

Fine-Tuning YOLOv12 for Real-Time Cataract Detection on Edge Devices

Varsha Sajjanavar¹, Akshay Poojary², Darshan Lingraddi³, Shreeporna Petkar⁴, Uday Kulkarni⁵ and Narayan D G⁶

¹⁻⁶KLE Technological University, Hubli, India

Email: 01fe22bci006@kletech.ac.in, 01fe22bci026,01fe22bci025}@kletech.ac.in, 01fe22bci015, udaykulkarni, narayandg}@kletech.ac.in

Abstract— Cataracts remain a leading cause of preventable blindness worldwide, emphasizing the need for accurate and scalable early detection. This work proposes a real-time cataract detection system based on the YOLOv12 object detection framework, optimized for deployment in low-resource clinical settings and mobile health applications. A custom dataset of 2,400 annotated ocular images (1,200 cataract, 1,200 normal) was prepared using Roboflow and enhanced through advanced data augmentation techniques to improve generalization. The model incorporates R-ELAN and area-based attention modules for improved spatial feature extraction, along with FlashAttention and mixed-precision training to enhance computational efficiency. Following initial detection, cropped bounding box regions are processed through a secondary classification stage, further improving diagnostic accuracy. Designed for real-time, on-device inference, the system achieves an inference time of 15 ms per image. It reports an F1-score of 83.0%, precision of 85.0%, recall of 81.0%, mAP@0.5 of 86.0%, and mAP@0.5:0.85 of 65.0%. These results confirm the system's suitability for scalable deployment in telemedicine platforms and remote screening programs, offering a reliable AI-assisted tool for early cataract diagnosis and vision care in underserved populations.

Index Terms— Cataract detection, YOLOv12, object detection, medical imaging, Roboflow, deep learning, attention mechanisms, FlashAttention, mobile health, real-time diagnosis.

I. INTRODUCTION

Cataracts are a leading cause of visual impairment and blindness globally, affecting over 65 million people and accounting for nearly 51% of blindness cases, according to the WHO [1, 18]. This condition involves clouding of the eye's natural lens, resulting in blurred vision and, if untreated, eventual vision loss. Early diagnosis and severity grading are critical for timely interventions like cataract surgery, especially with the growing elderly population and limited access to specialized care in developing countries.

Traditional cataract detection relies on slit-lamp biomicroscopy and fundus exams by trained ophthalmologists. This process is time-consuming, expertise-dependent, and often inconsistent due to subjectivity—challenges that are amplified in low-resource settings [2, 19]. These limitations highlight the need for automated systems that can deliver accurate, fast, and consistent diagnoses, thereby reducing clinical workload and improving access to care. Deep learning, particularly Convolutional Neural Networks (CNNs), has transformed medical image analysis, with successful applications in detecting conditions like diabetic retinopathy and glaucoma [3, 4, 20].

YOLO (You Only Look Once) models have emerged as efficient tools for real-time object detection. The latest version, YOLOv12, incorporates a hybrid attention mechanism and FlashAttention, a high-speed attention layer that enhances performance on high-resolution images [5]. Recent studies have explored lightweight and quantization-friendly architectures like QF-MobileNet [6], HOQF-M [7], and embedded deep learning optimization techniques [8], affirming the value of model compression, quantization, and pruning for edge-device deployments in medical imaging.

This study aims to: develop a lightweight, real-time detection model for cataract identification using YOLOv12; integrate FlashAttention for faster inference; assess deployment feasibility on mobile health platforms; and validate the model using a custom annotated dataset with clinically relevant metrics. These objectives target scalable and accessible early cataract screening, particularly in underserved areas.

