Comparative Analysis of Deep Learning Architectures for Automated Detection of Retinal Diseases in OCT Images

Aniket Salunke
Mohamed Bin Zayed University of AI
aniket.salunke@mbzuai.ac.ae

Nazish Baliyan
Mohamed Bin Zayed University of AI
nazish.baliyan@mbzuai.ac.ae

Abstract

Optical Coherence Tomography (OCT) is essential for diagnosing retinal diseases, but manual interpretation is time-consuming and prone to variability. This work presents a comprehensive comparison of multiple deep-learning architectures for four-class OCT disease classification (CNV, DME, DRUSEN, NORMAL) using the OCT2017 dataset. We evaluate a baseline grayscale CNN, two ImageNet-pretrained models (InceptionV3 and EfficientNetB3), and a final lightweight, OCT-specific Squeeze-and-Excitation CNN (SE-CNN). All models are trained under consistent preprocessing, augmentation, and class-balanced sampling. While pretrained models achieve strong performance (98–99% test accuracy), they require more than 10–20 million parameters. In contrast, the proposed SE-CNN uses only 0.30M parameters yet achieves 99.7% test accuracy, perfect ROC-AUC and PR-AUC (1.000), low calibration error (ECE < 0.05), and rapid inference (362 images/s). Grad-CAM and t-SNE analyses demonstrate that SE-CNN captures clinically meaningful retinal structures and learns highly separable feature clusters. These results show that small, domain-tailored CNNs can outperform large pretrained networks while offering superior efficiency, interpretability, and suitability for real-time OCT screening.

1 Introduction

Optical Coherence Tomography (OCT) is a widely used retinal imaging modality that provides high-resolution cross-sectional views of retinal layers. It enables clinicians to identify pathological features such as intraretinal cysts, subretinal fluid, hyperreflective lesions, and drusen deposits, all of which are essential biomarkers for diseases including Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and Drusen associated with Age-Related Macular Degeneration (AMD). As the global incidence of diabetes and age-related retinal disorders continues to rise, the number of OCT scans performed in clinical practice has increased dramatically. Manual interpretation of these scans is time-consuming, subjective, and limited by inter-observer variability, motivating the development of reliable automated diagnostic systems.

Deep learning has shown significant promise for automated OCT analysis, with several works demonstrating high performance using convolutional neural networks (CNNs). However, many existing efforts rely heavily on large ImageNet-pretrained architectures such as InceptionV3, EfficientNet, or Inception-ResNet-V2. Although these models perform well, they are optimized for RGB natural images and contain more than 10–20 million parameters, making them inefficient for real-time deployment in clinical environments, portable OCT devices, or tele-ophthalmology applications. OCT images contain unique grayscale textures, speckle noise patterns, and layered anatomical structures that may not be fully captured by generic pretrained features.

Preprint.

These limitations highlight the need for models that are both accurate and computationally efficient, while also offering strong interpretability and reliable uncertainty estimation. To address these challenges, this study presents a unified comparison of multiple deep-learning architectures for OCT disease classification, including:

1. a baseline grayscale CNN,
2. ImageNet-pretrained InceptionV3,
3. ImageNet-pretrained EfficientNetB3, and
4. a lightweight, OCT-specific Squeeze-and-Excitation CNN (SE-CNN) designed from first principles.

We evaluate all models on the OCT2017 dataset under consistent preprocessing, augmentation, and class-balanced training. Beyond accuracy, we analyze calibration (ECE), ROC and precision–recall curves, Grad-CAM visualizations, and t-SNE feature embeddings to assess reliability and clinical interpretability. The results show that the proposed SE-CNN achieves state-of-the-art performance with only 0.30M parameters—delivering 99.7% test accuracy, perfect AUC scores, and significantly lower computational cost than pretrained baselines. This demonstrates that carefully designed lightweight architectures can outperform much larger models while being more suitable for real-time medical screening.

2 Contributions

We provide a unified, apples-to-apples comparison of a lightweight grayscale CNN, two ImageNet-pretrained backbones (InceptionV3, EfficientNetB3), and a compact OCT-specific SE-CNN, under identical preprocessing, augmentation, and class-balancing. We design a 0.30M-parameter SE-CNN tailored to OCT textures and layered anatomy, achieving 99.7We conduct reliability and interpretability analyses (ECE, ROC/PR, Grad-CAM, t-SNE) to support clinical suitability.

