfortunately, no definite drug can treat or prevent any type of cataract. To date, surgical removal of the lens and implantation of intraocular lens are the only beneficial treatments [4]. In this sense, timely and accurate cataract diagnosis (i.e., grading), especially for cortical cataract, is critical for planning treatment to minimize visual impairment [5]. However, clinically, cataract diagnosis needs a face-to-face consultation with the slit-lamp, which imposes a huge accessibility burden on the rapidly aging populations, especially in rural and economically disadvantaged areas. In light of the powerful representation capabilities of recent deep learning (DL) techniques on medical image analysis, several DL methods have been developed for automatic cataract grading [6], [7].

Despite considerable progress, the current automatic cortical cataract grading still suffers from low practical performance due to two specific issues. (1) *Complex cortical opacities.* As shown in Fig. 1(a), cortical cataracts are often wedge-shaped radially oriented opacities originating from the peripheral edge of the lens. Note that the shadow of the examiner's glove and reflected light from the slit-lamp are disturbing features, whose color is similar to cortical cataract. Besides, since cortical cataracts may appear as dense sheets or diffuse blocks and may extend from the periocular to the center, the morphology can be varied and complex, making it hard to grade comprehensively. According to clinical research and experience, the severity of cortical cataract is determined by various key factors including opacity location [8], [9], opacity area [9], [10], and opacity density [8], [11]. Examples of these are shown in Fig. 1(b). It has been observed that wider areas [10], higher density [12], and central location of opacities [9] correspond to a more severe grade of cataract. Nevertheless, clinically, assessing and quantifying cortical opacities based on these specific factors is laborious and error-prone. Even



(a) A slit-lamp image of cortical cataract. The green area indicates cortical opacities. The red and pink areas indicate the shadow of the glove and reflect light respectively, which are disturbing features.

(b) Three examples of cortical cataract with grade 5, which are graded as level 5 mainly due to three different factors: (I) wide areas; (II) high density; (III) extension to the central area.

Fig. 1. Illustrating (a) clinical features and (b) grading factors of cortical opacities.

for computer-aided methods, it is still a challenge to extract and quantify (grade) cortical opacities. In the literature, only a few machine learning (ML) methods [13] were designed for cortical cataract extraction and grading. In these ML methods, cortical opacities were mainly extracted by edge detection [8], [14], [15] and the grading was mainly based on an ML classifier [16], [17] or artificial threshold [15], [18]. Hence, the performance of these methods is limited by the low quality of the edge detection, which can lose most key features of opacities. On the other hand, known DL methods for cataract grading neglected the specific diagnostic images and criteria for cortical cataracts that are different from other cataracts, and did not incorporate the above clinical experience into the model design or did not focus on extracting features of cortical opacities. As a result, these DL methods have limited generalization ability on cortical cataract grading. (2) *Uncertain annotations.* As discussed above, the combination of various factors decides cortical cataract grades, which largely enhances the difficulty of doctors' decisions. Consequently, for the same case, different doctors may provide different grades and even the same doctor may not give exactly the same grade at different annotation times [19] (see Fig. 2). The grading (dis-)agreement also leads to uncertain annotations. With low-quality annotations, training a credible DL model under supervised learning is hard. A common practice is to adopt labels obtained via either the majority vote or simply one annotation version from a preferred rater [20]–[22]. However, DL models trained by these strategies can be over-confident (*i.e.,* easily fall into specific distributions or individual subjectivity), resulting in poor generalization [23], [24]. Random sampling of uncertain labels in training may be a better calibrated strategy to some extent [23], [24]. But, a principled approach is still highly desirable that can directly deal with inconsistent information from uncertain labels [25].

In this paper, we propose a novel DL framework, called Transformer-based Knowledge Distillation Network (TKD-Net), for cortical cataract grading, which is capable of capturing complex cortical opacities with uncertain annotations. To better extract the complex cortical opacities, we propose to incorporate clinical experience into the model by first designing a zone decomposition strategy to consider the location influence of opacities, and then introducing two pairs of triplet sub-scores to explore the area and density influence of opacities. Specifically, we decompose the original imaging features into two zones: the central zone and periocular zone, and give each zone three sub-scores including (**A**): the area of typical cortical opacity, (**B**): the severity of typical cortical opacity, and (**C**): the severity of the rest background opacity. After that, we develop a multi-modal Transformer to fuse imaging features of the two zones and sub-scores prior clinical knowledge. As such, our model can focus more on the quantitative features of opacity and also the relations between them. However, the sub-scores modality may be missing since it is usually difficult to obtain. To address the modality-missing issue as well as uncertain data simultaneously, we develop a novel knowledge distillation method to (1) transfer the sub-score information from the teacher network with perfect data to the student network with modality-missing and uncertain data, and (2) utilize special soft labels to replace the uncertain labels for uncertain data training supervision. The soft labels are high-dimensional logit embeddings from the teacher network that possess more spatial information for reliable supervision. Specifically, we collect the grading annotations of a doctor at three different time instances, in which the unified opinion samples are regarded as certain data and the remaining samples as uncertain data. The inputs of the teacher network are only certain data including images and sub-scores, while the student network receives both certain and uncertain data that include images only. Finally, to train and validate our model in order to provide a more reliable clinical reference, we build a dataset with commonly-used slit-lamp images and the LOCS-III grading system.

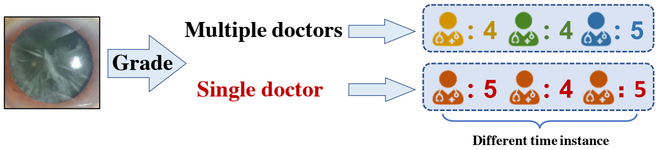


Fig. 2. Two types of uncertain annotations: (i) Multiple doctors may give different grades to the same image; (ii) the same doctor may give different grades to the same image at different annotation times.

In summary, our main contributions are as follows:

- We propose a novel Transformer-based knowledge distillation network, TKD-Net, to address the challenges and explore the potential of DL in cortical cataract analysis, and apply a clinical diagnosis-guided DL algorithm specifically for cortical cataract grading.
- We propose a zone decomposition strategy with additional sub-scores to masterly explore the critical factors of cortical opacity assessment, while further developing a Multi-modal Mix-Attention Transformer to efficiently fuse the multi-modal features.
- We design a knowledge distillation strategy to mitigate the problem of modality missing and uncertain annotations based on the Transformer module.

- We conduct extensive experiments to verify the effectiveness of TKD-Net based on a dataset of commonly-used slit-lamp images annotated by the LOCS III grading system.

II. Related Work

A. Slit-lamp Images and LOCS III System

Different types of source images were used for cataract grading in prior studies, including fundus images [26]–[28], digital camera images [29], ultrasonic images [30], retroillumination images [14], [31], OCT images [32], and slit-lamp images [33]. Further, the grading systems include lens opacity classification system (LOCS) I to III [3], [10], [34], Wisconsin grading system (WGS) [19], Oxford clinical cataract classification and grading system (OCCCGS) [35], and other rough manual classification standards [36]–[38]. Thus, it is difficult to make comparisons between various studies using different types of images and grading systems. Worse, most studies used low-quality fundus images and rough grading systems to train models, which cannot provide a reliable clinical reference.

LOCS III is an improved LOCS system for evaluating slit-lamp and retroillumination images of cataract, and slit-lamp images are widely used in clinical practice. To our best knowledge, cortical cataracts are more clear in slit-lamp images and the LOCS III grading system is more refined and referable. As shown in Fig. 1(a), in slit-lamp images, opacity features are exhibited as white radial lines from the edge to the center. In clinic, LOCS III can score slit-lamp images from 0.1 to 5.9 with the reference standard 1 through 5 as shown in Fig. 3, according to the area, location, and density of the opacity. Thus, by comparing the aggregate area of the opacity in the unknown images with that in the standard images of LOCS III, experts can regard the cortical cataract grading as a 7-class classification: 0 (transparent), 0.1-0.9, 1.0-1.9, 2.0-2.9, 3.0-3.9, 4.0-4.9, and 5.0-5.9.

Therefore, it is promising to study a unified clinical dataset and a corresponding DL model for cortical cataract grading based on slit-lamp images and the LOCS III grading system.

B. Cataract Grading

Many DL methods have been proposed for cataract grading. Zhang *et al.* [36] presented the first DL method for cataract grading that used a convolutional neural network (CNN) to automatically extract features and grade cataract into normal, mild (slight), medium, or severe ones. Some studies [28], [39] also showed that DL methods using CNN perform better in cataract grading compared to ML methods. A few methods [37], [38] proposed to combine DL and ML for six-level cataract grading (non-cataractous, slightly mild, mild, medium, slightly severe, and severe). A recent study [40] collected over 25000 retinal photograph images for automatic detection of visually significant cataracts using a DL algorithm. Some studies focused only on nuclear cataract. Xu *et al.* [41] proposed a fully DL method for nuclear cataract grading that first localized nuclear regions in slit-lamp images by Faster R-CNN, and then applied a ResNet-101 [41] based classification model. Zhang *et al.* [42] developed a CNN model, GraNet, for nuclear cataract classification on AS-OCT images.

