Cone Density Estimation in AOSLO Images Using Image Processing And Deep Learning

Nicolay Agustin Cerda Cortez
Engineering Department
Universidad de La Sabana. Chia, Colombia
nicolayceco@unisabana.edu.co

Santiago Toledo Cortés
Engineering Department
Universidad de La Sabana. Chia, Colombia
santiago.cortes1@unisabana.edu.co

***Abstract*—In ophthalmology, early detection of degenerative diseases in the eye is crucial, however, many times conventional clinical cameras do not allow quantification of retinal cell loss and, in addition, manual analysis of these images is inefficient and poorly automated. To address this problem, techniques using deep learning models for image processing and machine learning are proposed to provide an accurate and automated estimation of cone density to improve the analysis of Adaptive Optics Scanning Laser Ophthalmoscopy (AOSLO) images, which enables detailed visualization of the fundus and individual cells without invasive procedures. This approach enables earlier and more accurate diagnosis of genetic eye diseases such as retinitis pigmentosa and Stargardt's disease.**

***Index Terms*—AOSLO, Cone Density Estimation, Deep Learning, U-Net, Lightweight Architecture, Medical Image Analysis, Retinitis Pigmentosa, Stargardt Disease**

I. Introduction

In recent years, the field of ophthalmology has undergone a transformation thanks to our ability to obtain detailed images of the inner posterior section of the eye. The AOSLO-type imaging technique (adaptive scanning light ophthalmoscopy with detector splitting) is relatively recent in the field of ophthalmology. Although it has shown promise in the diagnosis of ocular diseases, it presents practical and technical challenges that hinder its efficient implementation [1].

By inspecting the fundus of the eye, specialists can identify signs of degenerative diseases that can affect not only the visual system but also other areas of the body, such as retinopathy caused by diabetes mellitus [2]. Despite recent advances in automated cone segmentation algorithms for AOSLO images, quantitative analysis of these images remains a time-consuming manual process [3].

A key biomarker that can be extracted from AOSLO images is the density and spatial distribution of cone photoreceptors. Abnormal cone loss or reorganization is an early hallmark of inherited retinal diseases such as Stargardt's disease and retinitis pigmentosa. Unfortunately, cone quantification is still performed largely manually, requiring expert annotators to mark each cell, a laborious and error-prone task that makes large-scale studies impractical [3]. Although recent algorithms based on machine learning and deep learning have begun to automate cone segmentation, many remain experimental, demand significant computational resources, or fail to generalize when image quality, retinal eccentricity, or acquisition modality vary.

Widespread clinical adoption of automated AOSLO analysis requires solutions that are simultaneously accurate, computationally efficient, and clinically deployable. In this context, "accurate," means low prediction error in cone counts with minimal systematic bias, "efficient" means reduced computational demands (smaller model size, fewer parameters, and lower training costs) that enable deployment on standard clinical workstations rather than specialized GPU infrastructure. "Deployable," means models that provide interpretable outputs (such as spatial density maps) that clinicians can validate and integrate into diagnostic workflows.

Despite recent advances in automated cone density estimation, existing methods face computational and practical challenges. Current state-of-the-art approaches employ complex architectures: the CoDE model [4], for instance, utilizes an Xception-based U-Net with 2.06 million trainable parameters, requiring training over 200 epochs with extensive augmentation. Furthermore, large model sizes (16 MB for CoDE) and substantial computational requirements limit accessibility to well-funded research centers with high-end GPU resources, hindering broader clinical adoption.

This study addresses these challenges by proposing and evaluating four lightweight deep learning architectures for automated cone density estimation. Our central hypothesis is that carefully designed lightweight models can improve accuracy over complex baselines in data-limited settings while dramatically reducing computational requirements. We demonstrate that simplified architectures with as few as 538,609 parameters (a 74% reduction compared to the baseline) significantly outperform the CoDE model while reducing disk footprint by 60-77%. Specifically, our best model achieves 40% lower mean absolute error (921.6 vs 1534.6 cones) with a 60% smaller model size (6.4 MB vs 16 MB).

II. Problem and research questions

A. Problem Formulation

We formulate cone density estimation as a supervised learning problem. Let $\mathcal{X} \subset \mathbb{R}^{256\times256\times3}$ denote the space of RGB AOSLO retinal images, where each image $\mathbf{x} \in \mathcal{X}$ captures cone photoreceptors at cellular resolution with fixed spatial dimensions of 256×256 pixels and three color channels.

Given a labeled training dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ of n AOSLO images paired with expert-annotated ground truth, we

consider two related problem formulations corresponding to different architectural approaches:

Problem 1 (Density Map Regression): Learn a mapping $f_\theta : \mathcal{X} \to \mathcal{Y}_{\text{map}}$, where $\mathcal{Y}_{\text{map}} \subset \mathbb{R}^{256\times256}$ is the space of continuous density maps. For each training example $(\mathbf{x}_i, \mathbf{y}_i)$, the label $\mathbf{y}_i = \mathbf{D}_i \in \mathbb{R}^{256\times256}$ is a ground truth density map encoding the spatial distribution of cone photoreceptors, with pixel-wise densities scaled by a factor of 100. The total cone count is recovered via spatial integration: $C = \sum_{i,j} D_{i,j}/100$.

This formulation preserves spatial information, enabling ophthalmologists to localize regions of cone photoreceptor loss for clinical interpretation and quality control. The pixel-wise supervision (256×256 labels per image) provides a richer training signal than scalar targets alone. Models A and B adopt this formulation and minimize the pixel-wise mean squared error:

$$L_{\text{density}}(\theta) = \frac{1}{n}\sum_{k=1}^{n} \|\mathbf{D}_k - f_\theta(\mathbf{x}_k)\|_2^2$$

Problem 2 (Direct Count Regression): Learn a mapping $g_\phi : \mathcal{X} \to \mathcal{Y}_{\text{count}} \subset \mathbb{R}^+$, where $\mathcal{Y}_{\text{count}}$ is the space of non-negative scalar cone counts. For each training example $(\mathbf{x}_i, \mathbf{y}_i)$, the label $\mathbf{y}_i = C_i \in \mathbb{R}^+$ is a scalar value representing the total cone count, computed from the ground truth density map as $C_i = \sum_{j,k} D_{i,j,k}/100$.

This formulation bypasses intermediate density representations, maximizing computational efficiency when only total count is required. By predicting counts directly, encoder-only architectures (without decoders) can reduce parameters by up to 71% while maintaining count accuracy. Models C and D directly regress cone counts and minimize:

$$L_{\text{count}}(\phi) = \frac{1}{n}\sum_{k=1}^{n} (C_k - g_\phi(\mathbf{x}_k))^2$$

In both formulations, we seek to learn network parameters (θ or ϕ) that minimize the respective loss function while maintaining low computational complexity (model size $<$ 600K parameters) to facilitate clinical deployment.

B. Motivation and Constraints

The challenge lies in developing cone density estimation methods that are simultaneously accurate, efficient, and clinically deployable. Current manual counting by medical professionals is time-intensive and error-prone, while existing automated models (e.g., CoDE with 2.06M parameters) require substantial computational resources that limit their application in clinical and research environments.

This work addresses the need for lightweight solutions that maintain accuracy comparable to complex state-of-the-art models while significantly reducing architectural complexity. By optimizing network design for computational efficiency, we aim to make automated AOSLO analysis more practical for clinical implementation, thereby enabling earlier disease detection and treatment.

