Refining Intraocular Lens Power Calculation: A Multi-modal Framework Using Cross-layer Attention and Effective Channel Attention

Qian Zhou[1], Hua Zou[1, ✉], Zhongyuan Wang[1], Haifeng Jiang[2], and Yong Wang[2]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] Aier Eye Hospital of Wuhan University, Wuhan University, Wuhan, China
zouhua@whu.edu.cn

Abstract. Selecting the appropriate power for intraocular lenses (IOLs) is crucial for the success of cataract surgeries. Traditionally, ophthalmologists rely on manually designed formulas like "Barrett" and "Hoffer Q" to calculate IOL power. However, these methods exhibit limited accuracy since they primarily focus on biometric data such as axial length and corneal curvature, overlooking the rich details in preoperative images that reveal the eye's internal anatomy. In this study, we propose a novel deep learning model that leverages multi-modal information for accurate IOL power calculation. In particular, to address the low information density in optical coherence tomography (OCT) images (*i.e.*, most regions are with zero pixel values), we introduce a cross-layer attention module to take full advantage of hierarchical contextual information to extract comprehensive anatomical features. Additionally, the IOL powers given by traditional formulas are taken as prior knowledge to benefit model training. The proposed method is evaluated on a self-collected dataset consisting of 174 samples and compared with other approaches. The experimental results demonstrate that our approach significantly surpasses competing methods, achieving a mean absolute error of just 0.367 diopters (D). Impressively, the percentage of eyes with a prediction error within ± 0.5 D achieves 84.1%. Furthermore, extensive ablation studies are conducted to validate each component's contribution and identify the biometric parameters most relevant to accurate IOL power calculation. Codes will be available at https://github.com/liyiersan/IOL.

Keywords: Intraocular lens power calculation · Multi-modal learning · Attention mechanism.

1 Introduction

Cataracts are the leading cause of blindness and vision impairment globally, responsible for approximately 45% of blindness in adults over the age of 50 [18]. Currently, cataract surgery stands as the cornerstone of treatment, in which choosing a proper refractive power of an intraocular lens (IOL) plays a pivotal role in determining the outcome and significantly impacts the postoperative visual acuity [7].

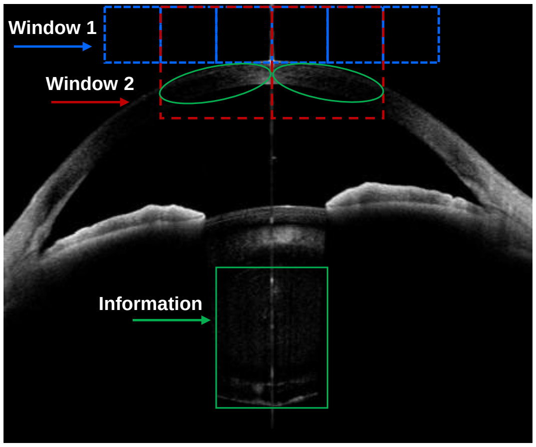

Fig. 1. Illustration of dumb windows, which contain no valid information. The red and blue boxes are sliding windows with different kernel sizes, and the areas within green are informative regions.

Over the past decades, many manually designed formulas such as Barrett Universal [2], Hoffer Q [9], Holladay [10], and SRK/T [15] have been developed to calculate IOL power. Despite widespread use, these formulas are still far from being perfect due to certain limitations even in normal unoperated eyes [16]. Firstly, they mainly focus on biometric measurements such as axial length (AL) and corneal curvature, neglecting the imaging information. Preoperative multi-view optical coherence tomography (OCT) images, for instance, can provide detailed insights into the retinal thickness, lens position, anterior chamber depth, and other anatomical structures, which are vital for accurate calculations [1]. Secondly, the reliance of these formulas on calculating the effective lens position (ELP) introduces variability, as different formulas estimate ELP diversely, making them suitable only for specific eye types. For example, Hoffer Q is best for short AL (< 22.0 mm) and SRK/T for long AL (> 26.0 mm), underscoring the need for more versatile approaches. Recently, some computer-aided approaches [22,3,12,13] have been proposed, but they still focus on single-modal data and are suboptimal because of simplistic model designs, *e.g.*, basic multi-layer perceptrons (MLPs). These easy models may fail to learn complex patterns within multi-modal data.