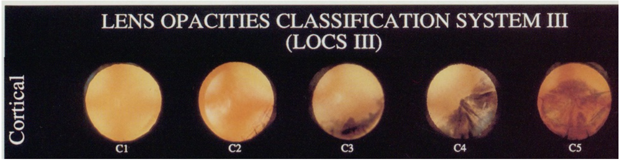


Fig. 3. The LOCS III classification standard reference.

For cortical cataract grading, previous methods mainly extracted opacity textures based on the Canny and Laplacian edge detection [14], [15]. Li *et al*. [8] improved the edge detection with non-linear least-square ellipse fitting. Chow *et al*. [31] proposed to use local entropy filtering to improve the robustness of the edge detection. Gao *et al.* [18] performed pterygium detection on cornea images to enhance the automatic detection and grading of cortical cataract. Gao *et al.* [17] proposed to fuse intensity histogram and texture information for automatic cortical cataract grading with the SVM classifier.

But, DL methods specifically designed for cortical cataract grading are still lacking. We postulate that this is due to the following issues: complex cortical opacities and uncertain annotations. Hence, we propose a novel DL method to tackle these issues, which is the first DL model for cortical cataract grading and outperforms all existing cataract grading methods.

C. Multi-modal Transformers for Medical Image Analysis

Transformer [43] was first applied in the natural language processing (NLP) field. With powerful capabilities in representation learning, many Transformer-based models were developed for multi-modal fusion in medical image analysis. For example, TranMed [44] leveraged a vision Transformer (ViT) to fuse multi-modal MRI images for classification. Zhang *et al*. [45] proposed TransFuse to effectively fuse multi-modal features for medical image segmentation (2D and 3D). Some studies [46], [47] used a ViT to fuse multi-modal MRI features for MRI reconstruction. Tulder *et al.* [48] proposed a cross-view Transformer to fuse multi-modal features from different views of X-ray images for registration. However, the cross-attention applied in this multi-modal Transformer could incur high computation costs. These studies demonstrated that Transformers possess a powerful ability for multi-modal fusion. A recent work [49] showed that using only a few tokens for attention can improve fusion performance and at the same time reduce computation cost. Inspired by this, we design a novel mix-attention Transformer, which uses only a fusion token for attention and can generate a scaleable weight to control the fusion ratio between different modalities.

D. Knowledge Distillation on Medical Image Analysis

Knowledge distillation [50] was first proposed to transfer knowledge from a cumbersome model to a small model which is more suitable for deployment. Researchers then found that knowledge distillation can learn multi-modal information with incomplete modalities [51], [52]. Hence, knowledge distillation has been widely applied to medical multi-modal analysis settings where some modalities may be missing. Hu *et al.* [53]

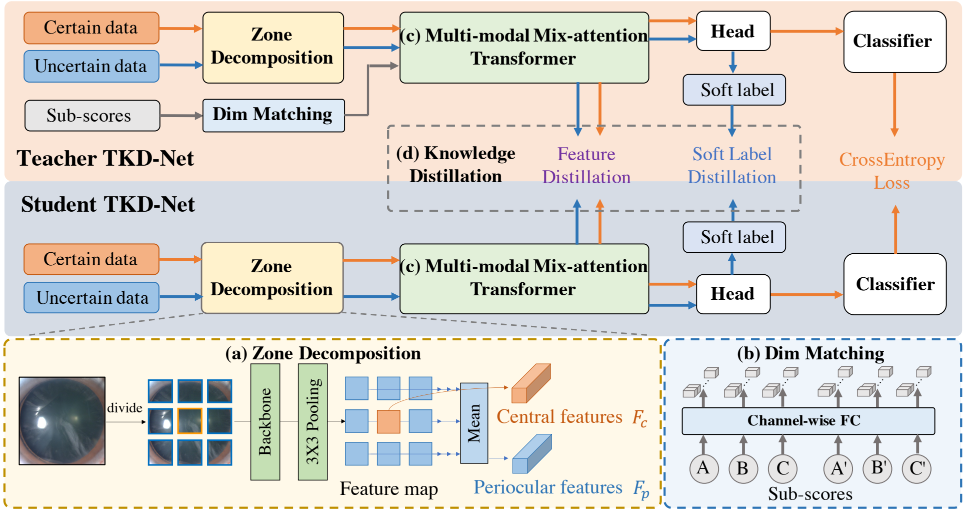


Fig. 4. An overview of our proposed Transformer-based Knowledge Distillation Network TKD-Net.

proposed to use knowledge distillation to transfer knowledge from a trained multi-modal network to a mono-modal one for medical image segmentation. Xing *et al.* [54] proposed a discrepancy and gradient-guided knowledge distillation framework to transfer privileged knowledge from a multi-modal teacher network to a student network for pathological glioma grading. Yang *et al.* [55] proposed an affinity-guided dense tumor-region knowledge distillation mechanism to align features for brain tumor segmentation with missing modalities.

In this paper, inspired by the previous knowledge distillation work, we explore how to adapt knowledge distillation in our Transformer module to mitigate the problem of modality missing and uncertain annotations simultaneously.

III. Method

A. Problem Formulation

We formulate the task of cortical cataract grading as an image classification problem. Given a set X of slit-lamp images and the clinical sub-scores C, our goal is to predict the cortical cataract grading scores $\tilde{Y} = f(X, C; \Theta)$, where f is the model with parameters Θ. Let $S = \{(x_1, c_1, y'_1), \ldots, (x_n, c_n, y'_n)\}$ be the training set of data, where x_i is the i-th image, c_i is the corresponding clinical score, and y'_i either is the corresponding certain label y_i or is an uncertain label set $\hat{y}_i$. That is because, in our task, some samples are annotated with different labels by the same doctor at three different time instances. We define these samples as uncertain data with an uncertain label set $\hat{y}_i$ for x_i. The remaining samples with consistent annotations y_i are certain data. Meanwhile, due to high acquisition cost, the sub-score modality c_i is only used for teacher model training.

In our approach, we only use certain data with full modality (i.e., (x, c, y)) to train the teacher network, which then guides the student network with both certain and uncertain data but missing sub-score modality (i.e., (x, y')) via knowledge distillation. Thus, our student network relies only on images and does not require the sub-score modality in inference.

B. Overview

In Fig. 4, we show the overall architecture of our proposed TKD-Net, a knowledge distillation network with a teacher-student structure. In a nutshell, in TKD-Net, we first carefully design *Zone Decomposition* and *Dim Matching* to capture key clinical features, and then build a multi-modal mix-attention Transformer to fuse multi-modal features and obtain more complex cortical opacity representations. Specifically, TKD-Net first splits a raw image into nine patches to decompose the central and periocular features based on location influence, while designing three quantitative sub-scores for area and density of cortical opacities. After that, to fuse image features between the zones with sub-scores modality, the multi-modal mix-attention Transformer module assigns different weights to different zones. After fusion, the obtained embeddings can be used for classification via cross-entropy loss by the last classifier. To further handle sub-scores modality missing and uncertain label problems, TKD-Net develops knowledge distillation through both token distillation and soft label distillation.

C. Zone Decomposition and Sub-scores

In clinical practice, the location, area, and density of opacity are three decisive factors for assessing the severity of cortical cataract [8], [9]. Thus, we incorporate these three factors into our model design, *i.e.,* we propose a zone decomposition strategy to separately model the location factor of cortical opacities, and introduce two pairs of triplet scores for assessing the area and density factors of cortical opacities. Below, we describe the zone decomposition and sub-scores.

1) Zone Decomposition: Doctors assign more importance to the central zone of the lens since it has bigger influence on the visual acuity [56], [57]. The central zone in most researches [9], [56]–[58] is defined as the central 3-mm diameter area of the pupil. More attention should be paid to lesions in the central zone. Correspondingly, the pupillary margin area is the periocular zone. By manual estimation, we find that when dividing an image equally into 9 patches, the size of the middle patch corresponds exactly to 3-mm diameter. Based on this observation, we propose a zone decomposition strategy to split the extracted central zone features and periocular zone features of an input image, as shown in Fig. 4(a). Specifically, we first divide the input image into 3 × 3 patches, and use a backbone network to extract features. The shape of the

TABLE I
DETAILED DESCRIPTIONS OF THE THREE SUB-SCORES.

	Description	Scores

A	The area of typical cortical opacity.	0 = N/A;
		1 = within 1 quarter;
		2 = within 2 quarters;
		3 = within 3 quarters;
		4 = full.

