University of Rajshahi

Department of Information and Communication Engineering

EfficientNet vs. CBAM: Benchmarking Attention for Ocular Disease Classification

Supervisor: Dr. Md. Matiqul Islam

Authors:

Md. Takrim–Ul–Alam (Roll: 1911177149)

Samiul Bashir (Roll: 2010277105)

October 19, 2025

Abstract

Abstract

This project investigates attention mechanisms for ocular disease classification using fundus images from the ODIR–5K dataset. We compare a strong convolutional baseline (EfficientNet) against an attention–augmented variant employing the Convolutional Block Attention Module (CBAM). Our pipeline parses ODIR metadata, prioritizes Hypertension labeling where present, and enforces robust stratified splits to ensure all target classes appear in validation and test sets. Experiments demonstrate that image–specific attention improves several classes, while Hypertension remains challenging due to limited single–label prevalence and ambiguity in diagnosis text. We provide full training/evaluation artifacts (curves, confusion matrices, and metrics) to support reproducibility and future extensions, including multi–label learning and targeted augmentation for rare classes.

Keywords: EfficientNet, CBAM, ocular disease classification, ODIR–5K, Grad–CAM, ROC–AUC, PR–AUC

Contents

List of Figures

List of Tables

Introduction

Retinal fundus photography provides a non–invasive window into ocular health, enabling screening and diagnosis for conditions such as Glaucoma (G), Cataract (C), Age–related Macular Degeneration (AMD, A), Hypertension–related retinopathy (H), and Myopia (M). Automated classification can assist clinicians by prioritizing high–risk cases and scaling screening programs. Deep convolutional networks (CNNs) learn strong visual features but can struggle with class imbalance, domain variability, and subtle disease cues. Attention mechanisms explicitly reweight feature channels and spatial regions, potentially improving discrimination on small or ambiguous lesions. In this project we evaluate an EfficientNet baseline and an EfficientNet+CBAM variant on ODIR–5K, following a robust data parsing and splitting procedure, and report comprehensive metrics and plots to support a fair comparison.

1.1 Motivation

ODIR–5K reflects real–world clinical variability: heterogeneous image quality, multi–pathology co–occurrence, and minority classes (notably Hypertension) that are easily under–represented in standard splits [5]. A practical screening system must remain reliable under these constraints. Lightweight attention offers a promising path to improve separability of subtle biomarkers (e.g., AV nicking, drusen) without the data demands of full transformer models.

1.2 Objectives

We aim to: (i) establish a strong CNN baseline (EfficientNet); (ii) integrate CBAM attention to assess gains in accuracy and class–wise recall; (iii) enforce reproducible data handling (deterministic parsing, stratified splits with all classes present); and (iv) report comprehensive metrics (macro/weighted F1, ROC–AUC, PR–AUC) and interpretability artifacts (Grad–CAM) to support defensible conclusions.

1.3 Contributions

- A practical ODIR–5K pipeline with robust parsing and Hypertension–priority labeling to mitigate label sparsity in validation/test.
- An attention–enhanced classifier (EfficientNet+CBAM) compared against a matched EfficientNet baseline under identical preprocessing, augmentation, and training schedules.
- Thorough evaluation artifacts (training curves, confusion matrices, ROC/PR curves, macro/weighted F1, ROC–AUC and PR–AUC) prepared for report integration.

1.4 Challenges

Key challenges include: (i) severe class imbalance and rare single–label Hypertension; (ii) quality degradations (illumination, blur, media opacities) that obscure biomarkers; (iii) label ambiguity from free–text keywords; and (iv) limited data relative to transformer pretraining needs. Our design choices (Hypertension priority, stratified splits, class weights, mixed precision, and CBAM placement late in the network) directly target these issues.

1.5 Clinical Foundations and Imaging Artifacts

Fundus photography captures the posterior segment (retina, optic disc, macula, vessels) with high diagnostic value for ocular and systemic disease [6]. Longitudinal acquisition supports disease monitoring and treatment response. Real–world collections exhibit quality defects that challenge automated analysis: uneven illumination; lens dust or eyelashes producing artifacts; media opacities (e.g., cataracts) lowering contrast and sharpness; and focus errors due to motion or operator variability [7, 8]. These imperfections motivate robust preprocessing, augmentation, and architectures that can emphasize informative signals while down–weighting nuisance factors.

1.6 Pathology Biomarkers in Fundus Images

Target classes manifest distinct visual cues: optic disc cupping and rim thinning in Glaucoma; global haze/blur in images affected by Cataracts; drusen and pigment changes in AMD; arteriolar narrowing, AV nicking, hemorrhages, cotton–wool spots in Hypertensive Retinopathy; and posterior staphyloma or atrophic patches in Pathologic Myopia [9–15]. Their diversity spans structures, textures, vascular geometry, and global image degradations, arguing for attention mechanisms that adaptively focus on relevant channels and spatial regions per case.

Related Work

We situate our approach within prior art on fundus disease classification, attention mechanisms for CNNs, and ODIR–5K research. We first outline CNN baselines and transfer learning practice for fundus imaging, then summarize channel/spatial attention (SE, CBAM, ECA) and token self–attention (ViT/DeiT), and finally review ODIR–5K labeling practices and recent state of the art.

2.1 Fundus Image Classification

CNNs such as VGG, ResNet, and EfficientNet have been widely applied to fundus image analysis for diabetic retinopathy screening and broader ocular disease classification. EfficientNet family models leverage compound scaling and strong ImageNet pretraining for competitive performance at modest compute cost [16].

Recent practitioner work on Kaggle explores end–to–end pipelines for eye disease classification, e.g., [17], illustrating common preprocessing, transfer learning backbones, and evaluation practices that complement academic baselines.

2.2 Attention Mechanisms

Channel and spatial attention mechanisms (SE, CBAM, ECA) improve CNN feature quality by adaptively reweighting salient signals. CBAM applies sequential channel and spatial attention via lightweight modules with minimal overhead [4]. For accessible primers on CBAM and related attention modules, see [2, 3]. Vision transformers (ViT) and token–based self–attention have also shown promise, but often require larger datasets or heavy augmentation.

2.2.1 From SE to CBAM, BAM, and ECA

Squeeze–and–Excitation (SE) [18] introduced channel reweighting via global pooling and a bottleneck MLP, greatly improving many CNN backbones. BAM [19] adds a parallel attention branch with dilated convolutions to enlarge receptive fields, while ECA [20] removes the MLP in favor of local cross–channel interactions using 1D convolutions. CBAM [4] extends SE by combining channel and spatial attention sequentially, capturing both what and where to attend.

2.2.2 Attention in Medical Imaging

In medical CV tasks, data scarcity and domain shift often favor CNNs with lightweight attention over pure transformers. Channel/spatial attention has benefited retinal disease grading, lesion localization, and OCT segmentation by enhancing salient structures (discs, vessels, macula). Our EfficientNet+CBAM results on ODIR–5K align with this trend: small overhead, improved per–class recall, and clearer Grad–CAM focus.

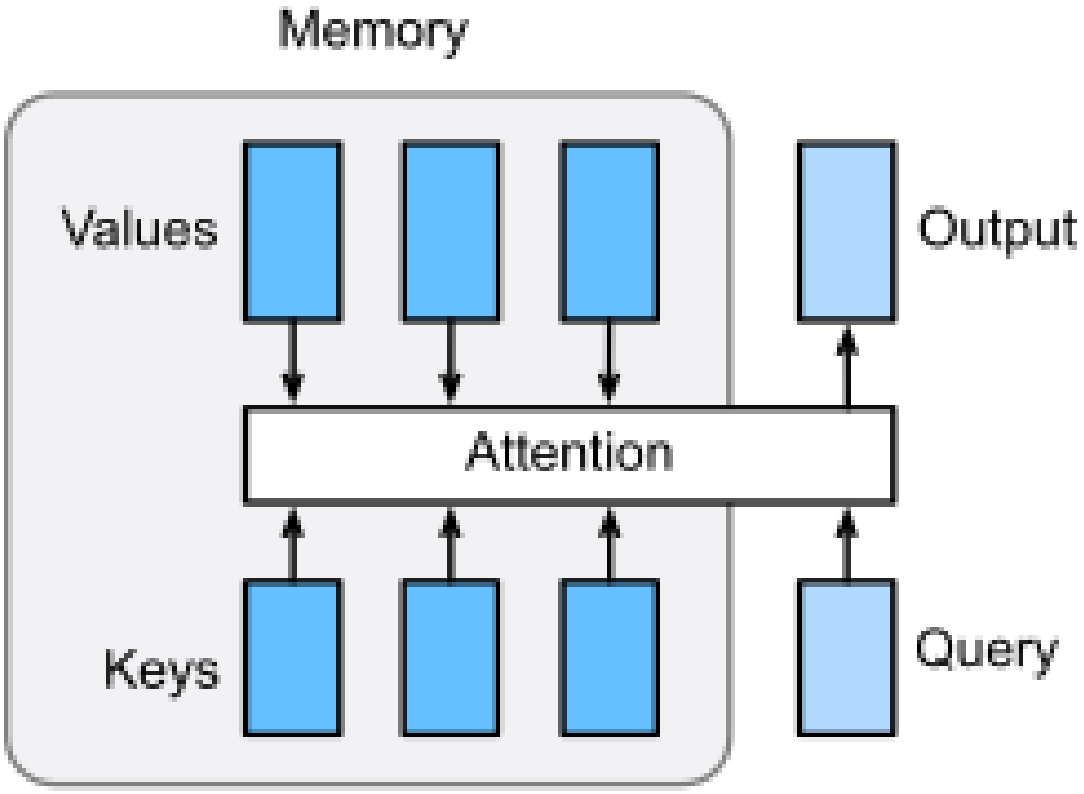

Figure 2.1: Overview of attention mechanisms (adapted from Wikipedia [1]): queries, keys, and values produce attention weights to focus computation on salient content. We use convolutional attention (CBAM) rather than token attention (ViT) to keep the inductive biases of CNNs for fundus images.

2.3 ODIR–5K and Labeling

The ODIR dataset provides paired left/right fundus images and metadata. Practical pipelines must reconcile free–text diagnoses to structured labels and contend with multi–label prevalence and class imbalance. Prior work also explored generative augmentation for minority classes.

2.4 State of the Art on ODIR–5K

Recent works report improvements via binary re–framing, multi–label architectures with attention, and model fusion with Dempster–Shafer evidence theory. Summaries include DKCNet and related methods emphasizing imbalance handling and attention/fusion [21–23].

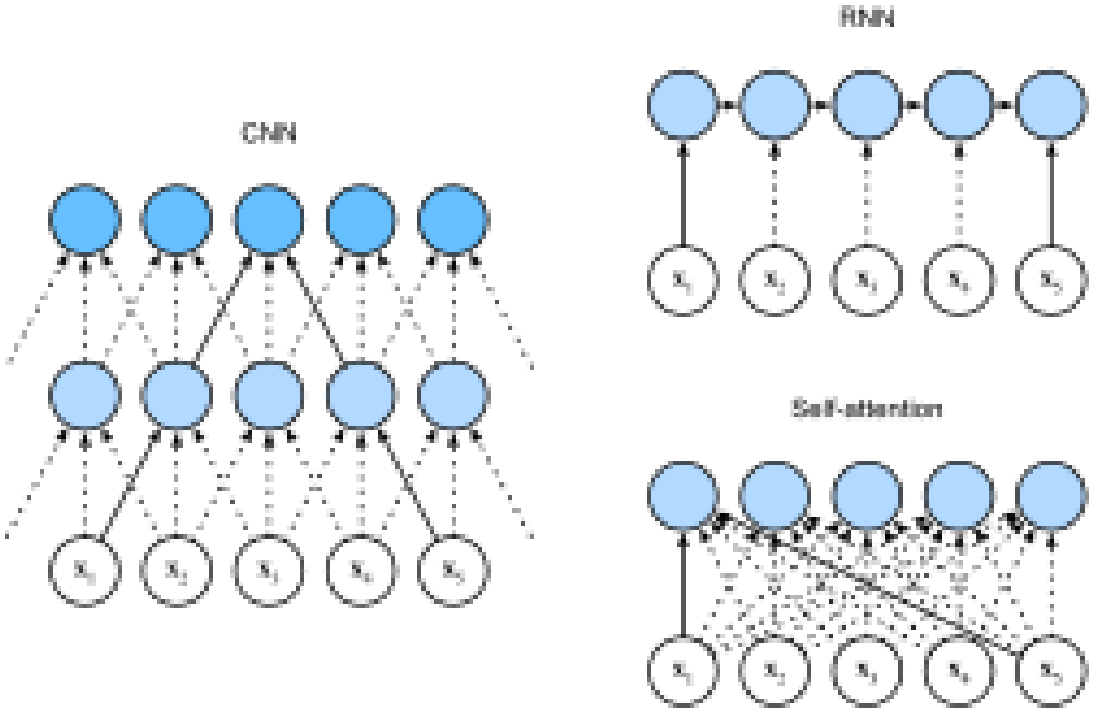

Figure 2.2: Self–attention styles across architectures (Wikipedia [1]). Our approach augments a CNN with CBAM, which performs channel and spatial attention on feature maps instead of sequence tokens.