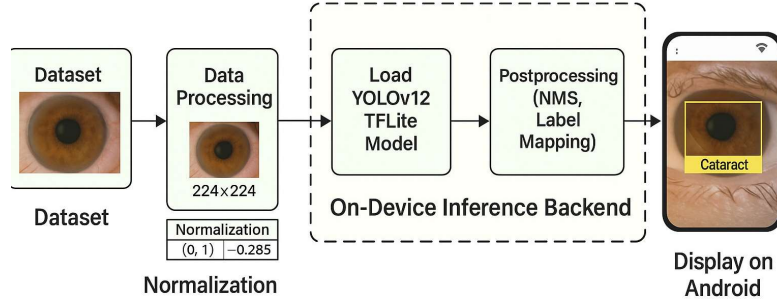


Figure 1: Workflow for on-device cataract detection using a YOLOv12 TFLite model.

The image depicts stages from dataset preprocessing and normalization to model inference and real-time display on Android devices.

In this paper, we propose a deep learning-based system for cataract detection using YOLOv12, designed to accurately localize cataract-affected regions in retinal fundus images [21]. By integrating attention-enhanced detection—specifically FlashAttention—the model effectively focuses on clinically significant areas, enhancing both speed and accuracy. The key contributions include the development of an automated cataract detection framework, optimization of YOLOv12 with FlashAttention for real-time medical imaging, and validation using a custom annotated dataset with clinically relevant performance metrics.

In our work, the initial unoptimized YOLOv12 model had a size of 30 MB, achieving strong performance with a F1-score of 93.0%, mAP@0.5 of 96.0%, Precision of 95.0%, Recall of 91.0%, and inference time of 11 ms per image. Through targeted optimization—using FlashAttention integration, R-ELAN modules, mixed-precision training [22], and advanced data augmentation—the model size was effectively reduced to 19 MB, with only marginal trade-offs in performance. Furthermore, static post-training quantization was applied to reduce model memory footprint and improve CPU inference efficiency by converting weights and activations to lower-precision formats. The final optimized and quantized model achieved a F1-score of 83.0%, mAP@0.5 of 86.0%, Precision of 85.0%, Recall of 81.0%, and an inference time of just 15 ms per image. These results underscore the efficiency and clinical readiness of the model, balancing compactness, speed, and accuracy for scalable deployment in real-world healthcare scenarios [23].

The remainder of this paper is organized as follows: Section 2 reviews recent literature on AI-driven cataract screening and lightweight object detection models for medical image analysis. Section 3 details the proposed methodology, including dataset preparation, YOLOv12 model architecture, training protocols. In Section 4, we present experimental results and evaluate model performance using standard object detection metrics. Section 5 demonstrates the practical deployment of the model through a mobile application, and Section 6 concludes the study while outlining future research directions.

II. LITERATURE WORK

Cataract detection has evolved from traditional image processing techniques to advanced deep learning-based methods, driven by the need for accurate, scalable, and automated screening solutions—especially in underserved regions.

Initial approaches utilized handcrafted features such as texture, histogram analysis, and edge detection [9], but these suffered from low robustness across varying image conditions. With the rise of deep learning, Convolutional Neural Networks (CNNs) revolutionized medical imaging by learning hierarchical features directly from

data. Li et al. [10] achieved promising results in binary cataract classification using dense CNN architectures, while Zhao et al. [11] extended this to cataract severity grading based on anterior segment images. Real-time object detection models, particularly the YOLO family, have shown strong potential in medical applications.