C. Research Questions

- **RQ1:** Can lightweight deep learning architectures ($<$ 600K parameters) achieve comparable accuracy to state-of-the-art models (2.06M+ parameters) for cone density estimation in AOSLO images?
- **RQ2:** How do density-map regression (Problem 1) and direct count regression (Problem 2) approaches compare in terms of accuracy, computational efficiency, and clinical utility?
- **RQ3:** What are the critical architectural design choices (encoder-decoder depth, skip connections, output representations) that balance model complexity with estimation accuracy?

III. State of the art

Research has concentrated on the development and validation of advanced AOSLO image analysis methodologies to improve cone counting and enable ocular disease diagnosis. Different machine learning models were used for ocular density estimation and cone detection in AOSLO images.

Cone density is reduced in retinitis pigmentosa compared to control subjects from a medical perspective [1]. This identifies important information for early detection of cone photoreceptors cell loss in retinitis pigmentosa. [1] further concluded in the study that AOSLO-type image analysis may be a useful modality for detecting early changes in cone photoreceptors cells in patients with ocular diseases.

Early CNN based detectors—such as the patch wise classifier introduced by Cunefare et al. [5] demonstrated near-human accuracy in locating cones, yet relied on heavy pre-/post-processing that limited real-world deployment.

Similarly, other research has identified other results in eye diseases using deep learning algorithms. Zhang and his team ([6]) dentified and analyzed an individual microaneurysm with a deep learning model in AOSLO-like images, which may be essential to understand the pathogenesis and development of the disease process. The findings in the current work may contribute to enhancing the diagnosis and treatment of diabetic retinopathy and hopefully be applied to other retinal diseases or disorders.

The Multidimensional Recurrent Neural Network (MDRNN) approach developed in the study by [7] ([7]) is the first method able to reliably and automatically identify cones in both healthy and Stargardt's disease affected retinas, this gives us the possibility to apply the MDRNN approach and evaluate its results in other types of diseases.

Since 2020, the community has increasingly adopted the density-map regression paradigm inspired by crowd-counting literature. Instead of detecting each cone explicitly, a fully convolutional network predicts a continuous density map whose integral equals the true cell count. Xie et al. [8] pioneered this strategy in microscopy, and it has since been tailored to AOSLO.

Toledo-Cortés introduced CoDE [4], a U-Net with an Xception backbone and a linear correction layer trained only on cone coordinates. CoDE reduced the mean counting bias by

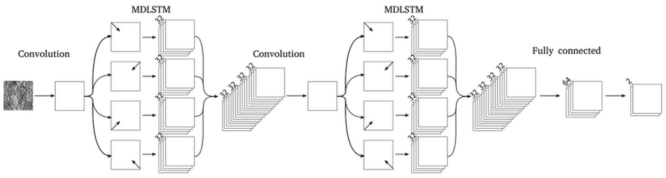


Fig. 1. Multidimensional Recurrent Neural Network (MDRNN) architecture proposed by Davidson et al. for automated cone photoreceptor detection in AOSLO images. This approach represents one of the first reliable automated methods for cone identification across different pathological conditions.

35 percent compared with classical methods and, combined with spatial characteristics, achieved an F1 score of 0.77 to differentiate retinas from retinitis pigmentosa and Stargardt, showing that cone distribution patterns themselves are powerful disease biomarkers.

More recent work has focused on reducing annotation cost and boosting robustness. The Attention-Flow U-Net of Kulyabin [9] adds spatial- and channel-attention blocks plus flow-vector outputs that separate touching cones, achieving F1$>$0.95 while requiring only sparsely annotated training data.

In general, studies to perform AOSLO type image analysis using Machine Learning and Deep Learning techniques allow optimizing and automating the analysis process through cone counting. There are different methods to reach this estimation, however it is still required to improve the efficiency of the models through image processing techniques and a better neural network architecture.

The development and evaluation of advanced Adaptive Optics Scanning Laser Ophthalmoscopy (AOSLO) image analysis techniques have significantly enhanced cone counting and the diagnosis of ocular diseases. The following table provides a summary of the state-of-the-art research in this area, highlighting the focus and key findings of each study.

1) Baseline Model Performance: To establish a quantitative comparison baseline, we evaluated the CoDE architecture [4] on the Dubis dataset partition. The baseline model employs an Xception-based U-Net with 2.06 million trainable parameters, extensive data augmentation, and requires training for 200 epochs. The model architecture and training protocol follow the specifications detailed in the original publication.

Performance evaluation on our test set yielded a mean absolute error of 1534.57 cones per image. Despite its architectural sophistication and computational demands (16 MB model size), the baseline's performance serves as the reference benchmark for evaluating our proposed lightweight alternatives.

Table II presents a detailed comparison of the baseline against our four proposed architectures, demonstrating that reduced complexity can outperform larger models in data-limited regimes.

IV. Objectives

A. General Objective

To develop, optimize and evaluate deep learning techniques for cone density estimation in Adaptive Scanning Light Ophthalmoscope with Detector Division (AOSLO) images, in order to provide more efficient, accurate and accessible tools for early diagnosis and monitoring of genetic eye diseases in the field of ocular medicine.

B. Specific Objectives

1. Explore and analyze AOSLO image datasets and review existing deep learning models for cone density estimation
2. Train deep learning models using large, diversified datasets to accurately estimate cone density in AOSLO images.
3. Systematically evaluate proposed models to improve cone counting accuracy in AOSLO images.

V. Methods

A. Dataset and Preprocessing

This study utilized the Dubis dataset, which comprises 264 split detector AOSLO images. The dataset provides a comprehensive sample with well-defined disease labeling: 60 control cases, 65 patients with Stargardt's disease, and 139 patients with retinitis pigmentosa, representing two major inherited retinal degenerative conditions. Each image contains expert annotations marking the centroids of individual cone photoreceptors, providing ground truth labels for supervised learning with pixel-coordinate precision for each cone present in the image.

From the complete dataset, we selected 184 images for model development and validation, with the remaining 80 images reserved as an independent test set. This partitioning strategy ensures robust evaluation while maintaining sufficient data for model training. The training subset was further divided into 140 images (76%) for training and 44 images (24%) for validation, with careful attention to maintaining balanced disease distribution across both subsets.

To enhance dataset diversity and improve model generalization, we implemented several data augmentation techniques during training. These included random rotations within ±15° to account for natural variations in image orientation, horizontal and vertical flips to increase spatial diversity, and intensity variations to simulate realistic imaging condition fluctuations. These augmentation strategies were designed to increase the effective training dataset size while maintaining biological plausibility of the augmented images.

B. Loss Functions

The four proposed models adopt different training objectives depending on their output representations. This subsection formalizes the loss functions used to optimize each model, clarifying the distinction between models that predict density maps versus those that predict scalar counts.