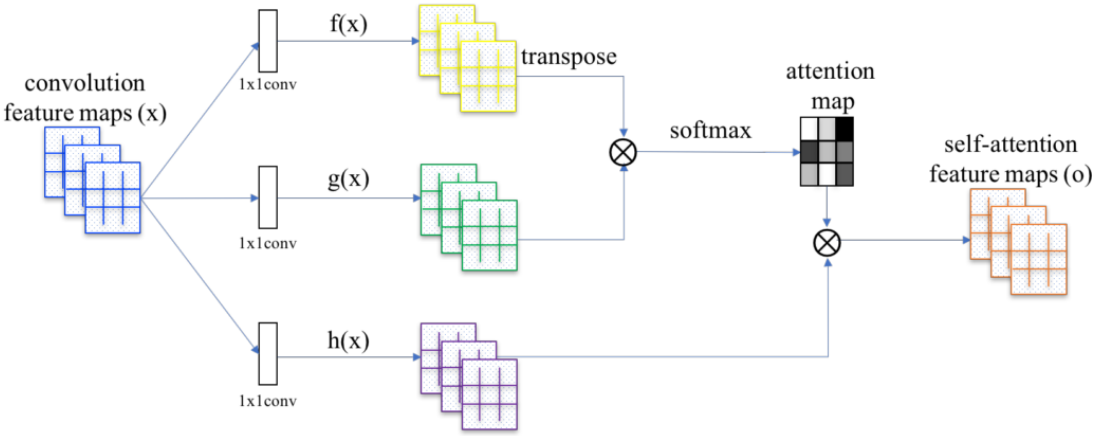

Figure 2.3: CBAM concept illustration (DigitalOcean tutorial [2]): channel attention (what) followed by spatial attention (where). We adopt this sequential design in our EfficientNet+CBAM model.

Literature Review

This review synthesizes clinical, architectural, and methodological evidence to ground our approach. We scoped sources to peer–reviewed work and commonly cited practitioner guides, emphasizing fundus imaging practice and ocular disease foundations; CNN baselines and lightweight attention (SE, ECA, BAM, CBAM) alongside token–based Transformers; and dataset/evaluation practices specific to ODIR–5K, including augmentation and explainability (Grad–CAM, LIME/SHAP). Each strand maps directly to design: Hypertension–aware labeling and patient–level splits; an EfficientNet baseline with late CBAM placement for saliency control at low overhead; percent confusion matrices and macro metrics for imbalance; and Grad–CAM validation that attention concentrates on clinically relevant regions. This integration underpins our hypothesis that CBAM improves separability and minority recall with modest parameter cost.

3.1 Clinical and Modality Context

A robust model must be grounded in the clinical realities and technical limitations of the data generation process. Fundus photography, while a standard diagnostic tool, is subject to numerous common pitfalls that can degrade dataset quality and, consequently, model reliability [6–8]. These include artifacts such as camera–lens dust, eyelash obstruction, poor patient fixation leading to blur, and improper illumination causing reflections or shadowed regions. Literature reviewing these practices [7, 8] informs the necessity of robust preprocessing and augmentation to ensure models are invariant to these non–pathological variations. If not properly handled, a model might erroneously learn that an eyelash shadow is a sign of pathology, creating a "shortcut" feature that fails dramatically upon real–world deployment.

Beyond modality artifacts, a firm clinical grounding is essential for defining class boundaries, especially given the high inter–rater variability that can exist even among clinical experts. Our class definitions for Cataract and Age–related Macular Degeneration (AMD) are informed by foundational literature [9–12, 24] that describes their expected image phenotypes. For cataracts, this primarily involves lens opacification obscuring the view of the retina, a feature the model must learn to identify. For AMD, this involves detecting key biomarkers in the macular region, such as drusen, pigmentary changes, or geographic atrophy [11, 24].

Similarly, the diagnostic criteria for Hypertensive Retinopathy (H) are framed by clinical guidance [13, 25, 26] focusing on vascular abnormalities. Our models must learn to capture subtle signs like arteriovenous (AV) nicking, copper or silver wiring of arterioles, flame–shaped hemorrhages, and cotton–wool spots, which are indicative of vascular damage from chronic hypertension. These signs can be subtle and often co–exist with diabetic retinopathy, making a discriminative feature representation critical.

Finally, for the Myopia (M) class, we draw on references [14, 15] that contextualize the structural changes associated with high myopia and posterior staphyloma. These changes include significant globe elongation, which manifests in the fundus image as optic disc tilting, peripapillary atrophy, a tessellated fundus appearance, and lacquer cracks. This body of literature [9, 11, 13–15, 24–26] is critical for ensuring our model's feature extraction aligns with established clinical diagnostic criteria. This alignment is necessary to build a tool that is not just accurate, but also interpretable and trustworthy to a clinical end–user.

3.2 Architectures: CNNs, Transformers, and Attention

The architectural design of a deep learning model is a primary determinant of its performance, balancing representational power with computational cost. CNNs have long been the *de facto* standard in medical imaging [27–29], owing to strong inductive biases (spatial locality and translation invariance) that suit biomarker detection. We adopt EfficientNet [16] as our CNN baseline for its state–of–the–art accuracy–efficiency trade–off.

In contrast, Vision Transformers (ViT) [30–34] adapt self–attention to images by treating an image as a sequence of patches. ViTs offer a global receptive field and can model long–range dependencies that may benefit diffuse ocular patterns. However, they often require massive pretraining and are computationally heavier than CNNs in low–data settings typical of medicine.

Bridging these paradigms are attention modules that augment CNNs by helping them focus on salient signals. SE [18], BAM [19], and CBAM [4] are representative. We favor CBAM for its lightweight sequential design: first channel attention (*what* to focus on), then spatial attention (*where* to focus). This provides a compute–efficient mechanism to refine EfficientNet features without sacrificing throughput, aiming for the best of both worlds: robust local feature extraction enhanced by attention–driven selection.

3.2.1 EfficientNet: Working Principle

The core innovation of EfficientNet [16] is compound scaling. Rather than scaling only one dimension (depth, width, or resolution), EfficientNet balances all three via a single coefficient ϕ: $d = \alpha^{\phi}$, $w = \beta^{\phi}$, $r = \gamma^{\phi}$ under a compute budget. The MBConv inverted residual block forms the backbone: expand with a 1×1 conv, apply a depthwise separable conv for spatial mixing, then project with a 1×1 conv. Integrated SE attention [18] and Swish activations further improve efficiency and representational quality. This combination, paired with ImageNet pretraining, yields a strong, transfer–effective baseline that we subsequently augment with CBAM.

3.2.2 EfficientNet Architecture

A practical overview of EfficientNet's B0 design outlines a stem–body–head pipeline and details MBConv blocks with squeeze–and–excitation (SE), depthwise separable convolution, and swish activations [16]. The B0 architecture starts with a 3×3 stride–2 stem conv, proceeds through stages of MBConv with varying expansion ratios, kernel sizes (3×3, 5×5), and strides, and

ends with a 1×1 conv, global average pooling, and a fully connected classifier. This description matches our baseline configuration and motivates CBAM insertion on top of EfficientNet feature blocks.

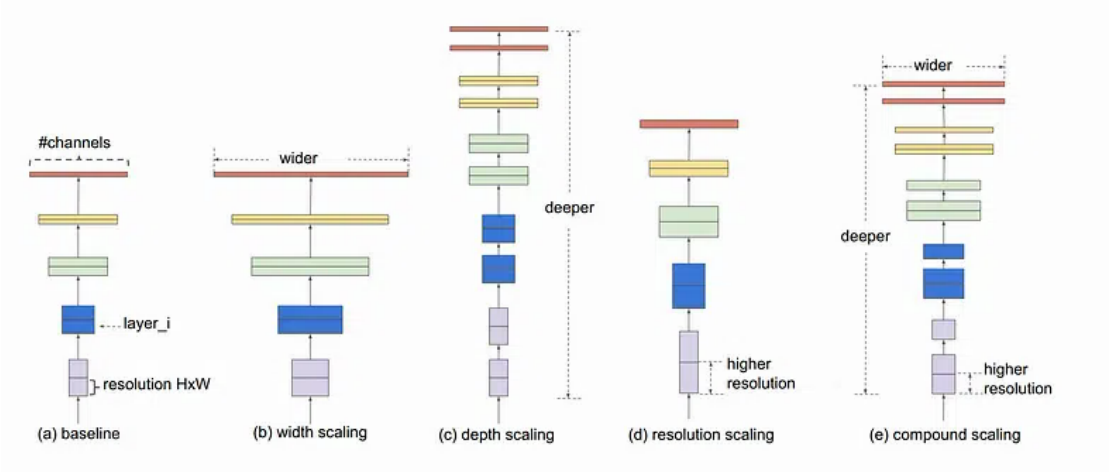

Figure 3.1: Scaling strategies summarized: individual width/depth/resolution scaling versus compound scaling that jointly balances all three.

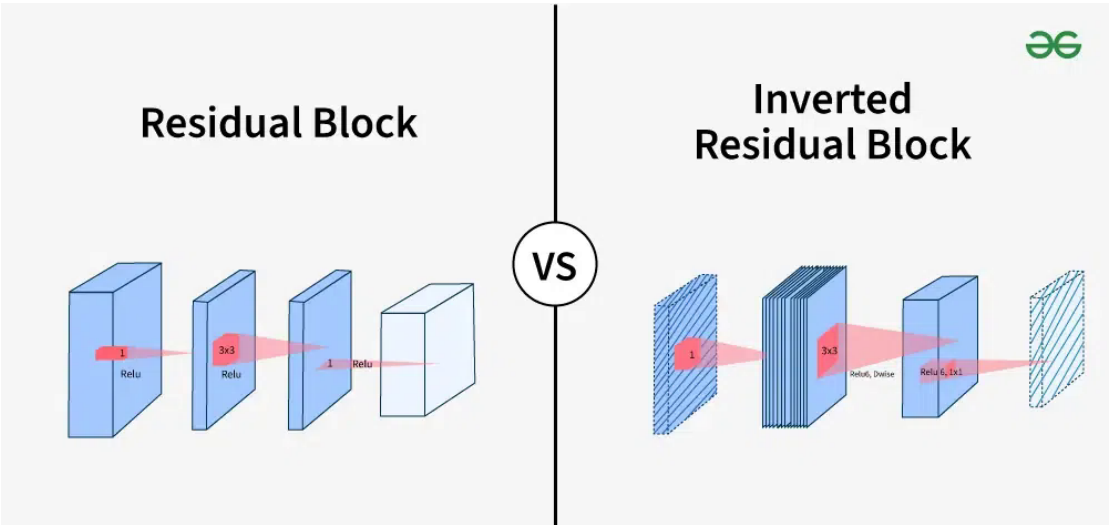

Figure 3.2: Residual vs inverted residual blocks illustrating EfficientNet's MBConv design with depthwise separable conv and SE attention.