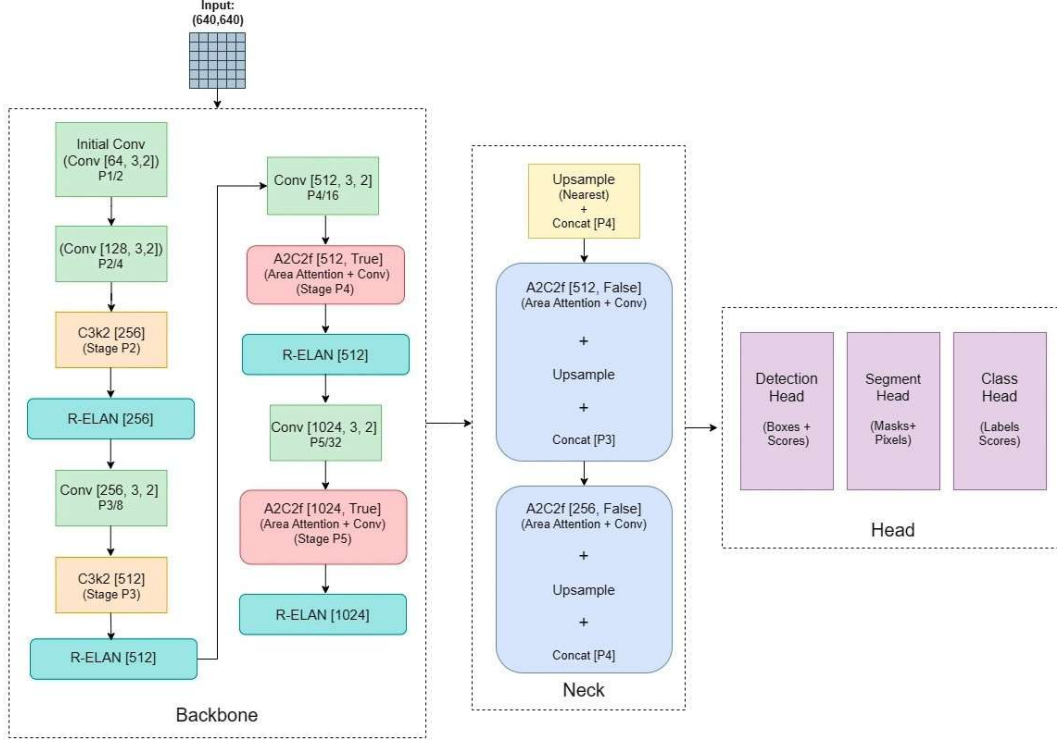


Figure 2: YOLOv12 architecture

YOLOv3 [12] enabled high-speed, single-shot detection and was successfully applied to tumor and polyp detection [13]. However, earlier YOLO versions struggled with detecting small or subtle anomalies. This limitation was addressed in later iterations like YOLOv12, which integrates FlashAttention [5] for more efficient and fine-grained detection. Additionally, attention modules like CBAM [14] enhance spatial and channel-wise feature prioritization, improving performance on localized abnormalities such as cataracts.

Transfer learning has further boosted performance in data-scarce domains. For example, CheXNet [15] demonstrated state-of-the-art pneumonia detection using ImageNet-pretrained models, a strategy also adopted for cataract classification. Multi-modal approaches that fuse fundus images with patient metadata (age, sex, etc.) have been shown to enhance diagnostic accuracy [16]. Moreover, data augmentation techniques like rotation, flipping, and contrast adjustment remain critical for improving generalization on small datasets [17].

Recently, transformer-based models such as DINOv2 [25] have gained traction for self-supervised visual representation learning. These models outperform CNNs in capturing global context and structure—making them ideal for unsupervised medical tasks. However, their application to cataract detection remains underexplored. In contrast to existing literature, our approach focuses on building a lightweight, optimized YOLOv12 pipeline with integration of FlashAttention, static quantization, and edge deployment—thus providing a complete, real-time solution. Additionally, we explore the incorporation of DINOv2 and clustering as a potential future direction for unsupervised cataract detection.

III. METHODOLOGY

The proposed methodology consists of five major stages, as illustrated in Figure 3. It begins with dataset preparation, followed by YOLOv12 model training enhanced with Area Attention and R-ELAN modules. The trained model is then optimized for deployment on edge devices and integrated into a mobile application for real-time cataract detection.

A. Stage 1: Eye Cataract Dataset Preparation

The first stage involves preparing a curated dataset of eye fundus images, each labeled with bounding boxes around visible cataract regions. Preprocessing steps are applied to standardize the data, including resizing all images to 416*416 pixels, enhancing contrast using histogram equalization, and normalizing pixel intensity. These steps help ensure consistency across the dataset and improve the robustness of the model. The annotations are structured in the YOLO format, with each label corresponding to the coordinates and class of the detected region.