B	The severity of typical cortical opacity (linear, cuneiform, or clustered opacity with high density).	0 = N/A;
		1 = light;
		2 = medium;
		3 = dense.

C	The severity of the rest background opacity (circumferential, nebula or diffuse opacity with low density).	0 = N/A;
		1 = light;
		2 = medium;
		3 = dense.

obtained feature map is (H, W, C), where $H = W = 3$ and $C = 2048$. Then, we take the feature map of the central patch as the central features, and obtain the periocular features by averaging those of the other 8 patches. Accordingly, both the central and periocular features are a vector of size 2048, and both these vectors will be divided into 4 tokens of size 512 each before being fed into the Transformer module.

Interestingly, our zone decomposition offers two advantages. (i) Averaging the feature maps of the 8 pupillary margin patches is equivalent to reducing the weight of the periocular zone features to about 1/8 of their sum, so that the weight of the central zone is relatively "lifted" to encourage the model to focus more on the opacity of the central zone. (ii) Zone decomposition is beneficial to separately extracting specific features in the zones, since the opacities of the central and periocular zones vary in shape (e.g., the central zone often has patchy opacity and the periocular zone has diffuse opacity).

2) Sub-scores: Considering that the LOCS III grading system offers only a single number for cataract, to explore more labeling possibilities for ophthalmologists' diagnosis and simultaneously incorporate more prior knowledge into the model design, we introduce several sub-scores to assess the area and density of cortical opacities as an additional modality of input data. All the sub-scores are labeled by an ophthalmologist who is an attending doctor in the cataract field. According to the standard in [9], we divide the zone within the pupil range into 2 parts: (1) the central zone (a central 3-mm diameter area); (2) the peripheral zone (the rest area within the pupil range). We classify the shape of cortical opacity into two types: (1) typical cortical opacity (wedge-shaped and radially oriented); (2) background cortical opacity (extending in a circumferential manner around the more peripheral cortex). Thus, both the central zone and peripheral zone are scored according to Table I with 3 labels, *i.e.*, (A, B, C) and (A', B', C'), respectively. Sub-score A (A') refers to the area of typical cortical opacity, which is assessed by the number of quarters that are involved by the opacity (we simplify the octant division of the circumference [9] to the quadrant division). Sub-scores B (B') and C (C') refer to the severity of typical cortical opacity and background cortical opacity, respectively. According to the standard in [59], we grade sub-scores B and C on 4 scales, with a higher score indicating a larger or denser opacity.

In order to let the dimension of the sub-scores match that of the image features for convenient concatenation, we use a channel-wise fully connected (FC) layer to expand each sub-score to 512 dimensions which are the same as the image feature tokens, as shown in Fig. 4(b).

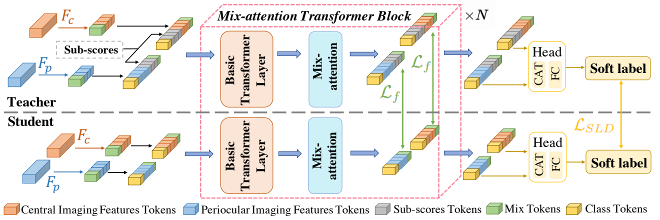


Fig. 5. The architecture of the Multi-modal Mix-Attention Transformer and Transformer-based knowledge distillation.

D. Multi-modal Mix-attention Transformer

Aiming to fuse the image and clinical score information and simultaneously interact with the central and periocular features, we propose a Mix-Attention Transformer (MAT) formed by several MAT blocks, as shown in Fig. 5. As the output of Zone Decomposition, the central feature vector and periocular feature vector are divided into $P + 1$ tokens. P tokens are image feature tokens and the remaining one is the mix token. Note that tokens are basic units in the Transformer that are equivalent to a one-dimensional vector of the same size. The input of each MAT block is a token sequence z formed by a CLS ("classification") token z_{cls} (randomly generated initially and its representation can be passed to a classifier for downstream classification tasks), P image feature tokens $\{x^i\}_{i=1}^P$, P sub-score tokens $\{c^i\}_{i=1}^P$, and a mix token z_m (used for fusion in the mix-attention function), as follows:

$$z_k = g(x_k, c_k, z_{cls_k}, z_{m_k}) = [z_{cls_k}, x_k^1, \ldots, x_k^P, c_k^1, \ldots, c_k^P, z_{m_k}], \tag{1}$$

where $k \in \{cen, per\}$. Thus, z_k denotes the central or periocular token sequence. Since there are 3 sub-scores for each zone, we set $P = 3$ to make the numbers of imaging tokens and sub-score tokens consistent to balance the image information and sub-score information. Specifically, each MAT block contains a basic Transformer encoder layer and a mix-attention operation, as presented below.

1) Basic Transformer Layer: The basic Transformer layer can deeply fuse the image information and sub-score information since the relationship between the tokens is efficiently learned during the multi-head self-attention process. Following the vanilla Transformer [43], each layer includes Multi-Headed Self-Attention (MSA), Layer Normalization (LN), and feed-forward network (FFN) layers applied using residual connections. These processes at the l^{th} layer can be formulated as:

$$\begin{aligned} z_{hidden}^l &= \text{LN}(\text{MSA}(z_{cen/per}^l)) + z_{cen/per}^l, \\ y_{cen/per}^{l+1} &= \text{LN}(\text{FFN}(z_{hidden}^l)) + z_{hidden}^l. \end{aligned} \tag{2}$$

Then, the mix-attention operation is applied to the central features and periocular features for delivering information from one zone to another zone, as follows:

$$z_{cen}^{l+1}, z_{per}^{l+1} = \text{Mix-attention}(y_{cen}^{l+1}, y_{per}^{l+1}). \tag{3}$$

After that, the delivered information is with other information at the next basic Transformer layer. Thus, obviously, the core insight of our MAT is mix-attention, *i.e.,* reweighted fusion between different zones.

2) Mix-attention: Inspired by [49], our proposed mix-attention utilizes only a single token to condense information in limited attention flows. However, considering the different importance of the central features and periocular features, the average strategy used in [49] cannot take account of that. In order to tackle this issue for effective fusion, we design a special mix-attention to not only avoid redundant attention flow but also assign different weights to the central features and periocular features. In our design, the mix-attention is performed only on mix tokens. Specifically, assume the output token sequences of the basic Transformer layer are as follows:

$$y_{cen}^{l+1} = [z_{cen'}^{l+1} || z_{m'_{cen}}^{l+1}],\ y_{per}^{l+1} \quad = [z_{per'}^{l+1} || z_{m'_{per}}^{l+1}]. \tag{4}$$

The mix-attention is designed to fuse mix tokens at a preset ratio, formulated as:

$$\begin{aligned} z_{m_{cen}}^{l+1} = z_{m_{per}}^{l+1} &= \lambda \times z_{m'_{cen}}^{l+1} + (1-\lambda) \times z_{m'_{per}}^{l+1}, \\ z_{cen}^{l+1} &= [z_{cen'}^{l+1} || z_{m_{cen}}^{l+1}], \\ z_{per}^{l+1} &= [z_{per'}^{l+1} || z_{m_{per}}^{l+1}], \end{aligned} \tag{5}$$

where λ is a hyper-parameter to assign different weights to the central features and periocular features. As one may see, the attention flow is only admitted to passing between the central and periocular features through the mix token. This gives three advantages. (1) Avoiding full (pair-wise) attention between the central and periocular features, thus reducing computation cost. (2) Most tokens do not interact directly with the features of another zone, and this allows to retain the specific features of each zone to the greatest extent for maintaining the integrity of information. (3) By presetting $\lambda > 0.5$, the mix-attention is able to force the model to focus more on the central features, and this is more in line with clinical prior knowledge.

E. Transformer-based Knowledge Distillation

1) Token Distillation: In clinical practice for cataract grading, images are the most accessible references made by machines. In contrast, sub-scores need manual evaluation that are difficult to obtain. Thus, for most patient samples in clinical practice, only images are included with no additional diagnostic information. This fact leads to lots of missing modality samples and the model cannot inference these samples if the model is trained with full modality samples. To address this issue, we propose token distillation based on transfer learning between the CLS tokens from the teacher network (z_{cls}^t) to the student network (z_{cls}^s). Specifically, for the i^{th} layer, the token distillation loss is:

$$\mathcal{L}_{f_i} = \frac{1}{2}[\text{Smoo-L1}(z_{cls_{cen}}^t, z_{cls_{cen}}^s) + \text{Smoo-L1}(z_{cls_{per}}^t, z_{cls_{per}}^s)], \tag{6}$$

where Smoo-L1 denotes the smooth-L$_1$ loss. Then, by averaging all the token distillation losses of the layers, the final token distillation loss is:

$$\mathcal{L}_{TD} = \frac{1}{n}\sum_{i=1}^{N} \mathcal{L}_{f_i}. \tag{7}$$

TABLE II
PERFORMANCE COMPARISON OF POPULAR BACKBONE MODELS.