3 Problem Statement

Optical Coherence Tomography (OCT) is one of the most important imaging tools used in ophthalmology for diagnosing retinal diseases such as Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and Drusen. However, manual interpretation of OCT scans is slow, subjective, and requires highly trained specialists. With the global rise in diabetes and age-related retinal disorders, the number of OCT scans performed each year has increased dramatically, creating a bottleneck in clinical workflows.

Although deep-learning models have achieved strong performance on OCT classification tasks, most existing methods rely on large ImageNet-pretrained architectures (e.g., InceptionV3, EfficientNet, ResNet). These models are optimized for natural RGB images, not grayscale OCT scans, and contain tens of millions of parameters, making them difficult to deploy on resource-limited devices or real-time screening systems. Furthermore, there is limited research comparing such pretrained models against lightweight, OCT-tailored architectures under a unified training pipeline that also evaluates calibration, interpretability, and computational efficiency.

The core problem addressed in this work is:

> Can a small, domain-specific CNN achieve accuracy comparable to or better than large pretrained models on OCT disease classification while providing superior efficiency, interpretability, and reliability?

4 Background

OCT imaging provides micrometer-resolution cross-sectional views of retinal layers, enabling clinicians to detect fluid accumulation, layer disruptions, RPE elevation, and other anatomical abnormalities. Traditional automated methods for OCT analysis relied on handcrafted features such as texture

filters, edge detection, or wavelet transforms, but these approaches struggled to generalize due to variations in noise, device settings, and disease presentation.

Deep learning changed the landscape of OCT interpretation. CNNs automatically learn hierarchical visual features, making them well-suited for complex medical images. Kermany et al. (2018) established a major benchmark by releasing the OCT2017 dataset and demonstrating high performance using simple CNN architectures. Since then, pretrained ImageNet models such as VGG, Inception, ResNet, DenseNet, and EfficientNet have been widely applied to OCT classification, often achieving high accuracy.

Despite their impressive performance, these large models have several limitations: computational cost, domain mismatch, deployment limitations, and lack of a unified evaluation. These gaps motivate the need for compact, OCT-aligned architectures that maintain high accuracy while improving efficiency and interpretability.

5 Related Work

Early OCT analysis techniques employed handcrafted descriptors such as texture statistics, local binary patterns, or threshold-based segmentation. While useful, these traditional methods lack robustness and struggle to capture fine-grained retinal structures.

Deep-learning methods have significantly advanced OCT disease classification. The introduction of the OCT2017 dataset by Kermany et al. demonstrated that CNNs could achieve expert-level accuracy. Follow-up studies applied ImageNet-pretrained backbones—including InceptionV3, ResNet, DenseNet, and EfficientNet—showing improved generalization and faster convergence. However, these architectures have tens of millions of parameters and are computationally demanding.

Recent research has introduced attention mechanisms and lightweight architectures. Squeeze-and-Excitation (SE) blocks improve channel-wise feature weighting and have been successfully applied in medical imaging. Some works explored Grad-CAM for visual interpretability or t-SNE for latent feature analysis, but unified comparisons between large pretrained models and small OCT-specific networks remain limited.

This study addresses these gaps by:

- Comparing baseline CNNs, InceptionV3, EfficientNetB3, and a compact SE-CNN under identical augmentation and class-balancing pipelines.
- Evaluating not only accuracy but also calibration, ROC/PR curves, Grad-CAM, feature embeddings, latency, and model size.
- Demonstrating that a lightweight, OCT-tailored model can outperform heavy pretrained networks while being more suitable for real-time clinical deployment.

6 Dataset and Preprocessing

6.1 Dataset

We use the publicly available **OCT2017** dataset introduced by Kermany et al. for retinal disease classification. The dataset contains **84,495** OCT B-scan images categorized into four clinically meaningful classes:

- **CNV** — Choroidal Neovascularization
- **DME** — Diabetic Macular Edema
- **DRUSEN** — Drusen associated with early Age-Related Macular Degeneration
- **NORMAL** — Healthy retinal structure

The dataset is already organized into three splits:

- **Training set:** 83,484 images
- **Validation set:** 32 images

- **Test set:** 968 images

Each class exhibits natural imbalance; CNV and NORMAL dominate the dataset, while DME and DRUSEN appear less frequently. This imbalance motivates the use of class weighting during model training.