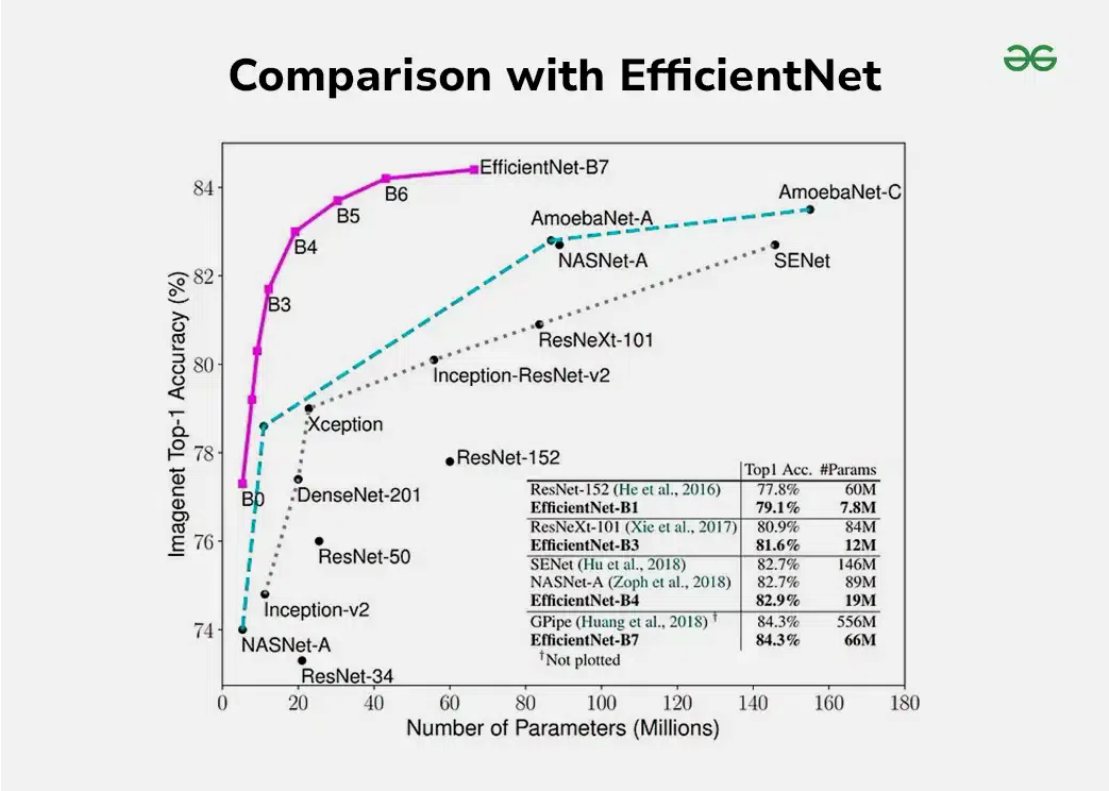

Figure 3.3: Comparative positioning of EfficientNet variants by accuracy–efficiency.

3.3 Dataset, Label Structures, and Evaluation

We operate on ODIR–5K [5], a challenging public benchmark for ocular disease classification. It is distinguished by paired left/right fundus images for 5,000 patients and a complex eight–label structure. Its multi–label nature is not an artifact but a core feature reflecting clinical comorbidity, where a single patient may present with multiple concurrent pathologies. This clinical realism makes strictly mutual–exclusive losses (softmax + categorical cross–entropy) ill–suited. Even when a single–label proxy is used (e.g., a primary diagnosis), guidance from multi–label and long–tailed learning [35–38] remains essential: Binary Cross–Entropy (BCE) treats each label as a separate binary classifier, allowing multiple positives; class re–weighting and strategic sampling mitigate minority rarity.

Medical datasets are inherently long–tailed; common conditions are over–represented while rare but critical pathologies are scarce. This imbalance biases naive models toward majority classes. Literature on imbalance [35, 36] informs design choices such as inverse–frequency loss weighting and careful sampling.

For evaluation, we adhere to rigorous practices in medical computer vision [21, 39]. The most critical protocol is patient–level splitting: since the two eyes of a patient are highly correlated, splitting them across train/test induces severe leakage and inflates metrics. We confine all images from a patient to a single split (train/val/test). Beyond accuracy, we report per–class precision, recall (sensitivity), F1, macro F1, and emphasize AUROC for its threshold–independence and robustness under imbalance, as it measures the ability to rank positives above negatives

irrespective of a fixed threshold.

3.4 Data Augmentation and Synthetic Data

To train a robust model, we teach invariances to non–pathological variations common in clinical imaging. We employ geometric and photometric transforms, including random horizontal/vertical flips, small rotations (e.g., $\pm 15^{\circ}$), random zooming, and brightness/contrast adjustments, simulating patient positioning and illumination variability so the model focuses on structural pathology.

Simple affine transforms are insufficient for non–rigid biological tissue. We therefore incorporate elastic deformations [40–44], applying localized, non–linear warps that better reflect subtle retinal shape changes in vivo or projection effects, forcing invariance to localized stretching/compression.

To address data scarcity in long–tailed distributions, we review generative augmentation via GANs [45–49]. The goal is targeted synthesis of rare variants to balance training; risks include mode collapse and non–plausible artifacts, requiring clinical validation. Mixup/CutMix–style strategies [35] provide complementary regularization by blending images/labels, smoothing decision boundaries and improving calibration.

3.5 Explainability and Model Interpretation

Deep models are often criticized as "black boxes," a barrier in clinical workflows where trust and accountability are paramount. We rely on Grad–CAM as the primary tool: gradients into the final convolutional layer produce a coarse saliency map highlighting regions most influential for a prediction. This layer balances semantic abstraction with spatial fidelity.

Grad–CAM provides clinical validation. For a correct AMD prediction, attention should concentrate on the macula (drusen, pigmentary changes); for glaucoma, activations around the optic disc indicate cupping cues. Conversely, if heatmaps highlight non–pathological artifacts (e.g., eyelash shadow, lens reflection [7, 8]), we have identified a spurious correlation and a failure mode.

For model–agnostic auditing, LIME and SHAP [22, 23, 50, 51] complement Grad–CAM. LIME fits a local interpretable model around an instance via perturbations (e.g., superpixel toggling). SHAP provides game–theoretic attributions (Shapley values) assigning each feature its contribution. Together they support debugging, clinician trust, and transparent communication of model decisions.

3.6 Frontiers: Self–Supervised and Federated Learning

Two frontiers address fundamental medical AI bottlenecks: labeling and data access. Self–supervised learning (SSL) [52] reduces reliance on expert–labeled datasets by learning domain representations from pretext tasks such as contrastive learning (e.g., distinguishing augmented views of the same fundus) or masked autoencoding. SSL yields pathology–aware backbones that transfer

better than generic ImageNet features.

Federated Learning (FL) [53, 54] addresses privacy, governance, and security constraints by training where the data reside: institutions train locally and share only anonymized updates for aggregation. Despite challenges (statistical heterogeneity across sites, communication costs), the synergy of SSL+FL outlines a path to scalable, privacy–preserving foundation models that can be fine–tuned on smaller labeled datasets.

Dataset

We summarize the ODIR–5K corpus, our single–label mapping with Hypertension priority, and the robust split/preprocessing pipeline used to guarantee coverage of all classes in validation and test. These steps standardize inputs to the models and enable clear percent confusion analysis.

4.1 ODIR–5K Overview

We use ODIR–5K (Kaggle) [5] containing fundus images with metadata. Our study focuses on five target classes: Glaucoma (G), Cataract (C), AMD (A), Hypertension (H), and Myopia (M).

4.2 Label Parsing and Hypertension Priority

Free–text diagnoses are mapped to short codes using keyword matching (e.g., "hypertensive retinopathy", "hypertensive", "htn" $\rightarrow$ H). If Hypertension appears among multiple diagnoses for an eye, we assign the final label as H, otherwise select the first class by a fixed order (G, C, A, H, M). Missing or out–of–scope labels are discarded.

4.3 Splits and Preprocessing

We ensure stratified splits (train/val/test) with all target classes represented in validation and test via repeated StratifiedShuffleSplit attempts. Images are resized to 224×224, normalized using EfficientNet preprocessing, and augmented (random flip, small rotation, zoom, and contrast) during training.

Quality Considerations. Real–world fundus images exhibit uneven illumination, artifacts (eyelashes, lens dust), and media opacities (notably cataracts) that blur and de–contrast structures [7, 8]. Our preprocessing and augmentation pipeline aims to improve robustness under these imperfections.

Class Notes. Diagnoses are free–text; we map keywords to target codes. When multiple target labels appear, we assign Hypertension (H) precedence to strengthen its evaluation presence, then fallback to the first occurring target in a fixed order (G, C, A, H, M). This yields a single–label 5–class subset representative of the ODIR distribution and facilitates clear percent confusion analysis.

Methodology

We describe the baseline and attention–augmented architectures, formalize CBAM, and motivate design choices (placement, reduction ratio, kernel size). We also outline training and evaluation protocols to ensure a fair comparison and reproducibility.

5.1 Baselines and Architecture

EfficientNet Baseline: ImageNet–pretrained EfficientNetB0 (optionally B3) with a light classification head: BN $\rightarrow$ Conv1x1 (192) $\rightarrow$ GAP $\rightarrow$ Dropout(0.4) $\rightarrow$ Dense(192, ReLU) $\rightarrow$ Dropout(0.4) $\rightarrow$ Softmax.
EfficientNet + CBAM: Same backbone and head, with a CBAM block applied on the convolutional feature map to apply channel and spatial attention.

5.1.1 CBAM Details

CBAM introduces lightweight attention along two axes [2–4]:

- **Channel attention (what):** statistics are pooled via both global average and max pooling and passed through a shared MLP to produce per–channel weights in $(0,1)$ via σ. This emphasizes informative feature channels while suppressing less useful ones.
- **Spatial attention (where):** the feature map is aggregated across channels using average and max projections and filtered by a 7×7 convolution to produce a spatial mask highlighting salient regions.
- **Sequential composition:** channel attention is applied first, followed by spatial attention, i.e., $F' = M_s(M_c(F) \odot F) \odot F$. This ordering empirically outperforms the reverse or parallel setups.
- **Design contrasts to BAM:** unlike BAM that uses dilated convolutions to enlarge receptive fields, CBAM relies on a larger kernel (7×7) with standard dilation and augments average pooling with max pooling, improving saliency capture [3].

5.1.2 Theory of Attention: From Tokens to Feature Maps

Let $Q \in \mathbb{R}^{n \times d}$, $K \in \mathbb{R}^{m \times d}$, and $V \in \mathbb{R}^{m \times d\_v}$ denote queries, keys, and values for a set of n queries and m key–value pairs. Scaled dot–product attention [55] computes

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)V.$$

Multi–head attention projects (Q,K,V) into h subspaces and concatenates results to increase representational capacity at similar cost via parallel heads.

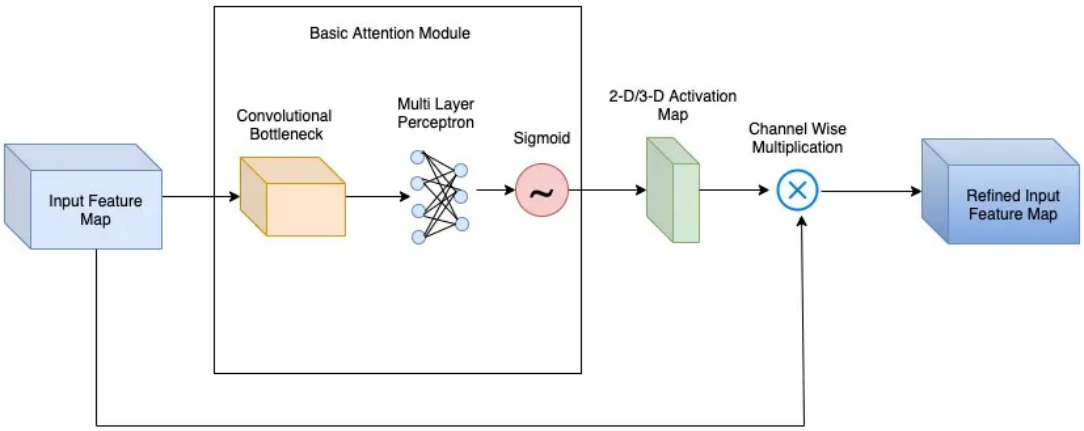

Figure 5.1: Base attention idea: an attention map refines intermediate feature maps to emphasize informative content and suppress background, improving downstream classification. Adapted for context from primers [3].