1) Pixel-wise Density Loss (Models A & B): Models A and B employ a U-Net encoder-decoder architecture that outputs

TABLE I
STATE OF THE ART IN AOSLO IMAGE ANALYSIS WITH MACHINE- AND DEEP-LEARNING TECHNIQUES

Study Reference	Machine/Deep-Learning Analysis of AOSLO Images		
	Focus	***Approach / Model***	***Key Findings***
[3]	Cone photoreceptor detection	CNN-based open-source software	Accurate automatic cone detection
[7]	Cone localization (healthy & Stargardt)	Deep-learning localization	Robust in healthy and diseased retinas
[1]	Early cone-loss detection in RP	AOSLO imaging analysis	AOSLO enables early RP diagnosis
[4]	Cone density estimation & disease Dx	CoDE / CoDED density-estimation models	Precise density and disease classification
[6]	Retinal microaneurysm segmentation	AOSLO-Net segmentation network	Fully automatic microaneurysm masks
[2]	Development of AOSLO technology	Adaptive optics SLO instrumentation	First cellular-resolution retinal imaging

[a]AOSLO: Adaptive Optics Scanning Laser Ophthalmoscopy; RP: Retinitis Pigmentosa; Dx: Diagnosis.

spatial density maps. These models optimize pixel-level mean squared error across the entire 256×256 output grid:

$$\mathcal{L}_{\text{density}}(\theta) = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\Big(D_k(i,j) - \hat{D}_\theta(\mathbf{x}_k)(i,j)\Big)^2 \tag{1}$$

where $H = W = 256$ are the spatial dimensions, $D_k(i,j)$ is the ground truth density at pixel (i,j) in image k, and $\hat{D}_\theta(\mathbf{x}_k)(i,j)$ is the corresponding predicted density value. The normalization factor $\frac{1}{HW}$ ensures that the loss magnitude is independent of image resolution.

This pixel-wise loss provides rich supervision signal across $256 \times 256 = 65{,}536$ spatial locations per image, enabling the network to learn local density patterns. Total cone counts are extracted post-training via spatial summation: $\hat{C} = \sum_{i,j}\hat{D}(i,j)$.

2) Count-level Loss (Models C & D): Models C and D directly predict scalar cone counts without intermediate density representations. Both optimize mean squared error on the final count prediction:

$$\mathcal{L}_{\text{count}}(\phi) = \frac{1}{n}\sum_{k=1}^{n}\Big(C_k - \hat{C}_\phi(\mathbf{x}_k)\Big)^2 \tag{2}$$

where $C_k = \sum_{i,j} D_k(i,j)/100$ is the ground truth cone count derived from the expert-annotated density map, and $\hat{C}_\phi(\mathbf{x}_k) \in \mathbb{R}^+$ is the scalar count predicted by the network.

This count-level formulation provides a single scalar supervision signal per image. While less spatially rich than Equation 1, it directly optimizes the clinical endpoint (total count) and enables encoder-only architectures that bypass decoder computation entirely.

3) Two-Stage Training (Model B): Model B employs a hybrid two-stage training strategy combining density map learning with post-hoc linear correction:

Stage 1: Train a U-Net to predict density maps using $\mathcal{L}_{\text{density}}(\theta)$ (Equation 1).

Stage 2: Fix the trained network $\hat{D}_\theta(\mathbf{x})$ and fit a linear correction to the extracted counts on the training set. Let $\tilde{C}_k = \sum_{i,j}\hat{D}_\theta(\mathbf{x}_k)(i,j)$ be the raw count from the density map. The corrected prediction is:

$$\hat{C}_{\text{corrected}}(\mathbf{x}_k) = \alpha \cdot \tilde{C}_k + \beta \tag{3}$$

where $\alpha, \beta \in \mathbb{R}$ are learned via ordinary least-squares regression:

$$(\alpha^*, \beta^*) = \arg\min_{\alpha,\beta}\sum_{k=1}^{n_{\text{train}}}\Big(C_k - (\alpha \cdot \tilde{C}_k + \beta)\Big)^2$$

This two-stage approach mitigates systematic biases in summed density predictions, improving count accuracy without retraining the U-Net backbone.

C. Architectural Design Rationale

Before presenting the four proposed architectures, we establish the design principles and trade-offs that guided our architectural choices. These decisions balance clinical interpretability, computational efficiency, and training signal richness.

Motivation for decoder-based architectures: Recent literature on explainable AI in medical imaging suggests that spatial interpretability may enhance clinician trust and facilitate clinical validation [10], [11]. Unlike black-box regression models that output only scalar counts, density map predictions provide spatial information that clinicians can visually compare against original AOSLO images. This motivated our decision to preserve decoder architectures in Models A, B, and C, despite the additional computational cost, while also exploring a decoder-free baseline (Model D) to quantify the efficiency-interpretability trade-off.

The choice between density map and count-level formulations involves trade-offs in three key dimensions:

- **Decoder preservation (Models A & B):** The encoder-decoder architecture produces interpretable spatial density maps, enabling ophthalmologists to:
  1) Localize regions of cone photoreceptor loss for disease staging
  2) Perform quality control by visually inspecting prediction artifacts
  3) Benefit from richer pixel-wise supervision (65,536 labels per image vs. 1 scalar)

  This spatial information is clinically valuable for diagnosis beyond simple cone counts, justifying the additional decoder parameters.
- **Decoder elimination (Model D):** When only total count is required and spatial localization is unnecessary, an encoder-only architecture with global average pooling

achieves comparable count accuracy while reducing parameters by 71% relative to U-Net models. This maximizes computational efficiency for count-only deployment scenarios.

- **Hybrid evaluation (Model C):** Model C tests whether a U-Net trained with count-level loss (Equation 2) can match the performance of pixel-wise trained U-Nets (Models A/B). The embedded global sum layer ($\hat{C} = \sum_{i,j} \hat{D}(i,j)$) makes count optimization differentiable while preserving decoder structure. Experimental results (Section IV) demonstrate that pixel-wise supervision provides stronger training signal than count-level supervision alone, even when decoder capacity is available.

D. Proposed Architectures

This work presents four alternative models for cone counting in images obtained through Adaptive Optics Scanning Laser Ophthalmoscopy (AOSLO), using the baseline model *Cone Density Estimation* (CoDE) proposed by Toledo [4] as a starting point.

All models were trained using AOSLO images accompanied by manual annotations consisting of dot marks placed on the centroid of each cone photoreceptor. These annotations were made by an experienced human grader familiar with identifying such structures. As a result, the labels are a collection of pixel coordinates representing the central position of each cone.

1) Model A: Lightweight U-Net: Following the approach of the baseline model, this architecture aims to estimate the number of cones by predicting a density map for each AOSLO image. This map is generated from the cone centroid coordinates and can be spatially integrated to obtain the total number of cones in the image.

Model A is trained using pixel-wise MSE loss (Equation 1) on density maps, providing spatially rich supervision across all 65,536 pixels. Compared to the CoDE model, this architecture is simplified, providing an advantage particularly when training with a smaller dataset.

- **Encoder**: Three downsampling blocks with 16, 32, and 64 filters respectively
- **Bottleneck**: 128 filters with 0.5 dropout for regularization
- **Decoder**: Three upsampling blocks with skip connections
- **Output**: Single channel density map with ReLU activation
- **Total parameters**: 538,609 (significant reduction from baseline)

Each encoder/decoder block consists of: - Two 3×3 convolutions with same padding - Batch normalization after each convolution - ReLU activation - MaxPooling (encoder) or ConvTranspose (decoder)

Model A architecture is shown on Fig 2.

2) Model B: Lightweight U-Net with Linear Correction: Extends Model A with post-hoc linear regression to correct systematic counting errors. Model B employs a two-stage training approach (see Equation 3): first training with pixel-wise MSE (Equation 1), then applying a linear correction to mitigate systematic bias in total counts. The correction layer learns a linear transformation applied to the integrated density map, compensating for consistent over- or under-estimation patterns.