Method	Accuracy	Recall	Precision	F1-score	Kappa	MCC	Params.
ViT [60]	63.7	55.6	58.4	56.5	51.9	51.7	86M
VGG-19 [61]	64.1	58.5	59.8	58.7	57.4	57.5	143M
ResNet-18 [62]	77.2	64.9	68.7	66.7	72.4	72.5	**12M**
ResNet-50 [62]	**78.2**	66.1	**69.5**	67.2	**74.2**	**74.3**	25M
ResNet-101 [62]	77.9	**66.3**	69.4	**67.5**	72.7	72.8	44M
DenseNet [63]	74.8	64.7	65.1	64.9	66.7	66.8	20M
InceptionV3 [64]	58.5	48.6	57.5	50.6	43.6	44.0	24M

Our token distillation between CLS tokens is specially designed using Transformer since the CLS tokens are used to aggregate all the tokens' information for the final classification. On the other hand, compared to the distillation between all the corresponding tokens of the teacher and student networks, our token distillation only through CLS tokens is more efficient since the distillation cost is significantly reduced.

2) Soft Label Distillation: In order to address the problem of uncertain labeled training data, we propose soft label distillation, which generates soft labels through the teacher-student network for supervision, rather than using uncertain labels. Since the teacher network is trained with certain data, we take the soft labels generated by the teacher network as "ground truth" for supervising the student network learning. The generation of soft labels is based on the embeddings encoded by both the CLS tokens using a fully connected (FC) layer. Specifically, the soft label distillation loss is as follows:

$$\mathcal{L}_{SLD} = KL(FC([z_{cls_{cen}}^t \| z_{cls_{cen}}^s]), FC([z_{cls_{per}}^t \| z_{cls_{per}}^s])), \tag{8}$$

where KL denotes the Kullback–Leibler divergence.

Since CLS tokens are used for final classification, the soft labels generated with CLS tokens could be viewed as a high dimensional transformation of true labels. By training the teacher model with only certain data, we regard the soft labels in the teacher model as true labels of the uncertain data. By soft label distillation, we force the soft labels of the student model to align with the soft labels of the teacher model so as to achieve the supervised training with uncertain data.

F. Overall Loss Functions

For both the teacher and student networks, their total losses include the following Cross-Entropy (CE) loss for certain data with images and certain labels:

$$\mathcal{L}_{CE} = \frac{1}{N_{certain}} \sum_{i=1}^{N_{certain}} \sum_{j \in C} y_i^j f^j(x_i, c_i; \Theta). \tag{9}$$

The student network's total loss also consists of the token distillation loss and soft label distillation loss, while the teacher network does not. Thus, the total losses of the teacher and student networks are as follows:

$$\mathcal{L}_{tot}^t = \mathcal{L}_{CE},\ \mathcal{L}_{tot}^s = \mathcal{L}_{CE} + \alpha(\mathcal{L}_{TD} + \mathcal{L}_{SLD}), \tag{10}$$

where α is a hyper-parameter that controls the importance of the terms. We set $\alpha = 0.5$ based on experiments.

IV. EXPERIMENTS

A. Experimental Settings

1) Dataset and Pre-processing: Our dataset contains 2150 samples of slit-lamp images, including 150 normal samples

TABLE III
PERFORMANCE COMPARISON OF TKD-NET WITH KNOWN METHODS.

Method	Year	Accuracy	Recall	Precision	F1-score	Kappa	MCC
Zhang *et al.* [38]	2019	77.2	64.9	68.7	66.7	72.4	72.5
Xu *et al.* [41]	2019	77.9	66.3	69.4	67.5	72.7	72.8
Khan *et al.* [39]	2021	64.1	58.5	59.8	58.7	57.4	57.5
Tham *et al.* [40]	2022	78.2	66.1	69.5	67.2	74.2	74.3
CataractNet [28]	2021	54.3	42.8	51.4	43.4	37.1	37.8
GraNet [42]	2020	78.8	66.9	70.9	68.9	74.5	74.5
DeepLensNet [65]	2022	74.4	63.5	65.6	64.0	66.7	66.8
Ensemble CNN [66]	2022	74.2	62.5	65.9	63.8	64.5	64.6
Stacking Ensemble [67]	2022	74.4	63.6	66.0	64.4	64.9	65.0
CataractEyeNet [68]	2023	63.0	52.4	58.7	54.0	47.3	47.4
TKD-Net (ours)	-	**82.1**	**71.4**	**73.0**	**71.8**	**78.7**	**78.8**
TKD-Net (Teacher)	-	**95.1**	**81.6**	**82.0**	**81.8**	**90.7**	**90.8**

(77 women and 73 men; 27.8±5.3 years) and 2000 cataract samples (1092 women and 908 men; 71.4±14.6 years). Only one eye of each patient and subject was included. The labels of 1670 samples are certain and the labels of the other 480 samples are uncertain. Based on the LOCS III standard, each sample of the certain data is graded by an experienced expert in 7 ranking categories. To avoid negative influence of the relatively complex background, we crop the lens region of each original image as an input image after the lens region is localized by Faster R-CNN [69]. All the cropped images are resized to the size of 224 $\times$ 224 and normalized by subtracting the ImageNet means and stds. We have obtained approval by the Medical Ethical Committee of our cooperated hospital (No. ChiCTR2300071279) for scientific research using these slit-lamp images.

2) Implementation: Our experiments use a computer with an Intel i7 processor and an NVIDIA GTX 3090 GPU. The code is built on the PyTorch platform. We adopt the SGD optimizer and set the batch size as 16 for training all the models. The initial learning rate is set as 0.01 and is reduced by a factor of 10 at 12, 24, and 36 epochs. The warm step is set to 50 iterations. For the hyper-parameters of the MSA, the number of heads is 3, the number of blocks is 12, the input dimensions are 512, the forward dimensions are 768, and the activation is GeLU. For data augmentation, we use random horizontal flipping. The images with certain labels are randomly divided by 75%, 5%, and 20% for training, validation, and testing, respectively. Especially, the teacher network is trained with only certain data while the student network is trained with both certain and uncertain data. For both the teacher and student networks, the validation and testing samples only contain certain data. The known methods are implemented using the original papers' codes or re-implemented based on the original papers. When training the known models, the labels of uncertain data are determined by a random strategy [23].

3) Evaluation Metrics: We conduct quantitative evaluation using several widely-used metrics: Accuracy, Recall, Precision, F1-score, Kappa, and MCC.

B. Backbone Analysis

To determine the best option for our backbone, we conduct extensive experiments and compare the accuracy and computational costs of a set of widely-used backbone models. Table II shows the results, from which we observe that although some other backbones use fewer parameters than ResNet-50, their performance is not as good as ResNet-50. ResNet-50 achieves generally the best performance with relatively low computational costs. Thus, we choose ResNet-50 as our backbone model.

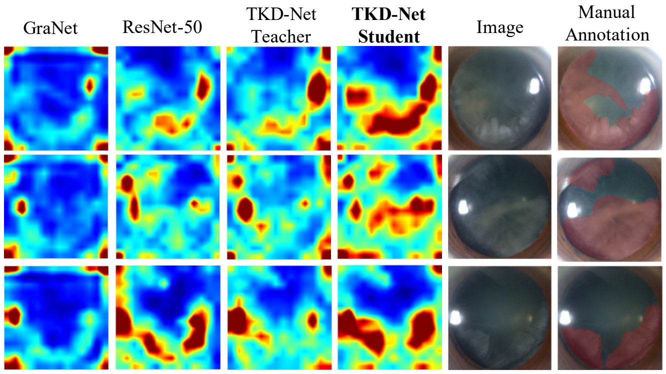


Fig. 6. Visualized feature maps of known methods and our TKD-Net.

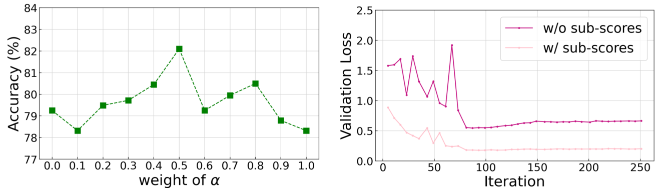
C. Comparison with Known DL Methods

We compare the performance of our proposed TKD-Net against the following known DL methods for cataract grading.