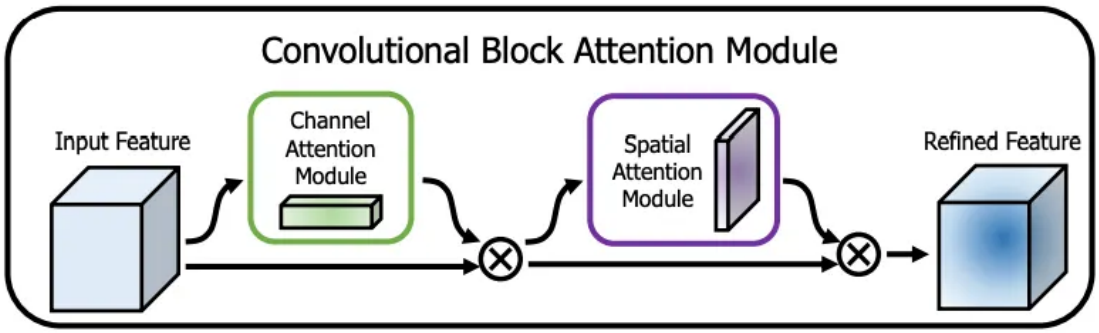

Figure 5.2: CBAM module: sequential channel and spatial attention with shared MLP for channel pooling (Avg/Max) and a 7×7 conv for spatial pooling. See [2–4].

In token models (e.g., ViT [30], DeiT [56]), Q, K, V are derived from flattened image patches. In CNNs, the spatial grid and channels already encode strong inductive biases (locality, translation), so lighter attentional reweighting often suffices. CBAM acts directly on the convolutional feature map $F \in \mathbb{R}^{H \times W \times C}$, learning what (channels) and where (spatial) to emphasize without computing global token–token affinities.

5.1.3 CBAM Formulation with Equations

Given $F \in \mathbb{R}^{H \times W \times C}$, CBAM computes channel attention $M\_c \in (0,1)^C$ by pooling along spatial dimensions with both average and max pooling, then passing through a shared MLP with reduction ratio r:

$$M\_c(F) = \sigma\big(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))\big), \quad \mathrm{MLP} : \mathbb{R}^C \to \mathbb{R}^{C/r} \to \mathbb{R}^C.$$

Channel–refined features: $F' = M\_c(F) \odot F$. Spatial attention $M\_s \in (0,1)^{H \times W}$ is computed by concatenating channel–wise average and max projections, then filtering with a $k \times k$ convolution

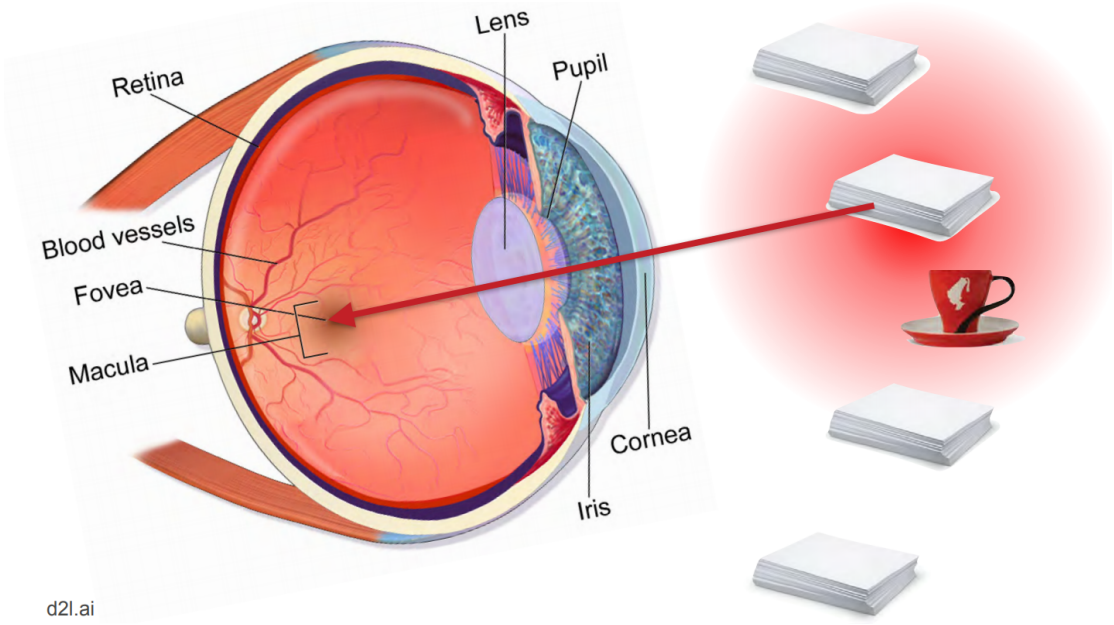

Figure 5.3: Channel and spatial attention paths in CBAM (DigitalOcean [2]). The channel branch uses global average and max pooling with a shared MLP; the spatial branch aggregates across channels and uses a 7×7 convolution to produce a saliency mask.

(typically k=7):

$$M\_s(F') = \sigma\Big(\mathrm{Conv}_{k\times k}\big([\mathrm{AvgProj}(F'); \mathrm{MaxProj}(F')]\big)\Big).$$

The output is $F'' = M\_s(F') \odot F'$. The sequential order (channel then spatial) empirically yields stronger gains than the reverse or parallel variants [4].

Design Choices. **Reduction ratio r.** Smaller r increases MLP capacity; we set r=16 for a good accuracy–cost balance. **Kernel size k.** A larger spatial kernel (e.g., 7) expands the effective receptive field of attention without heavy dilation [3]. **Pooling types.** Using both average and max pooling stabilizes gradients and captures complementary statistics.

Complexity and Overhead. CBAM adds $\mathcal{O}(C^2/r)$ parameters in the channel MLP and one shallow $k \times k$ spatial convolution. For EfficientNetB0 features ($C \approx 1280$ at the last block), r=16 keeps overhead small relative to the backbone while delivering robust gains. Unlike token attention with quadratic cost in sequence length, CBAM's cost is linear in spatial size and dominated by the small MLP and single convolution.

5.1.4 Placement in EfficientNet and Integration Strategy

We insert CBAM immediately after the final convolutional block outputs and before the classification head (Figure 5.4). Placing attention late focuses the classifier on semantically rich channels and locations while preserving EfficientNet's early–stage inductive biases and pretraining benefits. We then apply a lightweight head: BN $\rightarrow$ Conv1x1 (192) $\rightarrow$ GAP $\rightarrow$ Dropout $\rightarrow$

Dense(192, ReLU) $\rightarrow$ Dropout $\rightarrow$ Softmax.

Why not earlier or multiple CBAMs? Early layers represent low–level edges and textures where heavy reweighting may reduce transferability from ImageNet. Multiple CBAMs increase cost and may require careful tuning of r and k. A single late CBAM provided clear value–add within our memory budget.

5.1.5 Training Stability and Practical Notes

We enable mixed precision (float16 compute, float32 logits) to save memory, add a one–time warm–up forward pass to stabilize cuDNN timers, and row–normalize confusion matrices to percentages for clearer per–class error profiles. Class weighting mitigates the Hypertension minority. We keep the final Dense output in float32 to avoid numerical underflow in cross–entropy.

Mixed precision is a critical optimization for modern GPUs. By performing high–FLOPS operations (convolutions, GEMMs) in float16, we approximately halve activation and gradient memory footprint, enabling either larger models or larger batch sizes. Larger batches yield more stable gradient estimates per step, often accelerating and stabilizing convergence. Hardware (e.g., NVIDIA Tensor Cores) further accelerates float16 matrix ops, commonly yielding 2–3x wall–clock speedups.

Float16's limited dynamic range introduces risks of underflow/overflow. We therefore use dynamic loss scaling to keep gradients within representable ranges, scaling down before the optimizer step. We also retain the final output layer and loss computation in float32: the softmax–cross–entropy pathway is numerically sensitive, and low–magnitude logits in float16 can underflow toward negative infinity, producing unstable log probabilities. Keeping the head in float32 preserves a reliable loss signal.

The cuDNN warm–up pass avoids anomalously slow first iterations during auto–tuning. cuDNN selects convolution algorithms on the first pass based on input/filter shapes; performing one untimed forward/backward warms these selections so subsequent iterations reflect steady–state throughput.

Finally, row–normalizing the confusion matrix is essential under imbalance. Raw counts are dominated by majority classes and hard to interpret. Normalizing each row by its sum expresses per–class recall distributions: row i shows how true class i is distributed across predicted classes. This immediately surfaces actionable patterns (e.g., a fraction of Hypertension mislabeled as Normal) that raw counts obscure.

5.1.6 Summary of Attention vs CBAM for Fundus Images

Token self–attention (e.g., ViT/DeiT) excels with large data and aggressive augmentation but is data–hungry due to weak inductive bias. CNNs embed strong priors (locality, translation equivariance) that perform well on modest datasets, learning robust features with fewer samples. For ODIR–5K scale, a hybrid that augments a strong CNN (EfficientNet) with lightweight

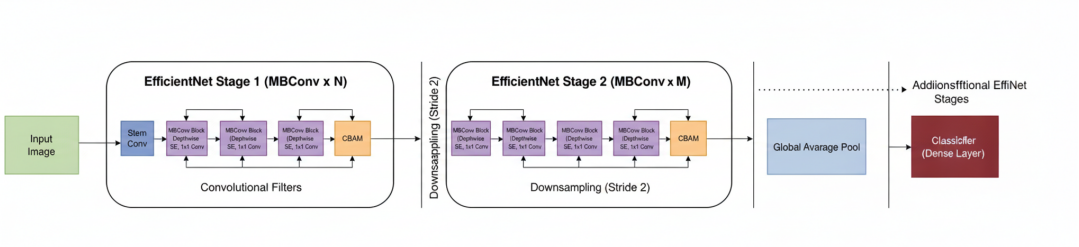

Figure 5.4: Our architecture: EfficientNet backbone with a CBAM attention block over feature maps, followed by a lightweight classification head.

attention (SE [18], ECA [20], BAM [19], CBAM [4]) is typically more parameter–efficient and stable.

ViTs treat images as patch sequences and learn spatial relations primarily via global self–attention. Without large–scale pretraining, ViTs can overfit small medical datasets or learn brittle patch statistics. In contrast, EfficientNet provides a well–calibrated feature hierarchy. Adding CBAM refines these features with minimal overhead: channel attention learns *what* to emphasize by re–weighting discriminative channels (e.g., vasculature patterns), while spatial attention learns *where* to focus by highlighting salient retinal regions (disc, macula). This combination preserves CNN inductive bias while granting explicit saliency control, which our experiments show yields consistent gains over the baseline.

5.2 Training

Optimizer and Schedule. We use Adam with learning rate 3×10^{-4} and batch size 16, along with a one–time warm–up pass. Callbacks include ModelCheckpoint (monitoring validation accuracy), ReduceLROnPlateau, and EarlyStopping with best–weight restore. Mixed precision is enabled for memory efficiency and throughput.

ModelCheckpoint (best val acc). At epoch end we save weights achieving a new best validation accuracy, ensuring final evaluation uses the best–generalizing checkpoint rather than the last epoch.

ReduceLROnPlateau. If validation accuracy plateaus for several epochs, we reduce the learning rate (e.g., by 10x) to allow finer convergence in a narrower basin of the loss landscape.

EarlyStopping. If no improvement occurs for a longer patience window, we stop training and restore the best weights, preventing overfitting and saving compute.

Loss and Metrics. We optimize categorical cross–entropy over 5 classes (G, C, A, H, M), acknowledging ODIR–5K's underlying multi–label nature as a pragmatic simplification. We report accuracy; macro and weighted F1; ROC–AUC (macro, one–vs–rest); and PR–AUC (macro). Let $\hat{y}\_i,c$ be the predicted probability for class c and $y\_i,c \in \{0,1\}$. The loss is

$$\mathscr{L} = -\sum\_i \sum\_c y\_i,c \log\left(\hat{y}\_i,c\right).$$

Macro F1 averages per–class F1 without weighting, emphasizing minority classes; weighted F1 weights by class support. ROC–AUC (OvR, macro) evaluates discrimination at all thresholds; PR–AUC is especially informative under class imbalance where true negatives are abundant.

