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

Qian Zhou, Hua Zou*, Zhongyuan Wang, Haifeng Jiang, and Yong Wang
Wuhan University, Wuhan, China
zouhua@whu.edu.cn

1. INTRODUCTION

Cataract surgeries rely heavily on selecting the appropriate intraocular lens (IOL) power, which directly impacts postoperative visual outcomes. Traditional methods, such as the Barrett and Hoffer Q formulas, primarily use biometric measurements while neglecting preoperative image data, leading to limited accuracy. This study introduces a novel multi-modal deep learning framework that integrates optical coherence tomography (OCT) images with biometric data. Using RepLKNet as the backbone, along with cross-layer attention (CLA) for multi-scale feature refinement and effective channel attention (ECA) for multi-modal feature fusion, the proposed framework achieves a mean absolute error of 0.367 diopters, outperforming other approaches significantly.

2. METHOD

2.1 Framework

As shown in Figure 1, the proposed framework has three components: a dual-branch encoder, a fusion network, and a prediction head. OCT images are processed using RepLKNet with cross-layer attention (CLA) , while biometric data is encoded by an MLP. Effective channel attention (ECA) are used for feature extraction and fusion. The fused features are then passed through fully connected layers for IOL power prediction.

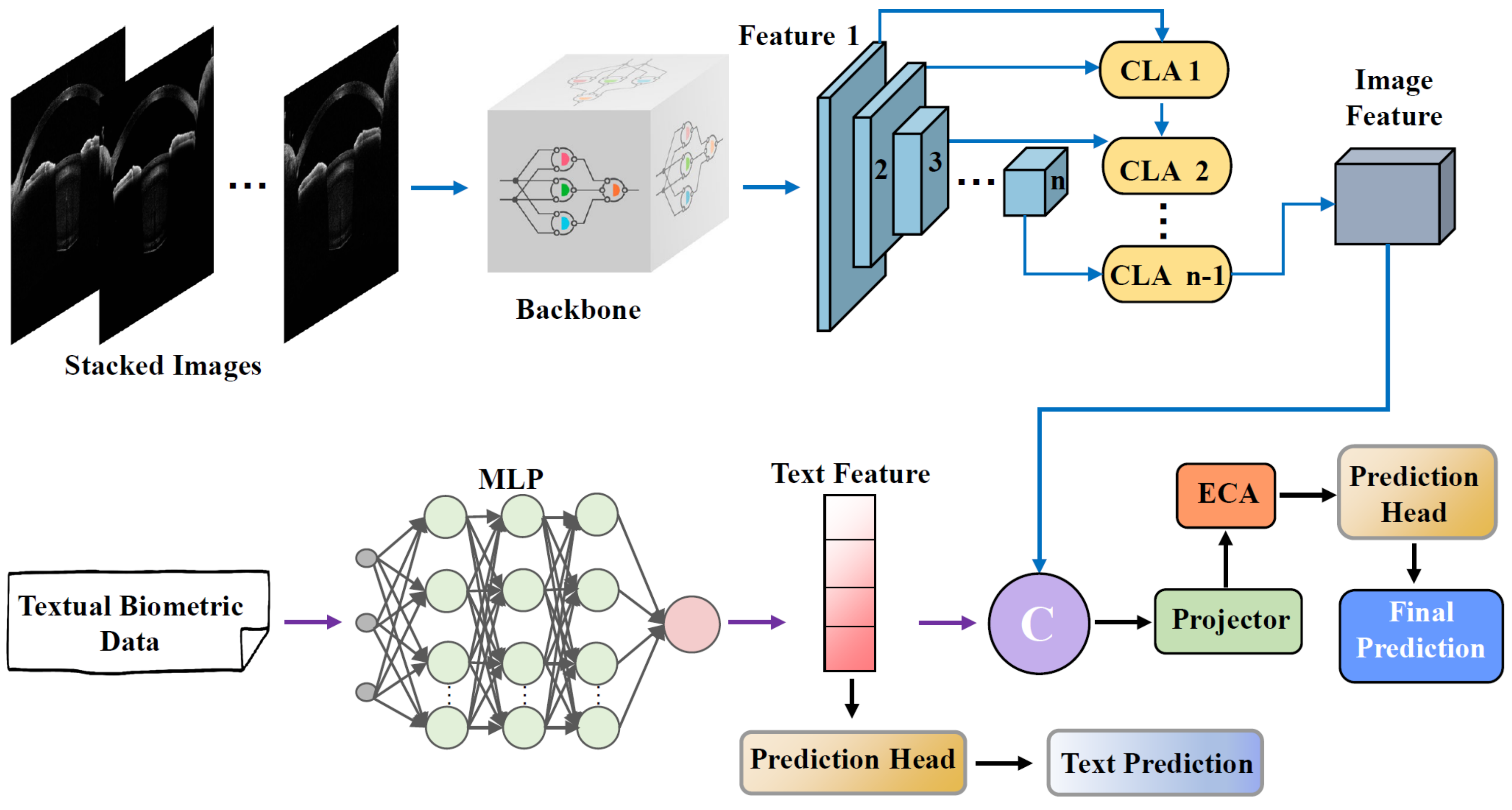

Figure 1. Overview of the proposed multi-modal framework. For OCT images, we use a large-kernel backbone with CLA to handle low-information density, while a simple MLP processes the biometric data.

2.2 Image-Encoder

As shown in Figure 2, OCT images have low-information density. To address this, the image encoder uses RepLKNet with large kernels to capture relevant details. Cross-layer attention (CLA) further enhances multi-scale feature extraction, preserving key structural information.

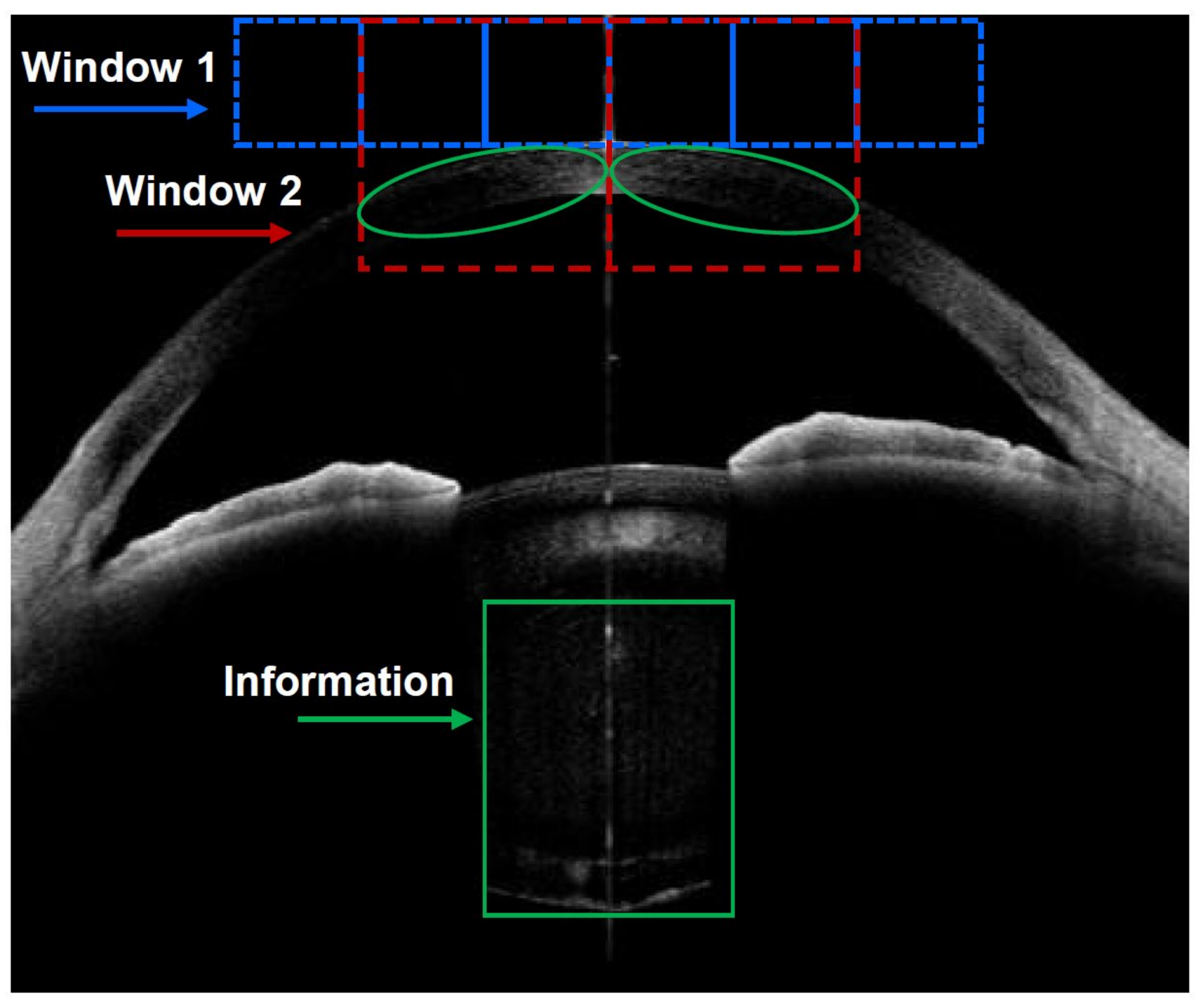

Figure 2. Illustration of low-information density in OCT images. Most regions are with sparse or no useful information.