- Khan *et al.* [39]: It applied a VGG-19 pre-trained model as the feature extractor, and only the fully connected layers were trained for cataract detection.
- Zhang *et al.* [38]: It applied ResNet-18 as the feature extractor and attached an SVM classifier for 6-level cataract grading.
- Tham *et al.* [40]: It applied ResNet-50 as the feature extractor and attached an XGBoost classifier for cataract detection.
- Xu *et al.* [41]: It localized the nuclear regions in slit-lamp images with Faster R-CNN, followed by a ResNet-101 based grading model.
- CataractNet [28]: It proposed a 16 layers dense CNN for feature extraction and cataract detection.
- GraNet [42]: It proposed a new grading block network based on ResNet-18, and used both focal loss and cross-entropy loss to train the model for nuclear cataract classification.
- CataractEyeNet [68]: It used VGG-19 and added additional 20 conventional layers for cataract detection.
- DeepLensNet [65]: It applied DenseNet for cataract grading.

TABLE IV
ABLATION STUDY FOR THE TEACHER NETWORK AND STUDENT NETWORK. ZD = ZONE DECOMPOSITION, MMT = MULTI-MODAL MIX-ATTENTION TRANSFORMER, SS = SUB-SCORES, TD = TOKEN DISTILLATION, AND SLD = SOFT LABEL DISTILLATION.

Method	Teacher			Student		Accuracy	Recall	Precision	F1-score
	ZD	MMT	SS	TD	SLD				
Teacher						78.2	66.1	69.5	67.2
	✓					80.0	68.6	71.0	69.7
			✓			85.9	72.3	74.9	73.8
	✓	✓				80.9	69.5	72.3	70.6
	✓		✓			86.1	72.6	75.2	74.2
	✓	✓	✓			**95.1**	**81.6**	**82.0**	**81.8**
Teacher + Student	✓	✓	✓			80.7	68.8	72.0	69.7
				✓		81.4	70.4	72.8	70.9
					✓	80.9	69.7	72.4	70.7
				✓	✓	**82.1**	**71.4**	**73.0**	**71.8**



(a) (b)

Fig. 7. (a) Performance of our TKD-Net with different weights of α. (b) The learning curves of the TKD-Net with and without the sub-scores guidance.

TABLE V
PERFORMANCE COMPARISON OF TKD-NET WITH DIFFERENT TRANSFORMER METHODS. IPS = IMAGE PER SECOND.

Method	Accuracy	Recall	Precision	F1-score	IPS
Baseline	86.1	72.6	75.2	74.2	**60**
Vanilla self -attention [43]	94.3	80.4	81.8	81.0	30
Vanilla cross -attention [43]	93.22	80.3	79.2	79.7	44
Bottleneck attention [49]	92.91	79.1	79.9	79.4	50
TKD-Net (mix-attention)	**95.1**	**81.6**	**82.0**	**81.8**	53

- Ensemble CNN [66]: It combined the models of AlexNet, Inception V3, Xception, and InceptionResNetV2 for cataractdetection and grading.
- Stacking Ensemble [67]: It combined the models of Inception V3, MobileNet-V2, and NasNet-Mobile for cataract grading.

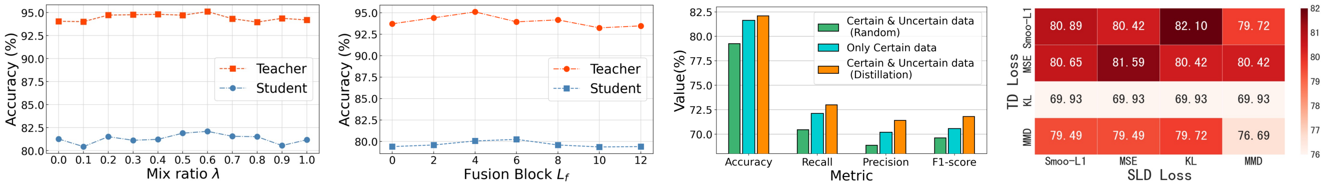
We present the results in Table III, from which several observations can be made. (1) On our task, the best performing models are all ResNet based ones (*i.e.,* [38], [40]–[42]), which further show the advantages of ResNet and it is effective to choose ResNet as our backbone model. (2) Compared to the previous DL methods, our Teacher TKD-Net improves the grading performance significantly in various metrics. This validates the effectiveness of our proposed TKD-Net for cortical cataract grading. This is because TKD-Net incorporates clinical diagnostic criteria into the model design and utilizes clinical prior knowledge as information supplement to assess cortical opacities comprehensively. (3) Compared to the state-of-the-art DL models for cataract grading [28], [42], when only images are used as input, our Student TKD-Net still outperforms them significantly, which demonstrates the superiority of our TKD-Net benefited from knowledge distillation.

In addition, we visualize some feature maps of the two best-performing baselines and our model in Fig. 6, to further show the effectiveness of our TKD-Net. These feature maps are from the Conv4 stages of the backbones (we omit the Conv2 and Conv3 stages, as they do not show much difference). For each such stage, we average the features in the channel dimension, and then apply a sigmoid function and upsample them to the original image size. From these feature maps, one can observe that compared to GraNet and ResNet-50, our student TKD-Net focuses more on opacities, demonstrating the effectiveness of our proposed feature zone decomposition and Transformer-based knowledge distillation strategies. By focusing more on the opacities, our method can better capture and quantify the opacity textures and obtain accurate grading performance, thus providing a more reliable clinical reference for doctors. Note that our teacher TKD-Net can pay more attention to opacities since the additional sub-scores may substitute part of opacities features. That is, the teacher TKD-Net utilizes sub-scores for opacity assessment and exhibits less dependence on images.

D. Ablation Study

1) Efficiency of Key Components: We conduct ablation experiments for both the teacher and student networks, to evaluate the role of each key component in our TKD-Net. The results are shown in Table IV.

The ablation experiments for the teacher network include the zone decomposition (ZD) strategy, the multi-modal mix-attention Transformer (MMT) module, and the sub-score utilization (SS). From Table IV, we can draw several observations. (1) Compared with the baseline model, ResNet-50, our ZD strategy improves in all the metrics, showing that our separated feature modeling of the central zone and periocular zone to focus more on the central zone is effective. (2) MMT without sub-scores slightly improves the model performance, showing that MMT can better fuse features of different regions. (3) Only adding the sub-scores as more references (like experts do in clinical practice), the grading performance is significantly improved. It suggests that the sub-scores can provide detailed guidance for extracting and summarizing opacity lesion features. (4) With the sub-scores, applying MMT achieves the best Accuracy of 95.1%, a further improvement by 9.0%. The Recall, Precision, and F1-score are



(a) (b) (c) (d)

Fig. 8. (a) Performances of the teacher and student networks with different values of the mix ratio $\boldsymbol{\lambda}$. It achieves the best performance when $\boldsymbol{\lambda}$ = 0.6. (b) The impact of using mix-attention for fusion that starts at different fusion blocks $\boldsymbol{L_f}$. (c) Performances of different model versions in the four metrics. (d) Performances of different model versions with different distillation losses.

TABLE VI
PERFORMANCE COMPARISON OF TKD-NET WITH OTHER KNOWLEDGE DISTILLATION METHODS.

Method	Accuracy	Recall	Precision	F1-score
Xiong *et al.* [70]	80.2	67.0	71.3	70.0
Choi *et al.* [71]	79.8	68.2	70.9	69.2
ProtoKD [72]	81.4	70.3	72.5	70.6
Xing *et al.* [73]	81.2	70.0	72.2	70.4
TKD-Net (ours)	**82.1**	**71.4**	**73.0**	**71.8**

also improved by 9.0%, 6.8%, and 7.6%, respectively. The significant improvements show that MMT is more effective for multi-modal fusion and is able to utilize the sub-scores to adjust the weight of the model's attention for each zone.

For the student network, the effect of knowledge distillation, including token distillation (TD) and soft label distillation (SLD), is evaluated based on the trained teacher network with the above schemes. Table IV reports the detailed results, from which several observations can be made. (1) Compared with the baseline teacher network, by loading the pre-trained model weights from the best teacher network without any distillation process, the baseline student network achieves an Accuracy of 80.7%, providing an improvement of 2.5%. This shows that the teacher network indeed learns relevant information from the sub-scores and retains a small part of the information in the student network even if the sub-scores are missing. (2) By using the TD scheme based on the Transformer, the model outperforms the baseline student network by 0.7% in Accuracy, 1.6% in Recall, 0.8% in Precision, and 1.2% in F1-score, validating the effectiveness of the feature alignment between the CLS tokens for modeling the missing modal. (3) SLD is effective for learning missing modal features, and the model obtains competitive performances, which shows that the soft labels can replace the ambiguous labels for uncertain data. (4) With both our distillation schemes TD and SLD, our TKD-Net has a final Accuracy of 82.1%, Recall of 71.4%, Precision of 73.0%, and F1-score of 71.8%, which outperforms the baseline model by a large margin and achieves the best performance, indicating the superiority of our approach.

2) Different Weights of Distillation Loss: We also vary the trade-off weight of the distillation loss and analyze its sensitivity to the hyper-parameter α in Eq. (10). Specifically, we use a range of $\alpha \in [0, 1]$ with a step size of 0.1, and observe the grading results. As shown in Fig. 7(a), our method achieves the best performance when $\alpha = 0.5$, while a larger weight of the distillation loss will overwhelm the supervised one and lead to underfitting.