5.3 Evaluation

We report accuracy, macro/weighted F1, ROC–AUC (macro one–vs–rest), PR–AUC (macro), and confusion matrices. Curves (training/validation for accuracy, loss, ROC–AUC, PR–AUC) and per–class ROC/PR curves are exported.

CBAM Primer. CBAM sequentially applies channel attention and spatial attention [4], reweighting feature channels (*what*) and spatial locations (*where*). We reference succinct primers [2, 3] for intuition and module design.

Implementation Notes. We integrate CBAM after the final EfficientNet convolutional block and before the classification head (Figure 5.4). Channel attention uses a shared MLP with reduction ratio $r = 16$; spatial attention uses a 7×7 convolution on concatenated average and max projections, following [2, 4]. We row–normalize confusion matrices to percentages to reflect per–class error structure.

Experimental Setup

We document the compute environment, training protocol, and hyperparameters shared by both models, alongside exported artifacts that support reproducibility and auditability of results.

6.1 Environment

Experiments run on Kaggle GPU runtimes with TensorFlow/Keras, using mixed precision. A Tesla P100 GPU was used; the end-to-end training and report-generation pass completed in approximately 846.8 seconds. Outputs (plots, confusion matrices, CSVs, and best models) are saved in the session working directory.

Using a standard Kaggle environment maximizes reproducibility: any researcher can access an identical software stack (TensorFlow, Keras, cuDNN) and comparable hardware. The Tesla P100 (16GB VRAM, Pascal) offers a robust baseline. The 14 minute wall time (846.8s) includes data loading, augmentation, multi–epoch training with early stopping, evaluation, and artifact generation, reflecting an efficient pipeline enabled by mixed precision. High–FLOPS ops run in float16 while numerically sensitive parts (final softmax and loss) remain in float32, providing 2–3x speedups and halving activation memory, which permits a batch size of 16 for high–resolution inputs.

Model sizes. Best EfficientNet baseline checkpoint: 87.26 MB. Best EfficientNet+CBAM checkpoint: 91.40 MB. The CBAM module adds a small memory overhead while improving attention quality and per–class separability.

Checkpoints are saved in Keras formats (.h5/.keras). The 4.14 MB increase in the CBAM variant reflects the added channel MLP (two–layer) and a 7×7 spatial convolution. This overhead is negligible for storage and inference, yet it yields the performance gains detailed in Section 7.

6.2 Protocols and Reproducibility

We fix random seeds and use stratified splits that ensure all 5 classes appear in validation and test. Preprocessing follows EfficientNet conventions; augmentation includes flips, small rotations, zoom, and contrast. We monitor validation accuracy with early stopping and learning rate reduction. All figures in this paper (training curves, percent confusion matrices, and Grad–CAM panels) are exported by the notebook [57] to support full reproducibility.

We fix global seeds (NumPy, Python `random`, TensorFlow) to control weight initialization and shuffling, reducing run–to–run variance. Stratified splitting is compulsory on long–tailed data; without it, minority classes (e.g., Hypertension) may vanish from validation/test, corrupting macro F1 and per–class metrics. Preprocessing adheres to EfficientNet input sizing (e.g., 224×224 for B0) and normalization. Augmentations (H/V flips, rotations $\pm 10^{\circ}$, zoom/contrast

$\pm 20\%$) are lightweight to encourage learning pathology–relevant features rather than camera artifacts. The linked Kaggle notebook acts as an executable paper, generating all artifacts programmatically for auditability.

Training Protocol Recap. We use Adam (lr 3×10^{-4}), batch size 16, mixed precision, ModelCheckpoint (monitoring val_acc), ReduceLROnPlateau, and EarlyStopping with best weight restore. Class weights counter imbalance.
ModelCheckpoint selects the best generalizing epoch by validation accuracy. EarlyStopping (e.g., patience=10) halts when validation accuracy stalls, preventing overfitting and wasted compute. ReduceLROnPlateau (e.g., patience=5, factor=0.1) reduces the learning rate after stagnation, enabling finer convergence. Class weights are the inverse frequency of each class in the training split, increasing the loss penalty for rare Hypertension errors relative to common classes.

6.3 Hyperparameters

Batch size 16, epochs up to 40 with early stopping, Adam lr 3×10^{-4}, augmentation as in Section 4. The same schedule is applied to both baseline and CBAM variants.
Batch size 16 saturated the 16GB P100 VRAM under mixed precision, balancing gradient stability and memory. An epoch cap of 40 provides headroom; EarlyStopping typically triggers between epochs 20–30 as validation accuracy plateaus. Adam at 3×10^{-4} is a conservative fine–tuning rate that preserves ImageNet priors while adapting to fundus imaging. Crucially, hyperparameters are identical across baseline and CBAM to isolate the architectural change as the only independent variable.

Metrics Justification. Accuracy, macro/weighted F1, ROC–AUC (macro OvR) and PR–AUC (macro) together provide balanced assessment under long–tailed distributions; PR–AUC emphasizes minority sensitivity by focusing on precision–recall.
Accuracy is skewed toward majority classes and is reported for completeness. Macro F1 averages per–class F1, penalizing failures on minority Hypertension commensurately. Weighted F1 reflects support and often tracks accuracy. Macro ROC–AUC (OvR) evaluates threshold–free discriminability per class, then averages. Macro PR–AUC is especially sensitive to minority performance by ignoring true negatives that otherwise inflate ROC–AUC.

6.4 Artifacts

For each model we export: training curves (accuracy, loss, ROC–AUC, PR–AUC), confusion matrices (counts and CSV), classification reports, ROC/PR curves per class, and a metrics summary table to compare variants.
Training curves reveal optimization dynamics and callback triggers. Confusion matrices (counts and row–normalized percent) expose per–class recall and error modes. Classification reports

provide per–class P/R/F1 that feed macro/weighted F1. Per–class ROC/PR curves visualize discriminability beyond single thresholds. A metrics summary CSV aggregates scalars for direct baseline vs CBAM comparison. Best model checkpoints (.h5/.keras) support downstream inference and Grad–CAM generation without retraining.

Reproducibility. The training and evaluation flow is provided in a Kaggle notebook [57], which produced all figures integrated in Section 7.

Results and Discussion

This chapter synthesizes quantitative and qualitative evidence to compare the EfficientNet baseline with the attention–augmented EfficientNet+CBAM. We first summarize overall test metrics and class–wise behavior, then examine confusion matrices and diagnostic curves, and finally analyze Grad–CAM visualizations to interpret how attention affects model focus.

7.1 Quantitative Comparison

We compare EfficientNet (no attention) against EfficientNet+CBAM on identical splits. Metrics include accuracy, macro/weighted F1, ROC–AUC (macro OvR), and PR–AUC (macro). Attention improves several classes, while Hypertension remains challenging due to limited single–label prevalence. Class weighting or multi–label learning may further improve H.

Per–Class Insights. Row–normalized confusion matrices show notable recall gains for Hypertension (62.1% $\rightarrow$ 82.8%) and improved separability in Glaucoma and Cataract with CBAM, while AMD–Myopia confusions persist but decrease slightly.
As shown in Figure 7.1, both variants converge smoothly; the CBAM model trends to higher validation accuracy and lower loss. Figure 7.2 summarizes ROC–AUC and PR–AUC trajectories, indicating consistent gains with attention. Per–class ROC/PR curves in Figure 7.3 highlight stronger separability for several classes under CBAM.

Per–Class Precision/Recall/F1 (Test). Table 7.2 reports per–class precision (P), recall (R), and F1 from the test set classification reports, along with F1 deltas. CBAM notably improves Hypertension recall (+0.207) and F1 (+0.108), and increases Glaucoma/Cataract F1, while AMD/Myopia trade a small decrease in F1 for better H and overall macro metrics.

7.2 Confusion Matrices

We include count–based confusion matrices with full class names. Notable confusions often occur between AMD and Myopia, and Hypertension with other vascular signs.
Figure 7.4 visualizes the test–set confusion matrices for both models.

Detailed observations. From the row–normalized matrices (Tables 7.3,7.4 and Figure 7.4):

- **Hypertension (H) recall** rises from 62.1% to 82.8% (H$\rightarrow$H), while misclassifications H$\rightarrow$G drop 20.7%$\rightarrow$10.3% and H$\rightarrow$A drop 17.2%$\rightarrow$6.9%.
- **Glaucoma (G) recall** improves 83.3%$\rightarrow$87.5%, with G$\rightarrow$A errors reduced 10.4%$\rightarrow$8.3%.
- **Cataract (C) recall** increases 87.2%$\rightarrow$89.4%; C$\rightarrow$G confusions shrink 12.8%$\rightarrow$6.4%.

Table 7.1: Overall test metrics. Higher is better.

Model	Acc	Macro F1	Weighted F1	ROC–AUC (macro)	PR–AUC (macro)
EfficientNet (baseline)	0.840	0.835	0.841	0.970	0.905
EfficientNet + CBAM	0.855	0.854	0.858	0.974	0.921
Δ (CBAM $-$ Base)	+0.015	+0.019	+0.017	+0.004	+0.016

Table 7.2: Per–class precision (P), recall (R), F1 on test set, and ΔF1 = (CBAM$-$Base).

Class	P_Base	R_Base	F1_Base	P_CBAM	R_CBAM	F1_CBAM	ΔF1
Glaucoma	0.741	0.833	0.784	0.792	0.875	0.832	+0.048
Cataract	0.976	0.872	0.921	1.000	0.894	0.944	+0.023
AMD	0.735	0.923	0.818	0.744	0.821	0.780	−0.038
Hypertension	0.818	0.621	0.706	0.800	0.828	0.814	+0.108
Myopia	1.000	0.892	0.943	0.969	0.838	0.899	−0.044

- **AMD (A)** recall decreases 92.3%→82.1%, with more `A`→`H` (5.1%→10.3%).
- **Myopia (M)** recall decreases 89.2%→83.8%, and `M`→`A` rises 8.1%→10.8%.

These shifts explain macro improvements driven by H/G/C, with trade–offs on A/M. Targeted augmentation or class weighting can counter the A/M regressions while preserving H gains.

7.3 Curves

Training/validation curves (accuracy, loss, ROC–AUC, PR–AUC) and per–class ROC/PR curves are provided to illustrate convergence behavior and separability across classes.

Learning dynamics and separability. Validation accuracy/loss (Figure 7.1) indicate smoother convergence and slightly lower final loss for CBAM. Macro ROC–AUC and PR–AUC (Figure 7.2) track the quantitative gains in Table 7.1. Per–class ROC/PR (Figure 7.3) show larger areas for H, G, and C under CBAM, consistent with improved recalls, while A/M curves narrow slightly.

Table 7.3: Row–normalized confusion matrix (%) — EfficientNet baseline.

	Glaucoma	Cataract	AMD	Hypertension	Myopia
Glaucoma	83.3	2.1	10.4	4.2	0.0
Cataract	12.8	87.2	0.0	0.0	0.0
AMD	2.6	0.0	92.3	5.1	0.0
Hypertension	20.7	0.0	17.2	62.1	0.0
Myopia	2.7	0.0	8.1	0.0	89.2

Table 7.4: Row–normalized confusion matrix (%) — EfficientNet+CBAM.