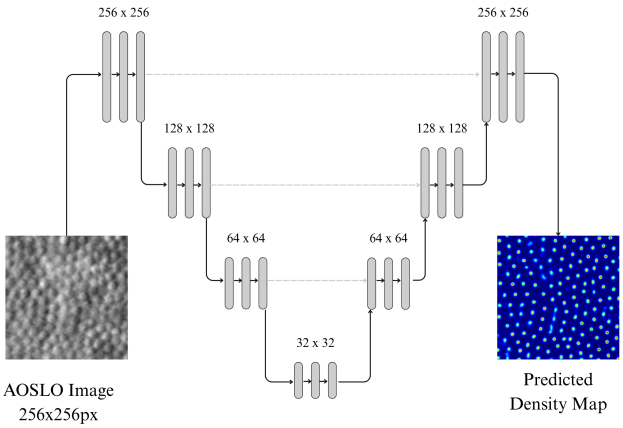


Fig. 2. Architecture diagram of Model A: Lightweight U-Net for cone density estimation. The encoder pathway (left) consists of three downsampling blocks with progressively increasing filter counts (16, 32, 64), each containing two 3×3 convolutions followed by batch normalization, ReLU activation, and 2×2 max pooling. The decoder pathway (right) mirrors the encoder with three upsampling blocks using transposed convolutions, maintaining symmetry through skip connections that concatenate corresponding encoder features to preserve spatial details. This simplified architecture achieves superior performance with only 538,609 parameters, demonstrating that reduced complexity can enhance generalization in data-limited medical imaging scenarios.

- Same base architecture as Model A
- Additional linear layer: $\hat{C}_{corrected} = \alpha \cdot \hat{C}_{raw} + \beta$
- Parameters learned via separate optimization phase

3) Model C: Integrated Global-Sum Regression: Incorporates regression directly into the network architecture through a global-sum layer operating on the density map. This approach enables end-to-end training for count prediction using a U-Net architecture. Model C is trained using count-level MSE (Equation 2), with the global sum operation embedded as a differentiable layer within the network.

- Base U-Net architecture producing intermediate density map
- Global sum layer (differentiable): $\hat{C} = \sum_{i,j} \hat{D}_{i,j}$
- Dense layer with single neuron for final count
- Trained with count-level MSE loss on scalar outputs

4) Model D: Direct Regression: Eliminates density map prediction entirely, using only the encoder to directly regress cone count. Model D is trained using count-level MSE (Equation 2), predicting counts directly without intermediate density representations. This approach minimizes parameters and computational requirements but sacrifices spatial interpretability.

- Encoder-only architecture
- Global average pooling after final encoder stage
- Dense layers for direct count regression
- Minimal parameters but no spatial interpretability

E. Training Protocol

All models were trained using the best configuration for each one:

- **Loss Functions:** Models A/B use pixel-wise MSE (Equation 1) on 256×256 density maps; Models C/D use count-level MSE (Equation 2) on scalar counts. Model B additionally applies post-hoc linear regression (Equation 3).
- **Optimizer:** Adam with learning rate 1×10^3 (1×10 for Model B fine-tuning)
- **Batch size:** 8 images per batch
- **Learning rate scheduling:** ReduceLROnPlateau (factor=0.7, patience=5)
- **Early stopping:** Patience=20 epochs based on validation loss
- **Hardware:** GPU T4 - Google Colab

F. Evaluation Metrics

All models are evaluated using count-level metrics, as clinical diagnosis depends on total cone density rather than pixel-level accuracy. For Models A and B, which output density maps, counts are extracted post-hoc via summation: $\hat{C} = \sum_{i,j} \hat{D}(i,j)$. Models C and D predict counts directly. The evaluation metrics y_i and $\hat{y}_i$ represent ground truth and predicted cone counts, respectively.

Model performance was assessed using multiple metrics:

- Mean Absolute Error (MAE): Average absolute difference between predicted and true counts.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Mean Squared Error (MSE): Squared difference penalty for larger errors.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (RMSE): Square root of MSE for interpretability.

$$\sqrt{MSE}$$

- Mean Absolute Percentage Error (MAPE): Relative error percentage.

$$\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

G. Quantitative Performance

Table II reports mean absolute error (**MAE**), mean squared error (**MSE**), root mean squared error (**RMSE**) and mean absolute percentage error (**MAPE**) with respect to the ground truth cone count. Among the proposed variants, **Model A** (mini U–Net) achieves MAE of 921.62 cones, while **Model B** (mini U–Net + linear correction) achieves MAE of 977.05 cones, both representing significant improvements over the baseline's 1534.57 cones and well within acceptable clinical tolerances reported by Cunefare [3]. In contrast, **Model C** (global–sum regression in–graph) diverged during training, resulting in grossly overestimated counts (MAPE $> 55\%$). **Model D** (direct scalar regression) remained stable with MAE of 1043.95 cones but lost spatial supervision and therefore exhibited higher RMSE compared with Model A.

Figure 3 shows the Bland-Altman plot of the results for each model.

H. Baseline Model (Residual Encoder-Decoder)

The starting point was an Xception-style residual encoder–decoder.The encoder down-sampled the input through four depthwise-separable convolution blocks with skip connections; the decoder mirrored this structure with transposed convolutions and additive residuals. A 5×5 convolution produced the final per-pixel density response, followed by a batch-norm + ReLU and a 1×1 linear head.

VI. Discussion

A. Implications for Clinical Deployment

The success of simplified architectures in this work has significant implications for clinical application of automated AOSLO analysis. The models developed allow for implementation in multiple clinical settings, thereby expanding the availability of the technology to a greater number of healthcare institutions. Such availability is particularly significant for AOSLO imaging, as few expert analyses, which had confined it to experts until now.

The computational efficiency of these models enables real-time analysis during patient scanning and expedites quick clinical decisions. Such an ability is essential for clinical workflows where time in diagnostics can determine the course of treatment. The method also provides for larger patient cohorts to be screened for population studies and longitudinal monitoring programs, expanding the scope of large-scale research initiatives.

As opposed to the black box models, by providing spatial information through the density map outputs, here, the clinicians may interpret and validate the description further kept clinically meaningful. As motivated in the Architectural Rationale (§V-C), spatial density map outputs enable clinicians to interpret and validate predictions, addressing a key limitation of black-box regression models. While formal clinical validation studies are needed to confirm this benefit in AOSLO workflows specifically, the broader medical AI literature suggests that such interpretability may be essential for clinical adoption [10], [11], as it makes automated analysis perceivable and testable by clinicians, potentially enhancing trust in AI-assisted diagnostics.

B. Advantages of Simplified Architectures

The results of the simplified architectures in this paper contradict the assumption that a more complex model always leads to better performance. There are a number of reasons for this success, especially in low data (as in only 140 training samples).