3) Learning Curves of Validation Loss: In Fig. 7(b), we plot the learning curves of the validation loss in the TKD-Net with and without the sub-scores guidance. It can be seen that the TKD-Net without sub-scores model faces a certain overfitting issue (the validation loss unexpectedly begins to increase as the model is trained with more iterations). But, the TKD-Net with sub-scores model can deal with the overfitting issue (the validation loss maintains a balanced fitting state as the model is trained with more iterations), which we think benefits from the more consistent information brought by sub-scores.

E. Multi-modal Transformer Analysis

Next, we compare our mix-attention Transformer with other multi-modal Transformer methods, and explore several different settings for the multi-modal mix-attention Transformer component.

1) Comparison with Other Multi-modal Transformer Methods: Our method provides an alternative approach to efficiently fuse the image and sub-score features simultaneously with mix-attention. To validate this capability, we compare our approach with the baseline and other multi-modal attention methods, as follows.

- Baseline: Use fully connected layers to fuse multi-model features.
- Vanilla self-attention [43]: Concatenate the multi-modal feature tokens and apply unrestricted pairwise self-attention between all the tokens at each layer.
- Vanilla cross-attention [43]: Concatenate the multi-modal feature tokens and use them to update each modality by the multi-head cross attention.
- Bottleneck attention [49]: Assign several bottleneck tokens to each modality and use self-attention within the modalities; then average bottleneck tokens at the end of each layer.

Table V shows performance results and inference time of the entire pipeline with different attention modules. The baseline method attains the fastest inference speed since the FC layers incur less computational costs, but its performances in Accuracy and other metrics are not very good. In contrast, the methods with Transformers gain great increases in Accuracy and other metrics, albeit at the expense of lower inference speeds. Among these Transformer-based methods, our method achieves the highest performances in all the metrics and remains competitive in inference speed with only a slight decline from that of the baseline method. This validates the comprehensive ability of our proposed mix-attention approach in both Accuracy and inference speed for multi-model fusion.

2) Different Settings for the Mix-attention Transformer: The hyper-parameter λ is used to assign different weights to the central and periocular features. To determine the best setting

TABLE VII
PERFORMANCE COMPARISON WITH DIFFERENT SUB-SCORE INPUTS.

Sub-score Input	Accuracy	Recall	Precision	F1-score
None	80.9	69.5	72.3	70.6
Only A	85.4	73.9	74.1	73.9
Only B	81.6	70.0	71.8	70.8
Only C	87.9	74.0	75.6	74.7
A + B	86.7	75.5	76.0	75.7
A + C	94.2	78.1	79.5	78.7
B + C	90.7	76.8	78.0	77.4
A + B + C	**95.1**	**81.6**	**82.0**	**81.8**

for λ and verify the clinical prior knowledge, we conduct ablation experiments for the value of λ varying from 0 to 1 with a step size of 0.1. The results are given in Fig. 8(a). It is observed that both the teacher and student networks of TKD-Net attain the best performances when $\lambda = 0.6$, showing that the central features are more important than the periocular features, which is in line with the clinical prior knowledge.

3) When to Start Mix-attention for Fusion?: In default, the mix-attention is applied in all the MAT blocks. However, in most cases, *mid fusion* is considered a better way for fusion. Hence, we investigate the impact of varying the starting block L_f in which mix-attention begins to be applied. We conduct experiments with $L_f = 0, 2, 4, 6, 8, 10, 12$ for both the teacher and student networks. Fig. 8(b) shows the results. One can see that for both the teacher and student networks, *mid fusion* outperforms both early ($L_f = 0$) and late ($L_f = 12$) fusion. For the student network, the best performance is attained by using fusion layer $L_f = 6$. This suggests that the model benefits from multiple blocks of cross-modal (mean central and periocular features) information flow in the later blocks, allowing earlier blocks to specialize in learning unimodal features. However, the best performance of the teacher network is obtained by using fusion layer $L_f = 4$. We hypothesize that this is due to the added sub-scores, from which the supplementary clinical information can accelerate the process of image feature extraction to advance the process of feature fusion. In order to obtain the best performance for the student network, we set $L_f = 6$.

F. Knowledge Distillation Analysis

We first compare our Transformer-based knowledge distillation approach with other known knowledge distillation methods. Then we explore the premises and necessity of knowledge distillation and also the best loss function settings.

1) Comparison with Other Knowledge Distillation Methods: To validate the effectiveness of our Transformer-based knowledge distillation approach, we compare TKD-Net with previous knowledge distillation methods used in medical disease diagnosis [70]–[73]. These known methods are all CNN-based knowledge distillation models while our method is the first knowledge distillation approach based on Transformer. Xiong *et al.* [70] and Choi *et al.* [71] proposed to apply knowledge distillation during the feature-extracting process. Wang *et al.* [72] and Xing *et al.* [73] proposed to apply knowledge distillation during both the feature-extracting process and the predicting head of the model. The comparison results are shown in Table VI. One can see that compared to the previous knowledge distillation methods, our proposed TKD-Net obtains the best performances in all the metrics, demonstrating the superiority of our proposed Transformer-based knowledge distillation approach.

TABLE VIII
PERFORMANCE COMPARISON OF REGRESSING SUB-SCORES AND OUR TRANSFORMER-BASED KNOWLEDGE DISTILLATION APPROACH.

Method	Accuracy	Recall	Precision	F1-score
Regressing sub-scores	80.5	67.4	71.6	70.6
Ours TKD-Net	82.1	71.4	73.0	71.8

2) Do the Uncertain Data Matter?: In our approach, the strategy is based on two premises that need to be verified: (1) training with additional uncertain data can be better than training with only certain data; (2) for the utilization of uncertain data, knowledge distillation and soft labels are better than randomly assigned labels. To verify premise (1), we conduct experiments that use only certain data to train TKD-Net. For premise (2), we conduct experiments that use both certain data and uncertain data in which the label of each uncertain sample is randomly chosen from the adjacent categories. From Fig. 8(c), we find that when uncertain data are assigned with labels from the adjacent categories randomly for training, the model yields the worst performance, which shows that randomly assigned labels may give massive error-prone information and a better strategy is not to use such random labels. Furthermore, compared to the using-only-certain-data model, our approach that uses knowledge distillation to provide soft labels to uncertain data achieves higher performance, validating that uncertain data possess useful information for guidance, and the knowledge distillation and soft labels help the model to become more effective.

3) Different Losses for Distillation: In our proposed TKD-Net, the losses for token distillation and soft label distillation are selected from several commonly-used losses: Smooth L_1 loss, MSE loss (L_2 loss), Kullback-Leibler divergence (KL), and Maximum Mean Discrepancy (MMD). Extensive experiments are conducted to select the best for these loss functions. The heatmap in Fig. 8(d) reports the results. From the heatmap, one can find that the combination of Smooth-L_1 loss for token distillation (TD) and KL for soft label distillation (SLD) achieves the best performance.

G. Sub-score Analysis

1) Different Combinations: Finally, we explore the impacts of different sub-scores and their combinations on the model performance. The experimental results are given in Table VII, from which several conclusions can be drawn. (1) All three sub-scores are beneficial to the model performance improvement, and this validates the clinical effectiveness and reliability of introducing the sub-scores. (2) The performance improvement by the addition of sub-score C is more significant while the improvement is much smaller when adding only sub-score B. This suggests that it may be difficult for the model to capture the low-density cortical opacity features of the images, but the model can capture high-density cortical opacity features very well. (3) When combining the sub-scores, the improvement is bigger than the sum of the improvements by using each individual sub-score alone which is involved in the combination. For example, compared to the baseline, the model with only sub-score C improves Accuracy by 7% and the model with only sub-score A improves Accuracy by 4.5%, but the model with both sub-scores A and C improves Accuracy by 13.3%, which is bigger than 7% + 4.5% = 11.5%. This shows that using sub-scores in combination can provide more information, and further illustrates the effectiveness and necessity of using the three sub-scores and their combination.

2) Regressing Sub-scores by Multi-Task Learning?: Note that in our design, we take sub-scores as additional features for enhancing the model performance. Another possible way is to utilize sub-scores as labels via multi-task learning. That is, here our goal is to regress these sub-scores and grade the cortical cataract simultaneously. In this way, the loss function contains the cataract grade classification loss (Cross Entropy Loss) and sub-scores regressing loss (MSE loss). We conduct experiments to explore the effectiveness of this scheme, as shown in Table VIII. We find that this multi-task learning based scheme does not show very good performance. We think this is due to that regressing these sub-scores is as difficult as cataract grading, and learning both sub-scores and grading labels increases the difficulty of model fitting.