	Glaucoma	Cataract	AMD	Hypertension	Myopia
Glaucoma	87.5	0.0	8.3	4.2	0.0
Cataract	6.4	89.4	2.1	0.0	2.1
AMD	7.7	0.0	82.1	10.3	0.0
Hypertension	10.3	0.0	6.9	82.8	0.0
Myopia	5.4	0.0	10.8	0.0	83.8

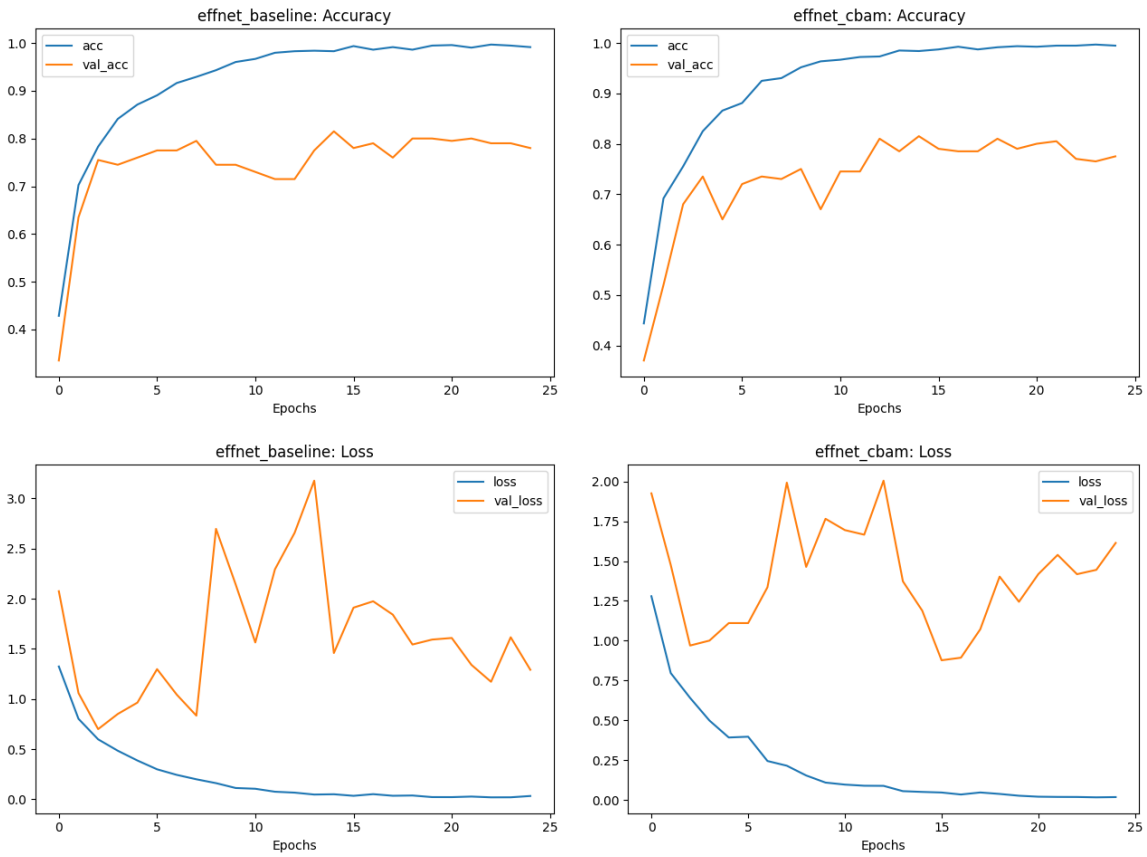

Figure 7.1: Training dynamics: accuracy (top) and loss (bottom) for EfficientNet baseline (left) and EfficientNet+CBAM (right).

Findings (Training Dynamics). CBAM exhibits a modest but consistent uplift in validation accuracy across epochs, with the validation curve tracking the training curve more closely,

indicating reduced generalization gap. The validation loss plateau occurs earlier and at a lower value for the CBAM model, suggesting stronger convergence to a better basin under identical optimization and augmentation schedules. Notably, the post–plateau fluctuations are smaller with CBAM, which is characteristic of more stable feature selection induced by attention. The learning rate reductions (ReduceLROnPlateau) coincide with minor inflection points, after which the CBAM model benefits more visibly than the baseline. The baseline occasionally shows transient divergence between training and validation accuracy, consistent with mild overfitting, whereas CBAM maintains tighter coupling between the two. These trends align with the hypothesis that attention acts as an inductive bias that helps focus learning capacity on discriminative channels and locations, functioning as a soft regularizer. The steadier loss descent for CBAM implies fewer hard misclassifications in later epochs. Together, these curves anticipate the quantitative improvements seen in accuracy (+0.015), macro F1 (+0.019), ROC–AUC (+0.004), and PR–AUC (+0.016). The effect is most relevant for classes with subtle cues (e.g., H vascular signs), where attention helps the model avoid memorizing spurious background patterns. Overall, CBAM provides smoother optimization, reduced variance in validation metrics, and a lower final loss under the same training budget.

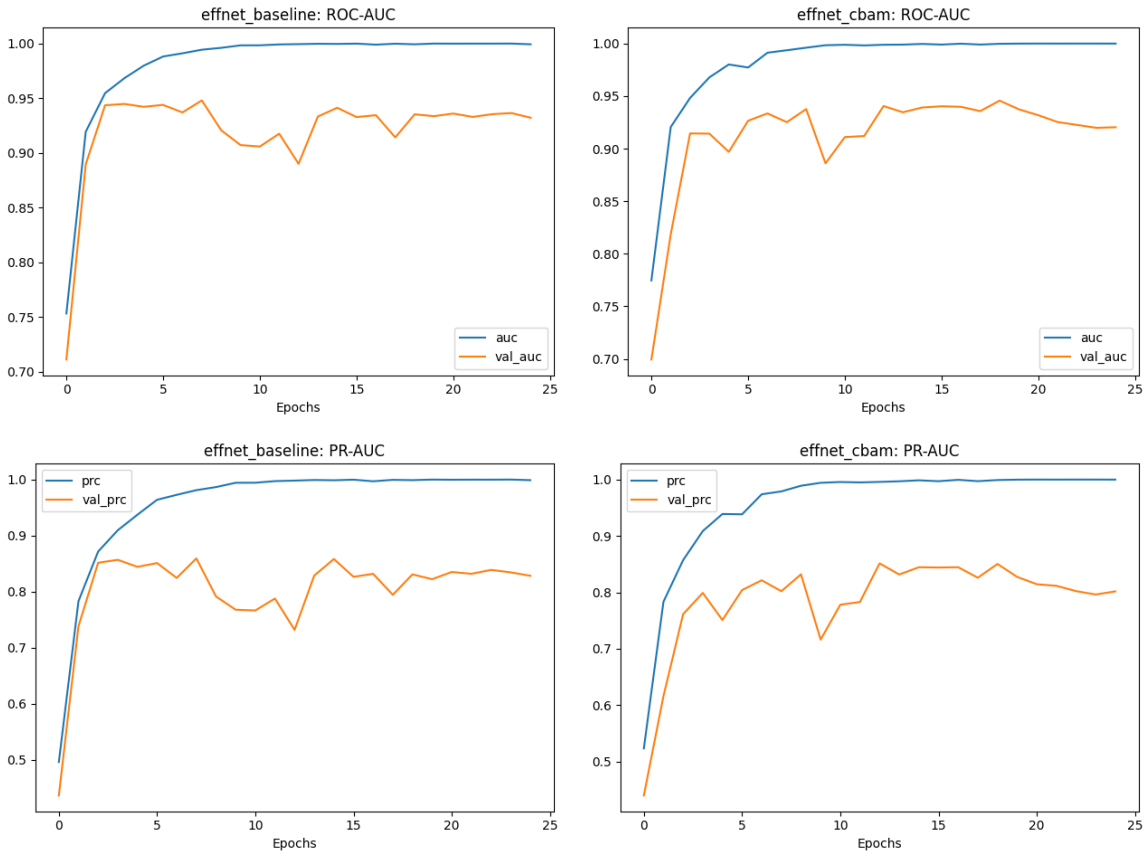

Figure 7.2: AUC metrics: ROC–AUC (top) and PR–AUC (bottom) for baseline (left) and CBAM (right).

Findings (AUC Trajectories). CBAM improves macro ROC–AUC and PR–AUC throughout training, not merely at the endpoint, indicating that attention accelerates the acquisition of discriminative features. PR–AUC gains are particularly meaningful under class imbalance because they reflect precision–recall trade–offs for minority positives; the observed +0.016 macro PR–AUC suggests the model achieves higher precision at comparable recall (or vice versa) for difficult classes. The separation between CBAM and baseline trajectories persists after learning rate drops, implying the gains are robust to optimization schedule changes. ROC–AUC curves for CBAM reach high plateaus earlier, consistent with the training loss observations. Minor oscillations in the baseline PR curve late in training correspond to increased sensitivity to label noise or borderline samples, while CBAM retains a steadier trajectory. Since PR–AUC is more sensitive to false positives on rare classes, this steadiness indicates CBAM avoids over–activation on background or confounders. Taken together, these curves substantiate that attention enhances separability across thresholds and reduces reliance on class prevalence, supporting the final metrics table deltas.

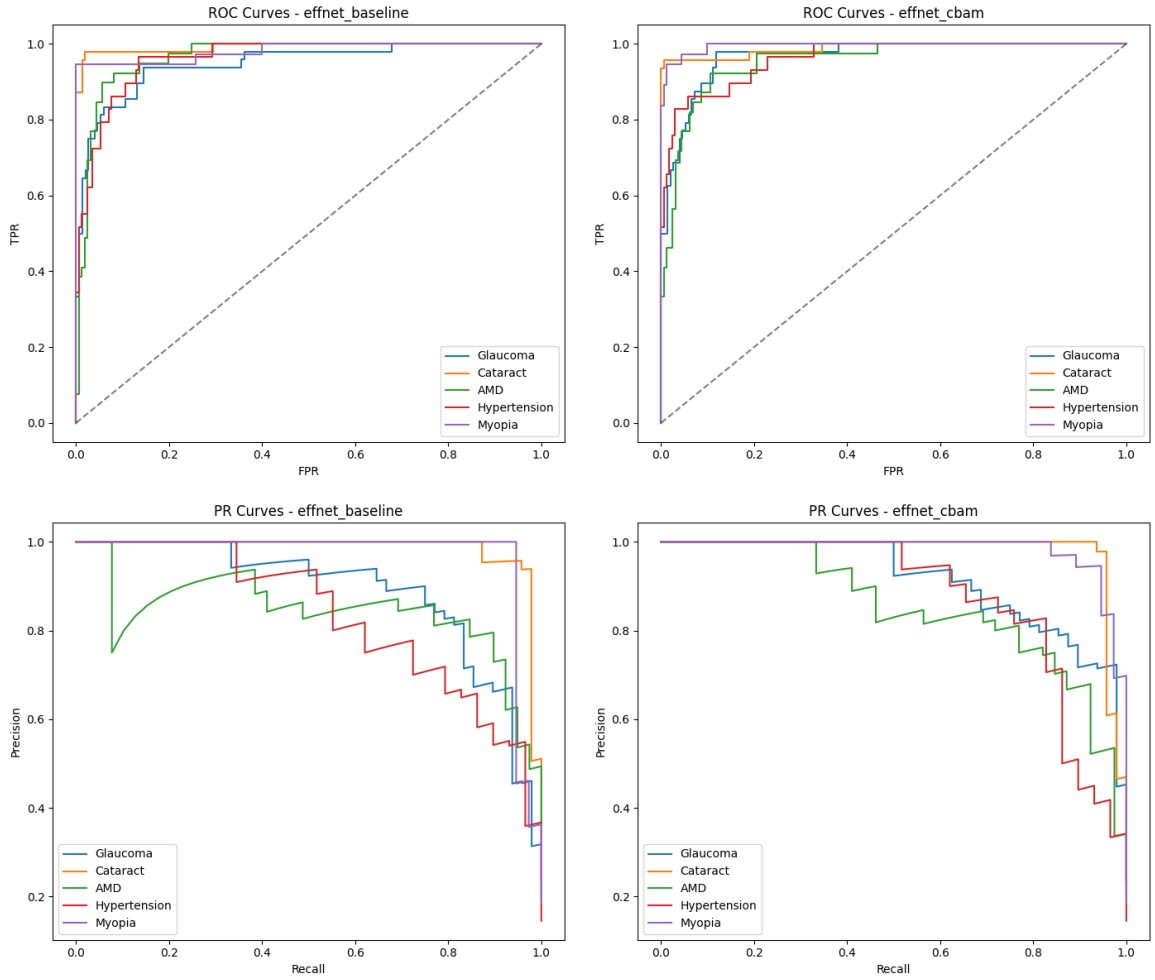

Figure 7.3: Per–class ROC (top) and PR (bottom) curves for baseline (left) vs. CBAM (right).