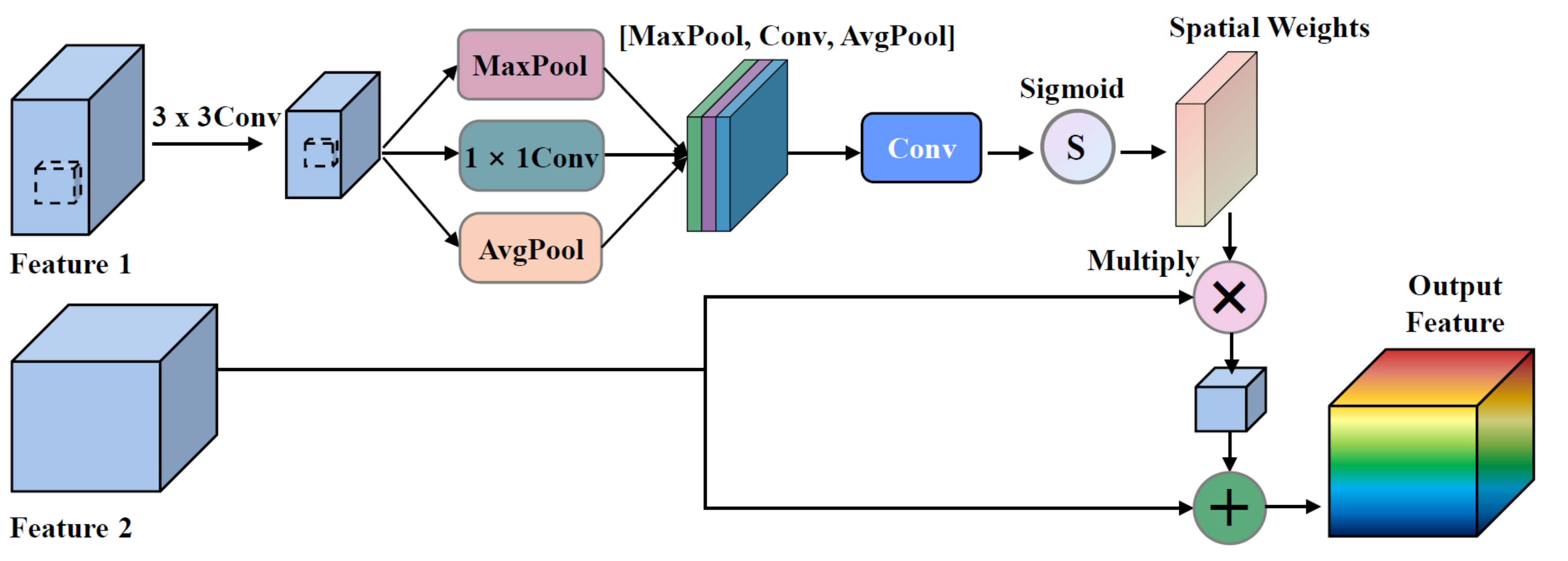

Figure 3. Structure of the CLA module. "Feature 1" from last layer generates spatial weights, while "Feature 2" represents the current layer's features.

2.3 Biometric Data Encoding

The biometric data encoding branch processes key measurements like axial length and corneal curvature using an MLP. An auxiliary loss is applied during training to enhance feature extraction. These features are then fused with OCT image data for multi-modal integration.

2.3 Biometric Data Encoding

The biometric data encoding branch processes key measurements like axial length and corneal curvature using an MLP. An auxiliary loss is applied during training to enhance feature extraction. These features are then fused with OCT image data for multi-modal integration.

2.4 Fusion Network

The fusion network combines the features extracted from both the OCT images and biometric data. After concatenation, the fused features are refined using effective channel attention (ECA) to capture important multi-modal correlations. This mechanism dynamically adjusts the feature importance across channels, enhancing the overall representation.

$$L = L_{MSE}(final\_preds, gts) + \alpha L_{MSE}(bio\_preds, gts)$$

Here, $final\_preds$ are the output given by the whole model, bio_preds are the predictions of biometric branch, and gts are the ground truths. α is set to 0.5, L is training loss of the whole model.