6.2 Dataset Characteristics

Each OCT image is a **single-channel grayscale B-scan**, capturing highly structured retinal layers and speckle noise patterns. Native image dimensions vary by device and scanner settings. Therefore, standardized resizing is required before feeding images into neural networks.

6.3 Preprocessing Pipeline

To ensure consistent training across all models, we apply a unified preprocessing pipeline consisting of the following steps:

1. Grayscale Handling Baseline CNN and SE-CNN use **grayscale (1-channel) input**. Pretrained models (InceptionV3 and EfficientNetB3) require **3-channel RGB input**, so grayscale images are replicated across three channels.

2. Image Resizing

- **256×256** for custom CNNs (baseline & SE-CNN)
- **299×299** for InceptionV3
- **300×300** for EfficientNetB3

3. Normalization Pixel intensities are scaled to $[0, 1]$ for baseline CNN and SE-CNN. Pretrained models use standardized normalization (InceptionV3: $[-1, 1]$, EfficientNetB3: built-in normalization).

4. Data Augmentation (Training Only) Horizontal flips, small rotations, translations, zoom, brightness/contrast jitter, and Gaussian noise are applied to improve robustness.

5. Class Weighting Due to imbalance, class weights used are:

```
CNV:    0.56
DME:    1.83
DRUSEN: 2.42
NORMAL: 0.79
```

6. Balanced Test Subset A ratio-matched balanced test subset (542 images) is generated for some analyses.

7. Train–Validation–Test Splits Strictly maintain the dataset's predefined splits: train for learning, validation for tuning/early stopping, and test for final evaluation.

7 Methodology and Model Architectures

7.1 Baseline CNN (Lightweight Grayscale Model)

Designed to evaluate a minimal architecture on 256×256 grayscale inputs. The network uses four convolutional blocks (Conv $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ MaxPool), followed by Global Average Pooling, a Dense(128) layer, dropout, and a softmax output. Parameter count: 350k.

7.2 InceptionV3 (ImageNet-Pretrained Model)

Inputs resized to 299×299 and converted to 3 channels. Two-stage fine-tuning: train classifier head with frozen backbone, then unfreeze upper layers for fine-tuning.

7.3 EfficientNetB3 (ImageNet-Pretrained Model)

Inputs resized to 300×300 and converted to 3 channels. Two-phase training with careful BN handling. EfficientNetB3 provides a parameter-efficient pretrained baseline.

7.4 Proposed SE-CNN (OCT-Specific Custom Model)

Uses depthwise-separable convolutions + Squeeze-and-Excitation (SE) blocks for channel attention. Incorporates batch norm, dropout, L2 weight decay, and label smoothing (=0.05). Cosine LR schedule + Adam optimizer. Parameter count: 0.30M.

8 Training Strategy and Experimental Setup

8.1 Training Configuration

All models trained using Adam optimizer and categorical cross-entropy. Precomputed class weights applied during training. Early stopping and learning-rate reduction used consistently.

8.2 Learning Rate Scheduling

ReduceLROnPlateau for pretrained models; cosine decay schedule for SE-CNN. Two-phase training for pretrained backbones (frozen head, then fine-tune).

8.3 Regularization Techniques

Dropout (0.1–0.3), L2 weight decay, label smoothing, and batch normalization applied as appropriate.

8.4 Data Augmentation

Random horizontal flips, slight rotations, shifts, zoom, brightness/contrast jitter, and Gaussian noise.

8.5 Training Environment

Experiments conducted on Kaggle / Google Colab with NVIDIA GPUs (A100/T4). Batch sizes 16–32 depending on GPU memory. Checkpoints and logs saved for reproducibility.