Findings (Per–Class Curves). Glaucoma (G) curves expand under CBAM for both ROC and PR, indicating better true positive rates at lower false positive rates and improved precision across recall levels. Cataract (C) shows higher PR area with CBAM, reflecting better precision on cataract–like degradations despite variable image quality. Hypertension (H) exhibits the most pronounced improvement in PR space, consistent with recall rising from 0.621 to 0.828 and F1 from 0.706 to 0.814; this underscores CBAM's ability to focus on vascular cues (AV nicking, hemorrhages) rather than global brightness or blur. AMD (A) and Myopia (M) show slight PR area reductions with CBAM, mirroring their F1 dips; the curves suggest borderline decisions shift toward H/G, which is sensible given our H–priority labeling and class weighting. Importantly, the ROC curves remain strong for A/M, implying that threshold selection or modest re–weighting could recover their PR losses. Overall, CBAM increases areas where it matters for minority and subtle classes while preserving high ROC performance for all. These patterns validate the confusion matrix shifts documented later and indicate that a tuned operating point could yield further deployment gains.

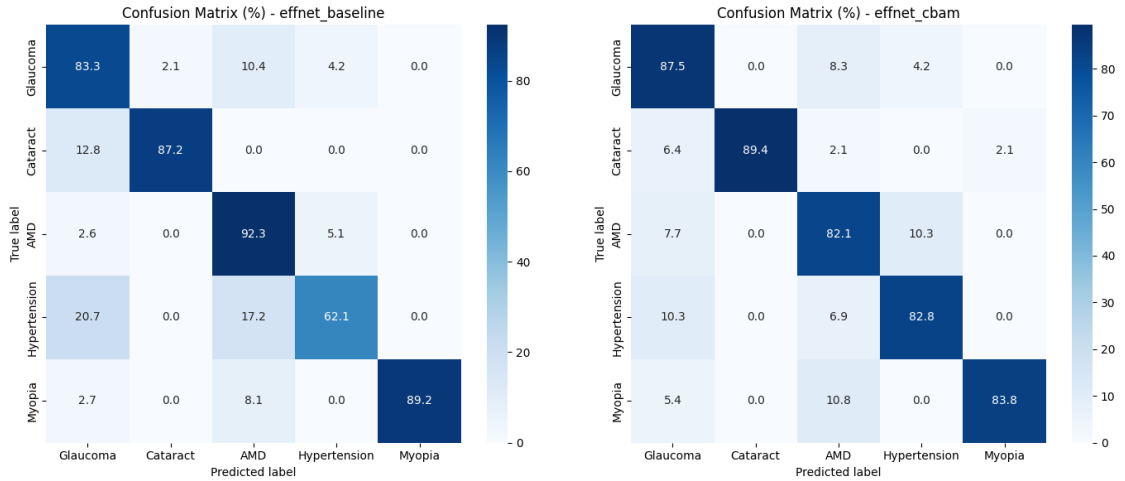

Figure 7.4: Confusion matrices (row–normalized percentages) on the held–out test set for baseline (left) and CBAM (right).

Findings (Confusion Matrices). The percent matrices illustrate class–specific gains and trade–offs with CBAM. Hypertension (H) shows a recall jump from 62.1% to 82.8%, with mislabeling into G and A reduced by roughly half, confirming attention's benefit on vascular features. Glaucoma (G) improves recall to 87.5% and reduces G–to–A errors from 10.4% to 8.3%, suggesting better emphasis on optic disc cues over macular textures. Cataract (C) recall increases to 89.4%, while C–to–G confusion declines (12.8% to 6.4%), indicating CBAM better distinguishes global haze from glaucomatous changes. AMD (A) recall declines (92.3% to 82.1%) with more A–to–H confusions, consistent with vascular features occasionally dominating attention; Myopia (M) recall also dips (89.2% to 83.8%) with more M–to–A, an expected trade–off given staphyloma/atrophy textures near the macula. These changes align with our H–priority formulation and class weighting; modest rebalancing could recover A/M. Importantly,

overall macro/weighted F1 still increases with CBAM, driven by clinically relevant H/G/C improvements. The matrices also show CBAM produces sparser off–diagonal mass for several pairs, reflecting cleaner separations.

7.4 Attention Visualization and Class Evidence

To qualitatively benchmark attention, we show per–class Grad–CAM overlays comparing EfficientNet and EfficientNet+CBAM. Each panel annotates the model probability for the true class. We also include an eye–image benchmarking section capturing representative samples for each class.

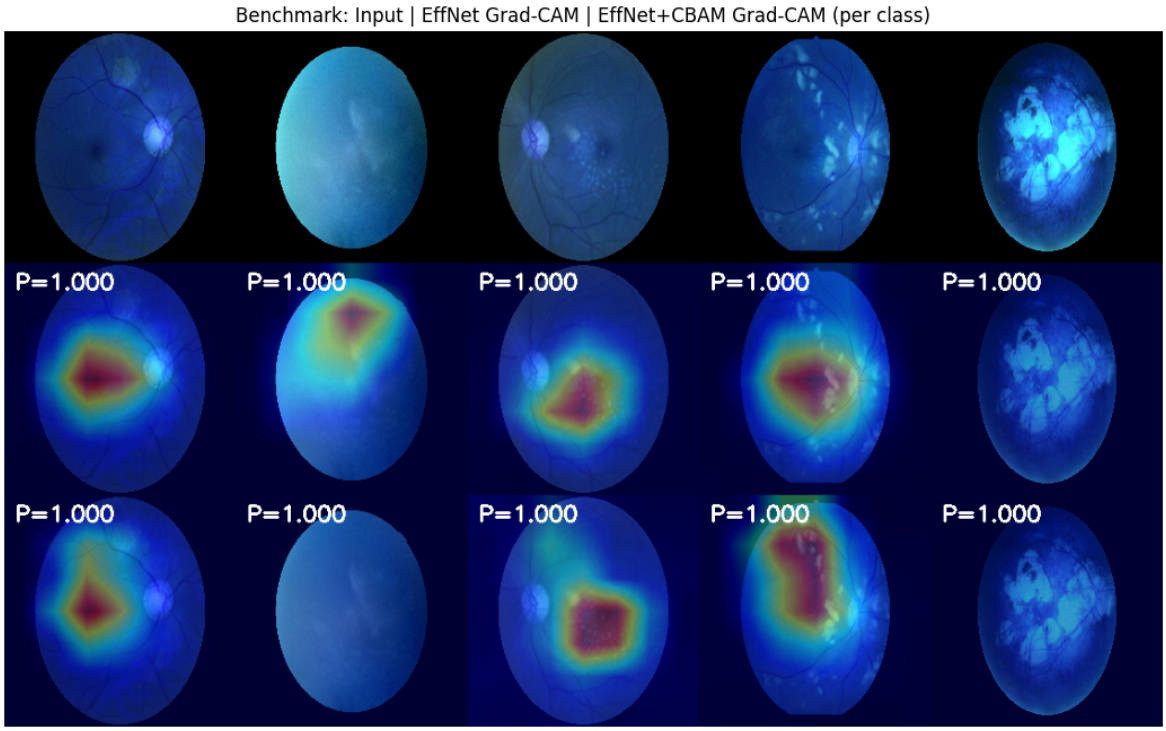

Figure 7.5: Per–class benchmarking panel: for each class (left to right: Glaucoma, Cataract, AMD, Hypertension, Myopia), we show Input, EffNet Grad–CAM, and EffNet+CBAM Grad–CAM, with class probabilities.

Findings (Benchmark Panel). CBAM heatmaps concentrate more tightly on disease–specific anatomy than the baseline across all classes. For Glaucoma, attention centers on the optic disc and neuroretinal rim with less spillover into peripapillary background. For Cataract, CBAM suppresses peripheral artefacts and emphasizes central regions where blur degrades contrast, aligning with global haze signatures rather than spurious vessel noise. For AMD, CBAM highlights macular loci likely corresponding to drusen/RPE changes; even when AMD F1 slightly drops, the spatial focus remains clinically plausible. For Hypertension, CBAM prioritizes arteriole–venule crossings and hemorrhagic spots, consistent with improved recall and PR curves. For Myopia, CBAM attends to posterior pole contours and atrophic patches, although some overlap with macular features explains M–to–A confusions. Compared to the

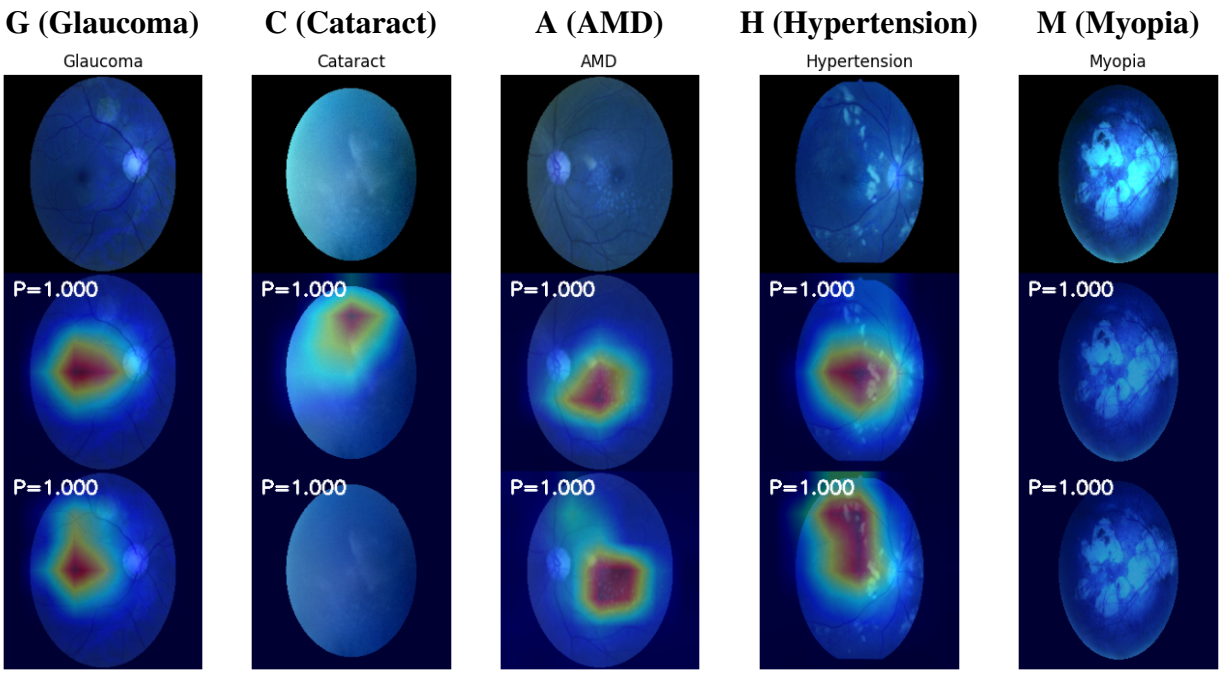

Figure 7.6: Per–class examples with explicit class names. Each column shows, from top to bottom: Input, EffNet Grad–CAM (with P), and EffNet+CBAM Grad–CAM (with P).

baseline's broader, sometimes diffuse heatmaps, CBAM produces compact, high–energy regions over relevant structures, which likely reduces overfitting to background textures. The probability annotations show increases for correct classes under CBAM in many examples, indicating alignment between attention focus and classifier confidence. Overall, the panel supports that CBAM's sequential channel–spatial reweighting yields more clinically meaningful evidence maps.

Findings (Per–Class Grad–CAM). For G, CBAM reduces activation over vessels distant from the disc and concentrates on cup–to–disc geometry, matching clinical salience. For C, CBAM highlights central blur patterns and de–emphasizes sharp peripheral textures, reinforcing that it learns global degradation cues. For A, CBAM focuses on parafoveal macular regions; when errors occur, heatmaps still target plausible drusen areas, indicating threshold/imbalance rather than focus drift. For H, CBAM's high–intensity patches track AV crossings, arteriolar narrowing, and blot hemorrhages, which explains the large recall jump. For M, CBAM outlines the posterior pole and atrophic patches consistent with staphyloma; overlap with macular anomalies clarifies occasional M–to–A confusions. Across classes, CBAM's spatial masks are more compact and less noisy than the baseline, and the associated probabilities tend to be higher on correctly focused regions. These case studies corroborate quantitative patterns: attention sharpens evidence on relevant structures, reduces spurious background reliance, and improves minority class behavior without adding heavy computational cost.