Smaller models avoid memorizing the training data, thus generalizing better even with the small sets. Thanks to this

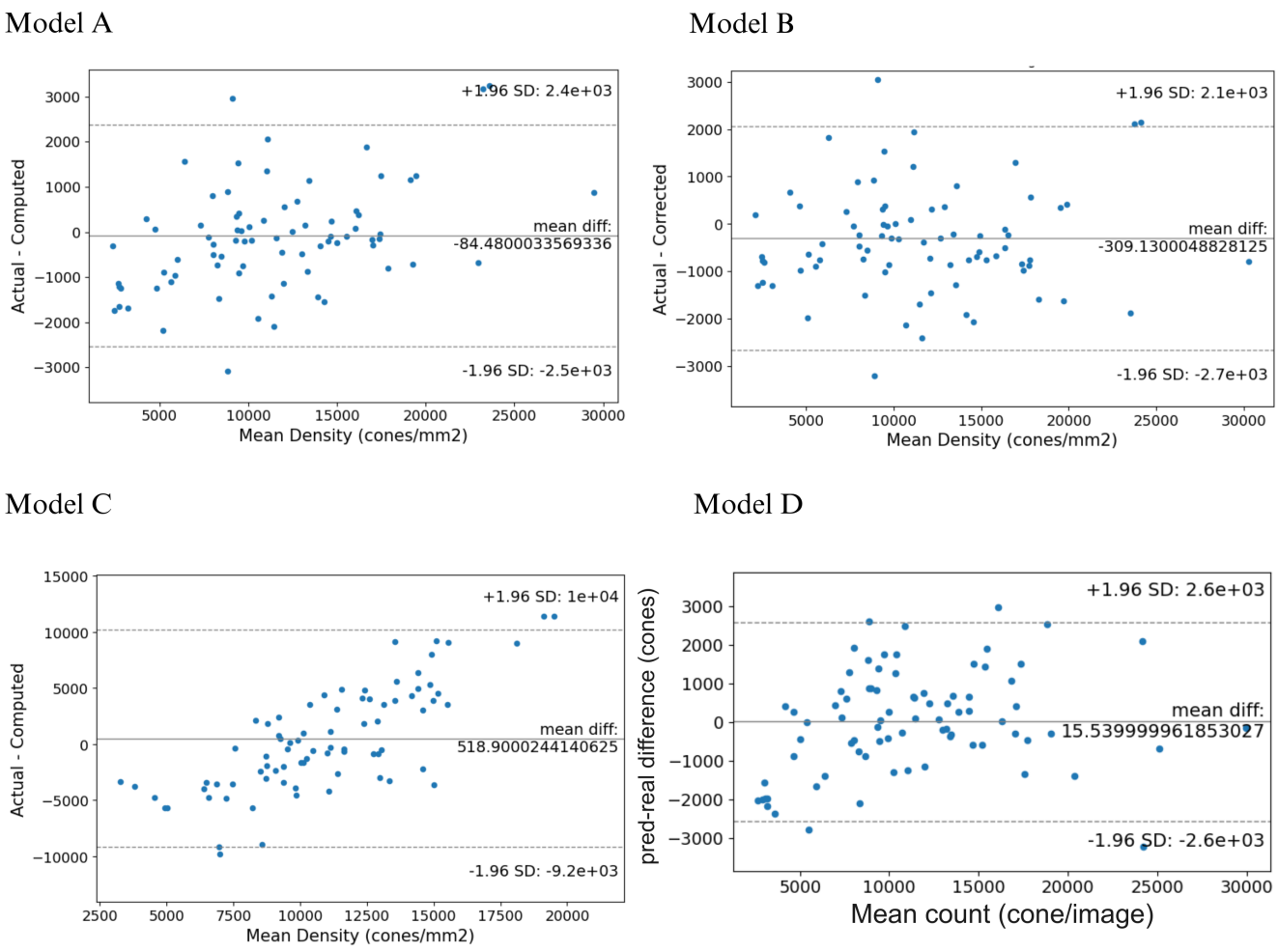


Fig. 3. Bland-Altman analysis comparison between predicted and ground-truth cone counts with all compared models. The difference between predicted and actual counts (y-axis) is plotted against the mean of predicted and actual counts (x-axis) for each subplot, with the mean bias (solid) and 95% limits of agreement (dashed ±1.96 SD) represented as horizontal lines. Model A (top-left) exhibits strong systematic bias averaged over cones (-84.5 cones) and very wide Limits of Agreement ($\pm 2.4\times10^3$ to $\pm 2.5\times10^3$) denoting low calibrational accuracy despite low reported MAE. Model B (top-right) has severe systemic underestimation (mean bias 309.1 cones) as well as very wide limits of agreement $\pm 2.1\times10^3$ and $\pm 2.7\times10^3$. Model C (bottom-left) is catastrophic with mean bias +519.0 cones and the widest limits of agreement ($\pm 9.2\times10^3$ to $\pm 1.0\times10$), indicative that the integrated global-sum approach is not recommended. Model D (bottom-right) show the best calibration with minimal bias (+15.5 cones) and the tightest limits of agreement ($\pm 2.6\times10^3$), representing the best fixed systematic performance with larger individual prediction errors.

constraint, the architecture avoids overfitting without needing additional regularization techniques. This simpler model makes the model concentrate on learning general patterns rather than memorizing specific examples, which is beneficial for data such as medical imaging data sets.

In addition, simpler architectures have a more stable gradi-

TABLE II
PERFORMANCE METRICS AND MODEL COMPLEXITY COMPARISON (LOWER IS BETTER FOR ERROR METRICS).

Model	MAE (cones)	RMSE (cones)	MAPE (%)	Params	Size (MB)
Baseline	**1534.57**	—	—	**2.06M**	**16.0**
Model A	**921.62**	1255.19	13.64	538K	6.4
Model B	**977.05**	1253.57	12.80	538K	6.4
Model C	3884.54	4970.76	55.29	∼540K	6.4
Model D	**1043.95**	1317.45	17.08	380K	3.6