In this paper, we propose a new deep learning framework that takes full advantage of multi-modal data for accurate and reliable IOL power calculation. The framework consists of a dual-branch encoder that takes both preoperative multi-view OCT images and biometric parameters as input, a feature fusion network for complementary information aggregation, and a predictive head for power calculation. Especially different from existing methods, we concentrate on imaging data. Notably, as shown in Fig. 1, OCT images display low information density, which is caused by the numerous zero-value pixels in background areas and affects the efficiency of feature extraction. To address this, we introduce a cross-layer attention (CLA) module to better aggregate contextual information among different layers. For biometric data, we incorporate the calculated result using classical formulas as prior knowledge. Moreover, an auxiliary prediction loss is introduced in the biometric encoder to benefit model training. The enhanced features from both multi-view OCT images and biometric data are then concatenated, and processed through the fusion network with the effective channel attention (ECA) module for multi-modal information fusion. Finally, they are fed into the predictive network for precise IOL power calculation.

The main contributions of this work are as follows: (1) We develop a pioneering deep learning approach that exploits multi-modal data for end-to-end IOL power calculation, eliminating the necessity for ELP estimation. To the best of our knowledge, this is the first application of multi-modal data in this domain. (2) The CLA module is proposed to effectively harness hierarchical correlations within OCT images. Moreover, we utilize prior knowledge provided by conventional formulas and introduce an auxiliary prediction loss for advanced biometric data encoding. (3) Experimental results show that our approach outperforms existing formula-based and machine learning-based methods by a large margin, demonstrating the potential for accurate and reliable IOL power calculation in clinical practice. Importantly, we have also identified key biometric parameters for IOL power calculation, which can provide critical insights for future formula development.

2 Methodology

2.1 Framework Overview

As can be seen in Fig. 2, the proposed framework is composed of three main parts: a dual-branch encoder for representation learning, a fusion network for multi-modal information integration, and MLPs for power prediction. For the image encoder, we employ the RepLKNet [4] as the backbone for its large kernel size and incorporate the CLA module for improved feature extraction. For biometric data encoding, an MLP with prediction loss is adopted. After that, features from both imaging and biometric data are then concatenated and fed into the fusion network, in which effective channel attention (ECA) [20] is used to explore the multi-modal correlations. Finally, the IOL power prediction is made through fully connected (FC) layers.

2.2 Dual-branch Encoder

Low Information Density In OCT images, as depicted in Fig. 1, a significant portion of the pixels are part of the background and have zero values. When applying convolutions on sliding windows, this results in windows that are completely zero-filled, contributing minimal informative value. We refer to these zero-value windows as "dumb windows" and describe the prevalent zero-value

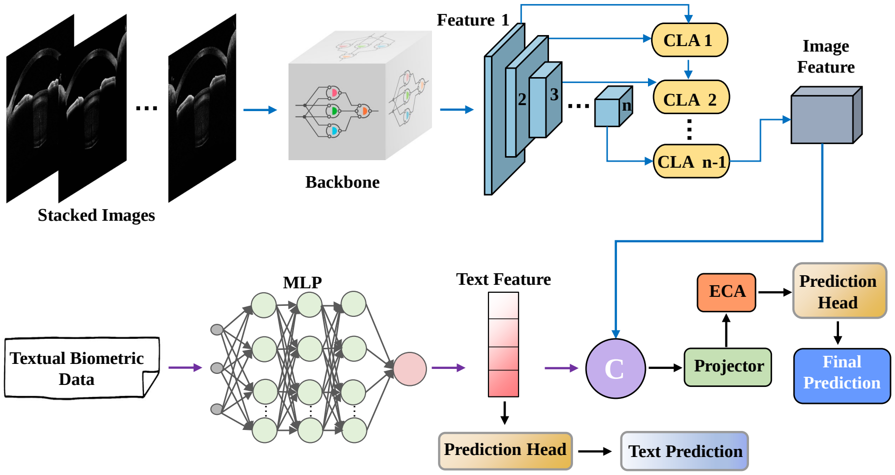

Fig. 2. Pipeline of the proposed framework. The two branches in encoders process biometric data and images, respectively. 'C' in the circle means channel-wise feature concatenation. The prediction head for the biometric encoder will be dropped and only the final prediction will be preserved during inference.

pixels as "low information density". These dumb windows may have negative impacts on backward gradients as they do not provide meaningful direction for parameter updates. Specifically, ReLU activation often results in zero responses to dumb windows and this may affect feature aggregation when pooling. Meanwhile, low information density makes it difficult for models to obtain meaningful representations.