V. Discussions

On our TKD-Net for cortical cataract grading, the following points are worth noting. (1) Feature zone decomposition based on positions is a new way for separate lesion area modeling and grading performance improvement. Extensions of this technique can be further explored and have a potential to be widely applicable in other ophthalmic disease diagnoses, since the disease severity of the periocular area and that of the center area are usually different and the center area is often much more important. (2) Multi-modal fusion is used widely in medical image analysis. In this work, our mix-attention Transformer provides support for efficient multi-modal fusion. But, how to improve multi-modal fusion is worth further exploration. (3) Our idea is motivated by the hypothesis that clinical diagnosis details can help deep learning to better capture features. The promising results show that developing deep learning models from a clinical perspective is highly beneficial and important. (4) Our method achieves significant performance improvement using sub-scores for multi-modal learning and knowledge distillation. However, a limitation is that sub-scores have not been used in the LOCS III grading system as a recognized evaluation standard. In the future, through further research and verification, we mainly focus on improving our proposed sub-scores and hope to develop the LOCS III grading system by adding the proposed sub-scores as a recognized evaluation standard for more reliable reference.

Another limitation is that the reflective shadows/lights are common in the images in our dataset, this will have a certain impact on model feature extraction. In fact, reflective shadows/lights often appear in other disease images taken by optical instruments as well. In order to eliminate the systematic error, how to remove the reflective shadows/lights is also worth exploring which is another major goal of our future work.

VI. Conclusions

In this paper, we studied the key issues in cortical cataract image analysis, including the difficulties of quantifying opacity features and uncertain labels for the common subjectivity of doctors. We presented a new Transformer-based knowledge distillation network, TKD-Net, for cortical cataract grading. Our proposed feature zone decomposition strategy decomposes image features based on two regions, *i.e.,* central features and periocular features, to analyze their relationship and adjust their weights. Our multi-modal Transformer can fuse image features and clinical information (sub-scores) and further adjust the weights of the central and periocular features. Our proposed knowledge distillation is able to utilize uncertain samples without hard labels, and at the same time maintains effective performance in the absence of sub-scores. Extensive experiments on a unified cataract image dataset validated the superiority of our new method over state-of-the-art methods.

References

[1] H. Li, J. H. Lim, J. Liu, T. Y. Wong, A. Tan, J. J. Wang, and P. Mitchell, “Image based grading of nuclear cataract by SVM regression,” in *Medical Imaging*, vol. 6915, 2008, pp. 985–992.

[2] S. R. Flaxman, R. R. Bourne, S. Resnikoff, P. Ackland, T. Braithwaite, M. V. Cicinelli, A. Das, J. B. Jonas, J. Keeffe, J. H. Kempen *et al.*, “Global causes of blindness and distance vision impairment 1990-2020: A systematic review and meta-analysis,” *The Lancet Global Health*, vol. 5, no. 12, pp. e1221–e1234, 2017.

[3] L. T. Chylack, M. C. Leske, R. Sperduto, P. Khu, and D. McCarthy, “Lens opacities classification system,” *Archives of Ophthalmology*, vol. 106, no. 3, pp. 330–334, 1988.

[4] J. Thompson and N. Lakhani, “Cataracts,” *Primary Care: Clinics in Office Practice*, vol. 42, no. 3, pp. 409–423, 2015.

[5] K. Pesudovs and D. Elliott, “Cataract morphology, classification, assessment and referral,” *CE Optometry*, vol. 4, no. 2, pp. 55–60, 2001.

[6] J. H. L. Goh, Z. W. Lim, X. Fang, A. Anees, S. Nusinovici, T. H. Rim, C.-Y. Cheng, and Y.-C. Tham, “Artificial intelligence for cataract detection and management,” *The Asia-Pacific Journal of Ophthalmology*, vol. 9, no. 2, pp. 88–95, 2020.

[7] D. Tognetto, R. Giglio, A. L. Vinciguerra, S. Milan, R. Rejdak, M. Rejdak, K. Zaluska-Ogryzek, S. Zweifel, and M. D. Toro, “Artificial intelligence applications and cataract management: A systematic review,” *Survey of Ophthalmology*, vol. 67, no. 3, pp. 817–829, 2022.

[8] H. Li, L. Ko, J. H. Lim, J. Liu, D. W. K. Wong, T. Y. Wong, and Y. Sun, “Automatic opacity detection in retro-illumination images for cortical cataract diagnosis,” in *IEEE International Conference on Multimedia and Expo*, 2008, pp. 553–556.

[9] B. Thylefors, L. Chylack Jr, K. Konyama, K. Sasaki, R. Sperduto, H. Taylor, and S. West, “A simplified cataract grading system / The WHO Cataract Grading Group,” *Ophthalmic Epidemiology*, vol. 9, no. 2, pp. 83–95, 2002.

[10] L. T. Chylack, J. K. Wolfe, D. M. Singer, M. C. Leske, M. A. Bullimore, I. L. Bailey, J. Friend, D. McCarthy, and S.-Y. Wu, “The lens opacities classification system III,” *Archives of Ophthalmology*, vol. 111, no. 6, pp. 831–836, 1993.

[11] V. Mehra and D. Minassian, “A rapid method of grading cataract in epidemiological studies and eye surveys,” *British Journal of Ophthalmology*, vol. 72, no. 11, pp. 801–803, 1988.

[12] K. Yao and J. Flammer, “Relationship cataract density and visual field damage,” *European Journal of Ophthalmology*, vol. 3, no. 1, pp. 1–5, 1993.

[13] X.-Q. Zhang, Y. Hu, Z.-J. Xiao, J.-S. Fang, R. Higashita, and J. Liu, “Machine learning for cataract classification/grading on ophthalmic imaging modalities: A survey,” *Machine Intelligence Research*, vol. 19, no. 3, pp. 184–208, 2022.

[14] H. Li, L. Ko, J. H. Lim, J. Liu, D. W. K. Wong, and T. Y. Wong, “Image based diagnosis of cortical cataract,” in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 3904–3907.

[15] S. Patange and A. Jagadale, "Framework for detection of cataract and gradation according to its severity," in *International Conference on Pervasive Computing*, 2015, pp. 1–3.

[16] X. Gao, H. Li, J. H. Lim, and T. Y. Wong, "Computer-aided cataract detection using enhanced texture features on retro-illumination lens images," in *IEEE International Conference on Image Processing*, 2011, pp. 1565–1568.

[17] X. Gao, D. W. K. Wong, T.-T. Ng, C. Y. L. Cheung, C.-Y. Cheng, and T. Y. Wong, "Automatic grading of cortical and PSC cataracts using retroillumination lens images," in *ACCV*, 2013, pp. 256–267.

[18] X. Gao, D. W. K. Wong, A. W. Aryaputera, Y. Sun, C.-Y. Cheng, C. Cheung, and T. Y. Wong, "Automatic pterygium detection on cornea images to enhance computer-aided cortical cataract grading system," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 4434–4437.

[19] B. E. K. Klein, R. Klein, K. L. P. Linton, Y. L. Magli, and M. W. Neider, "Assessment of cataracts from photographs in the Beaver Dam Eye Study," *Ophthalmology*, vol. 97, no. 11, pp. 1428–1433, 1990.

[20] G. Chen, D. Xiang, B. Zhang, H. Tian, X. Yang, F. Shi, W. Zhu, B. Tian, and X. Chen, "Automatic pathological lung segmentation in low-dose CT image using eigenspace sparse shape composition," *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1736–1749, 2019.

[21] G. Li, C. Li, C. Zeng, P. Gao, and G. Xie, "Region focus network for joint optic disc and cup segmentation," in *AAAI*, vol. 34, no. 01, 2020, pp. 751–758.

[22] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *MICCAI International Workshop*, 2019, pp. 311–320.

[23] M. H. Jensen, D. R. Jørgensen, R. Jalaboi, M. E. Hansen, and M. A. Olsen, "Improving uncertainty estimation in convolutional neural networks using inter-rater agreement," in *MICCAI*, 2019, pp. 540–548.

[24] A. Jungo, R. Meier, E. Ermis, M. Blatti-Moreno, E. Herrmann, R. Wiest, and M. Reyes, "On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation," in *MICCAI*, 2018, pp. 682–690.

[25] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *CVPR*, 2021, pp. 12 341–12 351.

[26] Y. Dong, Q. Wang, Q. Zhang, and J. Yang, "Classification of cataract fundus image based on retinal vascular information," in *International Conference on Smart Health*, 2016, pp. 166–173.

[27] Z. Qiao, Q. Zhang, Y. Dong, and J.-J. Yang, "Application of SVM based on genetic algorithm in classification of cataract fundus images," in *International Conference on Imaging Systems and Techniques*, 2017, pp. 1–5.