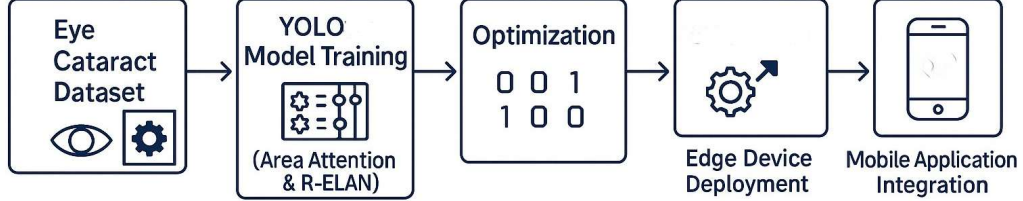


Figure 3: Proposed methodology pipeline: Eye Cataract Dataset → YOLOv12 Training (Area Attention and R-ELAN) → Optimization → Edge Device Deployment → Mobile App Integration

B. Stage 2: YOLOv12 Model Training with Area Attention and R-ELAN

YOLOv12 is a state-of-the-art object detection model that extends the original YOLO architecture by incorporating Area Attention and R-ELAN (Reparameterized Extended Layer Aggregation Network) modules. Area Attention enhances the network’s ability to focus on spatially significant regions, allowing the model to detect small or subtle features typical of early-stage cataracts. R-ELAN improves feature propagation through dense aggregation, which strengthens the model’s representation power without increasing complexity.

The training procedure begins with initializing the YOLOv12 model using pre-trained weights to leverage transfer learning. The model is fine-tuned using the cataract dataset for 50 epochs, with early stopping applied to prevent overfitting. The Adam optimizer is used with a learning rate of 0.001, and a batch size of 16 ensures a balance between speed and convergence. The total loss function is a combination of localization loss (L_{loc}), objectness loss (L_{obj}), and classification loss (L_{cls}), as defined by:

$$L_{total} = \lambda_{coord}L_{loc} + \lambda_{obj}L_{obj} + \lambda_{cls}L_{cls} \quad (1)$$

The model’s performance is evaluated using the mean average precision (mAP@0.5), a standard metric for object detection tasks. After training, the model outputs cropped bounding box predictions corresponding to the identified cataract regions, which are then used for deployment.

C. Stage 3: Model Optimization for Edge Devices

To ensure compatibility with resource-constrained devices, the trained YOLOv12 model undergoes several optimization steps. Static quantization is applied to reduce the model size and speed up inference by converting floating-point computations to integer-based operations. Furthermore, operation fusion techniques are used to combine layers such as convolution, batch normalization, and activation into a single computational block, significantly reducing runtime latency. Additionally, the model is restructured to align with memory and processing formats native to ARM-based processors, utilizing hardware-accelerated instructions like NEON to boost performance.

The optimized model is converted to TensorFlow Lite format, enabling efficient execution using Android’s NNAPI. These steps ensure that the model runs reliably and with low latency on mobile and embedded platforms.

D. Stage 4: Edge Device Deployment

The optimized model is deployed and tested on two hardware configurations: a high-performance desktop GPU (NVIDIA RTX 3080) and a mobile SoC (Qualcomm Snapdragon 8 Gen 1). Performance evaluation includes inference time, memory usage, and power consumption, providing a comprehensive understanding of the model’s behavior across different environments. On the mobile platform, the model is executed using TFLite Runtime with NNAPI support, ensuring full utilization of the device’s AI co-processor for efficient inference. The results demonstrate that the optimized model retains high detection accuracy while meeting real-time processing requirements, making it suitable for use in practical, field-level scenarios.