Large Kernel Backbone To reduce dumb windows, a larger kernel size can be a feasible solution since it enables the inclusion of more non-zero pixels in each window. When adding the kernel size from 3×3 to 31×31, the dumb windows decrease from 77% to 52%. Besides, a larger convolutional kernel implies a broader receptive field, allowing the model to access more information. This aids in accurately locating anatomical structures and exploring details within the images. Therefore, we adopt RepLKNet-31B [4] as the image backbone considering its high performance with depth-wise large kernel design.

Cross-layer Attention Module However, while effective against dumb windows, the large kernel design doesn't address the core issue of low information density in images. This characteristic is inherent to the image itself and remains unaffected by the model choice. That is, even with a large-kernel design, the extracted features may include many irrelevant characteristics, impacting the model's performance. In such a situation, we introduce the cross-layer attention (CLA) module to suppress the unnecessary features. Assuming two successive stages with features F_i and F_{i+1}, F_i contains detailed structural information from a shallower layer while F_{i+1} from a deeper layer holds higher-level features. Therefore, F_i is utilized to generate the spatial attention weights on F_{i+1} to enhance focus on relevant details and improve feature integration. A 3 × 3 convolution is introduced to reduce the spatial size of F_i to match F_{i+1} as Eq. (1). Detailed architectures of CLA can be found in the supplementary file (Fig.1).

$$\begin{aligned} F_{in} &= f^{3\times3}(F_i) \\ M_s &= \text{Sigmoid}\left(f^{7\times7}\left(\left[\text{Avg}(F_{in}), f^{1\times1}(F_{in}), \text{Max}(F_{in})\right]\right)\right) \\ M_{out} &= F_{i+1} + F_{i+1} \odot M_s \end{aligned} \tag{1}$$

where $f^{k\times k}$ represents the convolution with a filter size of $k \times k$ and $\odot$ means element-wise product. Avg and Max are average and max pooling on channels.

Biometric Encoder For biometric data encoding, we take biometric data and prior results (*i.e.*, power calculated by traditional formulas) as input and apply an MLP to extract features. Besides, an independent prediction head is introduced to guide the biometric encoding as Eq. (2).

$$L_{bio} = L_{\text{MSE}}(\text{bio\_preds}, \text{gts}) \tag{2}$$

where L_{MSE} is the mean square error loss between the predicted IOL power using biometric features and the ground truth.

2.3 Fusion Network

The image features and biometric features are then concatenated to form comprehensive representations of multi-modal data. After that, they are passed through a projector, further refining the multi-modal representations and decreasing the dimension. To effectively explore the correlations among channels, effective channel attention (ECA) [20] is employed, whose details are shown in the supplementary file (Fig.2). This strategy enables the model to focus on the most informative features by dynamically adjusting the importance of each channel based on the learned correlations. The final power prediction is given by the prediction head, which is a fully connected layer.

The model is trained end-to-end with weighted prediction loss as Eq. (3).

$$Loss = L_{\text{MSE}}(\text{final\_preds}, \text{gts}) + \alpha L_{bio} \tag{3}$$

where L_{MSE} is the mean square error loss between the final prediction and the ground truth. α is a hyper-parameter and set to 0.5 by default.

2.4 Implemantation Details

All experiments are implemented with Pytorch on 8× RTX 4090 GPUs. The images are resized to 512 × 512 and center cropped to 448 × 448. We adopt

Adam as the optimizer with an initial learning rate of 0.001 and $\beta_1 = 0.9$, $\beta_2 = 0.99$. The mini-batch size is set to 16. The models are trained for 100 epochs, and the learning rate will decay by 0.1 every 20 epochs. Besides, the dataset is randomly split to 80% for training and 20% for testing with 5-fold cross-validation to produce more solid results.