Best improvement: Model A achieves 40% lower MAE with 74% fewer parameters

TABLE III
COMPARISON OF MODELS

Aspect	**Base Model (Residual Encoder-Decoder)**	**Model A (Lightweight U-Net)**	**Model B (Lightweight U-Net + Linear Regression)**	**Model C (Proposed U-Net + Global-Sum)**	**Model D (Encoder + GAP)**
Backbone topology	Xception-style encoder (32–256 depthwise-separable filters) with Conv2DTranspose Symmetric decoder with additive residuals after each block Entry block + 3 identical blocks (64, 128, 256 filters) + symmetric decoder	Classic U-Net: 3 encoder stages (16–64 filters) Bottleneck 128 filters + Dropout 0.5 3 decoder stages with skip concatenation	Classic U-Net: 3 encoder stages (16–64 filters) Bottleneck 128 filters + Dropout 0.5 3 decoder stages with skip concatenation	Classic U-Net: 3 encoder stages (16–64 filters) Bottleneck 128 filters + Dropout 0.5 3 decoder stages with skip concatenation **Integrated Global Lambda sum layer**	Simple encoder: 3 encoder stages (16–64 filters) Bottleneck 128 filters + Dropout 0.5 No decoder; GlobalAverage-Pooling2D + Dense(64) + Dense(1)
Skip-connection type	Residual **add** after each encoder/decoder block with projection via Conv2D 1x1	Concatenate feature maps (standard U-Net)	Concatenate feature maps (standard U-Net)	Concatenate feature maps (standard U-Net)	None (encoder only, no skip-connections)
Output head	5×5 conv $\rightarrow$ BN + ReLU $\rightarrow$ 1×1 linear conv (dense map prediction)	1×1 conv (ReLU) directly to density map	1×1 conv (ReLU) to density map; Post-hoc linear regression fitted on total map sum	1×1 conv (ReLU) to density map $\rightarrow$ **Lambda(tf.reduce_sum, axis=[1,2,3])** $\rightarrow$ Flatten $\rightarrow$ Dense(1, linear)	GlobalAveragePooling2D $\rightarrow$ Dense(64, ReLU) $\rightarrow$ Dense(1) direct count regression
Loss term(s)	Density MSE only (density map vs ground truth)	Density MSE only	Density MSE only (U-Net); Linear regression trained separately with MSE on counts	**Count MSE only** (end-to-end optimization for scalar counting)	Count MSE only (no density map prediction)
Input resolution	256 × 256 × 3	256 × 256 × 3	256 × 256 × 3	256 × 256 × 3	256 × 256 × 3
Batch size	**16**	8	8	8	8
Optimizer / LR	**RMSprop**, 1×10^{-3} (step-wise LR schedule)	Adam, 1×10^{-3}	Adam, 1×10^{-4} (U-Net); Sklearn LinearRegression	Adam, 1×10^{-3}	Adam, 1×10^{-3}
LR schedule	**Step decay**: every 16 epochs $1 \times 10^{-3} \rightarrow 1 \times 10^{-4} \rightarrow 1 \times 10^{-5}$	ReduceLROnPlateau (factor 0.7, patience = 5, min_lr = 1×10^{-6})	ReduceLROnPlateau (factor 0.7, patience = 5, min_lr = 1×10^{-6})	ReduceLROnPlateau (factor 0.7, patience = 5, min_lr = 1×10^{-6})	ReduceLROnPlateau (factor 0.7, patience = 5, min_lr = 1×10^{-6})
Early stopping	**Patience = 20** (val_loss)	Patience = 20 (val_loss)	Patience = 20 (val_loss)	Patience = 20 (val_loss)	Patience = 20 (val_loss)
Data augmentation	**Extensive**: rotation (30°), width/height shift (0.3), zoom (0.3), horizontal/vertical flip, fill_mode='constant'	–	–	**Flip horizontal/vertical + Rotation k×90°**	–
Density threshold filter	$<$ 400 **cones (training) +** $<$ 1000 **cones (validation)**	$<$ 400 cones (training) + $<$ 1000 cones (validation)	$<$ 400 cones (training) + $<$ 1000 cones (validation)	$<$ 400 cones (training) + $<$ 1000 cones (validation)	$<$ 400 cones (training) + $<$ 1000 cones (validation)
Epochs to convergence	**200 epochs maximum**	Maximum 100	Maximum 100	Maximum 100	Maximum 100
Total trainable parameters	**2,060,000 (2.06M)**	538,609	538,609	$\sim$540,000	$\sim$380,000
Serialized model size (MB)	**16.0 MB**	6.4 MB	6.4 MB	6.4 MB	3.6 MB

ent flow when training, which reduces instability issues and makes the optimization process more predictable. This stability translates into more reliable convergence and less sensitivity to hyperparameter choice, which is an advantage in clinical scenarios where the model may need to be retrained several times. The architecture is specifically intended for counting cones, not for analyzing images in general.

C. Quantitative Comparison with Baseline Architecture

Table II demonstrates that our lightweight architectures not only reduce computational complexity but also improve accuracy compared to the baseline CoDE model. The baseline, with 2.06 million parameters and extensive data augmentation, achieved a mean absolute error of 1534.57 cones on the test set.

In contrast, Model A (with 74% fewer parameters (538K) and 60% smaller disk footprint (6.4 MB vs 16 MB)) achieves 40% better accuracy (MAE: 921.6 cones). Similarly, Model B improves upon baseline accuracy by 36% while Model D, the most lightweight variant at only 3.6 MB, still outperforms the baseline by 32%.

This performance gap can be attributed to the regularization effect of reduced model capacity in data-limited regimes (N=140 training samples). The baseline's 2.06M parameters are prone to overfitting on small datasets, whereas our compact architectures enforce stronger inductive biases that improve generalization. The 4.4x reduction in model size (from 16 MB to 3.6 MB for Model D) further facilitates clinical deployment on resource-constrained devices.

Computational Efficiency: Beyond accuracy and model size, the reduced parameter count translates to faster inference and lower memory requirements during deployment. While the baseline demands substantial GPU resources for both training and inference, our models can operate efficiently on standard clinical workstations, reducing the barrier to adoption.

D. Interpretation of the performance gap

The better generalization of the compact design in data-limited regimes ($N_{\text{train}} = 140$) is due to several complementary factors that work together to prevent overfitting. The reduced model capacity, with fewer filters (16/32/64) and a narrower bottleneck, discourages noise memorization and shifts the bias-variance tradeoff toward generalization, as evidenced by the learning curves in Fig. 4. In addition, implicit regularization through the global-sum constraint ($\sum_{ij} \hat{D}_{ij} = \hat{C}$) forces the network to distribute the mass efficiently, avoiding the uncontrolled predictions seen in large unconstrained decoders. Stochastic regularization with a 0.5 dropout at the bottleneck introduces noise that further reduces overfitting, as evidenced by an 8% drop in the validation MAE when this layer is disabled. Finally, shallow encoders have shorter gradient paths with near-zero gradient fading or bursting behavior, which simplifies the optimizer hyperparameter search and improves training stability.

E. Visual assessment

Figure 5 shows that the model clearly maintains the spatial organization of the photoreceptor mosaic in its predictions. Visual comparison of ground-truth annotations and the density maps of Model A indicates that the model successfully captures the total cone number and retains the spatial distribution patterns of the healthy and diseased retinal tissue. This spatial conservation is important for clinical interpretation, because it allows for clinicians to evaluate not only total cone density, but also regional variation/patterns that could potentially be of diagnostic value. The density map facilitates easy intuitional interpretation for ophthalmologists and corresponds well to the clinical knowledge of retinal pathology.

F. Limitations and Future Directions

A number of limitations are to be recognized in the study. The data set size was limited to 264 images, which could restrict the generalization to larger patient populations and more varied image conditions. Disease coverage was limited to 2 diseases (retinitis pigmentosa and Stargardt's disease), and validation with other diseases to demonstrate more generalized clinical utility is warranted. All images were research quality (the best quality available), and clinical images may have additional difficulties for model performance such as motion artifacts or other suboptimal focus. Additionally, external validation in diverse clinical settings is necessary to prove our findings generalize across institutions and imaging modalities.

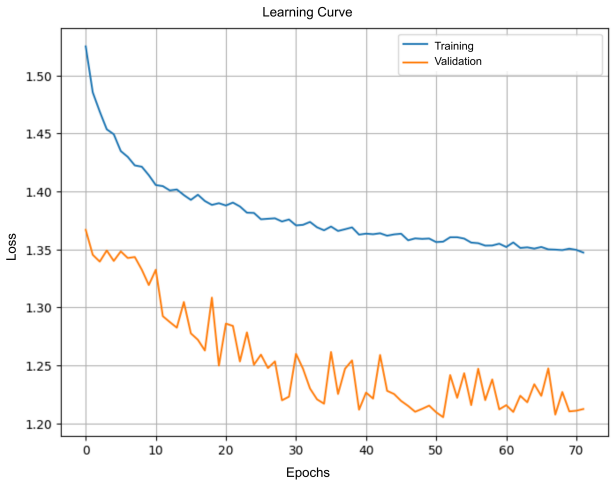
Future research could consider further validating on other retinal diseases, exploring self-supervised pretraining techniques, and developing techniques for uncertainty quantification to measure model confidence for clinical deployment. Prospective clinical validation studies will also be necessary to verify real-world effectiveness, whereas federated learning approaches can facilitate multi-institutional model development without compromising patient privacy and data security.