[28] M. S. Junayed, M. B. Islam, A. Sadeghzadeh, and S. Rahman, "Cataract-Net: An automated cataract detection system using deep learning for fundus images," *IEEE Access*, vol. 9, pp. 128 799–128 808, 2021.

[29] H. R. Tawfik, R. A. Birry, and A. A. Saad, "Early recognition and grading of cataract using a combined log Gabor/discrete wavelet transform with ANN and SVM," *International Journal of Computer and Information Engineering*, vol. 12, no. 12, pp. 1038–1043, 2018.

[30] H. Wu, J. Lv, and J. Wang, "Automatic cataract detection with multi-task learning," in *International Joint Conference on Neural Networks*, 2021, pp. 1–8.

[31] Y. C. Chow, X. Gao, H. Li, J. H. Lim, Y. Sun, and T. Y. Wong, "Automatic detection of cortical and PSC cataracts using texture and intensity analysis on retro-illumination lens images," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 5044–5047.

[32] Z. Xiao, X. Zhang, R. Higashita, Y. Hu, J. Yuan, W. Chen, and J. Liu, "Gated channel attention network for cataract classification on AS-OCT image," in *Neural Information Processing*, 2021, pp. 357–368.

[33] S. Hu, X. Luan, H. Wu, X. Wang, C. Yan, J. Wang, G. Liu, and W. He, "ACCV: Automatic classification algorithm of cataract video based on deep learning," *BioMedical Engineering OnLine*, vol. 20, pp. 1–17, 2021.

[34] L. T. Chylack, M. C. Leske, D. McCarthy, P. Khu, T. Kashiwagi, and R. Sperduto, "Lens opacities classification system II (LOCS II)," *Archives of Ophthalmology*, vol. 107, no. 7, pp. 991–997, 1989.

[35] J. Sparrow, A. Bron, N. Brown, W. Ayliffe, and A. Hill, "The Oxford clinical cataract classification and grading system," *International Ophthalmology*, vol. 9, pp. 207–225, 1986.

[36] L. Zhang, J. Li, H. Han, B. Liu, J. Yang, Q. Wang *et al.*, "Automatic cataract detection and grading using deep convolutional neural network," in *International Conference on Networking, Sensing and Control*, 2017, pp. 60–65.

[37] J. Ran, K. Niu, Z. He, H. Zhang, and H. Song, "Cataract detection and grading based on combination of deep convolutional neural network and random forests," in *International Conference on Network Infrastructure and Digital Content*, 2018, pp. 155–159.

[38] H. Zhang, K. Niu, Y. Xiong, W. Yang, Z. He, and H. Song, "Automatic cataract grading methods based on deep learning," *Computer Methods and Programs in Biomedicine*, vol. 182, p. 104978, 2019.

[39] M. S. M. Khan, M. Ahmed, R. Z. Rasel, and M. M. Khan, "Cataract detection using convolutional neural network with VGG-19 model," in *IEEE World AI IoT Congress*, 2021, pp. 0209–0212.

[40] Y.-C. Tham, J. H. L. Goh, A. Anees, X. Lei, T. H. Rim, M.-L. Chee, Y. X. Wang, J. B. Jonas, S. Thakur, Z. L. Teo *et al.*, "Detecting visually significant cataract using retinal photograph-based deep learning," *Nature Aging*, vol. 2, no. 3, pp. 264–271, 2022.

[41] C. Xu, X. Zhu, W. He, Y. Lu, X. He, Z. Shang, J. Wu, K. Zhang, Y. Zhang, X. Rong *et al.*, "Fully deep learning for slit-lamp photo based nuclear cataract grading," in *MICCAI*, 2019, pp. 513–521.

[42] X. Zhang, Z. Xiao, R. Higashita, W. Chen, J. Yuan, J. Fang, Y. Hu, and J. Liu, "A novel deep learning method for nuclear cataract classification based on anterior segment optical coherence tomography images," in *International Conference on Systems, Man, and Cybernetics*, 2020, pp. 662–668.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[44] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.

[45] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing Transformers and CNNs for medical image segmentation," in *MICCAI*, 2021, pp. 14–24.

[46] C.-M. Feng, Y. Yan, G. Chen, H. Fu, Y. Xu, and L. Shao, "Accelerated multi-modal MR imaging with Transformers," *arXiv preprint arXiv:2106.14248*, 2021.

[47] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision Transformers for multimodal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[48] G. v. Tulder, Y. Tong, and E. Marchiori, "Multi-view analysis of unregistered medical images using cross-view Transformers," in *MICCAI*, 2021, pp. 104–113.

[49] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 200–14 213, 2021.

[50] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[51] Q. Wang, L. Zhan, P. Thompson, and J. Zhou, "Multimodal learning with incomplete modalities by knowledge distillation," in *ACM SIGKDD*, 2020, pp. 1828–1838.

[52] Y. Yang, D.-C. Zhan, X.-R. Sheng, and Y. Jiang, "Semi-supervised multi-modal learning with incomplete modalities," in *IJCAI*, 2018, pp. 2998–3004.

[53] M. Hu, M. Maillard, Y. Zhang, T. Ciceri, G. La Barbera, I. Bloch, and P. Gori, "Knowledge distillation from multi-modal to mono-modal segmentation networks," in *MICCAI*, 2020, pp. 772–781.

[54] X. Xing, Z. Chen, M. Zhu, Y. Hou, Z. Gao, and Y. Yuan, "Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading," in *MICCAI*, 2022, pp. 636–646.

[55] Q. Yang, X. Guo, Z. Chen, P. Y. Woo, and Y. Yuan, "D2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities," *IEEE Transactions on Medical Imaging*, 2022.

[56] S. H. Feil, A. S. Crandall, and R. J. Olson, "Astigmatic decay following small incision, self-sealing cataract surgery," *Journal of Cataract & Refractive Surgery*, vol. 20, no. 1, pp. 40–43, 1994.

[57] U. Guthauser and J. Flammer, "Quantifying visual field damage caused by cataract," *American Journal of Ophthalmology*, vol. 106, no. 4, pp. 480–484, 1988.

[58] H. Miyashita, N. Hatsusaka, E. Shibuya, N. Mita, M. Yamazaki, T. Shibata, H. Ishida, Y. Ukai, E. Kubo, and H. Sasaki, "Association between ultraviolet radiation exposure dose and cataract in Han people living in China and Taiwan: A cross-sectional study," *PLoS One*, vol. 14, no. 4, p. e0215338, 2019.

[59] D. B. Elliott, J. Gilchrist, and D. Whitaker, "Contrast sensitivity and glare sensitivity changes with three types of cataract morphology: are these techniques necessary in a clinical evaluation of cataract?" *Ophthalmic and Physiological Optics*, vol. 9, no. 1, pp. 25–30, 1989.

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*,

"An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[63] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[64] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[65] T. D. Keenan, Q. Chen, E. Agrón, Y.-C. Tham, J. H. L. Goh, X. Lei, Y. P. Ng, Y. Liu, X. Xu, C.-Y. Cheng *et al.*, "DeepLensNet: Deep learning automated diagnosis and quantitative classification of cataract type and severity," *Ophthalmology*, vol. 129, no. 5, pp. 571–584, 2022.

[66] R. R. Maaliw, A. S. Alon, A. C. Lagman, M. B. Garcia, M. V. Abante, R. C. Belleza, J. B. Tan, and R. A. Maaño, "Cataract detection and grading using ensemble neural networks and transfer learning," in *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2022, pp. 0074–0081.

[67] Y. Elloumi, "Cataract grading method based on deep convolutional neural networks and stacking ensemble learning," *International Journal of Imaging Systems and Technology*, vol. 32, no. 3, pp. 798–814, 2022.

[68] A. Sohail, H. Qayyum, F. Hassan, and A. U. Rahman, "CataractEyeNet: A novel deep learning approach to detect eye cataract disorder," in *Proceedings of International Conference on Information Technology and Applications: ICITA 2022*. Springer, 2023, pp. 63–75.

[69] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[70] F. Xiong, C. Shen, and X. Wang, "Generalized knowledge distillation for unimodal glioma segmentation from multimodal models," *Electronics*, vol. 12, no. 7, p. 1516, 2023.

[71] Y. Choi, M. A. Al-Masni, K.-J. Jung, R.-E. Yoo, S.-Y. Lee, and D.-H. Kim, "A single stage knowledge distillation network for brain tumor segmentation on limited MR image modalities," *Computer Methods and Programs in Biomedicine*, p. 107644, 2023.

[72] S. Wang, Z. Yan, D. Zhang, H. Wei, Z. Li, and R. Li, "Prototype knowledge distillation for medical segmentation with missing modality," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[73] X. Xing, Z. Chen, Y. Hou, and Y. Yuan, "Gradient modulated contrastive distillation of low-rank multi-modal knowledge for disease diagnosis," *Medical Image Analysis*, p. 102874, 2023.