3 Experiments

3.1 Datasets and Evaluation Metrics

We have collected a multi-modal cataract dataset from a local eye hospital, comprising 174 eyes from 117 patients. This dataset encompasses OCT images with 16 different views of 2D scans acquired from the CASIA2 device and detailed biometric measurements. These measurements contain axial length (AL), corneal curvature (K1 and K2), anterior chamber depth (ACD), lens thickness (LT), and white-to-white (WTW) distance, as well as demographic information, including age, gender, and preoperative visual acuity. The actual IOL power (ground truth) is determined by 3 experienced ophthalmologists through an analysis of one-month post-surgery optometry refraction and the specific IOL power used in cataract surgery. We employ three metrics for evaluation following [3,19]: Mean Absolute Error (MAE), Median Absolute Error (MedAE), and overall prediction accuracy. For simplicity, a prediction is considered accurate if the MAE falls within a range of ± 0.5 Diopters (D). The 0.5 is chosen for two reasons: 1) Predictions with MAE $\leq$ 0.5 D are deemed clinically acceptable [3,19], and accuracy reflects the proportion of clinically useful predictions. 2) The ground truth in our dataset is accurate to 0.5 D increments.

Table 1. Quantitative prediction results on the collected dataset. AutoML means the AI-driven models using autogluon. MMT represents multi-modal transformers.

Type	Methods	MAE (↓)	MedAE (↓)	Accuracy (↑)
Formulas	Barrett Universal [2]	0.616 ± 0.267	0.406 ± 0.062	0.618 ± 0.077
	Hoffer Q [9]	0.932 ± 0.096	0.545 ± 0.043	0.447 ± 0.043
	Holladay [10]	0.508 ± 0.067	0.452 ± 0.049	0.547 ± 0.080
	SRK/T [15]	0.547 ± 0.074	0.466 ± 0.101	0.517 ± 0.061
AutoML	Tabular [5]	0.705 ± 0.281	0.457 ± 0.069	0.682 ± 0.063
	MultiModal [17]	0.942 ± 0.021	0.542 ± 0.062	0.452 ± 0.053
MMT	CLIP [14]	1.386 ± 0.245	1.325 ± 0.083	0.230 ± 0.096
	ViLT [11]	1.172 ± 0.413	1.045 ± 0.063	0.266 ± 0.095
	BEiT-3 [21]	2.727 ± 0.188	2.005 ± 0.124	0.180 ± 0.040
Ours	Full (image + text)	$\mathbf{0.367 \pm 0.040}$	$\mathbf{0.333 \pm 0.086}$	$\mathbf{0.841 \pm 0.052}$
	Variant-1 (image only)	0.459 ± 0.039	0.373 ± 0.055	0.706 ± 0.042
	Variant-2 (bio data only)	0.496 ± 0.054	0.417 ± 0.059	0.671 ± 0.051
	MLP (no prior)	0.542 ± 0.053	0.436 ± 0.071	0.624 ± 0.073

3.2 Quantitative Performance

We have compared our approach against traditional formulas (Barrett Universal [2], Hoffer Q [9], Holladay [10], and SRK/T [15]) and AI-driven models using autogluon. The autogluon models are TabularPredictor [5] and MultiModalPredictor [17] with "good_quality" and "high_quality", respectively. Besides, the emerging multi-modal transformers (MMTs) are also finetuned for IOL power prediction, including CLIP [14], ViLT [11], and BEiT-3 [21]. To further validate the efficacy, we also design two variants of our method: one (variant-1) employing only the biometric encoder and the other (variant-2) utilizing solely the image encoder. These variants are compared against a naive MLP prediction model without prior information to provide a comprehensive assessment.