3. EVALUATION

3.1 Datasets and Metrics

We use a self-collected dataset of 174 eyes from 117 patients, including OCT images and biometric measurements. The ground truth IOL power is determined by three ophthalmologists. Model performance is evaluated using Mean Absolute Error (MAE), Median Absolute Error (MedAE), and prediction accuracy, with an MAE within ±0.5 diopters considered clinically acceptable.

3.2 Quantitative Performance

The proposed method is compared with other approaches on the collected dataset.

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

3.3 Qualitative Analysis

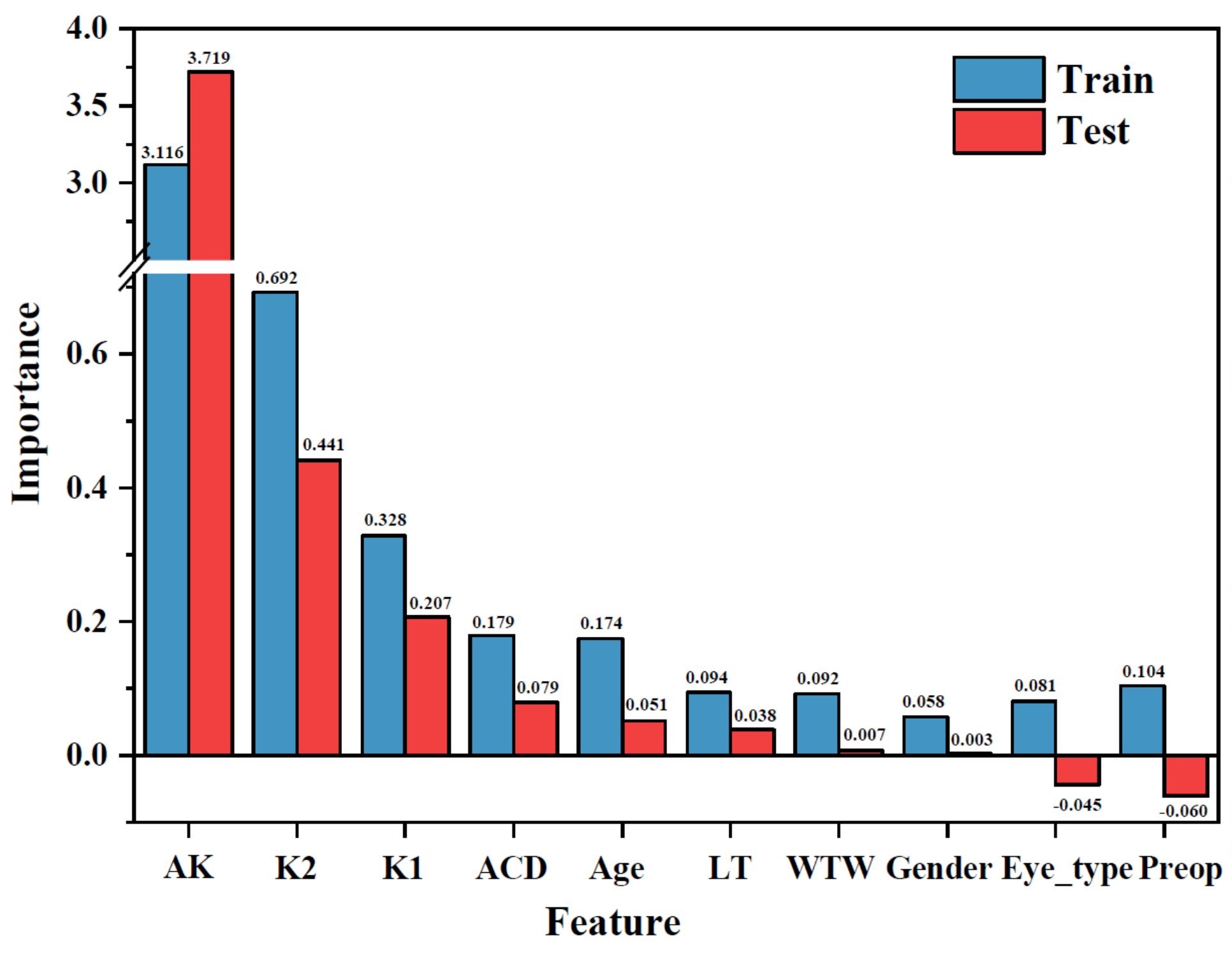

Figure 4. Visualization of feature importance of each biometric data. The train and test mean that the importance is calculated in the train dataset and test dataset, respectively.