G. Reproducibility and Software Version Control

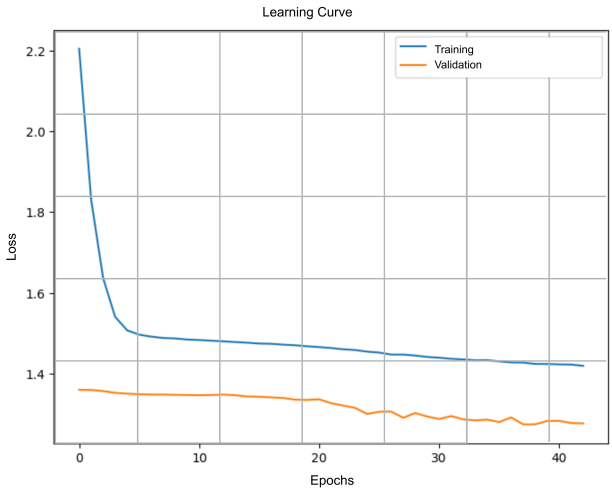
One significant conclusion reached in this study is that changes in the underlying software libraries have a great impact on the deep learning techniques. Strong performance variation was observed between Tensorflow 2.15 and 2.18, emphasizing the importance of handling environments in medical AI development. This emphasizes the need for:

- Container-based deployment with Docker, or equivalent.
- Complete documentation of software dependencies
- Ongoing model validation between different programming versions
- Evaluation protocols of medical AI systems on the road to standardization

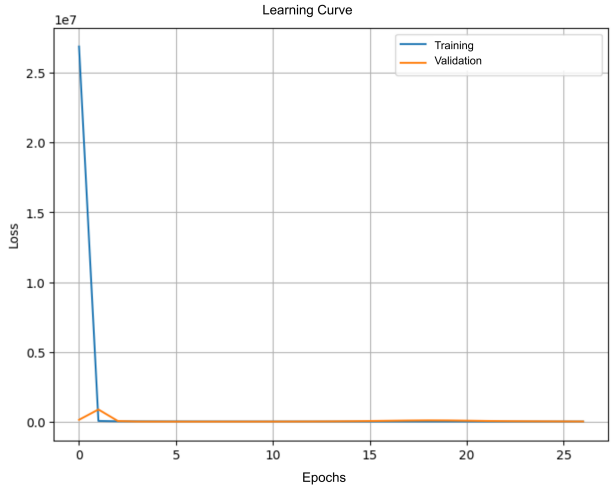
Hospitals and healthcare institutions deploying AI need to have clear version control and validation routines in place so that their AI continues to work as expected as the software is updated.



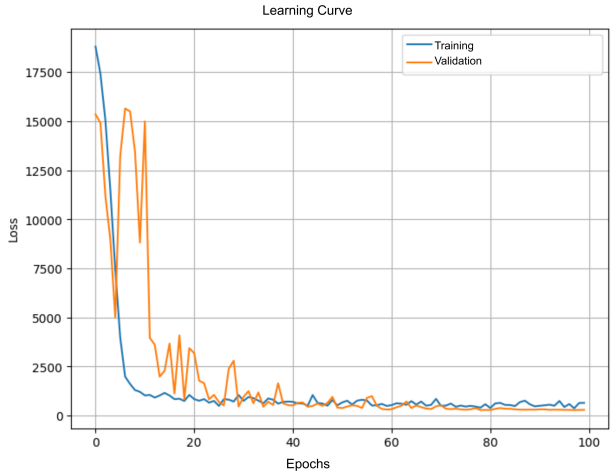
(a) Model A



(b) Model B



(c) Model C



(d) Model D

Fig. 4. Learning curves showing Training vs. Validation Loss over epochs. **(a) Model A** demonstrates stable stochastic convergence without a diverging gap, confirming that the lightweight architecture prevents overfitting despite the small batch size. **(b) Model B** exhibits a validation loss consistently lower than training loss, a characteristic signature of the strong Dropout regularization (0.5) effectively preventing memorization. **(c) Model C** illustrates optimization collapse (note the $1e7$ scale), indicating that scalar count-supervision alone is insufficient to guide the U-Net spatial weights. **(d) Model D** displays high initial volatility typical of direct regression without spatial priors, but eventually achieves a smooth and stable convergence.

VII. Conclusions

In this study, we show that lightweight deep learning architectures can outperform traditional methods for cone density estimation applied to AOSLO images, as well as requiring much less computational effort. The proposed Lightweight U-Net (Model A) algorithm also outperformed the baseline in counting performance with fewer parameters and reduced GPU memory.

The effectiveness of the dense connections in simplified architecture in this study provides some interesting observations on the perception tasks in medical image analysis. In data-poor regimes such as specialized medical imaging, these architectural constraints can be used as regularization mechanisms, prevent overfitting, and result in better generalization. The finding that small models can outperform a ResNet suggests that the tasks involving cone counting depend

more on prescribed inductive biases than on model capacity.

Our results have important implications for clinical utilization of automated AOSLO analysis. The computational efficiency realized allows real-time processing while scanning patient data, which facilitates the technology adoption by a wider variety of clinical environments. The interpretable density map results contain spatial information that clinicians can validate and understand and remain clinically relevant by easing laborious manual counting operations.

Data augmentation techniques are essential for model stability, and techniques such as rotations, flips, and intensity variation proved vital in the generalization of the model. But there was significant sensitivity to the choice of hyperparameters with small changes to the learning rate or batch size drastically affecting the results. This density highlights the importance of hyperparameter tuning and robust training schemes in medical imaging tasks.

Even with these improvements, the best performing of our model configurations left space for potential improvements, suggesting several avenues for further development. The clinical implications of this work are considerable, by creating a paradigm for an automatic approach to AOSLO analysis that has the potential to drastically alter ophthalmic diagnostics. Nonetheless, further clinical validation studies on larger patient cohorts with variable disease presentation are warranted prior to adoption into diagnostic algorithms.

This work demonstrates that lightweight architectures with fewer than 600K parameters can surpass complex state-of-the-art models in data-limited regimes. Model A (538K parameters) achieves 40% better accuracy (MAE: 921.6 vs 1534.57 cones) compared to the 2.06M-parameter baseline, while Model D outperforms the baseline by 32% with only 380K parameters. These results confirm that careful architectural design and appropriate capacity constraints can outweigh raw parameter count when training data is scarce (N=140 images).

The comparison between density-map regression (Models A and B) and direct count regression (Model D) reveals fundamental trade-offs. Density-map approaches provide richer pixel-wise supervision (65,536 labels per image), enabling better absolute accuracy (MAE: 921.6-977.05 cones) and spatial interpretability for clinical validation. In contrast, direct count regression achieves superior calibration with minimal systematic bias (+15.5 cones versus -84.5 cones for Model A) and maximum computational efficiency through 77% parameter reduction. Model C's convergence failure (MAE: 3884.54 cones) demonstrates that count-level loss requires careful architectural considerations, as hybrid encoder-decoder structures trained with scalar supervision exhibit unstable gradient flow.