From Table 1, it can be seen that our approach achieves the best performance with a significant margin over other methods. The naive MLP model achieves the worst performance in all variants, indicating the insufficiency of simple models to capture the intricate relationships. Variant-1 shows better performance than naive MLP, which demonstrates the effectiveness of introducing calculated powers by traditional formulas. Variant-2 secures performance closely trailing our method, underscoring the significant role of images. Interestingly, we find that TabularPredictor outperforms MultiModalPredictor. The primary reason for this is that the multi-view OCT images should be carefully treated as the numerous zero-value pixels in background areas may act as noise to representation learning. In addition, all multi-modal transformers perform worse on IOL power prediction. This may be attributed to three reasons: 1) Low information density results in many dump patches, which may be computationally invalid in optimizing the parameters of bottom transformer layers [6]. 2) The text encoders in MMTs trained with natural language texts may not be as effective when applied to biomedical data. 3) Transformers can easily overfit small-scale data since they are more data-hungry than CNNs.

Table 2. Ablation study results using multi-modal data. 'w/o' means omitting the corresponding module.

Model	MAE (↓)	MedAE (↓)	Accuracy (↑)
Full	$\mathbf{0.367 \pm 0.040}$	$\mathbf{0.333 \pm 0.086}$	$\mathbf{0.841 \pm 0.052}$
ResNet-50 [8] Backbone	0.395 ± 0.064	0.367 ± 0.074	0.771 ± 0.034
ResNet-101 [8] Backbone	0.472 ± 0.067	0.453 ± 0.066	0.653 ± 0.044
ResNet-152 [8] Backbone	0.583 ± 0.052	0.492 ± 0.041	0.582 ± 0.040
w/o CLA	0.426 ± 0.061	0.423 ± 0.067	0.735 ± 0.087
w/o ECA	0.397 ± 0.046	0.382 ± 0.087	0.784 ± 0.033
w/o auxiliary loss	0.388 ± 0.051	0.364 ± 0.054	0.806 ± 0.062

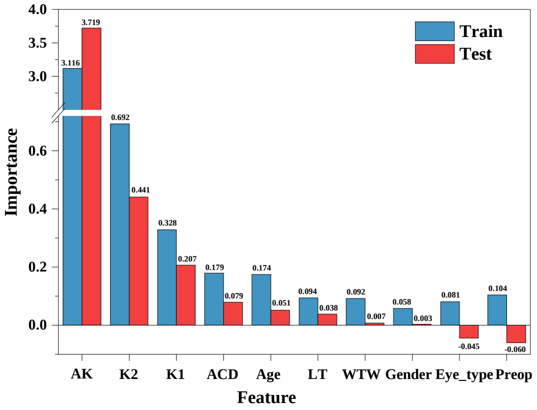

Fig. 3. Visualization of feature importance. The train and test mean that the importance is calculated in the train dataset and test dataset, respectively.

3.3 Ablation Study

Effectiveness of Each Component The ablation studies are conducted to verify the contribution of each component and the results are listed in Table 2. When replacing the large kernel design with ResNet-50 [8], the overall accuracy drops from 84% to 77%. Additionally, the ResNet backbones appear to be overfitting on the collected dataset like MMTs, as indicated by the performance drop observed with ResNet-101 and ResNet-152. Our employed RepLKNet backbone takes advantage of a large-kernel design, which can effectively capture the informative regions as shown in the supplementary file (Fig.3). The CLA module has a more significant impact on performance compared to the ECA module, as it is highly related to the quality of feature extraction. As for biometric auxiliary prediction loss, it has a slight influence on model performance.

Importance of Biometric Data To identify the most relevant biometric parameters for IOL power calculation, we compare their feature importance using the naive MLP without prior input. The importance is defined as the performance drop when one column's values are randomly shuffled[3] across rows during inference. Through the results shown in Fig. 3, we find that AL, K2, K1, and ACD emerge as essential elements, exerting a great influence. Conversely, eye type is identified as having a negligible effect on power predictions. It is worth noting that the results uncover a positive correlation between age and performance, a factor not fully appreciated in traditional formulas.

[3] https://explained.ai/rf-importance/#4

4 Conlusion

In conclusion, we present an end-to-end deep learning framework that significantly advances the accuracy of IOL power calculation without ELP estimation. Comprehensive and complementary representations can be obtained by ingeniously leveraging preoperative multi-view OCT images and biometric measurements. The integration of the CLA modules enables precise exploitation of cross-layer correlations in OCT images, effectively overcoming challenges of low information density. Additionally, we employ ECA modules for effective multimodal information aggregation. Extensive experiments have proved the effectiveness and superiority of our method. We also analyze the biometric parameters most relevant to IOL power calculation, offering invaluable insights for the development of future calculation formulas.