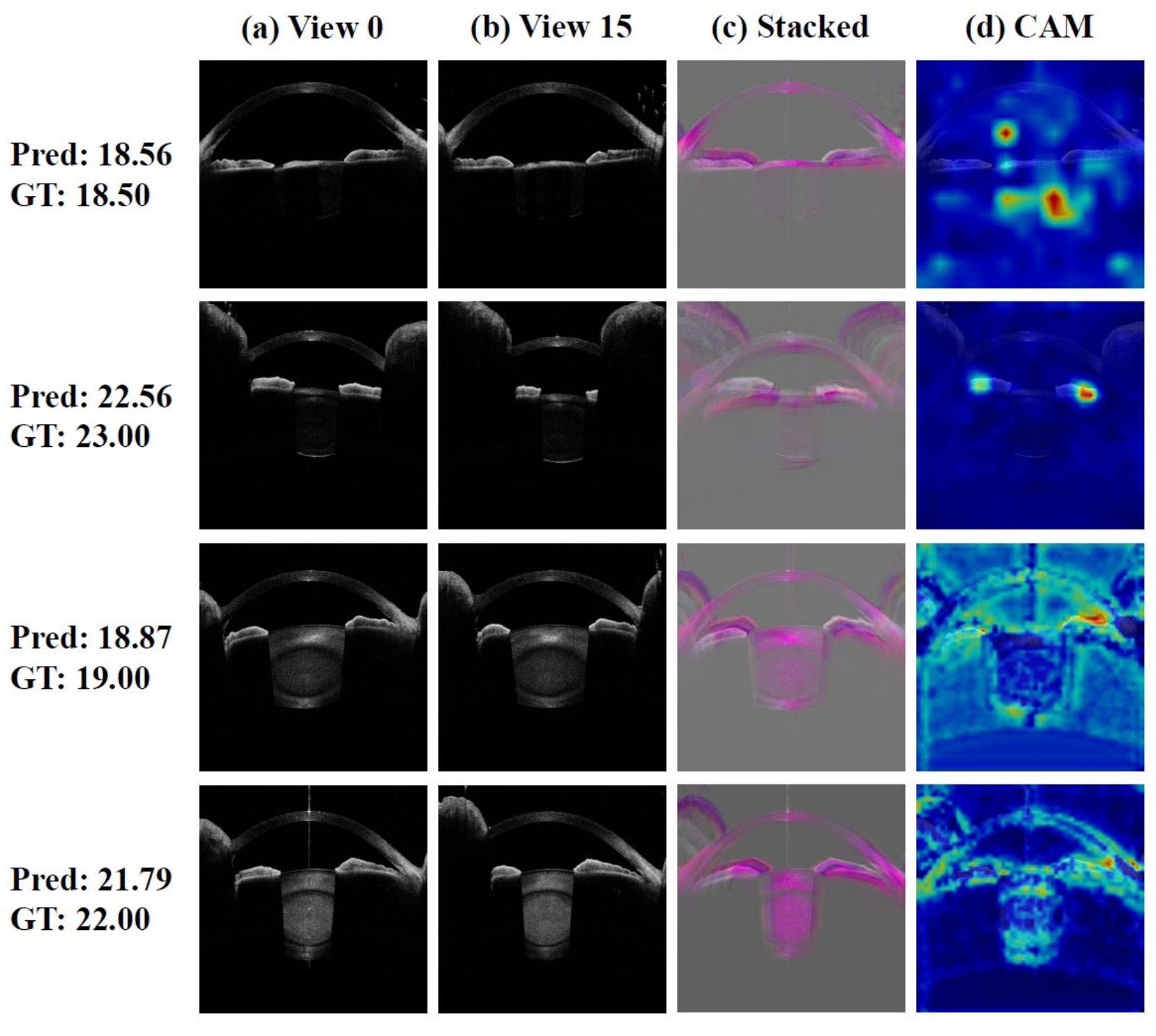

Figure 5. Visualizations of input images and CAMs. (a) and (b) are OCT images from views 0 and 15, respectively. (c) shows the stacked multi-view OCT output, and (d) displays CAMs for view 0.

4. CONCLUSION

Our framework leverages multi-modal data to enhance IOL power prediction accuracy. By combining OCT images and biometric data, it significantly outperforms traditional methods. This approach offers insights for future advancements in IOL power calculation and can be applied beyond this domain.

5. Acknowledgements

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).