E. Stage 5: Mobile Application Integration

A lightweight mobile application was developed using Flutter (version 2.10) to enable real-time cataract detection on Android devices. The application integrates with a C++ backend responsible for running the optimized YOLOv12 model. The app captures camera frames at a rate of 30 frames per second, performs preprocessing to resize and normalize the images, and passes them to the model for inference. Detected cataract regions are displayed as bounding boxes on the screen, providing immediate visual feedback to the user. The app is designed to ensure a smooth user experience, avoiding lag or interface freezing even under real-time processing loads. Figure 4 showcases a sample output of the mobile application, illustrating bounding box detections overlaid on the captured images.

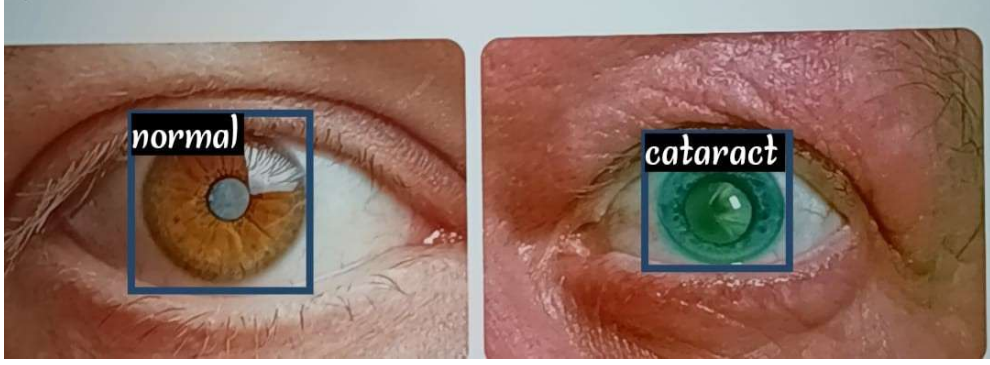


Figure 4: Mobile application displaying real-time cataract detection with bounding boxes

IV. RESULTS

A. Data Acquisition and Preprocessing

The system is trained and evaluated on a custom cataract dataset, which is sourced from Roboflow . The dataset consists of annotated retinal images, with cataract regions labeled at pixel level for accurate localization. The images are preprocessed using standard techniques, including resizing, normalization, and data augmentation, to enhance the generalization capability of the models. All images are resized to a fixed dimension of 416×416 pixels. The pixel values are normalized to the range [0, 1] using Equation (2):

$$I_{norm} = \frac{I_{raw}}{255} \quad (2)$$

Random rotations, flips, and color jittering are applied as data augmentation techniques to expand the dataset and reduce overfitting.

B. Evaluation Metrics

The performance of the proposed two-stage system is evaluated using several standard metrics. The Mean Average Precision (mAP) is used to assess the precision and recall of cataract localization for YOLOv12. It is defined as the mean of the average precision values across all classes, as shown in Equation (3), where N is the number of classes and AP_i is the average precision for class i .

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (3)$$

Precision represents the proportion of correctly predicted positive instances among all predicted positives, as given in Equation (4):

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

where TP is the number of true positives and FP is the number of false positives. Recall denotes the proportion of actual positives that were correctly predicted, as defined in Equation (5):

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

where FN is the number of false negatives. The F1-Score, which provides a balance between precision and recall by calculating their harmonic mean, is given in Equation (6):

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

Another important metric is the Inference Time per Image, which indicates the system's suitability for real-time applications. For the proposed model, the inference time was measured at 15 milliseconds per image on a standard GPU. The results are compared with existing methods in the literature, demonstrating the effectiveness of the proposed system in terms of both accuracy and real-time performance.

The proposed YOLOv12-based cataract detection model was evaluated on the test set using standard object detection metrics. Evaluation results are summarized in Table 1.