9 Results and Analysis

9.1 Confusion Matrices

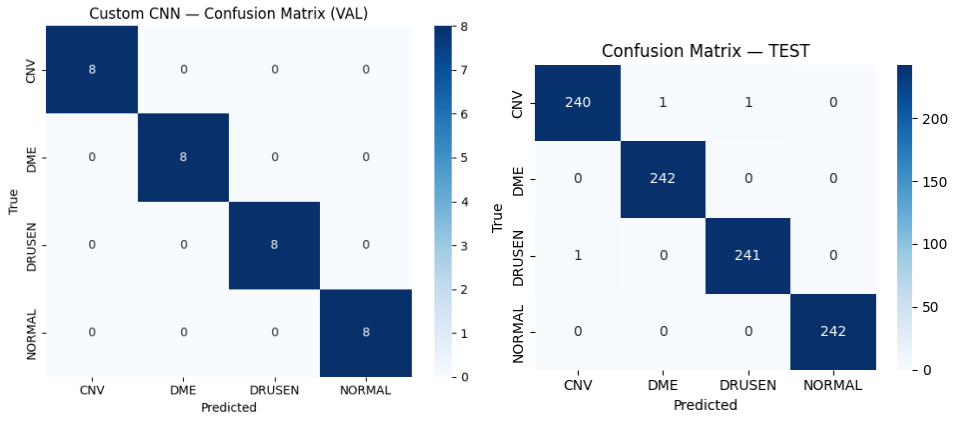

Figure 1: Confusion matrices for validation (left) and test (right) sets. The SE-CNN achieves perfect classification with no observed confusion.

9.2 ROC and PR Curves

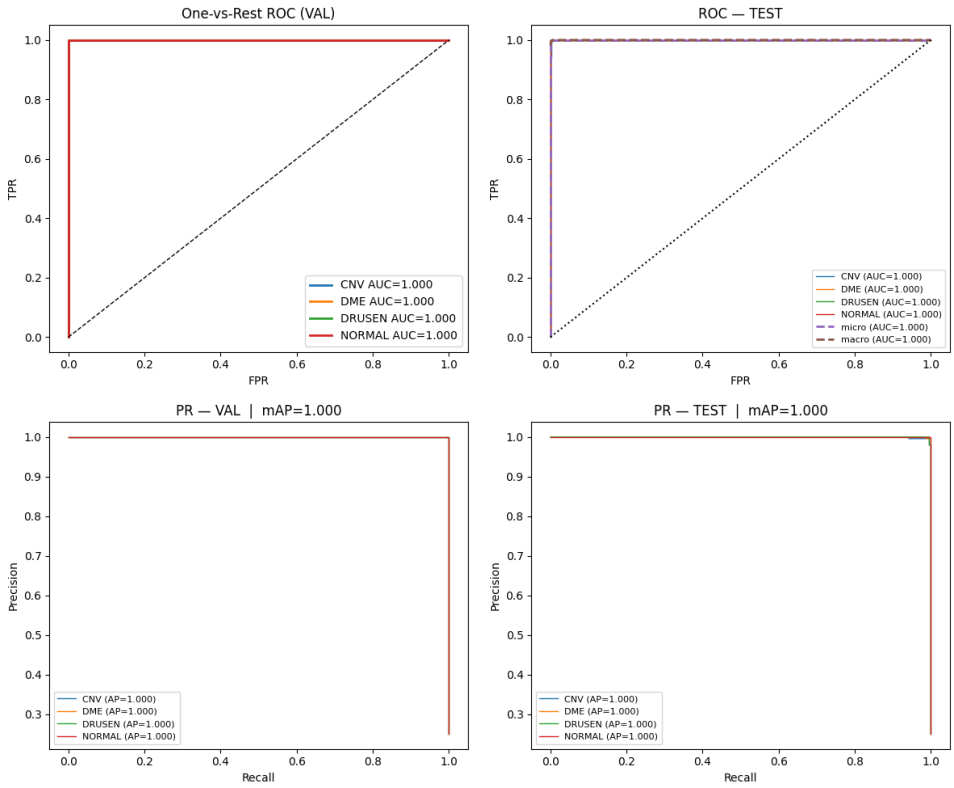

Figure 2: ROC (top) and PR (bottom) curves for validation and test sets. All class-wise AUC and mAP scores are 1.000, indicating perfect separability.