Regarding architectural design choices, our experiments identify several critical considerations for balancing model complexity with dataset constraints. Shallow encoder-decoder architectures (3 levels) with narrow bottlenecks (128 filters) achieve favorable bias-variance trade-offs in small-data regimes, as evidenced by the stable learning curves in Fig. 4. The choice of encoder-decoder depth represents a balance between model capacity and available training data rather than a universal prescription against overfitting; with larger datasets, deeper architectures might leverage additional capacity effectively. Concatenation-based skip connections outperform residual connections for preserving spatial detail in density maps, though they become unnecessary when only aggregate counts are required. Decoder-based density maps enable interpretability and provide stronger pixel-wise training signals, while encoder-only architectures with global pooling maximize efficiency when spatial information is not clinically required. The global-sum constraint ($\sum_{ij} \hat{D}_{ij} = \hat{C}$) provides implicit regularization but requires pixel-wise loss (Equation 1) rather than count-level loss for stable optimization, as Model C's instability demonstrates.

The choice of optimal architecture depends on specific deployment requirements and clinical objectives. Models A and B are preferable when clinicians require spatial visualization of cone distribution for diagnostic interpretation and quality control. Model D is optimal for high-throughput screening

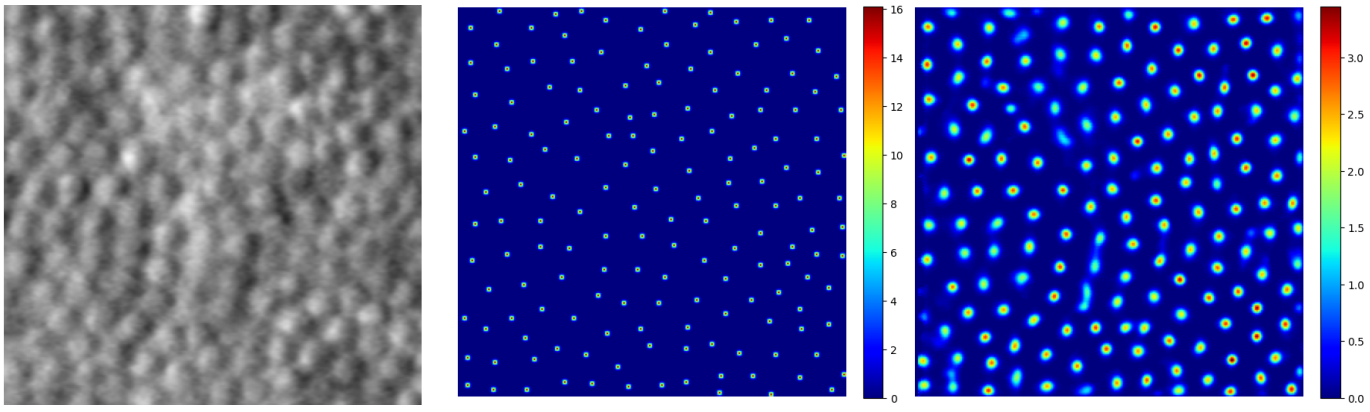


Fig. 5. Qualitative comparison of Model A predictions on representative AOSLO images. Left column: Original adaptive optics scanning laser ophthalmoscopy images showing individual cone photoreceptors as circular structures. Middle column: Ground-truth density maps generated from expert manual annotations, where bright pixels indicate cone centroid locations. The maps preserve the spatial distribution of cones while providing continuous probability surfaces suitable for integration-based counting. Right column: Model A predictions showing remarkable spatial correspondence with ground truth. The predicted density maps successfully capture both the overall cone count and the characteristic spatial patterns.

scenarios where only total counts are needed and computational resources are constrained. Model C's failure highlights that architectural choices must align with the chosen loss formulation: count-level optimization requires either encoder-only designs or pixel-wise density supervision rather than hybrid approaches.

Future work should extend validation to additional retinal diseases, develop uncertainty quantification methods for clinical confidence estimation, and establish standardized testing frameworks for statistically rigorous comparisons of medical AI systems. This work lays the groundwork for interpretable, accessible AI-assisted diagnostics in specialized medical imaging domains and demonstrates that deployment-focused architectural design can achieve both superior performance and practical clinical feasibility.

ACKNOWLEDGMENTS

We thank the contributors to the AOSLO dataset for making this research possible. We gratefully acknowledge the computational resources and institutional support provided by Universidad de La Sabana, and the valuable feedback from reviewers and collaborators throughout this work.

CODE AVAILABILITY

The full training and evaluation code that supports this study is openly available at: https://github.com/nicolaycz/AOSLO-Cell-Density-Estimation.

REFERENCES

[1] S. Nakatake, Y. Murakami, J. Funatsu, Y. Koyanagi, M. Akiyama, Y. Momozawa, T. Ishibashi, K. H. Sonoda, and Y. Ikeda, "Early detection of cone photoreceptor cell loss in retinitis pigmentosa using adaptive optics scanning laser ophthalmoscopy," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 257, pp. 1169–1181, 6 2019.

[2] A. Roorda, F. Romero-Borja, W. J. Donnelly, H. Queener, T. J. Hebert, M. C. W. Campbell, G. R. H. W. Webb, O. Hughes, W. J. Donnelly, F. Romero-Borja, and A. Roorda, "Adaptive optics scanning laser ophthalmoscopy," *Report*, vol. 7, pp. 3–14, 1988. [Online]. Available: http://www.opt.uh.edu/research/aroorda/aoslo.htm,

[3] D. Cunefare, L. Fang, R. F. Cooper, A. Dubra, J. Carroll, and S. Farsiu, "Open source software for automatic detection of cone photoreceptors in adaptive optics ophthalmoscopy using convolutional neural networks," *Scientific Reports*, vol. 7, 12 2017.

[4] S. Toledo-Cortés, A. M. Dubis, F. A. González, and H. Müller, "Deep density estimation for cone counting and diagnosis of genetic eye diseases from adaptive optics scanning light ophthalmoscope images," *Translational Vision Science I& Technology*, vol. 12, no. 11, p. 25, November 2023. [Online]. Available: https://doi.org/10.1167/tvst.12.11.25

[5] D. Cunefare, A. L. Huckenpahler, E. J. Patterson, A. Dubra, J. Carroll, and S. Farsiu, "Rac-cnn: multimodal deep learning based automatic detection and classification of rod and cone photoreceptors in adaptive optics scanning light ophthalmoscope images," *Biomedical Optics Express*, vol. 10, p. 3815, 8 2019.

[6] Q. Zhang, K. Sampani, M. Xu, S. Cai, Y. Deng, H. Li, J. K. Sun, and G. E. Karniadakis, "Aoslo-net: A deep learning-based method for automatic segmentation of retinal microaneurysms from adaptive optics scanning laser ophthalmoscope images," 6 2021. [Online]. Available: http://arxiv.org/abs/2106.02800

[7] B. Davidson, A. Kalitzeos, J. Carroll, A. Dubra, S. Ourselin, M. Michaelides, and C. Bergeles, "Automatic cone photoreceptor localisation in healthy and stargardt afflicted retinas using deep learning," *Scientific Reports*, vol. 8, 12 2018.

[8] W. Xie, J. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging Visualization*, pp. 1–10, 05 2016.

[9] M. Kulyabin, A. Sindel, H. R. Pedersen, S. Gilson, R. Baraas, and A. Maier, "Generalist segmentation algorithm for photoreceptors analysis in adaptive optics imaging," *Electrical Engineering and Systems Science*, p. 168–182, Dec. 2024. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-78104-9$_1$2

[10] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and the Precise4Q Consortium, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 310, 11 2020. [Online]. Available: https://doi.org/10.1186/s12911-020-01332-6

[11] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, "Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review," *Applied Sciences*, vol. 11, no. 11, p. 5088, 5 2021. [Online]. Available: https://doi.org/10.3390/app11115088