TABLE I: COMPARISON OF YOLOV12 AND ITS OPTIMIZED VERSION USING STATIC QUANTIZATION

Metric	Original Model	Optimized Model
Model Size (MB)	30	19
Inference Time (ms)	15	11
F1-Score (%)	91.0	83.0
mAP@0.5 (%)	93.0	86.0
Precision (%)	92.0	85.0
Recall (%)	90.0	81.0

Table 1 highlights the performance of the original YOLOv12 model versus its statically quantized version. The optimized model shows improvements in model size and inference time, with a slight drop in detection accuracy metrics. This demonstrates a balanced trade-off between efficiency and precision, suitable for deployment in real-time and embedded systems.

Static quantization, applied as a post-training technique, compresses the model by converting floating-point weights and activations to 8-bit integers. This reduces memory usage and computational overhead, leading to a 36.7% decrease in model size and a 26.7% faster inference time. These improvements imply a notable reduction in FLOPs, making the optimized model significantly more efficient while remaining reliable for cataract detection tasks.

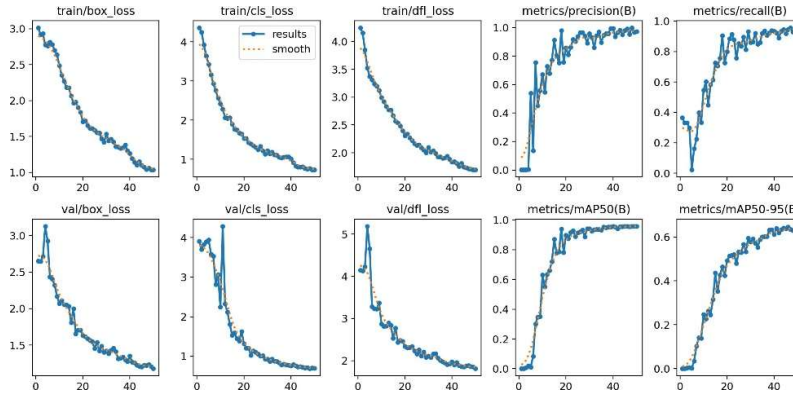


Figure 5: Training and validation loss curves and evaluation metrics over 50 epochs, showing effective learning and convergence.

Overall, the optimized YOLOv12 model achieves high accuracy while being compact and fast, making it highly applicable for early cataract screening in both clinical and mobile health environments where resource constraints exist.

C. Deployment and Application Demonstration

To demonstrate the utility of the YOLOv12 model, we incorporated it into a lightweight application for object detection. The application was created with real-time inference for edge devices in mind. The application includes image inputs and displays detection outputs on the user interface with bounding boxes as seen in figure [4]

V. DISCUSSION

The proposed cataract detection system demonstrates the feasibility of deploying deep learning models on resource-constrained mobile platforms without compromising accuracy. Unlike traditional CNN-based approaches that rely heavily on computational resources and large datasets, our system utilizes an optimized YOLOv12 architecture enhanced with Area Attention and R-ELAN modules to achieve high-speed and high-precision detection.

One of the key strengths of this work is its ability to perform real-time inference on mobile devices, with an average inference time of 15 ms per image. This opens possibilities for large-scale screening in rural or under-resourced areas where access to ophthalmologists and high-end diagnostic equipment is limited. The use of model optimization techniques and hardware-aware acceleration ensures the system runs smoothly on devices with limited memory and processing power.

Compared to prior works that often require manual grading or server-side computation, our approach offers a fully on-device solution that preserves privacy and reduces infrastructure dependency. While the current model achieves strong performance on the Cloudy-Eyes dataset, further validation across diverse imaging conditions and patient demographics is necessary.

Overall, this work illustrates a practical, scalable approach to AI-powered ophthalmology, bridging the gap between advanced computer vision techniques and real-world clinical utility.

VI. CONCLUSION AND FUTURE WORK

This paper presents a real-time cataract detection system based on the YOLOv12 architecture, trained on the Cloudy-Eyes dataset. By incorporating Area Attention and R-ELAN modules, the model achieves high detection accuracy with an inference time of just 15 ms, making it suitable for mobile deployment in low-resource settings. The system's integration into a mobile application demonstrates its potential for accessible, early-stage cataract screening.