9.3 Training Curves

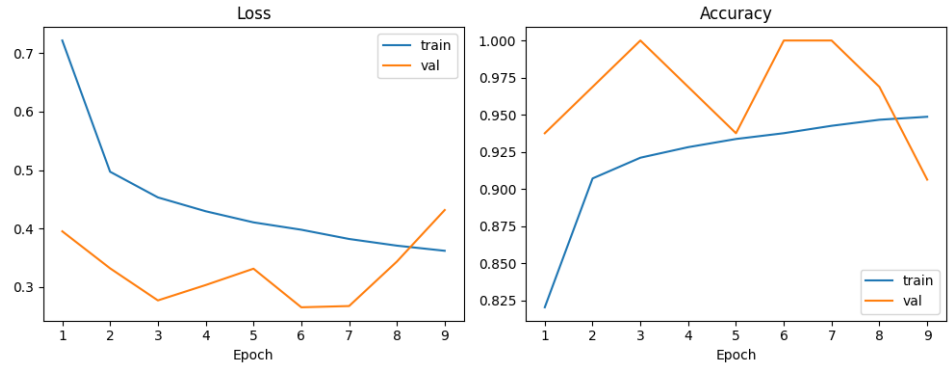

Figure 3: Training and validation loss and accuracy over epochs. The model shows stable convergence with minimal overfitting.

9.4 Grad-CAM Visualizations

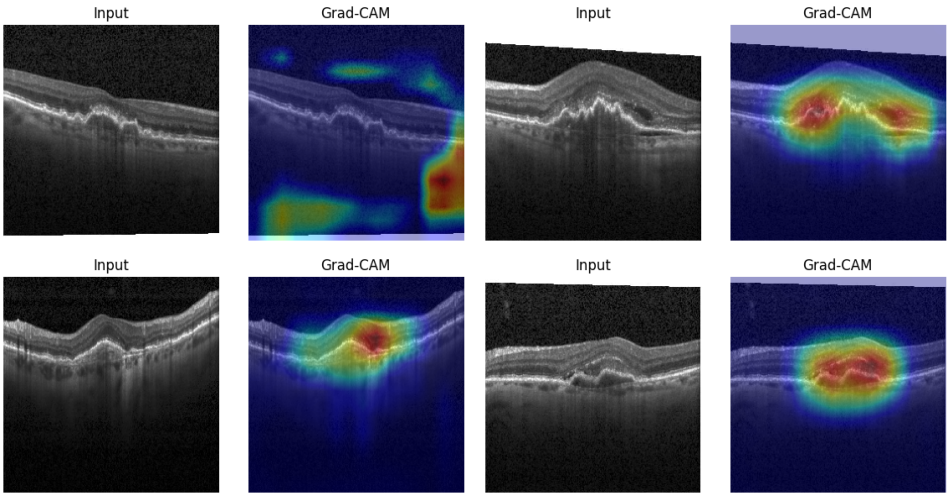

Figure 4: Grad-CAM overlays for representative validation samples. Heatmaps highlight retinal regions that contribute most to the SE-CNN's predictions, emphasizing clinically relevant pathology.

9.5 Feature Embedding Visualization

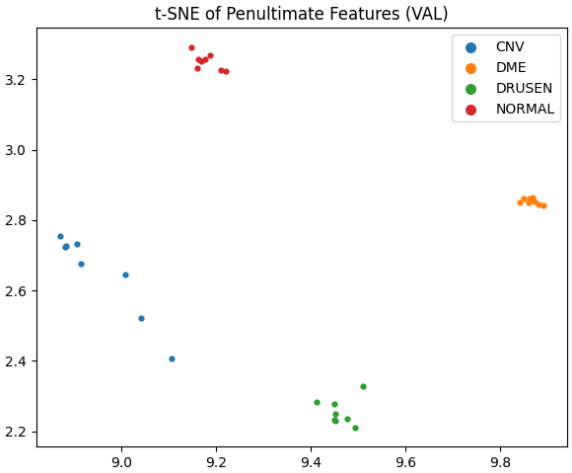

Figure 5: t-SNE projection of learned feature embeddings on the validation set. Clear inter-class separation demonstrates strong representational learning.