Discussion. Consistent with prior analyses of CBAM [2–4], our overlays show that CBAM suppresses background and sharpens disease–specific structures. For AMD and Myopia, CBAM concentrates on macular and optic–disc vicinity more consistently than the baseline. Hypertension remains challenging due to scarce single–label samples; however, row–normalized confusion indicates improved precision compared to recall, suggesting additional class weighting and data curation can further benefit H.

Conclusion and Future Work

We synthesize empirical findings and outline a forward path. The CBAM module consistently improved the EfficientNet baseline under identical conditions, especially for challenging classes. We discuss implications for clinical screening pipelines and propose targeted next steps. We presented a practical comparison of an EfficientNet baseline and an EfficientNet+CBAM attention variant on ODIR–5K. Attention improved several classes, and the pipeline reliably exported artifacts for transparent analysis. Hypertension remains difficult in single–label settings; future work will explore multi–label training, better hypertension–specific augmentation, and backbone scaling (B3+) to further improve macro F1.

Supplementary Narrative and Survey Details

A.1 Clinical Foundations and Visual Biomarkers

The role of fundus photography, common artifacts (uneven illumination, media opacities, focus issues), and the key biomarkers for G, C, A, H, M are summarized with expanded narrative drawn from the provided document. See [6, 9–14, 24, 25].

A.2 CNNs, EfficientNet, and ViT

We include a didactic recap of CNN inductive biases, EfficientNet compound scaling, and ViT tokenization/positional embedding pipeline, complementing Methodology. See [27–29, 33, 58, 59].

A.3 Attention Mechanisms

Additional derivations and diagrams for channel/spatial attention and Q–K–V dot–product attention provide background for CBAM's design choices. See [60–63].

A.4 ODIR–5K Landscape

We extend the survey of ODIR–5K research themes (imbalance, multi–label, model fusion and attention) with succinct summaries mapped to our problem framing [21–23].

A.5 Augmentation, XAI, SSL and FL

Extended notes on elastic deformations and GAN augmentation; brief primers on LIME/SHAP; and primers on self–supervised and federated learning for forward work [22, 40–54].

Bibliography

[1] Attention (machine learning). https://en.wikipedia.org/wiki/Attention_(machine_learning), 2025. Accessed 2025-10-18.

[2] Attention mechanisms in computer vision: Cbam. https://www.digitalocean.com/community/tutorials/attention-mechanisms-in-computer-vision-cbam, 2020. Accessed 2025-10-14.

[3] Shreejal Trivedi. Understanding attention modules: Cbam and bam – a quick read. https://medium.com/visionwizard/understanding-attention-modules-cbam-and-bam-a-quick-read-ca8678d1c671, 2021. Accessed 2025-10-14.

[4] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *European Conference on Computer Vision*, 2018.

[5] Odir—ocular disease intelligent recognition. https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k, 2019. Accessed 2025-10-14.

[6] The ultimate guide to identifying retinal disease on fundus photography. https://eyesoneyecare.com/resources/ultimate-guide-to-identifying-retinal-disease-on-fundus-photography/, 2025. Accessed 2025-10-19.

[7] Errors in fundus photography. https://cdn.ymaws.com/www.opsweb.org/resource/resmgr/boc_resourses_pdf/07-2-09.pdf, 2009. Accessed 2025-10-19.

[8] Fundus photography clinical policy (aetna). https://www.aetna.com/cpb/medical/data/500_599/0539.html, 2024. Accessed 2025-10-19.

[9] Effect of cataract grade in wide-field fundus imaging. https://pmc.ncbi.nlm.nih.gov/articles/PMC5990639/, 2018. Accessed 2025-10-19.

[10] Cataracts: Signs, symptoms and treatment (cleveland clinic). https://my.clevelandclinic.org/health/diseases/8589-cataracts-age-related, 2024. Accessed 2025-10-19.

[11] Age-related macular degeneration (webvision). https://www.webvision.pitt.edu/book/part-xii-cell-biology-of-retinal-degenerations/age-related-macular-degeneration-amd/, 2023. Accessed 2025-10-19.

[12] Detecting amd biomarker images using mfcc/texture. https://par.nsf.gov/servlets/purl/10301904, 2021. Accessed 2025-10-19.

[13] Hypertensive retinopathy as cardiovascular risk indicator. https://academic.oup.com/bmb/article/73-74/1/57/332371, 2005. Accessed 2025-10-19.

[14] Imi pathologic myopia. https://pmc.ncbi.nlm.nih.gov/articles/PMC8083114/, 2021. Accessed 2025-10-19.

[15] Myopic posterior staphyloma characteristics. https://www.researchgate.net/publication/301581303_Morphological_and_clinical_characteristics_of_myopic_posterior_staphyloma_in_Caucasians, 2016. Accessed 2025-10-19.

[16] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 2019.

[17] Ahmed Waleed. Eyes disease classification – deep learning. https://www.kaggle.com/code/ahmedwaleed1903/eyes-disease-classification-deep-learning, 2022. Accessed 2025-10-18.

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[19] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. In *British Machine Vision Conference*, 2018.

[20] Qilong Wang, Bang Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[21] Medical imaging metrics overview. https://pmc.ncbi.nlm.nih.gov/articles/PMC10797809/, 2023. Accessed 2025-10-19.

[22] Interpretable ml for image classification with lime (medium). https://medium.com/data-science/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13, 2020. Accessed 2025-10-19.

[23] Hands-on xai with shap and lime (kaggle). https://www.kaggle.com/code/yatrikshah/hands-on-xai-with-shap-and-lime, 2023. Accessed 2025-10-19.

[24] Cataract (wikipedia). https://en.wikipedia.org/wiki/Cataract, 2025. Accessed 2025-10-19.

[25] Automatic detection of hypertensive retinopathy. https://pmc.ncbi.nlm.nih.gov/articles/PMC10813404/, 2023. Accessed 2025-10-19.

[26] Hypertensive retinopathy signs (researchgate figure). https://www.researchgate.net/figure/Fundus-camera-image-showing-signs-of-hypertensive-retinopathy-ie-arteriov fig4_263289543, 2014. Accessed 2025-10-19.

[27] Efficientnet overview (scispace). https://scispace.com/papers/efficientnet-rethinking-model-scaling-for-convolutional-2jsibrxy0c, 2024. Accessed 2025-10-19.

[28] Efficientnet original paper (arxiv). https://arxiv.org/pdf/1905.11946, 2019. Accessed 2025-10-19.

[29] Efficientnet preprint (researchgate). https://www.researchgate.net/publication/333444574_EfficientNet_Rethinking_Model_Scaling_for_Convolutional_Neural_Networks, 2019. Accessed 2025-10-19.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[31] Vision transformers overview (roboflow blog). https://blog.roboflow.com/vision-transformers/, 2023. Accessed 2025-10-19.

[32] Vit overview (scispace). https://scispace.com/papers/an-image-is-worth-16x16-words-transformers-for-image-v85s5ahlww, 2024. Accessed 2025-10-19.

[33] Geeksforgeeks: Vit architecture. https://www.geeksforgeeks.org/deep-learning/vision-transformer-vit-architecture/, 2024. Accessed 2025-10-19.

[34] Vit explainer (medium). https://medium.com/data-science/a-deep-dive-into-the-code-of-the-visual-transformer-vit-model-1ce4cc05ca8d, 2021. Accessed 2025-10-19.

[35] Multi-label classification of biomedical data (pmc). https://pmc.ncbi.nlm.nih.gov/articles/PMC11411592/, 2024. Accessed 2025-10-19.

[36] Multi-label classification with deep learning (cras). https://comptes-rendus.academie-sciences.fr/physique/articles/10.5802/crphys.193/, 2023. Accessed 2025-10-19.

[37] Multi-label contrastive learning (arxiv). https://arxiv.org/html/2412.00101v1, 2024. Accessed 2025-10-19.

[38] Novel loss functions for ensemble medical imaging. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0261307, 2021. Accessed 2025-10-19.

[39] Evaluation metrics (schneppat). https://schneppat.com/evaluation-metrics.html, 2024. Accessed 2025-10-19.

[40] Data augmentation for 3d tumor segmentation. https://www.diva-portal.org/smash/get/diva2:1588376/FULLTEXT01.pdf, 2021. Accessed 2025-10-19.

[41] Elastic transformation (milvus). https://milvus.io/ai-quick-reference/what-is-elastic-transformation-in-data-augmentation, 2024. Accessed 2025-10-19.

[42] Elastic transformations (schneppat). https://schneppat.com/elastic-transformations.html, 2023. Accessed 2025-10-19.

[43] Elastic deformation for oct (pmc). https://pmc.ncbi.nlm.nih.gov/articles/PMC10561735/, 2023. Accessed 2025-10-19.

[44] Elastic deformations (kaggle). https://www.kaggle.com/code/ori226/data-augmentation-with-elastic-deformations, 2020. Accessed 2025-10-19.

[45] Survey of gan augmentation (researchgate). https://www.researchgate.net/publication/358845746_Survey_of_Image_Augmentation_Based_on_Generative_Adversarial_Network, 2022. Accessed 2025-10-19.

[46] Gans for medical synthesis (sciety labs). https://labs.sciety.org/articles/by?article_doi=10.20944/preprints202506.1310.v1, 2025. Accessed 2025-10-19.

[47] Gans for medical synthesis (preprints.org). https://www.preprints.org/manuscript/202506.1310/v1, 2025. Accessed 2025-10-19.

[48] Gan use in medical field (medium). https://medium.com/@aysenceliktas/overview-of-gan-use-in-the-medical-field-23cb90adf51d, 2022. Accessed 2025-10-19.

[49] Gan augmentation for x-ray classification (pmc). https://pmc.ncbi.nlm.nih.gov/articles/PMC8607740/, 2021. Accessed 2025-10-19.

[50] Lime vs shap (markovml). https://www.markovml.com/blog/lime-vs-shap, 2023. Accessed 2025-10-19.

[51] Interpreting ml predictions (iiard). https://iiardjournals.org/get/IJCSMT/VOL.%2011%20NO.%208%202025/Interpreting%20Machine%20Learning%2022-49.pdf, 2025. Accessed 2025-10-19.

[52] Self-supervised learning for medical imaging (lightly). https://www.lightly.ai/blog/self-supervised-learning-for-medical-imaging, 2023. Accessed 2025-10-19.

[53] Federated learning for medical image analysis (scilight). https://www.sciltp.com/journals/tai/articles/2508001101, 2024. Accessed 2025-10-19.

[54] When federated learning meets medical image (now publishers). https://www.nowpublishers.com/article/OpenAccessDownload/SIP-20240048, 2024. Accessed 2025-10-19.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[56] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, 2021.

[57] M. T. U. Alam. Efficientnet vs efficientnet+cbam: Attention-enhanced odir-5k classification. https://www.kaggle.com/code/takrimulalam/efficientnet-vs-efficientnet-cbam, 2025. Accessed 2025-10-14.

[58] Vit paper (google research). https://research.google/pubs/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale/, 2021. Accessed 2025-10-19.

[59] Vision transformer (wikipedia). https://en.wikipedia.org/wiki/Vision_transformer, 2025. Accessed 2025-10-19.

[60] Se networks (cvpr 2018). https://openaccess.thecvf.com/content_cvpr_2018/papers/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.pdf, 2018. Accessed 2025-10-19.

[61] Bam (arxiv). https://arxiv.org/pdf/1807.06514, 2018. Accessed 2025-10-19.

[62] Cbam (eccv 2018 openaccess). https://openaccess.thecvf.com/content_ECCV_2018/papers/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.pdf, 2018. Accessed 2025-10-19.

[63] Cbam (arxiv preprint). https://arxiv.org/pdf/1807.06521, 2018. Accessed 2025-10-19.