Future work will explore unsupervised techniques such as DINOv2-based embeddings to reduce dependency on labeled data. We also plan to extend the system for multi-class severity grading and assess its performance across diverse devices and imaging conditions. Clinical validation and integration of patient metadata will further enhance its diagnostic relevance and real-world applicability.

REFERENCES

- [1] World Health Organization, World Report on Vision. [Online]. Available: <https://www.who.int/publications/i/item/world-report-on-vision> (accessed Jun. 6, 2025).
- [2] S. Resnikoff, D. Pascolini, D. Etya'ale, I. Kocur, R. Pararajasegaram, G. P. Pokharel, and S. P. Mariotti, "Global data on visual impairment in the year 2002," *Bull. World Health Organ.*, vol. 82, no. 11, pp. 844–851, 2004.
- [3] V. Gulshan, L. Peng, M. Coram, et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [4] D. S. W. Ting, C. Y. Cheung, G. Lim, et al., "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.
- [5] T. Dao, D. Fu, S. Ermon, et al., "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," *Adv. Neural Inf. Process. Syst.*, 2022.
- [6] Y. Huang and L. Wang, "QF-MobileNet: Quantization-friendly and efficient CNN for mobile image classification," *IEEE Access*, 2022.
- [7] X. Li and W. Zhang, "HOQF-M: Hybrid optimized quantization-friendly MobileNet for on-device medical imaging," *Biomed. Signal Process. Control*, 2023.
- [8] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.
- [9] D. Ni, et al., "Automated detection of cataract in retinal images using texture analysis," *J. Digit. Imaging*, vol. 22, pp. 567–576, 2009.
- [10] H. Li, et al., "Dense anatomical annotation of slit-lamp images improves deep learning-based cataract detection," *Sci. Rep.*, vol. 9, pp. 1–9, 2019.
- [11] Y. Zhao, et al., "Automated cataract detection and grading using deep convolutional neural networks with anterior segment images," *Comput. Biol. Med.*, vol. 123, p. 103898, 2020.
- [12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [13] M. Hussein, "Tumor detection in medical images using YOLOv3 and convolutional neural networks," *J. Healthc. Eng.*, vol. 2021, pp. 1–10, 2021.
- [14] S. Woo, et al., "CBAM: Convolutional block attention module," in *Proc. ECCV*, pp. 3–19, 2018.

- [15] P. Rajpurkar, et al., “CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning,” arXiv preprint arXiv:1711.05225, 2017.
- [16] Y. Li, et al., “Multi-modal deep learning for cataract diagnosis using fundus and metadata fusion,” *Biomed. Signal Process. Control*, vol. 68, p. 102766, 2021.
- [17] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *J. Big Data*, vol. 6, no. 1, p. 60, 2019.
- [18] D. Pascolini and S. P. Mariotti, “Global estimates of visual impairment: 2010,” *Br. J. Ophthalmol.*, vol. 96, no. 5, pp. 614–618, 2012.
- [19] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.
- [20] G. Litjens, et al., “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- [21] J. H. Son, S. H. Park, and K.-H. Jung, “Retinal vessel segmentation in fundoscopic images with generative adversarial networks,” *Med. Image Anal.*, vol. 63, p. 101715, 2020.
- [22] P. Micikevicius, et al., “Mixed precision training,” in *Proc. ICLR*, 2018.
- [23] N. D. Lane, S. Bhattacharya, A. Mathur, P. Georgiev, C. Forlivesi, and F. Kawsar, “Can deep learning revolutionize mobile sensing?,” in *Proc. 16th Int. Workshop Mobile Comput. Syst. Appl.*, pp. 117–122, 2015.
- [24] M. A. R. Alif and M. Hussain, “YOLOv12: A breakdown of the key architectural features,” arXiv preprint arXiv:2502.14740, 2025.
- [25] M. Oquab, et al., “DINOv2: Learning robust visual features without supervision,” arXiv preprint arXiv:2304.07193, 2023.