Acknowledgments. This work is partially supported by Bingtuan Science and Technology Program (No. 2022DB005 and 2019BC008) and Key Research and Development Program of Hubei Province (2022BCA009).

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

References

1. An, Y., Kang, E.K., Kim, H., Kang, M.J., Byun, Y.S., Joo, C.K.: Accuracy of swept-source optical coherence tomography based biometry for intraocular lens power calculation: a retrospective cross–sectional study. BMC ophthalmology **19**, 1–7 (2019)
2. Barrett, G.D.: An improved universal theoretical formula for intraocular lens power prediction. Journal of Cataract & Refractive Surgery **19**(6), 713–720 (1993)
3. Carmona González, D., Palomino Bautista, C.: Accuracy of a new intraocular lens power calculation method based on artificial intelligence. Eye **35**(2), 517–522 (2021)
4. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11963–11975 (2022)
5. Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., Smola, A.: Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint arXiv:2003.06505 (2020)
6. Fan, C., Hou, S., Huang, Y., Yu, S.: Exploring deep models for practical gait recognition. arXiv preprint arXiv:2303.03301 (2023)
7. Gatinel, D., Debellemanière, G., Saad, A., Dubois, M., Rampat, R.: Determining the theoretical effective lens position of thick intraocular lenses for machine learning–based iol power calculation and simulation. Translational Vision Science & Technology **10**(4), 27–27 (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

9. Hoffer, K.J.: The hoffer q formula: a comparison of theoretic and regression formulas. Journal of Cataract & Refractive Surgery **19**(6), 700–712 (1993)
10. Holladay, J.T., Musgrove, K.H., Prager, T.C., Lewis, J.W., Chandler, T.Y., Ruiz, R.S.: A three-part system for refining intraocular lens power calculations. Journal of Cataract & Refractive Surgery **14**(1), 17–24 (1988)
11. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
12. Langenbucher, A., Cayless, A., Szentmáry, N., Weisensee, J., Wendelstein, J., Hoffmann, P.: Prediction of total corneal power from measured anterior corneal power on the iolmaster 700 using a feedforward shallow neural network. Acta Ophthalmologica **100**(5), e1080–e1087 (2022)
13. Nemeth, G., Kemeny-Beke, A., Modis Jr, L.: Comparison of accuracy of different intraocular lens power calculation methods using artificial intelligence. European Journal of Ophthalmology **32**(1), 235–241 (2022)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
15. Retzlaff, J.A., Sanders, D.R., Kraff, M.C.: Development of the srk/t intraocular lens implant power calculation formula. Journal of Cataract & Refractive Surgery **16**(3), 333–340 (1990)
16. Savini, G., Taroni, L., Hoffer, K.J.: Recent developments in intraocular lens power calculation methods—update 2020. Annals of translational medicine **8**(22) (2020)
17. Shi, X., Mueller, J., Erickson, N., Li, M., Smola, A.J.: Benchmarking multimodal automl for tabular data with text fields. vol. 35 (2021)
18. Steinmetz, J.D., Bourne, R.R., Briant, P.S., Flaxman, S.R., Taylor, H.R., Jonas, J.B., Abdoli, A.A., Abrha, W.A., Abualhasan, A., Abu-Gharbieh, E.G., et al.: Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an analysis for the global burden of disease study. The Lancet Global Health **9**(2), e144–e160 (2021)
19. Stopyra, W., Langenbucher, A., Grzybowski, A.: Intraocular lens power calculation formulas—a systematic review. Ophthalmology and Therapy **12**(6), 2881–2902 (2023)
20. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020)
21. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023)
22. Wei, L., Song, Y., He, W., Chen, X., Ma, B., Lu, Y., Zhu, X.: Accuracy improvement of iol power prediction for highly myopic eyes with an xgboost machine learning-based calculator. Frontiers in Medicine **7**, 592663 (2020)