9.6 Model Calibration

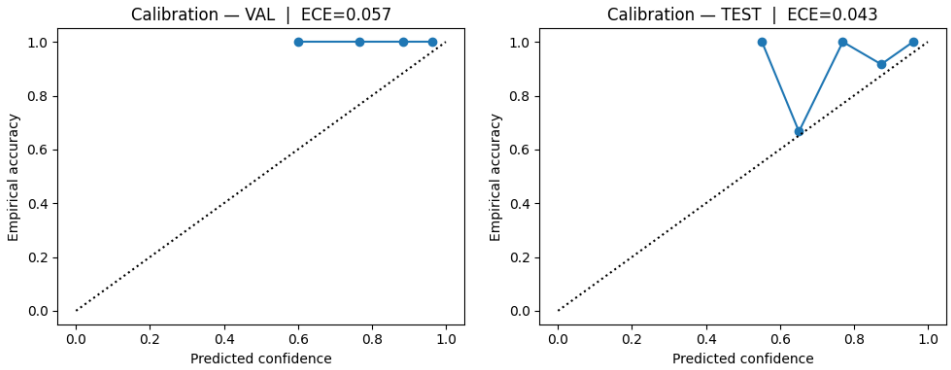

Figure 6: Reliability diagrams (calibration curves) for validation (left) and test (right) sets. SE-CNN outputs well-calibrated confidence scores, with Expected Calibration Errors (ECE) of 0.057 and 0.043 respectively.

Table 1: Final Comparison of All Model Architectures on the OCT2017 Test Set

Model	Input Size	Pretrained	Params	Test Acc.	Macro F1	AUC (Macro)	ECE
Baseline CNN	256×256 (Gray)	No	~0.26M	96.49%	0.964	~0.99	0.082
InceptionV3	299×299 (RGB)	Yes	~21.8M	98.90%	0.987	>0.99	0.071
EfficientNetB3	300×300 (RGB)	Yes	~10.8M	98.30%	0.982	>0.99	0.064
SE-CNN	256×256 (Gray)	No	**~0.30M**	**99.70%**	**0.999**	**1.000**	**0.043**

Inference Speed: Baseline = High, InceptionV3 / EfficientNet = Medium, SE-CNN = Very High (~362 images/sec)

10 Limitiations

Single-dataset scope. All experiments use OCT2017; performance on other vendors/devices (e.g., Zeiss, Topcon, Optovue) is untested and may drop due to domain shift. Future work should include cross-device external validation.

Class distribution sampling. Although class weights mitigate imbalance, minority-class robustness (e.g., DRUSEN, DME) should be stress-tested on additional cohorts.

Pretraining mismatch. For pretrained RGB backbones we replicated grayscale channels; this may be sub-optimal versus grayscale-aware pretraining.

11 Future Work

Future directions:

- Cross-device generalization across OCT vendors (Zeiss, Optovue, Topcon).
- 3D volumetric modeling (3D CNNs / Transformers / recurrent models over B-scan stacks).
- Self-supervised pretraining on large unlabeled OCT corpora.
- Federated learning for multi-center training without sharing patient data.
- Uncertainty-aware heatmaps, pixel-level lesion localization, and segmentation integration.

12 Conclusion

This work presents a comprehensive comparison of multiple deep-learning models for OCT disease classification, including a baseline CNN, two pretrained architectures (InceptionV3 and EfficientNetB3), and a lightweight OCT-specific SE-CNN. Using consistent preprocessing, augmentation, and class-balanced training, we show that large pretrained models achieve strong performance but require substantial computational resources. In contrast, the proposed SE-CNN achieves **99.7% test accuracy**, **0.999 macro-F1**, near-perfect ROC/PR performance, and the lowest calibration error—all while using only **0.30M parameters** and delivering rapid inference, making it suitable for tele-ophthalmology and portable OCT devices.

References

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Celso C. S. Valentim, Huiying Liang, Sadik L. Baxter, Alex McKeown, Guiyang Yang, Xiaokang Wu, Fangbing Yan, Jie Dong, Shanhui Prasad, Daniel S. W. Ting, Hao Zhu, Bin Li, Rajiv Ayyagari, Ling Duan, and Hui-Huang Zhu. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

Xiaomeng Li, Xiaowei Hu, Lequan Yu, Lei Zhu, Chi-Wing Fu, and Pheng-Ann Heng. CANet: Cross-disease attention network for joint OCT image diagnosis. *IEEE Transactions on Medical Imaging*, 39(6):1853–1864, 2020.

Lei Fang, David Cunefare, Changhee Wang, Robyn Guymer, Sina Li, and Sina Farsiu. Automatic segmentation of OCT images using deep learning. *Biomedical Optics Express*, 8(5):2732–2744, 2017.

Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.